

Composed Image Retrieval For Visual Localization: Evaluation For Architectural Contents

Emile Blettery, Valérie Gouet-Brunet, Livio de Luca

▶ **To cite this version:**

Emile Blettery, Valérie Gouet-Brunet, Livio de Luca. Composed Image Retrieval For Visual Localization: Evaluation For Architectural Contents. 7th Workshop on AnalySis, Understanding and ProMotion of HeritAge Contents (SUMAC '25), ACM Multimedia 2025, Oct 2025, Dublin, Ireland. pp.31-40, 10.1145/3746273.3760200 . hal-05335228

HAL Id: hal-05335228
https://hal.science/hal-05335228v1

Submitted on 28 Oct 2025

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

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



HAL Authorization

Composed Image Retrieval For Visual Localization: Evaluation For Architectural Contents

Emile Blettery
UPR CNRS 2002 MAP ; LASTIG, Univ. Gustave Eiffel, IGN-ENSG
Marseille, Champs-sur-Marne, France
emile.blettery@univ-paris1.fr

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

Livio De Luca
UPR CNRS 2002 MAP
Marseille, France
livio.deluca@map.cnrs.fr

Abstract

This article investigates the problem of visual localization with image retrieval in the context of heritage structuring and documentation, leveraging an extensive image dataset collected during the restoration of Notre-Dame de Paris. To address the challenges of image retrieval for localization, we present a retrieval pipeline, called CIR4Loc, based on the composed image retrieval (CIR) paradigm, introducing textual modifiers that refine retrieval towards configurations more suited for localization. By bridging the gap between visual and spatial retrieval, this approach ensures the selection of images that are both visually relevant and spatially distributed to improve localization, and more precisely, camera pose estimation. We demonstrate the effectiveness of this proposal in a real-world heritage context, specifically the scientific site related to the restoration of Notre-Dame de Paris, emphasizing the necessity of retrieval strategies explicitly tailored for spatially aware localization.

CCS Concepts

• **Information systems** → **Top-k retrieval in databases**; **Image search**; **Retrieval effectiveness**; • **Applied computing** → **Arts and humanities**.

Keywords

Composed image retrieval, Image retrieval, Localization, Digital humanities

ACM Reference Format:

Emile Blettery, Valérie Gouet-Brunet, and Livio De Luca. 2025. Composed Image Retrieval For Visual Localization: Evaluation For Architectural Contents. In *Proceedings of the 7th International Workshop on analySis, Understanding and proMotion of heritAge Contents (SUMAC '25), October 27–28, 2025, Dublin, Ireland.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3746273.3760200

1 Introduction

Adding location information to content becomes more and more common as it helps to retrieve, visualize in spatial context and understand them better. Images illustrating geographical areas are no exception to this rule, as most cameras now embed GPS coordinates in the images' metadata. Accessing image collections with some location information becomes easier, although location quality may vary or may not exist when considering older contents such as iconographic heritage. Localizing images, with the most precise information, leads to many potential applications including, for instance, modeling for urban planning, immersive visualization of the past for humanities, data integration in a digital twin or territory evolution observation for ecological studies. However, for applications requiring precise pose estimation, *i.e.* the estimation of the position and orientation of the camera, localization must go beyond simple geotagging.

Paper positioning. When an initial location is not known, localization methods often involve retrieving localized images similar in content to the unlocalized one, propagating their location to the query image, following various approaches from reasoning up to registration. Thus, a key challenge in image localization lies in this retrieval process, which serves as the initial step for many localization pipelines. Image retrieval, commonly optimized for visual similarity, does not always align with the needs of localization, where the viewpoint diversity and spatial distribution of the images manipulated are crucial. Although traditional retrieval methods have been extensively studied [12, 39], their role in localization deserves more attention. Furthermore, this work investigates retrieval-based localization in the specific and challenging context of heritage documentation, where spatial redundancy, visual repetition, morphological complexity, and complex acquisition patterns complicate traditional retrieval methods. This context explains why we have chosen to adapt the Composed Image Retrieval (CIR) paradigm, with the objective of reformulating retrieval with text modifiers in order to improve it for geolocalization. The objective is to retrieve images that are not only visually similar but also spatially distributed around the query image to enable better geolocalization.

As localization methods, we mainly focus on the camera position and orientation (pose) estimation, exploiting solely localized reference images and no other type of data, neither preexisting nor built using reference images. The approaches of this category have been extensively used, studied, and improved, but not in the application domain targeted here. Using such a pipeline serves as a basis to demonstrate our premise by improving an already tried-and-tested approach.

Contextualizing Image Spatial Integration within the Notre-Dame Project. The approach presented in this paper emerges from

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
SUMAC '25, Dublin, Ireland
© 2025 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 979-8-4007-2055-0/2025/10
https://doi.org/10.1145/3746273.3760200

a concrete need identified within a large-scale, multidisciplinary research effort dedicated to the scientific documentation and analysis of the Notre-Dame de Paris restoration process following the 2019 fire. One of the central pillars of this initiative is the creation of a comprehensive, structured, and spatially coherent corpus of visual data documenting both the monument and the restoration activities over time. Manual geotagging or metadata-based methods are often insufficient or unreliable, particularly in dense architectural environments such as Notre-Dame, where visual redundancy, symmetry, and repetitive patterns make spatial disambiguation a challenging task. Yet, accurate localization of each image — both in position and orientation — is crucial for multiple downstream applications: 3D reality-based models refinement, diachronic comparisons, automatic image-to-annotation linking, and spatial querying across the corpus. The visual corpus involved includes thousands of heterogeneous photographs captured by various actors (researchers, restorers, surveyors, photographers) using a wide range of devices and acquisition protocols. While some of these images are part of systematic photogrammetric campaigns, many others are isolated, context-specific photographs produced spontaneously in the field. In this context, visual localization becomes a key enabler of corpus interoperability and reusability. By automating the spatial integration of new images, researchers and heritage professionals can explore, analyze, and correlate visual data in a geometrically meaningful way, without requiring systematic reprocessing or re-annotation.

Contributions. They can be summarized as:

- A global overview of state-of-the-art approaches for image localization, as well as for image retrieval and its variant composed image retrieval (CIR),
- The evaluation of computer vision well-known pipelines to the service of image localization in an architectural image dataset,
- A novel proposal, CIR4Loc, exploiting a CIR perspective to guide the retrieval to ultimately improve localization.

The paper is classically organized: Section 2 revisits the state on the art of the approaches studied, while the dataset exploited, ND-dataset, is presented in Section 3, followed by the presentation of our proposal, CIR4Loc, and evaluating it in Sections 4 and 5, before concluding.

2 Related work

After defining the different types of location available for images in Section 2.1, this section briefly presents the state of the art on localization from images in Section 2.2, insisting on the solutions relying on image retrieval (revisited in Section 2.3), before addressing the literature on composed image retrieval in Section 2.4, which is at the core of our contribution.

2.1 Location information

The location information available or which can be estimated for an image can greatly vary between use cases and datasets. It may first refer to finding information of geolocation either of the content imaged or of the sensor at the origin of the image and second, can take several forms:

- A **textual annotation**, providing an information of geolocation with toponyms (department, city, name of a monument, etc.), coming from 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**, relative or absolute. Such information on images is natively provided by national mapping agencies, recent cameras equipped with GPS or geocoding techniques from toponyms, and can also be estimated by vision-based computational tools.
- A **6-DoF pose** (*i.e.* the 3D position and 3D 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.

Our selection. We focus on the last category, with the idea of targeting the composed image retrieval proposal for pose estimation in the context of architectural heritage documentation.

2.2 Visual localization

Localizing images can be performed in many ways, as investigated in [7] for example. In this work, however, we focus on automatic camera pose estimation, that is, estimating the 6-DoF pose of the query image camera. A brief overview of different approaches for pose estimation is presented. First, many approaches start with image retrieval (revisited later in Section 2.3) to find initial solutions before estimating the query's pose more precisely, based on the most similar images retrieved (Section 2.2.1) and/or by exploiting 3D available information (Section 2.2.2). Finally, trained, all-in-one approaches requiring nothing but the images, have also been developed to output a pose (Section 2.2.3).

2.2.1 Image retrieval-based approaches. The first group of methods exploits only a dataset of localized images as reference and computes a pose for the query after a single step of retrieval by content within this dataset. Here, the exploitation of the responses in the estimation varies from simple assignments up to more complex processes: pose assignment, where the pose of the first retrieved image is assigned to the query [6, 73], with evident advantages and drawbacks; pose averaging, where the poses of the first responses are averaged, with or without weighting schemes [53] [45] ; and finally pose estimation, where the poses of the most similar images are geometrically combined through more or less sophisticated triangulation processes taken in the Computer Vision and Photogrammetry domains [20] [59] [27] [79] [36].

Our selection. In this work, we turn our attention to these last solutions which conduce to a precise estimation of the pose at large scale - only with images - and provide more details about the best state-of-the-art approaches to employ in Section 5.1 dedicated to the implementations experimented and selected for our objectives.

2.2.2 3D-based approaches. A second, major group of approaches for localization are those exploiting 3D information in order to compute the pose using a spatial resection approach [21]. A first step of retrieval is performed to identify reference images similar to the

query. Those returned images are registered with the 3D data (they are sometimes used to create it), by determining 2D-3D matches between the query and 3D points. From these matches, a geometric Perspective-n-Point solver (PnP) is used to estimate the 6 degrees of freedom of the pose. In terms of PnP solver, multiple adaptations have been proposed to deal with more or less complex cases. A major difference between them is the previous knowledge (or not) of the camera's calibration (its intrinsic and extrinsic parameters) [26] [30] [18] [52].

Other approaches exploit an available 3D point cloud, either created on-the-fly by Structure-from-Motion (SfM) from the first retrieved images [54], or already existing (LiDAR points for instance). Then, 2D-2D-3D matches between the query image and the local point cloud can be found by passing through the retrieved images and the related 3D. Then, again a geometric PnP solver is used to estimate the precise pose [45].

A final group of approaches attempts to compute direct 2D-3D matches between the scene and the query, without passing through retrieved images in a 2D-2D-3D fashion.[9] proposes a network extracting and matching directly 2D and 3D features, using the Sinkhorn algorithm to estimate the most likely 2D-3D matches between the query image and the 3D scene. In [64], direct 2D-3D matches are estimated in a pyramidal fashion from coarse to fine by minimizing a volume displacement cost when aligning 2D and 3D points. [41] proposes Desc-Matcher, to directly extract and match 2D and 3D points between the query and the 3D scene, using a classification-based approach to select matches in a coarse to fine fashion. With DGC-CNN, [66] exploits color and geometric cues to directly estimate the 2D-3D matches, without using any 2D and 3D descriptors.

2.2.3 Trained, all-in-one approaches. A final family of localization approaches regroups those that need nothing but the images (localized or not) and directly output the pose of the image. Those approaches rely on trained networks and can sometimes be generalized to any scene, but not always. Out of this paradigm of approaches, Relative or Absolute Pose Regression methods (RPR/APR) have gained traction lately. The principle is for a network to learn from images (localized or not) the geometric representation of the scene. Thus, when a query image is given as input, the output is directly a pose, either absolute or relative to other images used to train the model. The main issue with such approaches is the fact that for APR, the network is trained solely for one scene and not generalizable to another. This is less true for RPR methods but they do not always generalize well, especially if the visual setting between scenes is too diverse. [53] presents a thorough state of the art of such approaches as they emerged, but the approaches are numerous and evolve regularly [70] [11] [51] [40]. Currently, we consider that they do not fully generalize well and are not scalable facing the size of the search area, while retrieval-based solutions (with 2D or 3D) still remain better on this last point.

A last, promising, type of approach is multi-task (often led by the multi-view reconstruction task objective), that performs multiple geometric 3D vision tasks at once. Currently, the best performing approach is the DUSt3R network [67], further improved by MASt3R [31] that takes as input an arbitrary image collection (no pose, no intrinsics) and performs simultaneously camera calibration, camera pose estimation, and dense 3D reconstruction to only cite some tasks; it was recently generalized within the DUNE framework [50]. Unifying all those vision tasks proves to be beneficial for all of them at the same time.

2.3 Image retrieval

As seen in the previous section, many of the visual-based localization techniques, relying on 2D or 3D data, start with a step of content-based retrieval to find initial locations in the scene. We briefly revisit this large research area here. Image retrieval commonly consists of two successive steps. First, the image description, used to compare two images in order to evaluate their visual similarity. Then, a search step that efficiently retrieves images most similar to a query one. Here, we focus on image descriptors, referring readers to [43] for a synthesis of search approaches in large databases.

In terms of recent image descriptors, most deep learning-based retrieval methods exploit network backbones originally designed for classification tasks, adapted for image retrieval. Notable architectures include VGG [56], ResNet [22], and ResNest [74]. The extracted features from these backbones are subsequently processed to form the final image descriptors. There exist **global descriptors**, which capture the overall image content, obtained through pooling operations, such as SPoC [2], GeM [47], CVNet [29]. Alternatively, **local descriptors** focus on salient regions within the image, highlighted by attention mechanisms, and can be aggregated and compared with local descriptors from other images, as seen in DeLF [42] and How [62]. The aggregation can be achieved through various methods, including the visual bag-of-words paradigm [13, 57], and more recently the efficient ASMK [61, 62] which proposes aggregated selective match kernels for image retrieval. More recent advancements exploit **vision transformers** as feature extractors, such as [68] [16] [32] [23]. Large multimodal models like CLIP [48] exploit very large vision transformers to efficiently describe complex images, helped in this by adaptations like the SPARO attention layer [63].

After retrieving the most similar images, a re-ranking step is usually added in order to refine the results using a different and more sophisticated similarity score. It can be operated according to various strategies, such as geometric verification [10, 29, 71], late fusion of responses [65, 72, 75], learning-based re-ranking with transformers [28, 78, 80], query expansion which refines descriptors from the initial retrieved list [2, 19, 34, 77], diffusion where the similarities are spread in the graph of neighbors [3, 14, 24, 44, 55, 76] and combinations of these approaches [8].

Our selection. In this work, several retrieval approaches were experimented to prepare the evaluation of the proposal for localization (Section 5): the image descriptor How [62] associated with ASMK [61, 62] proved to be the best combination for efficient retrieval on our dataset ; detector and descriptor SuperPoint [15] and LightGlue [35] also confirm their efficiency for post-retrieval matching dedicated to re-ranking or pose estimation.

2.4 Composed Image Retrieval

Composed Image Retrieval (CIR) aims at retrieving images based on an initial query image and textual information (or modifier). A

common use case is product suggestion; for instance, if the user selects the image of a blue dress whose shape they like but rather wants it in red, the query image would be the blue dress, and the modifier could be text like "The same dress in the color red". The CIR process should then return images of red dresses similar to the blue one, if existing in the dataset. Figure 1 illustrates the generic process of Composed Image Retrieval.



Figure 1: Overview of the Composed Image Retrieval process. Figure from [5].

For an extensive review on this topic, the reader may refer to survey [58]. In this short overview, we have selected methods relevant to our objective, which are detailed below.

Two main types of approaches exist: first, those like the CLIP4CIR proposal [5]. Here, the text input is taken as a modifier to apply to the visual descriptor before retrieval. A combiner network is leveraged to alter the input image's visual descriptor based on the text input. This altered visual descriptor is then used as a query for image retrieval in the dataset. This is the approach that inspired our CIR4Loc proposal. Second, approaches like RS4CIR [46] where both textual and visual features are combined but with a weighting scheme that could allow for either pure monomodal retrieval (either visual or textual). Multiple other methods have been proposed lately, of which a short overview is provided:

Feature modifying approaches. A first group of approaches, similar to [5] focuses on the textual feature to perform the composition. [60] only exploits text-to-image retrieval to get images, first describing the query image as text, then modifying this text, differentiating between global description and local requirements, increasing the specificity of the retrieval. [1] follows the same principle, by training a textual inversion network to represent the query image as text. [25] also exploits an image-to-text encoder to then simply modify the textual representation. However, they train the network to differentiate between the main visual element to keep during retrieval using masks on the image. With HyCIR, [25] trains the network in a contrastive fashion to ensure the coherence between the image and text features that describe the main query object. [4] keeps using an image-to-text conversion before modifying the text but shows that using a complete sentence as image description increases overall retrieval performance.

Composition-based approaches. Differently, [37] works using both image-to-text and text-to-image encoders to train the network in a bi-directional fashion, ensuring that it focuses better on what to keep and what to modify. [38] performs CIR in a two-step approach, first filtering the obvious incompatible images and then ranking the more similar images using finer textual information. [17] exploits at the same time image-to-image, image-to-text, and text-to-text retrieval in a successive fashion to modify the retrieval. However, text modifiers are labels rather than complex sentences, similarly to [46].

Generation-based approaches. Finally, [33] adopts a different strategy by using a generative approach on the textual modifier to create a novel image, supposedly as desired. The features of this novel image and the query's are then fused (averaging is a common choice) to obtain the features used for the final image-to-image retrieval.

3 Notre-Dame image dataset

This section presents the application context of the proposal, by focusing on the presentation of the dataset exploited. It stems from a major interdisciplinary digital documentation effort around the restoration of Notre-Dame de Paris, where systematic and high-resolution photographic coverage has created a uniquely structured and spatially redundant image corpus. The images used to create a dataset all come from a global photographic acquisition after the fire of 2019 and all along the restoration process. Images were obtained in a quite systematic process, in a grid-like fashion, in order to cover most surfaces and details of the cathedral. In this work, the spotlight is set on photographs depicting the outside of the cathedral (see examples in Figure 2).



Figure 2: Samples of the ND-dataset.

The dataset exploited here, afterwards called ND-Dataset, consists of **10,901** images quite harmoniously distributed around the architectural building, with visual overlap, either due to the grid acquisition or in a medium/long range fashion as depending on the angle, different parts of the building can be seen in one image.

With such a dataset, several challenges appear when considering image retrieval: the global visual resemblance of all the image contents due to the architectural harmony of the monument, as well as the fact that the building is by construction very symmetrical, which means repeated patterns that can be easily mistaken for one another during the retrieval process. Added to all this is the fact that current description approaches are based on learning strategies relying on general objects or landmarks training datasets, not optimal for this type of content. These difficulties, in turn, impact retrieval-based localization processes, which may be detrimental for the heritage documentation process [69]. Indeed, when it comes

to heritage preservation and documentation, many applications require a localization associated with the images, for example, for 3D reconstruction, immersive visualization or multimedia documents integration in a 3D digital twin.

4 Proposal: CIR4Loc

As revisited in Section 2.2, image retrieval proves to be an adequate base for image localization via pose estimation. However, it is our belief that image retrieval is not necessarily suited for pose estimation. Indeed, for pure visual retrieval only, an ideal image retrieved by visual similarity should be very similar in terms of content. Thus, visual-based retrieval will favor high overlap between images if available in the dataset. However, for pose estimation purposes, such a configuration is not ideal; it could lead to collinear disposition and prevent the pose estimation process from performing correctly. Thus, we propose to explore a composed image retrieval strategy to guide the retrieval step towards a more suited configuration of images for pose estimation. Indeed, an ideal configuration of images for pose estimation requires images spatially distributed around the image to localize. The focus of this work is thus set mainly on ensuring such retrieval, with images flanking the query image on all sides. The CIR4Loc pipeline is illustrated in Figure 3 and detailed in the three following sections.

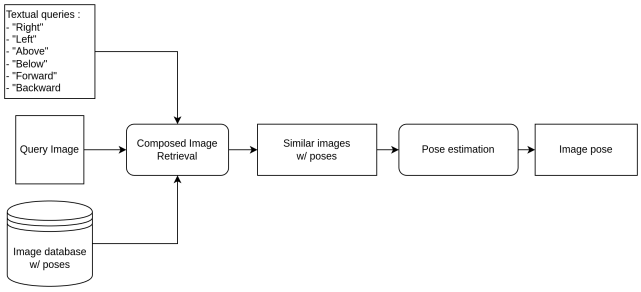


Figure 3: CIR4Loc: the CIR pipeline up to pose estimation.

4.1 Spatial modifiers

We have chosen to build our proposal from CLIP4CIR [5] revisited in Section 2.4, which gathers properties suited for our objectives. We propose an adaptation of this method dedicated to localization, called CIR4Loc. To do this, we define twelve different textual modifiers associated with spatial relative and absolute positions around an image location:

- Six relative positions: Above, Below, Left, Right, Forward, Backward. They are set with regards to the pose of the query image;
- Six absolute positions: Higher, Lower, Northward, Southward, Westward, Forward. They are set with regards to the poses reference system.

4.2 Training

Training models involved in CIR approaches, and then in CIR4Loc, require the creation of a specific dataset. Such dataset is created following several steps leveraging retrieval and 3D reconstruction methods as presented after.

Ground truth. First, an important step that must be highlighted is the creation of the ground truth, that confirms that several images depict the same scene. This is usually essential to estimate the quality of the retrieval in terms of instance retrieval and thus that it can be used to classify objects or estimate a localization for instance. In our case, it helps in finding pairs to be fed to the trained network. To create this ground truth, a global reconstruction of the scene (the cathedral) has first been performed with SfM (Structure from Motion) using Colmap [54], guided by the poses of the ND-dataset images (Section 3). Two images are then considered similar if a sufficient amount of 3D points of the reconstruction are in common between them. This threshold was set to 100 in our experiments, as a good trade-off, limiting both a too high number of "false positives" (images with only a few points in common) and a too high number of images with no similar images (never sharing enough points with any image). Then, a first step of retrieval has been performed, using the efficient combination How + ASMK descriptor (see Section 2.3), and the first 20 retrieved images were manually checked to ensure that the ground truth was as correct as possible.

Training set creation. In order to create the actual training set for CIR4Loc, for each image, the first 20 visually retrieved images are chosen. If they are considered a correct match, based on the previously established ground truth, their distance to the initial image is evaluated. Based on their distances following each direction of the modifiers presented in Section 4.1, the triplet {initial image, retrieved image, move} is created if the translation is said to be above 3 meters. The absolute translation is then computed, and the modifier is set as, for instance:

["Relative move: Right " , "Absolute move: Westward"]

Models training. The CLIP4CIR model is then modified and exploited to train models using the proposed modifiers, on the dataset ND-dataset. The choice was made to ensure that each model focuses on a specific move and its counterpart, rather than potentially combining moves which could lead to a fuzzier understanding of the modifiers. This induces three trained models, each able to process two opposing modifiers. Thus, each trained model focuses on one type of translation: either laterally (Left/Right), vertically (Up/Down), or longitudinally (Forward/Backward). These six models lead to six types of visual descriptors which allow six types of retrieval driven by these directions. The absolute move associated with the relative one is determined with regard to the pose of the image.

4.3 Retrieval

At test time, one query image conduces to six types of retrieval results, each following one relative move textual modifier. The first retrieved images in each "modified" retrieval setting are taken and added to the set of images exploited for the pose estimation process, with the objective of providing a set of similar images depicting the same area and sufficiently spatially distributed to satisfy the pose estimation process. This process of set creation is performed

in a naive way. The method iterates through the retrieval lists at random. For each list, the first image (in decreasing similarity order) that has not yet been included in the final set is added to it. This ensures that each new addition maintains some level of similarity while avoiding doubles. Although this method is relatively naive, it is designed to produce a diverse set of images by drawing from lists with opposing modifiers, making it likely that the final selection consists of images taken from different viewpoints.

Visual examples of the retrieval textually modified are shown in Figure 4, facing a traditional image retrieval. One can see how the retrieval is oriented with respect to the textual modification required. The retrieval results focus, in each case, on one side of the image rather than on the other one, and both lead to different images, displaying a larger variety of viewpoints than the initial non-modified retrieval of the first line. The whole pipeline CIR4Loc is evaluated in depth in Section 5.

5 Experiments and evaluation

We first present the framework of evaluation before the experiments and their results.

5.1 Evaluation framework

This section presents the necessary elements for understanding the evaluation of the proposed approach.

Localization process. Once similar images (with their poses) of a query image are retrieved, with or without the CIR4Loc strategy, the localization of the query is performed following the image retrieval + pose estimation paradigm introduced in Section 2.2.1. More precisely, we chose the following characteristics corresponding to state-of-the-art up-to-date solutions:

(1) The detection and matching of keypoints between images (the query and one similar image) is done with the couple SuperPoint [15] + LightGlue [35], well-known as efficient for post-retrieval matching for re-ranking or pose estimation;
(2) Relative pose estimation (that is between the query and one similar image in an arbitrary reference frame) is performed using MicMac, an open-source SfM software [49];
(3) Exploiting the different relative poses associated with the query and its similar images, the absolute pose of the query (with respect to a reference 3D world coordinate frame) is estimated. For this final pose estimation, the approach proposed in [59] is implemented.

Evaluation of the localization. To ensure coherence between all experiments, the dataset ND-dataset is split into three sets train/val/test of respective proportions 70/15/15%. Localization experiments are thus only performed on the test set, consisting of 1638 images selected randomly all around the cathedral. The evaluation of the pose quality is measured using three error metrics:

- The distance between actual and estimated camera 3D positions (in meter),
- The angle difference (between the two orientation quaternions) (in degree),
- The direction difference (in degree), which is similar to the angle difference without taking into account the rotation of the camera along its aiming direction. It thus estimates better whether or not the camera points towards the center of the scene.

For most of the following experiments and for each metric, the results will present the values of mean, median and the first and third quartile (Q1 and Q3).

In this work, the retrieval is performed without any particular accelerating retrieval strategy for efficient nearest neighbor search. By conducting 6 types of retrieval at run-time, the global retrieval time is necessarily increased facing a single-query retrieval, but we assume it can be notably reduced with appropriate indexes [43].

5.2 Experiments and results

In this section are conducted intermediate experiments and evaluations of the proposal facing the state of the art.

5.2.1 Influence of image descriptors on CIR4Loc. We begin by evaluating the performance of CIR4Loc, with respect to three different image descriptors, two of them directly coming from the literature:

- **CIR4Loc-CLIP**: the first one, using the classical CLIP4CIR setting, that is CLIP as the image descriptor (a global one) [5];
- **CIR4Loc-HowG**: the second one, aimed at exploiting the performance of the well-known image descriptor How [62], while remaining in the same setting of global description. The descriptor used is thus a pooled descriptor based on the feature map from which local descriptors are extracted.
- **CIR4Loc-HowL**: for this third one, we employ a local version of the descriptor How, by modifying the training process in order to provide a more local representation of the content. Indeed, extensive experiments (not reported here) have exhibited that using a global representation of the image content may limit the discriminative power during instance-based retrieval, with a potential consequence on localization. The proposed idea is to train the network using the locations of the local descriptors on the image and, based on the textual modifier, to estimate a score as to whether or not this local descriptor should be kept for the aggregation and thus for the retrieval process afterwards. It can be seen as an adaptation of CLIP4CIR where the descriptor is an embedding of the location of the local descriptor, where the modification leads to an associated binary score as to whether or not the descriptor will "attract" images from the right direction during the retrieval.

In order to quickly evaluate the impact of these descriptors on CIR4Loc in terms of pose estimation quality, we first set up a coarse evaluation where the pose is estimated by simply averaging the poses of four images. Those images are the first retrieved using CIR4Loc and its variants, on each of four directions ("left", "right", "above", "below"). The results are presented in Table 1: it involves the three variants of CIR4Loc with the averaged poses.

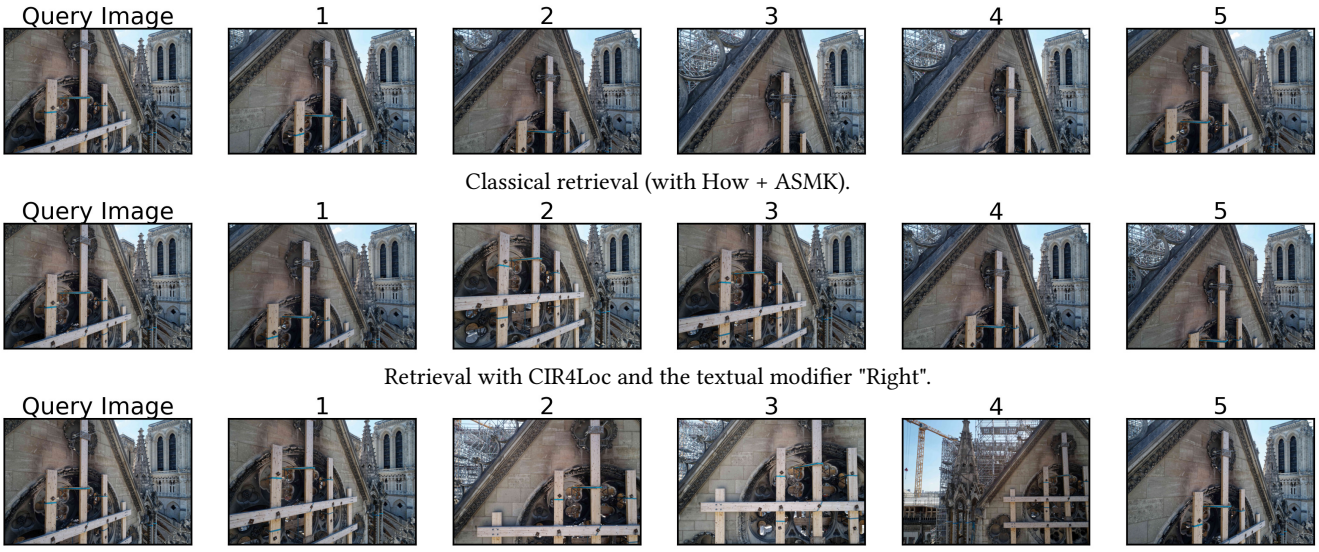


Figure 4: Visualization of the retrieval (first five closest responses) without and with CIR4Loc.

	Distance		Angle		Direction	
	Mean	Med.	Mean	Med.	Mean	Med.
CIR4Loc-CLIP	16.32	6.07	26.69	8.53	22.14	8.10
CIR4Loc-HowG	4.62	1.67	9.62	2.70	8.44	2.49
CIR4Loc-HowL	**4.41**	**1.53**	**9.17**	**2.69**	**8.01**	**2.45**

Table 1: Preliminary evaluation of the average first retrieved poses using the three variants of CIR4Loc on four directions ("left", "right", "above", "below"). Best results in bold.

The Table shows that the variants CIR4Loc-HowG and CIR4Loc-HowL are more efficient than CIR4Loc-CLIP to retrieve a set of well-located images whose average pose is close to the actual pose of the query image, whatever the metric considered. Furthermore, we also confirm that the local version of How, CIR4Loc-HowL, provides better results than the global one CIR4Loc-HowG. Thus, CLIP, as a descriptor for CIR4Loc, is discarded from the rest of the experiments, and only How-based approaches are kept. Although CIR4Loc-HowL appears slightly better than CIR4Loc-HowG, this evaluation is only preliminary, with only a rough pose estimation approach, not a finer one. This is why both approaches are kept in the following.

5.2.2 Localization baselines set up. As baseline of comparison, we have performed extensive experiments to determine the optimal state-of-the-art image retrieval strategies, that provide the best localization results. Two configurations were selected:

Retrieval-based localization. It relies on the best visual retrieval combination of techniques we have evaluated: How and ASMK as descriptor and retrieval strategy (see their introduction in Section 2.3).

Spatial-based localization. In order to improve the set of images retrieved for pose estimation, a second ad hoc approach is proposed. The selection of images is performed following these steps:

(1) From an initial visual-based retrieval (such as the previous one), the first five visually most similar images are retrieved;
(2) The poses of these images are averaged, excluding potential outliers (e.g. the poses of images further from the average pose by more than 10 meters). This leads to an average *a priori* pose of the query image;
(3) With this *a priori* pose, a spatial search is performed among all poses of the database, as follows:
   - Images whose poses are closest to the *a priori* viewpoint center are selected;
   - They are then filtered to check that their aiming direction is within 45 degrees of the aiming direction of the *a priori* pose;
   - The four images, both closest to the *a priori* viewpoint and respecting the direction constraint are selected.

For these two strategies, the selected images are then used in the pose estimation process presented in Section 5.1, similarly to the different variants of CIR4Loc.

5.2.3 Evaluation of CIR4Loc facing baselines. CIR4Loc with the descriptor How in both its global and local representations (CIR4Loc-HowG and CIR4Loc-HowL), is evaluated facing the two baselines previously presented. Preliminarily, several tests have consistently shown that using the six first images retrieved by CIR4Loc (with relative modifiers: "left", "right", "above", "below", "backward", "forward") rather than four (without "backward" and "forward") was detrimental to the final pose. Indeed, using "backward" and "forward" textual modifiers adds noise to the CIR4Loc retrieval process

Localization type	Distance				Angle				Direction			
	Mean	Median	Q1	Q3	Mean	Median	Q1	Q3	Mean	Median	Q1	Q3
Retrieval-based loc.	**3.56**	1.75	1.00	3.37	10.24	4.43	1.64	11.48	8.64	3.93	1.29	10.35
Spatial-based loc.	4.24	2.29	1.31	4.22	**9.08**	**2.81**	**0.82**	**9.42**	**7.42**	**2.38**	**0.62**	**7.99**
CIR4Loc-HowG	5.11	2.14	1.06	4.89	10.85	4.16	1.47	11.69	9.08	3.63	1.16	10.37
CIR4Loc-HowL	4.11	**1.45**	**0.79**	**2.91**	9.51	3.88	1.29	10.23	7.86	3.41	1.01	8.79

Table 2: Localization performances based on pure visual retrieval, spatial retrieval and composed image retrieval (CIR with How-Global and Local). In bold the best result and underlined the second best.

and thus to the pose estimation one. For the remainder of the experiments, the four images used for localization are those obtained by CIR4Loc using the following relative modifiers: "left", "right", "above", "below".

Table 2 presents the evaluation of CIR4Loc-HowG and CIR4Loc-HowL in this context, facing the baselines. The results show that the proposed CIR4Loc strategy is promising. First, these experiments still reinforce the fact that exploiting a local descriptor is essential in such a context. Indeed, the CIR4Loc-HowG performs worse than visual retrieval (using local descriptors but not set for pose estimation) in terms of camera location and is on par in terms of viewing angle. These experiments also demonstrate that a spatially guided composed image retrieval is better for pose estimation. Indeed, in terms of viewpoint location, although on average visual retrieval is better, up to the third quartile CIR4Loc-HowL is consistently better (40 cm better at Q3). This tends to show that when the CIR strategy performs poorly, it is very poor, and thus the pose estimation is missed by a lot. However, when it works, it sets up the pose estimation process in a better way.

In terms of viewing direction, although CIR4Loc-HowL does not reach the performance of the spatial retrieval, it consistently beats visual retrieval. This once more supports the idea that guiding the retrieval in terms of spatial directions leads to a better geometric disposition of images for pose estimation purposes.

To conclude, the experiments carried out with CIR4Loc tend to indicate that although image retrieval should be improved to perform as best as possible for visual similarity search, it is not always the best way to retrieve the optimal set of images, depending on the subsequent application. In this case of pose estimation, the proposed approach of composed image retrieval proves to be efficient for guiding the retrieval for this purpose.

6 Conclusions

This work has highlighted the crucial role of image retrieval in a localization process, while exposing its inherent limitations if not targeted to the application. Traditional retrieval pipelines are optimized for visual similarity, which does not always align with the requirements of pose estimation, where spatial image distribution and viewpoint diversity are critical. To alleviate this issue, we have proposed to exploit a composed image retrieval strategy, called CIR4Loc, involving the exploitation of a spatial knowledge between images. By incorporating textual modifiers that guide the retrieval of images spatially distributed in a structured manner around the query, CIR4Loc allows a more spatially aware selection of images. The results show that CIR4Loc improves pose estimation and is robust compared to other traditional retrieval-based approaches. It offers a new perspective on image selection for localization, reinforcing the idea that retrieval should not be limited to visual similarity but actively guided to suit specific spatial constraints.

Ultimately, this study reinforces the importance of thinking of image retrieval not only as a strategy to find similar images according to a given visual criterion - as it is usually done in several domains - but also as a tool that should be driven by the characteristics of the application. In this article where we consider localization, we have discussed and studied methods that balance visual similarity and spatial distribution.

As future work, we aim to fully integrate CIR4Loc within end-to-end localization pipelines. This involves coupling CIR4Loc with robust 6-DoF pose estimation frameworks, leveraging the spatial diversity of the retrieved image set to enhance multi-view geometry computation. CIR4Loc has been proposed to demonstrate the contribution of the text modifiers in geolocalization, but more generally, we will also have to experiment the global strength of the approach facing other approaches of geolocalization, such as those mentioned in Sections 2.2.2 and 2.2.3.

The structure ND-dataset made it particularly well-suited to the implementation and evaluation of the proposal, but we also plan to evaluate it on other heritage datasets, where repetitive structures, occlusions, and architectural symmetries also challenge conventional feature-based localization methods. Another avenue concerns the embedding of CIR4Loc within annotation-driven workflows, enabling spatially localized images to support visual documentation, and interactive navigation across structured cultural heritage datasets. Through these developments, we envision not merely a retrieval mechanism, but as a fundamental component of scalable, automated localization systems adapted to the complexities of real-world environments and cultural heritage applications.

Acknowledgments

All pictures of the Notre-Dame de Paris cathedral presented in this work are credited under the © Chantier Scientifique Notre-Dame de Paris / Ministère de la Culture / CNRS / EPRNDP, France. This work has been funded since 2019 by the CNRS MITI and the French Ministry of Culture within the framework of the national scientific action Notre-Dame de Paris, and since 2022 by the European Research Council (ERC) Advanced Grant nDame_Heritage: n-Dimensional analysis and memorization ecosystem for building cathedrals of knowledge in Heritage Science (Grant agreement ID:101055423).

References

[1] Lorenzo Agnolucci, Alberto Baldrati, Marco Bertini, and Alberto Del Bimbo. 2024. iSEARLE: Improving Textual Inversion for Zero-Shot Composed Image Retrieval. *arXiv* 2405.02951 (2024).

[2] Artem Babenko and Victor Lempitsky. 2015. Aggregating local deep features for image retrieval. In *International Conference on Computer Vision*. 1269–1277. doi:10.1109/ICCV.2015.150

[3] Song Bai, Peng Tang, Philip H S Torr, and Longin Jan Latecki. 2019. Re-ranking via metric fusion for object retrieval and person re-identification. In *Conference on Computer Vision and Pattern Recognition*. 740–749. doi:10.1109/CVPR.2019.00083

[4] Yang bai, Xinxing Xu, Yong Liu, Salman Khan, Fahad Khan, Wangmeng Zuo, Rick Siow Mong Goh, and Chun-Mei Feng. 2024. Sentence-level Prompts Benefit Composed Image Retrieval. In *International Conference on Learning Representations (ICLR)*.

[5] Alberto Baldrati, Marco Bertini, Tiberio Uricchio, and Alberto Del Bimbo. [n. d.]. Composed Image Retrieval using Contrastive Learning and Task-oriented CLIP-based Features. *ACM Transactions on Multimedia Computing, Communications and Applications* ([n. d.]).

[6] Gabriele Berton, Riccardo Mereu, Gabriele Trivigno, Carlo Masone, Gabriela Csurka, Torsten Sattler, and Barbara Caputo. 2022. Deep visual geo-localization benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5396–5407.

[7] 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*. 31–40.

[8] Emile Blettery and Valérie Gouet-Brunet. 2024. Heritage Iconographic Content Structuring: from Automatic Linking to Visual Validation. In *Journal on Computing and Cultural Heritage*.

[9] Dylan Campbell*, Liu Liu*, and Stephen Gould. 2020. Solving the Blind Perspective-n-Point Problem End-To-End with Robust Differentiable Geometric Optimization. In *ECCV*. * equal contribution.

[10] Bingyi Cao, André Araujo, and Jack Sim. 2020. Unifying Deep Local and Global Features for Image Search. In *European Conference on Computer Vision*, Vol. 12365. 726–743. doi:10.1007/978-3-030-58565-5_43

[11] Shuai Chen, Zirui Wang, and Victor Prisacariu. 2021. Direct-posenet: Absolute pose regression with photometric consistency. In *2021 International Conference on 3D Vision (3DV)*. IEEE, 1175–1185.

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

[13] Gabriella Csurka, C. Dance, Lixin Fan, J. Willamowski, and Cédric Bray. 2002. Visual categorization with bags of keypoints. In *European Conference on Computer Vision*.

[14] Agni Delvinioti, Hervé Jégou, Laurent Amsaleg, and Michael E Houle. 2014. Image retrieval with reciprocal and shared nearest neighbors. In *International Conference on Computer Vision Theory and Applications*, Vol. 2. 321–328. doi:10.5220/0004672303210328

[15] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. 2018. SuperPoint: Self-Supervised Interest Point Detection and Description. In *Conference on Computer Vision and Pattern Recognition Workshops*. 224–236. doi:10.1109/CVPRW.2018.00060

[16] Xiaoyi Dong, Jianmin Bao, Dongdong Chen, Weiming Zhang, Nenghai Yu, Lu Yuan, Dong Chen, and Baining Guo. 2022. Cswin transformer: A general vision transformer backbone with cross-shaped windows. In *Conference on Computer Vision and Pattern Recognition*. 12124–12134. doi:10.1109/CVPR52688.2022.01181

[17] Nikos Efthymiadis, Bill Psomas, Zakaria Laskar, Konstantinos Karantzalos, Yannis Avrithis, Ondřej Chum, and Giorgos Tolias. 2024. Composed Image Retrieval for Training-Free Domain Conversion. *arXiv* 2412.03297 (2024).

[18] Andrea Fusiello, Eleonora Maset, and Fabio 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. doi:10.5194/isprsarchives-xl-5-w1-81-2013

[19] Albert Gordo, Filip Radenovic, and Tamara Berg. 2020. Attention-based query expansion learning. In *European Conference on Computer Vision*, Vol. 12373. 172–188. doi:10.1007/978-3-030-58604-1_11

[20] Richard Hartley, Khurrum Aftab, and Jochen Trumpf. 2011. L1 rotation averaging using the Weiszfeld algorithm. In *CVPR 2011*. IEEE, 3041–3048.

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

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Conference on Computer Vision and Pattern Recognition*, Vol. 2016-Decem. 770–778. doi:10.1109/CVPR.2016.90

[23] Huaibo Huang, Xiaoqiang Zhou, Jie Cao, Ran He, and Tieniu Tan. 2023. Vision Transformer with Super Token Sampling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[24] Ahmet Iscen, Giorgos Tolias, Yannis Avrithis, Teddy Furon, and Ondrej Chum. 2017. Efficient diffusion on region manifolds: Recovering small objects with compact CNN representations. In *Conference on Computer Vision and Pattern Recognition*. 2077–2086. doi:10.1109/CVPR.2017.105

[25] Yingying Jiang, Hanchao Jia, Xiaobing Wang, and Peng Hao. 2024. HyCIR: Boosting Zero-Shot Composed Image Retrieval with Synthetic Labels. *arXiv* 2407.05795 (2024).

[26] 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 *Conference on Computer Vision and Pattern Recognition*. 2969–2976. doi:10.1109/CVPR.2011.5995464

[27] 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*. 929–938.

[28] Seongwon Lee, Suhyeon Lee, Hongje Seong, and Euntai Kim. 2023. Revisiting Self-Similarity: Structural Embedding for Image Retrieval. In *Computer Vision and Pattern Recognition (CVPR)*.

[29] Seongwon Lee, Hongje Seong, Suhyeon Lee, and Euntai Kim. 2022. Correlation Verification for Image Retrieval. In *Conference on Computer Vision and Pattern Recognition*. 5374–5384. doi:10.1109/CVPR52688.2022.00530

[30] 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. doi:10.1007/S11263-008-0152-6

[31] Vincent Leroy, Yohann Cabon, and Jérôme Revaud. 2024. Grounding image matching in 3d with mast3r. In *European Conference on Computer Vision*. Springer, 71–91.

[32] Wei Li, Xing Wang, Xin Xia, Jie Wu, Jiashi Li, Xuefeng Xiao, Min Zheng, and Shiping Wen. 2022. Sepvit: Separable vision transformer. *arXiv preprint arXiv:2203.15380* (2022).

[33] You Li, Fan Ma, and Yi Yang. 2024. Imagine and Seek: Improving Composed Image Retrieval with an Imagined Proxy. *arXiv* 2411.16752 (2024).

[34] Wei-Chao Lin. 2022. Block-based pseudo-relevance feedback for image retrieval. *Journal of Experimental and Theoretical Artificial Intelligence* 34, 5 (2022), 891–903. doi:10.1080/0952813X.2021.1938695

[35] Philipp Lindenberger, Paul-Edouard Sarlin, and Marc Pollefeys. 2023. LightGlue: Local Feature Matching at Light Speed. In *International Conference on Computer Vision*. https://arxiv.org/pdf/2306.13643.pdf

[36] Xingjian Liu, Wenyuan Chen, Harikrishnan Madhusudanan, Linghao Du, and Yu Sun. 2020. Camera orientation optimization in stereo vision systems for low measurement error. *IEEE/ASME Transactions on Mechatronics* 26, 2 (2020), 1178–1182.

[37] Zheyuan Liu, Weixuan Sun, Yicong Hong, Damien Teney, and Stephen Gould. 2024. Bi-Directional Training for Composed Image Retrieval via Text Prompt Learning. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*.

[38] Zheyuan Liu, Weixuan Sun, Damien Teney, and Stephen Gould. 2023. Candidate Set Re-ranking for Composed Image Retrieval with Dual Multi-modal Encoder. In *arXiv.org*.

[39] 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. doi:10.1007/s11263-020-01359-2

[40] Arthur Moreau, Nathan Piasco, Moussab Bennehar, Dzmitry Tsishkou, Bogdan Stanciulescu, and Arnaud de La Fortelle. 2023. Crossfire: Camera relocalization on self-supervised features from an implicit representation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 252–262.

[41] Uzair Nadeem, Mohammed Bennamoun, Roberto Togneri, Ferdous Sohel, Aref Miri Rekavandi, and Farid Boussaid. 2023. Cross domain 2D-3D descriptor matching for unconstrained 6-DOF pose estimation. *Pattern Recognition* 142 (2023), 109655.

[42] Hyeonwoo Noh, Andre Araujo, Jack Sim, Tobias Weyand, and Bohyung Han. 2017. Large-Scale Image Retrieval with Attentive Deep Local Features. In *International Conference on Computer Vision*, Vol. 2017-Octob. 3476–3485. doi:10.1109/ICCV.2017.374

[43] James Jie Pan, Jianguo Wang, and Guoliang Li. 2024. Survey of vector database management systems. *The VLDB Journal* 33, 5 (2024), 1591–1615.

[44] Shanmin Pang, Jin Ma, Jianru Xue, Jihua Zhu, and Vicente Ordonez. 2019. Deep Feature Aggregation and Image Re-Ranking With Heat Diffusion for Image Retrieval. *Transactions on Multimedia* 21, 6 (2019), 1513–1523. doi:10.1109/TMM.2018.2876833

[45] Noé Pion, Martin Humenberger, Gabriela Csurka, Yohann Cabon, and Torsten Sattler. 2020. Benchmarking image retrieval for visual localization. In *International Conference on 3D Vision*. 483–494. doi:10.1109/3DV50981.2020.00058

[46] Bill Psomas, Ioannis Kakogeorgiou, Nikos Efthymiadis, Giorgos Tolias, Ondřej Chum, Yannis Avrithis, and Konstantinos Karantzalos. 2024. Composed Image Retrieval for Remote Sensing. In *IGARSS 2024-2024 IEEE International Geoscience and Remote Sensing Symposium*. IEEE, 8526–8534.

[47] Filip Radenovic, Giorgos Tolias, and Ondrej Chum. 2019. Fine-Tuning CNN Image Retrieval with No Human Annotation. *Transactions on Pattern Analysis and Machine Intelligence* 41, 7 (2019), 1655–1668. doi:10.1109/TPAMI.2018.2846566

[48] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *International Conference on Machine Learning (ICML)*.

[49] E Rupnik, M Daakir, and M Pierrot Deseilligny. 2017. {MicMac} – a free, open-source solution for photogrammetry. *Open Geospatial Data, Software and Standards* 2, 1 (12 2017), 14. doi:10.1186/s40965-017-0027-2

[50] Mert Bulent Sariyildiz, Philippe Weinzaepfel, Thomas Lucas, Pau De Jorge, Diane Larlus, and Yannis Kalantidis. 2025. DUNE: Distilling a Universal Encoder from heterogenous 2D and 3D teachers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Nashville, Tennessee, USA.

[51] 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 *Conference on Computer Vision and Pattern Recognition*. 3247–3257. doi:10.1109/CVPR46437.2021.00326

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

[53] Torsten Sattler, Qunjie Zhou, Marc Pollefeys, and Laura Leal-Taixe. 2019. Understanding the limitations of cnn-based absolute camera pose regression. In *Conference on Computer Vision and Pattern Recognition*. 3302–3312. doi:10.1109/CVPR.2019.00342

[54] Johannes Lutz Schönberger and Jan-Michael Frahm. 2016. Structure-from-Motion Revisited. In *Conference on Computer Vision and Pattern Recognition*. 4104–4113. doi:10.1109/CVPR.2016.445

[55] Xi Shen, Yang Xiao, Hu Shell Xu, Othman Sbai, and Mathieu Aubry. 2021. Re-ranking for image retrieval and transductive few-shot classification. In *Advances on Neural Information Processing Systems*. 25932–25943. https://proceedings.neurips.cc/paper/2021/hash/d9fc0cdb67638d50f411432d0d41d0ba-Abstract.html

[56] Karen Simonyan and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. *International Conference on Learning Representations* (2015), 1–14. https://arxiv.org/pdf/1409.1556.pdf%E3%80%82

[57] Josef Sivic and Andrew Zisserman. 2003. Video Google: A text retrieval approach to object matching in videos. In *International Conference on Computer Vision*, Vol. 2. 1470–1477. doi:10.1109/iccv.2003.1238663

[58] Xuemeng Song, Haoqiang Lin, Haokun Wen, Bohan Hou, Mingzhu Xu, and Liqiang Nie. 2025. A Comprehensive Survey on Composed Image Retrieval. *arXiv* 2502.18495 (2025).

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

[60] Shitong Sun, Fanghua Ye, and Shaogang Gong. 2023. Training-free zero-shot composed image retrieval with local concept reranking. *arXiv preprint arXiv:2312.08924* (2023).

[61] Giorgos Tolias, Yannis Avrithis, and Hervé Jégou. 2016. Image search with selective match kernels: aggregation across single and multiple images. *International Journal of Computer Vision* 116 (2016), 247–261. doi:10.1007/S11263-015-0810-4

[62] Giorgos Tolias, Tomas Jenicek, and Ondřej Chum. 2020. Learning and Aggregating Deep Local Descriptors for Instance-Level Recognition. In *European Conference on Computer Vision*, Vol. 12346 LNCS. 460–477. doi:10.1007/978-3-030-58452-8_27

[63] Ankit Vani, Bac Nguyen, Samuel Lavoie, Ranjay Krishna, and Aaron Courville. 2024. SPARO: Selective Attention for Robust and Compositional Transformer Encodings for Vision. In *European Conference on Computer Vision (ECCV)*.

[64] Guangming Wang, Yu Zheng, Yanfeng Guo, Zhe Liu, Yixiang Zhu, Wolfram Burgard, and Hesheng Wang. 2023. End-to-end 2d-3d registration between image and lidar point cloud for vehicle localization. *arXiv preprint arXiv:2306.11346* (2023).

[65] Qi Wang, Weidong Min, Daojing He, Song Zou, Tiemei Huang, Yu Zhang, and Ruikang Liu. 2020. Discriminative fine-grained network for vehicle re-identification using two-stage re-ranking. *Science China Information Sciences* 63 (2020), 1–12. doi:10.1007/S11432-019-2811-8

[66] Shuzhe Wang, Juho Kannala, and Daniel Barath. 2024. DGC-GNN: Leveraging Geometry and Color Cues for Visual Descriptor-Free 2D-3D Matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 20881–20891.

[67] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. 2024. DUSt3R: Geometric 3D Vision Made Easy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 20697–20709.

[68] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. 2021. Pyramid Vision Transformer: A versatile backbone for dense prediction without convolutions. In *International Conference on Computer Vision*. 568–578. doi:10.1109/ICCV48922.2021.00061

[69] Laura Willot, Dan Vodislav, Valerie Gouet-Brunet, Livio De Luca, and Adeline Manuel. 2023. Clustering for the Analysis and Enrichment of Corpus of Images for the Spatio-temporal Monitoring of Restoration Sites. In *SUMAC '23: Proceedings of the 5th Workshop on AnalySis, Understanding and ProMotion of HeritAge Contents*. ACM Multimedia 2023, ACM, Ottawa, Canada, 39–47. doi:10.1145/3607542.3617353

[70] Fei Xue, Xin Wu, Shaojun Cai, and Junqiu Wang. 2020. Learning multi-view camera relocalization with graph neural networks. In *Conference on Computer Vision and Pattern Recognition*. 11372–11381. doi:10.1109/CVPR42600.2020.01139

[71] Min Yang, Dongliang He, Miao Fan, Baorong Shi, Xuetong Xue, Fu Li, Errui Ding, and Jizhou Huang. 2021. DOLG: Single-stage image retrieval with deep orthogonal fusion of local and global features. In *International Conference on Computer Vision*. 11772–11781. doi:10.1109/ICCV48922.2021.01156

[72] Guangnan Ye, Dong Liu, I-Hong Jhuo, and Shih-Fu Chang. 2012. Robust late fusion with rank minimization. In *Computer Vision and Pattern Recognition*. 3021–3028. doi:10.1109/CVPR.2012.6248032

[73] Mubariz Zaffar, Sourav Garg, Michael Milford, Julian Kooij, David Flynn, Klaus McDonald-Maier, and Shoaib Ehsan. 2021. Vpr-bench: An open-source visual place recognition evaluation framework with quantifiable viewpoint and appearance change. *International Journal of Computer Vision* 129, 7 (2021), 2136–2174.

[74] Hang Zhang, Chongruo Wu, Zhongyue Zhang, Yi Zhu, Haibin Lin, Zhi Zhang, Yue Sun, Tong He, Jonas Mueller, R. Manmatha, Mu Li, and Alexander Smola. 2020. ResNeSt: Split-Attention Networks. In *Conference on Computer Vision and Pattern Recognition Workshops*. 2736–2746. doi:10.1109/CVPRW56347.2022.00309

[75] Shaoting Zhang, Ming Yang, Timothee Cour, Kai Yu, and Dimitris N Metaxas. 2012. Query specific fusion for image retrieval. In *European Conference on Computer Vision*. 660–673. doi:10.1007/978-3-642-33709-3_47

[76] Xuanmeng Zhang, Minyue Jiang, Zhedong Zheng, Xiao Tan, Errui Ding, and Yi Yang. 2020. Understanding Image Retrieval Re-Ranking: A Graph Neural Network Perspective. (2020). https://arxiv.org/abs/2012.07620

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

[78] Zhongyan Zhang, Lei Wang, Luping Zhou, and Piotr Koniusz. 2023. Learning Spatial-context-aware Global Visual Feature Representation for Instance Image Retrieval. In *IEEE International Conference on Computer Vision (ICCV)*.

[79] Qunjie Zhou, Torsten Sattler, Marc Pollefeys, and Laura Leal-Taixe. 2020. To learn or not to learn: Visual localization from essential matrices. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 3319–3326.

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