Composed Image Retrieval For Visual Localization: Evaluation For Architectural Contents

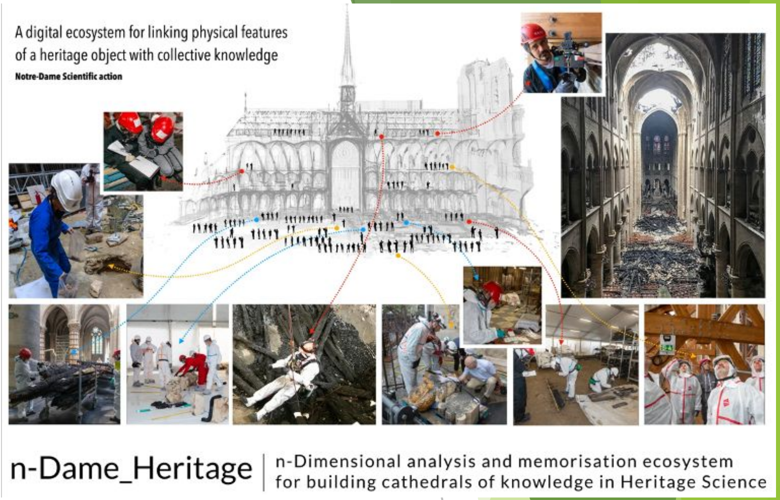
Emile Blettery[1 2], Valérie Gouet-Brunet[1] and Livio De Luca[2]

[1]LaSTIG, Univ Gustave Eiffel, IGN-Geodata Paris, France

[2]UPR CNRS 2002 MAP, Marseille, France

Context

- Part of the **N-Dame_Heritage ERC project**
- Scientific work alongside the restoration of the Notre-Dame cathedral after the fire



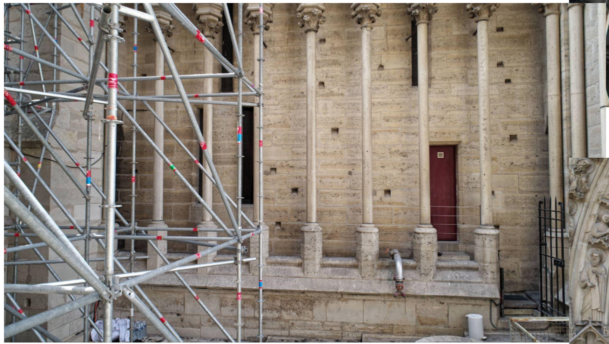
- Focus : **image localization** (position and pose) within a large and diverse corpus of localized images
- **No other data** (existing or built) than localized images

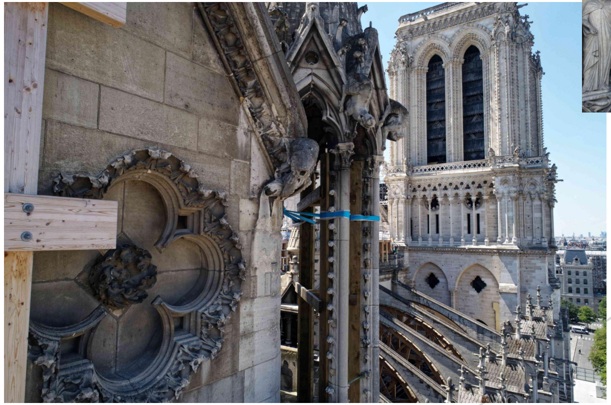
- Goal : **on-the-fly integration of novel images** within an ever growing collection

Dataset considered

- **10,901** images:
  - Exterior of the cathedral
  - Harmoniously distributed
  - High visual overlap

- **Challenges:**
  - Visual similarities
  - Repeated patterns
  - Multiple key elements in the background





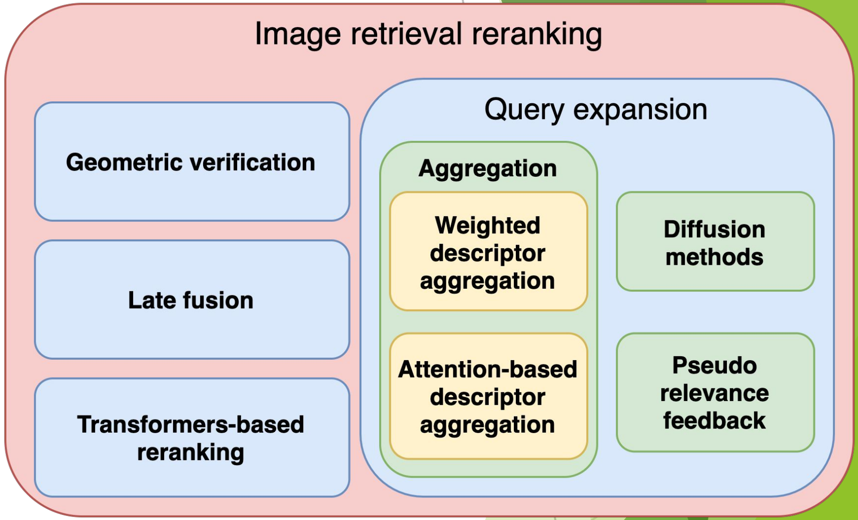
© AGP

Visual localization approaches

- Image retrieval-based approaches, our focus: *(Pion et al., 2020)*
  - **CBIR** in a reference dataset of localized images
  - Localization goes from **pose assignment to triangulation-based pose estimation**
- 3D-based approaches: *(Schönberger et al., 2016)*
  - CBIR identifies reference images and thus **associated/computed 3D points**
  - **PnP solver** computes the query image's pose *(Sattler et al., 2014)*
  - Novel trained approaches compute direct 2D-3D matches, without reference images *(Nadeem et al., 2023)*
- Trained, all-in-one approaches:
  - Take only i**mages as input and output a pose**
  - **RPR/APR** still do not generalize well (especially for large areas) *(Moreau et al., 2023)*
  - **Multi-task approaches** are promising but not adapted to such datasets *(Leroy et al., 2024)*

Image retrieval for visual localization

- Different image descriptors:
  - Most powerful ones are **trained**, with common backbones
  - **Global** descriptors exploit the whole visual context
  - **Local** ones focus on and aggregate salient elements
- An added re-ranking step:
  - Multiple options as seen here
  - Many potential **combinations**
- Our selection:
  - **How and ASMK** *(Tolias et al., 2020)* as image descriptor for retrieval
  - Point detector and descriptor **SuperPoint** *(DeTone et al. 2018)*, matched with **LightGlue** *(Lindenberger et al., 2023)* for geometric verification and subsequent pose estimation



Composed Image Retrieval (CIR)

- Retrieves an image based on an initial query image **and a textual modifier**
- Retrieval is guided both **visually and verbally**
- **Different types of approach** to tackle this

CIR main approaches

- Feature modifying approaches: *(Baldrati et al., 2023)*
  - **Textual input** is taken as **modifier** to the visual descriptor
  - The visual descriptor is modified via a **combiner network**
  - Image retrieval is initiated from this **modified visual descriptor**
  - **CLIP4CIR** is the method that inspired our proposal

- Composition-based approaches: *(Psomas et al., 2024)*
  - Uses both **textual and visual features** combined with a **weighting scheme**
  - Could allow for pure monomodal retrieval

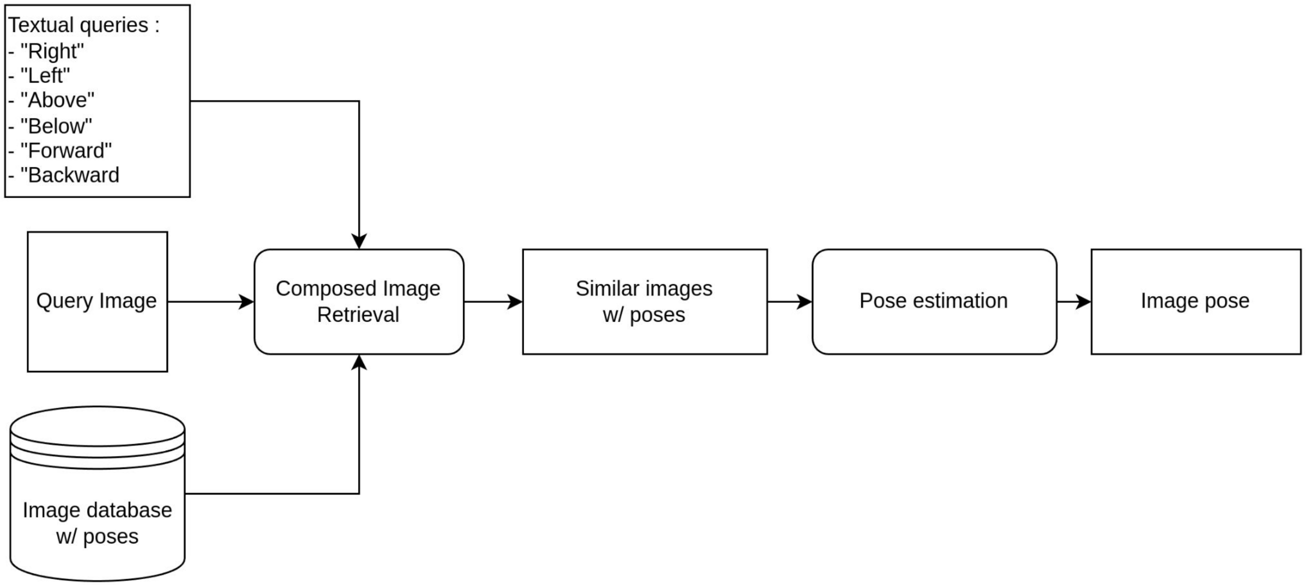
- Generation-based approaches: *(Li et al. 2024)*
  - **Generates a novel image** from the textual description
  - **Average** query and novel images descriptors for retrieval

Our proposal : CIR4Loc

- <u>Problem statement:</u>
  - Image retrieval is an **adequate base for pose estimation**
  - But its goal is to **maximize visual similarity, i.e. to retrieve images with similar viewpoints**
  - The **spatial configuration** of retrieved images may be **unsuited for pose estimation**

- <u>Proposed solution:</u>
  - Composed Image Retrieval with **spatial modifiers**
    - **relative** to the image:
      Above, Below, Left, Right, Forward, Backward
    - **absolute** in the reference system:
      Higher, Lower, Northward, Southward, Westward, Forward
  - To guide retrieval towards the best **spatially distributed** set of similar images

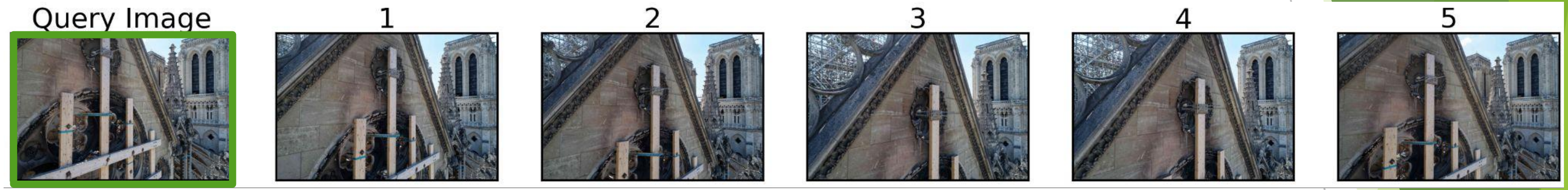
Our proposal : CIR4Loc



- **Three different models** are trained for each type of movement
- At retrieval time, the query image is associated to each spatial modifier
- The different lists are combined to obtain a spatially distributed set

CIR examples

- Classical image retrieval:



- Composed image retrieval with “Right” modifier:

- Composed image retrieval with “Left” modifier:

Evaluation framework

- The localization process:
  - *Keypoint detection and matching* with **SuperPoint + LightGlue**
  - *Relative pose estimation* (between query and each reference image) using open source library **Micmac**
  - *Final pose estimation* based on relative poses weighting from (Song et al., 2016)

- Evaluation 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), similar to the angle difference without the rotation of the camera along its aiming direction.
  - Mean, median, Q1 and Q3 values

Evaluation framework

- <u>Evaluation baselines to get a set of images for pose estimation:</u>
  - *Retrieval-based localization:*
    - Basic **CBIR-based** approach to get the set of images
    - Retrieval optimized for **visual similarity**
    - **How + ASMK** as descriptor
  - *Spatial-based localization:*
    - From the visual-based retrieval, poses of the **five most similar are averaged,** excluding outliers, to get an ***a priori* pose for the query**
    - A **spatial search** retrieves images closest to this *a priori* pose
    - Their poses are filtered so their **aiming direction is within 45 degrees** of the aiming direction of the a priori pose
    - The **four closest images respecting the angle constraint** are chosen
- Finally, the **localization process** estimates the query's pose

Experiments on CIR4Loc descriptors

- CIR4Loc variants based on descriptors:
  - using **CLIP** as an image descriptor (as used in the CLIP4CIR inspiration)
  - using **How as a global descriptor (HowG)**:
    - leverages **How's performance**
    - remains similar in terms of **descriptor type** (global)
  - using **How as a local descriptor (HowL)**:
    - the local aspect **increases retrieval performance** greatly
    - the network is modified to use the **locations of the local descriptors**
    - the network assigns a **binary score** to each descriptor based on location as to whether or not it should be used for retrieval
    - it will thus **"attract" images** from the direction of the spatial modifier

Results on CIR4Loc descriptors

- Preliminary results based on the average first retrieved poses using the variants of CIR4Loc on four directions ("left", "right", "above", "below")

	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**

- The **local version of How outperforms** all other variants
- The CLIP based-version is not at all suited for such type of contents

Experiments

- <u>CIR4Loc vs. baselines:</u>
  - *Retrieval-based* localization:
    - pure visual similarity
  - *Spatial-based* localization:
    - mostly spatial proximity
  - *CIR4Loc-HowG* based localization:
    - global representation of the How descriptor
  - *CIR4Loc-HowL* based localization:
    - local representation of the How descriptor

Results on CIR4Loc vs baselines

- Localization performances based on different retrieval

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

- A **local descriptor is essential** (CIR4Loc-HowG is worse than classical retrieval)
- For **viewpoint estimation**, up to Q3, CIR4Loc-HowL is the best, indicating a real improvement in cases where CIR performs correctly
- For **viewing direction**, spatial-based localization is better but CIR4Loc-HowL outperforms retrieval-based localization
- **CIR for localization is quite promising**

Conclusion & Perspectives

- For image based localization, the **retrieval step is crucial**
- **BUT** CBIR goals do not align with pose estimation requirements
- Proposed solution : **CIR4Loc,** composed image retrieval with spatial modifiers
- To guide retrieval towards a **spatially aware set of images**
- Promising results highlighting that **image retrieval should be driven by the characteristics of the application**
- Perspectives:
  - Integrate CIR4Loc in **end-to-end localization pipelines/systems**
  - Compare CIR4Loc to **other type of approaches** (3D, all-in-one)
  - Evaluate CIR4Loc on **other heritage datasets** challenging for

References

- ***Baldrati**, Alberto, et al. "Composed image retrieval using contrastive learning and task-oriented clip-based features." ACM Transactions on Multimedia Computing, Communications and Applications 20.3 (2023): 1-24.*
- ***DeTone**, Daniel, Tomasz Malisiewicz, and Andrew Rabinovich. "Superpoint: Self-supervised interest point detection and description." Proceedings of the IEEE conference on computer vision and pattern recognition workshops. 2018.*
- ***Leroy**, Vincent, Yohann Cabon, and Jérôme Revaud. "Grounding image matching in 3d with mast3r." European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2024.*
- ***Li**, You, Fan Ma, and Yi Yang. "Imagine and seek: Improving composed image retrieval with an imagined proxy." Proceedings of the Computer Vision and Pattern Recognition Conference. 2025.*
- ***Lindenberger**, Philipp, Paul-Edouard Sarlin, and Marc Pollefeys. "Lightglue: Local feature matching at light speed." Proceedings of the IEEE/CVF international conference on computer vision. 2023.*
- ***Moreau**, Arthur, et al. "Crossfire: Camera relocalization on self-supervised features from an implicit representation." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023.*
- ***Nadeem**, Uzair, et al. "Cross domain 2D-3D descriptor matching for unconstrained 6-DOF pose estimation." Pattern Recognition 142 (2023): 109655.*
- ***Pion**, Noé, et al. "Benchmarking image retrieval for visual localization." 2020 International Conference on 3D Vision (3DV). IEEE, 2020.*
- ***Psomas**, Bill, et al. "Composed image retrieval for remote sensing." IGARSS 2024-2024 IEEE International Geoscience and Remote Sensing Symposium. IEEE, 2024.*
- ***Sattler**, Torsten Chris Sweeney, and Marc Pollefeys. 2014. On sampling focal length values to solve the absolute pose problem. In European Conference on Computer Vision.*
- ***Schonberger**, Johannes L., and Jan-Michael Frahm. "Structure-from-motion revisited." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.*
- ***Tolias**, Giorgos, Tomas Jenicek, and Ondřej Chum. "Learning and aggregating deep local descriptors for instance-level recognition." European Conference on Computer Vision. Cham: Springer International Publishing, 2020.*