박 사 학 위 논 문
Ph.D. Dissertation

# 공간적, 위상학적 및 의미적 일관성을 갖춘 자기 적응형 시각 검색 및 QA 시스템 개발

# Developing a Self-Adaptive Visual Search and QA System with Spatial, Topological, and Semantic Consistency

2024

안 국 원 (安 國 元 An, Guoyuan)

한 국 과 학 기 술 원

Korea Advanced Institute of Science and Technology

박 사 학 위 논 문

# 공간적, 위상학적 및 의미적 일관성을 갖춘 자기 적응형 시각 검색 및 QA 시스템 개발

2024

안 국 원

한 국 과 학 기 술 원

전산학부

# 공간적, 위상학적 및 의미적 일관성을 갖춘 자기 적응형 시각 검색 및 QA 시스템 개발

안 국 원

위 논문은 한국과학기술원 박사학위논문으로
학위논문 심사위원회의 심사를 통과하였음

2024년 11월 19일

심사위원장 윤 성 의 (인)

심 사 위 원 김 명 호 (인)

심 사 위 원 오 혜 연 (인)

심 사 위 원 김 주 호 (인)

심 사 위 원 홍 승 훈 (인)

# Developing a Self-Adaptive Visual Search and QA System with Spatial, Topological, and Semantic Consistency

Guoyuan An

Advisor: Sung-Eui Yoon

A dissertation submitted to the faculty of
Korea Advanced Institute of Science and Technology in
partial fulfillment of the requirements for the degree of
Doctor of Philosophy in Computer Science

Daejeon, Korea
November 19, 2024

Approved by

---

Sung-Eui Yoon
Professor of Computer Science

The study was conducted in accordance with Code of Research Ethics[1].

[1] Declaration of Ethical Conduct in Research: I, as a graduate student of Korea Advanced Institute of Science and Technology, hereby declare that I have not committed any act that may damage the credibility of my research. This includes, but is not limited to, falsification, thesis written by someone else, distortion of research findings, and plagiarism. I confirm that my thesis contains honest conclusions based on my own careful research under the guidance of my advisor.

**초 록**

본 학위논문에서는 사용자 상호작용과 내적 일관성 검증을 통해 이미지에 대한 이해를 동적으로 개선하는 자기적응형 시각 검색 및 질의응답(Question Answering) 시스템을 제안한다. 기존의 정적으로 사전 학습된 지식에 의존하는 오픈-보캐블러리(Open-Vocabulary) 인식 및 VQA 모델과 달리, 본 시스템은 사용자 정의 개념, 개별 객체 특성, 변화하는 환경에 유연하게 대응할 수 있다.

이를 위해 본 연구는 데이터베이스 정보를 활용한 샘플 단위 적응(per sample-adaptation)과 새로운 의미적 · 시각적 정보를 획득함에 따라 시스템의 지식 베이스를 갱신하는 데이터베이스 적응(database-adaptation)을 결합한다. 사용자는 픽셀 단위 상호작용(pixel-level interaction)과 대화형 상호작용(conversational interaction)을 통해 직관적인 피드백을 제공할 수 있으며, 동시에 공간적 · 위상적 · 의미적 일관성 점검을 통해 시스템은 인간 개입 없이도 자동으로 오인식을 교정한다.

실험 결과, 본 접근법은 추론 정확도를 높이고 모델 재학습 필요성을 줄여, 시스템이 사용자 요구 변화에 더욱 긴밀히 대응할 수 있음을 보여준다. 궁극적으로 이 연구는 시각 검색 및 질의응답 시스템의 유연성, 개인화, 그리고 지속적 개선 능력을 한 단계 발전시킨다.

**핵 심 낱 말** 컴퓨터 비전, 이미지 검색, 정보 검색, 대형 언어 모델

**Abstract**

This thesis introduces a self-adaptive visual search and question answering system that dynamically refines its understanding of images through both user interaction and internal consistency checks. Unlike traditional open-vocabulary recognition and VQA models that rely on static, pre-trained knowledge, our system can adjust to user-defined concepts, personalized object properties, and changing environments.

We achieve this by combining per sample-adaptation—where each input is refined using database information and user feedback—with database-adaptation, which updates the system's knowledge base as it encounters new semantic and visual cues. User interactions occur through pixel-level and conversational modes, allowing for intuitive feedback. At the same time, implicit consistency checks (spatial, topological, and semantic) enable the system to self-correct without human intervention.

Experimental results show that this approach increases inference accuracy, reduces the need for model retraining, and aligns the system more closely with users' evolving needs. Overall, this work advances the flexibility, personalization, and continuous improvement of visual search and QA systems.

**Keywords** computer vision, image retrieval, information retrieval, large multimodal model

# Contents

# List of Tables

# List of Figures

# Chapter 1. Introduction

As intelligent vision systems continue to advance, the ability to recognize objects, answer queries, and adapt to new information stands at the forefront of research in computer vision and artificial intelligence [3, 125, 9]. Traditional open-vocabulary recognition and visual question answering (VQA) systems have made significant progress in labeling objects, understanding scenes, and interacting with users [49, 59]. However, these systems often operate in a fixed paradigm: they rely on pre-trained knowledge and struggle to adapt dynamically to new user-defined concepts, personalized object properties, and evolving environments. To address these limitations, this thesis proposes a novel self-adaptive visual search and question answering system designed to operate at a pixel level, seamlessly integrate new knowledge, and ensure consistent interpretation of visual and semantic content.

At the core of this self-adaptive system are two complementary components: per sample-adaptation and database-adaptation. Per sample-adaptation refers to dynamically refining inference results for a single input instance—adjusting initial predictions through post-processing steps to improve accuracy [68, 6, 8]. Inspired by the way humans reflect on their initial impressions to correct mistakes, this approach retrieves related information from a knowledge database and verifies it against the current input's attributes to enhance the fidelity of each prediction. By contrast, database-adaptation involves updating the underlying knowledge repository itself. This is essential for capturing evolving contextual information, such as a user's personalized naming conventions for familiar objects or the new visual characteristics introduced when an object is observed from a novel viewpoint.

The interplay between per sample-adaptation and database-adaptation forms the core of the system's continuous improvement process. From one perspective, enhancing per-sample accuracy relies on retrieving more relevant information from the database and verifying its consistency. From the other perspective, each adapted sample enriches the database, as user input and environmental observations introduce additional semantic and visual details. This closed-loop mechanism ensures that, over time, the system naturally aligns itself more closely with evolving user requirements, contextual nuances, and dynamic environmental changes.

To achieve this dynamic self-adaptation, the system leverages both explicit and implicit forms of feedback. Explicit feedback is facilitated through two primary interaction modes: pixel-level interaction and conversable interaction. Pixel-level interaction allows users to pinpoint exact regions of interest in an image, providing granular guidance on which objects to focus on or correct [7]. Conversable interaction enables the user and the system to engage in a dialogue, allowing for clarification, annotation, or the incorporation of new knowledge into the database through natural language instructions [3]. Although this explicit feedback requires a human in the loop, the system still qualifies as self-adaptive because it is the system itself that provides the pixel-level and conversable interfaces for interaction, and it also incorporates a consistency-based implicit feedback mechanism that does not rely on any human input.

Implicit feedback emerges through the concept of consistency. Visual and semantic consistency checks help the system self-correct and refine its predictions without direct user intervention. For instance, if a model misidentifies a landmark in an image, it can compare that image's spatial and topological features with related images in the database. By examining the geometry, arrangement, and semantic attributes of the known landmark, the system can detect inconsistencies and adjust its classification to the correct label—even when the user does not provide explicit corrections.

In this thesis, we present a comprehensive framework for realizing such a self-adaptive pixel-level visual search and QA system. We first describe the explicit methods for user interaction and adaptation in Chapters 2 and 3. Chapter 2 focuses on pixel-level interaction techniques [7], while Chapter 3 explores conversable interaction, illustrating how human-in-the-loop guidance can refine predictions and tailor the database to user-specific contexts. Subsequently, we address implicit self-adaptation based on consistency across various levels of analysis. Chapter 3 introduces semantic consistency checks, Chapter 4 delves into spatial consistency [6], and Chapter 5 investigates topological consistency mechanisms [8]. Together, these techniques demonstrate how the system can blend user guidance with internal consistency checks to achieve robust and evolving performance over time.

In summary, this thesis presents a unified approach that transcends traditional open-vocabulary recognition and VQA systems by embedding the principles of self-adaptation. Through both explicit human feedback (facilitated by system-provided interaction modes) and implicit consistency checks, the proposed system evolves to better understand its users, tailor its knowledge to their environment, and continually improve the accuracy and relevance of its predictions.

# Chapter 2. Towards Content-based Pixel Retrieval in Revisited Oxford and Paris

## 2.1 Introduction

Image retrieval is a long-standing and fundamental computer vision task and has achieved remarkable advances. However, because the retrieved ranking list contains false positive images and the true positive images contain complex co-occurring backgrounds, users may find it difficult to identify the query object from the ranking list. In this paper, we execute a user study and show that providing pixel-level annotations can help users better understand the retrieved results. Therefore, this paper introduces the pixel retrieval task and its first benchmarks. Pixel retrieval is defined as searching pixels that depict the query object from the database. More specifically, it requires the machine to recognize, localize, and segment the query object in database images in run time, as shown in Figure 2.1.

Similar to semantic segmentation, which works as an extension of classification and provides pixel-level category information to the machines, pixel retrieval is an extension of image retrieval. However, pixel retrieval differs from existing semantic segmentation [48, 157, 76] in two aspects: the fine-grained particular instance recognition and the variable-granularity recognition.

On the one hand, pixel retrieval asks the machine to consider the fine-grained information to segment the same instance with the query, *e.g.*, segment the particular query building in the street figures that contain many similar buildings. This is different from existing semantic segmentation [48] and instance segmentation [76, 157]. Semantic segmentation only requires the category level information, *e.g.*, to segment all the buildings in the street figures. On top of semantic segmentation, instance segmentation additionally requires demarcating individual instances, *e.g.*, segmenting all the buildings and giving the boundary of each building separately. However, instance segmentation does not distinguish the differences among the buildings [157, 76, 18].

On the other hand, pixel retrieval requires adjusting the recognition granularity as needed. The query image can be the whole building or only a part of the building. The search engine should understand the intention of the query and adjust the segmentation granularity in demand. This differs from existing segmentation benchmarks [157, 27, 69, 34, 39], where the recognition granularity is fixed in advance. Therefore, the pixel retrieval task is supplementary to semantic and instance segmentation by considering the recognition and segmentation featured with fine-grained and variable-granularity properties, which are also fundamental visual abilities of humans.

In order to promote the study of pixel retrieval, we create the pixel retrieval benchmarks Pixel-Revisited-Oxford (PROxford) and Pixel-Revisited-Paris (PRParis) on top of the famous image retrieval benchmarks Revisited-Oxford (ROxford) and Revisited-Paris (RParis) [91, 92, 97]. There are three reasons to use ROxford and RParis as our base benchmarks. Firstly, they are notoriously difficult and can better reflect the search engines' performance. Secondly, each query in these datasets has up to hundreds of positive images, so they are suitable for evaluating the fine-grained recognition ability. Thirdly, every positive image is guaranteed to be identifiable by people without considering any contextual visual information [97].

We provide the segmentation labels to a total of 5,942 images in ROxford and RParis. To ensure the label quality, three professional annotators independently label the query-index pairs and then refine and check the labels. The

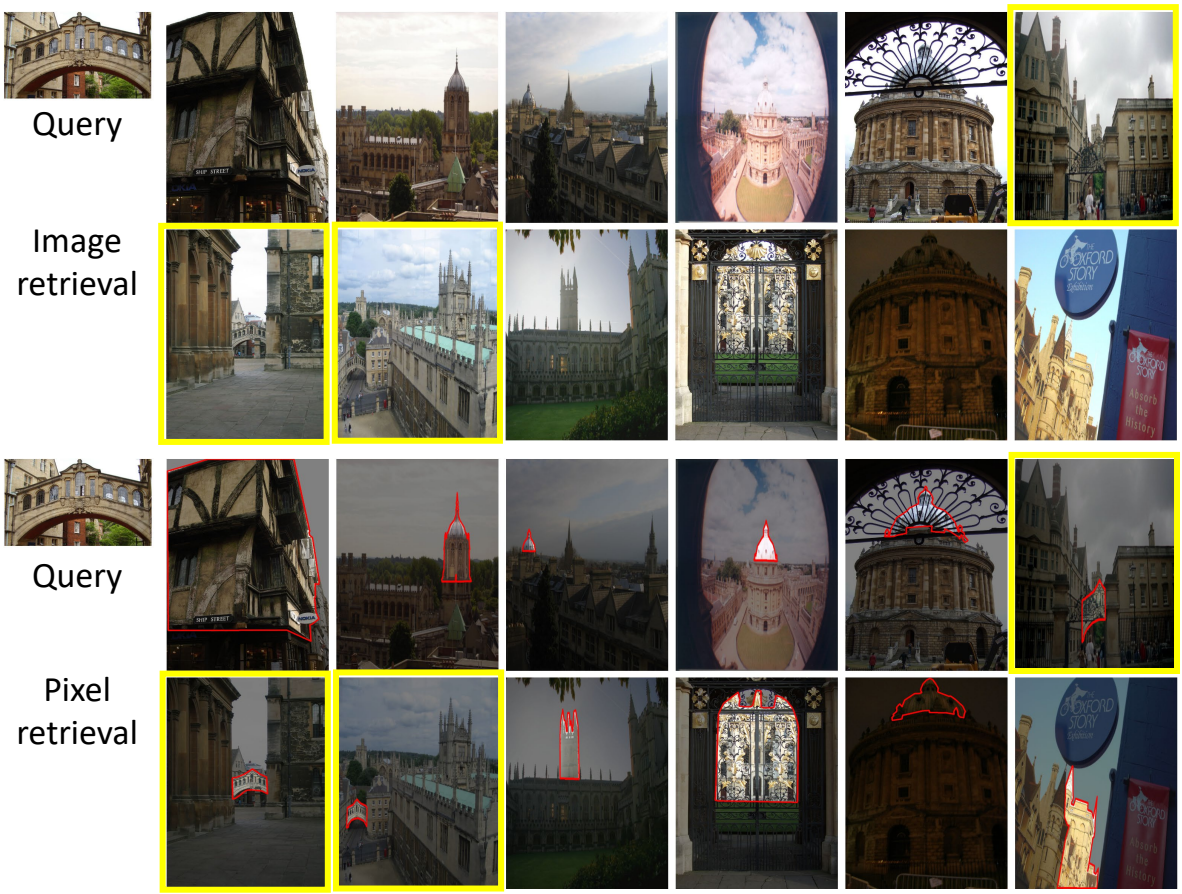



**Figure 2.1:** Example scenarios of image retrieval and pixel retrieval for the same query image. Pixel retrieval offers pixel-level annotation (red outlines) on the target object. Our user study shows that pixel retrieval can significantly improve the user experience (Sec. 2.3). Yellow boxes in the searched results indicate the ground truth ones. You can check our user study from this link. To start the user study, please enter any character into the "unique Prolific ID" blank.

annotators are aged between 26 to 32 years old and have worked full-time on annotation for over two years. We then design new metrics, mAP@50:5:95, and mAP, to evaluate the pixel retrieval performance (Section 2.2).

We provide an extensive comparison of State-Of-The-Art (SOTA) methods in related fields, including image search, detection, segmentation, and dense matching with our benchmarks. We have some interesting findings from the experiment. For example, we find the SOTA spatial verification methods [21, 88] give a high inlier number to some true query-index pairs but match the wrong regions. We find the dense and pixel-level approaches [82, 136] helpful for the pixel retrieval task. Most importantly, our results show that pixel retrieval is difficult, and further research is needed to advance the user experience on the content-based search task. `

Our contributions are as follows:

- We introduced the pixel retrieval task and provided the first two landmark pixel retrieval benchmarks, PROxford and PRParis. Three professional annotators labeled, refined, and checked the labels.
- We executed the user study and showed that the pixel level annotation could significantly improve user experience.
- We performed extensive experiments with SOTA methods in image search, detection, segmentation, and

dense matching. Our experiment results can be used as the baselines for future study.

## 2.2 Content-based pixel retrieval

### 2.2.1 Why Revisited Oxford and Paris?

We design the first content-based pixel retrieval benchmarks, PROxford and PRParis, directly on top of the famous image retrieval benchmarks Revisited-Oxford (ROxford) and Revisited-Paris (RParis) [91, 92, 97]. Oxford [91] and Paris [92] are introduced by Philbin *et al*. in 2007 and 2008, respectively. Their images are obtained from Flickr by searching text tags for famous landmarks in Oxford University and Paris. Radenovic *et al*. [97] refined the annotations and updated more difficult queries for them in 2018; the refined datasets are called ROxford and RParis.

We choose ROxford and RParis because they are among the most popular image retrieval benchmarks. Many well-known image retrieval methods are evaluated on them, from the traditional methods like RootSIFT [11], VLAD [55], and ASMK [130], to the recent deep learning based methods like R-MAC [130], GeM [98], and DELF [88].

These datasets are the ideal data sources for our pixel-retrieval, thanks to several properties. Firstly, compared to other famous datasets like image matching Phototourism [56] and dense matching Megadepth [67], the positive image pairs in ROxford and RParis have severe viewpoint changes, occlusions, and illumination changes. The new queries added by Radenovic *et al*. [97] have cropped regions that cause extreme zooms with the positive database images. These properties make the ROxford and RParis notoriously difficult. Secondly, each query image contains up to hundreds of positive database images, while other datasets, such as UKBench [87] and Holiday [54], only have 4 to 5 positive images for each query. A large amount of challenging positive images are suitable for evaluating fine-grained recognition ability.

The Google Landmark Dataset (GLD) [143] encompasses more landmarks than ROxford and RParis. However, ROxford and RParis outshine GLD in labeling quality. Notably, they stand as distinct benchmarks for contrasting machine and human recognition prowess.

It is known that people cannot easily recognize an object if it changes its pose significantly [94], but we do not know where the limit is. ROxford and RParis are **the only existing datasets that can reflect the human ability to identify objects** in the landmark domain to the best of our knowledge. Every positive image in ROxford and RParis is checked by five annotators independently based on the image appearance, and all the unclear cases are excluded [97]. This kind of annotation has two benefits. Firstly, although these benchmarks are difficult, the positive images are guaranteed to be identifiable by people without considering any contextual visual information [97]. This shows the possibility of enabling the machine to recognize these positive images by only analyzing the visual clue in the given query-index image pair. Secondly, these datasets can be used to compare human and machine recognition performance; human-level recognition performance should identify all the positive images. Although the classification performance (the top 5 accuracy) of machines on ImageNet has surpassed that of humans [105], the SOTA identification ability about the first-seen objects in ROxford and RParis is still far from human-level [63, 6, 21].

### 2.2.2 From image retrieval to pixel retrieval

In a similar spirit that semantic segmentation works as an extension of classification and provides pixel-level category information to the machines, pixel retrieval is an extension of image retrieval. It offers information about which

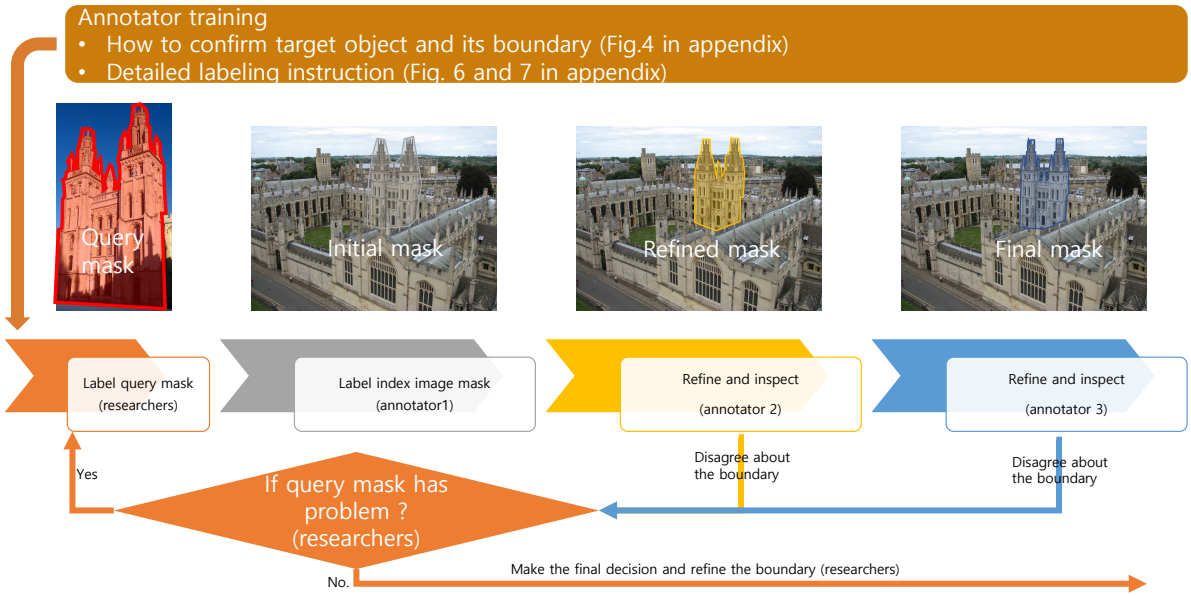



**Figure 2.2:** Labeling process.

pixels or regions are related to the query object. This task is very helpful when only a small region of the positive image corresponds to the query. Such situations frequently happen in many image retrieval applications, such as web search [97, 61, 70], medical image analysis [80, 20, 149], geographical information systems [155, 160, 116], and so on. We discuss the related applications in Section 2.3. Distinguishing and segmenting the first-seen objects is also one basic function of human vision system [120]; it is meaningful to understand and automate this ability.

Some previous works also noticed the importance of localizing the query object in the searched image. They have tried to combine image search and object localization [61, 70, 114]. However, due to the lack of a challenging pixel retrieval benchmark, they show only the qualitative result instead of the quantitative performance. Pixel-level labeling and quality assurance are arduous. In this work, 5,942 images are labeled, refined, and checked by three professional annotators. We hope this benchmark can boost and encourage future research on pixel-level retrieval.

We also compare our pixel-retrieval benchmark with segmentation, image matching, and dense matching benchmarks in the supplementary material.

### 2.2.3 Pixel-level annotation

**Images to annotate.** ROxford and RParis each contains 70 queries. The 70 queries are divided into 26 and 25 query groups in ROxford and RParis, respectively, based on the visual similarity; queries in the same query group share the same ground truth index image list. There are total 1,985 and 3,957 images to annotate for our PROxford and PRParis, respectively.

**Mask annotation.** Figure 2.2 shows our labeling process. Researchers with a computer vision background first annotate the target object in each query image. Each annotator for our new benchmark observes all the queries with masks in a query group and labels the segmentation mask for the images in the ground-truth list. Annotators are asked to identify the query object in the labeling image first and then label all the pixels depicting the target object. We show the query masks and the labeling instruction details in the supplementary materials.

**Objectivity.** To ensure the pixel retrieval task and our benchmark are objectively defined, we adopt two approaches. Firstly, we use query masks to distinctly identify the target objects and segregate them from the background (*e.g.,* the sky), occlusions (*e.g.,* other buildings), and the remaining part of the same building if the object is only a small part of it. These masks guide the removal of background and indicate the query boundary. Secondly, by examining the query with masks, our annotators reach a consensus on the target object and its boundary, thereby avoiding disagreement about our query intention. This consensus-based approach is a common method for reducing subjectivity in recognition tasks; it is also employed in the original ROxford and RParis benchmarks, where voting is used to determine the final ground truth for each query [97].

We retain small-sized occlusion objects, like windows and fences, during annotation. While this may involve subjective judgments regarding what qualifies as a small-sized occlusion, it is worth noting that well-known semantic segmentation datasets like VOC [34] and COCO [69] also involve subjective elements, such as identifying objects on a table as a table or the background behind the bike wheel as a bike. Such subjectivities are inevitable, given the difficulty of removing them. Nonetheless, they do not diminish the usefulness of benchmarks as reliable metrics for evaluating state-of-the-art methods. We include in the supplementary materials our mask rules, all the queries with masks, and our consensus checking.

**Quality assurance.** To improve the annotation quality, every query-index image pair labeling is performed by three professional annotators following the three steps: 1) anno- tate; 2) refine + inspect; 3) refine + insp, as shown in Figure 2.2. The three annotators are aged between 26 to 32 years old and have worked on annotation full-time for over 2 years. Their works have been qualified in many annotation projects.

### 2.2.4 Evaluation metrics

**Pixel retrieval from the database.** Pixel retrieval aims to search all the pixels depicting the query object from the large scale database images. An ideal pixel retrieval algorithm should achieve the image ranking, reranking, localization, and segmentation simultaneously. To the best of our knowledge, there is no existing pixel retrieval metric yet. Detection and segmentation tasks usually use mIoU and mAP@50:5:95 as the standard measurement [103]. Image retrieval methods commonly use mAP as the metric [97]. We combine them to evaluate the ranking, localization, and segmentation performance in pixel retrieval. Each ground-truth image in the ranking list is treated as a true-positive (TP), only if its detection or segmentation Intersection over Union (IoU) is larger than a threshold $n$. The other process of calculating AP and mAP follows the traditional image search mAP. Note that the mAP calculation methods in image search and traditional segmentation [34] are different; image search focuses more on ranking. Similar to detection and segmentation fields, the threshold $n$ is set from 0.5 to 0.95, with step 0.05. The average of scores under these thresholds are the final metric mAP@50:5:95. It is desirable to report both detection and segmentation mAP@50:5:95 for the methods that can generate pixel-level results; high segmentation performance does not necessarily lead to high localization performance, as shown in Sec 2.5. We follow the medium and hard protocols in ROxford and RParis [97] with and without 1 *M* distractors.

**Pixel retrieval from ground-truth query-index image pairs.** We can use existing ranking/reranking methods and treat the remaining process as one-shot detection/segmentation. In this case, the detection or segmentation performance is evaluated using the mean of mIoU of all the queries, where mIoU is the mean of the IoUs for all the ground-truth index images. We do not consider the false pairs because the ranking metric mAP well reflects the influence of false pairs in the ranking list.

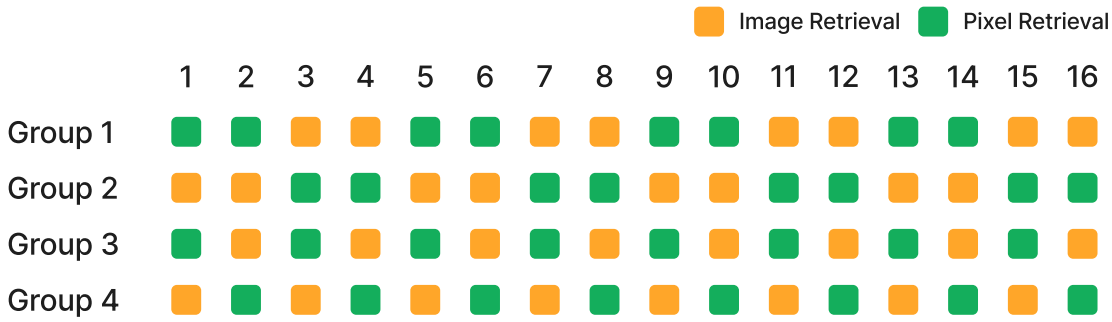



**Figure 2.3:** Design of the study on web search user experience. Image retrieval refers to a setting where no annotations are provided, whereas Pixel retrieval refers to a setting where pixel-level annotations are provided. 40 participants were divided into four groups and we counterbalanced the type of questions across the groups. Numbers 1 to 16 indicate the 16 questions.

## 2.3 Applications of pixel retrieval

Pixel retrieval requires the machine to recognize, localize, and segment a particular first-seen object, which is one of the fundamental abilities of the human visual system. It is useful for many applications. In this section, we first show that it can significantly improve the user experience in web search. We then discuss how pixel retrieval can help image-level ranking techniques. Finally, we introduce some other applications that may also benefit from pixel retrieval.

**Web search user experience improvement.** Modern image retrieval techniques focus on improving the image-level ranking performance of hard cases, such as images under extreme lighting conditions, novel views, or complicated occlusions. However, users may not easily perceive a hard case as a true positive, even if it is at the top of the ranking list. We claim that pixel-level annotation can significantly improve the user experience on the web search application.

To see how pixel-level annotation improves the user experience on image search, we ran a user study where users were asked to find images that contain a given target among candidate images in two different conditions; the one with pixel-level annotations (*i.e.,* Pixel retrieval) and the other with no annotations (*i.e.,* Image retrieval). We recruited 40 participants on Prolific[1] and compared the time taken to complete the task between the two conditions.

Participants were asked to complete 16 questions in total, where eight of them were Pixel retrieval and the other eight were Image retrieval. We divided the participants into four groups and counterbalanced the type of questions (Figure 2.3). For each question, participants were given a query image and 12 candidate images. There were three true positives and nine false positives in the candidate images, and we randomly chose ground truth images of other queries as false positives. We shuffled the order of the candidate images and asked participants to choose three images that contain the query image (*i.e.,* true positives) among them. Figure **??** shows one of the 16 questions. You can check our user study from this link. To start the user study, please enter any character into the "unique Prolific ID" blank. Anonymity is guaranteed.

Our results show that participants completed the task faster when the pixel-level annotations were presented (mean=37.07s, std=49.76s) than when no annotations were presented (mean=53.71s, std=80.08s). The difference between two conditions is statistically significant (T-test, p-value=0.00091), and participants responded that it was helpful to see annotations in completing the task (mean=6.375/7, std=0.89).

[1] prolific.co

**Other applications.** Image retrieval techniques have been applied to many applications, such as medical diagnosis and geographical information systems (GIS). The pixel-level retrieval is also desirable for these applications. For example, the size of medical and geographical images are usually huge, and the doctors and GIS experts are interested in retrieving regions of the particular structures or landmarks from the whole images in the database [149, 20, 80, 160].

Pixel retrieval can also help image matting [141, 148, 153]. Current image matting techniques rely on the user's click to confirm the target matting region [141, 148, 153]. Our pixel retrieval provides a new interaction method: deciding the target object based on the query example. This query-based interaction can significantly reduce user effort in situations where many images depict the same object [115].

## 2.4 Experiment

We evaluate the performance of state-of-the-art (SOTA) methods in multiple fields on our new pixel retrieval benchmarks. Our new pixel retrieval task is a visual object recognition problem. It requires the search engine to automate the human visual system's ability to identify, localize, and segment an object under illumination and viewpoint changes. It can be seen as a combination of image retrieval, one-shot detection, and one-shot segmentation. We introduce these related tasks and their SOTA methods in this section, and we implement these SOTA methods and discuss their results in Section 2.5.

### 2.4.1 Localization in retrieval

Some pioneering works [61, 70, 114] in image retrieval emphasized the importance of localization and tried to combine the retrieval and detection methods. However, due to the lack of a standard pixel retrieval benchmark, these pioneering works only showed qualitative results instead of quantitative comparisons. In this paper, we implement and compare the SOTA localization-related retrieval methods on our new benchmark dataset. They can be divided into two categories: spatial verification (SP) and detection.

SP [114, 79, 11, 88, 21] is one of the most popular reranking approaches in image retrieval. It is also known as image matching [56]; SP and stereo task in Image Matching Challenge (IMC) [56] share the same pipeline and theory except for the final evaluation step. In this work, we selected the local features and matching hyperparameters with the best retrieval performance on ROxford and RParis, which contain more challenging cases than datasets in IMC.

SP compares the spatial configurations of the visual words in two images. Theoretically, it can achieve verification and localization simultaneously. However, the image-level ranking performance cannot fully reflect the SP accuracy or localization performance. In the hard positive cases, *e.g.*, where many repeated patterns exist in the background, even though SP generates a high inlier number and ranks an image on top of the ranking list, the matched visual words can be wrong due to the repeated patterns. Our pixel retrieval benchmark can not only evaluate the localization performance, but also better reflect the SP accuracy and be helpful for future SP studies.

Researchers mainly focus on generating better local features to improve SP performance. The classical local features have SIFT [79], SUFT [14], and rootSIFT [11]. Recently, DELF [88] and DELG [21] local features, which are learned from the large landmarks training set [143], achieve the SOTA SP result. We evaluate the SP performance with SIFT, DELF, and DELG features on our new benchmark datasets in this paper.

Another localization-related image search approach is to directly apply the detection methods [61, 70, 102, 103, 137,

124, 126]. Faster-RCNN [103] and SSD detector [77] fine-tuned on a huge manually boxed landmark dataset [126] achieve the SOTA detect-related retrieval result [126]. Detect-to-retrieve (D2R) [126] uses these fine-tuned models to detect several landmark regions for a database image and uses aggregation methods like the Vector of Locally Aggregated Descriptors (VLAD) [55] and the Aggregated Selective Match Kernel (ASMK) [127] to represent each region. To better check the effect of the aggregation methods, we also implement the Mean aggregation (Mean), which simply represents each region using the mean of its local descriptors. The region with highest similarity can be seen as the target region for a given query. We evaluate the combination of different detectors and aggregation methods on our pixel retrieval benchmarks.

### 2.4.2 One-shot detection and segmentation

We can treat pixel retrieval as combining image retrieval and one-shot detection and segmentation. We test the performance of these approaches.

The Vision Transformer for Open-World Localization (OWL-ViT) [83] is a vision transformer model trained on the large-scale 3.6 billion images in LiT dataset [152]. It has shown the SOTA performance on several tasks including one-shot detection. The One-Stage one-shot Detector (OS2D) combines and refines the traditional descriptor matching and spatial verification pipeline in image search to do the one-shot detection. It achieves impressive detection performance in several domains, *e.g.*, retail products, buildings, and logos. We test these two detection methods on our new benchmarks.

The Hypercorrelation Squeeze Network (HSNet) [82] is one of the most famous few-shot segmentation methods. It finds multi-level feature correlations for a new class. The Mining model (Mining) [150] exploits the latent novel classes during the offline training stage to better tackle the new classes in the testing time. The Self-Support Prototype model (SSP) [35] generates the query prototype in testing time and uses the self-support matching to get the final segmentation mask. The self-support matching is based on one of the classical Gestalt principles [60]: pixels of the same object tend to be more similar than those of different objects. It achieves the SOTA few-shot segmentation results on multiple datasets. We evaluate these three methods on our new pixel retrieval benchmarks.

### 2.4.3 Dense matching

Different from image matching (SP in this paper), which calculates the transformation between two images of the same object from different views, dense matching focuses on finding dense pixel correspondence. We check if we can use the SOTA dense matching methods to correctly find the correspondence points for pixels in the query image and achieve our pixel retrieval target.

GLUNet [134] and RANSAC-flow [112] are popular among many famous dense matching methods. Recently, Truong *et al*. have shown that the warp consistency objective (WarpC) [136] and the GOCor module [133] can further improve the performance and achieve the new SOTA. Another popular method is PDC-Net [135]. It can predict the uncertainty of the matching pixels. The uncertainty can be useful for our pixel retrieval task, which is sensitive to the outliers. We test the origin GLUNet, GLUNet with WarpC (WarpC-GLUNet), GLUNet with GOCor module (GOCor-GLUNet), and PDC-Net in Table 2.1.

### 2.4.4 Experiment detail

We try our best to find the best possible result for each method on our novel benchmark. The retrieval localization methods employed in this study, including image matching (SP in this paper) and D2R, were configured to achieve

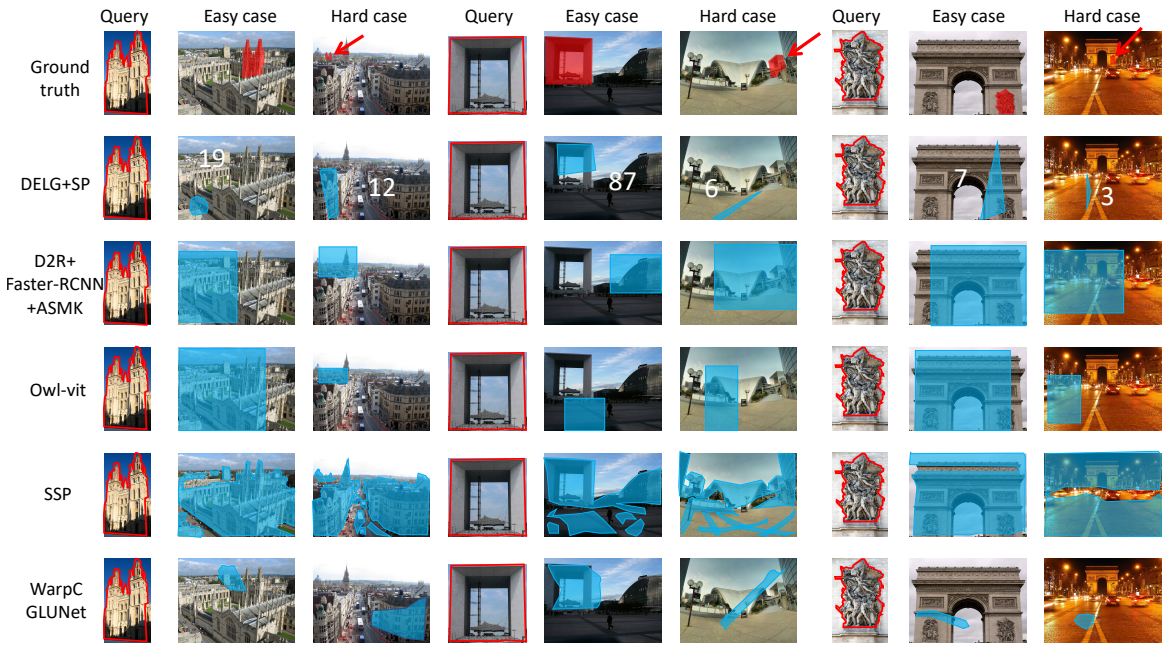



**Figure 2.4:** Qualitative comparison of the SOTA methods in different fields on the pixel retrieval benchmarks. Blue masks represent the prediction results of each method. For SP and WarpCGLUNet, we consider the union of all the matching points as the prediction masks. We also show the inlier numbers for the SP method. Pixel retrieval is challenging for existing methods and further research is needed.

optimal performance on ROxford and RParis. These methods rely on precise localization to enhance image retrieval performance. Thus, we adopt the same experimental configurations in our similar pixel retrieval benchmark. Similarly, dense matching methods, which encompass geometric and semantic matching tasks, are expected to operate directly on our pixel retrieval benchmark, as per task definitions. We evaluate its geometric models with the best performance on MegaDepth [67] and ETH3D [111], datasets that feature actual building images, rendering them the ideal valid sets for our benchmark. The difference is that our dataset contains more extreme viewpoints and illumination changes. Moreover, we evaluate the performance of semantic models to see if including semantic information can enhance rigid body recognition in our benchmarks. We refrained from fine-tuning the segmentation methods as there is no segmentation training set pertaining to the building domain to the best of our knowledge. Our comprehensive experimental findings can be employed as baseline metrics for future comparisons. We include the detailed experimental configurations for each method in the supplementary materials and intend to make them, along with their codes, publicly available.

## 2.5 Results and discussion

We report the results of pixel retrieval from ground-truth image pairs (mean of mIoU) for all the above mentioned methods in Table 2.1. We choose one to two representative methods for each field and show their qualitative results in Figure 2.4. To evaluate the performance of pixel retrieval from database, we combine these methods with SOTA image level ranking and reranking methods: DELG and hypergraph propagation (HP) [6]. We show their final mAP@50:5:95 in Table 2.2.

Although SP achieves impressive image-level retrieval results [21, 88], it shows suboptimal performance on pixel retrieval. We observe some true positive pairs where SP gives a high inlier number but matches the wrong regions.

**Table 2.1:** Results of pixel retrieval from ground truth query-index image pairs (% mean of mIoU) on the PROxf/PRPar datasets with both Medium and Hard evaluation protocols. D and S indicate detection and segmentation results respectively. **Bold** number indicates the best performance, and underline indicates the second one.

| Method | Medium | | | | Hard | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | PROxf | | PRPar | | PROxf | | PRPar | | avg. |
| | D | S | D | S | D | S | D | S | |
| Retrieval and localization unified methods | | | | | | | | | |
| SIFT+SP [91] | 26.1 | 10.9 | 24.2 | 9.7 | 18.2 | 7.3 | 19.3 | 7.8 | 15.44 |
| DELF+SP [88] | 43.7 | 20.0 | **40.7** | 16.7 | 33.2 | 13.9 | 32.2 | 12.4 | 26.60 |
| DELG+SP [21] | **44.1** | 19.7 | 40.1 | 16.5 | **34.8** | 14.5 | 31.2 | 11.7 | 26.57 |
| D2R [126]+Resnet-50-Faster-RCNN+Mean | 20.2 | - | 29.6 | - | 16.7 | - | 27.4 | - | - |
| D2R [126]+Resnet-50-Faster-RCNN+VLAD [55] | 25.8 | - | 37.5 | - | 21.6 | - | 35.5 | - | - |
| D2R [126]+Resnet-50-Faster-RCNN+ASMK [128] | 26.3 | - | 38.5 | - | 21.6 | - | **35.6** | - | - |
| D2R [126]+Mobilenet-V2-SSD+Mean | 19.7 | - | 25.9 | - | 20.1 | - | 27.9 | - | - |
| D2R [126]+Mobilenet-V2-SSD+VLAD [55] | 23.1 | - | 33. | - | 20.9 | - | 33.6 | - | - |
| D2R [126]+Mobilenet-V2-SSD+ASMK [128] | 22.4 | - | 34.0 | - | 20.8 | - | 33.1 | - | - |
| Detection methods | | | | | | | | | |
| OWL-VIT (LiT) [83] | 11.4 | - | 18.0 | - | 6.3 | - | 15.0 | - | - |
| OS2D-v2-trained [89] | 10.5 | - | 13.7 | - | 11.7 | - | 14.3 | - | - |
| OS2D-v1 [89] | 7.0 | - | 8.5 | - | 8.7 | - | 9.2 | - | - |
| OS2D-v2-init [89] | 13.6 | - | 15.4 | - | 14.0 | - | 15.1 | - | - |
| Segmentation methods | | | | | | | | | |
| SSP (COCO) + ResNet50 [35] | 19.2 | 34.5 | 31.1 | 48.7 | 15.1 | 25.3 | 29.8 | **41.7** | 30.68 |
| SSP (VOC) + ResNet50 [35] | 19.7 | 34.3 | 31.4 | **48.8** | 16.1 | **26.1** | 30.3 | 40.4 | 30.89 |
| HSNet (COCO) + ResNet50 [82] | 23.4 | 32.8 | 37.4 | 41.9 | 21.0 | 25.7 | 34.7 | 36.5 | 31.67 |
| HSNet (VOC) + ResNet50 [82] | 21.0 | 29.8 | 31.4 | 39.7 | 17.1 | 23.2 | 29.7 | 34.9 | 28.35 |
| HSNet (FSS) + ResNet50 [82] | 30.5 | **35.7** | 39.4 | 40.2 | 22.7 | 25.1 | 34.7 | 32.8 | **32.64** |
| Mining (VOC) + ResNet50 [150] | 18.3 | 30.5 | 29.6 | 42.7 | 15.1 | 21.4 | 28.1 | 34.3 | 27.50 |
| Mining (VOC) + ResNet101 [150] | 18.1 | 28.6 | 29.5 | 40.0 | 14.2 | 20.4 | 28.2 | 34.4 | 26.68 |
| Dense matching methods | | | | | | | | | |
| GLUNet-Geometric [134] | 18.1 | 13.2 | 22.8 | 15.2 | 7.7 | 4.6 | 13.3 | 7.8 | 12.84 |
| PDCNet-Geometric [135] | 29.1 | 24.0 | 30.7 | 21.9 | 20.4 | 15.7 | 20.6 | 12.6 | 21.87 |
| GOCor-GLUNet-Geometric [133] | 30.4 | 26.0 | 33.4 | 25.6 | 20.8 | 16.0 | 19.8 | 13.3 | 23.16 |
| WarpC-GLUNet-Geometric (megadepth) [136] | 31.3 | 25.4 | 36.6 | 27.3 | 21.9 | 15.8 | 26.4 | 17.3 | 25.25 |
| GLUNet-Semantic [134] | 18.5 | 14.4 | 22.4 | 15.6 | 8.7 | 5.6 | 12.8 | 7.8 | 13.22 |
| WarpC-GLUNet-Semantic [136] | 27.5 | 21.4 | 36.8 | 25.7 | 18.5 | 11.9 | 28.3 | 17.6 | 23.46 |

**Table 2.2:** Results of pixel retrieval from database (% mean of mAP@50:5:95) on the PROxf/PRPar datasets and their large-scale versions PROxf+1M/PRPar+1M, with both Medium (M) and Hard (H) evaluation protocols. **Red** indicates the best performance and **Bold** indicates the second best performance.

| | | PROxf | | PROxf+R1M | | PRPar | | PRPar+R1M | | Overhead per 100 pairs |
|---|---|---|---|---|---|---|---|---|---|---|
| | | M | H | M | H | M | H | M | H | |
| Image retrieval: DELG initial ranking [21] + HD reranking [6] | | | | | | | | | | |
| Pixel retrieval methods | DELG + SP [21] | **39.6** | **30.5** | **36.0** | 28.2 | **34.8** | 20.2 | **34.7** | 19.5 | 41.22s |
| | D2R+Faster-RCNN+ASMK [126] | 30.1 | 23.5 | 30.5 | 22.0 | 26.3 | 25.3 | 25.7 | 24.9 | 0.11 s |
| | OWL-VIT [83] | 12.3 | 6.6 | 12.1 | 13.6 | 7.9 | 7.6 | 7.9 | 7.8 | 296.21s |
| | SSP [35] | **33.0** | 29.7 | **35.7** | **30.5** | **46.4** | **37.2** | **45.6** | **37.2** | 62.33 s |
| | WarpCGLUNet [136] | 31.2 | **32.6** | 31.5 | **31.7** | 34.1 | **27.3** | 34.3 | **28.1** | 181.64s |

For example, in the first easy case in Figure 2.4, SP with DELG features generates 19 inliers, but none of the inliers are in the target object region. Note that 19 inlier number is high and only 4 false positive images are ahead of the this easy case in the final DELG reranking list [21]. This is not to say DELG is bad; in fact, its matching results are quite good in most cases. We choose this striking example only to show that the image-level ranking performance is not enough to reflect the SP accuracy. Our pixel retrieval benchmarks can be used to evaluate the matched features' locations of SP.

For SP, both deep-learning features DELF and DELG significantly outperform the SIFT features. Interestingly, although DELG shows better image retrieval performance [21] than DELF, it is slightly inferior to DELF in the pixel retrieval task. One reason might be that though DELG generates more matching inliers for the positive pairs than DELF, these inliers tend to exist in a small region and do not reflect the location or size of the target object. Improving SP performance in both image and pixel level can be a practical research topic.

Although the detect-2-retrieval [126] is inferior to SP in image retrieval [21, 88, 97], it shows better performance than SP in our pixel-level retrieval benchmarks. We conjecture that the detection models tend to cover the whole building more than SP. Our benchmark is helpful in checking this conjecture and designing a better pixel retrieval model for future works. The results of the region detector and the aggregation method are similar to the trend in image search [126]. The VLAD and ASMK aggregation methods significantly improve the Mean aggregation. A faster-RCNN-based detector shows better performance than SSD.

For dense matching methods, GLU-Net using warp consistency or GOCor module and PDC-Net show better results than other models. This trend is similar to that in the dense matching benchmark Megadepth [67].

The segmentation methods significantly outperform other methods in terms of the mean of segmentation mIoU. However, their detection mIoU results are not so impressive. They tend to predict the entire foreground, which contains the target building, as shown in the SSP line of Figure 2.4. Among the segmentation methods, SSP shows better segmentation than others, showing its self-support approach is helpful for finding more related pixels.

Another interesting finding is that better image ranking mAP does not necessarily brings better pixel retrieval mAP@50:5:95, as shown in Table 2.2. The reason might be that the image search techniques rank some hard cases high, but detection methods do not well localize the query object in them.

It is interesting to note that segmentation and dense matching methods have demonstrated superior mIoU results compared to matching-based and detection-based retrieval methods, despite not being originally designed for retrieval tasks. However, to effectively tackle the pixel retrieval task, these methods must work in conjunction with

image search techniques. While dense matching and segmentation methods are better suited for identifying target object areas, they may not achieve fine-grained recognition. In contrast, existing retrieval methods tend to identify certain textures or corners but lack the ability to capture the entire object's shape. Without a reliable benchmark, retrieval methods may simply associate an object and its context to improve image-level performance, leading to low localization and segmentation results, as we discussed above. We did our best to prepare our new benchmark so that it can provide a valuable evaluation for novel methods targeting pixel retrieval, which requires fine-grained and variable-granularity detection and segmentation. Moreover, we find pixel retrieval challenging. The current best mAP@50:5:95 in PROxford and PRParis at medium setting without distractors are only 37.3 and 47.0.

## 2.6 Future works

We present a novel task termed "Pixel Retrieval." This task mandates segmentation but transitions from a semantic directive to the content-based one, thus bypassing semantic vagueness. Concurrently, it demands large-scale, instance-level recognition—a subject frequently explored by the retrieval community. This innovative task poses several unique challenges, some of which we outline below:

### 2.6.1 Enhancing accuracy

For a superior user experience, it's vital to embrace methods, workflows, and datasets that bolster accuracy. Our findings illustrate that segmentation and dense matching methods are beneficial, especially when an image ranking list is provided using existing retrieval techniques. Beyond merely superimposing segmentation over retrieval, a compelling approach would be to rank images based on the results of the segmentation. Further insights and experimental outcomes in this regard are available on our website.

Although the introduction of new datasets, even those echoing the landmarks in our benchmarks, is commendable, it's pivotal to articulate their application to discern the sources of performance enhancements. If PROxford/PRParis and ROxford/RParis are employed as benchmarks, it's crucial to ensure the consistent usage of the same training set. Given the public accessibility of our ground truth files, it's imperative to prevent any unintended data leaks during training.

### 2.6.2 Scalability and speed

A major challenge lies in scaling the algorithms and augmenting the retrieval speed. Techniques like segmentation and dense matching, which compute for every pair, inherently lag in speed when compared to retrieval methods such as ASMK and D2R. Therefore, swift methods that can cater to extensive scales are highly sought after.

### 2.6.3 Innate visual recognition and The significance of training data

The prevalent trend in research is to amass expansive training or fine-tuning sets closely aligned with test instances—certainly a commendable approach. However, intriguingly, humans exhibit an innate ability to discern instances in query images. Our annotators, despite being unfamiliar with European landmarks, could effortlessly segment target objects in each positive image, even when subjected to extreme lighting and perspective alterations. What fuels this innate recognition? Is it purely due to extensive prior exposure, or are there underlying mechanisms at play? How pivotal is the training dataset in replicating human-like content-based segmentation, especially when semantic influences are excluded? These questions beckon exploration.

## 2.7 Conclusion

We introduced the first landmark pixel retrieval benchmark datasets, *i.e.,* PROxford and PRParis, in this paper. To create these benchmarks, three professional annotators labeled, refined, and checked the segmentation masks for a total of 5,942 image pairs. We executed the user study and found that pixel-level annotation can significantly improve the user experience on web search; pixel retrieval is a practical task. We did extensive experiments to evaluate the performance of SOTA methods in multiple fields on our pixel retrieval task, including image search, detection, segmentation, and dense matching. Our experiment results show that pixel retrieval is challenging and further research is needed.

# Chapter 3. Reflector, Connector, and Controller: Enhancing Large Multimodal Models for Fine-Grained Domains

## 3.1 Introduction.

Large Multimodal Models (LMMs), such as GPT-4V [3], Gemini-1.5 [125], and Claude-3 [9], have demonstrated remarkable advancements in visual understanding. However, recent studies [154, 147] reveal a notable performance gap between LMMs and their vision encoders in image classification. This gap arises due to the limited availability of vision instruction tuning data. For instance, on the ImageNet [29] classification task with 1,000 categories, GPT-4V achieves an accuracy of only 60.6%, significantly trailing behind CLIP-based models such as CLIP-L [99], which achieve 74.8%. This performance disparity is critical as image classification serves as the foundation for numerous advanced vision-language tasks.

To address this challenge, we decompose image classification within LMMs into two stages: initial inference and reflection. While LMMs often struggle to accurately infer the correct category in the initial stage, we find that they excel in verifying whether the initial inference aligns with the expected category—a process we term "reflection." While the concept of reflection has been explored in text-based decision-making and programming [118], its application in image recognition remains largely untapped.

To investigate why LMMs can perform reflection, we analyze their vision token representations. Specifically, we design a reverse embedding layer based on the linear text embedding layer to map vision tokens to their corresponding textual descriptors. Applying this reverse embedding layer to the vision tokens reveals that the key textual terms used by the LMM to describe an image are directly encoded in these vision tokens.

To validate this, we replace the original image input with the extracted key textual descriptors and observe that the LMM generates similar responses. This demonstrates that vision token features encapsulate much (though not all) of the detailed visual information in an explicit and interpretable manner, rather than relying solely on implicit representations.

This insight sheds light on the limitations of LMMs in classification tasks: the connector module fails to generate vision tokens that capture the fine-grained descriptions required for accurate classification. However, LMMs demonstrate strong reflection capabilities because their vision tokens encode descriptive terms from multiple perspectives, enabling the model to reason effectively based on these conceptual associations.

Inspired by this observation, we propose a training-free connector that seamlessly integrates several fine-tuned vision encoder across various domains. To manage this multi-domain integration, we introduce the Frontal Module, which dynamically controls the flow of information between vision encoders and the LMM. The Frontal Module determines when to adopt domain-specific vision encoder outputs and selectively applies reflection for uncertain predictions, thereby enhancing the overall performance.

In summary, our contributions are as follows:

1. Bridging the Vision-Language Gap: We address the critical challenge of aligning vision encoders with LMMs, a limitation highlighted in prior research.

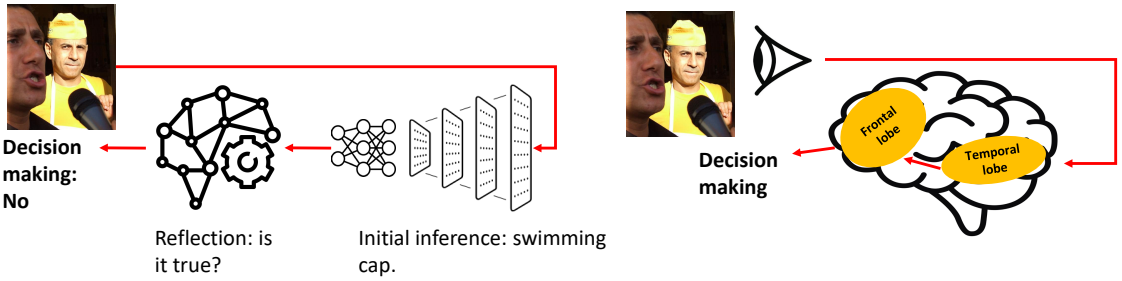



**Figure 3.1:** We divide the classification process into two stages: (1) initial inference and (2) reflection. The initial inference uses a conventional CNN- or ViT-based approach with a linear or MLP classifier. In the reflection stage, the model revisits the initial predictions to verify or correct potential misidentifications, mimicking human cognition, where the temporal lobe supports initial recognition and the frontal lobe facilitates reflective judgment.

2. Reflection for Improved Accuracy: We leverage the reflection capabilities of LMMs to improve classification accuracy, demonstrating for the first time that language models can assist in image recognition tasks.

3. Frontal Module for Dynamic Control: We propose the Frontal Module, a novel approach to managing multi-domain vision encoders and selectively enabling reflection, achieving state-of-the-art performance without additional training

## 3.2 Reflector: LMM itself.

### 3.2.1 LMMs for reflection.

Large Multimodal Models (LMMs) like GPT-4V [3], Gemini-1.5 [125], and Claude-3 [9] has demonstrated impressive visual understanding capabilities. However, recent studies highlight a key limitation: due to insufficient vision instruction tuning data when integrating the vision encoder and language model, LMMs struggle to effectively decode the information captured by the vision encoder [154, 147]. Consequently, current LMMs have difficulty recognizing fine-grained categories and, despite their significantly larger parameter counts, often underperform compared to widely used vision encoders like CLIP [99]. For instance, when tasked with selecting the most likely category from 1,000 candidates on the ImageNet [29] dataset, GPT-4V achieves only 60.6% accuracy, whereas CLIP-based models like CLIP-L reach 74.8%. This gap is crucial, as effective image classification is foundational for enabling more advanced VLM capabilities, such as visual question answering and reasoning.

While increasing the amount of visual instruction tuning data could help address this issue [154], such data is costly to collect, and training LMMs is resource-intensive [75, 131]. A training-free approach that can bridge the recognition gap between an LMM and its vision encoder—or even enhance the recognition performance of the vision encoder—would be highly desirable.

Zhang *et al.* [154] conducted valuable and extensive experiments on various training-free strategies, such as prompt engineering (e.g., adding "let's think step by step"), reducing label size (e.g., reducing ImageNet's 1000 categories to 100 or 20 while ensuring the ground truth remains), and modifying the inference algorithm (e.g., having LMMs predict the probability of each candidate category rather than directly generating the class name). While these approaches narrowed the performance gap between LMMs and CLIP models, the gap persists. Ultimately, improving the image recognition ability of LMMs without additional fine-tuning remains an open challenge.

To achieve this goal, we first divide the classification process into two stages: 1) initial inference and 2) reflection,

as shown in Figure 3.1. In the initial inference stage, the vision encoder predicts the category directly, bypassing the language model. The reflection stage then revisits this prediction to verify its accuracy.

Although LMMs struggle with quickly inferring the correct category from a set of candidates, we find that they excel in the reflection stage. While LLM reflection has been explored in tasks like text-based decision making and programming [118], to our knowledge, there are few, if any, previous works examining LMM reflection for image recognition. Our approach prompts LMMs to verify the accuracy of the initial prediction from the vision model by asking, "Does the picture show a/an {initially predicted category}?" This prompt directs the LMMs to focus on a single category at a time, in contrast to prior studies [154, 147], which often prompt LMMs to evaluate multiple categories simultaneously, using questions like "Rank the {list of categories} from most to least likely to describe the input image" or "What is the image? Choose one from {list of categories}. Let's think step-by-step."

Figure 3.2 illustrates examples where LMM reflection successfully identified incorrect initial inferences made by the vision model. LMMs appear to use common knowledge to reason and detect erroneous predictions. This common knowledge includes the familiar situations in which each category is commonly found and the definition of the object category. For instance, in the first column of Figure 3.2, a man in the background is wearing a cap. The cap resembles some training images of a bathing cap from ImageNet [29], as shown in Figure 3.3, leading ResNet50 [49] to mistakenly classify the image as a bathing cap without contextual information. However, by reasoning about the person's profession and role in the scene, GPT-4V correctly deduces that the cap is unlikely to be a bathing cap. Figure 3.2 shows both the ResNet50 + GPT-4V and CLIP-ViT-L/336px [99] + LLaVA-1.6 [75, 73] combinations. Notably, CLIP-ViT-L/336px serves as the vision encoder for LLaVA-1.6, underscoring that even with the same vision encoder, LMMs still have the potential to reflect on and identify inaccuracies in the initial inference effectively.

In the following sections, we will delve into the technical reasons behind this reflection capability and present methods to enhance overall recognition and VQA performance of LMMs across various domains. Before doing so, we introduce neuroscience insights that motivate us to divide classification into two stages.

### 3.2.2 Neuroscience insights.

Our approach of dividing classification into two stages is inspired by findings that visual stimuli are sequentially processed across multiple regions in the brain. As illustrated in Figure 3.4, visual stimuli are processed starting from the occipital lobe to frontal lobe, with the meaning and position of stimuli taking separate pathways [42, 107]. The meaning of visual stimuli is processed through the occipital and temporal lobes, which inspired key properties of CNNs, such as spatially localized receptive fields and pooling units that provide invariance to small shifts in feature positions [62, 15, 44]. The famous "grandmother cells" were discovered in the temporal lobe. These cells respond to specific concepts; they activate when a person sees a photo of their grandmother, a drawing of her, or even text referring to "grandmother" [96, 47].

In addition to the occipital and temporal lobes, neural activity is also observed in the frontal lobe during advanced visual recognition-related tasks [1], such as role-based relational reasoning [84, 139] and trial-unique visual discrimination [120]. Role-based relational reasoning [84, 139] involves drawing inferences about entities based on the roles they occupy (e.g., "go with") rather than relying on direct similarity. Trial-unique visual discrimination [120] involves distinguishing between similar but distinct objects encountered for the first time. Beyond these advanced visual recognition functions, the frontal lobe is also involved in many other executive processes, including planning, decision-making, and risk assessment [57, 12]. Notably, the size difference between the frontal lobe in humans and

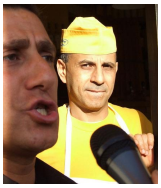
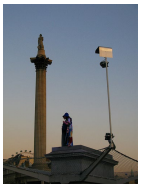
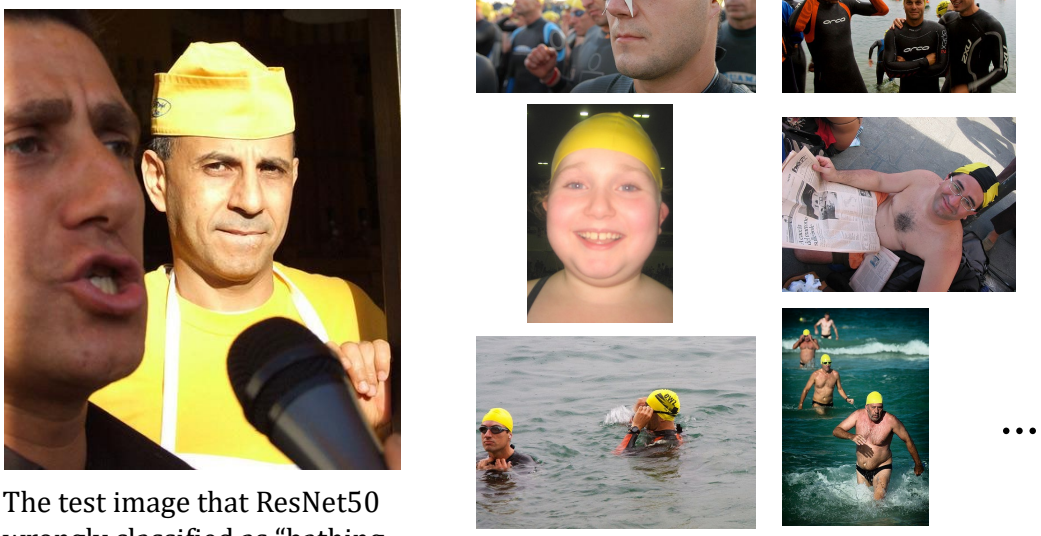
| | | | | | |
|---|---|---|---|---|---|
|  |  |  |  |  |  |
| **Ground truth:** microphone, mike | **Ground truth:** street sign | **Ground truth:** whistle | **Ground truth:** cassette | **Ground truth:** pedestal | **Ground truth:** bookcase |
| **ResNet50 prediction:** bathing cap, swimming cap | **ResNet50 prediction:** traffic light, traffic signal, stoplight | **ResNet50 prediction:** Pencil sharpener | **CLIP-ViT-L/336px prediction:** chest | **CLIP-ViT-L/336px prediction:** cloak | **CLIP-ViT-L/336px prediction:** entertainment center |
| **Prompt:** Does this picture have a "bath cap, swimming cap"? | **Prompt:** Is this a "traffic light, traffic signal, stop light"? | **Prompt:** Is it a pencil sharpener? | **Prompt:** What does this image have? Share your first impression, and then think about whether the image contains a chest? Please think about the definition of the chest, and then consider whether it is really a chest. Choose from the following options: [yes], [no], or [not sure]. | **Prompt:** What does this image have? Share your first impression, and then think about whether the image contains a cloak? Please think about the definition of the cloak, and then consider whether it is really a cloak. Choose from the following options: [yes], [no], or [not sure]. | **Prompt:** What does this image have? Share your first impression, and then think about whether the image contains an entertainment center? Please think about the definition of the chest, and then consider whether it is really an entertainment center. Choose from the following options: [yes], [no], or [not sure]. |
| **GPT-4V response:** The person in the background of the image is wearing a cap that might be mistaken for a bathing or swimming cap due to its color and shape. However, given the context – it seems to be a uniform of some sort – it's more likely to be a different type of cap, possibly related to the person's profession or role in the event being depicted. Swimming caps are typically used in aquatic environments and are made from materials like silicone, latex, or spandex to keep the hair dry and reduce drag in the water. This cap does not seem to be in such an environment. | **GPT-4V response:** No, the image does not show a traffic light, traffic signal, or stoplight. Instead, it features a stop sign, which is used to control traffic and instruct drivers to come to a complete stop before proceeding. Traffic signals or stoplights, on the other hand, are electronic signaling devices positioned at road intersections and pedestrian crossings to control the flow of traffic with colored lights (typically red, yellow/green). | **GPT-4V response:** No, the items in the picture are not pencil sharpeners. They appear to be whistle key chains based on their shape and the inclusion of a lanyard used to wear around the neck or attach to keys. Pencil sharpeners typically have a place to insert a pencil and a blade inside to shave the wood, which these items do not seem to have. | **LLaVA-1.6 response:** My first impression is that this is a nostalgic display of music media from a bygone era, likely representing the personal collection of someone who enjoys or collects cassette tapes. The container holding the tapes is a functional piece of luggage, designed to carry items securely and conveniently.<br>As for the chest, the container in the image does not fit the traditional definition of a chest, which is typically a large, enclosed storage container with a hinged lid and often made of wood or metal. The container in the image is more of a briefcase or suitcase, which is designed for carrying and transporting items rather than for storage.<br>Based on this analysis, I would choose [no] as the answer to whether the image contains a chest. The container in the image is a briefcase or suitcase, not a chest. | **LLaVA-1.6 response:** The first impression is that the person is wearing a cloak, as it drapes over their body and has a hood, which are characteristic features of a cloak. However, upon closer inspection, it's not entirely clear whether the garment is a cloak or a type of poncho or shawl. The garment does not seem to have the typical long, flowing shape of a cloak, but rather it appears to be more fitted around the person's body.<br>Based on the definition of a cloak, which is a large, loose garment that is worn over other clothing and is often used for warmth or as a fashion accessory, it is not entirely certain whether the garment in the image is a cloak. The garment does not appear to be loose or flowing, and it does not seem to be worn over other clothing.<br>Therefore, based on the analysis, I would choose [not sure] as the answer. | **LLaVA-1.6 response:** My first impression is that this is a simple, functional piece of furniture, possibly used for storing books and displaying decorative items. It does not have any visible electronic components or media storage spaces, which are typically associated with an entertainment center.<br>An entertainment center is generally defined as a piece of furniture designed to hold and display electronic devices such as televisions, gaming consoles, and audio equipment, as well as media storage like DVDs or Blu-rays. It often includes shelves, cabinets, and compartments to organize and store these items.<br>Based on the image and the definition of an entertainment center, I would say that this is not an entertainment center. It lacks the necessary features and compartments to accommodate electronic devices and media storage. Therefore, my answer is: [no]. |

**Figure 3.2:** Examples where LMM reflection successfully identified incorrect initial inferences made by the vision model. Yellow sentences highlight the reasons why the LMM deemed the initial inference incorrect. The first three columns demonstrate GPT-4V's reflection on ResNet50's initial predictions, where ResNet50's original linear classifier was used for ImageNet image inference. The last three columns show LLaVA-1.6's reflection on CLIP-ViT-L/336px's initial predictions, using the text embeddings of the 1,000 ImageNet category names as the parameters for its final linear classifier. Note that CLIP-ViT-L/336px is the vision encoder for LLaVA-1.6, highlighting that even with the same vision encoder, LMMs still have the potential to reflect on and identify inaccuracies in the initial inference effectively. Further discussion about the technical reason behind this reflection ability is provided in Section 3.3.



The test image that ResNet50 wrongly classified as “bathing cap”.

Some training images that belong to “bathing cap” category.

**Figure 3.3:** A man in the test image is wearing a cap; the cap resembles some training images of a bathing cap from ImageNet.

other animals is much larger compared to other brain regions [1]. The observation that different regions of the brain activate during various visual recognition tasks inspires us to divide recognition into two stages: initial inference and reflection, as introduced in Section 3.2.1

## 3.3 Connector: “Grandmother cells".

Previous works use extensive feature alignment data to train connectors that transform image features from vision encoders to vision tokens interpretable by the language model [65, 5, 75, 132, 101, 58]. These vision tokens are generally understood to convey visual information implicitly in the embedding space. However, in this work, we introduce a novel perspective, suggesting that these vision tokens may carry explicit text-related concepts. By leveraging only these explicit concepts—significantly reducing token size compared to the full set of vision tokens from an image—we find that the language model can still generate similar responses.

We begin by attempting to translate vision token features back into text tokens. While this might initially seem counterintuitive, it reveals some intriguing results. The text embedding layer of the language models is typically a linear projection from tokens to an embedding space: $\mathbf{y} = \mathbf{E}(\mathbf{x})$, where $\mathbf{x} \in R^{1\times v}$ is a one-hot text token vector, $\mathbf{y} \in R^{1\times e}$ is the corresponding text embedding, and $\mathbf{E} \in R^{v\times e}$ is a projection matrix, with $v$ as the vocabulary size and $e$ as the embedding dimension. Now given a vision token embedding $\mathbf{v}$, which shares same dimension with $\mathbf{y}$, we design a reverse embedding function to transfer it into a text token $\mathbf{x}$ as follows:

$$\mathbf{S} = \frac{\mathbf{v}}{\|\mathbf{v}\|_2} \cdot \left(\frac{\mathbf{E}}{\|\mathbf{E}\|_2}\right)^{\top}, \ \mathbf{x} = \text{one-hot}\left(\mathbf{S}\right). \tag{3.1}$$

Here $\mathbf{S} \in R^{1\times v}$ records the similarity between $\mathbf{v}$ and each text token. After obtaining the token $\mathbf{x}$, we use a pre-trained text tokenizer decoder to convert them back into text. As shown in Figure 3.5, when applying this process to all vision tokens from an image, the resulting text appears garbled, as expected; however, it surprisingly reveals key terms that the language model uses to inform its responses.

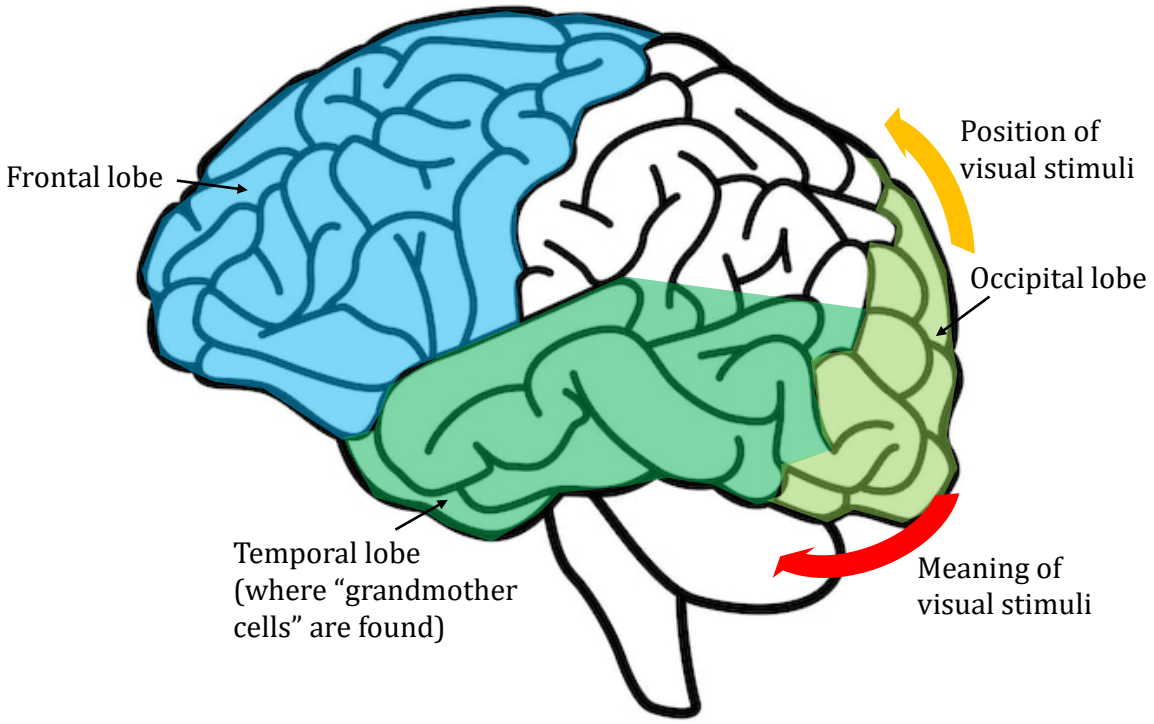



**Figure 3.4:** The human brain region that is related to visual recognition.

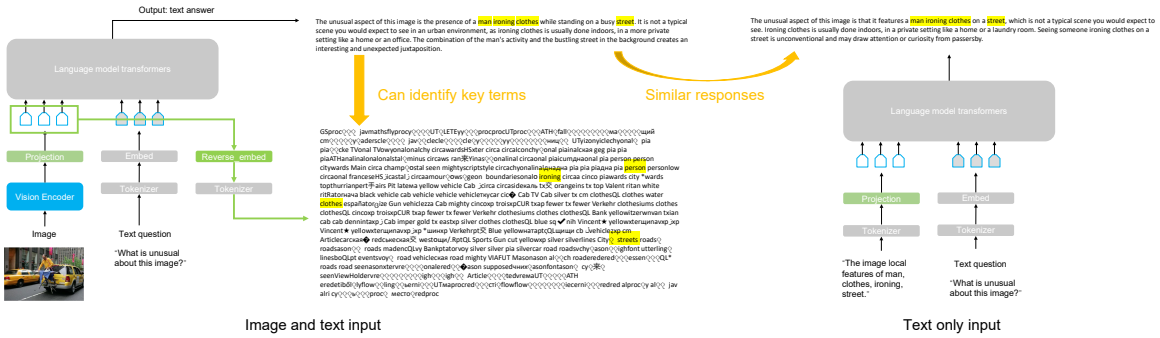



**Figure 3.5:** We implemented a reverse embedding function to convert image token features back into human-readable text. Although the resulting text appears garbled, we can still identify key terms that the language model uses to guide its responses. Interestingly, when we replace the image input with these identified text key terms, the language model generates similar answers.

Furthermore, we replace the image input with these key text terms, omitting the actual image. The key terms, formatted as "The image local features of ...," pass through the text tokenizer and embedding layer. Notably, the number of tokens is much smaller than the original vision tokens. Yet, the language model is still able to generate similar answers from when the image itself is provided as input, as shown in Figure 3.5.

While prior studies have examined the effectiveness of the linear projection layer as a connector [154, 131], none have demonstrated that image token features can align with explicit text concepts. This finding is intriguing because the vision connector is not explicitly trained to associate with specific text concepts. This phenomenon suggests a subtle resemblance to the "grandmother cells" in the brain, which activate when a person sees an image of their grandmother or encounters the word "grandmother" in text.

This observation offers insight into our findings in Section 3.2. LMMs may struggle with direct classification because the projection layer doesn't produce enough fine-grained, relevant terms needed for classification benchmarks. However, LMMs can do reflection because they receive a set of descriptive terms about the image, which the language model can then use for reasoning. For example, concepts like "bathing cap" typically don't co-occur with "uniform" or "someone speaking at a rally," enabling the model to reason effectively based on these conceptual associations.

Inspired by the observation that large-scale feature-alignment training ultimately maps image features to explicit text concepts, we propose a training-free approach that leverages more fine-grained explicit concepts to link a language model with pre-trained vision encoders, effectively capturing their detailed recognition capabilities. For any pre-trained vision encoder, once the image feature $x$ (of any dimension) is obtained, we pass it through a key matrix and value matrix. The process is as follows:

$$\hat{x} = A(W_{\text{key}}(x)), \quad y = W_{\text{value}}(\hat{x}), \tag{3.2}$$

where $A$ is an activation function, $W_{\text{key}} \in \mathbb{R}^{v \times e_x}$, and $W_{\text{value}} \in \mathbb{R}^{e_l \times v}$. Here, $e_x$ is the dimension of the image feature $x$, $v$ is the vocabulary size, and $e_l$ is the dimension of the token embedding that will be passed to the language model. $A$ can either be a softmax operation followed by one-hot encoding or the Frontal module, which we introduce in Section 3.4. Importantly, the vocabulary, as well as $W_{\text{key}}$ and $W_{\text{value}}$, can be set without additional training as following:

- Vocabulary: For a pre-trained vision encoder, we create a vocabulary with terms that align with the domain in which the vision encoder excels and the specific needs of the target task. For example, this could include ImageNet categories for image classification, landmark names when the LMM serves as a tour guide for visitors, or street names when the LMM functions as a city navigator. The size of the vocabulary can vary widely, ranging from thousands to millions of terms.

- $W_{\text{value}}$: For each term in the prepared vocabulary, we pass it through the tokenizer and token embedding layer of the language model. This results in $v$ embedding vectors, each with a shape of $1 \times e_l$, which are then set as the parameters of $W_{\text{value}}$.

- $W_{\text{key}}$: There are three methods to obtain the weights for $W_{\text{key}}$. For vision-language contrastive-trained encoders, such as CLIP, we use the text encoder embeddings of each term as the weights for $W_{\text{key}}$. For vision encoders trained with classification loss, the final layer that maps hidden states to category logits can be used as the weights for $W_{\text{key}}$. For vision encoders trained with other losses, we gather corresponding images for each category and use their vision encoder embeddings as the weights for $W_{\text{key}}$.

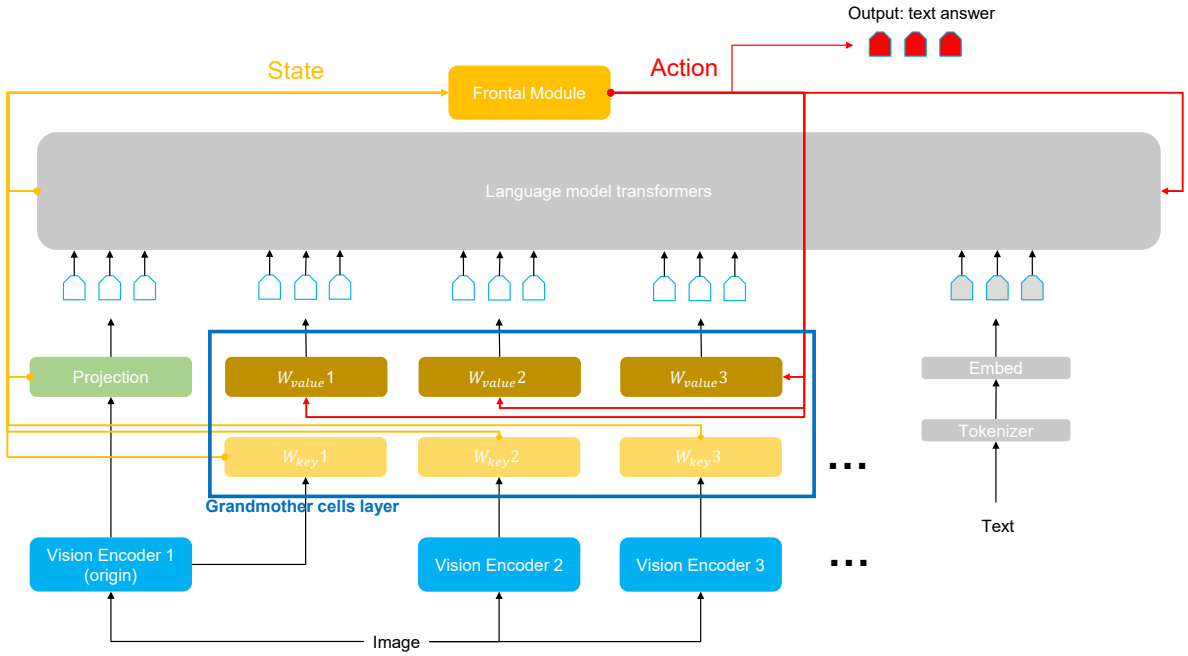



**Figure 3.6:** We implemented a reverse embedding function to convert image token features back into human-readable text. Although the resulting text appears garbled, we can still identify key terms that the language model uses to guide its responses. Interestingly, when we replace the image input with these identified text key terms, the language model generates similar answers.

We find that this training-free approach, though simple and straightforward, is both cost-efficient and effective in leveraging fine-grained domain knowledge from various pre-trained vision encoders compared with existing feature alignment methods [65, 5, 75].

## 3.4 Controller: Frontal Module.

In this section, we introduce a new module called the Frontal Module, designed to control other models within the LMM framework, as shown in Figure 3.6. The Frontal Module determines when to selectively use reflection and which outputs from additional vision encoders, connected through our grandmother cells' layer, should be included. This approach ensures that only relevant encoder outputs are integrated into the language model.

Not all vision encoders contribute meaningful information for every image; for example, encoders fine-tuned for landmark recognition might introduce noise when identifying fine-grained animal species. The Frontal Module assesses whether the output of each vision encoder should be utilized, balancing efficiency and accuracy.

Given that top-5 prediction accuracy generally surpasses top-1 accuracy, we leverage the reflection strategy we introduced in Section 3.2 to reflect on the initial top-5 predictions, selecting the most suitable outcome to improve top-1 accuracy. However, not all images benefit from LMM reflection due to the associated computational costs and the potential risk of reduced accuracy, as discussed in Section 3.5.3. To balance both accuracy and efficiency, reflection should be applied selectively to cases where initial predictions are uncertain—a decision managed by the Frontal Module.

Each vision encoder's feature is processed through $W_{\text{key}}$ in the grandmother cells' layer (Equation 3.3), producing an intermediate feature that is passed to the Frontal Module. Within the Frontal Module, this intermediate feature

first undergoes a softmax operation to generate a certainty signal for each prediction. Based on this signal, the Frontal Module makes the following decisions: if the certainty is high, the module directly accepts the vision encoder's output; if the certainty is moderate, it triggers additional LMM reflection; and if the certainty is low, it disregards the output from that vision encoder.

This decision process can be implemented using either a simple threshold or a multi-layer perceptron (MLP) trained with Proximal Policy Optimization (PPO). For PPO training, it is essential to customize the reward function based on the specific application to achieve an optimal balance. In high-stakes applications such as self-driving cars, where accuracy is paramount, prioritizing precision is critical. Conversely, in more casual scenarios, achieving a balance between accuracy and processing speed is more desirable.

## 3.5 Experiment.

We aim to investigate three key aspects through our experiments:

- Whether the final model performs well on standard fine-grained recognition and VQA benchmarks.
- Whether the proposed reflection method enhances the initial inference accuracy of the original vision encoder.
- Whether the proposed grandmother cell layer effectively integrates with other fine-tuned vision encoders in various fine-grained domains.

### 3.5.1 Performance on fine-grained recognition and VQA benchmarks.

We evaluated our model's performance using standard fine-grained benchmarks, including ImageNet [29] and ImageWikiQA [17].

For ImageNet, we evaluated our method on the entire set of 50,000 validation images, using both the original labels [29] and the revised 'Real' labels by Beyer et al. [17]. As shown in Table 3.1, our model achieved high scores, surpassing all previously reported LMM results. Impressively, this was accomplished without any additional fine-tuning on in-domain data. Our approach outperforms other inference-based methods aimed at enhancing recognition accuracy, such as prompt variation and alternative inference algorithms [154]. By employing our reflection strategy and incorporating the grandmother cells' layer, our method with LLaVA-1.6-34B outperformed the best existing proprietary LMM, GPT-4V, by 26.9%. Adding an additional ResNet-50 further improved the performance on the 'Real' labels from 77.5% to 79.0%.

We further evaluated our method on the ImageWikiQA benchmark, which consists of 2,000 questions, each paired with an image from ImageNet. The questions are generated using GPT-4 based on the Wikipedia pages of ImageNet classes. Answering questions in ImageWikiQA requires the LMM to recognize the specific types of an object instead of indentifying only the general category of an image. As shown in Table 3.2, our method improves the performance of LLaVANeXT-M7B by 15.5%, surpassing all existing public LMMs. Although our method does not match the performance of proprietary LMMs such as GeminiPro and GPT-4V, it operates with a significantly smaller model size. For example, GPT-4V has 1.76 trillion parameters [147], while our base model uses only 7 billion parameters. We think the main reason for this performance gap is that, although our LMM can accurately identify fine-grained object categories using the reflection strategy and grandmother cells' layer (notably outperforming all proprietary LMMs on ImageNet, as shown in Table 3.1), the base language model lacks sufficient domain knowledge to provide comprehensive answers to the questions.

| Model | Origin | Real |
|---|---|---|
| LMMs | | |
| BLIP2-2.7B [65] | 25.3 | N/A |
| IBLIP-7B [28] | 14.6 | N/A |
| IBLIP-13B [28] | 14.7 | N/A |
| LLaVA-1.5-7B [75] | 22.8 | N/A |
| LLaVA-1.6-7B [75] | 32.3 | N/A |
| LLaVA-1.5-13B [75] | 24.3 | N/A |
| Claude3 [9] | 53.6 | N/A |
| GeminiPro [125] | 56.0 | N/A |
| GPT4V [147] | 60.6 | N/A |
| Prompt engineering and inference algorithhms | | |
| LLaVA-1.5-7B+prompt 1 [154] | 19.7 | N/A |
| LLaVA-1.5-7B+prompt 2 [154] | 21.6 | N/A |
| LLaVA-1.5-7B+ sum tokens [154] | 34.8 | N/A |
| LLaVA-1.5-7B+ ave tokens [154] | 35.3 | N/A |
| LLaVA-1.5-7B+ CFG [154, 106] | 47.6 | N/A |
| BLIP2-2.7B+prompt 1 [154] | 27.6 | N/A |
| BLIP2-2.7B+prompt 2 [154] | 24.3 | N/A |
| BLIP2-2.7B+ sum tokens [154] | 21.0 | N/A |
| BLIP2-2.7B+ ave tokens [154] | 5.1 | N/A |
| BLIP2-2.7B+ CFG [154, 106] | 38.7 | N/A |
| Ours | | |
| Ours-LLaVA-1.5-7B (+ ResNet50) | 76.1 | 78.0 |
| Ours-LLaVA-1.6-34B | **76.9** | 77.5 |
| Ours-LLaVA-1.6-34B (+ ResNet50) | **76.9** | **79.0** |

**Table 3.1:** Performances on ImageNet origin [29] and real [17] labels. Our model surpasses all previously reported LMM results.

| Model | Acc | #parameters |
|---|---|---|
| Public LMMs | | |
| BLIP2-2.7B [65] | 21.7 | 2.7 billion |
| IBLIP-7B [28] | 36.3 | 7 billion |
| IBLIP-13B [28] | 37.5 | 13 billion |
| LLaVANeXT-V7B [74] | 37.0 | 7 billion |
| LLaVA1.5-13B [72] | 38.0 | 13 billion |
| LLaVANeXT-M7B [74] | 40.7 | 7 billion |
| Proprietary LMM | | |
| GeminiPro [125] | 49.1 | 540 billion |
| GPT4V [147] | 61.2 | 1760 billion |
| Ours | | |
| Ours-LLaVANeXT-M7B | **47.0** | 7 billion |

**Table 3.2:** Performances on ImageWikiQA. Our method outperforms all public LMMs.

### 3.5.2 The effectiveness of reflection.

Our work not only bridges the gap between the vision encoder and the LMM but also enhances the performance of the original vision encoder through LMM reflection. We evaluated the effectiveness of LMM reflection on ImageNet.

We observed that applying LMM reflection to all images directly led to a decrease in accuracy, as shown in Figure 3.8. When the LMM cannot establish clear reasoning, it can display randomness when choosing between "Yes," "Not sure," or "No," which ultimately reduces top-1 accuracy, as demonstrated in Figure 3.7. This aligns with cognitive research on intuitive versus reflective judgment [41], showing that while reflective reasoning can be more complex, intuitive judgments often yield more accurate and efficient outcomes [40].

However, with our Frontal Module, which selectively applies LMM reflection to uncertain images, performance improved significantly. As illustrated in Figure 3.8, using the Frontal Module, LMM reflection increased accuracy from 75.6 to 79.0, surpassing the baseline accuracy of the original vision encoder, CLIP-L/336p, which was 77.0. The impact was particularly pronounced for uncertain images selected for reflection; as shown in Figure 3.9, LLaVA-1.6-34B reflection improved the performance of the original ResNet-50 by an impressive 22.4%.

It is noteworthy that while CLIP-L/336p serves as the original vision encoder within the LLaVA-1.6-34B model, the reflection method consistently enhanced its baseline performance, achieving a 9.2% increase on uncertainty images as highlighted in Figure 3.9. This suggests that the observed improvements are not due to ensembling different vision encoders but are attributed to the combined effect of LMM reflection and the Frontal Module. Our method also is explainable; we can examine the LMM reflection content for each image, as shown in Figure 3.2.

### 3.5.3 Connect vision encoders from various domains.

To assess performance of LMM models in more fine-grained domains, we introduce two VQA benchmarks, GLD-VQA and Pittsburgh-VQA, derived from the Google Landmark Dataset (GLD) [143] and the Pittsburgh dataset [10], with 50 test images sampled from each. Drawing inspiration from MMVP [132] and ImageWikiQA [154], we develop multiple-choice questions for each image, such as "Which of the following places is the location for this

**Ground truth:**
'white wolf, Arctic wolf, Canis lupus tundrarum'

**Initial predicted top-5 categories:**
1 'white wolf, Arctic wolf, Canis lupus tundrarum',
2 'timber wolf, grey wolf, gray wolf, Canis lupus',
3 'coyote, prairie wolf, brush wolf, Canis latrans',
4 'dingo, warrigal, warragal, Canis dingo',
5 'Arctic fox, white fox, Alopex lagopus'

**LLaVA-1.6 reflection on prediction 1 ('white wolf, Arctic wolf, Canis lupus tundrarum'):**
The image shows two wolves standing on a rocky terrain with grass and rocks around. They appear to be in a natural, outdoor setting, possibly a wilderness area.
As for the identification of the species, the term "white wolf" is often used colloquially to describe a wolf with a predominantly white coat. However, it is important to note that there is no specific subspecies known as the "white wolf."
The Arctic wolf (Canis lupus arctos) is a subspecies of the gray wolf that is native to the Arctic regions of Canada and Greenland. It is known for its thick, white fur, which provides camouflage in the snowy environment.
Canis lupus tundrarum is another subspecies of the gray wolf, also found in the Arctic regions, but it is not necessarily associated with a white coat.
Based on the image, it is not possible to definitively identify the subspecies of the wolves. They could be Arctic wolves or another subspecies that has a white coat due to genetic variation. Without more information or a clearer view of their physical characteristics, it is not possible to accurately identify the subspecies. Therefore, my answer is: [not sure].

**LLaVA-1.6 reflection on prediction 5 ('white wolf, Arctic wolf, Canis lupus tundrarum'):**
The image shows two white foxes standing on a rocky terrain with vegetation in the background. They appear to be alert and looking in the same direction.
The Arctic fox (Alopex lagopus) is a small fox native to the Arctic regions of the Northern Hemisphere and is well-adapted to living in cold environments. It has a thick, white winter coat that allows it to blend in with the snow, and a thinner, darker summer coat.
The term "white fox" is not a specific species but can refer to any fox with a white coat, which can be found in various species of foxes, including the Arctic fox.
Based on the image, it is not possible to definitively identify the species of the foxes without more information. However, given their white coloration and the presence of rocks and vegetation, which are typical of Arctic environments, it is likely that these are Arctic foxes. Therefore, my answer is:[yes]

**Figure 3.7:** A failure case of LMM reflections on the initial top-5 predictions. We display only the reflection results for the 1st and 5th predictions, with the 2nd, 3rd, and 4th predictions all being "No." When the LMM is unable to find a definitive reason to accept a prediction, it exhibits a degree of randomness when choosing between "Yes," "Not sure," or "No." In this example, despite the correct initial top-1 prediction, the LMM reflection results in an incorrect reranking.

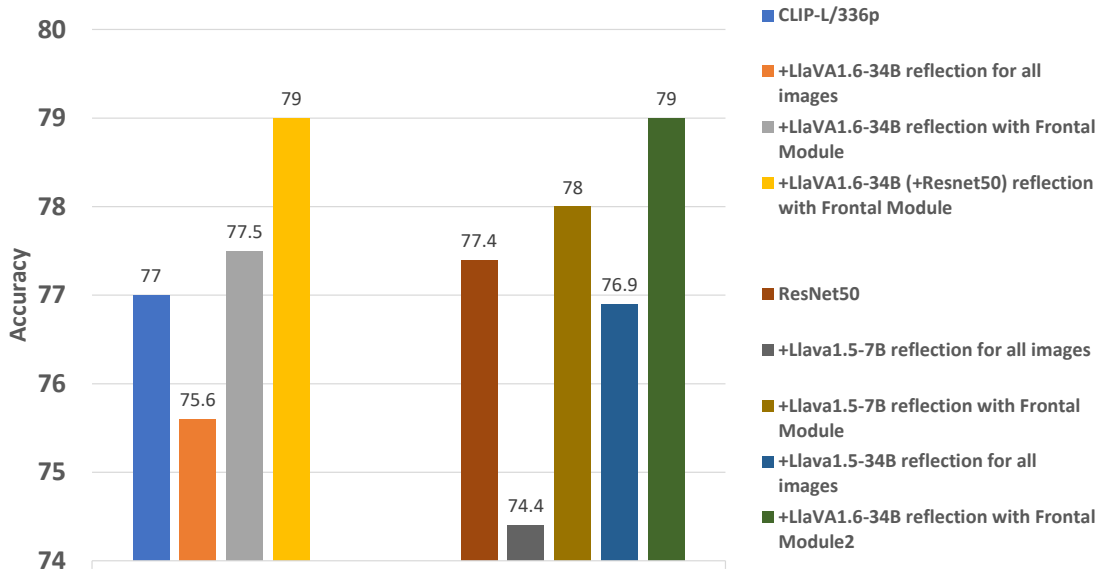



**Figure 3.8:** Effectiveness of LMM reflection for a vision encoder on ImageNet. Without the Frontal Module, reflection decreases performance, whereas incorporating the Frontal Module enhances performance.

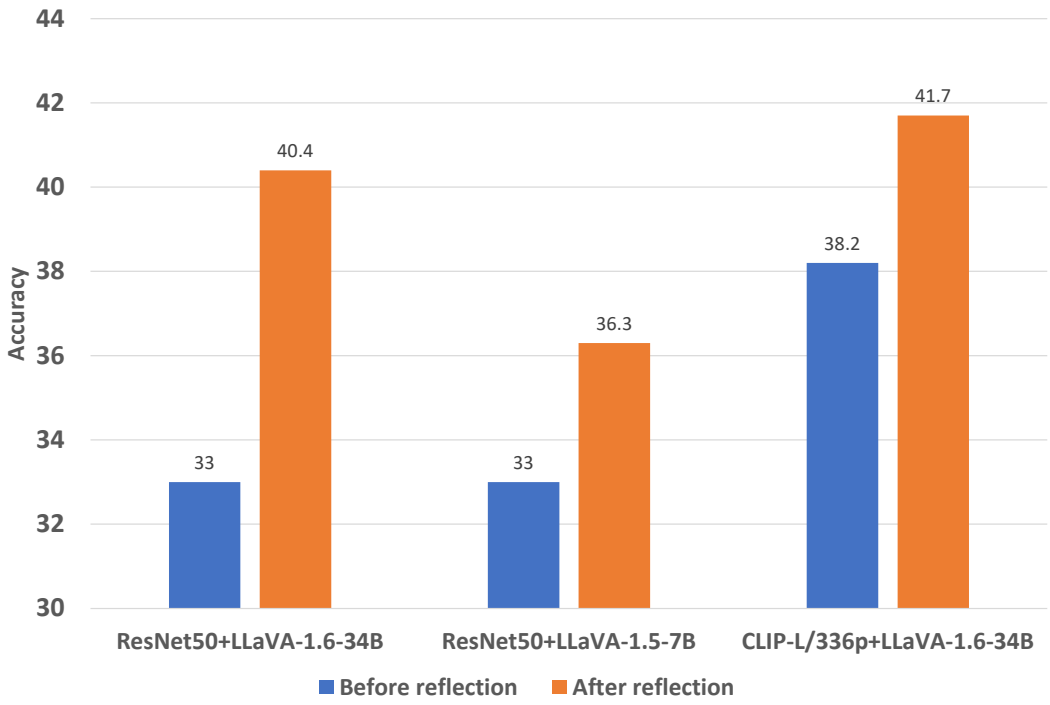



**Figure 3.9:** Image classification accuracy on ImageNet images with a certainty score below 0.5, before and after applying reflection.

| Model | GLD-VQA | Pitts-VQA |
|---|---|---|
| Origin | 38.0 | 12.0 |
| Ours (+DELG/EigenPlaces) | **46.0** | **68.0** |

**Table 3.3:** Performance of LlaVA-1.6-M7B on GLD-VQA and Pitts-VQA, with and without an additional vision encoder integrated through our grandmother cell layer.


**Figure 3.10:** A case demonstrating that integrating our grandmother cell layer, along with an additional DELG vision encoder, enhances the VQA performance of the original LMM.

image?" requiring the LMM to select from 10 possible answers. These benchmarks specifically assess the model's ability to recognize landmarks and urban streets, in contrast to MMVP's focus on spatial distinctions (e.g., left vs. right) and ImageWikiQA's emphasis on fine-grained species recognition.

To evaluate the effectiveness of the proposed grandmother cells layer, we integrate LMM with two more vision encoders: DELG [21], specialized in landmark recognition, and EigenPlaces [16], focused on visual place recognition. Specifically, we connected DELG-ResNet101 and EigenPlaces-ResNet50 to LlaVA-1.6-M7B. The vocabulary used for the connectors consisted of all item names from the target benchmarks.

As shown in Table 3.3, the original LlaVA-1.6-M7B model struggles with street place recognition, achieving only 12% accuracy, comparable to random guessing. However, integrating the new EigenPlaces vision encoder raises its performance to 68%, demonstrating a significant improvement. For landmark domain, the original LlaVA-1.6-M7B achieves 38% accuracy on our GLD-VQA benchmark. After adding the DELG vision encoder, this accuracy increases to 46%. Additionally, we observed that, without multi-selection test prompts, the original LMM sometimes produces unusual image descriptions; incorporating new in-domain vision encoders addresses this issue, as illustrated in Figure 3.10. These results highlight that our grandmother cell layer effectively connects new vision encoders to the pre-trained LMM, yielding substantial gains in recognition capabilities within the target domains.

# Chapter 4. Fast and Accurate Pixel Retrieval via Spatial and Uncertainty Aware Hypergraph Diffusion

## 4.1 Introduction

Image search, a fundamental task in computer vision, has seen significant advancements even before the advent of deep learning [97, 93, 26]. Despite these successes, challenges remain, particularly when the object of interest occupies only a small part of the true positive image, which often includes a complex background. This complexity can hinder users from recognizing the correct images easily, even when the search engine ranks them highly [7]. To address this problem, recent studies emphasize the importance of not just identifying but also precisely localizing and segmenting the query object within the retrieved images [7, 113]. This new task, known as pixel retrieval, aims to enhance the accuracy and user-friendliness of image search results.

There are several existing approaches for pixel retrieval. SPatial verification (SP) examines the spatial arrangement of matched local features to pinpoint the target object in database images [91]. Meanwhile, detection and segmentation methods treat the query image as a learning example to localize and segment the target in database images [89, 77]. Dense-matching methods involve training neural networks to establish pixel correspondences for query and database image pairs [136]. Although these methods show promise in pixel retrieval performance, they share a common limitation: the real-time computation of pixel correspondences for each image pair, leading to significant time consumption [7].

In real-world applications, pixel retrieval demands both speed and accuracy. Our paper proposes a novel approach to address this need: pre-computing spatial matching information of database images offline and swiftly propagating this information for any given online query. We introduce a spatial aware hypergraph diffusion model, which efficiently constructs a hypergraph for a given query and propagates spatial relationships through hyperedges across a sequence of images.

This novel technique excels in achieving rapid and accurate pixel retrieval, and it also notably enhances the performance of image-level retrieval of relation-based reranking methods like query expansion and diffusion. Query expansion [26, 45, 98, 11, 46] issues a new query by aggregating useful information of nearest neighbors of the prior query. Unfortunately, this approach only explores the neighborhood of similar images, resulting in a low recall [53]. On the other hand, diffusion [53, 159] explores more images on a neighborhood graph of the dataset. It, however, adopts false positive items in practice, leading to a low precision [97].

A key issue in these methods is the spatial ambiguity in propagation, as illustrated in Fig. 4.1. Since each database image may contain multiple objects, a sequence of images identified based on simple global feature similarity might not consistently contain the same object. Our hypergraph diffusion technique addresses these low recall and precision issues by effectively propagating to only corresponding local features along spatially aware hyperedges.

Another contribution of this paper is the development of a novel method for measuring the uncertainty of retrieval results. Predicting retrieval uncertainty is an essential but less explored topic in image retrieval; it is crucial for balancing accuracy and speed in image search engines [6]. We leverage the graph structure of the database, wherein each image is a node connected to similar nodes. Our hypothesis is that images in the same community

likely contain the same object. A high-quality retrieval for a query would typically yield images from a singular community. Conversely, a spread across diverse communities indicates higher uncertainty, prompting the search engine to employ more intensive reranking techniques. We term this approach "community selection."

Our experimental findings demonstrate that the spatial aware hypergraph diffusion method achieves outstanding performance in both image-level and pixel-level retrieval, surpassing current state-of-the-art methods. On the challenging ROxford dataset, our hypergraph-based approach records impressive mean Average Precisions (mAPs) of 73.0 and 60.5, with and without R1M distractors, respectively. Moreover, our method exhibits superior pixel retrieval accuracy at a significantly faster speed compared to existing techniques. Additionally, the community selection technique effectively reflects the actual quality of retrieval, confirming its efficiency.

In summary, our contributions are threefold.

1. We introduce a fast and accurate method for pixel retrieval, offloading the slow spatial matching process offline and leveraging a novel hypergraph model for real-time spatial information propagation.
2. We demonstrate that this hypergraph diffusion method substantially improves the image-level retrieval performance of existing relation-based reranking methods, achieving state-of-the-art accuracy on ROxford and RParis datasets.
3. We propose a unique community selection strategy to predict retrieval quality without user feedback, effectively balancing accuracy and speed in real-life applications.

## 4.2 Related works

### 4.2.1 Pixel retrieval

Pixel retrieval, an advanced variant of image retrieval, concentrates on identifying and segmenting pixels associated with a query object within database images [7]. It aims to offer more detailed and user-centric results compared to traditional image retrieval [113, 7].

In the realm of pixel retrieval, existing methods can be classified into spatial verification [26, 91, 21], detection and segmentation [35, 82], and dense matching [136]. Spatial verification evaluates the spatial location consistency of matched local features, such as SIFT [79], DELF [88], and DELG [21]. Originally developed to enhance image-level retrieval performance, it also effectively pinpoints the location of objects in images.

Techniques like Open-World Localization (OWL) [83], Hypercorrelation Squeeze Networks (HSNet) [82], and the Self-Support Prototype model (SSP) [35] represent the detection and segmentation approach. They treat the query image as a one-shot example to localize or segment the target object in database images.

Dense matching methods, including prominent ones like GLUNet [134] and PDCNet [135], focus on establishing dense pixel correspondences. They play a crucial role in identifying corresponding points in database images for pixels in the query image, thereby achieving pixel retrieval objectives.

However, these methods require identifying corresponding pixels between the database and query images during the query time, a process that can be time-consuming and negatively impact user experience in retrieval applications [97, 7]. To overcome this limitation, this paper introduces an innovative approach that significantly speeds up the pixel retrieval process while maintaining high accuracy, thereby enhancing efficiency and user experience in pixel retrieval applications.

### 4.2.2 Relation-based search: query expansion and diffusion

Numerous methods leverage the interrelations among database images to improve search results. These techniques are primarily categorized into query expansion (QE) [91] and diffusion [159].

QE [26] aggregates the top-ranked initial candidates into an expanded query, which is then used to search for more images in the database. The critical factor here is aggregating only helpful information and culling out unrelated information. Average query expansion (AQE) [26] mean-aggregates the top k retrieved images as the expanded query. Average query expansion with decay (AQEwD) [45] gives the top k images the monotonically decaying weights over their ranks. Alpha query expansion ($\alpha$QE) [98] uses the power-normalized similarity between the query and the top-ranked images as the aggregation weights. Discriminative query expansion (DQE) [11] also uses the weighted average, where the weight is the dual-form solution of an SVM, which classifies the positive (top-ranked) and negative (low-ranked) images of a query. Learnable attention-based query expansion (LAttQE) [46] uses self-attention [138] to share information between the query and the top-ranked items to compute better aggregation weights.

Diffusion [159, 53] ameliorates the abovementioned problem by propagating similarities through a pairwise affinity matrix. The works of Chang *et al*. [22] and Liu *et al*. [71] are two novel variants of this line. The critical issue in diffusion is how to calculate the affinity matrix. Some works [95, 53] use the reciprocal neighborhood relations to refine the search results, and other [93, 22] use the inliers number of SP to re-weight the similarity. However, these methods compress the relations between two images as a single scale. As a result, these prior methods lost the spatial matching detail. Although the propagation methods are elegant, they have difficulties recovering good relations to guide the diffusion direction.

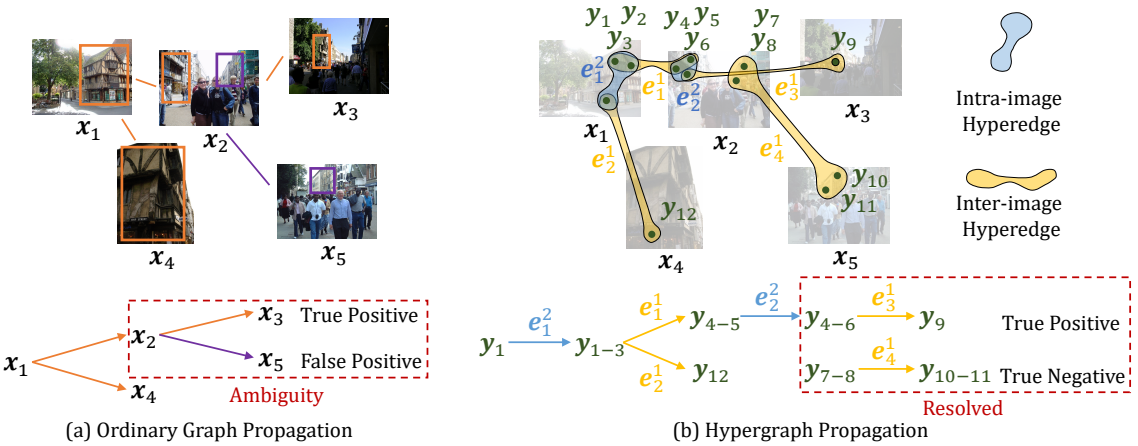
We propose a novel and straightforward hypergraph model to utilize the relations among database images better. It connects the corresponding local features among images using hyperedges; in a sequence of similar images, it detects the image from which the following images are no longer relevant to the query. This approach significantly enhances both image and pixel retrieval performance.

### 4.2.3 Local descriptors and spatial verification

Local descriptors are the representations of an image's important patches. Some popular local descriptors are RootSIFT [11], deep local feature (DELF) [88] and deep local and global feature (DELG) [21].

Local descriptors' location information is usually used in the spatial verification (SP) [91] stage. A standard SP pipeline is first using a KD-tree [91] to find the nearest neighbors between local descriptors in two images. Then it uses the RANSAC algorithm [37] to estimate the homography matrix [122] and the number of inlier correspondences between two images. Existing image search engines commonly apply SP on the shortlist, say the top 100 images ranked by a global descriptor.

SP is known to be crucial for verifying true positive images [97]. Performing SP before diffusion improves the final retrieval performance [97, 22]. However, it is also the slowest process [11, 93] in image search. In this paper, we put the heavy SP in an offline setting and propose the hypergraph diffusion method to propagate the spatial matching information online. In addition, we also introduce the community selection technique to reduce SP's computing overhead. Community selection frames the diffusion initialization task as selecting a community for the given query rather than verifying the top-ranked images individually. The experiment shows that our hypergraph diffusion and community selection reduce SP overhead without hurting accuracy.




(a) Ordinary Graph Propagation (b) Hypergraph Propagation

**Figure 4.1:** a) shows a part of an ordinary graph with scalar-weighted, i.e., similarity, edges. Orange frames are the common visible regions among images $\mathbf{x}_1$, $\mathbf{x}_2$, $\mathbf{x}_3$, and $\mathbf{x}_4$. Purple frames are the common visible regions between images $\mathbf{x}_2$ and $\mathbf{x}_5$. $\mathbf{x}_3$ and $\mathbf{x}_5$ are close neighbors to image $\mathbf{x}_2$. While $\mathbf{x}_3$ is related to $\mathbf{x}_1$ by sharing the orange frame, $\mathbf{x}_5$ is not. Utilizing scalar-weighted edges cannot propagate the query in the ordinary graph without this ambiguity issue. b) shows the corresponding hypergraph of a). Inter-image hyperedges $\mathbf{e}_s^1$ are shown in yellow, intra-image hyperedges $\mathbf{e}_k^2$ are in blue, and local features $\mathbf{y}_n$ are in green. A hypergraph path connects local features from $\mathbf{y}_1$ to $\mathbf{y}_9$ in $\mathbf{x}_1$ and $\mathbf{x}_3$, but no path connects local features in $\mathbf{x}_1$ and $\mathbf{x}_5$. A large version of this figure is in the appendix.

## 4.3 Overview

We first discuss issues of query expansions and diffusion process of ordinary graphs in Sec.4.3.1 and our novel hypergraph based propagation (Sec. 4.3.2) and initialization scheme utilizing the concept of community (Sec. 4.3.3) for achieving accurate diffusion process.

### 4.3.1 Propagating in ordinary graph

We can use a graph $\boldsymbol{G} = (\boldsymbol{X}, \boldsymbol{F})$ to represent the database images. $\boldsymbol{X} := \{\mathbf{x}_1, ..., \mathbf{x}_m\}$ is the node set, where item $\mathbf{x}_i$ could be either an image or an image region depending on a chosen retrieval method; $\boldsymbol{F}$ is the set of edges whose scalar weight represents the similarity between two nodes connected by each edge. Utilizing information in this ordinary graph is an important strategy to improve retrieval results, and, thus, widely used in various methods, such as query expansion [26, 25, 98, 46] and diffusion [53, 159]. Both diffusion and query expansion can be represented by the following function:

$$\mathbf{f}^* = I(\mathbf{f}^0), \tag{4.1}$$

where $I$ represents the post-processing process, $f_u \in \mathbb{R}$ is the ranking score for $\mathbf{x}_u$, $\mathbf{f}^0$ is the initial ranking vector with $f_u^0 = 1$ if $\mathbf{x}_u$ is a query and $f_u^0 = 0$ otherwise, and $\mathbf{f}^*$ is the ranking vector after query expansion or diffusion. Diffusion methods diffuse the label of the query node to its neighbors along the graph edges until reaching a stationary state. The stationary state is found by repeating the iteration function: $\mathbf{f}^{t+1} = \alpha \mathbf{O}\mathbf{f}^t + (1-\alpha)\mathbf{f}^0$, where $\mathbf{O}$ is the normalized affinity matrix that describes the weights of graph edges, $\alpha$ is a jumping probability. Query expansion gathers useful information in the query node's nearest neighbors to update the ranking list.

Diffusion and query expansion have different advantages and disadvantages. Diffusion can efficiently propagate information to the whole graph through a compact formulation. However, the propagation is conducted through the affinity matrix $\mathbf{O}$ [159, 53], whose scalar value cannot represent accurate spatial information, resulting in low

precision. Take Figure 4.1 as an example, where images of $\mathbf{x}_1$ and $\mathbf{x}_5$ are incorrectly connected through image $\mathbf{x}_2$ in the ordinary graph with scalar edge weights; note that they do not share common regions. This leads to false positives in the diffusion process. We denote this as the **ambiguity problem of propagation**. In contrast, query expansion uses sophisticated approaches, such as transformer [46] and spatial verification [26, 11], to guarantee that only shared information between query and its nearest neighbors are used to update the ranking list. This approach, however, only explores the neighborhood of very similar images due to the computational overhead [97, 11, 26], resulting in a low recall.

This paper proposes novel hypergraph propagation and community selection, settling the ambiguity problem separately in the propagation and initialization stages. We describe these two techniques next.

### 4.3.2 Extension to hypergraph

One meaningful direction to overcome the drawbacks of diffusion and query expansion is to handle the ambiguity problem of graph propagation by explicitly resolving the propagation in a more informative space. The local features contain informative spatial information that can help to distinguish the ambiguity through geometric verification [91]. However, considering the large scale of the image database, it is impractical to do the geometric verification on all the pairs. In practice, image search engines usually build the kNN graph instead of the complete graph, which means every image only connects with its top k similar images [53]. To utilize the spatial information on the kNN graph, we propose a novel hypergraph propagation model to efficiently propagate spatial information through hyperedges, connecting an arbitrary number of nodes. This technique is appropriate for the real-world situation where the same object in different images has a different number of local features. In the example of Figure 4.1, image $\mathbf{x}_1$ can connect image $\mathbf{x}_3$ by matching the local features with those of image $\mathbf{x}_2$, but will not connect image $\mathbf{x}_5$.

When a query is given, we define a hypergraph $\boldsymbol{H} = (\boldsymbol{Y}, \boldsymbol{E})$ on top of the ordinary image graph, where its node set $\boldsymbol{Y} = \{\mathbf{y}_1, ..., \mathbf{y}_n\}$ contains all the local features extracted, and the edge set $\boldsymbol{E}$ consists of hyperedges, which are used to connect the local features related to the query object in different images. Our model then propagates the label of local features of the query to their neighbors and uses an aggregation function to find the final ranking list of images, as follows:

$$\mathbf{l}^* = I_h(\mathbf{l}^0), \mathbf{f}^* = Agg(\mathbf{l}^*), \tag{4.2}$$

where $\mathbf{l}^0$ and $\mathbf{l}^*$ are the initial and final ranking lists of local features respectively, $\mathbf{f}^*$ is the ranking list of images, $I_h$ is the propagation function in a hypergraph, and $Agg$ is the aggregation function that converts ranking scores of local features to ranking scores of images. We will discuss the construction of hypergraph in Section 4.4.1 and the propagation process in Section 4.4.2.

### 4.3.3 Handling new queries: community selection

Diffusion methods [33, 158] usually consider the query point to be contained in the dataset. The standard approach to handle a new query is representing the query using its top-ranked images in the initial search to start the propagation [53]. However, it is hard to guarantee that these global descriptor-based nearest neighbors contain the same object as the query. We call it the **ambiguity problem of initialization**. Recent studies [97, 22] perform online spatial verification on 100 nearest neighbors of the query and delete the wrong neighbors to improve the quality of diffusion. Unfortunately, conducting spatial verification for 100 images in query time is a heavy computational load for a search engine.

We find that the structure of the image graph [93, 11] contains helpful information to measure the accuracy of the nearest neighbors. Following the term in graph theory, we call tightly connected groups communities, which are sets of nodes with many connections inside and few to outside. We observe that the items containing the same objects have a high probability of composing a community and having short geodesic distances. Based on this observation, we frame the diffusion initialization task as selecting a community for the given query instead of using spatial verification to delete the wrong neighbors. This graph-based approach improves both accuracy and speed. We will give the detail of community selection in Section 4.4.3.

## 4.4 Methods

### 4.4.1 Construction

Assume that we have an image dataset $\boldsymbol{X} = \{\mathbf{x}_1, ..., \mathbf{x}_m\}$, where $\mathbf{x}_u$ is an image, and its corresponding K-nearest neighbor (Knn) graph [159, 32] $\boldsymbol{G} = (\boldsymbol{X}, \boldsymbol{F})$, where $\boldsymbol{F}$ is the ordinary edge sets. For hypergraph propagation, we define a hypergraph $\boldsymbol{H} = (\boldsymbol{Y}, \boldsymbol{E})$, where $\boldsymbol{Y} = \{\mathbf{y}_1, ..., \mathbf{y}_n\}$ is the set of vertices and $\mathbf{y}_i$ represents a local feature in an image. $\boldsymbol{E} = \boldsymbol{E}^1 \cup \boldsymbol{E}^2$ is the union of the inter-image hyperedges set $\boldsymbol{E}^1 = \{\mathbf{e}_1^1, \mathbf{e}_2^1, ...\mathbf{e}_s^1\}$ and the intra-image hyperedges set $\boldsymbol{E}^2 = \{\mathbf{e}_1^2, \mathbf{e}_2^2, ...\mathbf{e}_k^2\}$. An inter-image hyperedge connects local features in two images; we also say it connect two images in this paper. An intra-image hyperedge connects local features in the same image. These hyperedges are shown by the orange and blue groups in Figure 4.1-b). The hypergraph has two incidence matrixes $\mathbf{S}^1 = (s_{ia}^1)$ and $\mathbf{S}^2 = (s_{ia}^2)$, where $s_{ia}^\alpha$ is 1 if $\mathbf{y}_i \in \mathbf{e}_a^\alpha$ and 0 otherwise. We build the ordinary Knn offline and build the hypergraph on top of the Knn graph online. We call the local features in the hypergraph **activated local features**. We now describe how to constructe the hypegraph.

In order to solve the ambiguity problem of graph propagation, hypergraph propagation propagates only among the spatially matched local features of images. This paper uses RANSAC [37], the most common features matching method in the spatial verification stage of image search [97, 88, 21], to pre-compute the homography matrix [37] between two images offline. We then use these pre-computed homography matrixes to verify if two features in two images are matched when building the hyperedges at query time.

We assume the query image can be represented by local features in $\boldsymbol{Y}$ and will discuss how to handle a new query in Section 4.4.3. Because a propagation process of a query is progressive, we expand hyperedges on-demand. To judge if an image is related to the query in a sequence of similar images, we only need to see if it has the activated local features. An inter-image hyperedge is then defined as follows:

**Definition 4.4.1 (Inter-image hyperedge)** *For two images $\boldsymbol{x}_u$ and $\boldsymbol{x}_v$, let $\boldsymbol{V}_u \subset \boldsymbol{Y}$ be the activated local features set of $\boldsymbol{x}_u$, and $\boldsymbol{I}_v \subset \boldsymbol{Y}$ be the local features set of $\boldsymbol{x}_v$. The inter-image hyperedge among features in $\boldsymbol{x}_u$ and $\boldsymbol{x}_v$ is: $\boldsymbol{e}^1 = (\ In = V_u, OUT = \{\ \boldsymbol{y}_j \in \boldsymbol{I}_v; \exists \boldsymbol{y}_i \in IN$ is matched with $\boldsymbol{y}_j$ by the pre-computed homography between $\boldsymbol{x}_u$ and $\boldsymbol{x}_v\ \}\ )$.*

If the matching inliers between two images are less than a threshold, they could be directly verified as irrelevant with each other. As the target of hypergraph is to search the related images for the query, we do not build the hyperedge for them. We set the threshold as 20 in this paper, which is the standard threshold in the spatial verification stage of the image search systems [93, 24]. Note that even if an image has more than 20 matching inliers with another one, its activated local features may not be matched.

Similar to object detection [103], we use bounding box to describe the object region. In an image, local features depicting an object locate in the object's bounding box. As a homography does not guarantee to find all the matched

features between two images, the inter-image hyperedges do not connect all the features in an object's bounding box. Unfortunately, the unmatched related features are also important for propagation. As shown in Figure 4.1-b), features $\mathbf{y}_1$, $\mathbf{y}_2$ and $\mathbf{y}_3$ are in the same region and describing the same building, but only $\mathbf{y}_1$ and $\mathbf{y}_3$ are connected within image $\mathbf{x}_2$ using an inter-image hyperedge to image $\mathbf{x}_1$. Without an intra-image hyperedge connecting $\mathbf{y}_1$, $\mathbf{y}_2$ and $\mathbf{y}_3$, we cannot propagate information from images $\mathbf{x}_1$ to $\mathbf{x}_3$. To recover the unmatched yet related local features, we define the intra-image hypergraph as follows:

**Definition 4.4.2 (Intra-image hyperedge)** *Given a feature set in an image $\boldsymbol{x}_u$, let $\boldsymbol{V}_u \subset \boldsymbol{Y}$ be the activated local features set of $\boldsymbol{x}_u$, and $\boldsymbol{I}_u \subset \boldsymbol{Y}$ be the set of all the local features of $\boldsymbol{x}_u$. Let the smallest rectangle covering all the features in $\boldsymbol{V}_u$ be the bounding box $b$. The intra-image hyperedge in $\boldsymbol{x}_u$ is defined as $\boldsymbol{e}^2 = (IN = \boldsymbol{V}_u, OUT = \{\boldsymbol{y}_j \in \boldsymbol{I}_u; \boldsymbol{y}_j$ exists in $b\})$.*

Due to the high computing cost, it is impractical to perform the propagation on a large-scale dataset consisting of millions of images. Inspired by the idea of truncation [53] in diffusion, we build inter-image hyperedges only among the hop-$N$ neighbors [1] of a given query image. Specifically, we build the inter-image hyperedges from image $\mathbf{x}_u$ to $\mathbf{x}_v$ only if both of them are in the hop-$N$ neighbors of the query.

### 4.4.2 Propagation

After building the hypergraph, we use a propagation model in the hypergraph to achieve the final ranking scores of retrieved images. Specifically, we define a propagation matrix $\mathbf{P} = (p_{ij})$ and set the propagation weight between $\mathbf{y}_i \in \mathbf{x}_u$ and $\mathbf{y}_j \in \mathbf{x}_v$ as $p_{ij} = d(\mathbf{x}_u, \mathbf{x}_v) \max_{abc}(s^1_{ia} s^1_{ba} s^2_{bc} s^2_{jc})$, where $d(\mathbf{x}_u, \mathbf{x}_v)$ is the Euclidean distance between the global features of $\mathbf{x}_u$ and $\mathbf{x}_v$, and $\max_{abc}(s^1_{ia} s^1_{ba} s^2_{bc} s^2_{jc})$ records whether $\mathbf{y}_i$ and $\mathbf{y}_j$ are connected in the hypergraph. The propagation matrix is normalized using the degree matrix $\mathbf{D} = diag(\mathbf{P}\mathbf{1}_n)$ as $\mathbf{P}' = \mathbf{P}\mathbf{D}$. For aggregation function, $a_{iu} \in \mathbf{A}$ is set as $\frac{1}{b_u}$ if $\mathbf{y}_i \in \mathbf{x}_u$, and 0 otherwise, where $b_u$ is the number of activated local features in $\mathbf{x}_u$. Then the propagation function in Equation 4.2 can be written as:

$$\begin{aligned} \mathbf{l}^* &= I_h(\mathbf{l}^0) = \mathbf{l}^0 + \mathbf{l}^0\mathbf{P}' + \mathbf{l}^0\mathbf{P}'^2 + \cdots + \mathbf{l}^0\mathbf{P}'^N, \\ \mathbf{f}^* &= Agg(\mathbf{l}^*) = \mathbf{l}^*\mathbf{A}. \end{aligned} \tag{4.3}$$

Intuitively, this propagation function gives high scores to the images with shorter hypergraph geodesic distances to the query while considering similarities of global features.

### 4.4.3 Community selection for new query

The hypergraph propagation described in Sec. 4.4 assumes the query is in the database, which may not be true in a general retrieval scenario. Now we consider how to find the good starting nodes for a new query.

As discussed in Section 4.3, we observe that the items containing the same object usually belong to the same community in $G$. By checking whether the top S items of the initial search are in the same community, we can evaluate the quality of the initial search without query time spatial verification. As shown in Figure 4.2, the quality of the global-feature-based search of Q1 is better than Q2 because its top-ranked items are in the same community.

Specifically, after the initial search, we use the top $S$ images and their connected edges in $G$ to build a subgraph $G_s$. $G_s$ consists of one or more components, and each component represents a community in $G$. We call the top-1 image of initial search the dominant image, its component in $G_s$ the dominant component, and its community

[1] Node A is in B's hop-$N$ neighbors if A can be connected with B using not more than $N$ different hyperedges.

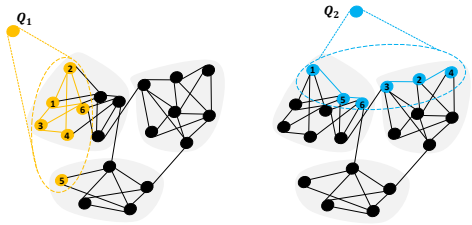



**Figure 4.2:** Two queries on an image graph with three communities. The numbers on the nodes are their rankings in the initial search. The uncertainty of the initial search result of Q1 is lower than that of Q2 as most retrieved items of Q1 distribute in the same community.

in $G$ the dominant community. We call other components in $G_s$ as the opposition components. Intuitively if all the top $S$ images in $G_s$ constitute a single component, all of them may contain the same object with the query image. Otherwise, the uncertainty will increase if the number and size of the opposition components increase. This observation also agrees with our experimental results. Inspired by Shannon entropy, we define the uncertainty index, $U$, of the initial search as:

$$U = -\sum_{i=1}^{n_G} P\left(C_i\right) \log P\left(C_i\right), \tag{4.4}$$

where $n_G$ is the number of component in $G_s$, $C_i$ is a component and $P\left(C_i\right)$ is the proportion of $C_i$ in $G_s$.

Community selection is the pre-stage of hypergraph propagation. The overall retrieval process with community selection is to perform the initial search firstly and then calculate the uncertainty of the initial search result. If the uncertainty index is low, say $U$ is lower than 1, hypergraph propagation uses images in the dominant community to represent the query image and starts the propagation process from them. If the uncertainty index is high, the search engine does the spatial verification on the top $T$ images of the initial search to find a new dominant community for the query. In this way, we significantly reduce the number of spatial verification while keeping very high accuracy.

## 4.5 Experiment

We first introduce the benchmarks and implementation detail in Section 4.5.1 and 4.5.2, respectively. We then compare the accuracy of our hypergraph diffusion and the SOTA image-level and pixel-level methods in Section 4.5.3 and 4.5.4. We show the effectiveness of community selection in Section 4.5.5. Finally, we report and discuss the time and memory costs in Section 4.6.

### 4.5.1 Test datasets and evaluation protocals

For image-level retrieval, we evaluate our methods on two well-known landmark retrieval benchmarks: revisited Oxford (ROxf) and revisited Paris (RPar) [97]. There are 4,993 (6,322) database images in the ROxf (RPar) dataset, and a different query set for each dataset, both with 70 images. We also report the result with the R1M distractor set, which contains 1M images. The accuracy is measured using mean average precision (mAP).

For the pixel-level retrieval, we evaluate our methods on pixel-retrieval benchmarks: pixel revisited Oxford (PROxford) and pixel revisited Paris (PRParis) [7]. These two datasets are designed based on the famous ROxford and RParis [97]. They have the same query, database, and distractor images as ROxford and RParis, and they provide pixel-level annotation. The accuracy is measured using mAP@50:5:95. Each ground truth image in the ranking list is treated as a true-positive only if its detection Intersection over Union (IoU) is larger than a threshold

n; the threshold n is set from 0.5 to 0.95, with step 0.05. As [7], we report the mean of IoU (mIoU) for all the methods and report the mAP@50:5:95 for some representative methods.

### 4.5.2 Implementation

We use the DELG model [21] trained on GLDv2 [143] to extract both global and local features for our Hypergraph Diffusion (HD) and all the baseline methods unless a specific explanation is given. We find the nearest neighbor number $K$ does not influence the final result of HD a lot, and we report the performance of setting $K$ as 200.

For the image-level comparison, we evaluate the performance of HD and baseline methods under the ordinary retrieval pipeline: performing the initial search, followed by conducting reranking methods like query expansion and diffusion. We reorder the initial ranking list using $\mathbf{f}^*$ after the HD as the final results. We do not use community selection and always start the HD from the initial dominant community for all the queries to make the comparison fair.

For the pixel-level comparison, we test all the methods using the same image-level ranking result, the image ranking result obtained by our HD, for a fair comparison. HD directly produces the pixel-level result when reranking the image-level result without having any extra time-consumption. For the SOTA baseline methods, we use them to calculate for each database and query pair to get the pixel-level result. The accuracy is reported in Section 4.5.4, and the time and memory cost is reported in Section 4.6.

For community selection, we set $S$, the number of images for calculating the uncertainty, as 20. When testing the performance of combining community selection and hypergraph propagation, we set the uncertainty threshold as 1, which we find to be a good balance between the accuracy and computation overhead. For the query with an uncertainty index higher than the threshold, we do the spatial verification for the top 100 items using SIFT features [78]. Once we find an item with more than 20 inliers, we treat its community as the new dominant community, stop the spatial verification, and apply hypergraph propagation in the new dominant community.

We conduct the offline process on 6 Intel(R) Core(TM) i9-9900K CPUs @ 3.60GHz and 896GB of RAM and implement the online hypergraph propagation and community selection on one CPU and 50GB of memory. The code of this work is publicly available on https://sgvr.kaist.ac.kr/~guoyuan/hypergraph_propagation/

### 4.5.3 Comparison to the image retrieval state-of-the-art

**Baselines**. We compare hypergraph propagation against the existing QE/diffusion methods: average query expansion [26], average query expansion with decay [45], $\alpha$ query expansion, and diffusion [33].

This manuscript builds upon our prior work [6]. During the preparation of this manuscript, several state-of-the-art (SOTA) methods have been introduced. We benchmark our approach against these new techniques. The new reranking methods are Guided Similarity Separation (GSS)[71], Learnable Attention-based Database-side Augmentation (LAttDBA), Query Expansion (LAttQE)[46], and Self-Attention Aggregation (SAA)[90]. We also evaluate against cutting-edge deep learning models, particularly those fine-tuned on Structure-from-Motion (SfM) datasets, namely HOW[129] and Feature Integration-based Retrieval (FIRe)[142], as well as those refined using the Google Landmark Dataset (GLD), including Detection-to-Retrieval (D2R)[126], Token-based model (Token)[145], Deep Orthogonal fusion of Local and Global features (DOLG)[151], Second-Order Loss and Attention for Image Retrieval (SOLAR)[86], Spatial-Context-Aware model (SpCa)[156], and Coarse-to-Fine Compact Discriminative model (CFCD) [161].

**Table 4.1:** Results (% mAP) on the ROxf/RPar datasets and their large-scale versions ROxf+1M/RPar+1M, with both Medium and Hard evaluation protocols. **Bold** number indicates the best performance, and underline indicate the second and third one.

| | ROxf | | ROxf+R1M | | RPar | | RPar+R1M | |
|---|---|---|---|---|---|---|---|---|
| Method | M | H | M | H | M | H | M | H |
| Using standard DELG features | | | | | | | | |
| Global search [21] | 76.3 | 55.6 | 63.7 | 37.5 | 86.6 | 72.4 | 70.6 | 46.9 |
| Spatial verification [21] | 81.2 | 64.0 | 69.1 | 47.5 | 87.2 | 72.8 | 71.5 | 48.7 |
| Average QE [26] | 77.2 | 57.1 | 68.5 | 43.0 | 87.6 | 74.3 | 75.4 | 54.8 |
| Average QE with decay [45] | 78.4 | 58.0 | 70.4 | 44.7 | 88.2 | 75.3 | 76.2 | 56.0 |
| $\alpha$ QE [98] | 65.2 | 43.2 | 57.0 | 30.2 | 91.0 | 81.2 | 81.0 | 64.1 |
| Diffusion [53] | 81.0 | 59.3 | 63.9 | 38.7 | 91.4 | 82.7 | 80.0 | 64.9 |
| Hypergraph Diffusion (Ours) | **85.7** | **70.3** | **78.0** | **60.0** | **92.6** | **83.3** | **86.6** | 72.7 |
| Advanced reranking methods | | | | | | | | |
| LAttQE [46] | 73.4 | 49.6 | 58.3 | 31.0 | 86.3 | 70.6 | 67.3 | 42.4 |
| LAttDBA+LAttQE [46] | 74.0 | 54.1 | 60.0 | 36.3 | 87.8 | 74.1 | 70.5 | 48.3 |
| SAA [90] | 78.2 | 59.1 | 61.5 | 38.2 | 88.2 | 75.3 | 71.6 | 51.0 |
| GSS [71]+SAA [90] | 79.3 | 62.2 | 62.1 | 42.3 | 90.7 | 80.0 | 85.1 | 70.3 |
| Diffusion [53]+SAA [90] | 76.3 | 57.8 | 66.2 | 42.4 | 90.2 | 81.2 | 86.3 | **75.4** |
| Advanced deep learning models fine-tuned using SfM | | | | | | | | |
| HOW (R50) [129] | 78.3 | 55.8 | 63.6 | 36.8 | 80.1 | 60.1 | 58.4 | 30.7 |
| FIRe (R50) [142] | 81.8 | 61.2 | 66.5 | 40.1 | 85.3 | 70.0 | 67.6 | 42.9 |
| Advanced deep learning models fine-tuned using GLD | | | | | | | | |
| D2R-DELF-ASMK (R50) [126] | 76.0 | 52.4 | 64.0 | 38.1 | 80.2 | 58.6 | 59.7 | 29.4 |
| Token (R101) [145] | 77.8 | 60.1 | 66.7 | 43.1 | 87.9 | 74.7 | 75.2 | 53.2 |
| DOLG (R101) [151] | 78.4 | 58.6 | 75.5 | 52.4 | 88.5 | 75.4 | 78.3 | 61.9 |
| SOLAR (R101) [86] | 81.6 | 63.3 | 71.8 | 45.3 | 88.2 | 75.2 | 72.9 | 51.3 |
| SpCa-cro (R50) [156] | 79.9 | 59.3 | 72.8 | 49.3 | 87.4 | 73.1 | 78.0 | 58.3 |
| SpCa-cat (R50) [156] | 81.6 | 61.2 | 73.2 | 48.8 | 88.6 | 76.2 | 78.2 | 60.9 |
| SpCa-cro (R101) [156] | 82.7 | 65.6 | 77.8 | 53.4 | 90.2 | 79.3 | 79.1 | 65.8 |
| SpCa-cat (R101) [156] | 83.2 | 65.9 | 77.8 | 53.3 | 90.6 | 80.0 | 79.5 | 65.0 |
| CFCD-3 scales (R50) [161] | 82.5 | 63.6 | 72.7 | 48.5 | 89.6 | 78.1 | 78.9 | 60.1 |
| CFCD-3 scales (R101) [161] | 84.1 | 67.8 | 74.7 | 54.1 | 91.0 | 81.2 | 82.2 | 65.5 |
| CFCD-5 scales (R50) [161] | 82.4 | 65.1 | 73.1 | 50.8 | 91.6 | 81.7 | 81.6 | 62.8 |
| CFCD-5 scales (R101) [161] | 85.2 | 70.0 | 74.0 | 52.8 | 91.6 | 81.8 | 82.8 | 65.8 |

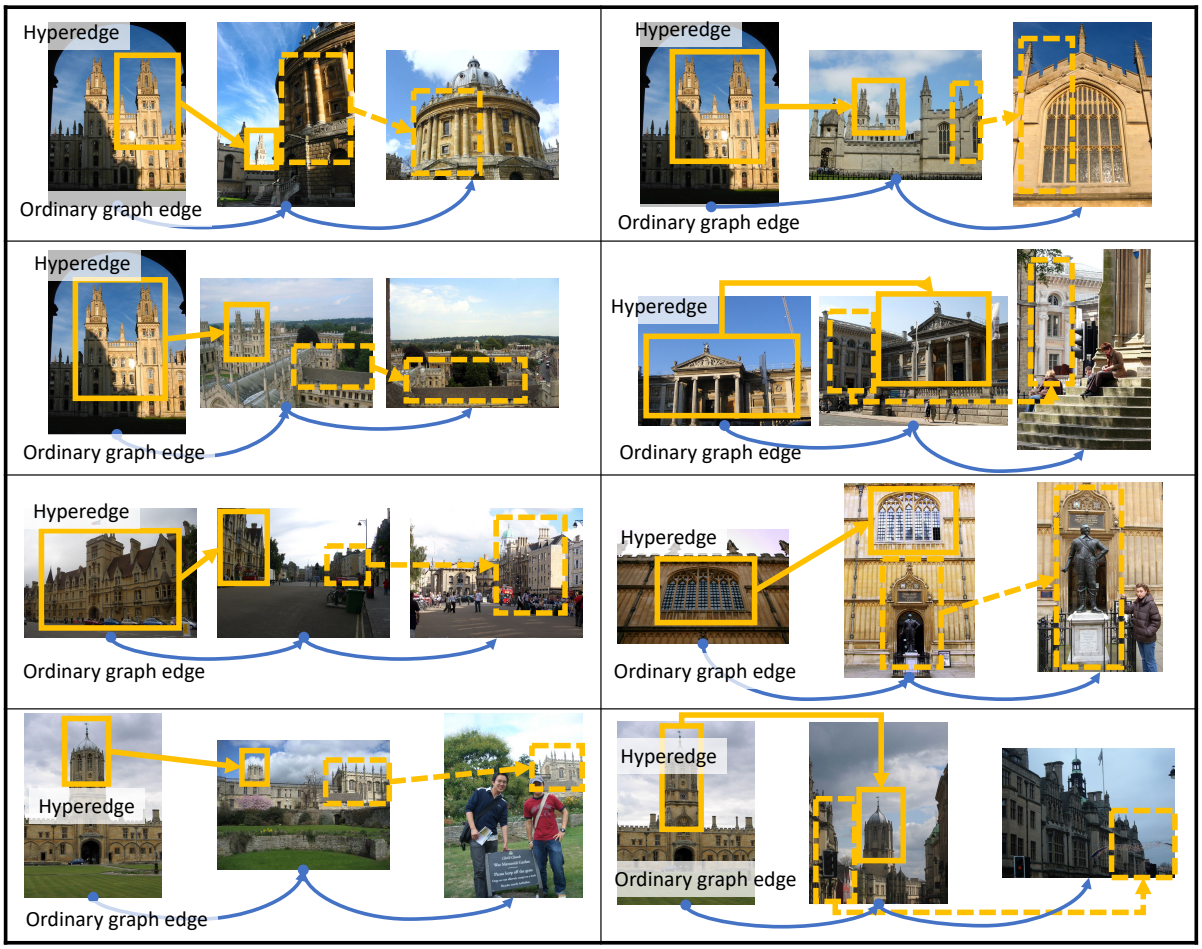



**Figure 4.3:** Illustration of hypergraph diffusion mechanism. The orange boxes with arrows represent the hyperedges, and the blue curved arrows are the ordinary graph edges. In each triplet, the first image and the third image are wrongly connected through the second image. While traditional diffusion and QE methods wrongly propagate the similarity score from the first to the third image through ordinary graph edge, our hypergraph diffusion does not diffuse the similarity scores from the first to the third image by solving the spatial ambiguity problem of propagation.

**Result**. Table 4.1 compares the accuracy of hypergraph diffusion (HD) and other retrieval methods. Among the QE/diffusion methods, HD obtains the best result on ROxford and RParis, both with and without the R1M distractor set. Compared to RParis, ROxford is more challenging as it contains more difficult retrieval cases such as the change of viewpoint [97]. The performance improvement of our method on ROxford hard cases is impressive; we get the mAP 70.3 and 60 with and without 1M distractors, outperforming the traditional diffusion method 18.5% and 55%, respectively. Our results are also better than other techniques in [97, 22], which integrate additional steps with more computation and memory requirements.

In Figure 4.3, we demonstrate the efficacy of hypergraph propagation using actual query examples. The figure utilizes highlighted regions, marked in yellow, to indicate areas where correct matches are identified through hyperedges in our model. The figure presents triplets of images to illustrate a common issue in traditional diffusion and QE methods. In each triplet, the first and third images are often mistakenly linked via the second image, although they do not contain the same object. This linkage exemplifies the famous false-positive problem in image retrieval, as discussed in [97]. Our hypergraph mechanism addresses this by resolving the spatial ambiguity commonly observed in propagation processes. By diffusing along the hyperedges instead of the ordinary graph

edge, our method significantly mitigates the false-positive issue in existing diffusion and QE techniques.

Ordinary diffusion is better than QE methods on all the datasets except ROxford with R1M distractors. We think that the texture and shape of medieval buildings in ROxford are common and tend to appear in other buildings around the world. When retrieving these objects with R1M distractors, its ambiguity problem explores more false positive items, leading to a low mAP than QE methods. Our hypergraph propagation keeps the spatial information during propagation, thus avoiding the false positive issue of ordinary diffusion, resulting in high accuracy.

While preparing this manuscript, several advanced and promising methods have emerged in the field. As shown in Table 4.1, our HD approach maintains a competitive accuracy compared to these recent methods. HD achieves the highest accuracy on the ROxford and RParis benchmarks. The only exception is in the challenging scenario of RParis with 1 million distractors under the hard setting, where the diffusion [53] combined with SAA [90] outperforms others. To the best of our knowledge, methods in Table 4.1 are the state-of-the-art (SOTA) techniques in the landmark retrieval field. Table 4.1 shows that our HD still gives the SOTA accuracy among fine-tuned models using standard GLD (both v1 and v2-clean) and SfM fine-tuning sets.

### 4.5.4 Comparison to the pixel retrieval state-of-the-art

**Baselines**. Besides the image level performance, we compare our hypergraph diffusion with the state-of-the-art (SOTA) methods on pixel retrieval benchmarks PROxford and PRParis in Table 4.3. To find the pixel correspondences in the given database image, we choose the SOTA methods for landmark search, detection, segmentation, and dense matching fields. SPatial verification (SP) with Scale-Invariant Feature Transform (SIFT), DEep Local Feature (DELF), and DEep Local and Global features (DELG) [21] are the SOTA matching approaches in the landmark retrieval field. We also test the SOTA detection method Detect-to-Retrieval (D2R) in the landmark retrieval field, where the detailed implementation is explained in the next paragraph. The Vision Transformer for Open-World Localization (OWL-ViT) [83] and One-Stage one-shot Detec- tor (OS2D) are the SOTA methods in detection. The Self-Support Prototype model (SSP) [35], Hypercorrelation Squeeze Network (HSNet), and Mining model (Mining) are the SOTA segmentation methods. The Global-Local Universal Network (GLUNet), Probabilistic Dense Correspondence Network (PDCNet), and GLUNet trained using Globally Optimized Correspondence (GOCor) and Warp Consistency objective (WarpCGLUNet) [136] are the SOTA methods in dense matching.

We additionally introduce the implementation detail about D2R [126] in Table 4.2 and 4.3 because this is the first time to use its fine-tuned detector for localization purposes. D2R was originally designed to increase the image-level retrieval performance; it detects many candidate regions for each database image to get a better global feature. In this paper, we do not change the image-level ranking order for comparison fairness; Table 4.1 shows that our HD gives better results on the image-level ranking. For each proposed candidate region detected by D2R, we use the Aggregated Selective Match Kernel (ASMK) to aggregate all the DELG local features inside it. The region with the highest matching score with the ASMK feature of the query image is chosen as the final region. When aggregating the local features, we also tried the Vector of Locally Aggregated Descriptors (VLAD) and the simple average (Mean) aggregation. We report its result of using Faster-RCNN [103] or SSD [77] backbones fine-tuned on a huge manually boxed landmark dataset [126].

**Result**. In Tables 4.2 and 4.3, we present a comparative analysis of pixel-retrieval accuracy between our Hypergraph Diffusion (HD) method and other state-of-the-art (SOTA) techniques. Unlike conventional approaches that compute direct correspondences for each query-database image pair, HD employs an innovative strategy. It diffuses spatial information from the query to the database images, which significantly cuts down on time and computational

**Table 4.2:** Results of pixel retrieval from ground truth query-index image pairs (% mean of mIoU) on the PROxf/PRPar datasets with both Medium and Hard evaluation protocols. D and S indicate detection and segmentation results respectively. **Bold** number indicates the best performance, and underline indicates the second one. Our HD achieves both the best image-level (Table 4.1 ) and pixel-level retrieval accuracy among the retrieval and localization unified methods. Note that HD is much faster than SP, detection, segmentation, and dense matching methods, as shown in Table 4.3 and Sec. 4.6.

| Method | Medium | | | | Hard | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | PROxf | | PRPar | | PROxf | | PRPar | | Avg. |
| | D | S | D | S | D | S | D | S | |
| Retrieval and localization unified methods | | | | | | | | | |
| SIFT+SP [91] | 26.1 | 10.9 | 24.2 | 9.7 | 18.2 | 7.3 | 19.3 | 7.8 | 15.44 |
| DELF+SP [88] | 43.7 | 20.0 | 40.7 | 16.7 | 33.2 | 13.9 | 32.2 | 12.4 | 26.60 |
| DELG+SP [21] | **44.1** | 19.7 | 40.1 | 16.5 | **34.8** | 14.5 | 31.2 | 11.7 | 26.57 |
| D2R [126]+Resnet-50-Faster-RCNN+Mean | 20.2 | - | 29.6 | - | 16.7 | - | 27.4 | - | - |
| D2R [126]+Resnet-50-Faster-RCNN+VLAD [55] | 25.8 | - | 37.5 | - | 21.6 | - | 35.5 | - | - |
| D2R [126]+Resnet-50-Faster-RCNN+ASMK [128] | 26.3 | - | 38.5 | - | 21.6 | - | 35.6 | - | - |
| D2R [126]+Mobilenet-V2-SSD+Mean | 19.7 | - | 25.9 | - | 20.1 | - | 27.9 | - | - |
| D2R [126]+Mobilenet-V2-SSD+VLAD [55] | 23.1 | - | 33. | - | 20.9 | - | 33.6 | - | - |
| D2R [126]+Mobilenet-V2-SSD+ASMK [128] | 22.4 | - | 34.0 | - | 20.8 | - | 33.1 | - | - |
| HD (ours) | 29.7 | 24.7 | **49.2** | 38.2 | 22.6 | 19.4 | **39.1** | 31.0 | 31.74 |
| Detection methods | | | | | | | | | |
| OWL-VIT (LiT) [83] | 11.4 | - | 18.0 | - | 6.3 | - | 15.0 | - | - |
| OS2D-v2-trained [89] | 10.5 | - | 13.7 | - | 11.7 | - | 14.3 | - | - |
| OS2D-v1 [89] | 7.0 | - | 8.5 | - | 8.7 | - | 9.2 | - | - |
| OS2D-v2-init [89] | 13.6 | - | 15.4 | - | 14.0 | - | 15.1 | - | - |
| Segmentation methods | | | | | | | | | |
| SSP (COCO) + ResNet50 [35] | 19.2 | 34.5 | 31.1 | 48.7 | 15.1 | 25.3 | 29.8 | **41.7** | 30.68 |
| SSP (VOC) + ResNet50 [35] | 19.7 | 34.3 | 31.4 | **48.8** | 16.1 | **26.1** | 30.3 | 40.4 | 30.89 |
| HSNet (COCO) + ResNet50 [82] | 23.4 | 32.8 | 37.4 | 41.9 | 21.0 | 25.7 | 34.7 | 36.5 | 31.67 |
| HSNet (VOC) + ResNet50 [82] | 21.0 | 29.8 | 31.4 | 39.7 | 17.1 | 23.2 | 29.7 | 34.9 | 28.35 |
| HSNet (FSS) + ResNet50 [82] | 30.5 | **35.7** | 39.4 | 40.2 | 22.7 | 25.1 | 34.7 | 32.8 | **32.64** |
| Mining (VOC) + ResNet50 [150] | 18.3 | 30.5 | 29.6 | 42.7 | 15.1 | 21.4 | 28.1 | 34.3 | 27.50 |
| Mining (VOC) + ResNet101 [150] | 18.1 | 28.6 | 29.5 | 40.0 | 14.2 | 20.4 | 28.2 | 34.4 | 26.68 |
| Dense matching methods | | | | | | | | | |
| GLUNet-Geometric [134] | 18.1 | 13.2 | 22.8 | 15.2 | 7.7 | 4.6 | 13.3 | 7.8 | 12.84 |
| PDCNet-Geometric [135] | 29.1 | 24.0 | 30.7 | 21.9 | 20.4 | 15.7 | 20.6 | 12.6 | 21.87 |
| GOCor-GLUNet-Geometric [133] | 30.4 | 26.0 | 33.4 | 25.6 | 20.8 | 16.0 | 19.8 | 13.3 | 23.16 |
| WarpC-GLUNet-Geometric (megadepth) [136] | 31.3 | 25.4 | 36.6 | 27.3 | 21.9 | 15.8 | 26.4 | 17.3 | 25.25 |
| GLUNet-Semantic [134] | 18.5 | 14.4 | 22.4 | 15.6 | 8.7 | 5.6 | 12.8 | 7.8 | 13.22 |
| WarpC-GLUNet-Semantic [136] | 27.5 | 21.4 | 36.8 | 25.7 | 18.5 | 11.9 | 28.3 | 17.6 | 23.46 |

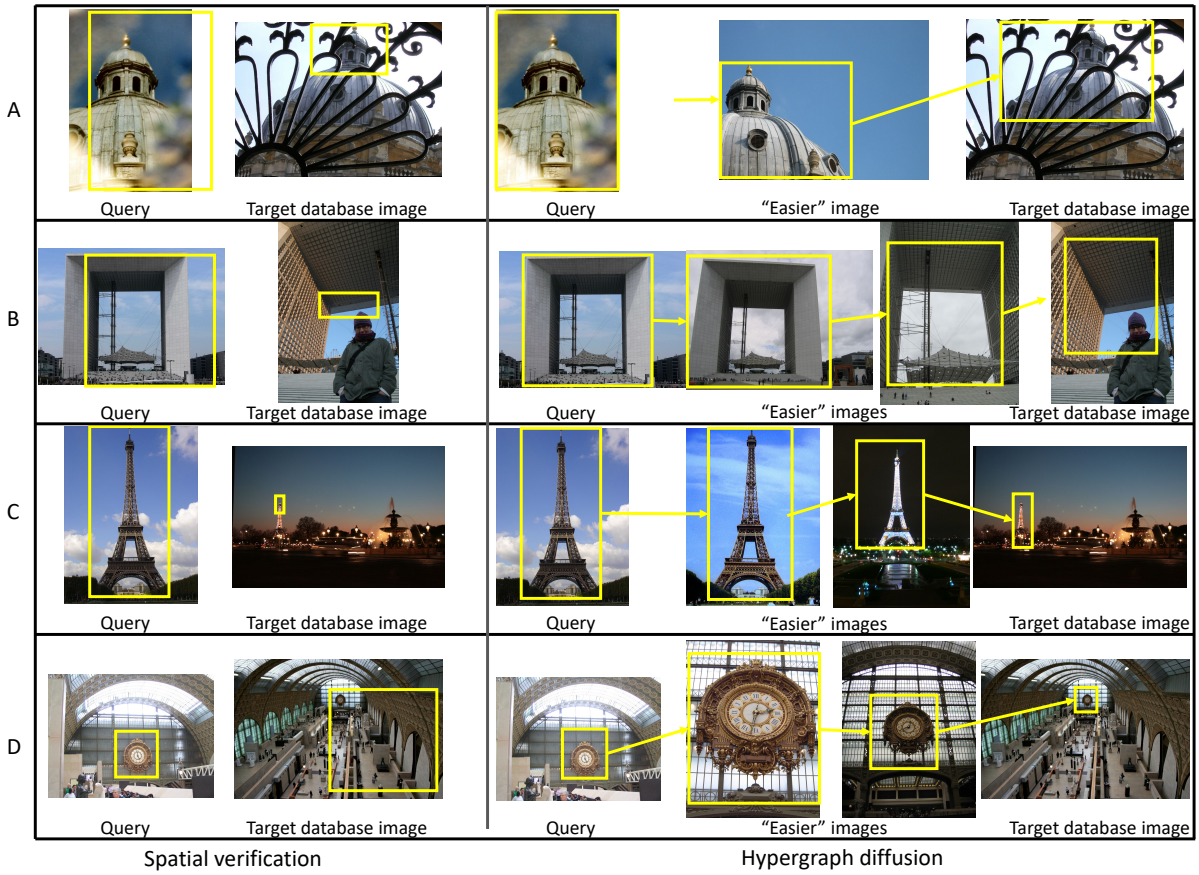



**Figure 4.4:** Visual examples of how Hypergraph Diffusion (HD) achieves better pixel retrieval results than direct SPatial verification (SP) using the same DELG features. The yellow boxes in query images are the cropped region given by the benchmark. Yellow boxes in SP are the minimum bounding box of the matching points. Yellow boxes with line arrows indicate the diffusion paths in HD from the query region to the correspondence region in the target database image. In A, an 'easier' database image offers a clear, unoccluded view of the target query landmark, demonstrating improved matching. B highlights how these database images provide viewpoints more aligned with the query image, enhancing object matching. C shows the advantage in scenarios with varying illumination, where 'easier' images assist in achieving more accurate matches. Lastly, D reveals that when the target database image has a more complex background than the query image, direct spatial verification can lead to outlier matchings. In contrast, the 'easier' database images selected by HD provide additional context, thus facilitating more precise pixel retrieval.

resources. Notably, despite initial expectations of a trade-off between speed and accuracy, our results reveal that HD's accuracy is not uniformly inferior to direct Spatial verification (SP) methods. Using the same DELG local features, HD demonstrates a reduced accuracy on the PROxford dataset but excels in the PRParis dataset compared to SP.

The underlying reason for this performance differential might lie in our diffusion approach. Rather than matching query and database images directly via SP, HD initially identifies correspondences in more straightforward database images and subsequently tackles more challenging ones. This stepwise diffusion process effectively mitigates the impact of viewpoint and illumination variations between image pairs, breaking down severe changes into manageable phases.

To validate this hypothesis, we conducted visualizations, as depicted in Fig. 4.4. The visual evidence suggests that while the direct SP struggles to localize correspondence objects accurately in complex cases, our HD method

**Table 4.3:** Results of pixel retrieval from database (% mean of mAP@50:5:95) on the PROxf/PRPar datasets and their large-scale versions PROxf+1M/PRPar+1M, with both Medium (M) and Hard (H) evaluation protocols. **Red** indicates the best performance and **Bold** indicates the second best performance. The speed of hypergraph propagation (HP) in the reranking stage is 0.23s per 100 image pairs; it does not add any additional time or computing cost for pixel-level matching/segmentation. Compared with DELG+SP, OWL-ViT, SSP, and WarpCGLUNet, HD is much faster while maintaining high accuracy. Compared with D2R+Faster-RCNN+ASMK, HD achieves higher accuracy on both pixel retrieval and image retrieval (Table 4.1).

| | | PROxf | | PROxf+R1M | | PRPar | | PRPar+R1M | | Overhead per 100 pairs |
|---|---|---|---|---|---|---|---|---|---|---|
| | | M | H | M | H | M | H | M | H | |
| Image retrieval: DELG initial ranking [21] + HD reranking [6] | | | | | | | | | | |
| Pixel retrieval methods | DELG + SP [21] | **39.6** | 30.5 | **36.0** | 28.2 | 34.8 | 20.2 | 34.7 | 19.5 | 41.22s |
| | D2R+Faster-RCNN+ASMK [126] | 30.1 | 23.5 | 30.5 | 22.0 | 26.3 | 25.3 | 25.7 | 24.9 | 0.11 s |
| | OWL-VIT [83] | 12.3 | 6.6 | 12.1 | 13.6 | 7.9 | 7.6 | 7.9 | 7.8 | 296.21s |
| | SSP [35] | **33.0** | 29.7 | **35.7** | 30.5 | **46.4** | **37.2** | **45.6** | **37.2** | 62.33 s |
| | WarpCGLUNet [136] | 31.2 | **32.6** | 31.5 | **31.7** | 34.1 | 27.3 | 34.3 | 28.1 | 181.64s |
| | HD | 26.9 | **39.7** | 25.3 | **33.1** | **43.7** | **39.6** | **43.8** | **38.9** | 0s |

achieves superior localization. It starts by detecting correspondences in less complex images, gradually extending to more challenging ones. Intriguingly, the 'easier' database images contain additional visual cues absent in the query images and significantly aid the RANSAC-based SP in matching objects in more complex scenarios. For example, Fig.4.4-A illustrates how an 'easier' database image provides a clearer, unoccluded view of the target query landmark. Similarly, Fig.4.4-B shows that these database images offer viewpoints more closely aligned with the query image, aiding in object matching. In scenarios with changing illumination, as shown in Fig. 4.4-C, 'easier' database images help mitigate this change, facilitating more accurate matching. Furthermore, in Fig. 4.4-D, we observe that the target database image has a more complex background than the query image. Note that the input query is only the yellow region. This complexity often leads to outlier matchings in direct SP. However, the 'easier' database images selected by our HD method provide additional background context, resulting in more accurate pixel retrieval.

Tables 4.2 and 4.3 illustrate that HD surpasses current SOTA detection methods capabilities and offers competitive, if not superior, performance than segmentation and dense-matching methods. Significantly, our approach merges image-level and pixel-level analyses, outpacing traditional two-step methods by delivering pixel-level results simultaneously with image-level rankings. This efficiency, crucially discussed in Section 4.6, positions HD as a robust and efficient solution for pixel retrieval tasks. It shows that shifting the spatial verification to an offline process not only enhances image-level performance but also drastically reduces retrieval time, achieving good pixel-retrieval performance. Considering accuracy and speed, HD emerges as a compelling choice for pixel retrieval.

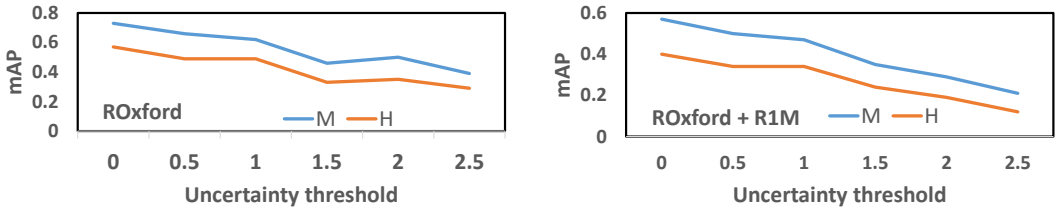



**Figure 4.5:** mAP of the query results above the uncertainty threshold. Uncertainty predicts the query quality well.

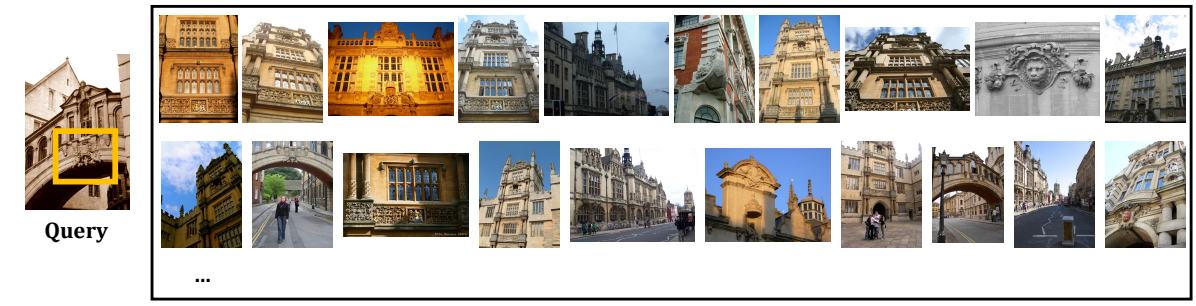



**Figure 4.6:** In this example, the top 11 images and the initial dominant community is unrelated to the query. Its uncertainty index is 1.77, which means the search engine is very uncertain about the accuracy of initial dominant community. So CS finds a new dominant community using spatial verification. After hypergraph propagation in this new dominant community, the retrieval performance is significantly improved.

### 4.5.5 Effectivness of community selection

Figure 4.5 shows the mAP of the initial search results in ROxford with different uncertainty thresholds. As the uncertainty increases, the mAP of the initial search result decreases. The uncertainty index predicts the quality of the initial search without any user feedback. Only for the high uncertainty retrieval results, we use the computationally heavy approach, i.e., spatial verification, to improve the precision of nearest neighbors before the hypergraph propagation.

Table 4.4 compares the performances between with and without community selection (CS) as the pre-stage when using spatial verification (SP) to initialize the hypergraph propagation (HP). With CS, the search engine improves the performance of hypergraph propagation in ROxford very efficiently by only conducting SP on 4 and 7 queries with and without R1M distractors, respectively. The performance improvement with R1M distractors is less than that without them. We think its reason is that the correct items with the R1M distractors are ranked outside the top

**Table 4.4:** Comparison between with and without community selection (CS) as the pre-stage when using spatial verification (SP) to initialize the hypergraph propagation (HP). This table shows both the mAP and the total number of queries which need SP when testing the corresponding dataset.

| | HP | CS + SP + HP | | SP + HP | |
|---|---|---|---|---|---|
| | mAP | mAP | total # SP | mAP | total # SP |
| ROxf(M/H) | 85.7/70.3 | 88.4/73.0 | **4** | 88.4/73.0 | 70 |
| ROxf+R1M(M/H) | 78.0/60.0 | 79.1/60.5 | **7** | 79.1 / 60.5 | 70 |
| RPar(M/H) | 92.6/83.3 | 92.6/83.3 | **0** | 92.6/83.3 | 70 |
| RPar+R1M(M/H) | 86.6/72.7 | 86.6/72.7 | **0** | 86.6/72.7 | 70 |

**Table 4.5:** Average data size per database image (in Byte), measured in Numpy array format. Spatial verification (SP) averagely requires 1040000 bytes, whereas Hypergraph Diffusion (HD) needs only 2678. HD demonstrates a significantly higher memory efficiency, being 388 times more efficient than SP.

| Global feature | Matching information for HD | Local features for SP |
|---|---|---|
| 8192 B | 2678 B | 1040000 B |

**Table 4.6:** Breakdown of average time per query.

| Initial search | Hypergraph diffusion | Uncertainty calculation |
|---|---|---|
| 0.62 s | 1.07 s | 0.0003 s |

100 images and thus are not detected by SP. Although CS does not improve the retrieval results of these cases, it successfully tags them as uncertain cases. We think this result leads to a promising future direction. Because once a search engine can automatically detect the failed query cases, it is possible to refine the representation ability by relabeling and training only for the failed cases instead of training on a large dataset.

Thanks to the improved representation ability of DELG [21] features, the nearest neighbors of all queries in RParis dataset have high precision. The average precisions of the top 20 items of the initial result of all the queries in RParis are 0.97 and 0.90 in medium and hard protocols, respectively. As the initial dominant communities for all the queries in RParis are correct, online SP is not helpful for the propagation process. CS avoids these unnecessary SPs in these cases as the uncertainties of all the queries are low, as shown in Table 4.4. CS significantly improves the retrieval performance for some hard cases. Figure 4.6 shows an impressive example in ROxford, in which the mAP increases from 0.077 and 0.062 to 0.988 and 0.987 in medium and hard protocols, respectively.

In summary, the main benefits of CS include: 1) CS quickly detects the low-quality retrieval result, which is either because of the difficult query or the insufficient feature representation ability; 2) CS avoids the unnecessary computationally heavy spatial verification process; 3) CS provides significant performance improvement for some hard cases.

## 4.6 Time and memory cost

**Table 4.7:** Comparison of time complexity and latency across different pixel-retrieval approaches. K indicates the number of images to calculate the pixel-level result. h (set as 3 in this paper) indicates the maximum iteration in hypergraph diffusion. r (set as 1000, the most common choice in this field) is the maximum RANSAC iteration time. l and d are the local features' numbers and dimensions, 1000 and 1024 for DELG. n is the image resolution or feature map size, f is the representation dimension, k is the kernel size of convolutions, and L is the number of Transformer or CNN layers. When testing the latency, we use an i9-9900K CPU @ 3.60GHz for HD and SP and a GeForce RTX™ 3090 Ti for deep learning methods. We use OWL [83] and WarpCGLUNet [136] as the representative transformer-based and CNN-based models, respectively. Our method is much faster than the existing approaches for the pixel retrieval task.

| Method | Hypergraph diffusion (ours) | Spatial verification | Transformer backbone | CNN backbone |
|---|---|---|---|---|
| Time complexity per query | $O(h)$ | $O(K \times r \times l^2 \times d^2)$ | $O(K \times L \times n^2 \times f)$ | $O(K \times L \times c \times n \times f^2)$ |
| Latency per 100 image pairs | 0.23s | 41.22s | 296.2s (OWL [83]) | 181.6s (WarpCGLUNet [136]) |

In this section, we focus on assessing the time and memory costs associated with Hypergraph Diffusion (HD) in retrieval systems. While the accuracy of HD has been addressed in previous sections, here we aim to shed light on its efficiency and resource utilization, essential factors for practical deployment.

For hypergraph construction during query execution, spatial matching information for each database image and its k-nearest neighbors (with k set to 200 as introduced in Sec. 4.5.2) is recorded. Table 4.5 presents a comprehensive breakdown of memory usage. The matching information occupies an average of 2,678 bytes per database image. Notably, this memory usage constitutes only about 33% of the memory required for the DELG global feature and a mere 0.2% of that for local features. Note that HD does not need to load the local features in query time. This comparison illustrates HD's memory efficiency.

In Table 4.6, we show the average query time during our experiments on the ROxford + R1M dataset. HD averages 1.07 s per query, while the initial search takes 0.62 s. This overhead, we argue, is reasonable given the benefits in accuracy and memory efficiency. Furthermore, the latency for uncertainty checks via our community selection technique is impressively minimal at 0.0003 s, bolstering our claim about the method's efficiency.

In Table 4.7, we compare the time complexity and latency of various pixel retrieval methods, including HD. Unlike Spatial Verification (SP) and neural network-based approaches, HD leverages pre-calculated matching information, leading to a time complexity linearly dependent on the diffusion process's maximum iteration number ($O(h)$). This simplicity starkly contrasts with the complexity of SP and the inference latency inherent in neural network methods, as shown in Table 4.7. To provide a tangible comparison, we document the latency experienced in processing 100 image pairs using different methods, with HD showing a marked advantage in speed. This comparative analysis underpins the practical viability of HD in time-sensitive applications. Our experiments used an i9-9900K CPU @ 3.60GHz for HD and SP and a GeForce RTX™ 3090 Ti for deep learning methods. This hardware setup reflects a high but realistic standard in computational resources, ensuring that our results represent potential real-world deployments.

## 4.7 Discussion and conclusion

Our research introduces a novel approach to both image-level and pixel-level retrieval, leveraging hypergraph diffusion and community selection. This method has demonstrated high accuracy and remarkable efficiency in both time and memory usage, positioning it as a strong candidate for advanced retrieval systems. Despite these promising results, we acknowledge several areas for future exploration and improvement:

**Reducing Offline Computing Overhead**: Currently, our system relies on multiple CPUs for offline spatial verification, requiring approximately 12 hours to process the ROXford dataset on a single CPU. A transition to a GPU-based implementation is anticipated to enhance offline computing speed significantly.

**Enhancing Online Processing Speed**: The online processing speed, directly related to the number of diffusion steps in hypergraph diffusion, presents an opportunity for optimization. While our current implementation uses Python, transitioning to a C implementation could substantially improve processing speed, which is particularly important for real-time applications.

**Optimizing for Large-Scale Queries**: Improving the system's performance in large-scale query scenarios remains a critical challenge. Efficient handling of extensive datasets and queries is essential for practical applications.

**Database Management Efficiency**: An important aspect to consider is the efficient addition and deletion of images

in the database. This process involves applying spatial verification for newly added database images, which requires efficient management to maintain system performance.

In conclusion, we believe that our approach combining hypergraph propagation with community selection takes a meaningful step in the realm of image and pixel search techniques. Notably, our method has achieved impressive mean Average Precision (mAP) scores of 73.0 and 60.5 on the ROxford dataset's hard protocol, with and without R1M distractors, respectively. To the best of our knowledge, these results represent state-of-the-art achievements in this field. We hope that our work lays a solid foundation for future research and development in efficient and accurate image retrieval systems.

# Chapter 5. Topological RANSAC for instance verification and retrieval without fine-tuning

## 5.1 Introduction and related work

Content-based image retrieval, a fundamental and long-standing challenge, has seen significant advancements due to the rise of deep learning [98, 142, 21, 88]. These data-driven approaches, although successful, often rely heavily on fine-tuning with data from the same domain [88, 143]. However, acquiring such a fine-tuning set can be impractical or costly, particularly in open-world or private scenarios. Moreover, these methods often fall short in providing explainability, a critical factor for real-world applications where search results are integral to decision-making.

In situations where a fine-tuning set is absent, the esteemed SPatial verification (SP) [91] method has demonstrated the highest accuracy in verifying an image pair [97]. However, its dependency on a spatial model for recognition exposes vulnerabilities, particularly to viewpoint changes in 3D objects and neglect of topological relations among key points (Sect. 5.2.1). To address these issues, we pioneer the replacement of the spatial model with a topological one during the RANdom SAmple Consensus (RANSAC) process. We introduce Homeomorphism Region (Def. 5.2.1), which circumvents the mentioned problems, providing a size metric superior to the original inlier counts used in SP. Drawing inspiration from human observation, we propose innovative saccade and fovea functions to validate topological relations among keypoints within RANSAC [37] iterations (Sect. 5.3).

Numerous studies have explored SP [4, 100, 66, 113], but none have made such a direct modification to the spatial model in RANSAC, possibly because earlier benchmarks presented the issues of the spatial model as less severe [97, 91, 92]. However, with the introduction of more challenging benchmarks like ROxford and RParis [97], these problems have become increasingly critical. While some recent deep-learning approaches have explored feature relations, their effectiveness diminishes without a fine-tuning set, and they often lack explainability [142, 123]. In contrast, by adopting the hypothesis-testing strategy with topological rules, our method outperforms even large-scale pre-trained methods in non-fine-tuning scenarios, maintaining high explainability and lightweight nature. Moreover, it can further enhance performance when used alongside fine-tuned features. These results show our novel attempt at topological-based RANSAC is a success.

Our contributions are summarized as follows:

- We propose a novel approach to directly modify the underlying common sense in hypothesis testing. Our method is pioneering in adopting topological common sense, diverging from traditional spatial approaches. This shift paves the way for numerous future research opportunities.
- Drawing inspiration from the human observation process, we introduce innovative saccade and fovea functions to verify topological relations among keypoints
- Our method has demonstrated a significant improvement in accuracy by adopting the hypothesis-and-test strategy with topological rules. Our method surpasses all other methods without fine-tuning. Moreover, our approach can be combined with fine-tuned features to achieve even better performance.
- Our method offers a highly explainable and lightweight solution, which is crucial for real-world applications.

## 5.2 Motivation and overview

### 5.2.1 Spatial verification

RANdom SAmple Consensus (RANSAC) is an effective method for handling noisy data [37]. It involves using a hypothesis and test strategy to optimize model parameters to accurately describe inliers iteratively. The model selected is crucial for achieving high inference accuracy. For instance recognition, Spatial verification (SP) [91] employs a spatial model to match features in two sets of points, $P_1$ and $P_2$. It finds applications in a variety of contexts, such as reranking in image search [21], facilitating loop closure detection [64], and serving as a preliminary step in 3D reconstruction [110]. By using RANSAC, SP calculates a transformation matrix $M$ to account for the spatial configuration change between different views. The similarity between the two images can then be determined using the following equation:

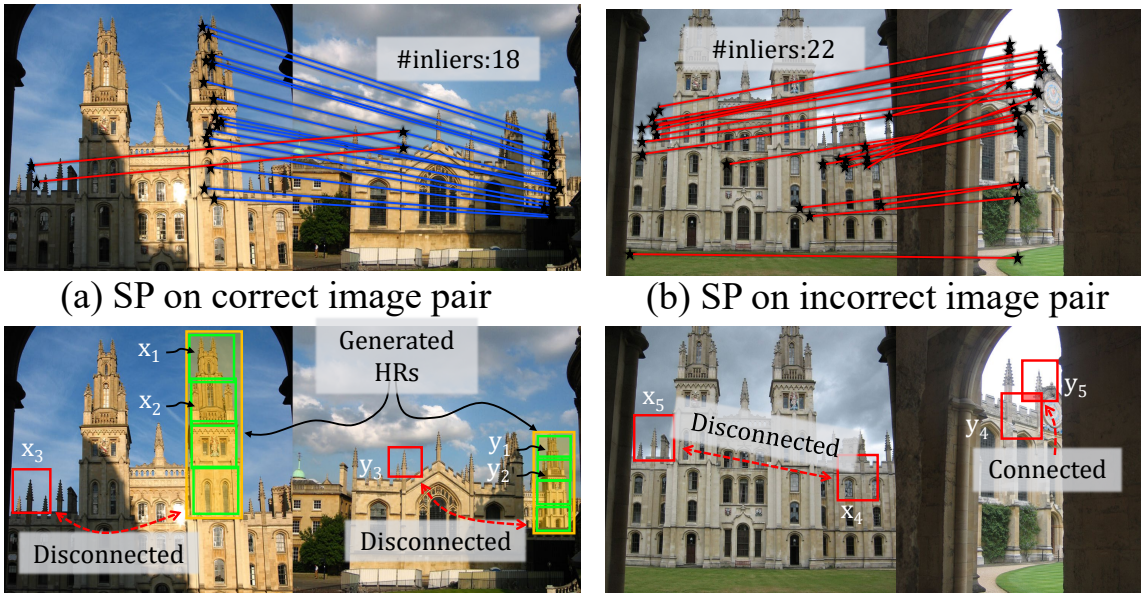
$$D_s(P_1, P_2|M) = \sum_{p_1^i \in P_1} [\![\|\mathbf{p}_1^i M - \mathbf{p}_2^i\| < \epsilon]\!], \tag{5.1}$$

where $\mathbf{p}_1^i$ and $\mathbf{p}_2^i$ represent the locations of $p_1^i \in P_1$ and $p_2^i \in P_2$, respectively. $p_2^i$ is the feature in $P_2$ that is most similar to $p_1^i$. The threshold for the location disparity is denoted by $\epsilon$.

The spatial model is intrinsically intertwined with RANSAC in the field of computer vision [30, 23], and it is particularly well-suited for tasks such as pose estimation where achieving an accurate homography or fundamental matrix is the primary objective [119, 85]. However, despite the wealth of research surrounding the use of RANSAC for pose estimation, its integration within rigid body recognition remains under-discussed in the existing literature. This paper endeavors to address this gap, offering insights into why a spatial model may not be ideally suited for recognition tasks.

SP has been the subject of numerous studies [4, 100, 66, 113]. However, none of the existing works have sought to modify the fundamental spatial model. On the other hand, while near-perfect results have been reported on the original Oxford and Paris benchmarks [91, 92], newer challenges such as complex positives and confusing distractors have recently emerged [97]. To address these challenges, contemporary research often bypasses the RANSAC process, focusing instead on enhancing performance through fine-tuning [98, 21, 88, 123]. Nevertheless, this strategy may not be feasible in private or dynamic scenarios where fine-tuning datasets are unavailable. As discussed in Section 5.5, tasks involving non-fine-tuning recognition are of significant importance. To the best of our knowledge, in the absence of a fine-tuning set, RANSAC-based SP still stands as the most effective method [97] in ROxford and RParis. This paper aims to delve into the often-overlooked aspect of enhancing the RANSAC process within rigid instance recognition tasks.

In our study, we identified two significant issues when spatial model is deployed for recognition. Firstly, SP assumes that two images depicting the same plane from different perspectives can be matched using an affine transformation—an assumption that may not always hold true. Real-world objects are often three-dimensional, not planar, which means the 3 by 3 affine matrix employed in SP may not accurately represent their variations. Secondly, SP's sparse inlier count of the predicted homography matrix neglects the topological relationships among features and the relevance of other image regions. This issue is evident in Fig. 5.1, which presents SP results for both a correct and an incorrect image pair. Notably, the incorrect image pair generates more inliers than the correct pair, an outcome that starkly illustrates the limitations of using inlier count as the sole measure of accuracy in image recognition tasks. It is important to note that the two issues we've identified with SP primarily originate from the spatial model itself rather than from the hypothesis-testing strategy. This realization motivates us to modify the spatial model while retaining the hypothesis-testing approach.




(a) SP on correct image pair (b) SP on incorrect image pair

(c) Ours on correct image pair (d) Ours on incorrect image pair

**Figure 5.1:** This figure presents our novel adaptation of the SPatial model (SP) [91] into the ToPological model (TP), marking the first such transformation in hypothesis-testing. Sub-figures (a) and (b) highlight an instance where SP incorrectly assigns more inliers to an erroneous image pair (22 inliers) compared to the correct pair (18 inliers). Conversely, sub-figures (c) and (d) demonstrate our TP model accurately discerns the right image pair and dismisses the wrong pair by iteratively refining the Homeomorphism Regions (HRs) based on patch connections.

### 5.2.2 Topological verification

In contrast to existing works that rely on fine-tuning to enhance performance [123, 21, 88], our study pioneers the use of a topological model in place of the spatial model within RANSAC. Importantly, this approach remains viable even when fine-tuning is not feasible. We introduce the concept of a Homeomorphism Region (HR) to address two primary issues associated with the spatial model. Its size is a new metric for RANSAC iteration and image pair similarity. The formal definition of HRs is in Definition 5.2.1.

**Definition 5.2.1 (Homeomorphism region)** *Let $r$ be a small image patch of image $I$, $r_1^n$ and $r_2^n$ be the corresponding patches in $I_1$ and $I_2$, $R_1 = \{r_1^1, ..., r_1^n\}$ and $R_2 = \{r_2^1, ..., r_2^n\}$ be the families of patches of images $I_1$ and $I_2$. $R_1$ and $R_2$ are Homeomorphism regions (HRs), if they satisfy the following conditions:*

*1. Local consistency: $r_1^n \in R_1$ and $r_2^n \in R_2$ are identified as the same patches based on their local descriptors.* [Two corresponding patches should be similar; they are supposed to depict the same region of an object.]

*2. Topological consistency: $\forall(r_1^n, r_1^m, r_2^n, r_2^m)$, if $r_1^n$ and $r_1^m$ overlap each other, $r_2^n$ and $r_2^m$ should also overlap each other, and vice versa.* [Parts of an object that are connected should remain so, even when the viewpoint changes.]

*3. Connectivity: $\forall r_1^n \in R_1, \exists r_1^m \cap r_1^n ! = \emptyset, r_1^m \in R_1$* [We do not consider similar but isolated patches. Note that wrong similar patches $x_3$ and $y_3$ in Fig. 5.1 are excluded from the HR in this way.]

Our novel topological perspective effectively circumvents the aforementioned problems, yielding more robust results than its spatial counterpart. In contrast to SP's planar assumption, our approach posits that parts of an object that are connected should remain so, even when the viewpoint changes. This assumption holds true for 3D objects.

Furthermore, our topological approach considers the topological relationships between patches and leverages more information in an image pair. For example, incorrect SP matches can occur when there are numerous repeated patterns, even if these matched pairs aren't topologically consistent with other features, as shown in Fig. 5.1 (b). Our model, however, correctly identifies these as different structures. As demonstrated in Fig. 5.1 (c) and (d), our method successfully discriminates between correct and incorrect image pairs, highlighting its potential in enhancing image recognition tasks.

## 5.3 Method detail

**Algorithm 1** Topological inference

**Input:** HypothesisSet: $\{(r_1^1, r_2^1),(r_1^2, r_2^2),...\}$.
**Output:** : Regions: list of HRs

```
Regions ← []
for (r_1^i, r_2^i) in HypothesisSet do
    π ← []
    π′ ← [(r_1^i, r_2^i)]
    while π′ is not empty do
        (r_1, r_2)= pop(π′)
        r̂_1, r̂_2, verified=F(r_1, r_2)
        if verified = True then
            Add (r̂_1, r̂_2) to π
            π′=S(π)
        end if
    end while
    add π to Regions.
end for
```

To find HRs, we are inspired by our brain's mechanism of comparing two first-seen objects. Our eye mostly captures low-resolution visual information except for a tiny patch called the fovea. The fovea only observes a very small area and the entire scene illusion is created by stitching several glimpses of the fovea during eye movements called saccade [51, 43]. Inspired by this process, we design our saccade and fovea functions to detect the HRs that satisfied the Def. 5.2.1.

Algorithm 1 shows the overall pipeline of our method. Traditionally, RANSAC samples a subset and computes a hypothesis based on this selection. We modify this approach by using each ratio test matching result as an individual hypothesis. Therefore, the size of our sample subset is effectively reduced to one, with each iteration's hypothesis being the translation, as opposed to the affine matrix. Instead of counting inliers based on the spatial constraint, we detect HRs for each hypothesis, the size of which is then used as the new iteration metric. We recognize that the sampling strategy and convergence proof are the focal points of ongoing RANSAC research. However, as the first study to adopt a topological perspective, this paper primarily seeks to demonstrate the feasibility and advantages of this paradigm shift.

### 5.3.1 Topological consistency and connection

To verify the connection and topological consistency, we progressively verify and select the patches one by one. It can be expressed as follows:

$$\pi'_t = S(\pi_t), \pi_{t+1} = F(\pi'_t), \tag{5.2}$$

where $S$ is the saccade function to guide the observation place, $F$ is the fovea function to verify two patches, $\pi_t$ records the verified patches at $t$-th iteration, $\pi_0$ is the given hypothesis about translation, and $\pi'_t$ records the candidate patches generated by $S$. $\pi_t$ and $\pi'_t$ are shown using green and purple boxes in Fig. 5.2. After finishing this verification progress, the final verified region $\pi_T$ will be treated as the HR defined in Section 3.1. The saccade function $S$ makes the new candidate patch intersect with one and only one verified patch in $\pi_t$, as shown in Fig. 5.2. In this way, a series of verified patches are connected with each other and also satisfy the topological consistency.

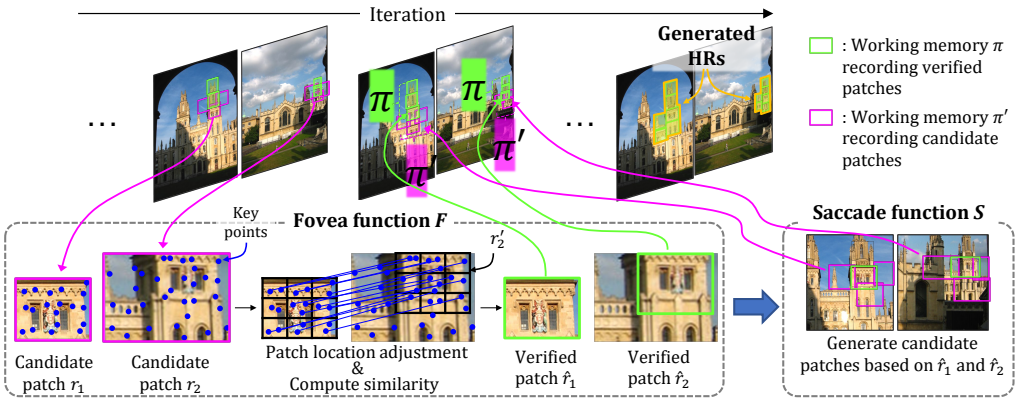



**Figure 5.2:** The process of finding HRs for each hypothesis. This is an iterative method. Fovea function F verifies the similarity between two candidate patches at each iteration, and saccade function S updates the candidate patches.

Under the assumption that two images are depicting the same object with some transformations (zoom, foreshortening, vertical shearing), the saccade function $S$ predicts the area around the verified patch pair will be similar. For example, $\hat{r}_1$ and $\hat{r}_2$ in Fig. 5.2 are two verified patches, and saccade function $S$ expects the patches around them are also similar. The correspondence patch location and size will be refined by $F$, as shown in the next section. Similar to [91], we do not consider in-plane image rotations because of the fact that images are usually displayed with the correct orientation.

### 5.3.2 Local consistency

The fovea function $F$ concentrates on two patch areas and verifies if the query patch $r_1$ and test patch $r_2$ are similar based on the local feature sets $P_1$ and $P_2$ of images $I_1$ and $I_2$.

**Feature matching.** The first step for verification is to match the key points in two images correctly. Traditional geometric verification directly searches the nearest neighbor of a keypoint in $I_1$ from keypoints in $I_2$ based on their descriptors. However, we observe this approach may have many wrong matching points in practice due to the repeated patterns in an image, as shown in Fig. 5.1. To address this problem, we restrict the search region when implementing the fovea function. That is, for each key point in patch $r_1$, we find its nearest neighbor among keypoints in patch $r_2$ instead of $I_2$. We find that this approach increases the matching performance.

**Patch location adjustment.** We need to adjust the location and size of the candidate patches here. Because $r_1$ and $r_2$ are candidate patches generated by saccade function $S$ only based on the verified region $\pi$, their contents may not be exactly same. As shown in Fig. 5.2, the scope of patch $r_2$ is different from patch $r_1$. After the feature matching, we treat the $r_2$ region containing the matched points as the adjusted corresponding patch $r_2'$, as shown in Fig. 5.2.

**Locally spatial constraint.** Two main challenges of object verification are the location disparity of the key points due to the view change of two images and the matching outliers. Fortunately, our saccade function $S$ has separated an object into patches. It offers three benefits. Firstly, because the patch size is smaller than the whole image, a patch is more likely to depict the 2D plane of a 3D object. Locally using the spatial constraint like Eq. (5.1) can be more robust than the origin SP. Secondly, we can assume the center points of patches $r_1$ and $r_2'$ are correct if they share the same pattern. That is we assume $\mathbf{c}_1 = \mathbf{c}_2 = \mathbf{c}$, where $\mathbf{c}_1$ and $\mathbf{c}_2$ are the center locations of $r_1$ and $r_2'$. Treating $\mathbf{c}$ as the reference point, the disparity of two correct matching key points $\mathbf{p}_2^i$ and $\mathbf{p}_1^i$ is $(M - I)(\mathbf{p}_1^i - \mathbf{c})$,

where $I$ is the identity matrix. So Eq. (5.1) can be rewritten as:

$$F = \sum_{p_1^i \in r_1} [\![ \| (\mathbf{p}_2^i - \mathbf{p}_1^i) - (M - I)(\mathbf{p}_1^i - \mathbf{c}) \| < \epsilon ]\!]. \tag{5.3}$$

Like [91], we do not consider the in-plane rotations and overturn. We try to infer a threshold of disparity $(\mathbf{p}_2^i - \mathbf{p}_1^i)$ based on $(\mathbf{p}_1^i - \mathbf{c})$ without calculating a correct $M$. Note that the upper bound of $\left|\mathbf{p}_1^i - \mathbf{c}\right|$ is half of the patch diagonal length.

Thirdly, because fovea function only focuses on the regional patterns, we can assume there is no unrelated object or occlusion in $r_1$ and $r_2'$. Thus, we use the inlier ratio to replace the inlier number of SP:

$$F = \frac{1}{|r_1|} \sum_{p_1^i \in r_1} [\![ \| (\mathbf{p}_2^i - \mathbf{p}_1^i) - (M - I)(\mathbf{p}_1^i - \mathbf{c}) \| < \epsilon ]\!], \tag{5.4}$$

where $|r_1|$ is the number of local features in $r_1$. This change is to use probability to deal with the uncertainty of feature matching. Although it is difficult to guarantee the accuracy of a specific local feature matching, we expect the clear different inlier ratio distributions between the correct patches and in-correct patches with enough matching samples.

**Implementation.** We simplify Eq. (5.4) for the easy of computing. For $p_1^i \in r_1$, the expectation location of its corresponding point $p_2^i \in r_2'$ is $\mathbf{p}_1^i + (M - I)(\mathbf{p}_1^i - \mathbf{c})$. Because photos are taken from a restricted range of canonical views, transformation $M$ varies in a restricted space. We try to predict the location $\mathbf{p}_2^i$ based on $(\mathbf{p}_1^i - \mathbf{c})$ without calculating a correct $M$. To do so, we divide $r_1$ and $r_2'$ into 3 by 3 sub-patches and each sub-patch contains some points. That is $r_1 = [a_1, a_2, ..., a_9]$ and $r_2' = [b_1, b_2, ..., b_9]$, where $a_k$ and $b_k$ are correspondence sub-patches. For a matching point pair $p_1^i \in r_1$ and $p_2^i \in r_2'$, if $p_1^i$ locates in $a_k$, we expect its corresponding point $p_2^i$ will be in $b_k$. In this way, we simplify the Eq. (5.4) as:

$$F(r_1, r_2') = \sum_{k=1}^{9} f(a_k, b_k), f(a_k, b_k) = \frac{1}{|a_k|} \sum_{p_n \in a_k} k(p_n, p_n'), \tag{5.5}$$

where $p_n$ and $p_n'$ are two matched points, $k(p_n, p_n') = 1$ if $p_n' \in b_k$, otherwise $k(p_n, p_n') = 0$.

## 5.4 Experiment

### 5.4.1 Experiment setting

In our study, we contrast our novel ToPological verification (TP) approach with established methods like SP, Vector of Locally Aggregated Descriptors (VLAD) [55], and Aggregated Selective Match Kernel (ASMK) [127], employing standard SIFT features for the comparison. For SP, we utilize both standard RANSAC and Graph-Cut RANSAC (GCRANSAC) [13], recognized as one of the superior RANSAC methods. The conventional comparison for SP involves reranking the top 100 images found by ASMK using the same local feature [97]. As advanced initial ranking methods can introduce more challenging images and elevate the difficulty [97], we opt for the state-of-the-art fine-tuned feature DELG [21] over SIFT for initial ranking. For fairness, all methods rerank the top 100. Though our primary focus is not on enhancing representative learning, we contrast our method with state-of-the-art (SOTA) pre-training and fine-tuning techniques to underscore the importance of this inference direction and its potential for future exploration. We benchmark our method against pre-trained methods CLIP [99] and Superpoint [31]. In terms of the fine-tuning approach, we choose methods fine-tuned on the extensive, cleaned Google Landmark Dataset (GLD) [143]. We notice that some recent studies claim that GLD is not a fair fine-tuning

**Table 5.1:** Results (% mAP) on the ROxf/RPar datasets and their large-scale versions ROxf+1M/RPar+1M, with both Medium and Hard evaluation protocols.

| Method | ROxf | | ROxf+R1M | | RPar | | RPar+R1M | |
|---|---|---|---|---|---|---|---|---|
| | M | H | M | H | M | H | M | H |
| compare inference ability using SIFT | | | | | | | | |
| VLAD [55] | 37.6 | 21.7 | 31.4 | 10.7 | 83.2 | 69.1 | 69.5 | 44.7 |
| ASMK [127, 146, 142] | 49.6 | 29.1 | 41.4 | 18.0 | 83.6 | 69.5 | 66.2 | 44.4 |
| SP (with ratio test) [97, 91] | 64 | 42 | 56.3 | 30.5 | 86.3 | 71.6 | 70.1 | 46.4 |
| SP (without ratio test) [97, 91] | 44.4 | 23.3 | 37.3 | 13.5 | 83.7 | 69.0 | 66.9 | 44.2 |
| GCRANSAC [13] | 63.3 | 40.2 | 55.7 | 29.7 | 85.9 | 71.0 | 69.6 | 45.7 |
| TP with SP (Ours) | **68.6** | **46.3** | **59.2** | **33.9** | **86.6** | **71.7** | **70.5** | **46.6** |
| TP without SP (Ours) | **69.5** | **46.9** | **60.4** | **34.4** | **86.5** | **71.6** | **70.6** | **46.8** |
| vs. pre-trained models | | | | | | | | |
| CLIP (RN101) [99] | 46.4 | 21.9 | 39.4 | 12.5 | 85.1 | 70.0 | 68.9 | 45.7 |
| CLIP (ViT-B/32) [99] | 47.5 | 24.0 | 38.0 | 13.2 | 83.9 | 68.9 | 69.2 | 46.1 |
| Superpoint [31] | 66.1 | 42.4 | 59.4 | 33.4 | 85.9 | 71.0 | 70.0 | 46.1 |
| vs. large-scale fine-tuned models | | | | | | | | |
| GeM [98]+CL [98] (GLD) [143] | 71.0 | 49.1 | 57.1 | 30.3 | 86.6 | 72.4 | 70.2 | 46.6 |
| GeM [98]+AP [104] (GLD) [143] | 69.9 | 48.2 | 55.1 | 29.1 | 86.8 | 72.7 | 70.5 | 47.1 |

set for ROxford and RParis because there are overlapped landmarks [142]. We agree with this claim. However, we still compare our approach with SOTA methods fine-tuned on GLD because our method does not need any fine-tuning, and the methods fine-tuned on GLD, although have fairness issues, show the best performance on ROxford and RPairs. We compare our method with the well-known GeM [98], fine-tuned on GLD using contrastive loss (CL) and AP loss (AP) [104]. To demonstrate the generalizability of our TP, we pair it with the SOTA fine-tuned features of DELG [21] and the SOTA re-ranking method Hypergraph Propagation (HP) [6]. We notice that there are existing claimed better results than DELG and HP on ROxford and RParis, but combining DELG and HP is the best baseline we can successfully reproduce by the submission date of this paper.

### 5.4.2 Quantitative result

As depicted in Table 5.1, our method notably surpasses SP, GCRANSAC [13], ASMK, and VLAD across all settings. It significantly outperforms SP by 8.6% and 11.7% on the challenging ROxford under middle and hard settings, respectively. This underlines the feasibility and effectiveness of our novel adaptation of the topological model in RANSAC. Given the widespread usage of SP [91, 108, 110, 109] and its competitiveness with other instance recognition approaches in terms of accuracy [97, 21, 88], this improvement is particularly noteworthy. Our results represent the top non-fine-tuning performance on ROxford and RParis. While GCRANSAC excels at predicting homography or fundamental matrices [13], it does not surpass standard RANSAC in our recognition task. This may be due to the spatial model's incompatibility with recognition tasks, suggesting that enhancing spatial-based RANSAC may not improve landmark recognition performance. In an ablation study conducted to assess the impact of using SP to filter the hypothesis set, we found that incorporating

**Table 5.2:** Results (% mAP) of HP on the ROxford.

| Method | M | H |
|---|---|---|
| SP (hop1) | 80 | 61 |
| Our TP (hop1) | **81** | **63** |
| SP (hop2) | 85 | 69 |
| Our TP (hop2) | **86** | **71** |

SP surprisingly led to a decrease in final accuracy. Despite our Python-based method averaging 1.23s per image pair, slower than C-implemented SP at 0.53s, we anticipate considerable speed improvements with a C-based implementation of our method.

Our TP outperforms notable pre-trained models, including CLIP [99] and Superpoint [31], without requiring any training. Although CLIP delivers fair performance, it falls short of both SP and TP. These results underscore the competitive edge of the hypothesis-testing strategy for landmark recognition in highly noisy situations, even against large-scale pre-training methods. Despite Superpoint's superiority to SIFT in SP and its enhancement of inference performance, it remains less effective than our TP. These results attest to our TP's ability to circumvent the inherent issues of the spatial model for recognition.

Our novel TP method demonstrates competitive results against the renowned GeM [98], fine-tuned on the large cleaned Google landmarked dataset (GLD)[143], without any pre-training or fine-tuning. This comparison, while unconventional due to TP being a reranking method and GeM an initial ranking method, is meaningful. GeM, a state-of-the-art retrieval baseline, marked a milestone as the first metric learning method to surpass non-fine-tuned SP approaches [97], following a series of improvements from selective search[124] to R-MAC [130]. Since then, non-fine-tuning techniques have seen limited advancements. However, TP breaks this trend, achieving comparable results to GLD-trained GeM. This is not to say SIFT is enough for this task but to emphasize the potential of hypothesis testing when combined with a topological model for non-fine-tuning image retrieval.

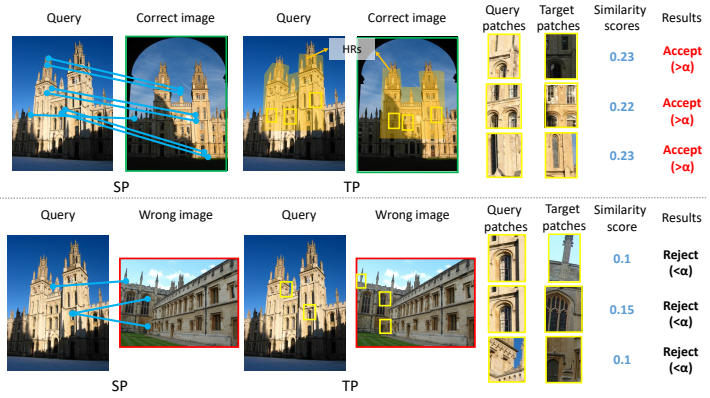



**Figure 5.3:** SP results, TP results, and some TP steps for verifying a correct image pair (upper) and a wrong image pair (down) using SIFT features. The threshold $\alpha$ is set as 0.2.

We acknowledge that advancements in fine-tuning features for ASMK or SP, such as SuperFeature and DELG [142, 21], have emerged recently. While it is possible to train features for our TP, we leave this for future work, as this work mainly focuses on the non-fine-tuning retrieval. However, to show our method's generalization capabilities, we still apply TP to fine-tuned DELG local features and Hypergraph Propagation (HP) [6]. HP is a fast reranking method that combines verification and diffusion on local features. It runs the heavy verification offline and thus achieves fast, low-memory, and high-accuracy reranking online. Table 5.2 demonstrates that our method outperforms SP on HP, irrespective of propagating among hop1 or hop2 neighbors; hop1 or hop2 is the truncation hyperparameter in diffusion [6]. This improvement is noteworthy, considering it does not necessitate any additional training. This is the SOTA retrieval performance on ROxford to the best of our knowledge.

We understand that readers may have concerns about the **practical value of a RANSAC-based method** in retrieval, given its slower processing time compared to using dot products with learned embedding vectors. However, it is important to note that if RANSAC can provide reliable labels without fine-tuning, these labels can be used directly to train the embedding space for any database, eliminating the need for a specially collected fine-tuning set. As we shown in Table 5.2, it can also directly improve HP accuracy. Because verification in HP is offline, the slower speed of RANSAC is acceptable. Furthermore, SP is employed not only in retrieval but also in other tasks on mobile


**Figure 5.4:** Hypergraph propagation results using traditional SP (upper) and our method (down). In each triplet, the first image and third image are different. Red lines show SP's wrongly connected local features between the first image and the third image. The orange and green regions show the verified HRs between each image pair. Our method correctly separates the local features of the first and the third images. HR is the union of many verified patches. We only show the outline of key points in HR for ease of observation.

devices where model size is crucial, such as loop-closure detection. Our TP, which shares the lightweight nature of SP without involving millions of parameters like models such as ResNet [49], offers enhanced robustness and accuracy. Most importantly, our method is highly explainable, which we will discuss in the following section.

### 5.4.3 Qualitative result

In addition to its accuracy, our method exhibits high explainability, a crucial aspect for real-life applications. The explainability of our method can be understood in three aspects:

1. We can observe how the HRs are constructed step by step, find a correspondence patch for each patch in HR, and check how each candidate patch pair is accepted or rejected by the fovea function $F$, as shown in Figs. 5.2 and 5.3. These characters offer a clear understanding of how our method works.

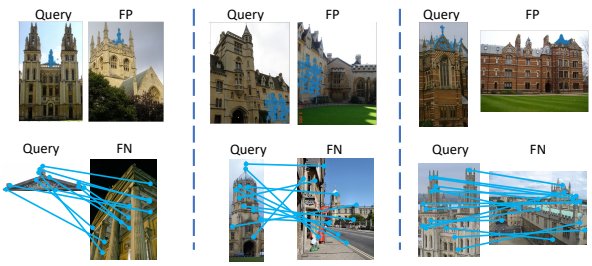



**Figure 5.5:** The false positive (FP) and false negative (FN) cases of our method. We draw the SIFT feature locations in the largest detected HRs for FP cases and the ratio test results for FN cases. Check more examples in the supplementary material. Check all the FP and FN cases from this link.

2. The TP detected HRs and matched patches are easy to comprehend, whether using SIFT (Fig. 5.3) or DELG (Figs. 5.1 and 5.4). As shown in the upper row of Fig. 5.3, although SP can also give a high inlier number for the correct image pair, its found matchings can be wrong. Our TP gives more reasonable results. This issue of high inlier counts with incorrect matchings is also observed in DELG, suggesting that fine-tuning can improve the similarity of local features for the same object, but cannot guarantee the accuracy of their locations. Additional examples are provided in the supplementary material. Fig. 5.4 compares the qualitative performance of SP and our TP on HP. The first query and third image in each triplet do not depict the same building. Due to SP's erroneous matching of some local features (indicated by red lines), a search engine using SP mistakenly identifies the third image as a positive match. In contrast, our method accurately matches the correct similar regions between two images and can correctly determine that the third image is unrelated to the query, thereby enhancing retrieval performance.

3. Our approach maintains explainability, even in cases of false positives and negatives, offering valuable insights into the effectiveness and potential issues at each stage. We have visualized all false negatives (FN) and positives (FP) in Fig. 5.5. For a more comprehensive review, we provide a link for readers to explore further. Interestingly,

without context, many FP cases are challenging for the authors to distinguish, signaling the effectiveness of our fovea function F in generating reasonable patch similarity results. Upon examining FN cases, we recognize areas for enhancement. The primary issues lie in incorrect ratio test results and significant size changes. While our method considers size changes, it struggles with handling extreme variations. An easy solution is to multi-rescale each image and traverse all possible matching scales. We refrained from this to avoid significantly reducing speed and to maintain a fair comparison with SP. However, future work could potentially apply this strategy or devise a better sampling method to manage extreme scale changes, thereby improving performance.

## 5.5 Discussion about the task value

This paper aims to advance explainable image retrieval in scenarios where a fine-tuning set is unavailable. This pursuit carries significant weight as acquiring a large fine-tuning set can often be impractical or costly, especially in open-world or private settings. Therefore, research in this domain has the potential to enhance the utility of AI systems across a variety of real-world contexts.

Our exploration into non-fine-tuned retrieval is further motivated by the observed gap between human-level performance and that of the state-of-the-art (SOTA) fine-tuned models [7, 123] on challenging datasets such as ROxford/ RParis [97] and PROxford/ PRParis [6]. Notably, these datasets provide invaluable user study results, as each positive image pair is verified to be identifiable by humans without the need for contextual visual information [97, 7]. Our recent study shows that annotators who are not familiar with European buildings can recognize, localize, and segment the target landmark of all the positive image pairs in ROxford and RParis. We propose the pixel-retrieval task and its first benchmarks PROxford and PRParis, to evaluate the pixel-level recognition [7]. Interestingly, humans achieve this remarkable recognition ability without extensive in-domain training, such as learning from the Google Landmarks Dataset (GLD) [143]. Yet, even the best fine-tuned models [21, 123] trained on the vast GLD with 4.1 million images have yet to reach human-level accuracy on ROxford/RParis and PROxford/PRParis. This discrepancy prompts us to question how humans excel in such recognition tasks without large-scale fine-tuning.

The process of recognition is a confluence of perception and cognition. The superior instance recognition of humans in the ROxford dataset might be more cognitively inclined, though concrete evidence remains elusive. Cognition entails reasoning based on existing knowledge for decision-making and the discovery of new knowledge [144, 38, 117]. When summarizing notable human instance recognition experiments [19, 50, 121], Greyson [2] defines perception as the direct link between an instance and its visual features, while cognition involves understanding the object's changes in viewpoint or state (*e.g.*, background and illumination) [144, 2, 36, 81]. Machine learning methods that align with this definition of cognition include ASMK [127] and SP [91], which further motivates us to explore improvements to SP without fine-tuning. Future research discerning the nuances between perception and cognition could potentially enhance non-fine-tuning methods for real-world applications.

## 5.6 Conclusion

Our study highlights the limitations of the SPatial verification (SP) method, particularly its reliance on a spatial model for recognition. As a potential alternative, we propose the use of topological common sense, which significantly outperforms SP and achieves unprecedented performance in non-fine-tuning instance recognition. This underscores the potential of hypothesis-testing inference.

The inspiration for our saccade and fovea function comes from the human process of comparing two newly seen

images. In this sense, our topological approach aligns more closely with the human visual system than the spatial approach. While humans don't compute transformation matrices like SP when comparing images, they intuitively understand the topological relations among observed objects—for instance, recognizing that a hand is connected to an arm. Notably, topological information is vital to the human visual system, as it is consistently preserved even when visual signals undergo significant deformation during transmission from the retina to the lateral geniculate nucleus and visual cortex [140, 52].

Importantly, our method doesn't compete with existing fine-tuning methods but complements them. It can directly integrate learned features, and future work could even fine-tune features specifically for this topological approach.

While our results are promising, they also pave the way for future exploration. Subsequent work could investigate the incorporation of more complex common sense concepts, both explicit, like our topological model, and implicit, derived from pre-training. Enhancements in these areas could potentially boost the robustness and adaptability of our method. Further study into the convergence properties of our method and refinement of the hypothesis sampling strategy are also warranted. Through addressing these areas, we aspire to advance object recognition and highlight the potential of hypothesis-testing approaches in artificial intelligence.

# Bibliography


[1] *Cognitive science*. Springer.

[2] G. Abid. Recognition and the perception–cognition divide. *Mind & Language*, 37(5):770–789, 2022.

[3] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[4] S. Agarwal, Y. Furukawa, N. Snavely, I. Simon, B. Curless, S. M. Seitz, and R. Szeliski. Building rome in a day. *Communications of the ACM*, 54(10):105–112, 2011.

[5] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

[6] G. An, Y. Huo, and S.-E. Yoon. Hypergraph propagation and community selection for objects retrieval. *Advances in Neural Information Processing Systems*, 34:3596–3608, 2021.

[7] G. An, W. J. Kim, S. Yang, R. Li, Y. Huo, and S.-E. Yoon. Towards content-based pixel retrieval in revisited oxford and paris. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20507–20518, 2023.

[8] G. An, J. Seon, I. An, Y. Huo, and S.-E. Yoon. Topological ransac for instance verification and retrieval without fine-tuning. *arXiv preprint arXiv:2310.06486*, 2023.

[9] Anthropic. Introducing the next generation of claude. 2024.

[10] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic. Netvlad: Cnn architecture for weakly supervised place recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5297–5307, 2016.

[11] R. Arandjelović and A. Zisserman. Three things everyone should know to improve object retrieval. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2911–2918. IEEE, 2012.

[12] D. Baldauf and R. Desimone. Neural mechanisms of object-based attention. *Science*, 344(6182):424–427, 2014.

[13] D. Barath and J. Matas. Graph-cut ransac. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6733–6741, 2018.

[14] H. Bay, T. Tuytelaars, and L. V. Gool. Surf: Speeded up robust features. In *European conference on computer vision*, pages 404–417. Springer, 2006.

[15] Y. Bengio, I. Goodfellow, and A. Courville. *Deep learning*, volume 1. MIT press Cambridge, MA, USA, 2017.

[16] G. Berton, G. Trivigno, B. Caputo, and C. Masone. Eigenplaces: Training viewpoint robust models for visual place recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11080–11090, 2023.

[17] L. Beyer, O. J. Hénaff, A. Kolesnikov, X. Zhai, and A. v. d. Oord. Are we done with imagenet? *arXiv preprint arXiv:2006.07159*, 2020.

[18] D. Bolya, C. Zhou, F. Xiao, and Y. J. Lee. Yolact: Real-time instance segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9157–9166, 2019.

[19] T. F. Brady, T. Konkle, G. A. Alvarez, and A. Oliva. Visual long-term memory has a massive storage capacity for object details. *Proceedings of the National Academy of Sciences*, 105(38):14325–14329, 2008.

[20] E. Breznik, E. Wetzer, J. Lindblad, and N. Sladoje. Cross-modality sub-image retrieval using contrastive multimodal image representations. *arXiv preprint arXiv:2201.03597*, 2022.

[21] B. Cao, A. Araujo, and J. Sim. Unifying deep local and global features for image search. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX 16*, pages 726–743. Springer, 2020.

[22] C. Chang, G. Yu, C. Liu, and M. Volkovs. Explore-exploit graph traversal for image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9423–9431, 2019.

[23] S. Choi, T. Kim, and W. Yu. Performance evaluation of ransac family. *Journal of Computer Vision*, 24(3):271–300, 1997.

[24] O. Chum et al. Large-scale discovery of spatially related images. *IEEE transactions on pattern analysis and machine intelligence*, 32(2):371–377, 2009.

[25] O. Chum, A. Mikulik, M. Perdoch, and J. Matas. Total recall ii: Query expansion revisited. In *CVPR 2011*, pages 889–896. IEEE, 2011.

[26] O. Chum, J. Philbin, J. Sivic, M. Isard, and A. Zisserman. Total recall: Automatic query expansion with a generative feature model for object retrieval. In *2007 IEEE 11th International Conference on Computer Vision*, pages 1–8. IEEE, 2007.

[27] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016.

[28] W. Dai, J. Li, D. Li, A. M. H. Tiong, J. Zhao, W. Wang, B. Li, P. N. Fung, and S. Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. In *NeurIPS*, 2023.

[29] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[30] K. G. Derpanis. Overview of the ransac algorithm. *Image Rochester NY*, 4(1):2–3, 2010.

[31] D. DeTone, T. Malisiewicz, and A. Rabinovich. Superpoint: Self-supervised interest point detection and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 224–236, 2018.

[32] W. Dong, C. Moses, and K. Li. Efficient k-nearest neighbor graph construction for generic similarity measures. In *Proceedings of the 20th international conference on World wide web*, pages 577–586, 2011.

[33] M. Donoser and H. Bischof. Diffusion processes for retrieval revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1320–1327, 2013.

[34] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010.

[35] Q. Fan, W. Pei, Y.-W. Tai, and C.-K. Tang. Self-support few-shot semantic segmentation. 2022.

[36] X. Fan, L. Wang, H. Shao, D. Kersten, and S. He. Temporally flexible feedback signal to foveal cortex for peripheral object recognition. *Proceedings of the National Academy of Sciences*, 113(41):11627–11632, 2016.

[37] M. A. Fischler and R. C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981.

[38] M. S. Gazzaniga. *The cognitive neurosciences*. MIT press, 2009.

[39] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3354–3361. IEEE, 2012.

[40] G. Gigerenzer. Gut feelings: The intelligence of the unconscious. *New York: Viking*, 2007.

[41] G. Gigerenzer and W. Gaissmaier. Heuristic decision making. *Annual review of psychology*, 62(1):451–482, 2011.

[42] M. A. Goodale and A. D. Milner. Separate visual pathways for perception and action. *Trends in neurosciences*, 15(1):20–25, 1992.

[43] I. Goodfellow, Y. Bengio, and A. Courville. *Deep learning*. MIT press, 2016.

[44] I. Goodfellow, D. Warde-Farley, M. Mirza, A. Courville, and Y. Bengio. Maxout networks. In *International conference on machine learning*, pages 1319–1327. PMLR, 2013.

[45] A. Gordo, J. Almazan, J. Revaud, and D. Larlus. End-to-end learning of deep visual representations for image retrieval. *International Journal of Computer Vision*, 124(2):237–254, 2017.

[46] A. Gordo, F. Radenovic, and T. Berg. Attention-based query expansion learning. In *European Conference on Computer Vision*, pages 172–188. Springer, 2020.

[47] C. G. Gross. Genealogy of the "grandmother cell". *The Neuroscientist*, 8(5):512–518, 2002.

[48] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[49] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[50] A. Hollingworth and J. M. Henderson. Accurate visual memory for previously attended objects in natural scenes. *Journal of Experimental Psychology: Human Perception and Performance*, 28(1):113, 2002.

[51] D. H. Hubel and T. N. Wiesel. Receptive fields of single neurones in the cat's striate cortex. *The Journal of physiology*, 148(3):574, 1959.

[52] A. D. Huberman, M. B. Feller, and B. Chapman. Mechanisms underlying development of visual maps and receptive fields. *Annual review of neuroscience*, 31:479, 2008.

[53] A. Iscen, G. Tolias, Y. Avrithis, T. Furon, and O. Chum. Efficient diffusion on region manifolds: Recovering small objects with compact cnn representations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2077–2086, 2017.

[54] H. Jegou, M. Douze, and C. Schmid. Hamming embedding and weak geometric consistency for large scale image search. In *European conference on computer vision*, pages 304–317. Springer, 2008.

[55] H. Jégou, M. Douze, C. Schmid, and P. Pérez. Aggregating local descriptors into a compact image representation. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 3304–3311. IEEE, 2010.

[56] Y. Jin, D. Mishkin, A. Mishchuk, J. Matas, P. Fua, K. M. Yi, and E. Trulls. Image Matching across Wide Baselines: From Paper to Practice. *International Journal of Computer Vision*, 2020.

[57] R. B. João and R. M. Filgueiras. Frontal lobe: functional neuroanatomy of its circuitry and related disconnection syndromes. *Prefrontal Cortex*, 41, 2018.

[58] O. F. Kar, A. Tonioni, P. Poklukar, A. Kulshrestha, A. Zamir, and F. Tombari. Brave: Broadening the visual encoding of vision-language models. *arXiv preprint arXiv:2404.07204*, 2024.

[59] D. Kim, A. Angelova, and W. Kuo. Region-centric image-language pretraining for open-vocabulary detection. In *European Conference on Computer Vision*, pages 162–179. Springer, 2025.

[60] K. Koffka. *Principles of Gestalt psychology*. Routledge, 2013.

[61] C. H. Lampert. Detecting objects in large image collections and videos by efficient subimage retrieval. In *2009 IEEE 12th International Conference on Computer Vision*, pages 987–994. IEEE, 2009.

[62] Y. LeCun, Y. Bengio, et al. The handbook of brain theory and neural networks, 1998.

[63] S. Lee, H. Seong, S. Lee, and E. Kim. Correlation verification for image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5374–5384, 2022.

[64] D. Li, X. Shi, Q. Long, S. Liu, W. Yang, F. Wang, Q. Wei, and F. Qiao. Dxslam: A robust and efficient visual slam system with deep features. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4958–4965. IEEE, 2020.

[65] J. Li, D. Li, S. Savarese, and S. Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023.

[66] X. Li, C. Wu, C. Zach, S. Lazebnik, and J.-M. Frahm. Modeling and recognition of landmark image collections using iconic scene graphs. In *Computer Vision–ECCV 2008: 10th European Conference on Computer Vision, Marseille, France, October 12-18, 2008, Proceedings, Part I 10*, pages 427–440. Springer, 2008.

[67] Z. Li and N. Snavely. Megadepth: Learning single-view depth prediction from internet photos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2041–2050, 2018.

[68] J. Liang, R. He, and T. Tan. A comprehensive survey on test-time adaptation under distribution shifts. *International Journal of Computer Vision*, pages 1–34, 2024.

[69] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[70] Z. Lin and J. Brandt. A local bag-of-features model for large-scale object retrieval. In *European conference on Computer vision*, pages 294–308. Springer, 2010.

[71] C. Liu, G. Yu, M. Volkovs, C. Chang, H. Rai, J. Ma, and S. K. Gorti. Guided similarity separation for image retrieval. 2019.

[72] H. Liu, C. Li, Y. Li, and Y. J. Lee. Improved baselines with visual instruction tuning, 2023.

[73] H. Liu, C. Li, Y. Li, B. Li, Y. Zhang, S. Shen, and Y. J. Lee. Llava-next: Improved reasoning, ocr, and world knowledge, 2024.

[74] H. Liu, C. Li, Y. Li, B. Li, Y. Zhang, S. Shen, and Y. J. Lee. Llava-next: Improved reasoning, ocr, and world knowledge, January 2024.

[75] H. Liu, C. Li, Q. Wu, and Y. J. Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024.

[76] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia. Path aggregation network for instance segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8759–8768, 2018.

[77] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016.

[78] D. G. Lowe. Object recognition from local scale-invariant features. In *Proceedings of the seventh IEEE international conference on computer vision*, volume 2, pages 1150–1157. Ieee, 1999.

[79] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.

[80] N. Mehta, A. Raja'S, and V. Chaudhary. Content based sub-image retrieval system for high resolution pathology images using salient interest points. In *2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 3719–3722. IEEE, 2009.

[81] E. K. Miller and J. D. Cohen. An integrative theory of prefrontal cortex function. *Annual review of neuroscience*, 24(1):167–202, 2001.

[82] J. Min, D. Kang, and M. Cho. Hypercorrelation squeeze for few-shot segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6941–6952, 2021.

[83] M. Minderer, A. Gritsenko, A. Stone, M. Neumann, D. Weissenborn, A. Dosovitskiy, A. Mahendran, A. Arnab, M. Dehghani, Z. Shen, et al. Simple open-vocabulary object detection with vision transformers. *arXiv preprint arXiv:2205.06230*, 2022.

[84] R. G. Morrison, L. A. Doumas, and L. E. Richland. The development of analogical reasoning in children: A computational account. In *Proceedings of the twenty-ninth annual conference of the cognitive science society*. Erlbaum Mahwah, NJ, 2006.

[85] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardos. Orb-slam: a versatile and accurate monocular slam system. *IEEE transactions on robotics*, 31(5):1147–1163, 2015.

[86] T. Ng, V. Balntas, Y. Tian, and K. Mikolajczyk. Solar: second-order loss and attention for image retrieval. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXV 16*, pages 253–270. Springer, 2020.

[87] D. Nister and H. Stewenius. Scalable recognition with a vocabulary tree. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 2161–2168. Ieee, 2006.

[88] H. Noh, A. Araujo, J. Sim, T. Weyand, and B. Han. Large-scale image retrieval with attentive deep local features. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[89] A. Osokin, D. Sumin, and V. Lomakin. OS2D: One-stage one-shot object detection by matching anchor features. In *proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

[90] J. Ouyang, H. Wu, M. Wang, W. Zhou, and H. Li. Contextual similarity aggregation with self-attention for visual re-ranking. *Advances in Neural Information Processing Systems*, 34:3135–3148, 2021.

[91] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Object retrieval with large vocabularies and fast spatial matching. In *2007 IEEE conference on computer vision and pattern recognition*, pages 1–8. IEEE, 2007.

[92] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Lost in quantization: Improving particular object retrieval in large scale image databases. In *2008 IEEE conference on computer vision and pattern recognition*, pages 1–8. IEEE, 2008.

[93] J. Philbin and A. Zisserman. Object mining using a matching graph on very large image collections. In *2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing*, pages 738–745. IEEE, 2008.

[94] J. Ponce, T. L. Berg, M. Everingham, D. A. Forsyth, M. Hebert, S. Lazebnik, M. Marszalek, C. Schmid, B. C. Russell, A. Torralba, et al. Dataset issues in object recognition. *Toward category-level object recognition*, pages 29–48, 2006.

[95] D. Qin, S. Gammeter, L. Bossard, T. Quack, and L. Van Gool. Hello neighbor: Accurate object retrieval with k-reciprocal nearest neighbors. In *CVPR 2011*, pages 777–784. IEEE, 2011.

[96] R. Q. Quiroga, L. Reddy, G. Kreiman, C. Koch, and I. Fried. Invariant visual representation by single neurons in the human brain. *Nature*, 435(7045):1102–1107, 2005.

[97] F. Radenović, A. Iscen, G. Tolias, Y. Avrithis, and O. Chum. Revisiting oxford and paris: Large-scale image retrieval benchmarking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5706–5715, 2018.

[98] F. Radenović, G. Tolias, and O. Chum. Fine-tuning cnn image retrieval with no human annotation. *IEEE transactions on pattern analysis and machine intelligence*, 41(7):1655–1668, 2018.

[99] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[100] R. Raguram, J. Tighe, and J.-M. Frahm. Improved geometric verification for large scale landmark image collections. In *BMVC*, pages 1–11, 2012.

[101] M. Ranzinger, G. Heinrich, J. Kautz, and P. Molchanov. Am-radio: Agglomerative vision foundation model reduce all domains into one. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12490–12500, 2024.

[102] A. S. Razavian, J. Sullivan, S. Carlsson, and A. Maki. Visual instance retrieval with deep convolutional networks. *ITE Transactions on Media Technology and Applications*, 4(3):251–258, 2016.

[103] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *arXiv preprint arXiv:1506.01497*, 2015.

[104] J. Revaud, J. Almazán, R. S. Rezende, and C. R. d. Souza. Learning with average precision: Training image retrieval with a listwise loss. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5107–5116, 2019.

[105] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

[106] G. Sanchez, H. Fan, A. Spangher, E. Levi, P. S. Ammanamanchi, and S. Biderman. Stay on topic with classifier-free guidance. *arXiv preprint arXiv:2306.17806*, 2023.

[107] G. E. Schneider. Two visual systems: Brain mechanisms for localization and discrimination are dissociated by tectal and cortical lesions. *Science*, 163(3870):895–902, 1969.

[108] J. L. Schönberger and J.-M. Frahm. Structure-from-motion revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[109] J. L. Schönberger, T. Price, T. Sattler, J.-M. Frahm, and M. Pollefeys. A vote-and-verify strategy for fast spatial verification in image retrieval. In *Asian Conference on Computer Vision (ACCV)*, 2016.

[110] J. L. Schönberger, E. Zheng, M. Pollefeys, and J.-M. Frahm. Pixelwise view selection for unstructured multi-view stereo. In *European Conference on Computer Vision (ECCV)*, 2016.

[111] T. Schops, J. L. Schonberger, S. Galliani, T. Sattler, K. Schindler, M. Pollefeys, and A. Geiger. A multi-view stereo benchmark with high-resolution images and multi-camera videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3260–3269, 2017.

[112] X. Shen, F. Darmon, A. A. Efros, and M. Aubry. Ransac-flow: generic two-stage image alignment. In *European Conference on Computer Vision*, pages 618–637. Springer, 2020.

[113] X. Shen, Z. Lin, J. Brandt, S. Avidan, and Y. Wu. Object retrieval and localization with spatially-constrained similarity measure and k-nn re-ranking. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*.

[114] X. Shen, Z. Lin, J. Brandt, and Y. Wu. Spatially-constrained similarity measurefor large-scale object retrieval. *IEEE transactions on pattern analysis and machine intelligence*, 36(6):1229–1241, 2013.

[115] X. Shen, X. Tao, C. Zhou, H. Gao, and J. Jia. Regional foremost matching for internet scene images. *ACM Transactions on Graphics (TOG)*, 35(6):1–12, 2016.

[116] Y. Shi and H. Li. Beyond cross-view image retrieval: Highly accurate vehicle localization using satellite image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17010–17020, 2022.

[117] Z. Shi. *Intelligence science: Leading the age of intelligence*. Elsevier, 2021.

[118] N. Shinn, F. Cassano, A. Gopinath, K. Narasimhan, and S. Yao. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 36, 2024.

[119] J. Shotton, B. Glocker, C. Zach, S. Izadi, A. Criminisi, and A. Fitzgibbon. Scene coordinate regression forests for camera relocalization in rgb-d images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2930–2937, 2013.

[120] Y. Shrager, J. J. Gold, R. O. Hopkins, and L. R. Squire. Intact visual perception in memory-impaired patients with medial temporal lobe lesions. *Journal of Neuroscience*, 26(8):2235–2240, 2006.

[121] L. Standing, J. Conezio, and R. N. Haber. Perception and memory for pictures: Single-trial learning of 2500 visual stimuli. *Psychonomic science*, 19(2):73–74, 1970.

[122] R. Szeliski. *Computer vision: algorithms and applications*. Springer Science & Business Media, 2010.

[123] F. Tan, J. Yuan, and V. Ordonez. Instance-level image retrieval using reranking transformers. In *proceedings of the IEEE/CVF international conference on computer vision*, pages 12105–12115, 2021.

[124] R. Tao, E. Gavves, C. G. Snoek, and A. W. Smeulders. Locality in generic instance search from one example. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2091–2098, 2014.

[125] G. Team, R. Anil, S. Borgeaud, Y. Wu, J.-B. Alayrac, J. Yu, R. Soricut, J. Schalkwyk, A. M. Dai, A. Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

[126] M. Teichmann, A. Araujo, M. Zhu, and J. Sim. Detect-to-retrieve: Efficient regional aggregation for image search. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5109–5118, 2019.

[127] G. Tolias, Y. Avrithis, and H. Jégou. To aggregate or not to aggregate: Selective match kernels for image search. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1401–1408, 2013.

[128] G. Tolias, Y. Avrithis, and H. Jégou. Image search with selective match kernels: aggregation across single and multiple images. *International Journal of Computer Vision*, 116(3):247–261, 2016.

[129] G. Tolias, T. Jenicek, and O. Chum. Learning and aggregating deep local descriptors for instance-level recognition. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pages 460–477. Springer, 2020.

[130] G. Tolias, R. Sicre, and H. Jégou. Particular object retrieval with integral max-pooling of cnn activations. *arXiv preprint arXiv:1511.05879*, 2015.

[131] S. Tong, E. Brown, P. Wu, S. Woo, M. Middepogu, S. C. Akula, J. Yang, S. Yang, A. Iyer, X. Pan, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *arXiv preprint arXiv:2406.16860*, 2024.

[132] S. Tong, Z. Liu, Y. Zhai, Y. Ma, Y. LeCun, and S. Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9568–9578, 2024.

[133] P. Truong, M. Danelljan, L. V. Gool, and R. Timofte. Gocor: Bringing globally optimized correspondence volumes into your neural network. *Advances in Neural Information Processing Systems*, 33:14278–14290, 2020.

[134] P. Truong, M. Danelljan, and R. Timofte. Glu-net: Global-local universal network for dense flow and correspondences. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6258–6268, 2020.

[135] P. Truong, M. Danelljan, R. Timofte, and L. Van Gool. Pdc-net+: Enhanced probabilistic dense correspondence network. *arXiv preprint arXiv:2109.13912*, 2021.

[136] P. Truong, M. Danelljan, F. Yu, and L. Van Gool. Warp consistency for unsupervised learning of dense correspondences. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10346–10356, 2021.

[137] J. R. Uijlings, K. E. Van De Sande, T. Gevers, and A. W. Smeulders. Selective search for object recognition. *International journal of computer vision*, 104(2):154–171, 2013.

[138] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017.

[139] J. A. Waltz, B. J. Knowlton, K. J. Holyoak, K. B. Boone, F. S. Mishkin, M. de Menezes Santos, C. R. Thomas, and B. L. Miller. A system for relational reasoning in human prefrontal cortex. *Psychological science*, 10(2):119–125, 1999.

[140] B. A. Wandell, S. O. Dumoulin, and A. A. Brewer. Visual field maps in human cortex. *Neuron*, 56(2):366–383, 2007.

[141] T. Wei, D. Chen, W. Zhou, J. Liao, H. Zhao, W. Zhang, and N. Yu. Improved image matting via real-time user clicks and uncertainty estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15374–15383, 2021.

[142] P. Weinzaepfel, T. Lucas, D. Larlus, and Y. Kalantidis. Learning super-features for image retrieval. *arXiv preprint arXiv:2201.13182*, 2022.

[143] T. Weyand, A. Araujo, B. Cao, and J. Sim. Google landmarks dataset v2-a large-scale benchmark for instance-level recognition and retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2575–2584, 2020.

[144] J. M. Wolfe, K. R. Cave, and S. L. Franzel. Guided search: an alternative to the feature integration model for visual search. *Journal of Experimental Psychology: Human perception and performance*, 15(3):419, 1989.

[145] H. Wu, M. Wang, W. Zhou, Y. Hu, and H. Li. Learning token-based representation for image retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 2703–2711, 2022.

[146] H. Wu, M. Wang, W. Zhou, and H. Li. Learning deep local features with multiple dynamic attentions for large-scale image retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11416–11425, 2021.

[147] W. Wu, H. Yao, M. Zhang, Y. Song, W. Ouyang, and J. Wang. Gpt4vis: what can gpt-4 do for zero-shot visual recognition? *arXiv preprint arXiv:2311.15732*, 2023.

[148] N. Xu, B. Price, S. Cohen, and T. Huang. Deep image matting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2970–2979, 2017.

[149] L. Yang, X. Qi, F. Xing, T. Kurc, J. Saltz, and D. J. Foran. Parallel content-based sub-image retrieval using hierarchical searching. *Bioinformatics*, 30(7):996–1002, 2014.

[150] L. Yang, W. Zhuo, L. Qi, Y. Shi, and Y. Gao. Mining latent classes for few-shot segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8721–8730, 2021.

[151] M. Yang, D. He, M. Fan, B. Shi, X. Xue, F. Li, E. Ding, and J. Huang. Dolg: Single-stage image retrieval with deep orthogonal fusion of local and global features. In *Proceedings of the IEEE/CVF International conference on Computer Vision*, pages 11772–11781, 2021.

[152] X. Zhai, X. Wang, B. Mustafa, A. Steiner, D. Keysers, A. Kolesnikov, and L. Beyer. Lit: Zero-shot transfer with locked-image text tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18123–18133, 2022.

[153] Y. Zhang, L. Gong, L. Fan, P. Ren, Q. Huang, H. Bao, and W. Xu. A late fusion cnn for digital matting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7469–7478, 2019.

[154] Y. Zhang, A. Unell, X. Wang, D. Ghosh, Y. Su, L. Schmidt, and S. Yeung-Levy. Why are visually-grounded language models bad at image classification? *arXiv preprint arXiv:2405.18415*, 2024.

[155] Z. Zhang, L. Wang, Y. Wang, L. Zhou, J. Zhang, and F. Chen. Dataset-driven unsupervised object discovery for region-based instance image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[156] Z. Zhang, L. Wang, L. Zhou, and P. Koniusz. Learning spatial-context-aware global visual feature representation for instance image retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11250–11259, 2023.

[157] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 633–641, 2017.

[158] D. Zhou, O. Bousquet, T. N. Lal, J. Weston, and B. Schölkopf. Learning with local and global consistency. In *Advances in neural information processing systems*, pages 321–328, 2004.

[159] D. Zhou, J. Weston, A. Gretton, O. Bousquet, and B. Schölkopf. Ranking on data manifolds. *Advances in neural information processing systems*, 16:169–176, 2004.

[160] S. Zhu, T. Yang, and C. Chen. Vigor: Cross-view image geo-localization beyond one-to-one retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3640–3649, 2021.

[161] Y. Zhu, X. Gao, B. Ke, R. Qiao, and X. Sun. Coarse-to-fine: Learning compact discriminative representation for single-stage image retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11260–11269, 2023.

# Guoyuan An

anguoyuan111@gmail.com
https://anguoyuan.github.io/


## Education

2020.09 – now **Ph.D. student in Computer Science, KAIST**
Advisor: *Sung-Eui Yoon*. Topic: Computer vision - Image search

2018.09 – 2020.09 **M.Sc. in Computer Science, KAIST**
Advisor: *Myoung-Ho Kim*. Topic: Database - Recommender system

2014.09 – 2018.09 **B.S. in Industrial Engineering, Inha University**
Minor in Software Engineering. Ranking: 1/108

## Main Research Publications

1 Inkyu An, **Guoyuan An**, Taeyound Kim, and Sung-Eui Yoon. "Microphone Pair Training for Robust Sound Source Localization with Diverse Array Configurations". In: *IEEE Robotics and Automation Letters (RA-L)*. 2023.

2 **Guoyuan An**, Ju-hyeong Seon, Inkyu An, Yuchi Huo, and Sung-eui Yoon. "Topological RANSAC for Instance Verification and Retrieval without Fine-tuning". In: *Thirty-Seventh Conference on Neural Information Processing Systems (NeurIPS)*. 2023.

3 **Guoyuan An**, Kim WooJae, Yang Salyne, Yuchi Huo, Li Rong, and Sung-eui Yoon. "Towards Content-based Segment Retrieval in Revisited Oxford and Paris". In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 2023.

4 **Guoyuan An**, Yuchi Huo, and Sung-eui Yoon. "Hypergraph Propagation and Community Selection for Objects Retrieval". In: *Thirty-Fifth Conference on Neural Information Processing Systems (NeurIPS)*. 2021.

5 **Guoyuan An**, Sung Eui Yoon, Jae Yoon Kim, Lin Wang, and Myoung Ho Kim. "GraphShop: Graph-based Approach for Shop-type Recommendation". In: *Proceedings of the 2021 SIAM International Conference on Data Mining (SDM)*. SIAM. 2021, pp. 55–63.

## Skills

Languages Chinese (native speaker), English, Korean (TOPIK 6)

Coding Currently fluent in Python. Used to be fluent in C, Java

## Miscellaneous Experience

### Awards and Achievements

2019 **First Place** (award $ 7500), INFORMS Student Competition, INFORMS, USA.

2015,16,17,18 **Outstanding Engineering Student Prize**, Inha University, South Korea.

### Activities

2014,15,16 As a member in the benevolent house repairing club.

2019,20 As the representative of the **KAIST Chinese Community**.