Towards Content-based Pixel Retrieval in Revisited Oxford and Paris

Guoyuan An[1], Woo Jae Kim[1], Saelyne Yang[1], Rong Li[3], Yuchi Huo[2,3], and Sung-Eui Yoon[1]

[1]School of Computing, KAIST
[2] State Key Lab of CAD&CG, Zhejiang University
[3]Zhejiang Lab

Abstract

This paper introduces the first two landmark pixel retrieval benchmarks. Pixel retrieval is segmented instance retrieval. Like semantic segmentation extends classification to the pixel level, pixel retrieval is an extension of image retrieval and offers information about which pixels are related to the query object. In addition to retrieving images for the given query, it helps users quickly identify the query object in true positive images and exclude false positive images by denoting the correlated pixels. Our user study results show pixel-level annotation can significantly improve the user experience. Compared with semantic and instance segmentation, pixel retrieval requires a fine-grained recognition capability for variable-granularity targets. To this end, we propose pixel retrieval benchmarks named PROxford and PRParis, which are based on the widely used image retrieval datasets, ROxford and RParis. Three professional annotators label 5,942 images with two rounds of double-checking and refinement. Furthermore, we conduct extensive experiments and analysis on the SOTA methods in image search, image matching, detection, segmentation, and dense matching using our pixel retrieval benchmarks. Results show that the pixel retrieval task is challenging to these approaches and distinctive from existing problems, suggesting that further research can advance the content-based pixel-retrieval and thus user search experience. The datasets can be downloaded from this link.

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 be 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 1.

Figure 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. 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.

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 [11, 62, 21] 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 [11] and instance segmentation [21, 62]. 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 [62, 21, 4].

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 [62, 7, 19, 8, 10], 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) [30, 31, 33]. 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 [33].

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 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).

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 [6, 28] give a high inlier number to some true query-index pairs but match the wrong regions. We find the dense and pixel-level approaches [25, 52] helpful for the pixel retrieval task. Most importantly, our results show that pixel retrieval is difficult and further research is needed for advancing 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. Content-based pixel retrieval

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) [30, 31, 33]. Oxford [30] and Paris [31] 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*. [33] 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 [2], VLAD [13], and ASMK [48], to the recent deep learning based methods like R-MAC [48], GeM [34], and DELF [28].

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 [14] and dense matching Megadepth [18], the positive image pairs in ROxford and RParis have severe viewpoint changes, occlusions, and illumination changes. The new queries added by Radenovic *et al*. [33] 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 [27] and Holiday [12], 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) [55] 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 [32], 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 [33]. 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 [33]. 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 [37], the SOTA identification ability about the first-seen objects in ROxford and RParis is still far from human-level [17, 1, 6].

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 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 [33, 16, 20], medical image analysis [24, 5, 57], geographical information systems [61, 63, 42], and so on. We discuss the related applications in Section 3. Distinguishing and segmenting the first-seen objects is also one basic function of human vision system [43]; 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 [16, 20, 40]. 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.

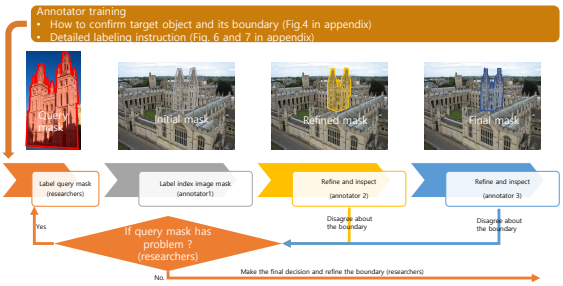

Figure 2. Labeling process (please zoom in for details).

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

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 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 [33].

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 [8] and COCO [19] 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) annotate; 2) refine + inspect; 3) refine + insp, as shown in Figure 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.

2.4. 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 [36]. Image retrieval methods commonly use mAP as the metric [33]. 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 [8] 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 5. We follow the medium and hard protocols in ROxford and RParis [33] 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.

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 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 1 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-

[1] prolific.co

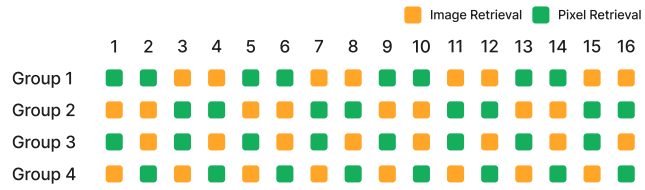

Figure 3. Design of the study on web search user experience. Image retrieval refers to a setting where no annotations are provided, whereas Pixel retrieval refers to a setting where pixel-level annotations are provided. 40 participants were divided into four groups and we counterbalanced the type of questions across the groups. Numbers 1 to 16 indicate the 16 questions.

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).

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 [57, 5, 24, 63].

Pixel retrieval can also help image matting [54, 56, 60]. Current image matting techniques rely on the user's click to confirm the target matting region [54, 56, 60]. 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 [41].

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 5.

4.1. Localization in retrieval

Some pioneering works [16, 20, 40] 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 [40, 23, 2, 28, 6] is one of the most popular reranking approaches in image retrieval. It is also known as image matching [14]; SP and stereo task in Image Matching Challenge (IMC) [14] 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 [23], SUFT [3], and rootSIFT [2]. Recently, DELF [28] and DELG [6] local features, which are learned from the large landmarks training set [55], 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 [16, 20, 35, 36, 53, 44, 45]. Faster-RCNN [36] and SSD detector [22] fine-tuned on a huge manually boxed landmark dataset [45] achieve the SOTA detect-related retrieval result [45]. Detect-to-retrieve (D2R) [45] 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) [13] and the Aggregated Selective Match Kernel (ASMK) [46] 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.

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) [26] is a vision transformer model trained on the large-scale 3.6 billion images in LiT dataset [59]. 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) [25] is one of the most famous few-shot segmentation methods. It finds multi-level feature correlations for a new class. The Mining model (Mining) [58] 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) [9] 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 [15]: 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.

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 [50] and RANSAC-flow [39] are popular among many famous dense matching methods. Recently, Truong *et al*. have shown that the warp consistency objective (WarpC) [52] and the GOCor module [49] can further improve the performance and achieve the new SOTA. Another popular method is PDC-Net [51]. 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 1.

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 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 [18] and ETH3D [38], 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.

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 1. We choose one to two representative methods for each field and show their qualitative results in Figure 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) [1]. We show their final mAP@50:5:95 in Table 2.

Although SP achieves impressive image-level retrieval results [6, 28], 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. For example, in the first easy case in Figure 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 [6]. 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 [6] 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 per-

Table 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 in each field; **red** number indicates the best performance throughout all fields.

Method	Medium				Hard			
	PROxf		PRPar		PROxf		PRPar	
	D	S	D	S	D	S	D	S
Localization methods in retrieval								
SIFT+SP [30]	10.5	3.9	14.0	5.1	7.1	2.4	12.4	4.3
DELF+SP [28]	14.5	**5.5**	21.3	**7.5**	9.4	**4.1**	16.7	**5.5**
DELG+SP [6]	13.8	5.2	18.6	7.2	8.9	2.9	13.6	4.9
D2R [45]+Resnet-50-Faster-RCNN+Mean	20.2	-	29.6	-	16.7	-	27.4	-
D2R [45]+Resnet-50-Faster-RCNN+VLAD [13]	25.8	-	37.5	-	**21.6**	-	35.5	-
D2R [45]+Resnet-50-Faster-RCNN+ASMK [47]	**26.3**	-	**38.5**	-	**21.6**	-	**35.6**	-
D2R [45]+Mobilenet-V2-SSD+Mean	19.7	-	25.9	-	20.1	-	27.9	-
D2R [45]+Mobilenet-V2-SSD+VLAD [13]	23.1	-	33.	-	20.9	-	33.6	-
D2R [45]+Mobilenet-V2-SSD+ASMK [47]	22.4	-	34.0	-	20.8	-	33.1	-
One-shot detection and segmentation methods								
OWL-VIT (LiT) [26]	11.4	-	18.0	-	6.3	-	15.0	-
OS2D-v2-trained [29]	10.5	-	13.7	-	11.7	-	14.3	-
OS2D-v1 [29]	7.0	-	8.5	-	8.7	-	9.2	-
OS2D-v2-init [29]	13.6	-	15.4	-	14.0	-	15.1	-
SSP (COCO) + ResNet50 [9]	19.2	34.5	31.1	48.7	15.1	25.3	29.8	**41.7**
SSP (VOC) + ResNet50 [9]	19.7	34.3	31.4	**48.8**	16.1	**26.1**	30.3	40.4
HSNet (COCO) + ResNet50 [25]	23.4	32.8	37.4	41.9	21.0	25.7	34.7	36.5
HSNet (VOC) + ResNet50 [25]	21.0	29.8	31.4	39.7	17.1	23.2	29.7	34.9
HSNet (FSS) + ResNet50 [25]	**30.5**	**35.7**	**39.4**	40.2	**22.7**	25.1	**34.7**	32.8
Mining (VOC) + ResNet50 [58]	18.3	30.5	29.6	42.7	15.1	21.4	28.1	34.3
Mining (VOC) + ResNet101 [58]	18.1	28.6	29.5	40.0	14.2	20.4	28.2	34.4
Dense matching methods								
GLUNet-Geometric [50]	18.1	13.2	22.8	15.2	7.7	4.6	13.3	7.8
PDCNet-Geometric [51]	29.1	24.0	30.7	21.9	20.4	15.7	20.6	12.6
GOCor-GLUNet-Geometric [49]	30.4	**26.0**	33.4	25.6	20.8	**16.0**	19.8	13.3
WarpC-GLUNet-Geometric (megadepth) [52]	**31.3**	25.4	36.6	**27.3**	**21.9**	15.8	26.4	17.3
WarpC-GLUNet-Geometric (megadepth_stage1) [52]	23.5	19.3	28.1	20.7	13.2	8.9	17.0	10.9
GLUNet-Semantic [50]	18.5	14.4	22.4	15.6	8.7	5.6	12.8	7.8
WarpC-GLUNet-Semantic [52]	27.5	21.4	**36.8**	25.7	18.5	11.9	**28.3**	**17.6**

formance in both image and pixel level can be a practical research topic.

Although the detect-2-retrieval [45] is inferior to SP in image retrieval [6, 28, 33], 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 [45]. 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 [18].

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 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. 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

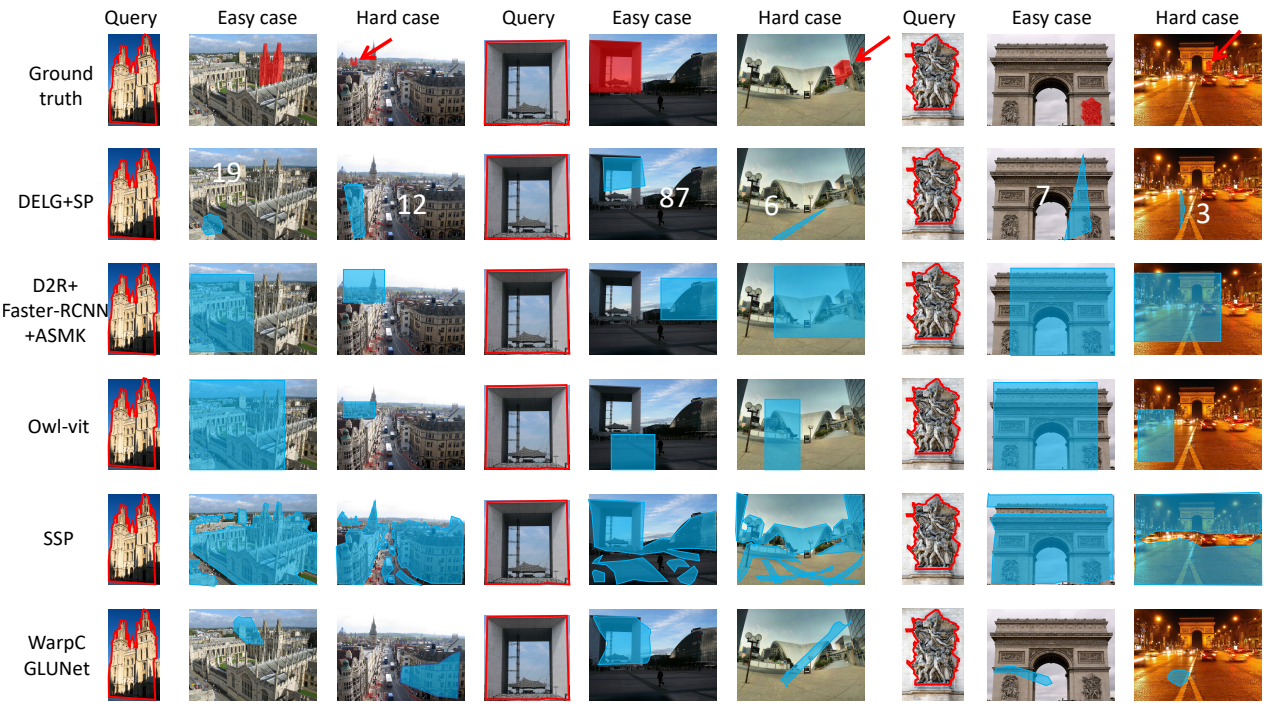

Figure 4. Qualitative comparison of the SOTA methods in different fields on the pixel retrieval benchmarks. Blue masks represent the prediction results of each method. For SP and WarpCGLUNet, we consider the union of all the matching points as the prediction masks. We also show the inlier numbers for the SP method. Pixel retrieval is challenging for existing methods and further research is needed.

Table 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. **Bold** indicates the best performance using the same image ranking list; **red** indicates the best performance in two ranking lists. Green lines show the image level mAPs of the ranking lists.

		PROxf		PROxf+R1M		PRPar		PRPar+R1M	
		M	H	M	H	M	H	M	H
Image retrieval: DELG initial ranking [6]									
Image level mAP		76.3	55.6	63.7	37.5	86.6	72.4	70.6	46.9
Pixel retrieval methods	DELG + SP [6]	6.1	6.3	5.8	6.7	10.9	8.0	10.5	7.8
	D2R+Faster-RCNN+ASMK [45]	29.6	22.5	28.8	19.1	26.3	25.6	23.7	20.5
	OWL-VIT [26]	13.1	8.1	12.8	7.2	8.3	12.7	7.6	11.4
	SSP [9]	**37.3**	34.6	**36.6**	29.9	**47.0**	**43.1**	**44.5**	**37.1**
	WarpCGLUNet [52]	34.3	**36.8**	33.9	**34.9**	33.9	28.8	32.9	27.1
Image retrieval: DELG initial ranking [6] + HP reranking [1]									
Image level mAP		85.7	70.3	78.0	60.0	92.6	83.3	86.6	72.7
Pixel retrieval methods	DELG + SP [6]	6.4	7.2	6.2	7.5	10.7	6.0	10.7	5.9
	D2R+Faster-RCNN+ASMK [45]	30.1	23.5	30.5	22.0	26.3	25.3	25.7	24.9
	OWL-VIT [26]	12.3	6.6	12.1	13.6	7.9	7.6	7.9	7.8
	SSP [9]	**33.0**	29.7	**35.7**	**30.5**	**46.4**	**37.2**	**45.6**	**37.2**
	WarpCGLUNet [52]	31.2	**32.6**	31.5	31.7	34.1	27.3	34.3	28.1

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.

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:

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.

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.

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.

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.

Acknowledgments and Disclosure of Funding

We would like to thank the anonymous reviewers for their valueable comments and insightful suggestions. Sung-Eui Yoon and Rong Li are the corresponding authors of the paper. This work was supported by the Institute of Information & communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT) (RS-2023-00237965, Recognition, Action and Interaction Algorithms for Open-world Robot Service) and the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (No. RS-2023-00208506).

References

[1] Guoyuan An, Yuchi Huo, and Sung-Eui Yoon. Hypergraph propagation and community selection for objects retrieval. *Advances in Neural Information Processing Systems*, 34:3596–3608, 2021. 3, 6, 8

[2] Relja Arandjelović and Andrew 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. 2, 5

[3] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. Surf: Speeded up robust features. In *European conference on computer vision*, pages 404–417. Springer, 2006. 5

[4] Daniel Bolya, Chong Zhou, Fanyi Xiao, and Yong Jae Lee. Yolact: Real-time instance segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9157–9166, 2019. 2

[5] Eva Breznik, Elisabeth Wetzer, Joakim Lindblad, and Nataša Sladoje. Cross-modality sub-image retrieval using contrastive multimodal image representations. *arXiv preprint arXiv:2201.03597*, 2022. 3, 5

[6] Bingyi Cao, André Araujo, and Jack Sim. Unifying deep local and global features for image search. In *European Conference on Computer Vision*, pages 726–743. Springer, 2020. 2, 3, 5, 6, 7, 8

[7] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt 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. 2

[8] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010. 2, 4

[9] Qi Fan, Wenjie Pei, Yu-Wing Tai, and Chi-Keung Tang. Self-support few-shot semantic segmentation. 2022. 6, 7, 8

[10] Andreas Geiger, Philip Lenz, and Raquel 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. 2

[11] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 1, 2

[12] Herve Jegou, Matthijs Douze, and Cordelia Schmid. Hamming embedding and weak geometric consistency for large scale image search. In *European conference on computer vision*, pages 304–317. Springer, 2008. 2

[13] Hervé Jégou, Matthijs Douze, Cordelia Schmid, and Patrick 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. 2, 5, 7

[14] Yuhe Jin, Dmytro Mishkin, Anastasiia Mishchuk, Jiri Matas, Pascal Fua, Kwang Moo Yi, and Eduard Trulls. Image Matching across Wide Baselines: From Paper to Practice. *International Journal of Computer Vision*, 2020. 2, 5

[15] Kurt Koffka. *Principles of Gestalt psychology*. Routledge, 2013. 6

[16] Christoph 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. 3, 5

[17] Seongwon Lee, Hongje Seong, Suhyeon Lee, and Euntai Kim. Correlation verification for image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5374–5384, 2022. 3

[18] Zhengqi Li and Noah 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. 2, 6, 7

[19] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 2, 4

[20] Zhe Lin and Jonathan Brandt. A local bag-of-features model for large-scale object retrieval. In *European conference on Computer vision*, pages 294–308. Springer, 2010. 3, 5

[21] Shu Liu, Lu Qi, Haifang Qin, Jianping Shi, and Jiaya Jia. Path aggregation network for instance segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8759–8768, 2018. 1, 2

[22] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016. 5

[23] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004. 5

[24] Neville Mehta, Alomari Raja'S, and Vipin 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. 3, 5

[25] Juhong Min, Dahyun Kang, and Minsu Cho. Hypercorrelation squeeze for few-shot segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6941–6952, 2021. 2, 6, 7

[26] Matthias Minderer, Alexey Gritsenko, Austin Stone, Maxim Neumann, Dirk Weissenborn, Alexey Dosovitskiy, Aravindh Mahendran, Anurag Arnab, Mostafa Dehghani, Zhuoran Shen, et al. Simple open-vocabulary object detection with vision transformers. *arXiv preprint arXiv:2205.06230*, 2022. 6, 7, 8

[27] David Nister and Henrik 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. 2

[28] Hyeonwoo Noh, Andre Araujo, Jack Sim, Tobias Weyand, and Bohyung Han. Large-scale image retrieval with attentive deep local features. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. 2, 5, 6, 7

[29] Anton Osokin, Denis Sumin, and Vasily Lomakin. OS2D: One-stage one-shot object detection by matching anchor features. In *proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 7

[30] James Philbin, Ondrej Chum, Michael Isard, Josef Sivic, and Andrew 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. 2, 7

[31] James Philbin, Ondrej Chum, Michael Isard, Josef Sivic, and Andrew 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. 2

[32] Jean Ponce, Tamara L Berg, Mark Everingham, David A Forsyth, Martial Hebert, Svetlana Lazebnik, Marcin Marszalek, Cordelia Schmid, Bryan C Russell, Antonio Torralba,

et al. Dataset issues in object recognition. *Toward category-level object recognition*, pages 29–48, 2006. 3

[33] Filip Radenović, Ahmet Iscen, Giorgos Tolias, Yannis Avrithis, and Ondřej 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. 2, 3, 4, 7

[34] Filip Radenović, Giorgos Tolias, and Ondřej Chum. Fine-tuning cnn image retrieval with no human annotation. *IEEE transactions on pattern analysis and machine intelligence*, 41(7):1655–1668, 2018. 2

[35] Ali S Razavian, Josephine Sullivan, Stefan Carlsson, and Atsuto Maki. Visual instance retrieval with deep convolutional networks. *ITE Transactions on Media Technology and Applications*, 4(3):251–258, 2016. 5

[36] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *arXiv preprint arXiv:1506.01497*, 2015. 4, 5

[37] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015. 3

[38] Thomas Schops, Johannes L Schonberger, Silvano Galliani, Torsten Sattler, Konrad Schindler, Marc Pollefeys, and Andreas 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. 6

[39] Xi Shen, François Darmon, Alexei A Efros, and Mathieu Aubry. Ransac-flow: generic two-stage image alignment. In *European Conference on Computer Vision*, pages 618–637. Springer, 2020. 6

[40] Xiaohui Shen, Zhe Lin, Jonathan Brandt, and Ying Wu. Spatially-constrained similarity measurefor large-scale object retrieval. *IEEE transactions on pattern analysis and machine intelligence*, 36(6):1229–1241, 2013. 3, 5

[41] Xiaoyong Shen, Xin Tao, Chao Zhou, Hongyun Gao, and Jiaya Jia. Regional foremost matching for internet scene images. *ACM Transactions on Graphics (TOG)*, 35(6):1–12, 2016. 5

[42] Yujiao Shi and Hongdong 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. 3

[43] Yael Shrager, Jeffrey J Gold, Ramona O Hopkins, and Larry R Squire. Intact visual perception in memory-impaired patients with medial temporal lobe lesions. *Journal of Neuroscience*, 26(8):2235–2240, 2006. 3

[44] Ran Tao, Efstratios Gavves, Cees GM Snoek, and Arnold WM 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. 5

[45] Marvin Teichmann, Andre Araujo, Menglong Zhu, and Jack 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. 5, 7, 8

[46] Giorgos Tolias, Yannis Avrithis, and Hervé 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. 5

[47] Giorgos Tolias, Yannis Avrithis, and Hervé Jégou. Image search with selective match kernels: aggregation across single and multiple images. *International Journal of Computer Vision*, 116(3):247–261, 2016. 7

[48] Giorgos Tolias, Ronan Sicre, and Hervé Jégou. Particular object retrieval with integral max-pooling of cnn activations. *arXiv preprint arXiv:1511.05879*, 2015. 2

[49] Prune Truong, Martin Danelljan, Luc V Gool, and Radu Timofte. Gocor: Bringing globally optimized correspondence volumes into your neural network. *Advances in Neural Information Processing Systems*, 33:14278–14290, 2020. 6, 7

[50] Prune Truong, Martin Danelljan, and Radu 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. 6, 7

[51] Prune Truong, Martin Danelljan, Radu Timofte, and Luc Van Gool. Pdc-net+: Enhanced probabilistic dense correspondence network. *arXiv preprint arXiv:2109.13912*, 2021. 6, 7

[52] Prune Truong, Martin Danelljan, Fisher Yu, and Luc 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. 2, 6, 7, 8

[53] Jasper RR Uijlings, Koen EA Van De Sande, Theo Gevers, and Arnold WM Smeulders. Selective search for object recognition. *International journal of computer vision*, 104(2):154–171, 2013. 5

[54] Tianyi Wei, Dongdong Chen, Wenbo Zhou, Jing Liao, Hanqing Zhao, Weiming Zhang, and Nenghai 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. 5

[55] Tobias Weyand, Andre Araujo, Bingyi Cao, and Jack 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. 3, 5

[56] Ning Xu, Brian Price, Scott Cohen, and Thomas Huang. Deep image matting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2970–2979, 2017. 5

[57] Lin Yang, Xin Qi, Fuyong Xing, Tahsin Kurc, Joel Saltz, and David J Foran. Parallel content-based sub-image retrieval using hierarchical searching. *Bioinformatics*, 30(7):996–1002, 2014. 3, 5

[58] Lihe Yang, Wei Zhuo, Lei Qi, Yinghuan Shi, and Yang Gao. Mining latent classes for few-shot segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8721–8730, 2021. 6, 7

[59] Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas 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. 6

[60] Yunke Zhang, Lixue Gong, Lubin Fan, Peiran Ren, Qixing Huang, Hujun Bao, and Weiwei 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. 5

[61] Zhongyan Zhang, Lei Wang, Yang Wang, Luping Zhou, Jianjia Zhang, and Fang Chen. Dataset-driven unsupervised object discovery for region-based instance image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 3

[62] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 633–641, 2017. 1, 2

[63] Sijie Zhu, Taojiannan Yang, and Chen 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. 3, 5