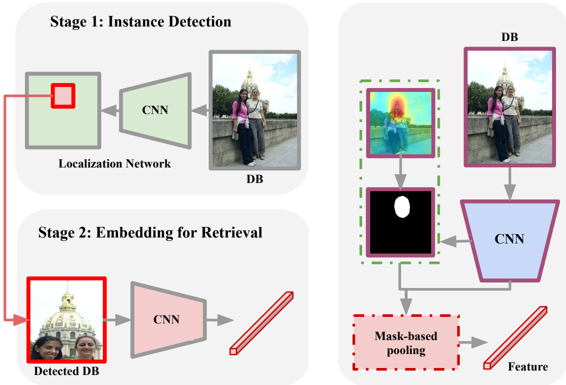
On Train-Test Class Overlap and Detection for Image Retrieval

Chull Hwan Song[1] Jooyoung Yoon[1] Taebaek Hwang[1] Shunghyun Choi[1]
Yeong Hyeon Gu[2*] Yannis Avrithis[3]
[1]Dealicious Inc. [2]Sejong University [3]Institute of Advanced Research on Artificial Intelligence (IARAI)

Abstract

How important is it for training and evaluation sets to not have class overlap in image retrieval? We revisit Google Landmarks v2 clean [56], the most popular training set, by identifying and removing class overlap with Revisited Oxford and Paris [34], the most popular evaluation set. By comparing the original and the new $\mathcal{R}$GLDv2-clean on a benchmark of reproduced state-of-the-art methods, our findings are striking. Not only is there a dramatic drop in performance, but it is inconsistent across methods, changing the ranking.

What does it take to focus on objects or interest and ignore background clutter when indexing? Do we need to train an object detector and the representation separately? Do we need location supervision? We introduce Single-stage Detect-to-Retrieve (CiDeR), an end-to-end, single-stage pipeline to detect objects of interest and extract a global image representation. We outperform previous state-of-the-art on both existing training sets and the new $\mathcal{R}$GLDv2-clean. Our dataset is available at https://github.com/dealicious-inc/RGLDv2-clean.



(a) Two-stage (b) One-stage

Figure 1. It is beneficial for image retrieval to detect objects of interest in database images and only represent those. (a) *Two-stage* pipeline. Previous works involve two-stage embedding extraction at indexing, or a two-stage training process, and they may use location supervision or not. (b) *One-stage* pipeline. We use a single-stage embedding extraction at training and indexing; training is end-to-end and uses no location supervision.

1. Introduction

Instance-level image retrieval is a significant computer vision problem, attracting substantial investigation before and after deep learning. High-quality datasets are crucial for advancing research. Image retrieval has benefited from the availability of landmark datasets [2, 8, 36, 28, 56]. Apart from depicting particular landmarks, an important property of training sets [8, 36] is that they do not contain landmarks overlapping with the evaluation sets [31, 32, 34]. *Google landmarks* [56] has gained widespread adoption in state of the art benchmarks, but falls short in this property [55].

At the same time, a fundamental challenge in image retrieval is to find a particular object among other objects or background clutter. In this direction, it is common to use attention [15, 27, 46] but it is more effective use object detection [41, 40] in order to represent only objects of interest for retrieval. These *detect-to-retrieve* (D2R) [48] methods however, necessitate complex two-stage training and indexing pipelines, as shown in Figure 1(a), often requiring a separate training set with location supervision.

Motivated by the above challenges, we investigate two directions in this work. First, in the direction of *data*, we revisit GLDv2-clean dataset [56]. We analyze and remove overlaps of landmark categories with evaluation sets [34], introducing a new version, $\mathcal{R}$GLDv2-clean. We then reproduce and benchmark state-of-the-art methods on the new dataset and compare with the original. Remarkably, we find that, although the images removed are only a tiny fraction, there is a dramatic drop in performance.

Second, in the direction of the *method*, we introduce CiDeR, a simple attention-based approach to detect objects of interest at different levels and obtain a global image representation that effectively ignores background clutter. Importantly, as shown in Figure 1(b), this is a streamlined end-to-end approach that only needs single-stage training, single-stage indexing and is free of any location supervision.

In summary, we make the following contributions:

1. We introduce $\mathcal{R}$GLDv2-clean, a new version of an established dataset for image retrieval.

*Corresponding author

2. We show that it is critical to have no class overlap between training and evaluation sets.
3. We introduce CiDeR, an end-to-end, single-stage D2R method requiring no location supervision.
4. By using exisiting components developed outside image retrieval, we outperform more complex, specialized state-of-the-art retrieval models on several datasets.

2. Related Works

Instance-level image retrieval Research on image retrieval can be categorized according to the descriptors used. *Local descriptors* [28, 44, 7] have been applied before deep learning, using SIFT [23] for example. Given that multiple descriptors are generated per image, aggregation methods [31, 14, 49] have been developed. Deep learning extensions include methods such as DELF [28], DELG [3], and extensions of ASMK [48, 50]. DELF is similar to our work in that it uses spatial attention without location supervision, but differs in that it uses it for local descriptors.

Global descriptors [2, 56, 46, 59, 47] are useful as they only generate a single feature per image, simplifying the retrieval process. Research has focused on spatial pooling [38, 36, 1, 15, 51, 8, 36] to extract descriptors from 3D convolutional activations. Local descriptors can still be used in a second re-ranking stage after filtering by global descriptors, but this is computationally expensive.

Detect-to-Retrieve (D2R) It is beneficial for image retrieval to detect objects of interest in database images and ignore background clutter [26, 43, 4, 16, 18, 39, 45]. Following Teichmann *et al.* [48], we call these methods *detect-to-retrieve* (D2R). In most existing studies, either training or indexing are two-stage processes, for example learn to detect and learn to retrieve; also, most rely on location supervision in learning to detect.

For example, DIR [8] performs 1-stage indexing but 2-stage training for a region proposal network (RPN) and for retrieval. Its location supervision does not involve humans but rather originates in automatically analyzing the dataset, hence technically training is 3-stage. Salvador *et al.* [43] performs 1-stage end-to-end training, but is using human location supervision, in fact from the *evaluation set*. R-ASMK [48], involves 2-stage training and 2-stage indexing. It also uses large-scale human location supervision from an independent set.

Table 1 shows previous studies organized according to their properties. We can see that, unlike previous studies, we propose a novel method that supports 1-stage training, indexing and inference, as well as allowing end-to-end D2R learning without location supervision. Compared with the previous studies, ours more thus efficient.

METHOD	LD	GD	D2R	E2E	SELF	LAND
DELF [28]	✓					✓
DELG [3]	✓	✓				✓
Tolias *et al.* [50]	✓					✓
DIR [8]		✓				✓
AGeM [9]		✓				✓
SOLAR [27]		✓				✓
GLAM [46]		✓				✓
Kucer *et al.* [16]		✓	✓			
PS-Net [18]		✓	✓			
Peng *et al.* [30]		✓	✓			
Zhang *et al.* [62]		✓	✓			✓
Liao *et al.* [22]		✓	✓			✓
R-ASMK [48]	✓		✓			✓
Salvador *et al.* [43]		✓	✓	✓		✓
CiDeR (Ours)		✓	✓	✓	✓	✓

Table 1. Related work on instance-level image retrieval. LD: local descriptors; GD: global descriptors. [O]: off-the-shelf (pre-trained on ImageNet); D2R: detect-to-retrieve; E2E (D2R only): end-to-end (single-stage) training for detection and retrieval; SELF (D2R only): self-localization (no location supervision); LAND: landmark datasets.

3. Revisiting Google Landmarks v2

Motivation A key weakness of current landmark retrieval datasets is their fragmented origins: training and evaluation sets are often independently collected and released by different studies. Initial datasets contained tens of thousands of images, a number that has now grown into the millions.

Evaluation sets such as Oxford5k (Ox5k) [31] and Paris6k (Par6k) [32], as well as their more recent versions, Revisited Oxford ($\mathcal{R}$Oxford or $\mathcal{R}$Oxf) and Paris ($\mathcal{R}$Paris or $\mathcal{R}$Par) [34], are commonly used for benchmarking. Concurrently, *training sets* such as *Neural Codes* (NC) [2], *Neural Codes clean* (NC-clean) [8], SfM-120k [36], Google Landmarks v1 (GLDv1) [28], and Google Landmarks v2 (GLDv2 and GLDv2-clean) [56] have been sequentially introduced and are widely used for representation learning.

These training sets are typically curated according to two criteria: first, to depict particular landmarks, and second, to not contain landmarks that overlap with those in the evaluation sets. They are originally collected by text-based web search using particular landmark names as queries. This often results in *noisy* images in addition to images depicting the landmarks. Thus, NC, GLDv1 and GLDv2 are *noisy* datasets. To solve this problem, images are filtered in different ways [8, 35] to ensure that they contain only the same landmark (instance). Accordingly, NC-clean, SfM-120k, and GLDv2-clean are *clean* datasets.

The *clean* datasets are also typically filtered to remove overlap with the evaluation sets. However, while NC-clean and SfM-120k adhere to both criteria, GLDv2-clean falls short of the second criterion. This discrepancy is not a lim-

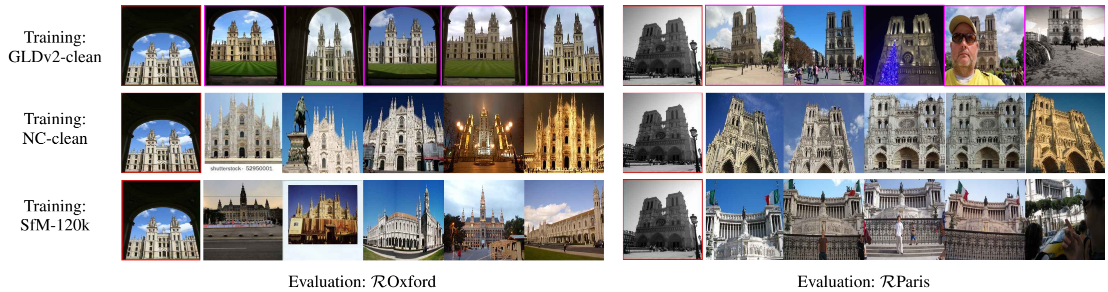


Figure 2. *Confirming overlapping landmark categories* between training sets (GLDv2-clean, NC-clean, SfM-120k) and evaluation sets ($\mathcal{R}$Oxford, $\mathcal{R}$Paris). Red box: query image. The query image from the evaluation set in each box/row is followed by top-5 most similar images from the training set. Pink box: training image landmark identical with query (evaluation) image landmark. More examples can be found in the Appendix.

itation of GLDv2-clean per se, because the dataset comes with its own split of training, index and query images. However, the community is still using the $\mathcal{R}$Oxford and $\mathcal{R}$Paris evaluation sets, whose landmarks have not been removed from GLDv2-clean. Besides, landmarks are still overlapping between the GLDv2-clean training and index sets.

This discrepancy is particularly concerning because GLDv2-clean is the most common training set in state-of-the-art studies. It has been acknowledged in previous work [55] and in broader community discussions[1]. The effect is that results of training on GLDv2-clean are not directly comparable with those of training on NC-clean or SfM-120k. Results on GLDv2-clean may show artificially *inflated performance*. This is often attributed to its larger scale but may in fact be due to overlap. Our study aims to address this problem by introducing a new version of GLDv2-clean.

Identifying overlapping landmarks First, it is necessary to confirm whether common landmark categories exist between the training and evaluation sets. We extract image features from the training sets GLDv2-clean, NC-clean, and SfM-120k, as well as the evaluation sets $\mathcal{R}$Oxf and $\mathcal{R}$Par. The features of the training sets are then indexed and the features of the evaluation sets $\mathcal{R}$Oxf and $\mathcal{R}$Par are used as queries to search into the training sets.

Figure 2 displays the results. Interestingly, none of the retrieved images from NC-clean and SfM-120k training sets depict the same landmark as the query image from the evaluation set. By contrast, the top-5 most similar images from GLDv2-clean all depict the same landmark as the query. This suggests that using GLDv2-clean for training could lead to artificially *inflated performance* during evaluation, when compared to NC-clean and SfM-120k. A fair comparison between training sets should require no overlap with the evaluation set.

Verification Now, focusing on GLDv2-clean training set, we verify the overlapping landmarks. Each image in this set belongs to a landmark category and each category is identified by a GID and has a landmark name. We begin by visual matching. In particular, we retrieve images for each query image from the evaluation set as above and we filter the top-k ranked images by two verification steps.

[1] https://github.com/MCC-WH/Token/issues/1

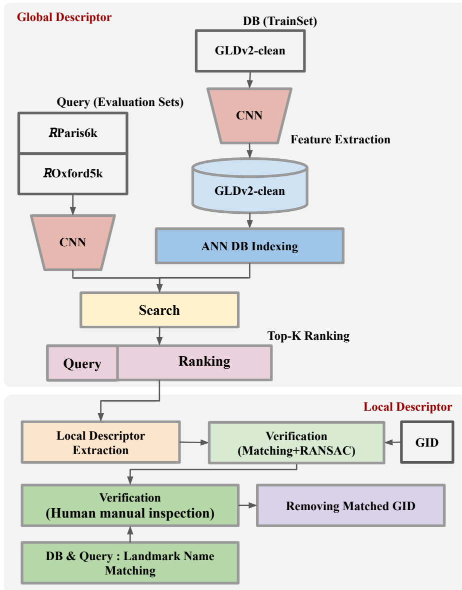


Figure 3. *Ranking and verification pipeline* to remove landmark categories from GLDv2-clean that overlap with those of the $\mathcal{R}$Oxf and $\mathcal{R}$Par evaluation sets and obtain the revisited version, $\mathcal{R}$GLDv2-clean.

First, we automatically verify that the same landmark is depicted by using robust spatial matching on correspondences obtained by local features and descriptors. Second, since automatic verification may fail, three human evaluators visually inspect all matches obtained in the first step. We only keep matches that are confirmed by at least one human evaluator. For every query from the evaluation set, we collect all confirmed visual matches from GLDv2-clean and we remove the entire landmark category of the GID that appears more frequently in this image collection.

EVAL	#EVAL IMG	#DUPL EVAL	#DUPL GLDV2 GID	#DUPL GLDV2 IMG
$\mathcal{R}$Par	70	36 (51%)	11	1,227
$\mathcal{R}$Oxf	70	38 (54%)	6	315
TEXT			1	23
TOTAL	140	74	18	1,565

Table 2. Statistical information about duplicate images/categories with ($\mathcal{R}$Oxford, $\mathcal{R}$Paris) and GLDV2. EVAL:Evaluation Sets. DUPL:duplicated. IMG:Image. GID:GLDV2 category.

TRAINING SET	#IMAGES	#CATEGORIES
NC-clean	27,965	581
SfM-120k	117,369	713
GLDv2-clean	1,580,470	81,313
$\mathcal{R}$GLDv2-clean (ours)	1,578,905	81,295

Table 3. Statistics of clean landmark training sets for image retrieval.

Independently, we collect all GIDs where the landmark name contains "Oxford" or "Paris" and we also mark them as candidate for removal. The entire landmark category of a GID is removed if it is confirmed by at least one human evaluator that it is in one the evaluation sets. This is the case for "Hotel des Invalides Paris". Figure 3 illustrates the complete ranking and verification process.

Revisited GLDv2-clean ($\mathcal{R}$GLDv2-clean) By removing a number of landmark categories from GLDv2-clean as specified above, we derive a revisited version of the dataset, which we call *$\mathcal{R}$GLDv2-clean*. As shown in Table 2, $\mathcal{R}$Par and $\mathcal{R}$Oxf have landmark overlap with GLDv2-clean respectively for 36 and 38 out of 70 queries, which corresponds to a percentage of 51% and 54%, respectively. This is a very large percentage, as it represents more than half queries in both evaluation sets. In the new dataset, we remove 1,565 images from 18 GIDs of GLDv2-clean.

Table 3 compares statistics between existing clean datasets and the new $\mathcal{R}$GLDv2-clean. We observe that a very small proportion of images and landmark categories are removed from GLDv2-clean to derive $\mathcal{R}$GLDv2-clean. Yet, it remains to find what is the effect on retrieval performance, when evaluated on $\mathcal{R}$Oxf and $\mathcal{R}$Par. For fair comparisons, we exclude from our experiments previous results obtained by training on GLDv2-clean; we limit to NC-clean, SfM-120k and the new $\mathcal{R}$GLDv2-clean.

4. Single-stage pipeline for D2R

Motivation From the perspective of instance-level image retrieval, the key challenge is that target objects or instances are situated in different contexts within the image. One common solution is to use object localization or detection, isolating the objects of interest from the background. The detected objects are then used to extract an image representation for retrieval, as shown in Figure 1(a). This *two-stage* process can be applied to the indexed set, the queries, or both.

This approach comes with certain limitations. First, in addition to the training set for representation learning, a specialized training set is also required that is annotated with location information for the objects of interest [41, 40]. Second, the two stages are often trained separately rather than end-to-end. Third, this approach incurs higher computational cost at indexing and search because it requires two forward passes through the network for each image.

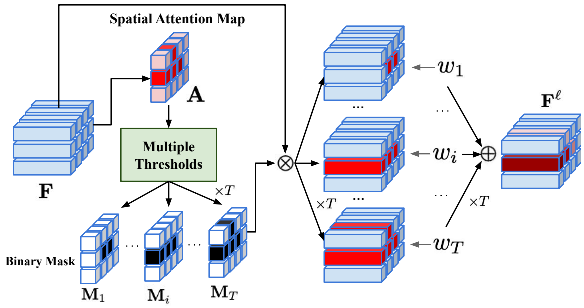


Figure 4. *Attentional localization* (AL). Given a feature tensor $\mathbf{F} \in \mathbb{R}^{w\times h\times d}$, we obtain a spatial attention map $A \in \mathbb{R}^{w\times h}$ (1) and we apply multiple thresholding operations to obtain a sequence of masks $M_1, \dots M_T$ (3). The masks are applied independently to $\mathbf{F}$ and the resulting tensors are fused into a single tensor $\mathbf{F}^\ell$ by a convex combination with learnable weights $w_1, \dots, w_T$ (4).

In this work, we attempt to address these limitations. We replace the localization step with a *spatial attention* mechanism, which does not require location supervision. This allows us to solve for both localization and representation learning through a single, end-to-end learning process on a single network, as illustrated in Figure 1(b). This has the advantage of eliminating the need for a specialized training set for localization and the separate training cycles.

Attentional localization (AL) This component, depicted in Figure 1(b) and elaborated in Figure 4, is designed for instance detection and subsequent image representation based on the detected objects. It employs a spatial attention mechanism [15, 28, 57], which does not need location supervision. Given a feature tensor $\mathbf{F} \in \mathbb{R}^{w\times h\times d}$, where $w \times h$ is the spatial resolution and d the feature dimension, we obtain the *spatial attention map*

$$A = \eta(\zeta(f^\ell(\mathbf{F}))) \in \mathbb{R}^{w\times h}. \quad (1)$$

Here, f^ℓ is a simple mapping, for example a 1×1 convolutional layer that reduces dimension to 1, $\zeta(x) := \ln(1 + e^x)$ for $x \in \mathbb{R}$ is the softplus function and

$$\eta(X) := \frac{X - \min X}{\max X - \min X} \in \mathbb{R}^{w\times h} \quad (2)$$

linearly normalizes $X \in \mathbb{R}^{w\times h}$ to the interval $[0, 1]$. To identify object regions, we then apply a sequence of thresholding operations, obtaining a corresponding sequence of masks

$$M_i(\mathbf{p}) = \begin{cases} \beta, & \text{if } A(\mathbf{p}) < \tau_i \\ 1, & \text{otherwise} \end{cases} \quad (3)$$

for $i \in \{1, \ldots, T\}$. Here, T is the number of masks, $\mathbf{p} \in \{1, \ldots, w\} \times \{1, \ldots, h\}$ is the spatial position, $\tau_i \in [0, 1]$ is the i-th threshold, β is a scalar corresponding to background and 1 corresponds to foreground.

Unlike a conventional fixed value like $\beta = 0$, we use a dynamic, randomized approach. In particular, for each $\mathbf{p}$, we draw a sample ϵ from a normal distribution and we clip it to $[0, 1]$ by defining $\beta = \min(0, \max(1, \epsilon))$. The motivation is that randomness compensates for incorrect predictions of the attention map (1), especially at an early stage of training. This choice is ablated in Table 8.

Figure 5 shows examples of attentional localization. Comparing (a) with (b) shows that the spatial attention map generated by our model is much more attentive to the object being searched than the pretrained network. These results show that the background is removed relatively well, despite not using any location supervision at training.

The sequence of masks $M_1, \ldots, M_T$ (3) is applied independently to the feature tensor $\mathbf{F}$ and the resulting tensors are fused into a single tensor

$$\mathbf{F}^\ell = \mathtt{H}(M_1 \odot \mathbf{F}, \ldots, M_T \odot \mathbf{F}) \in \mathbb{R}^{w \times h \times d}, \quad (4)$$

where $\odot$ denotes Hadamard product over spatial dimension, with broadcasting over the feature dimension. Fusion amounts to a learnable convex combination

$$\mathtt{H}(\mathbf{F}_1, \ldots, \mathbf{F}_T) = \frac{w_1 \mathbf{F}_1 + \cdots + w_T \mathbf{F}_T}{w_1 + \cdots + w_T}, \quad (5)$$

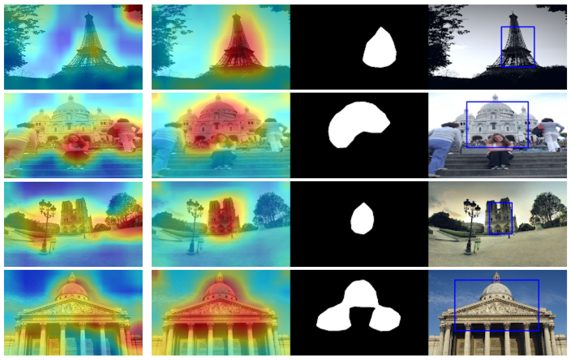
where, for $i \in \{1, \ldots, T\}$, the i-th weight is defined as $w_i = \zeta(\alpha_i)$ and α_i is a learnable parameter. Thus, the importance of each threshold in localizing objects from the spatial attention map is implicitly learned from data, without supervision. Table 9 ablates the effect of the number T of thresholds on the fusion efficacy.

5. Experiments

5.1. Implementation

Components Most instance-level image retrieval studies propose a kind of head on top of the backbone network that performs a particular operation to enhance retrieval performance. The same is happening independently in studies of category-level tasks like localization, even though the operations may be similar. Comparison is often challenging, when official code is not released. Our focus is on detection for retrieval in this work but we still need to compare with SOTA methods, which may perform different operations. We thus follow a neutral approach whereby we reuse existing, well-established components from the literature, introduced either for instance-level or category-level tasks.

In particular, given an input image $x \in \mathcal{X}$, where $\mathcal{X}$ is the image space, we obtain an embedding $\mathbf{u} = f(x) \in \mathbb{R}^d$,



(a) A, pre-trained (b) A, ours (c) Mask M_i (d) Bounding box

Figure 5. *Attentional localization (AL).* (a) Spatial attention map A (1) learned on frozen ResNet101, as pre-trained on ImageNet. (b) Same, but with the network fine-tuned on $\mathcal{R}$GLDv2-clean. (c) Binary mask M_i (3) for $i = 2$, with $\beta = 0$ for visualization. (d) Detected regions as bounding boxes of connected components of M_i, overlaid on input image (in blue).

where d is the embedding dimension and

$$f = f^p \circ f^\ell \circ f^c \circ f^e \circ f^b \quad (6)$$

is the composition of a number of functions. Here,

- $f^b : \mathcal{X} \to \mathbb{R}^{w \times h \times d}$ is the *backbone network*;
- $f^e : \mathbb{R}^{w \times h \times d} \to \mathbb{R}^{w \times h \times d}$ is *backbone enhancement* (BE), including non-local interactions like ECNet [53], NLNet [54], Gather-Excite [12] or SENet [13];
- $f^c : \mathbb{R}^{w \times h \times d} \to \mathbb{R}^{w \times h \times d}$ is *selective context* (SC), enriching contextual information to apply locality more effectively like ASPP [5] or SKNet [21];
- $f^\ell : \mathbb{R}^{w \times h \times d} \to \mathbb{R}^{w \times h \times d}$ is our *attentional localization* (AL) (section 4), localizing objects of interest in an unsupervised fashion;
- $f^p : \mathbb{R}^{w \times h \times d} \to \mathbb{R}^d$ is a *spatial pooling* operation, such as GAP or GeM [36], optionally followed by other mappings, *e.g.* whitening.

In the Appendix, we ablate different options for f^e, f^c and we specify our choice for f^p; then in subsection 5.5 we ablate, apart from hyperparameters of f^ℓ, the effect of the presence of components f^e, f^c, f^ℓ on the overall performance. By default, we embed images using f (6), where for each component we use default settings as specified in subsection 5.5 or in the Appendix.

Settings Certain existing works [8, 28] train the backbone network first on classification loss without the head corresponding to the method and then fine-tune including the head. We refer to this approach as "fine-tuning" (FT). To allow for comparisons, we train our model in two ways. *Without fine-tuning*, referred to as CiDeR, everything is trained in a single stage end-to-end. *With fine-tuning*, referred to as CiDeR-FT, we freeze the backbone while only training the head in the second stage. We give more details in the Appendix, along

Method	Train Set	Base		Medium				Hard				Mean	Diff
		Ox5k	Par6k	$\mathcal{R}$Oxf		$\mathcal{R}$Par		$\mathcal{R}$Oxf		$\mathcal{R}$Par			
		mAP	mAP	mAP	mP@10	mAP	mP@10	mAP	mP@10	mAP	mP@10		
Yokoo *et al.* [46]	GLDv2-clean	91.9	94.5	72.8	86.7	84.2	95.9	49.9	62.1	69.7	88.4	79.5	-5.4
Yokoo *et al.* [60]†	$\mathcal{R}$GLDv2-clean	86.1	93.9	64.5	81.0	84.1	95.4	35.6	51.5	68.7	86.4	74.1	
SOLAR [58]	GLDv2-clean	–	–	79.7	–	88.6	–	60.0	–	75.3	–	75.9	-8
SOLAR [27]†	$\mathcal{R}$GLDv2-clean	90.6	94.4	70.8	84.6	84.1	95.4	48.0	62.3	68.7	86.4	67.9	
GLAM [46]	GLDv2-clean	94.2	95.6	78.6	88.2	88.5	97.0	60.2	72.9	76.8	93.4	83.4	-4.1
GLAM [46]‡	$\mathcal{R}$GLDv2-clean	90.9	94.1	72.2	84.7	83.0	95.0	49.6	61.6	65.6	87.6	79.3	
DOLG [47]	GLDv2-clean	–	–	78.8	–	87.8	–	58.0	–	74.1	–	74.7	-7.4
DOLG [59]†	$\mathcal{R}$GLDv2-clean	88.3	93.9	70.8	85.3	83.2	95.4	47.4	60.0	67.9	87.4	67.3	
Token [58]	GLDv2-clean	–	–	82.3	–	75.6	–	66.6	–	78.6	–	75.8	-18.2
Token [58]†	$\mathcal{R}$GLDv2-clean	84.3	90.0	61.4	76.4	75.8	94.0	36.9	55.2	54.4	81.0	57.6	

Table 4. Comparison of the original GLDv2-clean training set with our revisited version $\mathcal{R}$GLDv2-clean for a number of SOTA methods that we reproduce with ResNet101 backbone, ArcFace loss and same sampling, settings and hyperparameters. †/‡: official/our code.

with all experimental setings.

5.2. Revisited *vs*. original GLDv2-clean

We reproduce a number of state-of-the-art (SOTA) methods using official code where available, we train them on both the original GLDv2-clean dataset our revisited version $\mathcal{R}$GLDv2-clean and we compare their performance on the evaluation sets. To ensure a fair evaluation, we use the same ResNet101 backbone [8, 15, 36, 9, 27, 60, 46, 59, 58] and ArcFace loss [60, 46, 59, 58, 47] as in previous studies.

Table 4 shows that using $\mathcal{R}$GLDv2-clean leads to severe performance degradation across all methods, ranging from 1% up to 30%. Because the difference between the two training sets in terms of both images and landmark categories is very small (Table 3), this degradation can be safely attributed to the overlap of landmarks between the original training set, GLDv2-clean, and the evaluation sets, Oxford5k and Paris6k, as discussed in section 3. In other words, this experiment demonstrates that existing studies using GLDv2-clean as a training set have artificially inflated accuracy metrics comparing with studies using other training sets with no overlap, such as NC-clean and SfM-120k.

5.3. Comparison with state of the art

Existing clean datasets Table 5 compares different methods using global or local descriptors, with or without a D2R approach, on existing *clean datasets* NC-clean and SfM-120k, which do not overlap with the evaluation sets.

Comparing with methods using global descriptors without D2R, our method demonstrates SOTA performance and brings significant improvements over AGeM [9], the previous best competitor. In particular, 2.9%, 0.6% mAP on Ox5k, Par6k Base, 9.2%, 18.2% on $\mathcal{R}$Oxf, $\mathcal{R}$Par Medium, and 6.4%, 9.5% on $\mathcal{R}$Oxf, $\mathcal{R}$Par Hard.

Comparing with methods using global descriptors without D2R, our method outperforms the highest-ranking approach by DIR+RPN [8], which was trained on the SfM-120k dataset. Specifically, our method improves mAP by 7.4% on Ox5k dataset and by 1.1% on Par6k. Interestingly, methods in the D2R category employ different training sets, as no single dataset provides annotations for both D2R tasks. Our study is unique in being single-stage, end-to-end (E2E) trainable and at the same time requiring no location supervision (Loc), thereby eliminating the need for a detection-specific training set.

New clean dataset, distractors Table 6 provides complete experimental results, including the impact of introducing 1 million distractors ($\mathcal{R}$1M) into the evaluation set, on our new clean training set, $\mathcal{R}$GLDv2-clean, as well as the previous most popular clean set, SfM-120k. Contrary to previous studies, we compare methods trained on the same training and evaluation sets to ensure fairness.

Without fine-tuning, we improve 1.3% mAP on $\mathcal{R}$Oxf +$\mathcal{R}$1M (medium), 5.1% on $\mathcal{R}$Oxf+$\mathcal{R}$1M (hard), 1.7% on $\mathcal{R}$Paris+$\mathcal{R}$1M (medium), and 0.8% on $\mathcal{R}$Paris+$\mathcal{R}$1M (hard) compared to DOLG [59] on $\mathcal{R}$GLDv2-clean. With fine-tuning, our CiDeR-FT establishes new SOTA for nearly all metrics. In particular, we improve 4.5% mAP on $\mathcal{R}$Oxf +$\mathcal{R}$1M (medium), 5.3% on $\mathcal{R}$Oxf+$\mathcal{R}$1M (hard), 4.3% on $\mathcal{R}$Paris+$\mathcal{R}$1M (medium), and 3.1% on $\mathcal{R}$Paris+$\mathcal{R}$1M (hard) compared to DOLG [59] on $\mathcal{R}$GLDv2-clean.

5.4. Visualization

Ranking and spatial attention Figure 6 shows examples of the top-5 ranking images retrieved for a number of queries by our model, along with the associated spatial attention map. The spatial attention map A (1) focuses exclusively on the object of interest as specified by the cropped area provided by the evaluation set, essentially ignoring the background.

Embedding space Figure 7 shows t-SNE visualizations of image embeddings of the $\mathcal{R}$Paris dataset [34], obtained by the off-the shelf network as pre-trained on ImageNet *vs*. our method with fine-tuning on SfM-120k [36]. It indicates superior embedding quality for our model.

Method	Train Set	Net	Pooling	Loss	FT	E2E	Self	Dim	Base: Oxf5k	Base: Par6k	$\mathcal{R}$Medium: $\mathcal{R}$Oxf	$\mathcal{R}$Medium: $\mathcal{R}$Par	$\mathcal{R}$Hard: $\mathcal{R}$Oxf	$\mathcal{R}$Hard: $\mathcal{R}$Par	Mean
Local Descriptors															
HesAff-rSIFT-ASMK*+SP [34]	SfM-120k	R50	–	–	✓	–	–	–	–	–	60.6	61.4	36.7	35.0	–
DELF-ASMK*+SP [34]	SfM-120k	R50	–	CLS	✓	–	–	–	–	–	**67.8**	**76.9**	**43.1**	**55.4**	–
Local Descriptors+D2R															
R-ASMK* [48]	NC-clean	R50	–	CLS,LOCAL	✓			–	–	–	69.9	**78.7**	45.6	**57.7**	–
R-ASMK*+SP [48]	NC-clean	R50	–	CLS,LOCAL	✓			–	–	–	**71.9**	78.0	**48.5**	54.0	–
Global Descriptors															
DIR [47]	SfM-120k	R101	RMAC	TP	✓	–	–	2048	79.0	86.3	53.5	68.3	25.5	42.4	59.2
Radenovic *et al.* [36, 34]	SfM-120k	R101	GeM	SIA		–	–	2048	87.8	**92.7**	64.7	77.2	38.5	56.3	69.5
AGeM [9]	SfM-120k	R101	GeM	SIA		–	–	2048	–	–	**67.0**	**78.1**	**40.7**	**57.3**	–
SOLAR [47]	SfM-120k	R101	GeM	TP,SOS	✓	–	–	2048	78.5	86.3	52.5	70.9	27.1	46.7	60.3
GLAM [46]	SfM-120k	R101	GeM	AF		–	–	512	**89.7**	91.1	66.2	77.5	39.5	54.3	**69.7**
DOLG [47]	SfM-120k	R101	GeM,GAP	AF		–	–	512	72.8	74.5	46.4	56.6	18.1	26.6	49.2
Global Descriptors+D2R															
Mei *et al.* [26]	[O]	R101	FC	CLS				4096	38.4	–	–	–	–	–	–
Salvador *et al.* [43]	Pascal VOC	V16	GSP	CLS,LOCAL		✓		512	67.9	72.9	–	–	–	–	–
Chen *et al.* [4]	OpenImageV4 [17]	R50	MAC	MSE		✓		2048	50.2	65.2	–	–	–	–	–
Liao *et al.* [22]	Oxford,Paris	A,V16	CroW	CLS,LOCAL				768	80.1	90.3	–	–	–	–	–
DIR+RPN [8]	NC-clean	R101	RMAC	TP	✓			2048	**85.2**	**94.0**	–	–	–	–	–
CiDeR (Ours)	SfM-120k	R101	GeM	AF		✓	✓	2048	**89.9**	92.0	**67.3**	**79.4**	**42.4**	57.5	**71.4**
CiDeR-FT (Ours)	SfM-120k	R101	GeM	AF	✓	✓	✓	2048	**92.6**	**95.1**	**76.2**	**84.5**	**58.9**	**68.9**	**79.4**

Table 5. Properties and mAP comparison of SOTA on existing training sets with no overlap with evaluation sets. FT: fine-tuning; E2E (D2R only): end-to-end (single-stage) training for detection and retrieval; Self (D2R only): self-localization (no location supervision). *Network*: R50/101: ResNet50/101; V16: VGG16; A: AlexNet. *Pooling*: GAP: global average pooling; GSP: global sum pooling. *Loss*: AF: ArcFace; TP: triplet; CLS: softmax; SIA: siamese; SOS: second-order similarity; MSE: mean square error; LOCAL: Localization Loss; SP: spatial verification. [O]: Off-the-shelf (pre-trained on ImageNet). Red: best result; blue: our results higher than previous methods; black bold: best previous method per block.

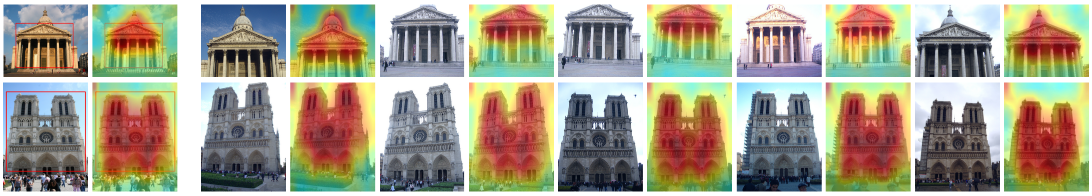
Method	Base: Ox5k mAP	Base: Par6k mAP	Medium: $\mathcal{R}$Oxf mAP	Medium: $\mathcal{R}$Oxf mP@10	Medium: $\mathcal{R}$Oxf +$\mathcal{R}$1M mAP	Medium: $\mathcal{R}$Oxf +$\mathcal{R}$1M mP@10	Medium: $\mathcal{R}$Par mAP	Medium: $\mathcal{R}$Par mP@10	Medium: $\mathcal{R}$Par +$\mathcal{R}$1M mAP	Medium: $\mathcal{R}$Par +$\mathcal{R}$1M mP@10	Hard: $\mathcal{R}$Oxf mAP	Hard: $\mathcal{R}$Oxf mP@10	Hard: $\mathcal{R}$Oxf +$\mathcal{R}$1M mAP	Hard: $\mathcal{R}$Oxf +$\mathcal{R}$1M mP@10	Hard: $\mathcal{R}$Par mAP	Hard: $\mathcal{R}$Par mP@10	Hard: $\mathcal{R}$Par +$\mathcal{R}$1M mAP	Hard: $\mathcal{R}$Par +$\mathcal{R}$1M mP@10
Global Descriptors (SfM-120k)																		
DIR [47]	79.0	86.3	53.5	76.9	–	–	68.3	97.7	–	–	25.5	42.0	–	–	42.4	83.6	–	–
Filip *et al.* [36, 34]	87.8	92.7	64.7	**84.7**	**45.2**	**71.7**	77.2	**98.1**	**52.3**	**95.3**	38.5	**53.0**	**19.9**	**34.9**	56.3	**89.1**	24.7	**73.3**
AGeM [9]	–	–	**67.0**	–	–	–	**78.1**	–	–	–	**40.7**	–	–	–	57.3	–	–	–
SOLAR [47]	78.5	86.3	52.5	73.6	–	–	70.9	**98.1**	–	–	27.1	41.4	–	–	46.7	83.6	–	–
GeM [47]	79.0	82.6	54.0	72.5	–	–	64.3	92.6	–	–	25.8	42.2	–	–	36.6	67.6	–	–
GLAM [47]	**89.7**	91.1	66.2	–	–	–	77.5	–	–	–	39.5	–	–	–	54.3	–	–	–
DOLG [47]	72.8	74.5	46.4	66.8	–	–	56.6	91.1	–	–	18.1	27.9	–	–	26.6	62.6	–	–
CiDeR (Ours)	**89.9**	92.0	**67.3**	**85.1**	**50.3**	**75.5**	**79.4**	97.9	51.4	**95.7**	**42.4**	**56.4**	**22.4**	**35.9**	57.5	87.1	22.4	69.4
CiDeR-FT (Ours)	**92.6**	**95.1**	**76.2**	**87.3**	**60.5**	**78.6**	**84.5**	98.0	**56.9**	**95.9**	**58.9**	**71.1**	**36.8**	**55.7**	**68.9**	**91.3**	**30.1**	**73.9**
Global Descriptors ($\mathcal{R}$GLDv2-clean)																		
Yokoo *et al.* [60]† (Base)	86.1	93.9	64.5	81.0	51.3	72.1	84.1	**95.4**	54.2	90.3	35.6	51.5	22.2	42.9	**68.7**	86.4	27.4	66.9
SOLAR [27]†	90.6	**94.4**	70.8	84.6	55.8	76.1	80.3	94.6	57.6	**92.0**	48.0	**62.3**	30.3	45.3	61.8	83.9	30.7	71.6
GLAM [46]‡	**90.9**	94.1	**72.2**	84.7	**58.6**	76.1	83.0	95.0	**58.6**	91.7	**49.6**	61.6	**34.1**	**50.9**	65.6	**87.6**	**33.3**	72.1
DOLG [59]†	88.3	93.9	70.8	**85.3**	57.3	**76.8**	**83.2**	**95.4**	57.3	**92.0**	47.4	60.0	29.5	46.2	67.9	87.4	32.7	**72.4**
Token [58]†	81.2	89.6	60.8	77.7	44.0	60.9	75.8	94.3	44.1	86.9	37.3	54.1	23.2	37.7	54.8	81.3	19.7	54.4
CiDeR (Ours)	89.8	**94.6**	**73.7**	**85.5**	**58.6**	76.3	**84.6**	**96.7**	**59.0**	**95.1**	**54.9**	**66.6**	**34.6**	**54.7**	**68.5**	**89.1**	**33.5**	**76.9**
CiDeR-FT (Ours)	**90.9**	**96.1**	**77.8**	**88.0**	**61.8**	**78.0**	**87.4**	**97.0**	**61.6**	**94.3**	**61.9**	**70.4**	**39.4**	**56.8**	**75.3**	**90.0**	**35.8**	**72.7**

Table 6. Large-scale mAP comparison of SOTA on training sets with no overlap with evaluation sets. In the new $\mathcal{R}$GLDv2-clean, settings are same as in Table 4. In the existing SfM-120k, results are as published. †/‡: official/our code. Red: best results; blue: our results higher than previous methods; black bold: best previous method per block. FT:fine-tuning.

5.5. Ablation study

Design ablation We study the effect of the presence of components f^e, f^c, f^ℓ (6) on the overall performance of the proposed model. Starting from the baseline, which is ResNet101 backbone (f^b) followed by GeM pooling (f^p), we add selective context (SC, f^c), attentional localization (AL, f^ℓ) and backbone enhancement (BE, f^e). Table 7 provides the results, illustrating the performance gains achieved by the proposed components.

Mask background β We study the effect of setting the background value β in masks (3) to a fixed value *vs.* clipping



Query Top-1 Top-2 Top-3 Top-4 Top-5

Figure 6. Examples of top-5 ranking images retrieved by our CiDeR model from evaluation sets Ox5k/Par6k and associated spatial attention map A (1). The red rectangle within the query on the left is the cropped area provided by the evaluation set and is actually used as the query image.

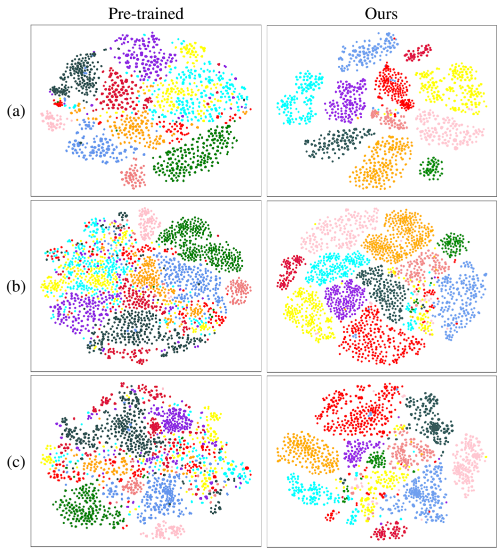


Figure 7. T-SNE visualization of image embeddings of the *revisited Paris* ($\mathcal{R}$Par) evaluation set under (a) *easy*, (b) *medium*, and (c) *hard* protocols [34]. Pre-trained: ResNet101 off-the shelf as pre-trained on ImageNet. Ours: our CiDeR-FT with fine-tuning on SfM-120k [36]. Positive images for each protocol are colored based on their query landmark category.

SC	AL	BE	Oxf5k	Par6k	Medium		Hard	
					$\mathcal{R}$Oxf	$\mathcal{R}$Par	$\mathcal{R}$Oxf	$\mathcal{R}$Par
			80.2	83.2	55.1	67.7	25.8	40.7
✓			87.6	90.7	64.7	76.6	38.2	52.7
✓		✓	89.4	91.1	66.1	76.7	40.6	53.3
	✓	✓	88.2	91.5	66.0	78.4	40.8	55.9
✓	✓		89.7	**92.0**	67.0	**79.4**	41.0	57.4
✓	✓	✓	**89.9**	**92.0**	**67.3**	**79.4**	**42.4**	**57.5**

Table 7. Effect of different components on mAP performance. Training on SfM-120k. Baseline: ResNet101 with GeM pooling. SC: selective context; AL: attentional localization; BE: backbone enhancement.

a sample ϵ from the normal distribution. Table 8 indicates that our dynamic, randomized approach is superior when $\epsilon \sim \mathcal{N}(0.1, 0.9)$, which we choose as default.

Number of masks T We study the effect of the number of masks T (3) in our attentional localization, obtained by thresholding operations on the spatial attention map A (1). Table 9 shows that optimal performance is achieved for $T =$

β setting	Oxf5k	Par6k	Medium		Hard	
			$\mathcal{R}$Oxf	$\mathcal{R}$Par	$\mathcal{R}$Oxf	$\mathcal{R}$Par
Fixed (0.0)	87.4	91.6	64.9	77.5	39.1	53.8
Fixed (0.5)	87.5	91.7	64.8	77.7	38.8	54.3
$\mathcal{N}(0.1, 0.5)$	**90.2**	90.5	**67.4**	78.1	40.2	55.2
$\mathcal{N}(0.1, 0.9)$	89.9	**92.0**	67.3	**79.4**	**42.4**	**57.5**

Table 8. Effect on mAP of different *mask background* β (3) settings in our attentional localization. Training on SfM-120k.

T	Oxf5k	Par6k	Medium		Hard	
			$\mathcal{R}$Oxf	$\mathcal{R}$Par	$\mathcal{R}$Oxf	$\mathcal{R}$Par
1	87.5	91.7	64.8	77.7	38.8	54.3
2	**89.9**	92.0	67.3	**79.4**	**42.4**	**57.5**
3	89.4	**92.2**	**67.5**	78.5	**42.4**	55.3
6	89.4	91.6	66.5	78.1	40.5	55.0

Table 9. Effect on mAP of *number of masks* T (3) in our attentional localization. Training on SfM-120k.

2, which we choose as default.

6. Conclusion

We confirm that training and evaluation sets for instance-level image retrieval really should not have class overlap. Our new $\mathcal{R}$GLDv2-clean dataset makes fair comparisons possible with previous clean datasets. The comparison between the two versions reveals that class overlap indeed brings inflated performance, although the relative difference in number of images is small. Importantly, the ranking of SOTA methods is different on the two training sets.

On the algorithmic front, D2R methods typically require an additional object detection training stage with location supervision, which is inherently inefficient. Our method CiDeR provides a single-stage training pipeline without the need for location supervision. CiDeR improves the SOTA not only on established clean training sets but also on the newly released $\mathcal{R}$GLDv2-clean.

Acknowledgment

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) under the metaverse support program to nurture the best talents (IITP-2024-RS-2023-00254529) grant funded by the Korea government (MSIT).

References

[1] Artem Babenko and Victor Lempitsky. Aggregating Local Deep Features for Image Retrieval. In *ICCV*, 2015.

[2] Artem Babenko, Anton Slesarev, Alexandr Chigorin, and Victor Lempitsky. Neural codes for image retrieval. In *ECCV*, 2014.

[3] Bingyi Cao, André Araujo, and Jack Sim. Unifying deep local and global features for image search. In *ECCV*, 2020.

[4] Bor-Chun Chen, Zuxuan Wu, Larry Davis, and Ser-Nam Lim. Efficient object embedding for spliced image retrieval. In *CVPR*, 2021.

[5] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. In *arXiv preprint arXiv:1706.05587*, 2017.

[6] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. ArcFace: Additive Angular Margin Loss for Deep Face Recognition. In *CVPR*, 2019.

[7] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superpoint: Self-supervised interest point detection and description. In *CVPRW*, 2018.

[8] Albert Gordo, Jon Almazan, Jerome Revaud, and Diane Larlus. Deep image retrieval: Learning global representations for image search. In *ECCV*, 2016.

[9] Yinzheng Gu, Chuanpeng Li, and Jinbin Xie. Attention-aware generalized mean pooling for image retrieval. In *arXiv preprint arXiv:1811.00202*, 2018.

[10] Jiedong Hao, Jing Dong, Wei Wang, and Tieniu Tan. What is the best practice for cnns applied to visual instance retrieval? In *ICLR*, 2017.

[11] Tong He, Zhi Zhang, Hang Zhang, Zhongyue Zhang, Junyuan Xie, and Mu Li. Bag of tricks for image classification with convolutional neural networks. In *CVPR*, 2018.

[12] Jie Hu, Li Shen, Samuel Albanie, Gang Sun, and Andrea Vedaldi. Gather-excite: Exploiting feature context in convolutional neural networks. In *NeurIPS*, 2018.

[13] Jie Hu, Li Shen, Samuel Albanie, Gang Sun, and Enhua Wu. Squeeze-and-Excitation Networks. In *CVPR*, 2018.

[14] H. Jegou, F. Perronnin, M. Douze, J. Sanchez, P. Perez, and C. Schmid. Aggregating local image descriptors into compact codes. In *PAMI*, 2011.

[15] Yannis Kalantidis, Clayton Mellina, and Simon Osindero. Cross-dimensional weighting for aggregated deep convolutional features. In *ECCV*, 2016.

[16] Michal Kucer and Naila Murray. A detect-then-retrieve model for multi-domain fashion item retrieval. In *CVPRW*, 2019.

[17] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, Tom Duerig, and Vittorio Ferrari. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. In *International Journal of Computer Vision*, 2020.

[18] Yining Lang, Yuan He, Fan Yang, Jianfeng Dong, and Hui Xue. Which is plagiarism: Fashion image retrieval based on regional representation for design protection. In *CVPR*, 2020.

[19] Juho Lee, Yoonho Lee, Jungtaek Kim, Adam Kosiorek, Seungjin Choi, and Yee Whye Teh. Set transformer: A framework for attention-based permutation-invariant neural networks. In *ICML*, 2019.

[20] Seongwon Lee, Hongje Seong, Suhyeon Lee, and Euntai Kim. Correlation verification for image retrieval. In *CVPR*, 2022.

[21] Xiang Li, Wenhai Wang, Xiaolin Hu, and Jian Yang. Selective kernel networks. In *CVPR*, 2019.

[22] Kaiyang Liao, Bing Fan, Yuanlin Zheng, Lin Guangfeng, and Congjun Cao. Bow image retrieval method based on ssd target detection. In *IET Image Processing*, 2020.

[23] D. Lowe. Distinctive image features from scale-invariant keypoints. In *IJCV*, 2004.

[24] TorchVision maintainers and contributors. Torchvision: Pytorch's computer vision library. https://github.com/pytorch/vision, 2016.

[25] Sachin Mehta and Mohammad Rastegari. Mobilevit: Light-weight, general-purpose, and mobile-friendly vision transformer. In *arXiv preprint arXiv:2110.02178*, 2021.

[26] Shuhuan Mei, Weiqing Min, Hua Duan, and Shuqiang Jiang. Instance-level object retrieval via deep region cnn. In *Multimedia Tools and Applications*, 2019.

[27] Tony Ng, Vassileios Balntas, Yurun Tian, and Krystian Mikolajczyk. SOLAR: Second-Order Loss and Attention for Image Retrieval. In *ECCV*, 2020.

[28] Hyeonwoo Noh, Andre Araujo, Jack Sim, Tobias Weyand, and Bohyung Han. Large-scale image retrieval with attentive deep local features. In *ICCV*, 2017.

[29] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An imperative style, high-performance deep learning. In *NeurIPS*, 2019.

[30] Yingshu Peng and Yi Wang3. Leaf disease image retrieval with object detection and deep metric learning. In *Frontiers in Plant Science*, 2022.

[31] James Philbin, Ondrej Chum, Michael Isard, Josef Sivic, and Andrew Zisserman. Object retrieval with large vocabularies and fast spatial matching. In *CVPR*, 2007.

[32] James Philbin, Ondrej Chum, Michael Isard, Josef Sivic, and Andrew Zisserman. Lost in quantization:Improving particular object retrieval in large scale image databases. In *CVPR*, 2008.

[33] Bill Psomas, Ioannis Kakogeorgiou, Konstantinos Karantzalos, and Yannis Avrithis. Keep it simpool: Who said supervised transformers suffer from attention deficit? In *ICCV*, 2023.

[34] Filip Radenović, Ahmet Iscen, Giorgos Tolias, Yannis Avrithis, and Ondřej Chum. Revisiting Oxford and Paris: Large-Scale Image Retrieval Benchmarking. In *CVPR*, 2018.

[35] Filip Radenović, Giorgos Tolias, and Ondřej Chum. CNN image retrieval learns from BoW: Unsupervised fine-tuning with hard examples. In *ECCV*, 2016.

[36] Filip Radenović, Giorgos Tolias, and Ondřej Chum. Fine-tuning cnn image retrieval with no human annotation. In *TPAMI*, 2019.

[37] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

[38] A. Razavian, J. Sullivan, A. Maki, and S. Carlsson. Visual instance retrieval with deep convolutional networks. In *CoRR*, 2015.

[39] Konda Reddy Mopuri and R Venkatesh Babu. Object level deep feature pooling for compact image representation. In *CVPRW*, 2015.

[40] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *CVPR*, 2016.

[41] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *NIPS*, 2015.

[42] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. In *International booktitle of Computer Vision*, 2015.

[43] Amaia Salvador, Xavier Giro-i Nieto, Ferran Marques, and Shin'ichi Satoh. Faster r-cnn features for instance search. In *CVPRW*, 2016.

[44] Oriane Simeoni, Yannis Avrithis, and Ondréj Chum. Local Features and Visual Words Emerge in Activations. In *CVPR*, 2019.

[45] Oriane Siméoni, Ahmet Iscen, Giorgos Tolias, Yannis Avrithis, and Ondřej Chum. Graph-based particular object discovery. *Machine Vision and Applications*, 30(2):243–254, 2019.

[46] Chull Hwan Song, Hye Joo Han, and Yannis Avrithis. All the attention you need: Global-local, spatial-channel attention for image retrieval. In *WACV*, 2022.

[47] Chull Hwan Song, Jooyoung Yoon, Shunghyun Choi, and Yannis Avrithis. Boosting vision transformers for image retrieval. In *WACV*, 2023.

[48] M. Teichmann, A. Araujo, M. Zhu, and J. Sim. Detect-to-retrieve: Efficient regional aggregation for image search. In *CVPR*, 2019.

[49] Giorgios Tolias, Yannis Avrithis, and Hervé Jégou. To aggregate or not to aggregate: Selective match kernels for image search. In *ICCV*, 2013.

[50] Giorgos Tolias, Tomas Jenicek, and Ondřej Chum. Learning and aggregating deep local descriptors for instance-level recognition. In *ECCV*, 2020.

[51] Giorgos Tolias, Ronan Sicre, and Hervé Jégou. Particular object retrieval with integral max-pooling of CNN activations. In *ICLR*, 2016.

[52] Hugo Touvron, Matthieu Cord, Alaaeldin El-Nouby, Piotr Bojanowski, Armand Joulin, Gabriel Synnaeve, and Hervé Jégou. Augmenting convolutional networks with attention-based aggregation. In *arXiv preprint arXiv:2112.13692*, 2021.

[53] Qilong Wang, Banggu Wu, Pengfei Zhu, Peihua Li, Wangmeng Zuo, and Qinghua Hu. ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks. In *CVPR*, 2020.

[54] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local Neural Networks. In *CVPR*, 2018.

[55] Weinzaepfel, Philippe and Lucas, Thomas and Larlus, Diane and Kalantidis, Yannis. Learning Super-Features for Image Retrieval. In *ICLR*, 2022.

[56] Tobias Weyand, Andre Araujo, Bingyi Cao, and Jack Sim. Google Landmarks Dataset v2 - A Large-Scale Benchmark for Instance-Level Recognition and Retrieval. In *CVPR*, 2020.

[57] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. CBAM: Convolutional Block Attention Module. In *ECCV*, 2018.

[58] Hui Wu, Min Wang, Wengang Zhou, Yang Hu, and Houqiang Li. Learning token-based representation for image retrieval. 2022.

[59] Min Yang, Dongliang He, Miao Fan, Baorong Shi, Xuetong Xue, Fu Li, Errui Ding, and Jizhou Huang. Dolg: Single-stage image retrieval with deep orthogonal fusion of local and global features. In *ICCV*, 2021.

[60] Shuhei Yokoo, Kohei Ozaki, Edgar Simo-Serra, and Satoshi Iizuka. Two-stage discriminative re-ranking for large-scale landmark retrieval. In *CVPRW*, 2020.

[61] Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling vision transformers. In *CVPR*, 2022.

[62] Zhongyan Zhang, Lei Wang, Yang Wang, Luping Zhou, Jianjia Zhang, and Fang Chen. Dataset-driven unsupervised object discovery for region-based instance image retrieval. In *TPAMI*, 2023.