ILIAS: Instance-Level Image retrieval At Scale

Giorgos Kordopatis-Zilos Vladan Stojnić Anna Manko Pavel Šuma
Nikolaos-Antonios Ypsilantis Nikos Efthymiadis Zakaria Laskar
Jiří Matas Ondřej Chum Giorgos Tolias

VRG, FEE, Czech Technical University in Prague

Abstract

This work introduces ILIAS, a new test dataset for Instance-Level Image retrieval At Scale. It is designed to evaluate the ability of current and future foundation models and retrieval techniques to recognize particular objects. The key benefits over existing datasets include large scale, domain diversity, accurate ground truth, and a performance that is far from saturated. ILIAS includes query and positive images for 1,000 object instances, manually collected to capture challenging conditions and diverse domains. Large-scale retrieval is conducted against 100 million distractor images from YFCC100M. To avoid false negatives without extra annotation effort, we include only query objects confirmed to have emerged after 2014, i.e. *the compilation date of YFCC100M. An extensive benchmarking is performed with the following observations: i) models fine-tuned on specific domains, such as landmarks or products, excel in that domain but fail on ILIAS ii) learning a linear adaptation layer using multi-domain class supervision results in performance improvements, especially for vision-language models iii) local descriptors in retrieval re-ranking are still a key ingredient, especially in the presence of severe background clutter iv) the text-to-image performance of the vision-language foundation models is surprisingly close to the corresponding image-to-image case.*
website: https://vrg.fel.cvut.cz/ilias/

1. Introduction

The ability to recognize and differentiate every unique object instance in the physical world represents one of the ultimate goals for foundation representation models [5, 34, 41, 75]. This work aims to assess this capability through the lens of instance-level image retrieval at a very large scale. Instance-level image retrieval corresponds to searching for images of particular objects within large collections. All images of a particular object form their own instance-level class. This is an important information retrieval task due to its numerous real-world applications in robotics [30, 45], e-commerce [74, 76], and cultural heritage [16, 50], to name just a few. The task faces challenges because of the substantial variations among positive examples, such as illumination/viewpoint [21, 62] changes and background clutter [4, 31]. An additional difficulty is the high similarity among negatives, which is driven by the extremely fine granularity in the class definitions. It becomes even more challenging at a real-world scale, where searching through millions or even billions of images requires handling an open-world setup with countless unseen objects spanning diverse and complex domains.

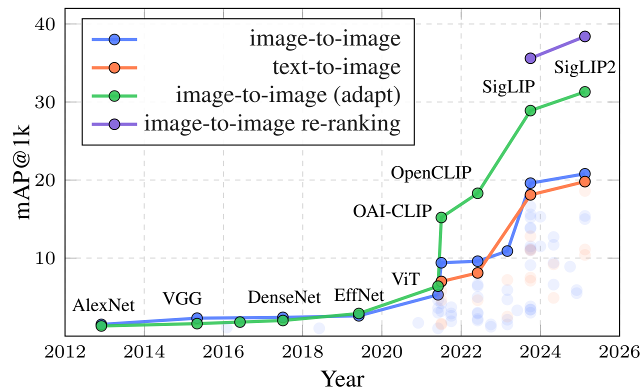

Figure 1. **Performance timeline on ILIAS**. Curves indicate best performance in chronological order for **image-to-image** and **text-to-image** retrieval, showing a significant boost with the release of foundation models. Representations are **linearly adapted** via multi-domain learning on UnED [72]. **Re-ranking with local descriptors** achieves the best results by a significant margin.

Benchmarking instance-level retrieval under real-world challenges is currently limited by the lack of suitable datasets. Constructing a dataset with instance-level class definitions necessitates huge development effort, reflected by the many shortcomings of existing datasets. Shortcomings exist in several key aspects, such as dataset size [66], domain diversity [39, 53], and ground-truth accuracy [67], which suffers from both false positives and false negatives. Popular datasets are typically limited to landmarks [39], and as dataset scale increases, ground-truth quality tends to decline [52, 67]. This is a consequence of automating the ground-truth creation process to facilitate scaling up. To address such limitations, we introduce the Instance-Level Image retrieval At Scale (ILIAS) evaluation dataset.

The creation of our dataset has two key elements. First, query and positive images are manually captured to ensure challenging variations, covering 1,000 objects across diverse domains. Second, to expand the dataset size without ground-truth errors or additional annotation effort, we leverage a key technique: distractor images, collected in 2014 from YFCC100M, are combined with query objects verified not to have publicly existed until after 2014. This distractor set includes 100 million images, two orders of magnitude larger than the largest existing dataset [39]. Notably, all images have a permissive license, allowing us to ensure long-term online availability to the full extent.

ILIAS includes both image and text queries. The latter is in the form of detailed descriptions of objects and their distinctive features. The dataset is designed to support future research in image-to-image and text-to-image retrieval for particular objects, and additionally serves as a large-scale benchmark for evaluating representations of foundation vision and language models (VLM) [41, 75]. To facilitate faster experimentation, we provide a mini, but challenging, version (5M) of the distractor set.

We perform an extensive evaluation comparison, including many foundation image-to-image and text-to-image models, and establish a comprehensive testbed that enables future comparisons. The provided evaluation includes retrieval with global image representation but also re-ranking techniques that use local representations [37, 43, 56] and query expansion [10, 40]. We observe the following:

- Performance of standard 10-year-old models, measured by mean Average Precision, is as low as 1.3%, while the best-performing model achieves 31.3%, as shown in Fig. 1. This points out the vast progress of representation models and the high challenging factors of ILIAS.
- VLMs are the top-performing models.
- Smaller (ViT-B) models trained/tested on large resolution (512/724) outperform larger models (ViT-L) trained/tested on small resolution (256/384).
- Using Universal Embedding Dataset (UnED) [72] to learn a linear adaptation layer on top of frozen models improves performance of most models, making it a candidate training set to couple with ILIAS. Notably, VLMs demonstrate the largest benefits, presumably because their training stage does not optimize image-to-image relations.
- In contrast to the current belief [48], local representation is a key ingredient, while global representation, despite being efficient and compact, performs much lower.
- The performance gap between image-to-image and text-to-image models is surprisingly small. Therefore, detailed text queries are a reasonable proxy in the absence of image queries, even at the instance level.

2. Related work

In this section, we review the related work in terms of existing datasets and benchmarks in the literature.

Datasets. Tab. 1 presents the datasets from the image retrieval literature related to ILIAS. The datasets can be compared based on five main axes: (i) *Class definition adopted*. Many datasets [20, 32, 33, 39, 67] adopt a strict definition very similar to ours, satisfying instance-level requirements. Others [2, 53, 66, 76] adopt a more relaxed definition, where some minor variations are permitted, *e.g*. color changes in objects of the same class. Even more relaxed are the fine-grained definitions [74], where the object of the very same type is considered related, *e.g*. same product with different variant. (ii) *Domain of the dataset*. Most datasets are tailored for a specific domain. Landmarks are among the most popular domains [2, 20, 33, 39, 67]. Other domains include products [32, 36, 76] and fashion [28, 53]. Some datasets cover multiple domains, either being standalone [66] or bundle of repurposed datasets [47, 72]. (iii) *Scale of database*. Most of the datasets are small-scale, counting a few thousand images [28, 66, 76]. Larger ones [39, 72, 74] expend slightly above a million. None satisfies large-scale requirements. (iv) *Noise in ground truth*. Most datasets consist of clean annotations, except for a few cases that contain inaccuracies, including false positives [67], *i.e*. images wrongly annotated as relevant, false negatives [32, 39], *i.e*. relevant images that have not been annotated as positives, or the possibility of false positives [39]. (v) *Availability*. Most datasets are publicly available with permissive licenses, with few exceptions of partial [2, 20] or no [33, 74] availability. To this end, no publicly available dataset fits the strict instance-level definition, contains objects from multiple domains, ensures error-free labeling and is large scale. This gap is filled with ILIAS satisfying all the aforementioned requirements.

Evaluation benchmarks. Benchmarking [51] tracks the progress in the field, which is even more necessary with the emergence of foundation models. Several benchmarks papers [24, 25, 77] exists in the instance-level retrieval literature, investigating the impact of learning scheme, post-processing, model ensembling, query expansion, and whitening. The most relevant benchmark to ILIAS is UnED [72] that combines existing datasets to create a union that evaluates models performance across various domains. Due to its wide variety, UnED serves as the training dataset for linear adaptation.

Regarding the evaluation of foundation models, the most common practice [14, 34, 57] is measuring classification performance on top of frozen models on ImageNet [12]. This is performed either with or without the training of a classifier via linear probing or k-NN search. Furthermore, models are usually evaluated on dense prediction tasks [34] and several multiple-downstream single-domain tasks [1].

datasets	year	objects	query	positives	database	gt	class def.	domain	bbox	online	license
UKB [32]	2006	2.5K	10K	10K	10K	FN	IL	product	✗	Fully	N/A
Holidays [20]	2008	500	500	991	1M	Clean	IL	landmark	✗	Partially	CC
Sculptures [2]	2011	10	70	3.1K	3.1K	Clean	Partial IL	sculpture	✗	Partially	Flickr TC
INSTRE [66]	2015	200	1250	27.3K	27.3K	Clean	Partial IL	multi	✓	Fully	Flickr TC
SOP [53]	2015	11.3K	60.5K	60.5K	60.5K	Clean	Partial IL	product	✗	Fully	MIT License
InShop [28]	2016	3.9K	14.2K	12.6K	12.6K	Clean	Partial IL	fashion	✓	Fully	N/A
R-Oxford [39]	2018	11	70	5K	1M	FN?	IL	landmark	✗	Fully	Flickr TC, CC
R-Paris [39]	2018	11	70	6.3K	1M	FN?	IL	landmark	✗	Fully	Flickr TC, CC
GLDv1 [33]	2018	30K	N/A	N/A	1.1M	Clean	IL	landmark	✗	Partially	Multiple
GLDv2 [67]	2020	318	1.1K	3.1K	762K	FP	IL	landmark	✗	Fully	CC/ Public-domain
Product1M [76]	2021	392	6.5K	40K	40K	Clean	Partial IL	product	✗	Partially	N/A
RP2K [36]	2021	1.2K	10.9K	10.9K	10.9K	Clean	Partial IL	product	✓	Fully	N/A
GPR1200 [47]	2021	1.2K	12K	12K	12K	Mix	IL+FG	multi	✗	Fully	Multiple
eProduct [74]	2021	206	10K	N/A	1.1M	N/A	FG	product	✗	No	N/A
FORB [69]	2023	N/A	13.9K	4.5K	49.8K	Clean	IL	planar	✗	Fully	Snap Inc.
UnED [72]	2023	21K	241K	244K	1.4M	Mix	IL+FG	multi	✗	Fully	Multiple
ILIAS	2025	1,000	1,232	4,715	100M	Clean	IL	multi	✓	Fully	CC

Table 1. **Comparison with other instance-level datasets.** Datasets are compared based on their size (object, query, positives, database), the accuracy of the ground truth (gt), type of class definition, domain, supplementary annotations (bbox) and accessibility (online, license). N/A: not available. FP/FN: false positives/negatives. FN?: possibility of false negatives. Mix: combination of clean and noisy datasets. IL: instance-level. FG: fine-grained. Partial IL: instance-level with subtle variations among same class objects. CC: Creative Commons.

For VLMs, zero-shot classification and retrieval serve as the primary benchmarks [41, 70, 75], utilizing class text labels. In this work, we provide similar evaluation protocols tailored for instance-level retrieval. One can test the raw model capabilities or adapt for the instance-level task via linear adaptation on UnED. Text-to-image retrieval is also facilitated for the evaluation of VLMs.

3. ILIAS dataset

3.1. Composition and collection

Instance-level class definition. Following an instance-level class definition [39, 71], we consider *all indistinguishable object instances of the real world to form their own class*. Nevertheless, we add a restriction to consider a pair of images as relevant to each other only if there is a view overlap. Other cases are explicitly not included in the dataset, contrasting the existing work [28, 53, 67]. Therefore, models should mostly rely on estimating the visual similarity and less on shortcuts through semantics.

Overview. ILIAS supports both image-to-image (i2i) and text-to-image (t2i) retrieval and follows the standard setup for retrieval datasets, consisting of two main parts: (i) *query* images and text, and (ii) *database* (db) images. The objective is to rank *positives* – db images relevant to the query – at the top ranks. The collected objects cover a wide range of categories and are not restricted to specific domains. An overview of some collected objects is provided in Fig. 2.

Queries and positives are created/collected by a group of *collectors* that are well-informed about the task objectives. In addition to positives, in the database, we include numerous *distractors* – irrelevant (negative) images to the queries – that make retrieval more challenging. Following previous work [39], adding a large, uncurated set of random images achieves this. The larger the set, the higher the chances of hard negatives – images with similar appearance or semantics to the queries. To this end, we select the YFCC100M [61] dataset to serve as the source of distractors due to its size and permissive license.

Selected objects. Ensuring that distractor images include no false negatives cannot be performed in a scalable or accurate way if one relies on human annotation or metadata. Instead, we take advantage of the fact that YFCC100M was crawled from Flickr in 2014. Hence, an object qualifies in ILIAS if it could not have appeared on Flickr before 2014. To verify this, we rely either on publicly available information, *e.g.* objects known to be created/manufactured after 2014, or on the collector's knowledge about the object not being publicly available. Additionally, we opt for objects with distinctive and unique features that set them apart from others within the same category. For example, we avoid recent smartphones that look like plain black screens or new objects with distinctive parts closely resembling older ones.

Queries and positives. Query images depict the instance on a clean or uniform background. When this is not feasible (*e.g.* buildings or statues), background blurring or cropping is applied. This is performed to avoid including background objects in the query that do not have corresponding positives in our ground truth information. Positives are images featuring the query object in challenging conditions, such as clutter, scale changes, occlusions, and partial views. Prior work [39] reveals that easy positives dominate performance metrics. Thus, we specifically opt for challenging cases that cannot be easily retrieved by the models. To avoid taking advantage of camera identification, most query and positive images are captured with at least two different camera models to introduce diversity. We also incorporate older camera models that are used in YFCC100M.

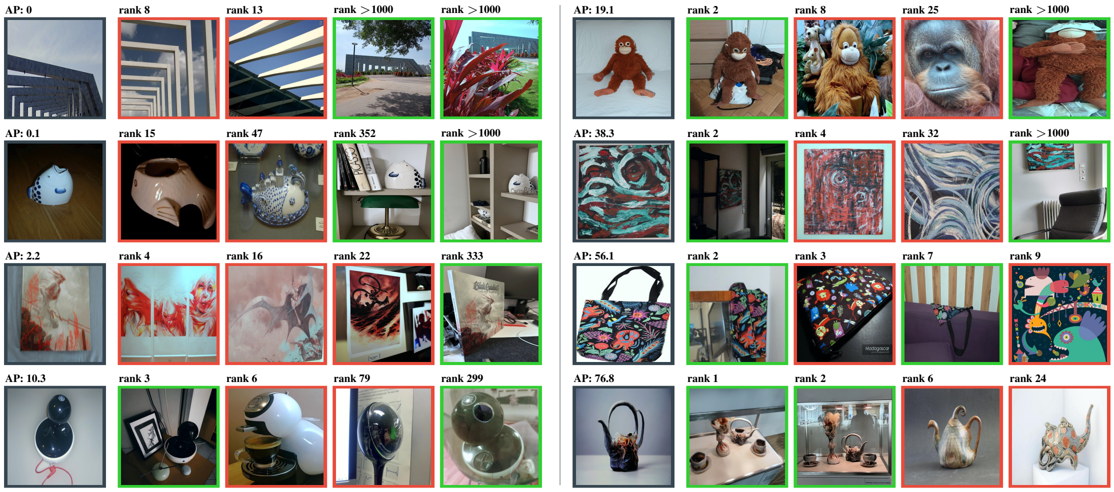

Figure 2. **Examples of query, positive and hard negatives within the distractor set.** Average Precision per query and rank of the negatives and positives is reported using SigLIP [75] model. **Gray:** queries. **Green:** positives. **Red:** distractors.

Each text query consists of a detailed and fine-grained textual description of an object. Descriptions are initially created by a large language model prompted to provide highly detailed depictions of the object shown in query images. Generated descriptions are manually edited to fix errors, insufficient descriptions, or nuances of the model.

Distractors. The YFCC100M dataset was chosen for the distractor set due to its large scale and diverse range of concepts. It consists of 100 million Flickr images, collected without specific filtering, aside from being shared under a permissive CC-BY license.

Bounding box annotation. We include supplementary bounding boxes that specify the precise location of objects in query and positive images. They provide statistics about the position and size of object areas, assist our analysis of the dataset challenges, and support future research in instance-level localization.

Evaluation metric. Retrieval performance is evaluated via mean Average Precision (mAP), a widely used metric in instance-level image retrieval [37, 38, 40]. Specifically, we adopt mAP@1k [67], which assesses the ranking of the top-1k nearest neighbors for each query, treating any positive not ranked among the top-1k as not retrieved. We estimate the area under the curve using rectangles and not trapezoids.

3.2. Statistics

Dataset size. The final ILIAS dataset includes 1,000 object instances captured in 5,947 images, of which 1,232 are queries and 4,715 are positives. Fig. 3a shows the distribution of positives per object. Also, 99,144,315 images from YFCC100M are downloaded. All images (queries, positives, distractors) are transferred through Flickr to ensure the same pre-processing.

Taxonomy. A hierarchical 3-level taxonomy is composed for ILIAS. All instances are assigned across one to three categories of different granularity levels. The taxonomy consists of 8 categories on the coarser level, 42 on the mid level, and 38 on the finer level. The categories are derived through manual labeling of the objects based on their semantic content. To form the coarser-level categories, we use domain definitions borrowed from prior work [28, 53, 67] to align with the literature, *i.e*. art, landmarks, products, fashion. We also define novel categories based on the objects that do not fit into any existing domain. The distribution of objects across categories is uneven, *e.g*. ranging from 168 and 162 for art and landmarks to 83 for products. Each mid- and finer-level category contains at least 4 instances. Note that taxonomy is given to provide statistics about the domains of objects and assist our analysis instead of being leveraged as ground truth. The distribution of taxonomy categories can be inferred by Fig. 5, and a detailed figure is provided in the supplementary material.

Bounding box analysis. A total number of 6,117 bounding boxes are annotated for both queries and positives. Note that positives may display multiple objects of near identical appearance to the query; in such cases, bounding boxes are drawn on all indistinguishable objects. There are 235 images with more than one bounding box. Based on the annotated bounding boxes, we compute the area covered in the image by the object instances to derive its relative scale. Fig. 3c shows the distribution of the scale ratio for queries and positives. Most objects in queries cover the largest area of the images; while in the vast majority of positives, the object covers a small area of less than half the image. It is a result of the severe scale changes and partial views. Moreover, we use the Segment Anything Model (SAM) [23, 42] to extract object segments from positives. The number of detected segments outside the query object's bounding boxes is computed. This indicates clutter from other items in the positives. Fig. 3b shows the segment number distribution, with most images containing multiple segments due to clutter.

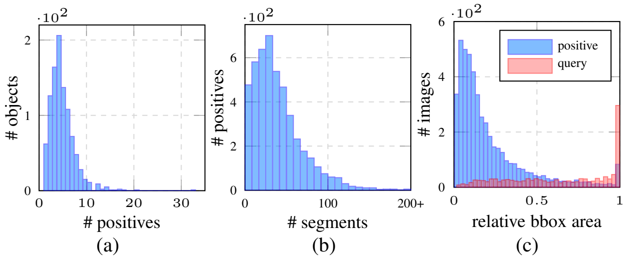

Figure 3. **ILIAS statistics**. (a) number of positives per object, (b) positive distribution by the SAM segments outside the bounding box, (c) image distribution by the relative bounding box area.

3.3. *mini*-ILIAS

We provide a small version of ILIAS, called *mini*-ILIAS, to facilitate quick experimentations. It consists of the query and positive images collected for ILIAS, and a subset of the YFCC100M distractors. Instead of randomly subsampling YFCC100M, we construct a challenging subset with the help of VLMs. We aim at selecting distractors displaying objects of similar categories as the query objects. We use the text category labels of the taxonomy as text queries. We also extend them with standard templates used for zero-shot recognition [41], which resulted in several thousands. T2i similarity between each text query and each distractor image is estimated. A similarity score for each distractor is derived based on its maximum similarity over the text queries. We ensemble the scores of 3 VLMs to rank images. The top-5M ranked distractors compose the final *mini*-ILIAS. Our experiments indicate that this subset is significantly more challenging than a random subset of the same size.

4. Benchmark methods

We describe the methods and foundation models we evaluate on ILIAS, which are grouped according to their type of representations used for retrieval in global (i2i) representations, re-ranking with global (i2i) representations, re-ranking with local (i2i) representations, and text-to-image. A detailed list of models, their performance, and implementation details are in the supplementary materials.

Image-to-image retrieval with global representations. Global representation methods use image encoders to map images to global descriptors and rank db images based on cosine similarity. We evaluate legacy and recent foundation models, varying in architecture, descriptor dimensionality, training scheme, training data, and input resolution. Foundation models [5] are the models trained with a training set on the scale of a hundred million. Particularly, 23 CNN [17, 26, 29, 49, 58, 59, 78] and 45 ViT [13, 14] models, trained with supervision [22, 54, 64, 68], self-supervision [7, 8, 18, 34], distillation [1, 46, 64], or visual-language alignment [9, 41, 57, 65, 70, 75] are benchmarked. Most of the non-foundation models are trained on ImageNet [12]. There are models trained on single specific domains [27, 35, 48], *i.e.* landmarks or products on GLDv2 or SOP. Universal models [1, 46, 72, 73] trained on multi-domains or multi-task schemes are included. The full list of models and results is provided in supplementary materials.

To mitigate the differences in training resolution, we use three widely-used resolutions, *i.e.* 384, 512, and 724 and resize images so that their larger dimension matches one of the three. The test resolution is defined to be one resolution above the one used for training, *e.g.* a network trained with 224 or 384 is tested with 384 or 512, respectively. The vast majority of models achieve best performance under this rule. Similar behavior is observed in the literature [55, 63].

Linear adaptation for i2i retrieval. Pre-trained foundation models, as well as legacy models, are trained to extract representations that are applicable to various tasks; not all encoded features are directly relevant to instance-level retrieval. To adapt the representation to the task at hand, we propose to train a single linear layer (projection) on top of frozen backbones. The recently introduced Universal Embeddings (UnED) dataset [72] is used for learning the linear adaptation. UnED contains images from 8 different domains with fine-grained and/or instance-level class annotation. In our experiments, the linear layer that converts the backbone output to a 512D descriptor is trained on a uniformly sampled subset of 1M images from UnED. The linear adaptation layer is trained with the UJCDS [72] method.

Text-to-image retrieval. Text-to-image retrieval is performed using Vision-Language Models (VLMs) trained to align the two modalities. Retrieval is performed based on cosine similarity between the text query and db image descriptors that are extracted using the textual and visual encoder, respectively. We evaluate 17 VLM models.

Re-ranking with global representations. Such methods rely on global descriptors for exhaustive search during the initial ranking, but also for a second refinement stage that issues a new query. We experiment with αQE [40], the adaptive variant of average Query Expansion [10]. After the initial ranking, the descriptors of the top-ranked images are aggregated with the query via weighted average pooling. The weights are derived from the similarity to the query in the power of α. We don't have a validation set; hence, we use a fixed value $\alpha = 1$.

Re-ranking with local representations. These re-ranking methods rely on global descriptors for exhaustive search during the initial ranking but estimate query-to-db image similarity based on local descriptors for a second refinement stage of the ranked list of images. We experiment with three methods: (i) Chamfer Similarity (CS) [3, 44] on the similarity matrix between local descriptors across the image pair. We use the asymmetric variant of CS with max over db descriptors and sum over query descriptors. (ii) Spatial Verification (SP) [6, 15, 37], a common re-ranking method where point correspondences are processed with a RANSAC-like

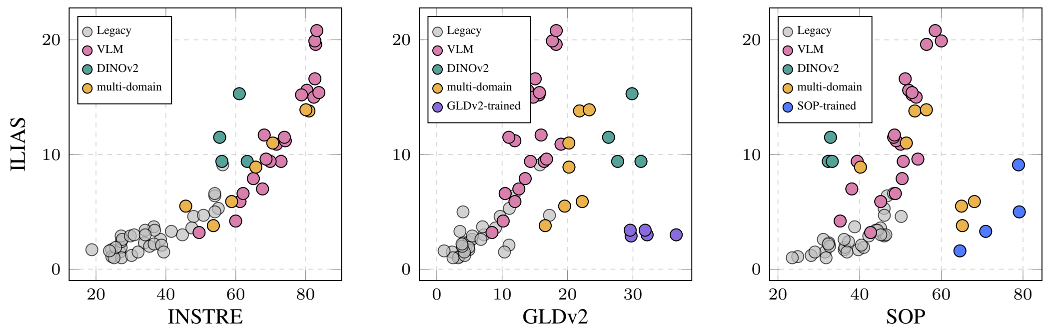

Figure 4. **Comparison with other instance-level retrieval datasets** via reporting mAP@1k. INSTRE: 27.3K db size, multi-domain. GLDv2: 762K db size, single-domain. SOP: 60.5K db size, single-domain. Different network types are color-coded. For GLDv2 and SOP, models fine-tuned on these domains with the corresponding training sets are highlighted. No linear adaptation is used.

process and the number of inliers is used for re-ranking. (iii) AMES [56], a recent transformer-based network to estimate the similarity between sets of local descriptors. Due to the scale of ILIAS database, we use only 100 binary local descriptors for each database image and 600 for the query image. Local descriptors are extracted using the base variant of DINOv2 with registers [11, 34] and selected based the local descriptor detector used in AMES [56]. Top-1k retrieved images are re-ranked.

5. Experiments

We evaluate all the above models and methods, extracting useful insights regarding the factors that boost retrieval performance. ILIAS is compared with other existing datasets for instance-level image retrieval. We analyze the performance of selected models[1] to break down the impact of different ILIAS attributes, such as domains, clutter, and scale. Unless stated otherwise, we use the large model variants with the largest resolution available, *e.g.* in our analysis we use SigLIP ViT-L trained with 384 resolution.

5.1. Comparison with other instance-level datasets

In Fig. 4, ILIAS is compared with other instance-level retrieval datasets based on evaluation of the same models. Linear adaptation is not applied as parts of GLDv2 and SOP are included in the UnED dataset. Only for the sake of this comparison, and for no other experiment in this work, we use models fine-tuned on specific domains (in-domain models), *i.e.* on the training sets of SOP and GLDv2. INSTRE, which is also multi-domain, shows a correlation to ILIAS, but its performance is saturated due to its small size. For single-domain datasets, in-domain models outperform others by a large margin, with few exceptions, *i.e.* DINOv2, which includes the trainset of GLDv2 in its training data. Several multi-domain models perform well on SOP since their training set is usually included in the training data. However, in-domain and multi-domain models face challenges on ILIAS, highlighting the diversity of our dataset.

[1]Although the very recent SigLIP2 is the best-performing model, we conduct most experiments with SigLIP.

5.2. Method comparison

Image-to-image retrieval with global representations. Tab. 2 presents the performance of global descriptor models on ILIAS. Selected models are presented to highlight useful comparisons, while many other models are included in the supplementary material. The main factors that improve performance are the size of the training set, training resolution, and model architecture, which aligns with the literature. The impact of dataset size is apparent in various model combinations, *e.g.* CLIP with `openai` and `laion2b`. This is also pronounced by the dominance of foundation models. Training with large resolution brings significant gains and consistently improves mAP@1k. In some cases of SigLIP, smaller models trained with large resolutions outperform larger ones trained with small resolutions. For the models of the same resolution, it is a common trend for larger model variants to bring corresponding performance gains. In general, VLMs perform the best. From non-VLMs, only DINOv2 and Unicom achieve competitive performance. Masked Image Modeling (MIM) and supervised models are not performing well. Our linear adaptation scheme is very effective, improving most models. The boost is more pronounced in the case of VLMs. A possible explanation for such improvements is that image-to-image relations are not optimized during the training of VLMs.

Text-to-image retrieval. Following results in Tab. 2, similar conclusions are derived for the t2i case. Retrieval performance improves with the scaling of the training data. The larger model achieves significantly better results, *i.e.* compare the base with large variants. Finally, it is noteworthy that the best performance achieved by SigLIP2 is very close to the i2i performance when no adaptation is used. Note that t2i includes 1k text queries in total, with one query per object, while i2i 1,232 image queries.

Evaluation of *mini*-ILIAS selection. Tab. 3 shows performance on *mini*-ILIAS for five models with linear adaptation. The selected subset is significantly more challenging than a random selection of 5M images. More precisely, a set of approximately ∼26M random images matches the performance of *mini*-ILIAS.

							image-to-image			text-to-image	
model	**arch**	**train**	**dataset**	**data size**	**train res**	**test res**	**5M**†	**100M**†	**100M**	**100M**	**5M**
ResNet50 [17]	R50	sup	in1k	1M	224	384	2.5	1.8	1.7	-	-
DINO [8]	R50	ssl	in1k	1M	224	384	4.1	2.9	2.9	-	-
ConvNext [29]	CN-L	sup	in1k	1M	288	384	4.2	2.9	2.2	-	-
OAI-CLIP [17]	R50	vla	openai	400M	224	384	8.5	6.0	3.2	1.5	2.3
OpenCLIP [19, 29]	CN-B	vla	laion2b	2B	256	384	18.1	14.0	7.9	4.6	7.0
OpenCLIP [19, 29]	CN-L	vla	laion2b	2B	320	512	22.9	18.3	9.6	8.1	11.5
ViT [13, 54]	ViT-B	sup	in1k	1M	224	384	1.9	1.3	1.0	-	-
EVA-MIM [14]	ViT-B	ssl	in22k	142M	224	384	4.7	3.2	2.1	-	-
ViT [13, 54]	ViT-B	sup	in21k	14M	224	384	6.2	4.4	3.0	-	-
DINO [8]	ViT-B	ssl	in1k	1M	224	384	6.6	4.8	3.7	-	-
UDON-CLIP [73]	ViT-B	sup	uned	2.8M	224	384	9.2	6.7	5.9	-	-
OAI-CLIP [41]	ViT-B	vla	openai	400M	224	384	10.7	7.9	4.2	1.6	2.7
EVA-CLIP [57]	ViT-B	vla	merged2b	2B	224	384	11.7	8.7	5.9	2.5	4.4
MetaCLIP [70]	ViT-B	vla	2pt5b	2.5B	224	384	12.7	9.4	6.6	4.9	7.6
DINOv2 [34]	ViT-B	ssl	lvd142m	142M	518	724	15.0	12.1	11.5	-	-
SigLIP [75]	ViT-B	vla	webli	10B	256	384	20.6	16.7	11.5	7.5	10.3
SigLIP [75]	ViT-B	vla	webli	10B	384	512	26.2	21.5	15.6	11.0	14.4
SigLIP [75]	ViT-B	vla	webli	10B	512	724	27.5	23.0	16.6	11.1	14.6
SigLIP2 [65]	ViT-B	vla	webli	10B	512	724	28.6	23.5	15.4	10.4	14.6
EVA-MIM [14]	ViT-L	ssl	in22k	14M	224	384	3.9	2.7	1.5	-	-
ViT [13, 54]	ViT-L	sup	in21k	14M	224	384	7.3	5.3	4.6	-	-
EVA-MIM [14]	ViT-L	ssl	merged38m	38B	224	384	8.8	6.1	4.7	-	-
OAI-CLIP [41]	ViT-L	vla	openai	400M	224	384	15.8	11.9	7.0	4.6	6.7
OpenCLIP [9, 19]	ViT-L	vla	laion2b	2B	224	384	17.5	13.7	9.4	7.0	9.4
Unicom [1]	ViT-L	dist	laion400m	400M	336	512	18.6	14.6	13.9	-	-
OAI-CLIP [41]	ViT-L	vla	openai	400M	336	512	19.9	15.2	9.4	5.8	8.4
DINOv2 [34]	ViT-L	ssl	lvd142m	142M	518	724	18.8	15.3	15.3	-	-
EVA-CLIP [57]	ViT-L	vla	merged2b	2B	336	512	20.9	16.0	10.9	7.2	10.6
MetaCLIP [70]	ViT-L	vla	2pt5b	2.5B	224	384	21.7	16.9	11.7	9.2	13.1
SigLIP [75]	ViT-L	vla	webli	10B	256	384	26.3	21.8	15.2	12.8	16.4
SigLIP [75]	ViT-L	vla	webli	10B	384	512	34.3	28.9	19.6	18.1	22.2
SigLIP2 [65]	ViT-L	vla	webli	10B	512	724	37.3	31.3	20.8	19.8	24.7

Table 2. **Performance comparison using mAP@1k on ILIAS and *mini*-ILIAS for global representation models for i2i and t2i.** Comparison of model architecture (arch), training scheme (train), training data, and train/test resolution. † indicates results with the linear adaptation. 5M and 100M correspond to the mini and full versions of the dataset, respectively. sup, ssl, dist, vla: supervised learning, self-supervised learning, distillation and vision-language alignment. R50, CN: ResNet50 and ConvNext.

model	**100M**	**5M-mini**	**5M-rand**
DINOv2† [34]	15.3	18.8	22.7±0.2
EVA-CLIP† [57]	16.0	20.9	28.8±0.2
MetaCLIP† [70]	16.9	21.7	29.2±0.1
OpenCLIP† [19, 29]	18.3	22.9	30.9±0.2
SigLIP† [75]	28.9	34.3	41.8±0.1

Table 3. **A challenging distractor subset for *mini*-ILIAS.** mAP@1k evaluated for different distractor sets, 100M: the full dataset, 5M-mini: *mini*-ILIAS subset, 5M-rand: random subset. We report the mean and std of 3 randomly sampled subsets. † indicates results with the linear adaptation.

reranking	**SigLIP** [75]		**SigLIP**† [75]	
	mAP@1k	**oracle**	**mAP@1k**	**oracle**
global	19.6	48.7	28.9	56.0
αQE1 [10, 40]	22.1	44.7	33.7	56.9
αQE2 [10, 40]	20.4	40.8	31.5	54.4
αQE5 [10, 40]	14.3	34.9	23.5	49.3
CS [44]	22.9	48.7	32.5	56.0
SP [37]	21.8	48.7	30.5	56.0
AMES [56]	26.4	48.7	35.6	56.0

Table 4. **Performance comparison for re-ranking methods.** Oracle represents the performance of perfect re-ranking at the top-1k images. Top: query expansion with global descriptors. Bottom: re-ranking with local descriptors. †: results with linear adaptation.

Retrieval with re-ranking. Tab. 4 shows the performance of re-ranking methods applied on top of SigLIP with and without linear adaptation on ILIAS. Complementary to mAP@1k, an oracle-based top-1k re-ranking metric is reported as the upper bound of a re-ranking method that processes the top 1k images. Local similarity estimated by a learned model proves to be very effective for re-ranking. Nevertheless, the oracle re-ranking performance indicates that there is a lot more space for improvements. Re-ranking with QE is useful when the number of aggregated neighbors is low and drops below the baseline when the number of neighbors is increased. Notably, global re-ranking affects and, interestingly, decreases oracle performance since the whole db is re-ranked; while local re-ranking does not affect it since it is performed only on a shortlist of images.

5.3. Analysis

Performance per domain. Fig. 5 shows the performance per taxonomy categories. The taxonomy annotations allow a fine-grained view of the results, which can possibly allow us to capture imbalanced improvements in future work. For example, DINOv2, despite being overall inferior to SigLIP, is outperforming it in categories like architecture and sculptures or is quite similar in categories like public art and paper art. This is possibly attributed to the curation and com-

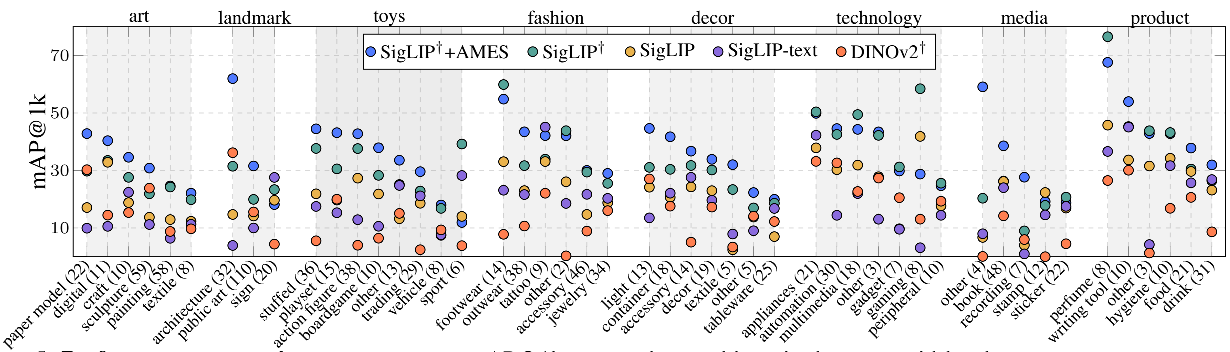

Figure 5. **Performance comparison per category.** mAP@1k averaged over objects in the same mid-level taxonomy category, organized by their primary-level category size, with sorting within each group by SigLIP†+AMES performance. Comparison between SigLIP with and without adaptation, SigLIP combined with AMES reranking, SigLIP t2i, and DINOv2. † indicates results with the linear adaptation.

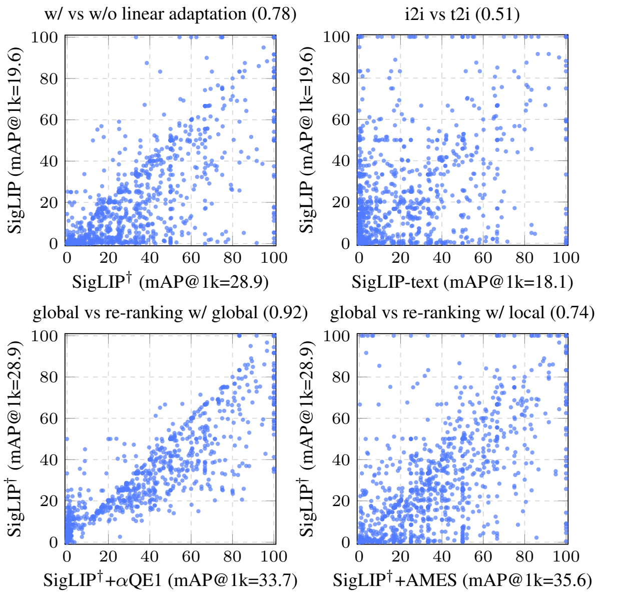

Figure 6. **Performance comparison reporting AP per query for different approaches with SigLIP.** Pearson correlation reported in parenthesis. † indicates results with the linear adaptation.

position of the DINOv2 training set, which includes artwork and landmark datasets. Also, some categories are hurt by re-ranking with AMES, with some demonstrating big drops, *i.e*. sport, gaming, perfume. These categories deviate significantly from the domain AMES is trained, *i.e*. landmarks, which could potentially justify such drops.

Per query comparisons. Fig. 6 shows the AP per query for various methods. Linear adaptation boosts most queries, *i.e*. performance drop only for 192 queries. Image- and text-based retrieval are not strongly correlated despite performing similarly, which is good evidence [60] for the effectiveness of model ensembles. Indeed, ensembling i2i and t2i by averaging similarities brings +6.1 improvement over i2i retrieval. Query expansion improves the queries with at least some positives at top positions, *i.e*. AP greater than 20. However, it harms many low-performing queries by aggregating descriptors irrelevant to the query. AMES improves the majority of the queries; however, many are harmed, indicating that there is plenty of room for improvement.

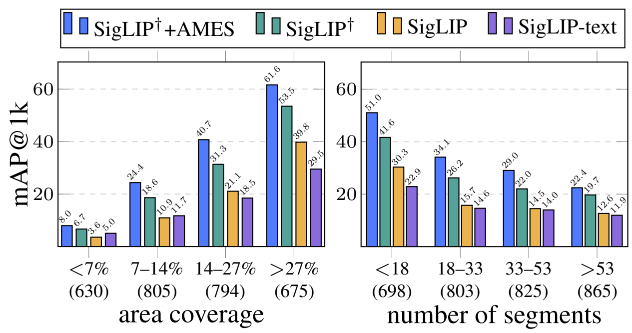

Figure 7. **Performance evaluation (mAP@1k) across different amounts of object area coverage and background clutter.** Positives across all queries are jointly ranked based on coverage or clutter and split into 4 equal size groups. Queries with no positive in the corresponding group are discarded. No. of queries per group is in parentheses. † indicates results with the linear adaptation.

Impact of clutter and scale in positives. To quantify the impact of background clutter and scale changes, Fig. 7 presents the performance for different groups of positives. Dealing with small objects and multi-object scenes form major weaknesses of existing models. Notably, t2i beats i2i without adaptation in small-scale groups.

6. Conclusions

We introduce ILIAS and conduct an extensive evaluation of current foundational models and retrieval methods, highlighting that instance-level retrieval remains an unsolved problem. Our results indicate that off-the-shelf application of foundational models leaves considerable room for improvement, particularly in handling small objects and complex backgrounds. While specialized retrieval methods leveraging local descriptors are effective in these cases, their high memory and computational costs become impractical at the scale of ILIAS or beyond. ILIAS is designed to become a standard benchmark for evaluating foundational representation models and retrieval methods, accommodating both global and local representations, and advancing the field of instance-level recognition.

Acknowledgments

Funding acknowledgement. This work was supported by the Junior Star GACR (grant no. GM 21-28830M), Horizon MSCA-PF (grant no. 101154126), Programme Johannes Amos Comenius (grant no. CZ.02.01.01/00/22_010/0003405), CTU in Prague (grant no. SGS23/173/OHK3/3T/13), the CTU institutional support (Future fund). JM was supported by Ministry of the Interior of the Czech Republic (project no. VJ02010041). We acknowledge VSB – Technical University of Ostrava, IT4Innovations National Supercomputing Center, Czech Republic, for awarding this project access to the LUMI supercomputer, owned by the EuroHPC Joint Undertaking, hosted by CSC (Finland) and the LUMI consortium through the Ministry of Education, Youth and Sports of the Czech Republic through the e-INFRA CZ (grant ID: 90254).

Contributors acknowledgments. We want to thank Larysa Ivashechkina for her work on the annotation of object bounding boxes and masks and the initial filtering of text queries. We also want to thank all contributors for the collection of ILIAS dataset. Additionally to all authors, the list of external contributor in alphabetical order: Aggeliki Tserota, Anna Nisyraiou, Celeste Abreu, Charalambos Tzamos, Christina Tserota, Dimitris Karageorgiou, Dmytro Mishkin, Eleni Karantali, Eleni Papadopoulou, Eva Tsiliakou, Kelly Kordopati-Zilou, Markos Zampoglou, Noa Garcia, Panagiotis Tassis, Paraskevas Kordopatis, Pavlos Alexantonakis, Ruslan Rozumnyi, Snežana Ćurguz, Tereza Nejedlá, Tomáš Jelínek, Yannis Kalantidis, Vasilis Alexiadis, Yankun Wu.

References

[1] Xiang An, Jiankang Deng, Kaicheng Yang, Jaiwei Li, Ziyong Feng, Jia Guo, Jing Yang, and Tongliang Liu. Unicom: Universal and compact representation learning for image retrieval. In *ICLR*, 2023. 2, 5, 7

[2] Relja Arandjelović and Andrew Zisserman. Smooth object retrieval using a bag of boundaries. In *ICCV*, 2011. 2, 3

[3] Harry G Barrow, Jay M Tenenbaum, Robert C Bolles, and Helen C Wolf. Parametric correspondence and chamfer matching: Two new techniques for image matching. In *IJCAI*, 1977. 5

[4] Wissam Bejjani, Wisdom C Agboh, Mehmet R Dogar, and Matteo Leonetti. Occlusion-aware search for object retrieval in clutter. In *IROS*, 2021. 1

[5] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. In *arXiv*, 2021. 1, 5

[6] Bingyi Cao, André Araujo, and Jack Sim. Unifying deep local and global features for image search. In *ECCV*, 2020. 5

[7] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *NeurIPS*, 2020. 5

[8] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, 2021. 5, 7

[9] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *CVPR*, 2023. 5, 7

[10] Ondrej Chum, James Philbin, Josef Sivic, Michael Isard, and Andrew Zisserman. Total recall: Automatic query expansion with a generative feature model for object retrieval. In *ICCV*, 2007. 2, 5, 7

[11] Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. In *ICLR*, 2024. 6

[12] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 2, 5

[13] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICML*, 2021. 5, 7

[14] Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. EVA: Exploring the limits of masked visual representation learning at scale. In *CVPR*, 2023. 2, 5, 7

[15] Martin A. Fischler and Robert C. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 1981. 5

[16] Cristopher Flagg and Ophir Frieder. Reconstruction of artifacts from digital image repositories. *JOCCH*, 2022. 1

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 5, 7

[18] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020. 5

[19] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, 2021. 7

[20] Hervé Jégou, Matthijs Douze, and Cordelia Schmid. Hamming embedding and weak geometric consistency for large scale image search. In *ECCV*, 2008. 2, 3

[21] Tomas Jenicek and Ondrej Chum. No fear of the dark: Image retrieval under varying illumination conditions. In *CVPR*, 2019. 1

[22] Sungyeon Kim, Boseung Jeong, and Suha Kwak. HIER: Metric learning beyond class labels via hierarchical regularization. In *CVPR*, 2023. 5

[23] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *CVPR*, 2023. 4

[24] Byungsoo Ko, Minchul Shin, Geonmo Gu, HeeJae Jun, Tae Kwan Lee, and Youngjoon Kim. A benchmark on tricks for large-scale image retrieval. In *arXiv*, 2019. 2

[25] Tarun Krishna, Kevin McGuinness, and Noel O'Connor. Evaluating contrastive models for instance-based image retrieval. In *ICMR*, 2021. 2

[26] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *NeurIPS*, 2012. 5

[27] Seongwon Lee, Hongje Seong, Suhyeon Lee, and Euntai Kim. Correlation verification for image retrieval. In *CVPR*, 2022. 5

[28] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *CVPR*, 2016. 2, 3, 4

[29] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. *CVPR*, 2022. 5, 7

[30] Shiyang Lu, Haonan Chang, Eric Pu Jing, Abdeslam Boularias, and Kostas Bekris. OVIR-3D: Open-vocabulary 3d instance retrieval without training on 3d data. In *CoRL*, 2023. 1

[31] Claudio Michaelis, Matthias Bethge, and Alexander Ecker. One-shot segmentation in clutter. In *ICML*, 2018. 1

[32] David Nister and Henrik Stewenius. Scalable recognition with a vocabulary tree. In *CVPR*, 2006. 2, 3

[33] Hyeonwoo Noh, Andre Araujo, Jack Sim, Tobias Weyand, and Bohyung Han. Large-scale image retrieval with attentive deep local features. In *ICCV*, 2017. 2, 3

[34] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *TMLR*, 2024. 1, 2, 5, 6, 7

[35] Yash Patel, Giorgos Tolias, and Jiří Matas. Recall@k surrogate loss with large batches and similarity mixup. In *CVPR*, 2022. 5

[36] Jingtian Peng, Chang Xiao, and Yifan Li. RP2K: A large-scale retail product dataset for fine-grained image classification. In *arXiv*, 2020. 2, 3

[37] James Philbin, Ondřej Chum, Michael Isard, Josef Sivic, and Andrew Zisserman. Object retrieval with large vocabularies and fast spatial matching. In *CVPR*, 2007. 2, 4, 5, 7

[38] James Philbin, Ondřej Chum, Michael Isard, Josef Sivic, and Andrew Zisserman. Lost in quantization: Improving particular object retrieval in large scale image databases. In *CVPR*, 2008. 4

[39] Filip Radenović, Ahmet Iscen, Giorgos Tolias, Yannis Avrithis, and Ondřej Chum. Revisiting oxford and paris: Large-scale image retrieval benchmarking. In *CVPR*, 2018. 1, 2, 3

[40] Filip Radenović, Giorgos Tolias, and Ondřej Chum. Fine-tuning cnn image retrieval with no human annotation. *PAMI*, 2019. 2, 4, 5, 7

[41] 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. 1, 2, 3, 5, 7

[42] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. SAM 2: Segment anything in images and videos. In *ICLR*, 2025. 4

[43] Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. CNN features off-the-shelf: An astounding baseline for recognition. In *CVPRW*, 2014. 2

[44] Ali S Razavian, Josephine Sullivan, Stefan Carlsson, and Atsuto Maki. Visual instance retrieval with deep convolutional networks. *ITE Trans. on Media Technology and Applications*, 2016. 5, 7

[45] Taichi Sakaguchi, Akira Taniguchi, Yoshinobu Hagiwara, Lotfi El Hafi, Shoichi Hasegawa, and Tadahiro Taniguchi. Object instance retrieval in assistive robotics: Leveraging fine-tuned simsiam with multi-view images based on 3d semantic map. In *IROS*, 2024. 1

[46] Mert Bulent Sariyildiz, Philippe Weinzaepfel, Thomas Lucas, Diane Larlus, and Yannis Kalantidis. UNIC: Universal classification models via multi-teacher distillation. In *ECCV*, 2024. 5

[47] Konstantin Schall, Kai Uwe Barthel, Nico Hezel, and Klaus Jung. GPR1200: a benchmark for general-purpose content-based image retrieval. In *MMM*, 2022. 2, 3

[48] Shihao Shao, Kaifeng Chen, Arjun Karpur, Qinghua Cui, André Araujo, and Bingyi Cao. Global features are all you need for image retrieval and reranking. In *ICCV*, 2023. 2, 5

[49] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015. 5

[50] Ivan Sipiran, Patrick Lazo, Cristian Lopez, Milagritos Jimenez, Nihar Bagewadi, Benjamin Bustos, Hieu Dao, Shankar Gangisetty, Martin Hanik, Ngoc-Phuong Ho-Thi, et al. SHREC 2021: Retrieval of cultural heritage objects. *Computers & Graphics*, 2021. 1

[51] Arnold WM Smeulders, Marcel Worring, Simone Santini, Amarnath Gupta, and Ramesh Jain. Content-based image retrieval at the end of the early years. *PAMI*, 2000. 2

[52] Chull Hwan Song, Jooyoung Yoon, Taebaek Hwang, Shunghyun Choi, Yeong Hyeon Gu, and Yannis Avrithis. On train-test class overlap and detection for image retrieval. In *CVPR*, 2024. 1

[53] Hyun Oh Song, Yu Xiang, Stefanie Jegelka, and Silvio Savarese. Deep metric learning via lifted structured feature embedding. In *CVPR*, 2016. 1, 2, 3, 4

[54] Andreas Steiner, Alexander Kolesnikov, Xiaohua Zhai, Ross Wightman, Jakob Uszkoreit, and Lucas Beyer. How to train your vit? data, augmentation, and regularization in vision transformers. *TMLR*, 2021. 5, 7

[55] Pavel Suma and Giorgos Tolias. Large-to-small image resolution asymmetry in deep metric learning. In *WACV*, 2023. 5

[56] Pavel Suma, Giorgos Kordopatis-Zilos, Ahmet Iscen, and Giorgos Tolias. AMES: Asymmetric and memory-efficient similarity estimation for instance-level retrieval. In *ECCV*, 2024. 2, 6, 7

[57] Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. EVA-CLIP: Improved training techniques for clip at scale. In *arXiv*, 2023. 2, 5, 7

[58] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *CVPR*, 2015. 5

[59] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *ICML*, 2019. 5

[60] Ryan Theisen, Hyunsuk Kim, Yaoqing Yang, Liam Hodgkinson, and Michael W Mahoney. When are ensembles really effective? In *NeurIPS*, 2024. 8

[61] Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. YFCC100M: The new data in multimedia research. *Communications of the ACM*, 2016. 3

[62] Akihiko Torii, Relja Arandjelovic, Josef Sivic, Masatoshi Okutomi, and Tomas Pajdla. 24/7 place recognition by view synthesis. In *CVPR*, 2015. 1

[63] Hugo Touvron, Andrea Vedaldi, Matthijs Douze, and Herve Jegou. Fixing the train-test resolution discrepancy. In *NeurIPS*, 2019. 5

[64] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *ICML*, 2021. 5

[65] Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. SigLIP 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. In *arXiv*, 2025. 5, 7

[66] Shuang Wang and Shuqiang Jiang. INSTRE: A new benchmark for instance-level object retrieval and recognition. *TOMM*, 2015. 1, 2, 3

[67] Tobias Weyand, André Araujo, Bingyi Cao, and Jack Sim. Google Landmarks Dataset v2 – A large-scale benchmark for instance-level recognition and retrieval. In *CVPR*, 2020. 1, 2, 3, 4

[68] Ross Wightman, Hugo Touvron, and Hervé Jégou. Resnet strikes back: An improved training procedure in timm. In *NeurIPSW*, 2021. 5

[69] Pengxiang Wu, Siman Wang, Kevin Dela Rosa, and Derek Hu. FORB: A flat object retrieval benchmark for universal image embedding. In *NeurIPS*, 2024. 3

[70] Hu Xu, Saining Xie, Xiaoqing Ellen Tan, Po-Yao Huang, Russell Howes, Vasu Sharma, Shang-Wen Li, Gargi Ghosh, Luke Zettlemoyer, and Christoph Feichtenhofer. Demystifying clip data. In *ICLR*, 2024. 3, 5, 7

[71] Nikolaos-Antonios Ypsilantis, Noa Garcia, Guangxing Han, Sarah Ibrahimi, Nanne Van Noord, and Giorgos Tolias. The MET dataset: Instance-level recognition for artworks. In *NeurIPS*, 2021. 3

[72] Nikolaos-Antonios Ypsilantis, Kaifeng Chen, Bingyi Cao, Mário Lipovský, Pelin Dogan-Schönberger, Grzegorz Makosa, Boris Bluntschli, Mojtaba Seyedhosseini, Ondřej Chum, and André Araujo. Towards universal image embeddings: A large-scale dataset and challenge for generic image representations. In *ICCV*, 2023. 1, 2, 3, 5

[73] Nikolaos-Antonios Ypsilantis, Kaifeng Chen, André Araujo, and Ondřej Chum. UDON: Universal dynamic online distillation for generic image representations. In *NeurIPS*, 2024. 5, 7

[74] Jiangbo Yuan, An-Ti Chiang, Wen Tang, and Antonio Haro. eProduct: A million-scale visual search benchmark to address product recognition challenges. In *CVPRW*, 2021. 1, 2, 3

[75] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *ICCV*, 2023. 1, 2, 3, 4, 5, 7

[76] Xunlin Zhan, Yangxin Wu, Xiao Dong, Yunchao Wei, Minlong Lu, Yichi Zhang, Hang Xu, and Xiaodan Liang. Product1m: Towards weakly supervised instance-level product retrieval via cross-modal pretraining. In *ICCV*, 2021. 1, 2, 3

[77] Liang Zheng, Yi Yang, and Qi Tian. Sift meets cnn: A decade survey of instance retrieval. *PAMI*, 2017. 2

[78] Barret Zoph, Vijay Vasudevan, Jonathon Shlens, and Quoc V Le. Learning transferable architectures for scalable image recognition. In *CVPR*, 2018. 5