Self-Supervised Incremental Learning of Object Representations from Arbitrary Image Sets

George Leotescu[1*] Alin-Ionut Popa[1*] Diana Grigore[1] Daniel Voinea[1] Pietro Perona[1,2]
[1]Amazon Inc.
{leoteg,popaaln,digrigor,dvoinea}@amazon.com
[2]California Institute of Technology
perona@caltech.edu

Abstract

*Computing a comprehensive and robust visual representation of an arbitrary object or category of objects is a complex problem. The difficulty increases when one starts from a set of uncalibrated images obtained from different sources. We propose a self-supervised approach, **M**ulti-**I**mage **L**atent **E**mbedding (**MILE**), which computes a single representation from such an image set. MILE operates incrementally, considering one image at a time, while processing various depictions of the class through a shared gated cross-attention mechanism. The representations are progressively refined as more available images are incorporated, without requiring additional training. Our experiments on Amazon Berkeley Objects (ABO) and iNaturalist demonstrate the effectiveness in two tasks: object or category-specific image retrieval and unsupervised context-conditioned object segmentation. Moreover, the proposed multi-image input setup opens new frontiers for the task of object retrieval. Our studies indicate that our models can capture descriptive representations that better encapsulate the intrinsic characteristics of the objects. Our code is available at* https://github.com/amazon-science/mile.

1. Introduction

Traditional object retrieval applications [2,61] and image search engines [14, 22] operate by capturing a single representation of the search subject and matching it against the model's internal category representation. Although this process is generally straightforward, the system may face difficulties when confronted with irregular structures, unconventional angles or viewpoints, and intricate scenes. Moreover, the appearance of objects can undergo transformations over time. While robustness and generalizability are essential qualities of an embedding, it is imperative to have a system capable of adapting its internal representation to accommodate natural shifts or variations in the distribution of object appearances. Most works that appeared as a solution to this ambiguous visual disentanglement task propose using large amounts of data to obtain such qualities. However, continuously updating the representations implies additional training along with the difficulty of finding supervised data, especially for low-resource categories.

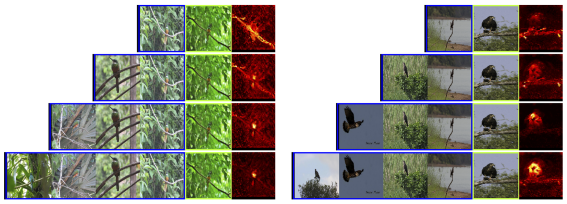

Figure 1. The MILE framework demonstrates a progressive capability to perceptually understand the common object within the input sequence. As the number of available images increases (blue rectangle), the framework is able to assimilate complementary visual attributes w.r.t. the specific image of interest (green rectangle). This, in turn, enables the discovery and representation of the common object across the diverse input images.

In this paper we propose a framework capable of producing holistic embeddings of visual concepts from multi-image inputs by leveraging the self-supervised learning (SSL) paradigm [4, 6, 17]. The key novelties we introduce are: *(a)* a single structured multi-image embedding achieved via a sequential latent space modelling flow operating on sets of images of undefined lengths, *(b)* an implicit mechanism to incrementally update the object representation (see Figure 1) during inference time as more images depicting the object are provided, *(c)* an SSL training framework specifically adapted to handle multi-image inputs and *(d)* a novel formulation of the object retrieval task based on image sets. We leverage the visual information present in multiple images by learning the embedding of objects or categories of objects through an implicit disentanglement of their structural characteristics. One of the key challenges we encountered was the non-trivial extension of the SSL paradigm to the multi-image setup, as previous approaches

[*]Authors contributed equally.

focused around the single-image scenario. Moreover, it is worth mentioning that rather than aiming to advance the SSL paradigm itself, our focus was on utilizing the capabilities of the SSL approach and integrating it within our proposed computational framework.

An overview for the proposed processing framework is illustrated in Figure 2, while the SSL flow can be observed in Figure 3. The incremental embedding refinement capability is achieved dynamically during inference time as more visual information is fed to the model, without the necessity to perform any additional training on larger sequence sizes. Another important remark regarding our work is the ability to overcome the implicit geometrical constraints imposed by the rigid nature of the analyzed object when processing it from a multi-image setup. Our proposed framework is validated on the challenging ABO [7] and iNaturalist [52] datasets, outperforming competitive self and weakly supervised frameworks on the task of object retrieval. This was achieved using the ViT family [9] backbones, however any image encoder can be employed. We specifically emphasize that our proposed solution is multi-image and not necessarily multi-view, as it can incorporate sources of information depicting the same object in various visual distributions. The multi-image setup is inherently more generic compared to stereo or multi-view, since the only requirement is that the input samples are related to the object of interest, without any constraints on their correlation or viewpoint consistency.

Utilising the constraint of having as input an image set depicting a single object, one might argue about the usability of contrastive learning approaches [27, 31, 39, 51, 59]. However, this implies a weakly supervised learning setup, necessitating at least 2 pairs of matching and non-matching items during the learning process. We acknowledge that we could potentially leverage objects with a significant number of associated images to obtain multiple image sets per object. However, this would significantly prune the overall training set, which we wanted to avoid. Our goal was to develop an approach that could effectively learn from a broader and more diverse training set.

2. Related Work

In recent years, the rapid growth of unlabeled visual data has fueled a surge of interest in self-supervised learning (SSL) techniques. This is particularly relevant for tasks like image and object retrieval [16, 24, 35, 54, 58], where constructing robust, pattern-rich representations is crucial - especially for industry-scale applications.

Within the SSL landscape, researchers have explored various approaches [3, 11, 28, 29, 32, 33, 49, 50, 53, 56, 57] that leverage multi-view image data to learn more comprehensive visual representations. A common theme among these methods is the use of contrastive learning, which aims to maximize the similarity between embeddings of the same object seen from different viewpoints, while simultaneously increasing the distance between embeddings of distinct objects. One such approach, introduced in [30] , is closely related to the work presented in this paper. The authors' technique uses contrastive learning to capture cross-view similarities of depicted objects, while also increasing the separation between embeddings of different objects. However, a key limitation of their method, as well as the other multi-view SSL approaches [29, 33, 53, 56], is that they provide a single embedding per image, without a dedicated strategy for aggregating the information contained in an entire image set. This lack of a comprehensive image-set aggregation mechanism is a notable shortcoming, as real-world applications often involve working with diverse collections of images depicting the same object or scene. Capturing the holistic, multi-faceted nature of such visual information requires a more sophisticated approach than simply producing independent embeddings for each individual image.

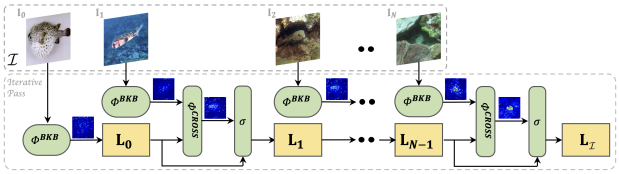

Figure 2. **Overview of MILE.** Given an arbitrary uncalibrated set of images $\mathcal{I} = (\mathbf{I}_i)_{i=0}^{N}$ depicting an object (top row), our method produces a holistic latent representation $\mathbf{L}$ which gets incrementally refined by processing each image input, $\mathbf{I}_i$. The representation is viewed as the state $(\mathbf{L}_i)_{i=0}^{N}$ of a dynamic system (a recurrent network). Each image is independently processed by a shared image encoding backbone Φ^{BKB}. The first state $\mathbf{L}_0$ is initialised with the embedding of the first image, $\Phi^{\text{BKB}}(\mathbf{I}_0)$. The following states $\mathbf{L}_i$, $i > 0$ are computed using a sequence of shared gated cross-attention operations (bottom row), σ and Φ^{CROSS}. The final latent state $\mathbf{L}_{\mathcal{I}} = \mathbf{L}_N$ is taken as the holistic object encoding of the multi-image input $\mathcal{I}$. Detailed algorithmic steps are described in Section 3.1. The entire ensemble of computational units (*i.e.* Φ^{BKB}, σ and Φ^{CROSS}) is trained end-to-end in a self-supervised fashion (see Section 3.2 and Figure 3).

Our work seeks to address this gap by introducing the MILE framework, a self-supervised technique that can construct a single, comprehensive embedding from a set of multi-view images. Inspired by a more dynamic form of SSL, exemplified by recent works [4, 15, 36], the MILE approach employs a student-teacher distillation framework to learn generic visual attribute representations in a self-supervised manner. Additionally, the MILE framework draws inspiration from the tasks of salient object co-segmentation [20, 44, 45], co-detection [12, 37, 60], and object/landmark matching [13, 21, 42, 46, 55, 62]. However, unlike these approaches, which rely on implicit foreground/background statistics and are constrained by the un-

derlying object/scene geometry, the MILE framework operates in an incremental fashion, adaptively refining the object representation as more images are processed.

3. Methodology

Let there be an image set $\mathcal{I} = \{\mathbf{I}_i \mid \mathbf{I}_i \in \mathbb{R}^{h_i \times w_i \times 3},\ i = 0 \dots N\}$ with depictions of an object, where N is the total number of images. Our objective is to obtain a latent representation $\mathbf{L} \in \mathbb{R}^d$ corresponding to the common object surfaced across $\mathcal{I}$. This is achieved through a sequential processing flow (inspired from [26, 41]) of the ordered sequence of images, $(\mathbf{I}_i)_{i=0}^{N}$. Essentially, we iterate over multiple state representations of the latent variable $\mathbf{L}$ by projecting the embedding of each image $\mathbf{I}_i$, retrieved using a shared image encoding backbone, denoted as $\Phi^{\texttt{BKB}} : \mathbb{R}^{h \times w \times 3} \rightarrow \mathbb{R}^d$. In our experimental setup, we use architecture variations of the ViT [10] for $\Phi^{\texttt{BKB}}$, however the method is not bound in any way to this design choice therefore any other encoding backbones [18, 48] can be employed. Intuitively, $\mathbf{L}$ transitions over different states, $(\mathbf{L}_i)_{i=0}^{N}$, as it incorporates complementary cues of the object. We use the final state $\mathbf{L}_N$ containing holistic sequence encoding, referred as $\mathbf{L}_{\mathcal{I}}$.

3.1. MILE

We illustrate the algorithmic steps of our proposed framework in Figure 2. For each image, we retrieve the embedding associated with the `[CLS]` token, as it contains information from the entire image space. *Firstly*, we initialize the latent variable with the embedding of the first image, $\mathbf{L}_0 = \Phi^{\texttt{BKB}}(\mathbf{I}_0)$. The *second* step consists in sequentially parsing the items from $\mathcal{I}$, thus obtaining the associated states of the latent embedding by cross-attending to the embedding of the image from step i, $\Phi^{\texttt{BKB}}(\mathbf{I}_i)$. Therefore, we utilize the cross-attention [5, 25] operation denoted with $\Phi^{\texttt{CROSS}}$. More specifically, the output of $\Phi^{\texttt{BKB}}$ represents the *keys* and *values* and the latent state from the previous step represents the *queries*. Inspired by [1, 19], we incorporate a learnable gating operation, denoted with $\sigma : \mathbb{R} \rightarrow \mathbb{R}$, which combines the $\Phi^{\texttt{CROSS}}$ output with the previous state of the latent variable. In essence, this acts as a gradual shifting mechanism in terms of visual information assimilation, which helps maintaining the right balance between the already processed latent embedding and the visual information processed at the current step. Another implicit benefit of this approach, is the ability to incrementally scale to sequences of undefined lengths with no additional training. Consequently, the state of the latent variable $\mathbf{L}_i$ at iteration $i > 0$ is defined as

$$\mathbf{L}_i = \sigma \cdot \Phi^{\texttt{CROSS}}(\Phi^{\texttt{BKB}}(\mathbf{I}_i), \mathbf{L}_{i-1}) + (1 - \sigma) \cdot \mathbf{L}_{i-1} \quad (1)$$

Lastly, we utilise the final state (*i.e.* $\mathbf{L}_N$) of the sequential pass denoted as $\mathbf{L}_{\mathcal{I}}$ containing the holistic embedding of the common object within the image-set.

3.2. SSL Training Procedure

To train the entire ensemble we use SSL through knowledge distillation (SSLKD). This paradigm is inspired by prior works such as [15], [4], and [49], which we have extended to handle multi-image inputs. An overview of the process is illustrated in Figure 3.

The learning pipeline is initialised by setting up the MILE framework in a Siamese configuration, with $\texttt{MILE}_{\texttt{T}}$ and $\texttt{MILE}_{\texttt{S}}$ representing the teacher and the student. We use an MLP head for each instance, denoted with $\Phi^{\texttt{SSL}}$, performing a succession of higher and lower dimensional projections of the latent embedding representations, $\mathbf{L}_{\mathcal{I}}^{\texttt{S}}$ and $\mathbf{L}_{\mathcal{I}}^{\texttt{T}}$. It is worth mentioning that this head is used solely for the purpose of the SSL training, after which it is discarded prior to the inference stage. Next, a softmax normalization is applied over the output of each head, $\Phi_{\texttt{S}}^{\texttt{SSL}}$ and $\Phi_{\texttt{T}}^{\texttt{SSL}}$, thus obtaining the teacher and student probabilities $\mathbf{\Omega}_{\texttt{S}}$ and $\mathbf{\Omega}_{\texttt{T}}$.

The learning objective is to align the student - teacher probability distributions by minimising the cross-entropy loss w.r.t. the student weights. Practically, we search the optimal joint parameter configuration $\Phi_{\texttt{S}}^{\texttt{SSL}} \circ \texttt{MILE}_{\texttt{S}}$ via cross-entropy loss. Operator $\circ$ denotes the composition between the two modules, $\Phi_{\texttt{S}}^{\texttt{SSL}}$ and $\texttt{MILE}_{\texttt{S}}$. This happens in a distillation context, with a fixed teacher configuration.

Starting from the original image set $\mathcal{I}$, we randomly generate a collection of image sets, having a one-to-one correspondence at individual image level with $\mathcal{I}$, containing crops from $\mathcal{I}$. The obtained collection is split into two, with $(\mathcal{I}_i^{\texttt{G}})_{i=0}^{\texttt{N}_g}$ containing N_g image sets of *global* view crops and $(\mathcal{I}_j^{\texttt{L}})_{j=0}^{\texttt{N}_l}$ containing N_l image sets of *local* view crops. The distribution alignment imposed by loss $\mathcal{L}_{\texttt{MI}}$ is performed while constraining the student network to analyse micro and macro context through $(\mathcal{I}_i^{\texttt{G}})_{i=0}^{\texttt{N}_g} \cup (\mathcal{I}_j^{\texttt{L}})_{j=0}^{\texttt{N}_l}$ while the teacher networks analyse the bigger picture only via $(\mathcal{I}_i^{\texttt{G}})_{i=0}^{\texttt{N}_g}$. By doing this, we encourage the networks to discover implicitly *local-to-global* correspondences within the latent embedding, thus highlighting discriminative aspects of the object across the multi-image setup.

The collection of image sets containing global views are fed to the teacher ensemble, $\Phi_{\texttt{T}}^{\texttt{SSL}} \circ \texttt{MILE}_{\texttt{T}}$, while the collection with all views to the student ensemble, $\Phi_{\texttt{S}}^{\texttt{SSL}} \circ \texttt{MILE}_{\texttt{S}}$. The resulting sets of latent embeddings are denoted with $\mathbb{L}_{\texttt{T}}^{\texttt{GBL}}$ and $\mathbb{L}_{\texttt{S}}^{\texttt{LCL} \cup \texttt{GBL}}$. In essence, the resulting latent sets which we aim to align correspond to the N_g global and N_l local crops generated from the original image set $\mathcal{I}$. Thus, we add a new complexity dimension to the original global to local single-image alignment task described in [4], allowing us to learn a more discriminative representation of an image set (*i.e.* $\mathcal{I}$). However, it comes with the cost of increasing

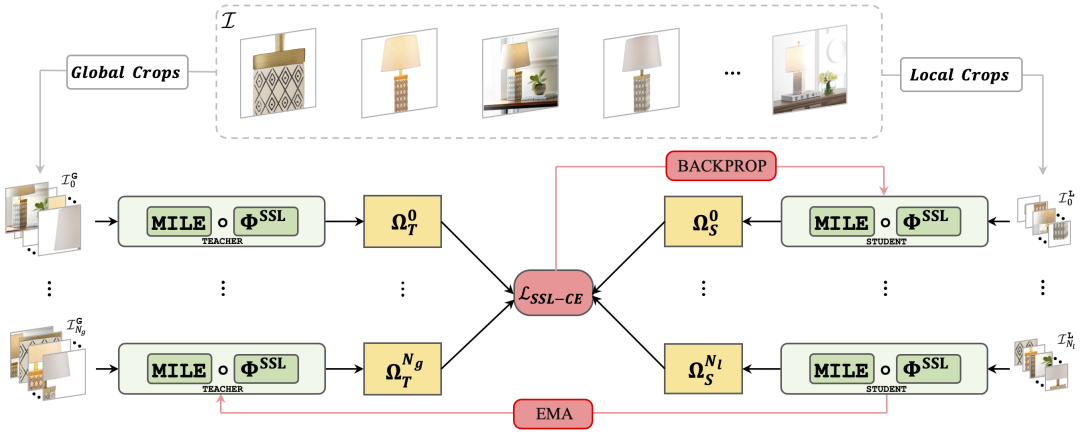

Figure 3. **Self-supervised training flow of MILE framework.** The module obtained from the composition of MILE, described in Section 3.1, and the SSL projection head Φ^{SSL} is cloned into a Siamese student-teacher setup. The outputs of both are normalized using *softmax*, thus producing two probability maps, $\mathbf{\Omega}_\text{S}$ and $\mathbf{\Omega}_\text{T}$. For simplicity, we illustrate the SSL flow by feeding *global* crops $(\mathcal{I}_i^\text{G})_{i=0}^{\text{N}_g}$ to the teacher and *local* crops $(\mathcal{I}_j^\text{L})_{j=0}^{\text{N}_l}$ to the student, however, in reality the student processes both *local* and *global* crops. The student is optimised using gradient backpropagation (see Equation 2) while the teacher ensemble is updated using an exponential-moving average of the student weights.

the instability of the SSL training process. More in depth training details are described in the ***supplementary***.

Therefore, the multi-image $\mathcal{L}_\text{MI}$ loss is defined as

$$\mathcal{L}_\text{MI} = \min_{\Phi_\text{S}^\text{SSL} \circ \text{MILE}_\text{S}} \sum_{\mathbf{L}_i \in \mathbb{L}_\text{T}^\text{LCL}} \sum_{\mathbf{L}_j \in \mathbb{L}_\text{S}^\text{LCL} \cup \text{GBL}} -\mathbf{\Omega}_\text{S}^j \; log(\; \mathbf{\Omega}_\text{T}^i \;) \quad (2)$$

This loss steers the student towards the optimal configuration via stochastic gradient descent. The teacher weights are updated during the training procedure using an *exponential moving average* over the student weights following a momentum encoder λ based on a cosine scheduler.

$$\Phi_\text{T}^\text{SSL} \circ \text{MILE}_\text{T} = \lambda(\Phi_\text{T}^\text{SSL} \circ \text{MILE}_\text{T}) + (1-\lambda)(\Phi_\text{S}^\text{SSL} \circ \text{MILE}_\text{S})$$

Using this learning paradigm, the teacher network stabilizes and simultaneously guides the student towards an optimal exploration through the embedding as the training progresses. This is also encouraged by the fact that we randomly sample different *global* and *local* views with each iteration, along with feeding a different input sequence permutation at each iteration. Practically, we decouple the Φ_T^SSL head and use only the MILE_T module. This can be used as pretrained backbone for any other downstream tasks requiring as input an indefinitely long sequences of images. In our evaluation setup, we experiment with the object retrieval task and, as a byproduct we illustrate regions segmentations of the implicitly inferred common object.

4. Experiments

Experimental Setup. We are interested in the object retrieval task. Traditionally, this is achieved using single-image support which we extend in our setup to multi-image support. This implies that the evaluation set is split in two categories - query, containing image sets depicting the searched objects with one set per object, and gallery, containing a collection of image sets representing the object search space. The objects from query have a one-to-one correspondence to the objects from gallery. This is a non-trivial task, as the search space encodes a high variety of object categories, usually following a fine-grained distribution. Moreover, for a single object the corresponding images are uncalibrated, showing the object in different poses. The objective is to build a ranking of the elements from the gallery according to their similarity against each query item.

We experiment on the challenging ABO [7] and iNaturalist [52] datasets. ABO contains products (objects) showcased in the Amazon catalogue from a wide range of categories. The iNaturalist dataset is comprised of a dense collection of macro categories from the flora and fauna kingdoms, each structured as a wide range of sub-categories denoting animals, insects, plants or other living creatures observed in various natural habitats.

Evaluation Protocol & Baseline Setup. We use the Recall@K (R@K) metric which measures if the relevant item is within the top-K retrieved items. Intuitively, by increasing or decreasing the K threshold, the metric should be more relaxed or enforced, respectively. To understand the robustness of our approach, we measure the average recall at 10 (AR@10). Since there was no readily available baseline comparison that could generate a single embedding for each image-set, regardless of the sequence length, it be-

	Model	Architecture	AR@10
Single-Image	DINO [4]	ViT-S	47.38
	DINO [4]	ViT-S	57.9
	EvaClip [47]	ViT-B	48.13
	EvaClip [47]	ViT-L	55.43
	DINOv2 [36]	ViT-S	57.40
	DINOv2 [36]	ViT-B	61.5
	DINOv2 [36]	ViT-L	66.93
	ft-DINOv2 [36]	ViT-L	80.49
	DINO-REG [8]	ViT-L	70.81
	***ft*-DINO-REG** [8]	**ViT-L**	**84.72**
	MAWS [43]	ViT-B	47.81
	MAWS [43]	ViT-L	48.54
	MAWS [43]	ViT-H	47.04
Multi-Image	MILE w. DINO	ViT-S	63.25
	MILE w. DINOv2	ViT-S	65.8
	MILE w. DINOv2	ViT-B	76.61
	MILE w. DINOv2	ViT-L	79.66
	MILE w. DINO-REG	**ViT-L**	**86.33**
	MILE w. MAWS	ViT-B	72.51
	MILE w. MAWS	ViT-L	82.51
	MILE w. MAWS	ViT-H	81.65

Table 1. **ABO Object Retrieval Results.** Prefix *ft* denotes the fine-tuned version of the single-image model setup on ABO. We fine-tuned only to the best performing pretrained backbones. All the approaches are self-supervised except EvaClip which uses weak-supervision in the form of image and text pairs. Our approach achieves superior performance for the AR@10 metric, thus proving robustness to setups where the structure of the matched object suffers changes between query and gallery visual content.

came necessary to utilize an aggregation technique. Thus, in the single-image setup, for each pair of query and gallery items (image-sets) we compute the pairwise cosine similarity between all the image combinations between the two multi-image sets and aggregate the final similarity result using MAX pooling. Practically, for a gallery and query item pair with n and m elements, respectively, we obtain a matrix of cosine similarities of size $n \times m$ collapsed to a single scalar value via the MAX pooling operation. This value corresponds to the cosine similarity of the best aligned pair of images in the embedding space from the two sets. For evaluation consistency, the results are obtained with $n = 4$ and $m = 4$ for query and gallery for both single and multi image setups.

4.1. ABO

Originally built for 3D object representation, ABO includes an object retrieval benchmark perfectly fit for our setup. For training, given that our approach is self-supervised and does not require external supervision, we use the joint training and validation sets containing $325,075$ images which correspond to a total of $50,756$ objects. For evaluation we use the proposed test set which is split by default in query and target (gallery), containing $23,328$ and $4,313$ image samples, respectively, both with a one-to-one correspondence (836 unique objects) in terms of image-sets.

Numerical results are available in Table 1. To ensure a fair comparison, we fine-tuned the best performing pre-

	Model	Architecture	AR@10
Single-Image	DINO [4]	ViT-S	67.2
	EvaClip [47]	ViT-B	80.85
	EvaClip [47]	ViT-L	84.71
	DINOv2 [36]	ViT-S	77.21
	DINOv2 [36]	ViT-B	82.79
	DINOv2 [36]	ViT-L	84.81
	DINO-REG [8]	ViT-L	84.15
	ft-MAWS [43]	ViT-B	83.90
	ft-MAWS [43]	ViT-L	85.27
	***ft*-MAWS** [43]	**ViT-H**	**86.40**
Multi-Image	MILE w. DINOv2	ViT-S	77.77
	MILE w. DINOv2	ViT-B	86.57
	MILE w. DINOv2	ViT-L	90.96
	MILE w. DINO-REG	ViT-L	93.2
	MILE w. MAWS	ViT-B	85.85
	MILE w. MAWS	ViT-L	90.71
	MILE w. MAWS	**ViT-H**	**94.02**

Table 2. **iNaturalist Category Retrieval Results.** Our best approach, MILE w. MAWS with a ViT-Huge backbone is able to outperform all single-image baselines by a large margin. For this setup, we use the *ft* prefix for the MAWS baselines as they are already fine-tuned on iNaturalist. Our method is superior for all reporting metrics, as the nature of the data enforces robustness of the embeddings with respect to structural variations of the matched classes of object and large shift in background distributions.

Dataset	Single-Image **MAX-POOL**	4-Images Stitch	MILE
ABO	84.72	83.22 ± 0.20	$\mathbf{86.33 \pm 0.23}$
iNaturalist	86.4	87.77 ± 0.25	$\mathbf{94.02 \pm 0.27}$

Table 3. **Learnable Aggregated Single-Image Baseline Results.** Using the best performing single-image models from each dataset, *ft*-DINO-REG for ABO and *ft*-MAWS w. ViT-Huge for iNaturalist we learned a 4-images stitched baseline. For the learnable baseline, we apply 10 random shuffles and report the average AR@10 and the standard deviation. Our approach emerged superior on both evaluation setups, thus ensuring a fair comparison against a learning-based baseline. Furthermore, the standard deviation values highlight the permutation invariance capability of MILE.

trained single-image variants on ABO. It is worth mentioning that MILE achieves superior performance over all backbone sizes. This highlights the robustness of the proposed embedding technique across different image encoding backbones. The MILE approach was able to capture more comprehensive and distinctive visual representations of the objects, even when working with diverse sets of uncalibrated images. The best overall performing approach is MILE w. DINO-REG ViT-Large.

4.2. iNaturalist

For the experiments on iNaturalist, we use a similar query / gallery retrieval setup as for ABO. This is obtained by sampling 100 random classes from all 13 super-classes resulting in 992 unique classes of objects. For each class, the associated images are randomly split in two subsets, one for query and the other for gallery, thus ensuring a one-to-one correspondence between searched classes. Experimental results are illustrated in Table 2. In this scenario, the

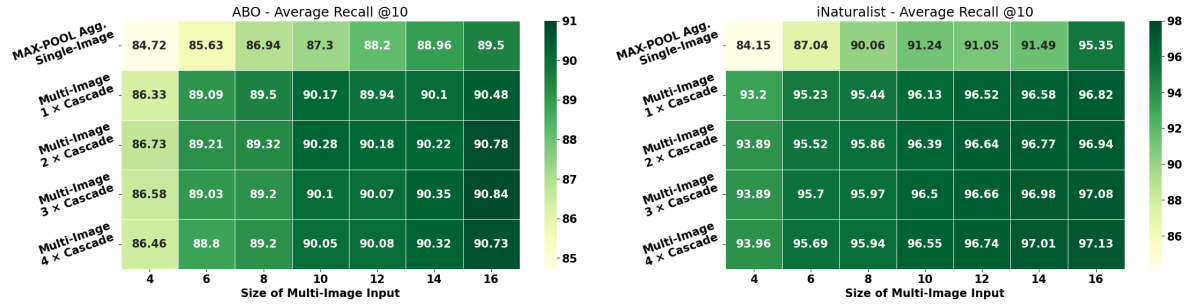

Figure 4. **Ablation on impact of input size on model performance.** We considered two ways of increasing the input size: *(a)* larger image sets with more depictions of the object, corresponding to X-axis and *(b)* cloning the input sequence multiple times, thus simulating a cascading effect, corresponding to the Y-axis. The cascading effect cannot be applied for the single-image setup. The impact is visibly higher on the X-axis, as there is a higher chance for the latent to incorporate unseen (complementary) characteristics of the analysed object. The cascading effect is the equivalent of a second opinion heuristic which recovers missing attributes omitted from the previous passes. Its impact is more visible on iNaturalist due to its more complex nature.

Metric	AR@10			AR@M · 10	AR@M · ⌈10/M⌉
Agg. Function	MEAN	MAX	MAX	MAX	MAX
Agg. Source	Embedding		Pairwise Cosine Similarities		
ABO	82.67	74.67	84.72	**92.20**	85.60
iNaturalist	83.73	80.47	86.40	**96.04**	90.59

Table 4. **Additional Single-Image Baselines.** Columns 2 and 3 illustrate the performance achieved by applying MAX and MEAN pooling operations across the embeddings of the input image set. In contrast, columns 5 and 6 contain the upper bound performance when joining all the the top-K retrieval results for each image query. For consistency, we use $M = 4$ and the best single-image models from Tables 1 and 2, respectively, highlighted in column 4.

Dataset	SI-Query / SI-Gallery	MI-Query / SI-Gallery	MI-Query / MI-Gallery
ABO	75.07	77.07	**86.33**
iNaturalist	87.75	89.36	**93.2**

Table 5. **Ablation on Query / Gallery setup configurations.** Rather than focusing solely on multi-image items in both the query and gallery collections, we assess the impact of incorporating multi-image inputs in the context of a single-image retrieval. The results show that using multi-image queries still outperforms the single-image to single-image search setup (columns 2 and 3), however, both inferior to multi-image to multi-image (column 2). We use best the MILE configs from Tables 1 and 2 for this study.

MAWS [43] single-image baseline is fine-tuned with iNaturalist data. Similar to the trends observed on the ABO dataset, the MILE approach consistently outperformed the associated single-image baseline counterparts. However, for this setup, the wide spectrum of visual diversity present in iNaturalist seems to have favored the high generalizability of the MILE w. MAWS ViT-Huge configuration.

4.3. Additional Studies

Incremental embedding refinement. During the multi-image training, the models are fed with image sets with length equal to 4 which are randomly sampled from the pool of images available for that particular object. Nevertheless, this does not limit the size of the input image set during inference. Therefore, one *significant* feature of our approach is its ability to process indefinitely long sequences of images by using the built-in capacity enabled by the $\sigma(\cdot)$ and $\Phi^{\text{CROSS}}(\cdot)$ functions. This effect is visually emphasized in Figure 1. By analysing MILE w. DINO-REG for both datasets, we empirically discovered (see Figure 4) that the addition of iterative cascading passes over the same image-sequence brings a surplus in performance. Allegedly, this heuristic helps to paint a more complete picture embedding-wise, by incrementally refining the latent representation and picking up missing visual cues of the object.

Search for a better baseline. One might argue that the comparison from Tables 1 and 2 is unfair as MILE is a learned aggregation technique operating at item level, compared to the single-image baselines which perform a MAX-pool operation over off-the-shelf independently obtained single-image embeddings. Consequently, we considered extending the best performing single-image baselines to a learnable 4-images stitch baseline. This strategy involves the usage of fixed sequences with 4 images, followed by a canvas stitching in a 2×2 grid. Next, during the self-supervised learning algorithm, the *global* and *local* views are sampled using crops from each cell of the grid.

The resulting embedding contains aggregated information from all 4 views and the self-supervised learning strategy enforces visual correlation across the images within the grid. Numeric results are illustrated in Table 3. For both datasets, our approach showcases superior results. Moreover, the 4-images stitch strategy imposes a limitation of the multi-image sequence to a length of 4 images, whereas

MILE can be extended to an arbitrarily large number.

We also consider several aggregated single-image baselines. In Table 4, column 2 and 3, we apply a MAX and MEAN pooling operation across the individual image embeddings from the input multi-image set. This operation was perfomed over the best single-image baselines from Tables 1 and 2. The obtained performance was lower suggesting that a simple pooling operation is not a viable solution towards obtaining an aggregated object level embedding. Interestingly, the MEAN-pooling operation performed better than MAX-pooling, likely capturing the common representation of the depicted object across the image set. In columns 4 and 5, we searched the upper bound performance of the single-image baselines. We came with a different heuristic where for each image within the query item, we selected the top-K closest gallery items. Next, we measured the total $AR@M \cdot 10$ metric. Since this is unfair compared to MILE, we measured alternatively $AR@M \cdot \lceil 10/M \rceil$, by keeping the first $\lceil 10/M \rceil$ items for each image within the query item.

In Table 5 we explore the impact of treating the gallery items as single images rather than image sets. Thus, we progressively measure the $AR@10$ for the query / gallery setup considered as single-image / single-image (column 2), multi-image / single-image (column 3) and multi-image / multi-image (column 4). For this experiment we use the best MILE model configuration on both ABO and iNaturalist. It is worth noting that we achieved superior performance with the multi / single configuration compared to the single / single setup.

Extended capabilities of MILE. The main body of experiments are performed around the task of object retrieval as it is a natural fit for the proposed setup, however as a result of the SSL flow we obtained other derived capabilities, anchored in the visual exploratory domain. In this regard, one observation is that the latent variable enforces an incremental holistic representation of the common object with respect to the number of images being processed. Consequently, this should be visually encoded within the cross-attention map generating the latent states, thus sketching a soft segmentation mask of the object of interest. This is achieved by fine-tuning the cross-attention head to attend all the image tokens including the [CLS] token, to anchor the latent information directly in the image space. We illustrate this effect in Figure 6 (*right* heatmap) by inspecting the cross-attention map between the last image (green rectangle) and the previous images (blue rectangle) from the same input sequence. Using this implicit conditioning, we obtain an almost complete representation of the common object, compared to the single-image setup, where the attention is dispersed throughout the image space (*left* heatmap), not focusing on a clear element.

***Keys* and *Values* of $\Phi^{\texttt{CROSS}}$**	$\Phi^{\texttt{BKB}}$	**AR@10**
ABO		
$[\texttt{CLS}]$	*Learned*	82.51
$[\texttt{CLS}] \cup [\texttt{T}_\texttt{i}]_{\texttt{i=0}}^{\texttt{N}_\texttt{Tokens}}$	*Frozen*	83.11
$[\texttt{T}_\texttt{i}]_{\texttt{i=0}}^{\texttt{N}_\texttt{Tokens}}$	*Frozen*	**85.05**
iNaturalist		
$[\texttt{CLS}]$	*Learned*	90.71
$[\texttt{CLS}] \cup [\texttt{T}_\texttt{i}]_{\texttt{i=0}}^{\texttt{N}_\texttt{Tokens}}$	*Frozen*	92.47
$[\texttt{T}_\texttt{i}]_{\texttt{i=0}}^{\texttt{N}_\texttt{Tokens}}$	*Frozen*	**92.47**

Table 6. **Performance impact on multiple $\Phi_{\texttt{CROSS}}$ variations for MILE w. MAWS using `ViT-L`.** By learning a student configuration from a frozen teacher configuration, we unlock granular level information encoded within patch-level embeddings bringing a performance gain in AR@10 for both ABO and iNaturalist.

Another derived capability of MILE is implicit correspondence among the semantics of the regions across the multi-image input sequence. We selected different sample image sets from ABO and iNaturalist and generated token (image-patch) level embeddings obtained with the image encoding backbone $\Phi^{\texttt{BKB}}$. We applied principal components analysis (PCA) on the resulting collection of patch-level embeddings and projected the 3 most dominant dimensions. By plotting the normalized projected dimensions, we notice a strong semantic region correspondence at image set level. This acts as an implicit soft-clustering mechanism, which groups image regions based on their similarity structure. Visual results with this correspondence effect are shown in Figures 7 for both ABO and iNaturalist. These successfully highlight the derived capabilities of the framework for handling other downstream tasks such as region correspondence or semantic region segmentation.

Impact of Macro vs. Micro embeddings for MILE. The MILE models variations from the main Tables with results were trained using the embedding associated with the [CLS] token as *keys* and *values* within the cross-attention module $\Phi^{\texttt{CROSS}}$. We made several training attempts using the entire image information obtained with $\Phi^{\texttt{BKB}}$, specifically the embedding associated with the [CLS] (macro level) plus the granular embeddings (micro level) from all the image tokens, $[\texttt{T}_i]_{i=0}^{N_{Tokens}}$. The EMA student-teacher distillation framework manifested overfitting behaviour from the beginning due to the high instability of the information flow passing from $\Phi^{\texttt{BKB}}$ to $\Phi^{\texttt{CROSS}}$. Therefore, once the training using the [CLS] token converged to an optimal configuration, we applied a frozen distillation learning setup to the best performing MILE configurations by keeping frozen the teacher configuration using [CLS], while distilling it with a student configuration using $[\texttt{CLS}] \cup [T_i]_{i=0}^{N_{Tokens}}$ as *keys* and *values*, resulting in a performance increase on AR@10 for both iNaturalist and ABO (see Table 6). Our hypothesis is that although the cross-attention ensemble managing all image information manifested instability during simultane-

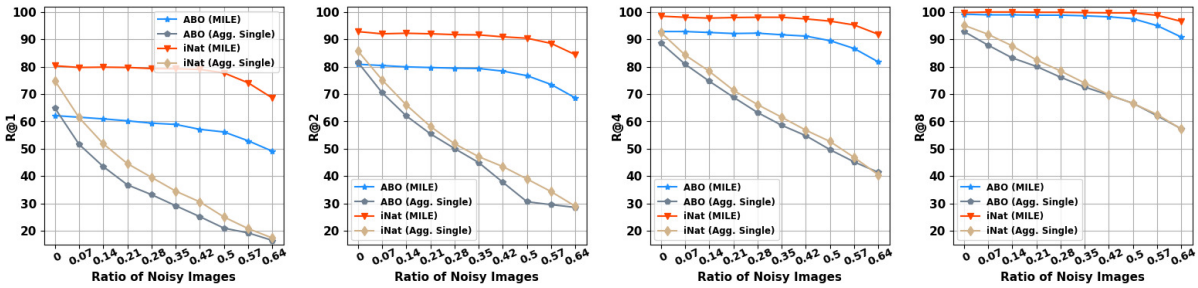

Figure 5. **Sensitivity of object recall performance w.r.t. input noise.** We measure the robustness of the generated object representation when the input sequence is altered on inputs with length 14 using MILE w. DINO-REG and its single-image counterpart. We progressively altered between 0% and 63% of the query input on both setups, single and multi, by replacing the query image from the best matched (via cosine similarity) query-gallery image pairs from single-image setup with random ones. Our approach showcases a much slower degradation rate on the overall evaluation board, demonstrating noise stability. Furthermore, notice the significant performance drop (≈ 15 points) for R@1 and R@2 for ABO on the single-image evaluation setup when the input sequence is altered by a single image, while for MILE there is almost no degradation.

Figure 6. **Visualisations of latent cross-attentions on iNaturalist.** We highlight different attentions maps obtained via different mechanisms for the image within the green rectangle. The heatmap to the *left* is the average attention provided by Φ^{BKB}. The *middle* heatmap comes from the in-place cross-attention (replacing the latent with `[CLS]` embedding), thus not having any previous information. The heatmap to the *right* is the attention resulting from the cross between the latent built from images within the blue rectangle and patch embeddings of the image in question.

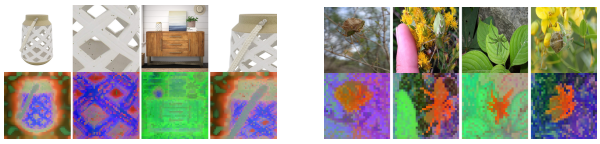

Figure 7. **Color coded region correspondence (*left* ABO and *right* iNaturalist) emphasised via PCA.** We apply PCA decomposition on the patch-level embeddings from all images within the sequence and project in RGB space the 3 most relevant components. The region correspondence is highlighted by the same color across all input images. Despite pose changes, out-of-focus placement within the scene, foreground-background color distribution similarities or placement in the proximity of similar class of objects, the common objects share the same PCA-induced color representation.

ous student-teacher training via distillation, by learning the student from a *frozen* teacher we were able to break the bottleneck flow caused by using only the `[CLS]` embedding and enabled harnessing granular information which resides at image patch level.

Robustness to noise. Another important property of our approach is the resilience to noise within the multi-image input, as illustrated in Figure 5. To assess this, we altered the image content in the query items responsible for the `MAX` similarity scores by replacing them with random images. This allowed us to measure the noise resilience of our approach compared to the best single-image baselines. Our approaches showcased more stable performance across the entire evaluation board (*i.e.* R@1, R@2, R@4, R@8) compared to the `MAX-POOL` aggregated single-image baselines. This demonstrated the robustness of our approach and the capability to self-stabilize even when the input image sequence does not contain exclusively images of the same object. It is worth mentioning that the positions of the replaced images within the original sequences are random, thus showcasing implicit permutation invariance capabilities of our approach. Interestingly, for the zero-noise setup, we noticed a slight increase in R@1 and R@2 scores, which we believe was caused by the presence of near-identical image pairs across the query-gallery pairs.

5. Conclusions

This paper introduces the MILE framework, designed to create a single holistic embedding visually descriptive elements from a multi-image input sequence depicting an object. This is achieved through a self-distillation learning framework, thus enabling the scaling capability of our framework to huge data corpuses. Through extensive ablation studies, we demonstrate the key capabilities of MILE, including its incremental embedding construction mechanism, noise resilience, and implicit permutation invariance. Importantly, the MILE approach outperforms single-image baselines that leverage both learnable and non-learnable embedding aggregation techniques, across multiple architectural configurations evaluated on ABO and iNaturalist.

References

[1] Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al.: Flamingo: a visual language model for few-shot learning. NeurIPS **35**, 23716–23736 (2022) 3

[2] Arandjelović, R., Zisserman, A.: Three things everyone should know to improve object retrieval. In: CVPR. pp. 2911–2918. IEEE (2012) 1

[3] Arica, S., Rubin, O., Gershov, S., Laufer, S.: Cuvler: Enhanced unsupervised object discoveries through exhaustive self-supervised transformers. In: CVPR. pp. 23105–23114 (2024) 2

[4] Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: CVPR (2021) 1, 2, 3, 5

[5] Chen, C.F.R., Fan, Q., Panda, R.: Crossvit: Cross-attention multi-scale vision transformer for image classification. In: ICCV. pp. 357–366 (2021) 3

[6] Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International conference on machine learning. pp. 1597–1607. PMLR (2020) 1

[7] Collins, J., Goel, S., Deng, K., Luthra, A., Xu, L., Gundogdu, E., Zhang, X., Vicente, T.F.Y., Dideriksen, T., Arora, H., et al.: Abo: Dataset and benchmarks for real-world 3d object understanding. In: CVPR. pp. 21126–21136 (2022) 2, 4

[8] Darcet, T., Oquab, M., Mairal, J., Bojanowski, P.: Vision transformers need registers. arXiv preprint arXiv:2309.16588 (2023) 5

[9] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020) 2

[10] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: ICLR (2021), https://openreview.net/forum?id=YicbFdNTTy 3

[11] El Hajjar, S., Dornaika, F., Abdallah, F., Barrena, N.: Consensus graph and spectral representation for one-step multi-view kernel based clustering. Knowledge-Based Systems **241**, 108250 (2022) 2

[12] Fan, Q., Fan, D.P., Fu, H., Tang, C.K., Shao, L., Tai, Y.W.: Group collaborative learning for co-salient object detection. In: CVPR. pp. 12288–12298 (2021) 2

[13] Germain, H., Lepetit, V., Bourmaud, G.: Visual correspondence hallucination: Towards geometric reasoning. arXiv preprint arXiv:2106.09711 (2021) 2

[14] Gordo, A., Almazán, J., Revaud, J., Larlus, D.: Deep image retrieval: Learning global representations for image search. In: ECCV. pp. 241–257. Springer (2016) 1

[15] Grill, J.B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., et al.: Bootstrap your own latent-a new approach to self-supervised learning. NeurIPS **33**, 21271–21284 (2020) 2, 3

[16] Han, X., Ng, K.W., Nag, S., Qu, Z.: Large-scale product retrieval with weakly supervised representation learning. arXiv preprint arXiv:2208.00955 (2022) 2

[17] He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: CVPR. pp. 16000–16009 (2022) 1

[18] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: NIPS (2017) 3

[19] Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation **9**(8), 1735–1780 (1997) 3

[20] Hsu, K.J., Lin, Y.Y., Chuang, Y.Y., et al.: Co-attention cnns for unsupervised object co-segmentation. In: IJCAI (2018) 2

[21] Jiang, W., Trulls, E., Hosang, J., Tagliasacchi, A., Yi, K.M.: COTR: Correspondence Transformer for Matching Across Images. In: ICCV (2021) 2

[22] Johnson, J., Douze, M., Jégou, H.: Billion-scale similarity search with gpus. IEEE Transactions on Big Data **7**(3), 535–547 (2019) 1

[23] Krubiński, M., Matcovici, S., Grigore, D., Voinea, D., Popa, A.I.: Watermark text pattern spotting in document images. arXiv preprint arXiv:2401.05167 (2024) 1

[24] Kumar, K., Arici, T., Neiman, T., Yang, J., Sam, S., Xu, Y., Ferhatosmanoglu, H., Tutar, I.: Unsupervised multi-modal representation learning for high quality retrieval of similar products at e-commerce scale. In: Proceedings of the 32nd ACM International Conference on Information and Knowledge Management. pp. 4667–4673 (2023) 2

[25] Lee, K.H., Chen, X., Hua, G., Hu, H., He, X.: Stacked cross attention for image-text matching. In: ECCV. pp. 201–216 (2018) 3

[26] Leotescu, G., Voinea, D., Popa, A.I.: Bidirectional long-range parser for sequential data understanding. arXiv preprint arXiv:2404.05210 (2024) 3

[27] Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: ICML. pp. 19730–19742. PMLR (2023) 2

[28] Li, Z., Tang, C., Liu, X., Zheng, X., Zhang, W., Zhu, E.: Consensus graph learning for multi-view clustering. IEEE Transactions on Multimedia **24**, 2461–2472 (2021) 2

[29] Lin, Y., Gou, Y., Liu, X., Bai, J., Lv, J., Peng, X.: Dual contrastive prediction for incomplete multi-view representation learning. PAMI **45**(4), 4447–4461 (2022) 2

[30] Lin, Y., Gou, Y., Liu, Z., Li, B., Lv, J., Peng, X.: Completer: Incomplete multi-view clustering via contrastive prediction. In: CVPR. pp. 11174–11183 (2021) 2

[31] Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. NeurIPS **36** (2024) 2

[32] Liu, J., Liu, X., Yang, Y., Guo, X., Kloft, M., He, L.: Multiview subspace clustering via co-training robust data representation. IEEE Transactions on Neural Networks and Learning Systems **33**(10), 5177–5189 (2021) 2

[33] Lu, J., Nie, F., Dong, X., Wang, R., Li, X.: Bidirectional attentive multi-view clustering. IEEE Transactions on Knowledge and Data Engineering (2023) 2

[34] Matcovici, S., Voinea, D., Popa, A.I.: K-nn embeded space conditioning for enhanced few-shot object detection. In: WACV Workshops. pp. 401–410 (January 2023) 1

[35] Muhamed, A., Srinivasan, S., Teo, C.H., Cui, Q., Zeng, B., Chilimbi, T., Vishwanathan, S.: Web-scale semantic product search with large language models. In: Pacific-Asia Conference on Knowledge Discovery and Data Mining. pp. 73–85. Springer (2023) 2

[36] Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023) 2, 5

[37] Peng, Z., Jie, Q., Shuo, W., Tian-Zhu, X., Huan, X.: Memory-aided contrastive consensus learning for co-salient object detection. In: AAAI (2023) 2

[38] Popa, A.I., Sminchisescu, C.: Parametric image segmentation of humans with structural shape priors. In: Computer Vision–ACCV 2016: 13th Asian Conference on Computer Vision, Taipei, Taiwan, November 20-24, 2016, Revised Selected Papers, Part II 13. pp. 68–83. Springer (2017) 1

[39] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: ICML. pp. 8748–8763. PMLR (2021) 2

[40] Sandu, I.C., Voinea, D., Popa, A.I.: Consent: Context sensitive transformer for bold words classification. arXiv preprint arXiv:2205.07683 (2022) 1

[41] Sandu, I.C., Voinea, D., Popa, A.I.: Large sequence representation learning via multi-stage latent transformers. In: COLING. International Committee on Computational Linguistics, Gyeongju, Republic of Korea (Oct 2022) 3

[42] Sarlin, P.E., DeTone, D., Malisiewicz, T., Rabinovich, A.: Superglue: Learning feature matching with graph neural networks. In: CVPR (2020) 2

[43] Singh, M., Duval, Q., Alwala, K.V., Fan, H., Aggarwal, V., Adcock, A., Joulin, A., Dollár, P., Feichtenhofer, C., Girshick, R., et al.: The effectiveness of mae pre-pretraining for billion-scale pretraining. arXiv preprint arXiv:2303.13496 (2023) 5, 6

[44] Stoian, I.S., Sandu, I.C., Voinea, D., Popa, A.I.: Unstructured object matching using co-salient region segmentation. In: CVPR Workshops. pp. 5051–5060 (June 2022) 1, 2

[45] Subramaniam, A., Vaidya, J., Ameen, M.A.M., Nambiar, A., Mittal, A.: Co-segmentation inspired attention module for video-based computer vision tasks. Computer Vision and Image Understanding **223**, 103532 (2022) 2

[46] Sun, J., Shen, Z., Wang, Y., Bao, H., Zhou, X.: Loftr: Detector-free local feature matching with transformers. In: CVPR. pp. 8922–8931 (2021) 2

[47] Sun, Q., Fang, Y., Wu, L., Wang, X., Cao, Y.: Eva-clip: Improved training techniques for clip at scale. arXiv preprint arXiv:2303.15389 (2023) 5

[48] Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A.: Inception-v4, inception-resnet and the impact of residual connections on learning. In: AAAI. vol. 31 (2017) 3

[49] Trosten, D.J., Løkse, S., Jenssen, R., Kampffmeyer, M.C.: On the effects of self-supervision and contrastive alignment in deep multi-view clustering. In: CVPR. pp. 23976–23985 (2023) 2, 3

[50] Tsai, Y.H.H., Wu, Y., Salakhutdinov, R., Morency, L.P.: Self-supervised learning from a multi-view perspective. arXiv preprint arXiv:2006.05576 (2020) 2

[51] Tschannen, M., Kumar, M., Steiner, A., Zhai, X., Houlsby, N., Beyer, L.: Image captioners are scalable vision learners too. Advances in Neural Information Processing Systems **36** (2024) 2

[52] Van Horn, G., Mac Aodha, O., Song, Y., Cui, Y., Sun, C., Shepard, A., Adam, H., Perona, P., Belongie, S.: The inaturalist species classification and detection dataset. In: CVPR. pp. 8769–8778 (2018) 2, 4

[53] Wang, R., Sun, H., Ma, Y., Xi, X., Yin, Y.: Metaviewer: Towards a unified multi-view representation. In: CVPR. pp. 11590–11599 (2023) 2

[54] Wang, W., Sun, Y., Yang, Z., Yang, Y.: Leveraging vision and vision-language models into large-scale product retrieval. arXiv preprint arXiv:2207.12994 (2022) 2

[55] Wiles, O., Ehrhardt, S., Zisserman, A.: Co-attention for conditioned image matching. In: CVPR (2021) 2

[56] Xu, J., Ren, Y., Tang, H., Pu, X., Zhu, X., Zeng, M., He, L.: Multi-vae: Learning disentangled view-common and view-peculiar visual representations for multi-view clustering. In: ICCV. pp. 9234–9243 (2021) 2

[57] Yan, W., Zhang, Y., Lv, C., Tang, C., Yue, G., Liao, L., Lin, W.: Gcfagg: Global and cross-view feature aggregation for multi-view clustering. In: CVPR. pp. 19863–19872 (2023) 2

[58] Yuan, J., Chiang, A.T., Tang, W., Haro, A.: eproduct: A million-scale visual search benchmark to address product recognition challenges. arXiv preprint arXiv:2107.05856 (2021) 2

[59] Zhai, X., Mustafa, B., Kolesnikov, A., Beyer, L.: Sigmoid loss for language image pre-training. In: ICCV. pp. 11975–11986 (2023) 2

[60] Zhao, J.X., Liu, J.J., Fan, D.P., Cao, Y., Yang, J., Cheng, M.M.: Egnet:edge guidance network for salient object detection. In: ICCV (Oct 2019) 2

[61] Zheng, L., Yang, Y., Tian, Q.: Sift meets cnn: A decade survey of instance retrieval. IEEE TPAMI **40**(5), 1224–1244 (2017) 1

[62] Zhou, Q., Sattler, T., Leal-Taixe, L.: Patch2pix: Epipolar-guided pixel-level correspondences. In: CVPR (2021) 2