Reason from Context with Self-supervised Learning

Anonymous Authors[1]

Abstract

Self-supervised learning (SSL) leads to capturing discriminative features that are useful for knowledge transfer. To better accommodate the object-centric nature of current downstream tasks such as object recognition and detection, various methods have been proposed to suppress contextual biases or disentangle objects from their contexts. Nevertheless, these methods often prove inadequate in situations where object identification benefits from contextual cues, such as when inferring tiny, poorly illuminated or occluded objects. Here we investigate whether and how contextual associations can be enhanced for visual reasoning within SSL regimes, by (a) proposing a new **Se**lf-supervised method with external memory for **Co**ntext Reasoning (SeCo) , and (b) introducing two new downstream tasks, lift-the-flap and object priming, addressing the problems of "what" and "where" in context reasoning. In both tasks, SeCo outperformed state-of-the-art (SOTA) SSL methods by a significant margin. Our network analysis revealed that the proposed external memory in SeCo learns to store prior contextual knowledge, facilitating target identity inference in the lift-the-flap task. Moreover, we conducted psychophysics experiments and introduced a **H**uman benchmark in **O**bject **P**riming dataset (HOP). Our results demonstrate that SeCo exhibits human-like behaviors.

1. Introduction

Self-supervised learning (SSL) aims to capture discriminative visual representations from unlabeled images, which could be transferred to downstream tasks such as object recognition and object detection.

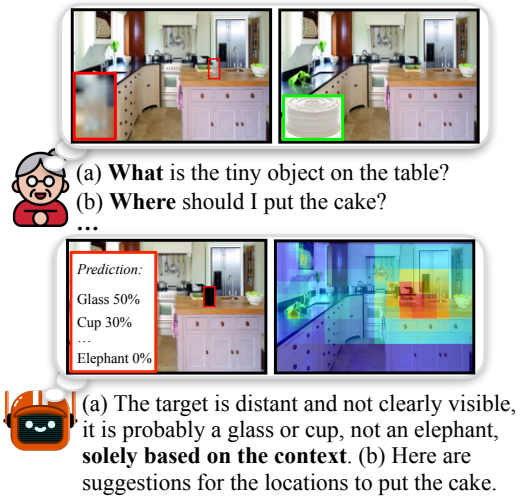


Figure 1. **Schematic illustration of lift-the-flap ("what") and object priming ("where") tasks of context reasoning in real-world applications**. An assistive robot has to solely rely on context to perform two reasoning tasks: (a) Lift-the-flap: infer the identity of a hidden object and (b) Object priming: infer appropriate locations to put an object.

Recent works (Singh et al., 2020; Mo et al., 2021) showed that mitigating contextual biases, caused by co-occurrences of objects and context in a complex scene, can improve the generalization ability of SSL to these downstream tasks. However, these object-centric methods dissociate the objects and contexts and thus fail to address scenarios where contextual information is crucial, such as recognizing/inferring small, blurred, or occluded distant objects (**Fig. 1a**). Objects and contexts always come as pairs in a natural scene. In this light, humans are adept at exploiting contextual cues to fill in information gaps in their sensory input. For example, in **Fig. 1a**, based on the scene context, one can infer that the occluded object inside the red box on the table can be a glass or a cup but not an elephant. To date, context reasoning capacity has been studied with supervised learning methods (Zhang et al., 2020; Bomatter et al., 2021), but there is a lack of SSL methods for contextual reasoning in the literature. Therefore, in this paper, we delve into the question of whether and how contextual associations can be enhanced for visual reasoning in a self-supervised manner.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

To bridge the above gaps, we propose a **Se**lf-Supervised Learning Method for **Co**ntext Reasoning (SeCo), where the pre-training objective is to learn to associate objects and their contexts in the embedding space. Briefly, SeCo first uses unsupervised methods to discover region proposals containing potential target objects of interest. Next, the target object of interest and its surrounding context are processed separately by two independent image encoders. Humans rely on prior knowledge of various objects and their mutual relationships to establish contextual associations. Inspired by human behavioral experiments, we introduce a learnable external memory to store learned contextual priors.

Here we establish a framework to utilize contextual knowledge for context-aware SSL. Given unlabeled images containing multiple objects in natural scenes, the objective of context-aware SSL is to learn object-context associations. To showcase the use of context in practical applications, such as tiny object recognition and placing items in context-appropriate locations for assistive robots, and to evaluate the context reasoning capabilities of different computational models, we introduce two evaluation protocols, lift-the-flap and object priming, addressing the problems of "what" and "where" in context reasoning. Specifically, the lift-the-flap task (**Fig. 1a**) requires all the models to utilize the scene context to infer the class of the hidden target object behind a flap (a black patch). In the object priming task (**Fig. 1b**), given an image and a target object (not already present in the image), models are expected to predict contextually correct image regions for placing the target object.

We stress-tested SeCo and state-of-the-art (SOTA SSL methods on in- and out-of-domain test sets of three datasets in lift-the-flap and object priming tasks. SeCo achieved remarkable performance and beats SOTA SSL methods in all the experiments. To benchmark the model performance in object priming, we conducted human psychophysics experiments. Our results show that SeCo exhibits human-like behaviors. Moreover, we gain insights into the role of our external memory from intensive network analysis. We summarize our key contributions below:

(1) To the best of our knowledge, this is *the first work* to investigate whether and how contextual associations can be enhanced within the SSL regime. We establish a new framework for the SSL community to study context reasoning, where lift-the-flap and object priming protocols are introduced to benchmark the contextual reasoning ability of SSL methods.

(2) We propose a *simple yet effective* SSL method (SeCo) to learn contextual associations. SeCo outperforms SOTA SSL methods on in-domain and out-of-domain test sets in three datasets in lift-the-flap and object priming tasks.

(3) We contribute a *new* object priming dataset (HOP) and human benchmarks on HOP with psychophysics experiments. Our SeCo achieves human-level performance and exhibits human-like behaviors.

2. Related Work

Given that ground truth labels are costly to obtain for supervised learning and that much larger datasets can be used without labels, SSL has become an emerging trend in ML. Past handcrafted pretext tasks have been designed to improve the quality of learned scene representations such as "inpainting" randomly masked regions of an image (Pathak et al., 2016). Another group of works (Hjelm et al., 2018; Misra & Maaten, 2020; He et al., 2020; Chen et al., 2020) use contrastive learning techniques for SSL by pulling positive samples together and pushing negative samples away. However, mining negative examples is not always feasible; thus, current research has shifted focus to representation learning solely from positive samples (Chen & He, 2021; Grill et al., 2020; Bardes et al., 2022; Caron et al., 2021). With the success of transformer-based models in NLP and vision tasks (Dosovitskiy et al., 2020), there has also been a trend in SSL to reconstruct images from randomly masked image patches (He et al., 2022; Chen et al., 2022). However, all these previous methods focus on learning image-level representations from monotonously large, salient, and centered objects (Deng et al., 2009).

Recent studies by Wang et al. (Wang et al., 2021) and Xie et al. (Xie et al., 2021) continue to concentrate on acquiring object-centric representations in self-supervised learning (SSL) settings, emphasizing the learning of such representations from intricate scenes. These works introduce diverse methods for extracting object patches from scenes, such as retrieving object patches from two contextually similar images or applying contrastive losses on local and global views of objects within the same image. However, a common limitation in these works is their struggle to capture associations at the instance level within a scene. Unlike all these works, our SeCo is capable of learning object-context associations from complex images where there could be multiple objects in the scene.

Several SSL methods (Caron et al., 2020; Li et al., 2020) introduce external memories to store trainable object prototypes and use them to assign similar images to distinct clusters. In contrast to these methods, our external memory stores prior knowledge on object-context associations so that our SeCo can flexibly retrieve useful object information from context cues in the visual scenes.

The context of a scene (Torralba et al., 2010; Hoiem et al., 2005; Desai et al., 2011; Lin et al., 2013; Divvala et al., 2009) is crucial to computer vision tasks, such as object

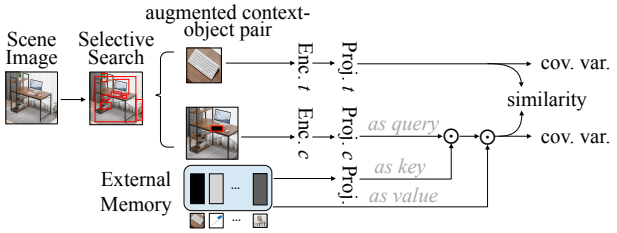


Figure 2. **The architecture of our proposed Self-supervised learning for Context reasoning (SeCo)**. Our SeCo consists of three critical components: an object discovery module, a two-stream visual processor, and an external memory. See **Sec. 3** for design motivation and implementation details, and **Fig. S7** in **Sec. S3.3** for architecture comparison with state-of-the-art SSL methods.

recognition (Zhang et al., 2020; Bomatter et al., 2021), place recognition (Wu et al., 2018), and object detection (Liu et al., 2018; Chen et al., 2018). However, numerous works (Shetty et al., 2019; Singh et al., 2020; Mo et al., 2021) found that models suffer from contextual biases caused by co-occurrences and try to improve the object-centric generalization ability by removing such biases. Breaking away from these works, we investigated the problem of whether and how to leverage contextual cues in the SSL setting. Although previous works introduced datasets with context variations, such as ImageNet-9 (Xiao et al., 2021), these datasets often contain very few objects, discarding the useful information of object co-occurrences in complex scenes. As we aim to study context reasoning abilities in "what" and "where" problems, we introduce lift-the-flap and object priming protocols, focusing on datasets with multiple objects and rich context (Caesar et al., 2018).

3. Method

We propose a Self-Supervised Learning Method for Context Reasoning (SeCo) which learns associations between objects and their contexts in natural images (**Fig. 2**). SeCo consists of three components: (a) object discovery module, (b) two-stream visual processor, and (c) external memory. First, the target discovery module uses unsupervised region proposal methods to locate potential objects of interest on a full image I_f. Each region proposal together with the full image I_f is subsequently converted to pairs of target images I_t and context images I_c. Second, the two-stream visual processor consists of two independent convolutional neural network (CNN) encoders and projectors, extracting information from I_t and I_c, respectively. Third, SeCo employs a trainable external memory to store knowledge priors about contextual associations learned during training phase. Features from I_c serve as queries to retrieve context-relevant prior knowledge from the external memory with an attention mechanism. The retrieved information provides the complementary signal to the context stream and gets compared with the target features from I_t of the object stream to maximize the agreement between the stored prior knowledge and the context-relevant object in the embedding space (see **Algo. S1** in **Sec. S3.5** for the PyTorch-style pseudocode of SeCo's training algorithm).

3.1. Context-Object Pair Discovery

Objects play an important role in context reasoning (Draschkow & Võ, 2017). To learn object-object and object-context associations, we propose a context-object pair discovery module to exploit regions containing objects of interest. We adopt the selective search algorithm (Uijlings et al., 2013) to generate regions of interest (RoI) that potentially contain objects. It is worth noting that selective search is an unsupervised learning algorithm. It performs heuristic searches on hundreds of anchor boxes and proposes RoIs by hierarchically grouping similar regions based on color, texture, size, and shape compatibility. To reduce false positives among many RoIs, we filter out resultant regions according to their area ratio (with a maximum of 0.1) and aspect ratio (within 0.2 and 5). Moreover, we merge RoIs with heavy overlaps by setting the threshold of IoU (intersection over union) as 0.3. For each selected RoI, we generate a pair of target images I_t and context image I_c. I_t is cropped out of full image I_f. The entire image with the RoI blacked out with zeros forms the context image I_c.

3.2. Feature Extraction with CNN

Due to eccentricity dependence, human vision has the highest acuity at the fovea and the resolution drops sharply in the periphery with increasing eccentricity. For example, while we are fixating on the mug on the table, the mug is often perceived in high resolution while the context gist of the kitchen scene is processed at low resolution in the periphery. Taking inspiration from this observation, we propose a two-stream visual processor, with one object stream dedicated to encoding the target image I_t and the other context stream dedicated to encoding the context image I_c. The encoded representations are denoted as $h_c = E_c(I_c)$ and $h_t = E_t(I_t)$, where $E_t(\cdot)$ and $E_c(\cdot)$ are target and context encoders and h_t and $h_c \in \mathbb{R}^D$. Since the features useful for reasoning and perception are different, we do not enforce weight sharing between the encoders. We demonstrate the benefit of this approach in the ablation study.

3.3. Training With An External Memory

As suggested by cognitive and neuroscience works (Zhang et al., 2020; Riesenhuber & Poggio, 1999; Thorpe et al., 1996), context processing often happens very fast in the brain. The perceived scene gist serves as a query to retrieve

prior knowledge from the semantic memory to modulate object recognition in a top-down manner. To mimic this underlying mechanism of context modulation, we introduce an external memory with trainable parameters, accumulating prior knowledge of contextual associations. Different from the well-established cross-attention mechanism (Vaswani et al., 2017), the objective of our external memory focuses on dynamically retrieving and updating the prior knowledge.

We define the external memory as a 2D matrix with trainable parameters, which consists of K memory slots of H dimension, denoted as $M = \{m_1, ..., m_K\}, M \in \mathbb{R}^{H \times K}$. Each memory slot is associated with a key, where $\phi_k(\cdot) : \mathbb{R}^H \rightarrow \mathbb{R}^H$ defines the linear mapping from the memory content to the keys $\phi_k(M)$. The encoded representation h_c from the context stream serves as queries to the external memory after a linear projection operation $\phi_c(\cdot) : \mathbb{R}^D \rightarrow \mathbb{R}^H$. The retrieved prior knowledge $s_c \in \mathbb{R}^H$ from M can then be represented as

$$s_c = \text{SOFTMAX}(\frac{\phi_c(h_c)\phi_k(M)^T}{\sqrt{H}})M \tag{1}$$

where SOFTMAX$(\cdot)$ is the standard softmax operation.

3.4. Loss Components

To encourage M to learn rich and meaningful context-object associations, we introduce three types of losses. Ideally, given only the scene gist, the retrieved prior s_c from M should represent useful object information related to the given context (i.e., "what could be the target object given the scene gist" versus "the actual object seen in the scene"). Thus, we apply a mean squared error loss l_{mse} to maximize the agreement between s_c and h_t. To make the vector dimension comparable, h_t is projected to $s_t \in \mathbb{R}^H$ in the embedding space via $\phi_t(\cdot)$.

As shown by previous works in non-contrastive learning (Bardes et al., 2022; Chen & He, 2021), maximizing the agreement between two-stream visual processors alone may lead to model collapses (e.g., where the external memory stores and outputs trivial knowledge of all zeros, while the visual processor encodes images to representations of all zeros). In this case, s_c and s_t align perfectly, but the encoded object representations and content in M are meaningless.

Thus, to prevent model collapses, we follow (Bardes et al., 2022) to enforce covariance L_{cov} and variance L_{var} regularization on both object and context streams. L_{var} maintains the variance of batch-wise representations, encouraging object class diversities, while L_{cov} de-correlates channel-wise variables to diversify attributes of an embedding, i.e., maximize independent attributes to represent objects. SeCo is jointly trained with the total loss:

$$\begin{aligned} L_{total} = \alpha L_{mse}(s_c, s_t) + \beta[L_{var}(s_c) + L_{var}(s_t)] \\ +\gamma[L_{cov}(s_c) + L_{cov}(s_t)] \end{aligned} \tag{2}$$

where $\alpha = 25$, $\beta = 25$ and $\gamma = 1$ are hyper-parameters weighting different loss components (see **Sec. S3.6** and **Tab. S3** for the hyper-parameter analysis).

3.5. Implementation Details

Augmentations. Data augmentation techniques are widely used at image levels in SSL. We applied standard image augmentations on both I_t and I_c, including color jitter, grayscale, horizontal flip, gaussian blur, and color normalization. Moreover, the random resized crop is another effective technique in SSL. However, directly applying this approach is not feasible in our case. Thus, we extended the standard approach to context-object image pairs with context-aware crops by ensuring that the relative locations among objects are preserved and the bounding box encompassing the target object is always intact and present on I_c after geometric transformations.

Network architecture. We use ResNet-50 (He et al., 2016) with $D = 2048$ output units as our encoders. We set the size of M as $K \times H = 200 \times 512$ and initialize M by the Xavier uniform initializer (Glorot & Bengio, 2010). We demonstrate the benefit of external memory and vary its sizes in the ablation study.

Training. We set the base learning rate to $lr = 0.2 * \text{batch\_size}/256$ (Goyal et al., 2017). The learning rate grows linearly from 0 to base value during the first 10 epochs and then decays with a cosine scheduler (Loshchilov & Hutter, 2016) for the rest of epochs with a minimum value of 0.0002. All our codes and data will be made publicly available upon publication.

4. Experiments

4.1. Datasets

To study contextual associations, we use datasets with multiple objects and rich context: COCO-Stuff (Caesar et al., 2018), PASCAL VOC07 (Everingham et al., 2010) and OCD (Bomatter et al., 2021) (see **Sec. S2.1**). To evaluate whether the learned contextual knowledge from SSL methods can generalize well in out-of-domain settings, we propose two custom regimes on pretext training, fine-tuning, and testing. **COCO-VOC** and **COCO-OCD** contain COCO-Stuff images with their object classes overlapping with VOC07 and OCD datasets respectively. There are 20 classes in COCO-VOC and 15 classes in COCO-OCD (see **Sec. S2.1** for lists of selected classes). We used the training set of COCO-VOC/COCO-OCD for pre-training and fine-tuning and then tested all the models

on the test set of COCO-VOC/COCO-OCD (in-domain) and VOC07/OCD datasets (out-of-domain).

4.2. Baselines

We compared our SeCo against other SSL methods, including Context Encoder(Pathak et al., 2016), SimCLR (Chen et al., 2020), SimSiam (Chen & He, 2021), DINO (Caron et al., 2021), and VICReg (Bardes et al., 2022). For all the methods, we used standard ResNet-50 backbones, with weights pre-trained on ImageNet obtained from their own public checkpoints. We used the same implementations from their original papers. For Context Encoder, since it was originally trained with AlexNet (Krizhevsky et al., 2012), we re-implemented it with the standard ResNet-50 backbone (**Fig. S1**). Moreover, we included a supervised learning baseline that takes I_c as inputs, given the ground truth target labels (see **Sec. S2.2** for further details). To showcase that learning visual representations solely through contrasts among local and global patches or between two contextually similar target objects is insufficient for visual reasoning tasks, we also compared SeCo with patch-wise contrastive learning methods such as DenseCL (Wang et al., 2021), ORL (Xie et al., 2021) in **Sec. S3.4**.

4.3. Evaluation Protocols for Context Reasoning

Lift-the-Flap. We introduce the lift-the-flap task to address the problem of "what" in context reasoning. In the task, all models are required to rely only on context information to infer the class identity of the hidden target object. To adapt the pre-trained model to this task, we freeze the model weights for feature extraction and then only train a linear classifier to predict the hidden target object. We report the performance in Top-1 accuracy of all methods in **Tab. 1**.

Object Priming. We introduce the object priming task to address the problem of "where" in context reasoning. Specifically, the model is given an image and a target object as inputs and has to predict contextually correct locations for placing the target object. As there was no object priming dataset in the literature, we curated our own dataset.

[Stimulus design.] We curated semantically relevant 864 unique image-object pairs on 206 images from the test set of the COCO-OCD dataset. To avoid "crowding" effects that could bias humans and models to place the same target objects in the same locations (e.g., images with eggs near other eggs), for each image-object pair in object priming, we made sure that there were no object instances present on the context image whereby these object instances belong to the same class as the given target object (see **Sec. S2.3** for details about selecting these image-object pairs).

[Human response collection.] We followed standard approved Institutional Review Board protocols and used Amazon Mechanical Turk (AMT) to collect responses from a total of 437 human subjects with their consent. For each subject, we randomly sampled 20 image-object pairs and presented the 800×800 image along with the question "Where would you put this *[obj]*?" where *[obj]* corresponds to the sampled target object. The subjects were required to make 10 non-repeated mouse clicks at relevant regions of the image. For each image-object pair, we collected responses from 3 human subjects, producing 30 unique clicks in total per image-object pair. We show the schematic for the human psychophysics experiment in **Fig. S2** and AMT interface in **Fig. S3**. For each image, we consolidated all 30 click coordinates and generated the click probabilistic map of size 25^2. After post-processing steps (see **Sec. S2.3**), we produced final human priming maps (Column. 3 of **Fig. 3** & **Fig. S5**).

[Model-human comparisons.] To predict priming maps for all the models, we converted the object priming task to a series of lift-the-flap tasks with the following steps: (1) we divide the context image into patches. (2) We covered a single image patch with a flap (black pixels) while the remaining patches remain intact. (3) We tested all models fine-tuned on COCO-OCD from the lift-the-flap task in (2) and recorded the predicted classification probability of the model for the given target object class in the object priming task. (4) We iterated through (2) and (3) until we exhaustively performed "lift-the-flap" tasks over all the image patches. (5) For each image patch, we then have a classification score indicating how confidently the model would put the given target object in that patch. We consolidated all the probabilities for all the patches and generated the priming map for each model. As the model predictions were sensitive to the patch sizes, we varied the patch sizes and normalized the final priming map over all patch sizes (see **Algo. S2** in **Sec. S2.3** for details). We compared the similarity between human priming maps and the priming maps generated by all models using root-mean-squared errors (RMSE) and reported the results in **Tab. 3**.

5. Results

5.1. Lift-the-flap task

We report the top-1 target inference accuracy of all models in the lift-the-flap task (**Tab. 1**). SeCo achieves an overall accuracy of 52.31% and 52.43% on the test sets of COCO-VOC and COCO-OCD, surpassing all the baselines by a large margin. The Context Encoder (Pathak et al., 2016) is trained with the hand-crafted pretext task by reconstructing the masked region at the pixel level. However, its performance is inferior to other baselines and our SeCo, implying that pixel-level reconstruction focuses on details of visual features, discarding the local

Table 1. **SeCo outperforms all baselines in the lift-the-flap task.** We test all the baselines on in- and out-of-domain images from COCO-VOC and COCO-OCD regimes (see **Sec. 4.1** section for data splits). We report the network size in pre-training and top-1 accuracy averaged over 5 runs. See **Sec. 4.2** for details.

	Method	#Param	In Domain	Out Of Domain
COCO-VOC	*Supervised*	*24M*	*48.59*	*53.69*
	Context Encoder	28M	15.78	14.82
	SimCLR	28M	32.78	37.65
	SimSiam	38M	39.79	45.76
	DINO	133M	42.06	48.07
	VICReg	175M	44.89	52.58
	SeCo (Ours)	50M	**52.31**	**57.27**
COCO-OCD	*Supervised*	*24M*	*42.51*	*20.17*
	Context Encoder	28M	20.55	10.68
	SimCLR	28M	35.78	15.51
	SimSiam	38M	42.46	19.36
	DINO	133M	43.21	15.34
	VICReg	175M	44.34	24.31
	SeCo (Ours)	50M	**52.43**	**31.37**

contextual associations, such as object co-occurrences. Contrastive methods like SimCLR (Chen et al., 2020) performed worse compared with non-contrastive methods like SimSiam (Chen & He, 2021). This observation suggests that multiple objects could co-occur in the same context and making a selection of negative samples is non-trivial and challenging in context-aware SSL. Interestingly, DINO (Caron et al., 2021) and VICReg(Bardes et al., 2022) have almost 3 times more parameters, but still underperform SeCo, indicating a larger capacity does not guarantee better reasoning ability. Moreover, SeCo even surpasses the supervised learning baseline, suggesting that SeCo learns to capture meaningful contextual associations in the scenes, beneficial for downstream reasoning tasks.

Contextual associations should be invariant to domain shifts of visual features (e.g., a bird flying in the sky regardless of whether the scene is depicted in Picasso or Monet styles). We tested all models in out-of-domain datasets, PASCAL VOC07 and OCD. Without any fine-tuning, SeCo outperforms previous approaches on out-of-domain images, with top-1 accuracy of 57.27% and 31.37% on PASCAL VOC07 and OCD respectively. Compared across domains, we noted that all methods achieve slightly better performance in PASCAL VOC07 than COCO-VOC, because both COCO-VOC and PASCAL VOC07 contain natural images, and the context-associated object pairs on these images are more prevalent on VOC. On the contrary, domain shift from natural images in COCO-OCD to synthetic images in OCD leads to a big performance drop for all the models. Yet, our model gets less impaired due to domain shifts, highlighting that SeCo learns contextual associations rather than correlations of visual features.

One critical challenge in the lift-the-flap task is to identify small, blurred, or occluded distant objects. To demonstrate this point, all the baseline SSL methods leverage contextual information in the lift-the-flap task as priors to modulate their predicted probability distribution in the object recognition task on COCO-OCD dataset. See **Sec. S3.1** for implementation details. We report the top-1 recognition accuracy in **Tab. 2**. Compared to the case when all SSL baselines recognize objects based on I_t alone, we observe higher top-1 accuracy after incorporating context. This suggests that context enhances object recognition. Moreover, after substituting the prior distribution predicted by all SSL baselines themselves in the lift-the-flap tasks with our SeCo, we saw another significant boost in object recognition accuracy. This emphasizes the superiority in the context reasoning ability of our SeCo against all SSL baselines. Consistent with previous works (Zhang et al., 2020; Bomatter et al., 2021), we also break down the results according to the target object sizes and we find that the effect of contextual cues is more prominent in recognizing smaller target objects (see **Sec. S3.1** and **Fig. S6** for results and more analysis).

Table 2. **SeCo enhances object recognition abilities of all baselines.** We report top-1 accuracy averaged over 5 runs on COCO-OCD dataset in object recognition tasks under three conditions: (1) without contextual priors; (2) with contextual priors predicted by the baselines and (3) by our SeCo.

Object	Context	Accuracy	Object	Context	Accuracy
SimCLR	-	55.38	SimSiam	-	67.12
SimCLR	SimCLR	57.33	SimSiam	SimSiam	**70.93**
SimCLR	SeCo	**58.29**	SimSiam	SeCo	70.72
DINO	-	70.84	VICReg	-	74.52
DINO	DINO	73.35	VICReg	VICReg	75.53
DINO	SeCo	**74.17**	VICReg	SeCo	**76.46**

5.2. Object priming task

We compare human priming maps with the maps predicted by all models and report RMSE scores in **Tab. 3**. As an upper bound, we calculated the between-human RMSE score (0.17) by comparing maps from pairs of humans. SeCo achieves the lowest RMSE of 0.32 compared to all baselines, emphasizing that SeCo predicts more human-like priming maps than all the baselines. In general, we also noticed that there still exists a big gap between model-human and human-human agreement in object priming. This gap could be due to several reasons: (1) models are not finetuned on the HOP dataset; (2) discrete priming maps have different-sized grids from the ones used in human experiments; and (3) it is still challenging for machines to capture how humans incorporate context, given that humans have decades of daily experience with context.

To assess the quality of the predicted priming maps by all

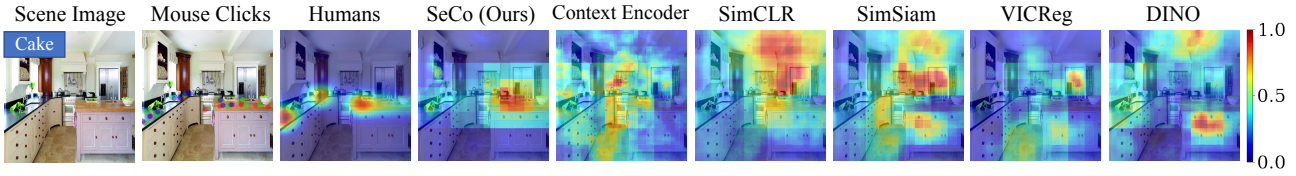


Figure 3. **SeCo priming maps highlight contextually relevant regions of the image and closely approximate human choices in the object priming task**. The leftmost column shows the input scene image and the given target object class label used for priming. The rest of the columns from left to right are the consolidated click map from 3 human subjects with 10 non-overlapping mouse clicks (dots) each in different colors, ground truth priming maps generated from human mouse clicks, and priming maps predicted by our SeCo and predicted by all baselines. See more qualitative examples in **Sec. S2.3** and **Fig. S5**.

Table 3. **Root mean square error (RMSE) between human priming maps and maps predicted by computational models in object priming task**. Lower is better. RMSE for the human agreement was calculated by comparing priming maps across the 3 human subjects for individual image-object pairs.

Method	RMSE	Method	RMSE
Supervised	*0.37*	Human	0.17
Context Enc.	0.41	SimCLR	0.44
SimSiam	0.43	DINO	0.42
VICReg	0.40	SeCo (Ours)	**0.32**

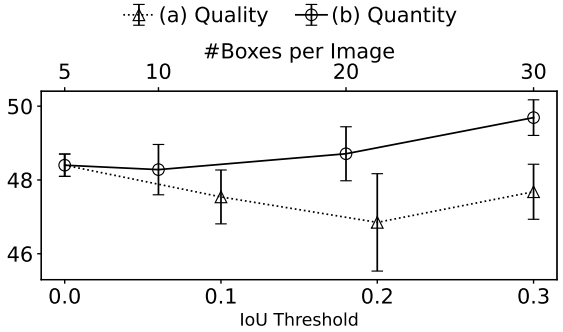


Figure 4. **Analysis of the effect of the quantity and quality of object proposals predicted by selective search in the lift-the-flap task.** We report the top-1 accuracy for varying quality of proposals (a) and varying quantity of proposals per image (b) in the lift-the-flap task. See **Sec. S3.2** for more details.

models, we also visually examined qualitative examples (**Fig. 3** and **Fig. S5**). In contrast to all the baselines which tend to generate relatively uniform flat priming maps, our SeCo manages to predict semantically reasonable locations to place target objects. Note that we do not train or fine-tune any methods to fit human priming maps. It is quite remarkable that our SeCo can transfer the knowledge in contextual associations to infer target-relevant semantically-correct locations in the scene.

5.3. Ablation and memory analysis

We assessed the importance of design choices by training and testing ablated versions of SeCo on COCO-OCD.

First, to demonstrate the effectiveness of the object discovery module, we replaced the object-context image pairs proposed by selective search (Uijlings et al., 2013) with randomly generated object-context image pairs (**Tab. 4**, III, RG). As expected, RG acts as the lower bound of the discovery module, and the top-1 accuracy drops by 16%. This highlights that the "objectiveness" in generated regions helps learn contextual associations. We also trained SeCo on the object-context image pairs from annotated ground truth bounding boxes (**Tab. 4**, II). Surprisingly, SeCo performs better with SS by 3%.

To investigate how SS affects pre-training, we looked into both the quantity and quality of the proposals from SS. See **Sec. S3.2** for experimental setups. We observe that in **Fig. 4 (a)**, raising the Intersection of Union (IoU) threshold compromises the performance of SeCo, while there is a slight gain in **Fig. 4 (b)** when more proposals are included. It indicates that the quantity of the proposals determines the target diversity and matters more than the quality of the proposals. This observation can attribute to SeCo performing better with selective search than with ground truth. Next, to further stress-test that our external memory dedicates to storing context-object associations, rather than a general form of "inpainting" buffer for filling in any missing pixels on I_c, we substituted I_c and I_t with two standard augmented views of the full image I_f (**Tab. 4**, IV). The inferior performance to our SeCo highlights: (1) context-object pair discovery module is essential, and (2) external memory works best in reasoning on object identity from context.

Next, we prepended object-discovery modules to feed object-context pairs to SimSiam and VICReg, denoted as SimSiam-SS and VICReg-SS (**Sec. S3.3** and **Fig. S7**). We also included the downsized VICReg with comparable network sizes as SeCo (VICReg-SS_{Tiny}). From **Tab. S1**, SeCo significantly surpasses VICReg-SS_{Tiny} and SimSiam-SS and performs competitively well as VICReg-SS although SeCo is 7 times smaller than VICReg-SS, which indicates object-discovery module

Table 4. **Ablation Study**. Top-1 accuracy in lift-the-flap on COCO-OCD for ablated models, where SS denotes Selective Search, GT denotes Ground Truth, RG denotes Random Generating, Standard denotes standard augmented view input, NSA denotes Non-Shared Architecture. See **Sec. 5.3** for descriptions. Default settings of SeCo are highlighted.

	Discovery	NSA	Memory	#Param	Accuracy
I	SS	✓	✓	50M	**52.43**
II	GT	✓	✓	50M	49.61
III	RG	✓	✓	50M	36.95
IV	Standard	✓	✓	50M	43.01
V	GT	✗	✓	25M	37.48
VI	GT	✓	✗	49M	44.07
VII	SS	✓	✗	50M	39.95

works best with the external memory.

Morever, we trained two separate encoders $E_t(\cdot)$ and $E_c(\cdot)$ in SeCo. Here, we enforced weight-sharing encoders (**Tab. 4**, V). SA achieved a lower top-1 accuracy than SeCo, suggesting that the same features for both target and context streams are insufficient to reason about context.

To study the effect of the external memory in context reasoning, we remove the external memory from our default SeCo (**Tab. 4**, VI), which leads to 5% drop in performance. To validate that the performance gain from external memory is not simply due to additional capacity, we remove external memory and increase the capacity of SeCo until its network size becomes comparable with the original SeCo (**Tab. 4**, VII). Compared to the original SeCo (**Tab. 4**, I), the performance drops by 12.5%. Inferior results in these two ablation studies demonstrate that external memory enhances the reasoning ability of SeCo. We also vary the number of memory slots and feature dimension of the external memory respectively (see **Sec. S3.5**). We observe that the performance of SeCo saturates when the external memory is oversized (**Fig. S8**). It suggests that larger memory capacity in general helps learn and store richer contextual associations; however, an overly large-sized memory may hurt context reasoning abilities, as the memory fails to generalize the learned contextual knowledge due to over-fitting.

We further probe what the external memory has learned by visualizing the pairwise KL divergence of attention score over memory slots for object categories in COCO-VOC. Each cell in the matrix denotes the distance of attended memory slots to retrieve information from, given the pair of contexts where the two object classes are present. The darker grids denote that object classes are more likely to share the same context. See **Sec. S3.6** and **Algo. S3** for implementation details. We highlighted several context-relevant pairs of object classes from various supercategories, such as vehicles, animals, and indoor objects. For example, though the tv and the potted plants are not visually similar, they are contextually relevant. This suggests the external memory in SeCo learns meaningful object-context associations.

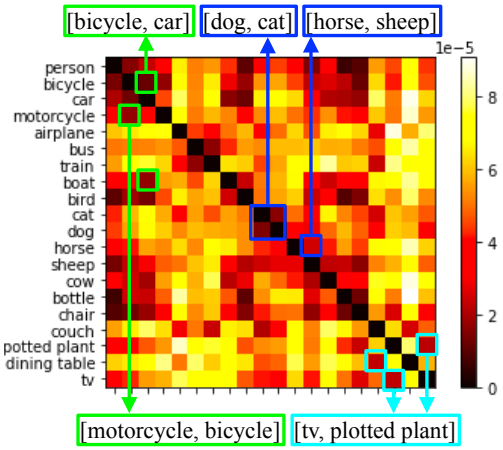


Figure 5. **Pairwise KL div. of attention scores over memory slots of the external memory in SeCo for object categories in COCO-VOC**. Dark grids show that targets sharing similar contexts in both categories retrieve information from similar memory slots. Colored boxes pointed by arrows denote different supercategories in VOC07, e.g. *vehicle*, *animal*, *indoor*. See **Sec. S3.6** for implementation details.

6. Discussion

We set out to determine whether and how SSL methods can capture the statistics of associations in natural images. To this end, we introduced SeCo, a simple yet effective self-supervised learning method for context reasoning, which learns object-context associations from unlabeled images. Like humans, SeCo relies on external memory to develop knowledge priors through repeated encounters with objects and their contexts during learning. SeCo subsequently reasonsby retrieving information from these learned priors.

We speculate that humans also learn context in a largely self-supervised fashion, similar to the learning protocol in SeCo. It is interesting that the SeCo model can extrapolate across lift the flap and object priming tasks from different domains. Our SeCo also significantly outperforms SOTA SSL methods, closing the gap in reasoning abilities between humans and AI models. Relying too much on context can be harmful in some corner cases. Thus, in the future, it will be important to investigate the trade-off between identifying objects and reasoning from context. Moreover, as our proposed external memory in SeCo can bootstrap reasoning ability, it is also worth investigating the generic memory functionality in object-centric SSL settings.

Impact Statements

SeCo relies on contextual information for decision-making, necessitating careful consideration during development and deployment to address potential biases. Concerns include the potential for falsifying context to manipulate SeCo into unfair decisions and the risk of unfair biases stemming from contextual reasoning in the training set.

References

Bardes, A., Ponce, J., and Lecun, Y. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. In *ICLR 2022-10th International Conference on Learning Representations*, 2022.

Bomatter, P., Zhang, M., Karev, D., Madan, S., Tseng, C., and Kreiman, G. When pigs fly: Contextual reasoning in synthetic and natural scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 255–264, 2021.

Caesar, H., Uijlings, J., and Ferrari, V. Coco-stuff: Thing and stuff classes in context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1209–1218, 2018.

Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., and Joulin, A. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in Neural Information Processing Systems*, 33:9912–9924, 2020.

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., and Joulin, A. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9650–9660, 2021.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PMLR, 2020.

Chen, X. and He, K. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15750–15758, 2021.

Chen, X., Li, L.-J., Fei-Fei, L., and Gupta, A. Iterative visual reasoning beyond convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7239–7248, 2018.

Chen, X., Ding, M., Wang, X., Xin, Y., Mo, S., Wang, Y., Han, S., Luo, P., Zeng, G., and Wang, J. Context autoencoder for self-supervised representation learning. *arXiv preprint arXiv:2202.03026*, 2022.

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

Desai, C., Ramanan, D., and Fowlkes, C. C. Discriminative models for multi-class object layout. *International journal of computer vision*, 95(1):1–12, 2011.

Divvala, S. K., Hoiem, D., Hays, J. H., Efros, A. A., and Hebert, M. An empirical study of context in object detection. In *2009 IEEE Conference on computer vision and Pattern Recognition*, pp. 1271–1278. IEEE, 2009.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2020.

Draschkow, D. and Võ, M. L.-H. Scene grammar shapes the way we interact with objects, strengthens memories, and speeds search. *Scientific reports*, 7(1):1–12, 2017.

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

Glorot, X. and Bengio, Y. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pp. 249–256. JMLR Workshop and Conference Proceedings, 2010.

Goyal, P., Dollár, P., Girshick, R., Noordhuis, P., Wesolowski, L., Kyrola, A., Tulloch, A., Jia, Y., and He, K. Accurate, large minibatch sgd: Training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*, 2017.

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. *Advances in neural information processing systems*, 33:21271–21284, 2020.

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

He, K., Fan, H., Wu, Y., Xie, S., and Girshick, R. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9729–9738, 2020.

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., and Girshick, R. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16000–16009, 2022.

Hjelm, R. D., Fedorov, A., Lavoie-Marchildon, S., Grewal, K., Bachman, P., Trischler, A., and Bengio, Y. Learning deep representations by mutual information estimation and maximization. In *International Conference on Learning Representations*, 2018.

Hoiem, D., Efros, A. A., and Hebert, M. Geometric context from a single image. In *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, volume 1, pp. 654–661. IEEE, 2005.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, volume 25, 2012. URL https://proceedings.neurips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf.

Li, J., Zhou, P., Xiong, C., and Hoi, S. Prototypical contrastive learning of unsupervised representations. In *International Conference on Learning Representations*, 2020.

Lin, D., Fidler, S., and Urtasun, R. Holistic scene understanding for 3d object detection with rgbd cameras. In *Proceedings of the IEEE international conference on computer vision*, pp. 1417–1424, 2013.

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

Liu, Y., Wang, R., Shan, S., and Chen, X. Structure inference net: Object detection using scene-level context and instance-level relationships. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6985–6994, 2018.

Loshchilov, I. and Hutter, F. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016.

Misra, I. and Maaten, L. v. d. Self-supervised learning of pretext-invariant representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6707–6717, 2020.

Mo, S., Kang, H., Sohn, K., Li, C.-L., and Shin, J. Object-aware contrastive learning for debiased scene representation. In *35th Conference on Neural Information Processing Systems, NeurIPS 2021*. Neural Information Processing Systems, 2021.

Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T., and Efros, A. A. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2536–2544, 2016.

Riesenhuber, M. and Poggio, T. Hierarchical models of object recognition in cortex. *Nature neuroscience*, 2(11): 1019–1025, 1999.

Shetty, R., Schiele, B., and Fritz, M. Not using the car to see the sidewalk–quantifying and controlling the effects of context in classification and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8218–8226, 2019.

Singh, K. K., Mahajan, D., Grauman, K., Lee, Y. J., Feiszli, M., and Ghadiyaram, D. Don't judge an object by its context: Learning to overcome contextual bias. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11070–11078, 2020.

Thorpe, S., Fize, D., and Marlot, C. Speed of processing in the human visual system. *nature*, 381(6582):520–522, 1996.

Torralba, A., Murphy, K., and Freeman, W. Using the forest to see the trees: Object recognition in context. *Comm. of the ACM*, 2, 2010.

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

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Wang, X., Zhang, R., Shen, C., Kong, T., and Li, L. Dense contrastive learning for self-supervised visual pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3024–3033, 2021.

Wu, K., Wu, E., and Kreiman, G. Learning scene gist with convolutional neural networks to improve object recognition. In *2018 52nd Annual Conference on Information Sciences and Systems (CISS)*, pp. 1–6. IEEE, 2018.

Xiao, K., Engstrom, L., Ilyas, A., and Madry, A. Noise or signal: The role of image backgrounds in object recognition. In *International Conference on Learning Representations*, 2021.

Xie, J., Zhan, X., Liu, Z., Ong, Y. S., and Loy, C. C. Unsupervised object-level representation learning from scene images. *Advances in Neural Information Processing Systems*, 34:28864–28876, 2021.

Zhang, M., Tseng, C., and Kreiman, G. Putting visual object recognition in context. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12985–12994, 2020.

S1. Method

We provide PyTorch-style pseudocode for SeCo in **Algo. S1**. In practice, we randomly sample 4 target-context pairs for each image in each iteration and average the loss value over these sampled pairs. We resize the context images to 224×224 and the target images to 96×96. All our experiments were conducted on Ubuntu with NVIDIA RTX A5000 GPUs of 24GB memory. Our code is implemented based on the public repository of each baseline, with following core packages: PyTorch 1.11.0, opencv-python 4.6.0, numpy 1.22.3. All our codes and data will be made publicly available upon publication.

Algorithm S1 PyTorch-style pseudocode for SeCo

```
# Ec, Et:  context and target encoders
# pc, pt:  context and target projectors
# M: external memory shaped in K-by-H
# pk:  key projection of external memory
# mse:  mean square error loss
# var_loss:  variance loss
# cov_loss:  covariance loss
# alpha, beta, gamma:  weightage of each
loss component

# load a batch of N images
for x in loader:

   # randomly augmented target and context
   t, c = augment(x)

   # encode and project context, target
   stream
   hc, ht = Ec(x), Et(x) # N x D
   sc, st = pc(hc), pt(ht) # N x H
   # compute keys of memory
   m = pk(M) # K x H
   # retrieve memory
   p = softmax(dot(sc, m))/sqrt(H) # N x K
   sc = p * M # N x H
   # calculate loss and update
   loss = alpha * mse(sc,st) + beta *
   (var_loss(sc) + var_loss(st)) / 2 + gamma
   * (cov_loss(sc)+ cov_loss(st))
   loss.backward()
```

S2. Experiments

S2.1. Datasets

COCO-Stuff Dataset (Caesar et al., 2018) contains 160K natural images from MSCOCO (Lin et al., 2014) with 80 thing classes and 91 stuff classes in total. Importantly, this dataset captures complex relationships between multiple objects and carries rich contextual information.

PASCAL VOC07 Dataset (Everingham et al., 2010) contains 9,963 images of realistic scenes with a total of 20 object classes.

Out-of-Context Dataset (OCD) (Bomatter et al., 2021) contains 15,773 synthetic test images of indoor scenes with 36 classes under 6 different contextual conditions. In our work, we only consider *normal context* condition with 2,309 test images.

To evaluate whether the learned contextual knowledge from SSL methods can generalize well in out-of-domain settings, we design two custom regimes for our experiments COCO-VOC and COCO-OCD. Overlapped classes are as follows:

COCO-VOC contains the same 20 classes in hierarchy of *superclass* and subclass as defined in PASCAL VOC07 (Everingham et al., 2010).

- *Person*: person
- *Animal*: bird, cat, cow, dog, horse, sheep
- *Vehicle*: aeroplane, bicycle, boat, bus, car, motorbike, train
- *Indoor*: bottle, chair, dining table, potted plant, sofa, tv/monitor

COCO-OCD contains the same 15 classes as in OCD dataset (Bomatter et al., 2021): wine glass, cup, knife, bowl, apple, cake, mouse, remote, keyboard, cell phone, microwave, book, toothbrush, pillow, towel.

S2.2. Baselines

We use ResNet-50 (He et al., 2016) as the encoder in Context Encoder (Pathak et al., 2016) for fair comparisons with other baselines (see **Sec. 4.2**). Following its original work, we use an asymmetric decoder with five up-convolution layers to reconstruct the masked central region. See (**Fig. S1**) for the architecture design. We pre-trained the model on ImageNet-1K(Deng et al., 2009) with mean square error loss for 100 epochs. We set the learning rate as 0.001. Starting from weights obtained on ImageNet-1K, we further fine-tuned the model on COCO-VOC and COCO-OCD respectively.

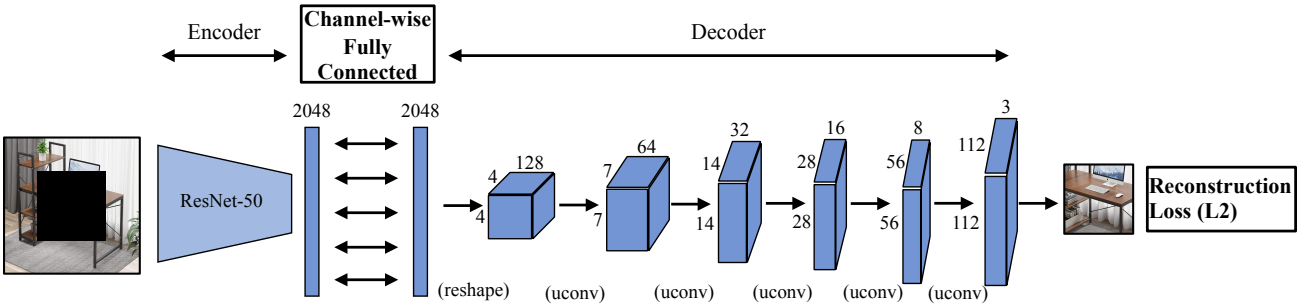


Figure S1. **The architecture of Context Encoder (Pathak et al., 2016) with ResNet-50 (He et al., 2016) as backbone encoder.** Aligned with its original work, we use a channel-wise fully connected layer followed by a five-layer decoder to reconstruct the masked central region from the encoder output.

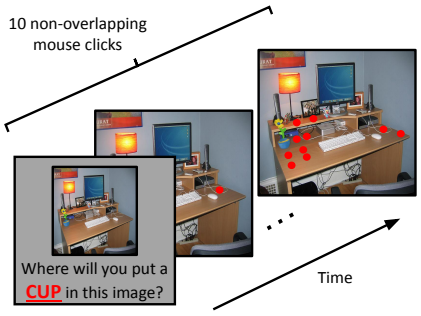


Figure S2. **Schematic for human psychophysics experiments in object priming task.** Subjects were first presented with a natural image and a target object. They were then asked to put the object at appropriate locations by making 10 non-overlapping mouse clicks (red dots).

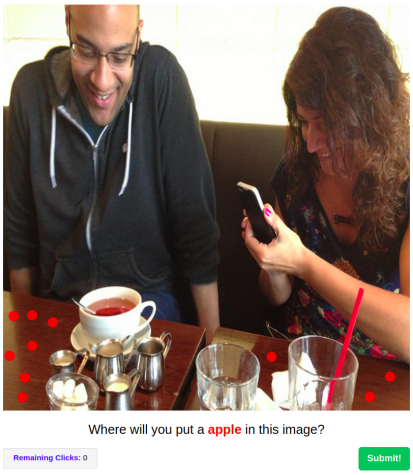


Figure S3. **AMT user interface for human object priming experiment.** Red dots indicate the past click locations.

S2.3. Object Priming

[Stimulus designs] Here, we describe the steps to curate semantically relevant image-object pairs for the object priming experiment. First, we wanted to select images that were semantically relevant to the 15 classes of the COCO-OCD dataset. To accomplish this, we sampled images from the test set of the COCO-OCD dataset that contained at least 3 object classes from the 15 objects classes. Next, for each image i in the sampled images, we manually select a subset C_i of semantically meaningful target classes from the 15 classes ensuring that the target class is not already present in the image. Following the above steps, we produce 206 images and 864 unique image-object pairs.

[Human response collection] we show the schematic for the human psychophysics experiment in **Fig. S2** and a screenshot of the AMT interface in **Fig. S3** used for human object priming experiments. All the psychophysics experiments were conducted with the subjects' informed consent and according to the protocols approved by our Institutional Review Board. For quality controls, we only recruited participants with *master* qualification and a minimum of 95% approval rate. Each participant is compensated for participation in the experiments, which typically took 6 mins to complete.

[Post-processing] Here, we describe the post-processing of human object priming responses in detail. We first created a 32×32 attention map by dividing the 800×800 stimuli image into 1,024 individual grids of size 25×25. We then aggregate the clicks made in each grid such that the pixel intensity in the attention map corresponds to the number of clicks. On this 32×32 attention map, we then apply Gaussian smoothing using an 11×11 filter, followed by resizing to 224×224, and min-max normalization to generate final human priming maps (**Fig. S4**).

[Model-human comparisons] We briefly introduce the process of generating priming maps for computer vision models in **Sec. 4.3** and provide its pseudocode in **Algo. S2**.

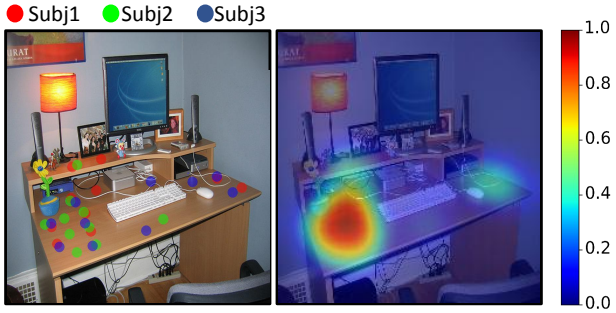


Figure S4. **Human priming map.** The left image shows the different mouse clicks made by 3 human subjects (colored dots) for *cup* as the target object. On the right, we show the corresponding human priming map from consolidated clicks. A higher density of clicks translates to a higher probability in the priming map. See the color bar for probability values.

We use 5 grid sizes to generate priming maps in different scales (8×8, 14×14, 28×28, 56×56, 112×112) and normalize these maps to obtain the final map. We provide more qualitative examples of model-human comparison in **Fig. S5**.

S3. Experiments, Ablations, and Analysis

S3.1. SeCo Enhances Object Recognition Abilities

In **Tab. 2**, we incorporated contextual information into the recognition task. Specifically, for the baseline methods, we trained a linear classifier ϕ_t on the top of the freezed backbone given cropped-out objects and corresponding labels from the COCO-OCD dataset. Then, we leverage linear classifiers ϕ_c trained in the lift-the-flap task to infer the target identity from the surrounding context of a given target object. We obtain the final prediction by multiplying the probabilities generated by ϕ_t and ϕ_c.

We break down the results according to the object sizes in **Fig. S6**. As observed, when the context-object ratio is larger than 2 on a logarithmic scale, incorporating contextual information learned with the lift-the-flap task constantly helps with recognizing smaller objects for all baselines (compare dotted line versus solid line). However, the effect of context impairs the recognition performance when the object is extremely small (the context-object ratio is less than 2). It is possible that the extremely small objects blend in the context and all recognition models fail to locate where the target objects are on the complex images.

S3.2. Analysis of Object Proposals Predicted by Selective Search

In **Tab. 4** we observed that SeCo pre-trained with selective search (SS) outperforms that with ground truth. To investigate how SS affects pre-training, we looked into both the quantity and quality of the proposals. Firstly, we scored each region proposal by IoU (intersection over union) against ground truth bounding boxes. We keep images in COCO-OCD containing at least 10 proposals with the IoU score larger than 0.3, which results in a dataset of 19.7K images. We use the following protocols to benchmark SeCo in terms of the quantity and quality of the proposals by SS.

[Quality]. We filtered out proposals according to an IoU threshold γ resulting in a proposal pool $\mathbf{B}_\gamma = \{b|IoU(b) > \gamma\}$. We keep the number of object proposals the same for every image I^i and only vary the IoU thresholds to study the quality of proposals. Specifically, we randomly selected 5 proposals from $\mathbf{B}^i_\gamma$ and varied $\gamma \in \{0, 0.1, 0.2, 0.3\}$. We applied the same training procedure described in **Sec. 3.5** and **Sec. S1**. We report the Top-1 accuracy in **Fig. 4(a), dash line**.

Table S1. **Baseline variations**. We tailored SimSiam and VICReg by prepending the object discovery module (SimSiam-SS and VICReg-SS). † denotes that original baselines use shared encoders. ‡ denotes that SeCo and all altered methods use selective search and non-shared encoders.

Method	#Param	Accuracy
SimSiam†	38M	42.46
SimSiam-SS‡	76M	45.45
VICReg†	175M	44.34
VICReg-SS‡	349M	52.70
VICReg-SS$_{\text{Tiny}}$‡	49M	40.95
SeCo‡	50M	52.43

[Quantity]. We fix the IoU threshold γ as 0. For each image I^i, we vary the number of proposals in $\{5, 10, 20, 30\}$ and randomly sample the proposals from $\mathbf{B}^i_{\gamma=0}$. After this, we applied the same training procedure described in **Sec. 3.5** and **Sec. S1**. We report the Top-1 accuracy in **Fig. 4 (b), solid line**.

We observe that in **Fig. 4 (a)**, raising the IoU threshold does not lead to the performance gain for SeCo. On the

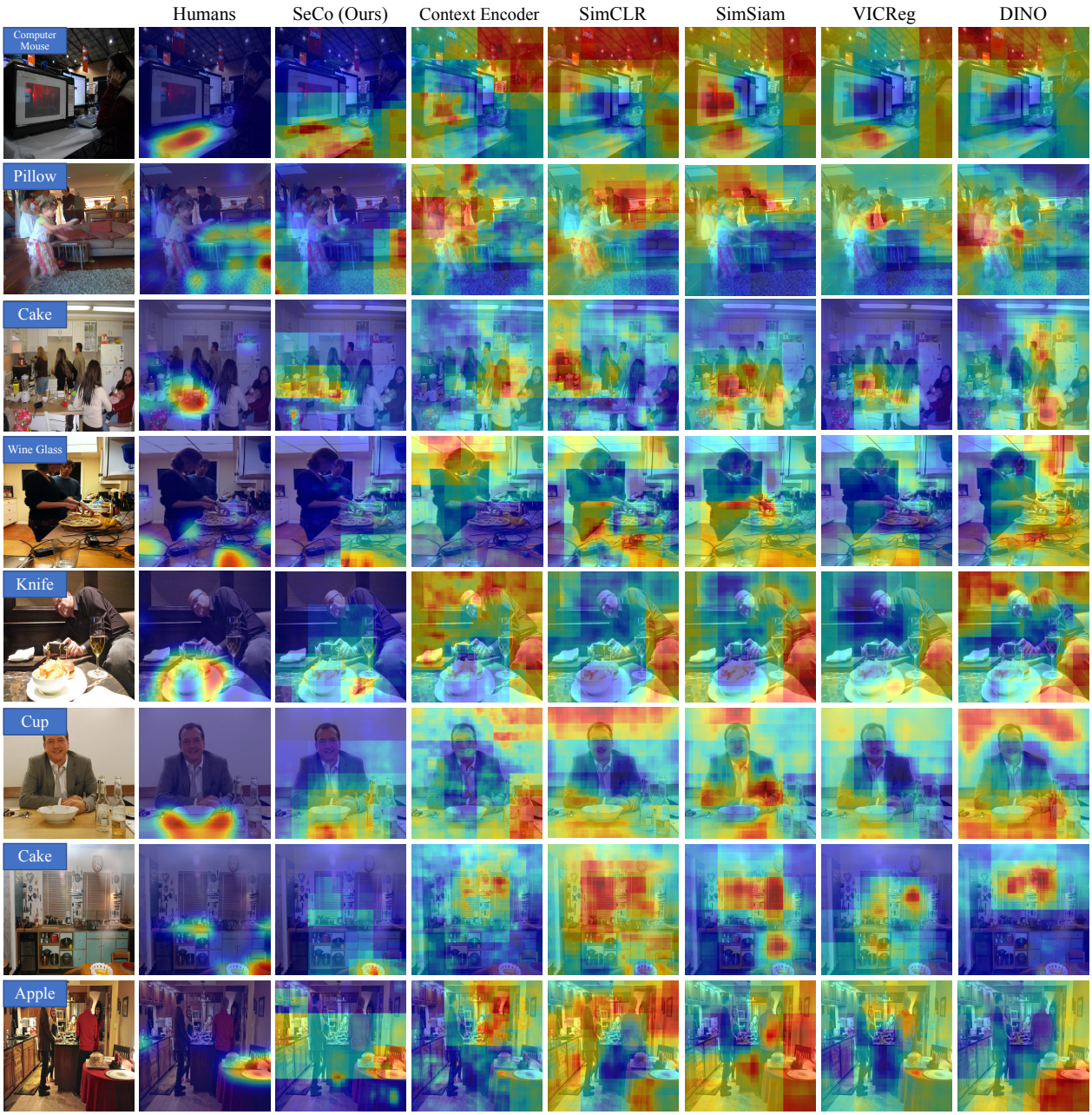


Figure S5. **SeCo priming maps highlight contextually relevant regions of the image and closely approximate human choices in the object priming task**. The leftmost column shows the input image and the given target object class label used for priming. The rest of the columns from left to right are priming maps from humans, predicted by our SeCo and predicted by all baselines. See **Fig. 3** for the color bar.

Algorithm S2 PyTorch-style pseudocode for generating priming maps.

```
# Ec:  trained context network with an
encoder and a linear classifier
# patch_sizes:  patch sizes when making
erased contexts

# load a batch of N images
for x, label in loader:

   maps = []

   # calculate priming maps in multiple
   scales
   for patch_size in patch_sizes:

      # iteratively erase a patch from
      image
      contexts = make_context(x, patch_size)

      # retrieve probability w.r.t location
      for a given object category
      p = softmax(Ec(x)[:,label])

      # normalize so that priming maps in
      different scales can add up
      p = (p - p.min()) / (p.max() -
      p.min())

      # upsample to the size of input image
      patch_num = x.size[1] // patch_size
      p = p.view((patch_num,patch_num))
      p = upsample(p)
      maps.append(p)


   # finalize priming maps by averaging and
   normalizing over different scales
   maps = torch.stack(maps).mean(0)
   maps = (maps - maps.min()) / (maps.max()
  - maps.min())
```

Algorithm S3 PyTorch-style pseudocode for calculating pairwise KL divergence of attention score over memory slots for object categories in COCO-VOC.

```
# Ec:  context encoders
# pc:  context projector
# M: external memory shaped in K-by-H
# F: frequency matrix shaped in C-by-K
# D: pair-wise KL-divergence matrix shaped
in C-by-C
# product:  cartesian product of two sets
# kld:  KL-divergence function

for x, label in loader:

   # obtain erased context
   c = erase(x)

   # encode and project context stream
   hc = Ec(x) # 1 x D
   sc = pc(hc) # 1 x H
   # compute keys of memory
   m = pk(M) # K x H

   # retrieve attention score over memory
   slots
   p = softmax(dot(sc, m))/sqrt(H) # 1 x K
   # sharpen the distribution
   top1 = p.max(0)[1]
   F[label, top1] += 1

# calculate pairwise KL-divergence
for i,j in product(range(C), range(C)):

   F[i] = (F[i] - F[i].min()) / (F[i].max()
   - F[i].min())
   F[j] = (F[j] - F[j].min()) / (F[j].max()
   - F[j].min())
   pi, pj = softmax(F[i]), softmax(F[j])
   D[i,j] = kld(pi, pj)
```

contrary, there is a slight increase in top-1 accuracy when we increase the number of proposals (**Fig. 4 (b)**). It indicates that the diversity of the proposals contributes more to the performance boost in SeCo+SS (**Tab. 4, I**) than the quality of the proposals in SeCo+GT (**Tab. 4, I**).

S3.3. Baseline Variations

We prepended object-discovery modules to feed "object" and "context" patches to SimSiam (Chen & He, 2021) and VICReg (Bardes et al., 2022) (SimSiam-SS and VICReg-SS). We also included the downsized VICReg with comparable network sizes as SeCo (VICReg-SS_{Tiny}). We visualize the architecture of SeCo, SimSiam, VICReg, and their altered versions in **Fig. S7**. We report top-1 accuracy on COCO-OCD in **Tab. S1**. As we observed, SeCo significantly surpasses VICReg-SS_{Tiny} and SimSiam-SS and performs competitively well as VICReg-SS although SeCo is 7 times smaller than VICReg-SS.

S3.4. Patch-Wise SSL

We compared our SeCo to existing patch-wise SSL methods, DenseCL (Wang et al., 2021) and ORL (Xie et al., 2021). Both methods rely on the augmented views of the same object instances from the same input image or the different object instances from similar contextual images. In contrast, our SeCo relies on the retrieved object representations from the learnable external memory and compares them against proposed regions. Thus, this enforces the context encoder of our SeCo to learn the context representations to retrieve the correct target object representations from the memory. The introduction to the external memory fundamentally changes the objectives from object-centric representation learning to object-context associative learning. For fair comparisons, we directly used the public checkpoints of DenseCL and

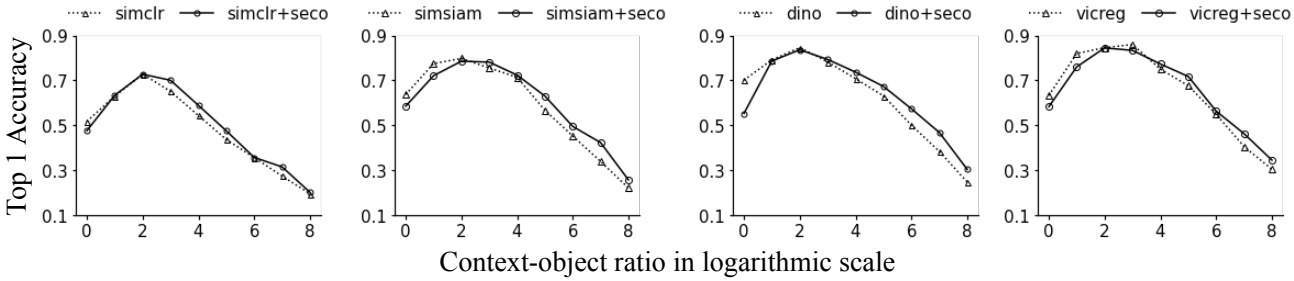
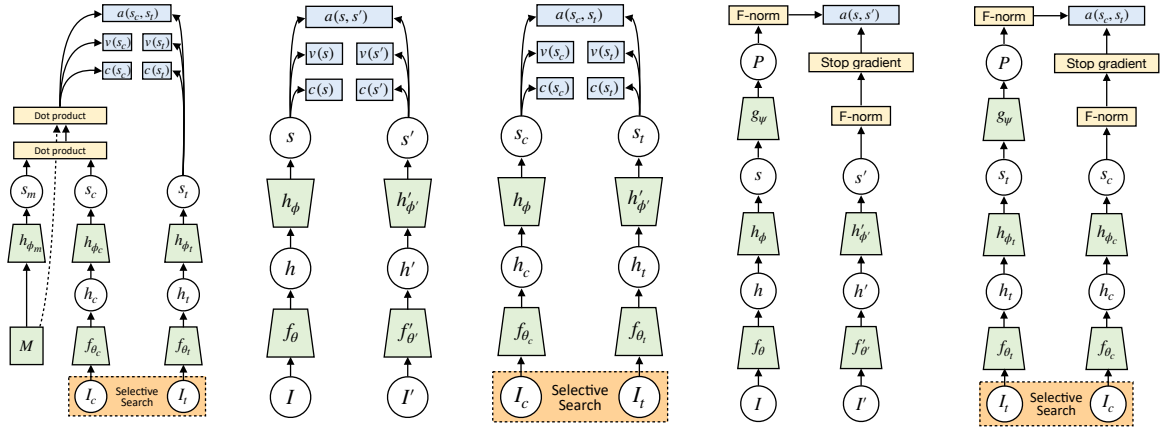


Figure S6. **Contextual cues improve recognition of small target objects.** We report the curves of Top 1 Accuracy on COCO-OCD versus context-object ratio in logarithmic scale.



(a) SeCo (b) VICReg (c) VICReg with object-discovery module (d) SimSiam (e) SimSiam with object-discovery module

Figure S7. **Architecture comparisons between SeCo, baselines, and their altered versions in Sec. S3.3** and **Tab. S1**. We use the same design conventions in (Bardes et al., 2022), where green blocks denote parametric functions, yellow boxes denote non-parametric functions, and blue boxes denote objective functions. In all methods, the input is either a pair of augmented views I and I' from the same image (b)(d), or a pair of context I_c and target I_t sampled from proposals generated by selective search (Uijlings et al., 2013) (a)(c)(e). The representations h are processed by a projector (narrowing trapezoid) to reduce the dimensionality (a)(d)(e) or an expander (widening trapezoid) to increase the dimensionality (b)(c). SeCo (a) applies learnable external memory M to store and retrieve contextual knowledge. The same variance, invariance, and covariance regularization objectives are applied on both branches as in VICReg (b)(c). SimSiam (d)(e) uses a predictor on one branch and the stop-gradient on another.

Table S2. The performance on the COCO-OCD In- & Out-of-Domain dataset in the lift-the-flap (Top-1 Accuracy) and Object Priming (RMSE).

Method	OCD-ID	OCD-OD	Object Priming
DenseCL	41.10	17.22	0.44
ORL	44.73	17.06	0.42
SeCo	**52.43**	**31.37**	**0.32**

ORL and compared them with SeCo on COCO-OCD in the lift-the-flap task.

S3.5. Analysis of External Memory Size

We also vary the number of memory slots (**Fig. S8**, left) from 100 to 800. There is a moderately positive increase of 2.5% in Top-1 accuracy in lift-the-flap. However, we observed a non-monotonic trend in Top-1 accuracy, when we vary the feature dimension of the external memory (**Fig. S8**, right). The top-1 accuracy peaks when the feature dimension equals 512. It suggests that larger memory capacity in general helps learn and store richer context-object associations; however, an overly large-sized memory may hurt context reasoning abilities, as the memory fails to generalize the learned contextual knowledge due to over-fitting.

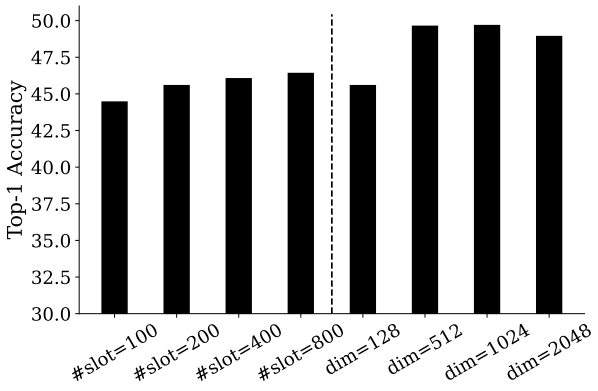


Figure S8. **Analysis of external memory of SeCo.** We report the top-1 accuracy for varying numbers of slots (left) and varying memory dimensionality per slot (right) in the lift-the-flap task.

Table S3. **Ablation study on loss components**. α, β, and γ are weightages of MSE loss, variance loss, and covariance loss respectively.

α	β	γ	Accuracy
25	25	1	**49.61**
1	1	0	47.72
0	25	1	41.72
25	0	1	collapse
1	0	0	collapse

S3.6. Analysis of Loss Components

SeCo has a joint loss of MSE loss, covariance loss, and variance loss. Here, we remove one loss at a time to analyze its effectiveness on pretraining. We report top-1 accuracy on COCO-OCD in **Tab. S3**. The result demonstrates that without variance loss, SeCo reached information collapse, aligning with the trend in VICReg (Bardes et al., 2022). Without covariance loss, performance drops 2% in accuracy. Different from the observations made in VICReg (Bardes et al., 2022), without MSE loss, SeCo manages to achieve 41.72% in accuracy without collapses. One possible reason is that starting from weights obtained on ImageNet, the encoder has captured useful visual features. Thus, adding information regularization during pre-training on COCO-OCD can avoid collapse even without enforcing association between contexts and targets.

S3.7. Probing External Memory

In the ablation study, we probe what the external memory has learned by visualizing the pairwise KL divergence of attention score over memory slots for object categories in COCO-VOC. Here, we provide the pseudocode of obtaining the matrix in **Algo. S3**.