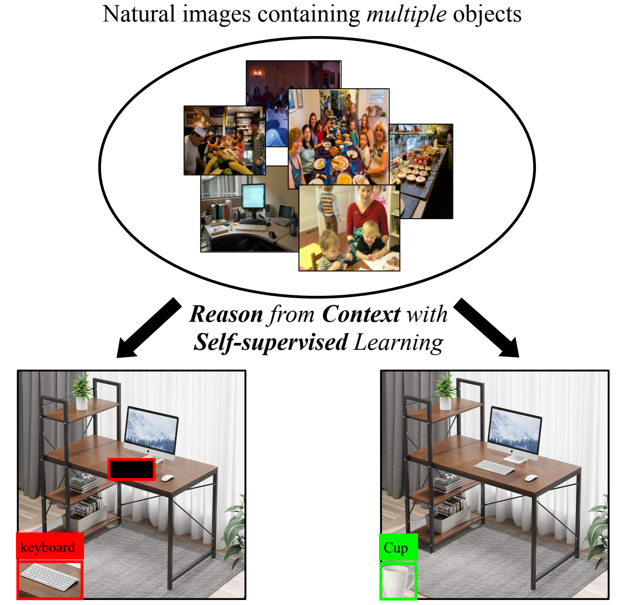
Reason from Context with Self-supervised Learning

Xiao Liu[1,2], Ankur Sikarwar[1,2], Joo Hwee Lim[1,2,3], Gabriel Kreiman[4,5], Zenglin Shi[1,2], Mengmi Zhang[1,2]

[1]I[2]R and [2]CFAR, Agency for Science, Technology and Research, Singapore,
[3]SCSE, Nanyang Technological University, Singapore
[4] Children's Hospital, Harvard Medical School, USA, [5] Center for Brains, Minds and Machines, USA
Address correspondence to mengmi@i2r.a-star.edu.sg

Abstract

*A tiny object in the sky cannot be an elephant. Context reasoning is critical in visual recognition, where current inputs need to be interpreted in the light of previous experience and knowledge. To date, research into contextual reasoning in visual recognition has largely proceeded with supervised learning methods. The question of whether contextual knowledge can be captured with self-supervised learning regimes remains under-explored. Here, we established a methodology for context-aware self-supervised learning. We proposed a novel **Se**lf-supervised Learning Method for **Co**ntext Reasoning (SeCo), where the only inputs to SeCo are unlabeled images with multiple objects present in natural scenes. Similar to the distinction between fovea and periphery in human vision, SeCo processes self-proposed target object regions and their contexts separately, and then employs a learnable external memory for retrieving and updating context-relevant target information. To evaluate the contextual associations learned by the computational models, we introduced two evaluation protocols, lift-the-flap and object priming, addressing the problems of "what" and "where" in context reasoning. In both tasks, SeCo outperformed all state-of-the-art (SOTA) self-supervised learning methods by a significant margin. Our network analysis revealed that the external memory in SeCo learns to store prior contextual knowledge, facilitating target identity inference in lift-the-flap task. Moreover, we conducted psychophysics experiments and introduced a **H**uman benchmark in **O**bject **P**riming dataset (HOP). Our quantitative and qualitative results demonstrate that SeCo approximates human-level performance and exhibits human-like behavior. All our source code and data are publicly available here.*


Natural images containing *multiple* objects
***Reason** from **Context** with*
***Self-supervised** Learning*
keyboard
Cup

(a) lift-the-flap

(b) object priming

Figure 1. **Schematic illustration of learning to reason via self-supervised training on natural images with multiple objects present in the scene**. Two protocols are introduced to evaluate contextual reasoning ability: (a) Lift-the-flap task (left branch) and (b) Object priming task (right branch). In (a), the task is to reason about the scene context and infer what a target object hidden behind a flap (black patch) is in a natural image. The original image (bottom left) reveals the target object (”keyboard”) which was not shown in the actual experiments. In (b), the task is to decide where to put the given object in the scene, e.g. to put the cup (at the bottom left) in the study room scene.

1. Introduction

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 hidden object on the table can be a keyboard or a book but never a private jet. To establish these kinds of associations between a hidden or occluded object and scene context, humans rely on prior knowledge and experiences with various objects and their mutual relationships.

To date, context reasoning capacity has been studied with supervised learning methods [4, 60]. However, the question of whether contextual knowledge can be captured in a self-supervised way remains under-explored. In the self-supervised learning (SSL) literature, most works [2, 8, 9, 11, 13, 26, 27, 47, 59] focus on learning image-level representations by pre-training neural networks on natural images, such as ImageNet [16], where objects of interest are monotonously large, salient, and centered. In contrast, natural scenes always contain multiple objects with complex context. Recent research [40, 56] draws attention to instance-level pre-training on natural images containing multiple objects in the scene, such as COCO [39] datasets. Yet, there are still missing links in associating scene-level representations with instance-level representations in a scene. Studies on establishing these missing links would unleash the reasoning capabilities of SSL methods in complex context.

As an initial effort in this direction, we established a methodology to study 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 evaluate the context reasoning capabilities of all computational models, we introduced two evaluation protocols, lift-the-flap and object priming, addressing the problems of "what" and "where" in context reasoning. Specifically, 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 given 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.

To tackle these problems, 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 interests. Next, the target object of interest and its surrounding context are processed separately by two independent image encoders. To store learned contextual priors, we introduced a learnable external memory in SeCo. We gain insights about the role of our external memory from intensive network analysis. We stress-tested SeCo and SOTA SSL methods on in-domain 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 on the same dataset we used for testing the models. Our results suggest that SeCo achieves human-level performance and exhibits human-like behaviors.

We summarize our key contributions below:

- We established a methodology for the community to study context-aware SSL. We introduced lift-the-flap and object priming protocols to benchmark contextual reasoning ability of current and future SSL methods.
- We proposed a novel SSL method (SeCo) to learn associations between objects and their context. SeCo beats SOTA SSL methods by a large margin on in-domain and out-of-domain test sets of three datasets in lift-the-flap and object priming tasks.
- We contributed a new object priming dataset (HOP) and human benchmarks on HOP with psychophysics experiments. Our achieves human-level performance and exhibits human-like behaviors.

2. Related works

2.1. Role of context in computer vision

The context of a scene can be described in various ways, including global scene context [52], geometric context [31], relative location [18], 3D layout [38] and spatial support and geographical information [19]. The ability to reason about objects and relations in context is crucial to computer vision, such as object recognition [4, 60], place recognition [54], scene recognition [24, 35, 57], object detection [14, 41], semantic segmentation [44], and visual question answering [49]. To tackle the problem of context-aware computer vision tasks, statistical optimization tools [10, 24, 35, 57], graph neural networks [3, 15, 17, 32], and transformer-based methods [4, 6] are proposed in the literature. Breaking away from all these previous works which require supervised training on labelled images, we investigated the problem of context reasoning in the SSL setting.

2.2. Self-supervised learning

Since supervised learning requires ground truth labels which are labor-intensive and costly to collect, SSL has recently become popular. To improve the quality of learned object representations, various handcrafted pretext tasks have been designed [36, 45–47]. Another group of works [11, 28, 30, 43, 55, 58] highlight the advantage of contrastive learning in SSL by pulling positive samples together and pushing negative samples away. Several non-contrastive methods are proposed to learn image-level representations solely based on positive samples [2, 9, 13, 26, 51, 59]. With the success of transformer-based models in NLP and vision tasks [20], there has also been a trend in SSL of reconstructing images from randomly masked image patches [12,27]. However, all these previous methods focus on learning image-level representations from monotonously large, salient, and centered objects. They often fail to capture instance-level association in the 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.

Another line of research in SSL is deep clustering methods [1, 7, 8, 33, 37, 61]. Several methods [8, 37] use external memories to store trainable object prototypes and use them to assign similar images to distinct clusters. In contrast to these models with external memories, our external memory specifically addresses the problem of context reasoning as it serves as a memory buffer storing prior knowledge on object-context associations and flexibly retrieves useful object information from context cues in the visual scenes.

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) target 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 the 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 (**Sec. 3.1**). Second, the two-stream visual processor consists of two independent pairs of CNN encoders and projectors, extracting information from I_t and I_c respectively (**Sec. 3.2**). Third, SeCo employs a trainable external memory (**Sec. 3.3**) to store knowledge priors about target-context associations learned during the training phase (**Sec. 3.4**). Features from I_c serves as queries to retrieve context-relevant prior knowledge from the external memorywith an attention mechanism. The retrieved information provides 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. Refer to **Supp.** for PyTorch-style pseudocode of SeCo's training algorithm.

3.1. Context-Object Pair Discovery

Objects play an important role in context reasoning [21]. 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 [53] to generate regions of interest (RoI) that potentially contain objects. It is worth noting that the 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 maximum as 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 image 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 zero pixels, forms the context image I_c.

3.2. Feature Extraction with CNN

Due to the eccentricity dependence, human vision has the highest acuity at the fovea and the resolution drops sharply in the far periphery with increasing eccentricity. For example, while we are fixating at 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. Seeking inspirations from this, we propose a two-stream visual processor, with one object stream dedicated at encoding the target image I_t and the other context stream dedicated at 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 respectively and h_t and $h_c \in \mathbb{R}^D$. Assuming that the features useful for context reasoning and object recognition are different, we do not enforce weights sharing between the encoders. We demonstrate its benefit in **Sec. 5.3**.

3.3. Training With An External Memory

As suggested by cognitive and neuroscience works [48, 50, 60], the context processing often happens very fast in the brains. The perceived scene gist serves as queries to retrieve prior knowledge from the semantic memory to module object recognition in a top-down manner. To mimic this underlying mechanism of context modulation in the biological brains, we introduce an external memory with trainable parameters, accumulating prior knowledge of context-object associations. Here, we introduce math notations and formulations of our memory mechanism in knowledge retrieval. See the subsection below for how prior knowledge are learned with self-supervised losses.

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

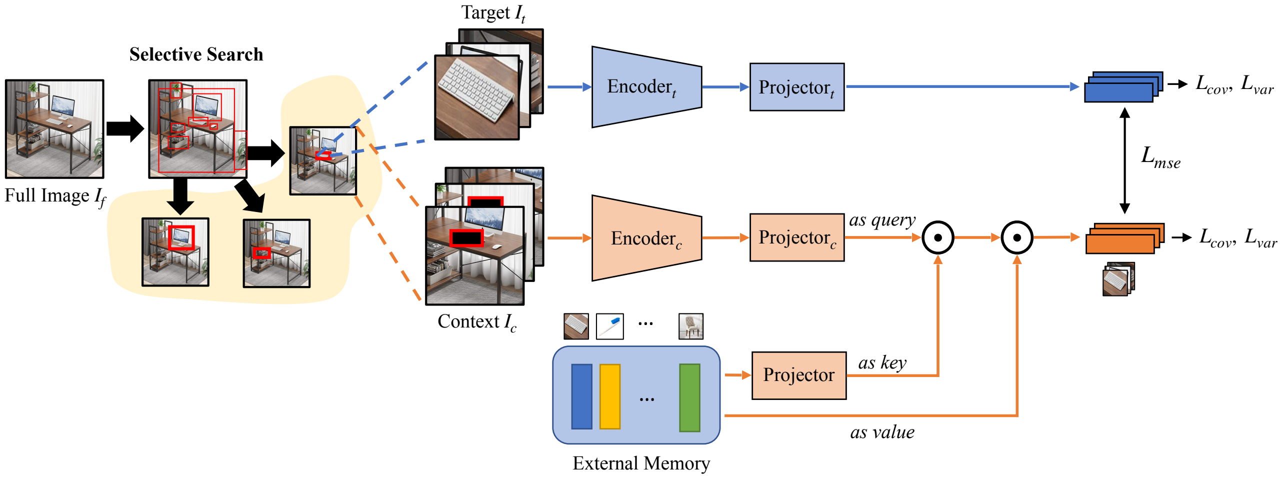


Figure 2. **The architecture overview of our Self-supervised learning for Context reasoning (SeCo).** The architecture is comprised of three components: target discovery module, two-stream visual processor, and external memory. SeCo uses an unsupervised method to discover potential targets. The discovered targets are then converted to multiple context-object pairs (**Sec.** 3.1). SeCo leverages non-shared networks to encode targets and contexts separately (**Sec.** 3.2). In addition, SeCo uses a trainable external memory to store the contextual priors learned during the training phase, which can then be used as a complementary signal during inference (**Sec.** 3.3). Finally, we use a joint loss which maximizes agreement between target and context with L_{mse} in the embedding space and regularizes the learned representations with L_{cov} and L_{var}, increasing diversity (**Sec.** 3.4).

$\phi_k(\cdot) : \mathbb{R}^H \rightarrow \mathbb{R}^H$ 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 [2], maximizing the agreement between two-stream visual processors alone may lead to model collapses. For example, the external memory stores and outputs trivial knowledge of all zeros, while the visual processor trivially encodes all the images to representations of all zeros. In this case, the agreement between s_c and s_t aligns perfectly; however, the encoded object representations and the content in M are meaningless.

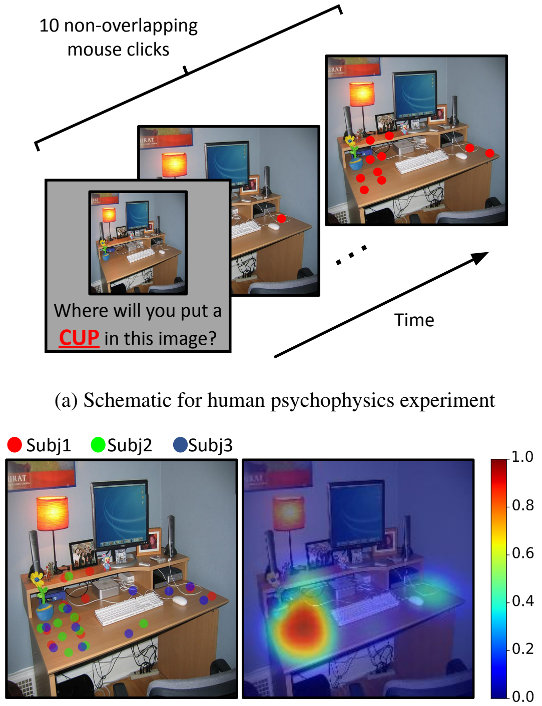
Thus, to prevent model collapses, we enforce covariance L_{cov} and variance L_{var} regularization losses on both object and context streams respectively. While L_{var} maintains the variance of different representations of data samples within a batch (e.g. a batch of images should represent diversified object classes), L_{cov} de-correlates channel-wise variables to diversify attributes of an embedding (e.g. use as many attributes as possible to represent objects and each attribute should be independent from each other). Our 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 **Supp.** 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, 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.


10 non-overlapping mouse clicks
Where will you put a CUP in this image?
Time
Subj1
Subj2
Subj3
1.0
0.8
0.6
0.4
0.2
0.0

(a) Schematic for human psychophysics experiment

(b) Visualization for human priming maps

Figure 3. **Human psychophysics experiments in object priming task.** (a) Subjects were presented with a natural image along with a target object and were asked to put the object at appropriate locations by making 10 non-overlapping mouse clicks (red dots). (b) 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. Higher density of clicks translates to higher probability in the priming map. See the colorbar for probability values.

Network architecture. We use ResNet-50 [29] with $D = 2048$ output units as our encoders $E_c(\cdot)$ and $E_t(\cdot)$. We set the size of M as $K \times H = 200 \times 512$ and initialize M by the Xavier uniform initializer [23]. We demonstrate the benefit of the external memory and vary its sizes in the ablation study (**Sec. 5.3**). For fair comparison with all baselines (**Sec. 4.2**) which are pre-trained on ImageNet [16], we initialize our encoders with weights pre-trained on ImageNet by VICReg [2].

Training. We set the base learning rate to $lr = 0.2 * \text{batch\_size}/256$ [25]. The learning rate grows linearly from 0 to base value during first 10 epochs and then decays with a cosine scheduler [42] for the rest epochs with a minimum value of 0.0002.

4. Experiments

4.1. Datasets

COCO-Stuff Dataset [5] contains 160K natural images from MSCOCO dataset [39] with 80 thing classes (e.g. car, person) and 91 stuff classes (e.g. grass, sky) in total. Importantly, this dataset captures complex relationships between multiple objects, and their contexts.

PASCAL VOC07 Dataset [22] contains 9,963 images of realistic scenes with total 20 object classes.

Out-of-Context Dataset (OCD) [4] 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 come up with two custom regimes on pretext training, fine-tuning, and testing:

COCO-VOC contains those images from COCO-Stuff where the object classes present in the scene overlap with the object classes from PASCAL VOC07 dataset. Overall, 20 classes overlap between COCO-Stuff and PASCAL VOC07 dataset. Refer to the **Supp.** for the list of selected 20 classes. We used the training set of COCO-VOC for pre-text training and fine-tuning and then tested all the models on the test set of COCO-VOC (in-domain) and PASCAL VOC07 dataset (out-of-domain).

COCO-OCD includes those images from COCO-Stuff where the object classes in the scene overlap with the object classes from OCD dataset. Here, in total 15 classes overlap between COCO-Stuff and OCD dataset. Refer to the **Supp.** for the list of selected 15 classes. We used the training set of COCO-OCD for pretext training and fine-tuning and then tested all the models on the test set of COCO-OCD (in-domain) and OCD dataset (out-of-domain).

4.2. Baselines

We compare our SeCo against a list of SSL methods introduced below: Context Encoder [47], SimCLR [11], SimSiam [13], DINO [9], and VICReg [2]. Context Encoder reconstructs randomly masked image regions to learn representations. SimCLR learns useful visual features using contrastive learning, while SimSiam, DINO, and VICReg use non-contrastive methods for representation learning. For Context Encoder, since it was originally trained with AlexNet [34], we re-implement it with the standard ResNet-50 backbone for fair comparison with SeCo and other baselines. We use the default settings for the remaining baselines. See **Supp.** for further details.

4.3. Evaluation Protocols for Context-aware SSL

Lift-the-Flap. We introduce 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 model trained with SSL to this task, we freeze the model weights for feature extraction and then only fine-tune a linear classifier to output predicted class labels for the hidden target object. We report the performance in Top-1 accuracy of all baselines 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 the model 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 designs] We curated semantically relevant 864 unique image-object pairs on 206 images from the test set of MSCOCO-OCD dataset (**Sec. 4.1**). Eggs tend to be nearby other eggs. To avoid this "crowding" effect that could bias humans and models in placing same target objects in the same locations, for each image-object pair in object priming, we made sure that there are no object instances present on the context image whereby these object instances belong to the same class as the given target object. See **Supp.** for details about selecting these image-object pairs.

[Human response collection] We followed standard Institutional Review Board protocols and used Amazon Mechanical Turk (AMT) to collect responses from total 437 human subjects. For quality controls, we only recruited participants with *master* qualification and a minimum of 95% approval rate. For each subject, we randomly sample 20 image-object pairs and present the 800×800 image along with the question "Where would you put this *[obj]*?" where *[obj]* corresponds to the sampled target object. The subjects are required to make non-repeated 10 mouse clicks at relevant regions of the given image. For each image-object pair, we collect responses from exactly 3 human subjects which gives us 30 unique clicks in total per image-object pair. We show the schematic for the human psychophysics experiment in **Fig. 3a** (see **Supp.** for AMT interface and further details).

[Post-processing] For each image, we consolidated all 30 click coordinates and generated the attention map of size 25×25, where the intensity at each location of the attention map denotes the click counts. We then applied Gaussian smoothing, upscaling, and min-max normalization on these maps to generate the final human priming maps (**Fig. 3b**). See **Supp.** for further details.

[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 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 **Supp.** for details). We compare 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. 2**.

	Method	**In Domain**	**Out Of Domain**
COCO-VOC	Context Encoder [47]	15.78	14.82
	SimCLR [11]	32.78	37.65
	SimSiam [13]	39.79	45.76
	DINO [9]	42.06	48.07
	VICReg [2]	44.89	52.58
	SeCo (Ours)	**52.31**	**57.27**
COCO-OCD	Context Encoder [47]	20.55	10.68
	SimCLR [11]	35.78	15.51
	SimSiam [13]	42.46	19.36
	DINO [9]	43.21	15.34
	VICReg [2]	44.34	24.31
	SeCo (Ours)	**52.43**	**31.37**

Table 1. **SeCo outperforms all baselines in lift-the-flap task.** We test all the SSL methods (**Sec. 4.2**) on in-domain and out-of-domain images over 3 datasets (**Sec. 4.1**) and report top-1 accuracy averaged over 5 runs (**Sec. 4.3**).

5. Results

5.1. Lift-the-flap task

We report the top-1 target inference accuracy of all models in 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. Context Encoder [47] 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 contextual associations, such as object co-occurrences. Next, we observed that contrastive methods like SimCLR [11] performed worse compared with non-contrastive methods like SimSiam [13], DINO [9], and VICReg [2]. It suggests that multiple objects could co-occur in the same context and making selection of negative samples is non-trivial and challenging in context-aware SSL.

Bird flies in the sky regardless of whether the scene is depicted in Picasso or Monet styles. Contextual associations should be invariant to domain shifts of visual

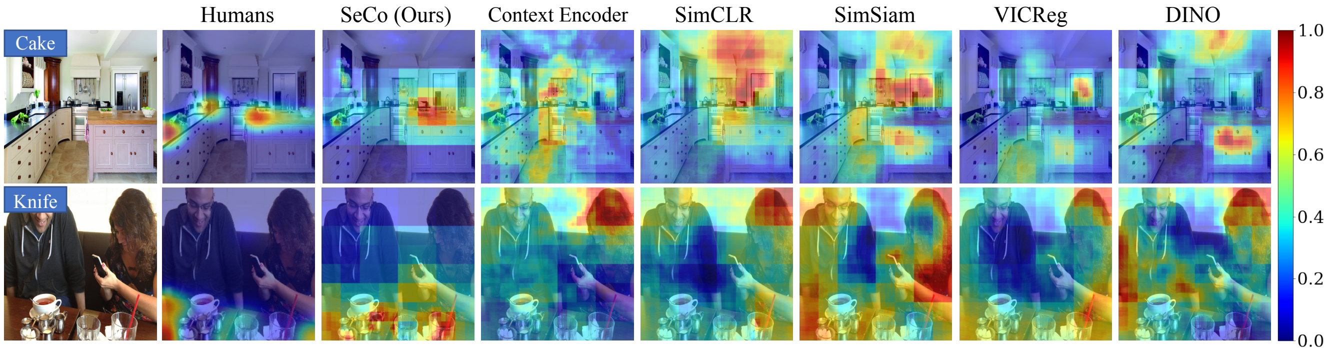


Figure 4. **SeCo priming maps highlight contextually relevant regions of the image and closely approximate human choices in the object priming task**. The leftmost column shows the input image and the given target object class label used for priming. The rest of the columns from left to right are priming maps from humans, predicted by our SeCo and predicted by all baselines (**Sec.** 4.2).

Method	RMSE
Human Agreement	0.17 ± 0.05
Context Encoder [47]	0.41 ± 0.06
SimCLR [11]	0.44 ± 0.07
SimSiam [13]	0.43 ± 0.09
DINO [9]	0.42 ± 0.07
VICReg [2]	0.40 ± 0.09
SeCo (Ours)	**0.32 ± 0.06**

Table 2. **Root mean square error (RMSE) between human priming maps and maps predicted by computational models in object priming task**. Lower is better. Error bars show standard deviation calculated across samples. RMSE for human agreement was calculated by comparing priming maps across the 3 human subjects for individual image-object pairs.

features. We test all models in out-of-domain datasets, PASCAL VOC07 and OCD. 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, when the domain shifts from natural images in COCO-OCD to the synthetic images in OCD, we saw a big performance drop for all the models. Yet, our model gets less impaired due to domain shifts, highlighting that our SeCo learns context associations rather than correlations of visual features.

5.2. Object priming task

We compare human priming maps with the maps predicted by all models and report RMSE scores in **Tab.** 2. As a lower 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 and human agreements in object priming.

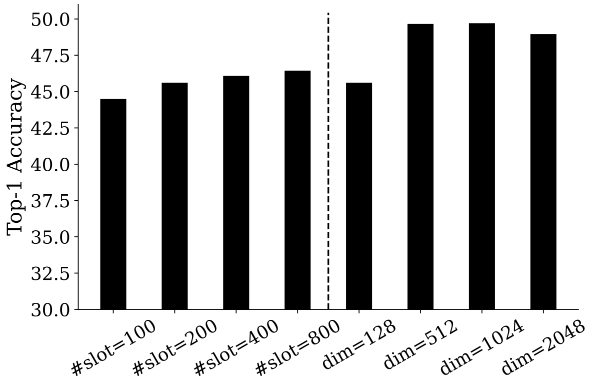


Figure 5. **Analysis on external memory of SeCo.** We report the top-1 accuracy for varying number of slots (left) and varying memory dimensionality per slot (right) in the lift-the-flap task.

To assess the quality of the predicted priming maps by all models, we also visually examined qualitative examples (**Fig.** 4). 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 the 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 context-object 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, we replaced the object-context image pairs generated by selective search [53] with randomly generated object-context image pairs (**Tab.** 3, RG). There is a

	Ablations	Accuracy
I	SS	**52.43**
	GT	49.61
	RG	36.95
II	NSA	**49.61**
	SA	37.48
III	w/ Mem.	**49.61**
	w/o Mem.	44.07

Table 3. **Ablation Study**. Top-1 accuracy in lift-the-flap on COCO-OCD of (I) different RoI generation strategies: selective search (SS), annotated bounding boxes (GT), random generation of boxes (RG); (II) different architecture: non-shared architectures of two-stream visual processors (NSA), shared architecture (SA); and (III) with or without the external memory. In experiments (II) and (III), we directly use annotated bounding boxes as RoIs (GT) to remove randomness in selecting target objects for better controls. Default settings of our SeCo are highlighted.

decrease of 16% in top-1 accuracy, highlighting 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.** 3, GT). Surprisingly, SeCo performs better with GT by 3%. It is possible that the exploited RoIs by the selective search contain small objects which are hardly labelled by human annotators but useful for context-aware SSL.

Next, we trained two separate encoders $E_t(\cdot)$ and $E_c(\cdot)$ in SeCo (**Sec.** 3.2). Here, we enforced weight-sharing encoders (**Tab.** 3, SA). SA achieved a lower top-1 accuracy of 37.48% 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.** 3, w/o Mem). The performance of w/o Mem drops by around 5% on average over all object sizes, demonstrating that the external memory enhances the reasoning ability of SeCo.

We also vary the number of memory slots (**Fig.** 5, 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 non-monotonic trend in Top-1 accuracy, when we vary the feature dimension of the external memory (**Fig.** 5, 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.

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 (**Sec.** 4.1). 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 **Supp.** 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 but they are contextually relevant. This suggests that the external memory in SeCo learns meaningful object-context associations.

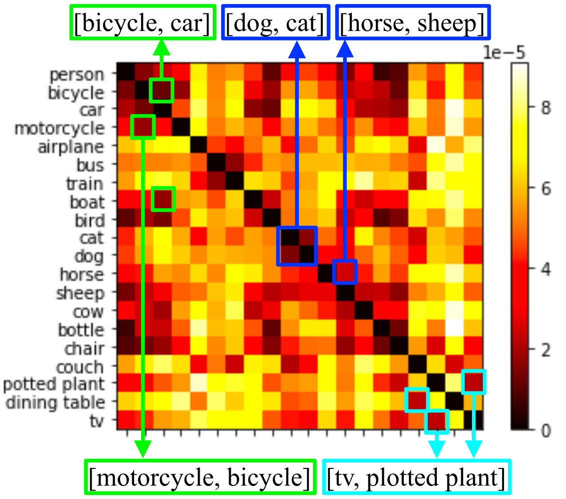


Figure 6. **Pairwise KL divergence of attention scores over memory slots of the external memory in SeCo for object categories in COCO-VOC**. Dark grids show that targets sharing similar context in both categories retrieve information from similar memory slots. Colored boxes pointed by arrows denote different supercategories in PASCAL VOC07, such as *vehicle*, *animal*, *indoor*.

6. Conclusions

We set out to determine whether SSL methods can capture the statistics of associations in natural images. To this end, we introduced SeCo, a novel self-supervised learning method for context reasoning, which learns object-context associations from unlabeled images. Like humans, while learning, SeCo relies on external memory to develop priors through repeated encounters with objects and their contexts, which it subsequently uses for reasoning by retrieving information from these knowledge priors. To evaluate the contextual associations learned by all models, we introduced two testing protocols, Lift-the-Flap and Object Priming. In addition, we conducted human psychophysics experiments and introduced HOP, a human benchmark dataset for contextual reasoning. We then used it to quantitatively and qualitatively evaluate different models on the object priming task. Finally, we performed a series of ablations and analytic experiments to assess the relevance of different components of our model. Our work provides new insights into how to perform contextual reasoning via self-supervised learning.

References

[1] YM Asano, C Rupprecht, and A Vedaldi. Self-labelling via simultaneous clustering and representation learning. In *International Conference on Learning Representations*, 2019. 3

[2] Adrien Bardes, Jean Ponce, and Yann Lecun. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. In *ICLR 2022-10th International Conference on Learning Representations*, 2022. 2, 4, 5, 6, 7, 17

[3] Peter Battaglia, Razvan Pascanu, Matthew Lai, Danilo Jimenez Rezende, et al. Interaction networks for learning about objects, relations and physics. *Advances in neural information processing systems*, 29, 2016. 2

[4] Philipp Bomatter, Mengmi Zhang, Dimitar Karev, Spandan Madan, Claire Tseng, and Gabriel Kreiman. When pigs fly: Contextual reasoning in synthetic and natural scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 255–264, 2021. 2, 5, 13

[5] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Coco-stuff: Thing and stuff classes in context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1209–1218, 2018. 5

[6] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. 2

[7] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *Proceedings of the European conference on computer vision (ECCV)*, pages 132–149, 2018. 3

[8] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in Neural Information Processing Systems*, 33:9912–9924, 2020. 2, 3

[9] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9650–9660, 2021. 2, 5, 6, 7

[10] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017. 2

[11] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 2, 5, 6, 7

[12] Xiaokang Chen, Mingyu Ding, Xiaodi Wang, Ying Xin, Shentong Mo, Yunhao Wang, Shumin Han, Ping Luo, Gang Zeng, and Jingdong Wang. Context autoencoder for self-supervised representation learning. *arXiv preprint arXiv:2202.03026*, 2022. 3

[13] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15750–15758, 2021. 2, 5, 6, 7

[14] Xinlei Chen, Li-Jia Li, Li Fei-Fei, and Abhinav Gupta. Iterative visual reasoning beyond convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7239–7248, 2018. 2

[15] Wongun Choi and Silvio Savarese. A unified framework for multi-target tracking and collective activity recognition. In *European Conference on Computer Vision*, pages 215–230. Springer, 2012. 2

[16] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 2, 5, 14

[17] Zhiwei Deng, Arash Vahdat, Hexiang Hu, and Greg Mori. Structure inference machines: Recurrent neural networks for analyzing relations in group activity recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4772–4781, 2016. 2

[18] Chaitanya Desai, Deva Ramanan, and Charless C Fowlkes. Discriminative models for multi-class object layout. *International journal of computer vision*, 95(1):1–12, 2011. 2

[19] Santosh K Divvala, Derek Hoiem, James H Hays, Alexei A Efros, and Martial Hebert. An empirical study of context in object detection. In *2009 IEEE Conference on computer vision and Pattern Recognition*, pages 1271–1278. IEEE, 2009. 2

[20] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2020. 2

[21] Dejan Draschkow and Melissa L-H Võ. Scene grammar shapes the way we interact with objects, strengthens memories, and speeds search. *Scientific reports*, 7(1):1–12, 2017. 3

[22] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010. 5, 13

[23] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on*

artificial intelligence and statistics, pages 249–256. JMLR Workshop and Conference Proceedings, 2010. 5

[24] Josep M Gonfaus, Xavier Boix, Joost Van de Weijer, Andrew D Bagdanov, Joan Serrat, and Jordi Gonzalez. Harmony potentials for joint classification and segmentation. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 3280–3287. IEEE, 2010. 2

[25] Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large minibatch sgd: Training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*, 2017. 5

[26] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020. 2

[27] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022. 2, 3

[28] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020. 2

[29] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5, 12, 14

[30] R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. In *International Conference on Learning Representations*, 2018. 2

[31] Derek Hoiem, Alexei A Efros, and Martial Hebert. Geometric context from a single image. In *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, volume 1, pages 654–661. IEEE, 2005. 2

[32] Hexiang Hu, Guang-Tong Zhou, Zhiwei Deng, Zicheng Liao, and Greg Mori. Learning structured inference neural networks with label relations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2960–2968, 2016.

[33] Jiabo Huang, Qi Dong, Shaogang Gong, and Xiatian Zhu. Unsupervised deep learning by neighbourhood discovery. In *International Conference on Machine Learning*, pages 2849–2858. PMLR, 2019. 3

[34] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, volume 25, 2012. 5

[35] Lubor Ladicky, Chris Russell, Pushmeet Kohli, and Philip HS Torr. Graph cut based inference with co-occurrence statistics. In *European conference on computer vision*, pages 239–253. Springer, 2010. 2

[36] Gustav Larsson, Michael Maire, and Gregory Shakhnarovich. Learning representations for automatic colorization. In *European conference on computer vision*, pages 577–593. Springer, 2016. 2

[37] Junnan Li, Pan Zhou, Caiming Xiong, and Steven Hoi. Prototypical contrastive learning of unsupervised representations. In *International Conference on Learning Representations*, 2020. 3

[38] Dahua Lin, Sanja Fidler, and Raquel Urtasun. Holistic scene understanding for 3d object detection with rgbd cameras. In *Proceedings of the IEEE international conference on computer vision*, pages 1417–1424, 2013. 2

[39] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 2, 5

[40] Songtao Liu, Zeming Li, and Jian Sun. Self-emd: Self-supervised object detection without imagenet. *arXiv preprint arXiv:2011.13677*, 2020. 2

[41] Yong Liu, Ruiping Wang, Shiguang Shan, and Xilin Chen. 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*, pages 6985–6994, 2018. 2

[42] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016. 5

[43] Ishan Misra and Laurens van der Maaten. Self-supervised learning of pretext-invariant representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6707–6717, 2020. 2

[44] Roozbeh Mottaghi, Xianjie Chen, Xiaobai Liu, Nam-Gyu Cho, Seong-Whan Lee, Sanja Fidler, Raquel Urtasun, and Alan Yuille. The role of context for object detection and semantic segmentation in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 891–898, 2014. 2

[45] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European conference on computer vision*, pages 69–84. Springer, 2016. 2

[46] Mehdi Noroozi, Hamed Pirsiavash, and Paolo Favaro. Representation learning by learning to count. In *Proceedings of the IEEE international conference on computer vision*, pages 5898–5906, 2017. 2

[47] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2536–2544, 2016. 2, 5, 6, 7, 12, 14

[48] Maximilian Riesenhuber and Tomaso Poggio. Hierarchical models of object recognition in cortex. *Nature neuroscience*, 2(11):1019–1025, 1999. 3

[49] Damien Teney, Lingqiao Liu, and Anton van Den Hengel. Graph-structured representations for visual question

answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2017. 2

[50] Simon Thorpe, Denis Fize, and Catherine Marlot. Speed of processing in the human visual system. *nature*, 381(6582):520–522, 1996. 3

[51] Yuandong Tian, Xinlei Chen, and Surya Ganguli. Understanding self-supervised learning dynamics without contrastive pairs. In *International Conference on Machine Learning*, pages 10268–10278. PMLR, 2021. 2

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

[53] Jasper RR Uijlings, Koen EA Van De Sande, Theo Gevers, and Arnold WM Smeulders. Selective search for object recognition. *International journal of computer vision*, 104(2):154–171, 2013. 3, 7

[54] Kevin Wu, Eric Wu, and Gabriel Kreiman. Learning scene gist with convolutional neural networks to improve object recognition. In *2018 52nd Annual Conference on Information Sciences and Systems (CISS)*, pages 1–6. IEEE, 2018. 2

[55] Enze Xie, Jian Ding, Wenhai Wang, Xiaohang Zhan, Hang Xu, Peize Sun, Zhenguo Li, and Ping Luo. Detco: Unsupervised contrastive learning for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8392–8401, 2021. 2

[56] Jiahao Xie, Xiaohang Zhan, Ziwei Liu, Yew Soon Ong, and Chen Change Loy. Unsupervised object-level representation learning from scene images. *Advances in Neural Information Processing Systems*, 34:28864–28876, 2021. 2

[57] Jian Yao, Sanja Fidler, and Raquel Urtasun. Describing the scene as a whole: Joint object detection, scene classification and semantic segmentation. In *2012 IEEE conference on computer vision and pattern recognition*, pages 702–709. IEEE, 2012. 2

[58] Mang Ye, Xu Zhang, Pong C Yuen, and Shih-Fu Chang. Unsupervised embedding learning via invariant and spreading instance feature. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6210–6219, 2019. 2

[59] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *International Conference on Machine Learning*, pages 12310–12320. PMLR, 2021. 2

[60] Mengmi Zhang, Claire Tseng, and Gabriel Kreiman. Putting visual object recognition in context. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12985–12994, 2020. 2, 3

[61] Chengxu Zhuang, Alex Lin Zhai, and Daniel Yamins. Local aggregation for unsupervised learning of visual embeddings. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6002–6012, 2019. 3

List of Supplementary Sections

List of Supplementary Figures

List of Tables

Algorithm 1: 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()
```

S1. Method

We provide PyTorch-style pseudocode for SeCo in **Algo. 1**. 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.

S2. Experiments

S2.1. Datasets

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 in **Sec. 4.1**. Overlapped classes are as follows:

COCO-VOC contains the same 20 classes in hierarchy of *superclass* and subclass as defined in PASCAL VOC07 [22].

- *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 [4]: wine glass, cup, knife, bowl, apple, cake, mouse, remote, keyboard, cell phone, microwave, book, toothbrush, pillow, towel.

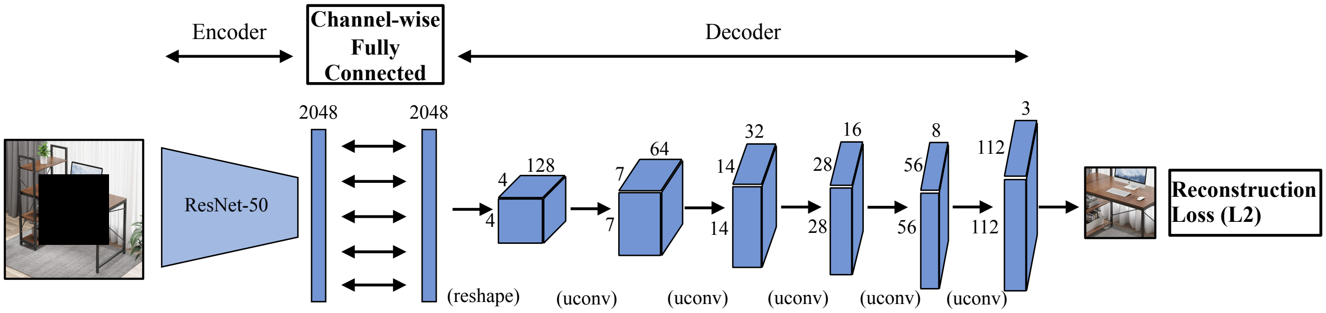


Figure S1. **The architecture of Context Encoder [47] with ResNet-50 [29] as backbone encoder.** Aligned with its original work, we use a channel-wise fully connected layer followed by a five-layer decoder to reconstruct masked central region from the encoder output.

S2.2. Baselines

We use ResNet-50 [29] as encoder for Context Encoder [47] for fair comparisons with other works (**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 [16] 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 (**Sec. 4.1**) respectively.

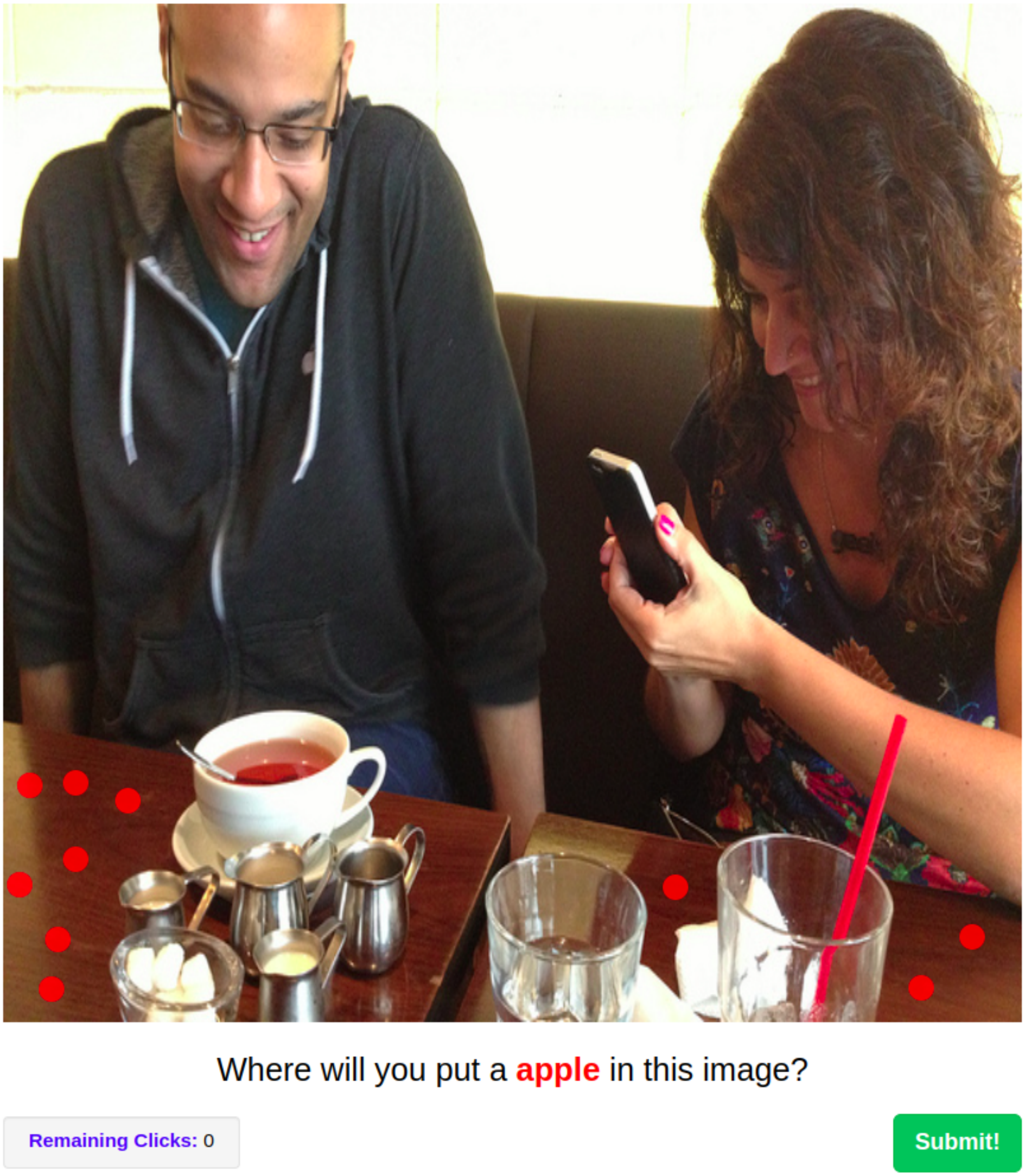
S2.3. Object Priming

[Stimulus designs] Here, we describe the steps we used 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 (**Sec. 4.1**). To accomplish this, we sampled images from the test set of 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. Finally, using the above steps, we ended up with 206 images and 864 unique image-object pairs.

[Human response collection] In **Fig. S2**, we show a screenshot of the AMT interface 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. Each participant received 15 USD per hour 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. Specifically, 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. To this 32×32 attention map, we then apply Gaussian smoothing using a 11×11 filter, followed by resizing to 224×224, and min-max normalization to generate final human priming maps.

[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. 2**. We use 5 grid sizes to generate priming maps in different scales (8×8, 14×14, 28×28, 56×56, 112×112) and normalize over these maps to obtain the final map. We provide more qualitative examples of model-human comparison in **Fig. S3**.



Where will you put a **apple** in this image?

Remaining Clicks: 0

Figure S2. **AMT user interface for human object priming experiment.** Red dots indicate the past click locations.

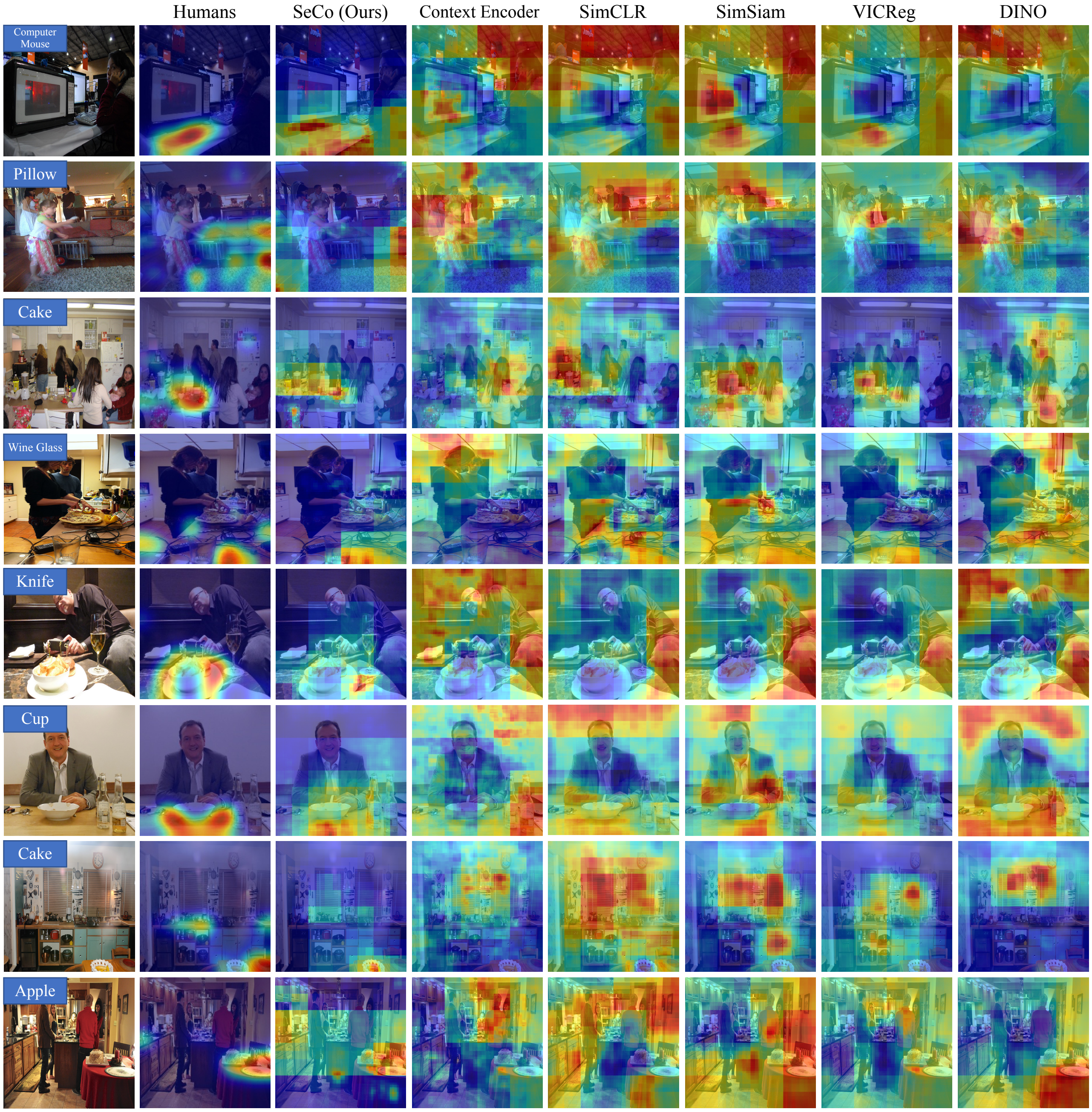


Figure S3. **SeCo priming maps highlight contextually relevant regions of the image and closely approximate human choices in the object priming task**. The leftmost column shows the input image and the given target object class label used for priming. The rest of the columns from left to right are priming maps from humans, predicted by our SeCo and predicted by all baselines (**Sec.** 4.2). See **Fig.** 4 for the colorbar.

S3. Ablation and memory analysis

S3.1. Analysis of Loss Components

As defined in **Sec. 3.4**, SeCo has a joint loss of MSE loss, covariance loss and variance loss. Here, we remove each component respectively to analyze its effectiveness on pretraining. We report top-1 accuracy on COCO-OCD in **Tab. S1**. The result demonstrates that without variance loss, SeCo reached information collapse, aligning with the trend in VICReg [2]. Without covariance loss, performance drops 2% in accuracy. Different from the observations made in VICReg [2], 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.

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

Table S1. **Ablation study on loss components**. α, β, γ are weightages of MSE loss, variance loss, covirance loss respectively.

S3.2. Probing External Memory

In **Sec. 5.3**, 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. 3**.

Algorithm 2: 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 scale 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 3: 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)
```