AN IMAGE IS WORTH K SLOTS: DATA-EFFICIENT SCALING OF SELF-SUPERVISED VISUAL PRE-TRAINING

Anonymous authors
Paper under double-blind review

ABSTRACT

Scaling up data and computing has become the norm for pre-training powerful visual encoders. Current algorithms, when scaled up, often require training on large-scale datasets that are unlikely to be object-centric. However, these algorithms were typically developed and validated on the object-centric ImageNet. This discrepancy may suggest sub-optimal scalability and underutilized data potential. Non-object-centric (NOC) data, with its multiple objects and complex layouts, tends to be more information-dense. To better leverage this underlying structure, we introduce a semantic bottleneck to MIM, which reduces the number of prototypes to encourage the emergence of objectness at patch-level token representation. Further, cross-view consistency regularization is applied to encourage multiview invariance. Together, this induces semantic object discovery and allows instance discrimination to be applied between object-level features (slots). Our experiments encompass pre-training on object-centric, scene-centric, web-crawled, and ego-centric data. Across all settings, our approach learns transferrable representations and achieves significant improvements over prior work in image recognition, scene understanding, and robot learning evaluations. When scaled up with million-scale datasets, our method also demonstrates superior data efficiency and scalability. We will make our code and model artifacts publicly available.

1 INTRODUCTION

Self-supervised representation learning from visual data has seen significant progress, evolving from contrastive learning (Chen et al., 2021; Caron et al., 2021) to masked image modeling (MIM) (Bao et al., 2022; He et al., 2022; Xie et al., 2022; Wei et al., 2022) and hybrid methods (Zhou et al., 2022; Oquab et al., 2024), which have benefited numerous downstream tasks. A key advantage of self-supervised learning is its ability to learn representations from unlabeled data, eliminating the need for human annotations and making it easier to scale up training datasets. Despite this advantage in utilizing diverse types of data, most research has focused on object-centric datasets like ImageNet for model development, leaving large volumes of non-object-centric (NOC) data, such as Open Images (Kuznetsova et al., 2020), SA-1B (Kirillov et al., 2023), LAION (Schuhmann et al., 2022), and Ego4D (Grauman et al., 2022), underutilized. However, many primary application domains of self-supervised learning – such as object detection, image segmentation, and robotics – often require handling NOC data.

This motivates us to explore the potential of NOC data for self-supervised learning, which is rich in information, offers new opportunities for data scaling, and could bridge the data-domain gap between self-supervised learning and real-world applications. While some research has investigated scene-centric data for self-supervised dense representation learning by developing pixel-level (Xie et al., 2021; Wang et al., 2021; Zhou et al., 2022) and object-level (Hénaff et al., 2021; 2022; Wen et al., 2022) contrastive learning objectives using learned or handcrafted objectness, these studies have primarily relied on ResNet-based backbones. As a result, it remains unclear how well these methods translate to modern architectures like vision transformers (see Tab. 1). Although one might argue that the state-of-the-art self-supervised model, DINOv2 (Oquab et al., 2024), already utilizes NOC data with a vision transformer backbone, its success heavily depends on data curation techniques that leverage the object-centric ImageNet dataset to select neighboring data, keeping its data distribution closely tied to object-centric approaches. Our preliminary experiments also suggest unsatisfactory results of DINOv2 using the same NOC data setup (Figs. 3 and 5).

To this end, we begin by conducting a comprehensive evaluation of existing self-supervised learning approaches on four datasets: object-centric (Deng et al., 2009) and non-object-centric (Lin et al., 2014; Changpinyo et al., 2021; Grauman et al., 2022). While the performance of current methods on non-object-centric data is suboptimal from a representation learning perspective, our study reveals that several insights from object-centric learning remain applicable to NOC data. Specifically, cross-view learning (Tian et al., 2020b) encourages semantic learning by enforcing feature invariance to data augmentations, while MIM, suitable for pre-training transformer-based architectures, is particularly effective at capturing fine-grained, low-level representations (see Fig. 1 for visualizations).

Motivated by these insights, we propose SlotMIM, a method that repurposes and integrates masked image modeling (MIM) and contrastive learning for effective representation learning from NOC datasets. The core idea of SlotMIM is to group patch-level image tokens into object-level feature abstractions, referred to as "slots", thereby decomposing NOC data into object-centric slots so that object-centric techniques can be effectively applied. To make patch-level tokens more semantically aware for subsequent grouping, we enhance MIM with cross-view consistency regularization. Additionally, we introduce a semantic bottleneck, which reduces the number of prototypes to encourage the emergence of semantic and objectness at patch-level token representations (see Fig. 1). Building on these semantically enriched patch tokens, we apply attentive pooling over the learned patch-level features, using prototypes to initialize object representations, thereby grouping patches into object-level slots and decomposing an image into object representations. Contrastive learning (Chen et al., 2021) is further applied to these slots to promote the discriminativeness of the learned representations. Together, these designs enable us to perform effective representation learning from NOC data.

In our experiments, we pretrain on a diverse range of datasets, including object-centric, scene-centric, web-crawled, and ego-centric data. We evaluate the pre-trained models on various tasks such as ImageNet linear probing and fine-tuning, ADE20K semantic segmentation, COCO detection/instance segmentation, and visuomotor control for robotics. Across all these evaluations, our method consistently outperforms existing approaches by a significant margin, showing 1) Data efficiency: It maximizes the utility of available data, reducing the dependency on continually scaling up data collection; 2) Domain adaptability: SlotMIM shows superior adaptability to datasets that are closer to the downstream application domains and potentially richer in information; and 3) Scalability: The method not only performs well at smaller scales but also scales efficiently with increasing data size.

In summary, our contributions can be summarized as follows:

- We conducted a comprehensive revisiting study across three non-object-centric datasets. Our findings reveal that non-object-centric (NOC) data is rich in information with vast potential, yet it remains underutilized in current approaches (Secs. 3.2 and 3.3).
- We formalize representation learning from NOC data into two key sub-tasks: decomposition and object-centric representation learning. By repurposing established techniques to target these specific sub-tasks, we developed a unified approach that effectively works for both non-object-centric and object-centric data, offering a robust solution that bridges the two domains (Fig. 3).
- Our method maximizes data utilization, achieving both data efficiency and excellent scalability. This contributes to the field of pre-training by exploring a new avenue for scaling up models using NOC data. At the same time, our approach delivers pre-trained models that are better suited for downstream tasks, including robotics, providing more relevant and effective solutions for real-world applications (Secs. 3.4 and 3.5).

2 METHOD

2.1 PRELIMINARIES

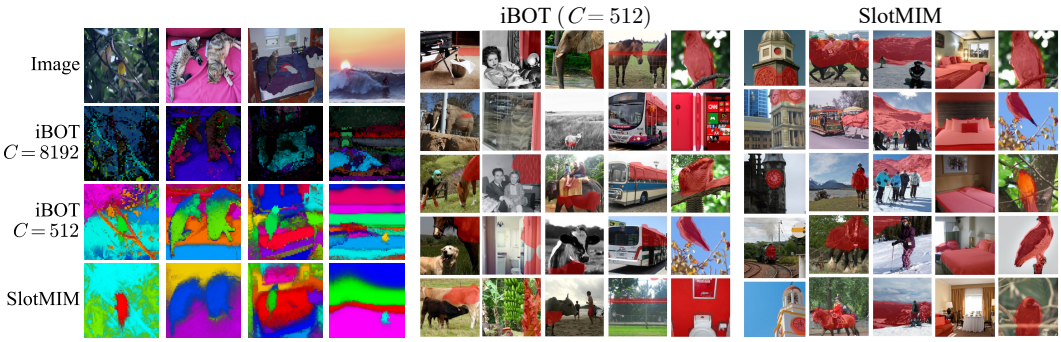
Deep clustering as self-distillation. DINO (Caron et al., 2021) is a discriminative self-supervised learning approach that learns a set of C prototypes online that clusters image embeddings. Let $\boldsymbol{x} \in \mathbb{R}^{H\times W\times 3}$ be an input image, and $f_\theta, f_\xi : \mathbb{R}^{H\times W\times 3} \to \mathbb{R}^d$ be student and teacher encoder networks parameterized by θ and ξ respectively. Let $\boldsymbol{z}_\theta = f_\theta(\boldsymbol{x})$ and $\boldsymbol{z}_\xi = f_\xi(\boldsymbol{x})$ denote the embeddings produced by the student and teacher networks (we omit the projector for simplicity). The cluster assignments are computed as $p_\theta(\boldsymbol{x}) = \text{softmax}(\boldsymbol{z}_\theta \cdot \mathcal{C}/\tau)$, where $\mathcal{C} = \{\boldsymbol{c}_c\}_{c=1}^C$ are the

prototypes, and τ is a temperature parameter. Then the loss is computed as cross-entropy between predictions of student model and teacher model: $\mathcal{L}_{\text{DINO}}(\boldsymbol{v}^1, \boldsymbol{v}^2) = -\sum_{c=1}^{C} q_\xi(\boldsymbol{v}^2)_c \log p_\theta(\boldsymbol{v}^1)_c$, where $\boldsymbol{v}^1$ and $\boldsymbol{v}^2$ are two augmented views of the same image. Since it resembles knowledge distillation with soft labels produced by the model itself, it is also dubbed as self-distillation.

DINO on image patches with MIM. iBOT (Zhou et al., 2022) extends the DINO objective from global image embeddings to local image patches with masked image modeling (MIM). Let $\mathcal{M} \in \{0, 1\}^N$ be a binary mask indicating which patches are masked, where N is the total number of patches. The masked input $\tilde{\boldsymbol{v}}$ is defined as $\tilde{\boldsymbol{v}}_i = \boldsymbol{m}$ if $\mathcal{M}_i = 1$, and $\tilde{\boldsymbol{v}}_i = \boldsymbol{v}_i$ otherwise, where $\boldsymbol{m}$ is a mask token. The iBOT loss predicts the clustering assignments of masked patches given unmasked patches: $\mathcal{L}_{\text{iBOT}}(\boldsymbol{v}) = \sum_{i:\mathcal{M}_i=1} \mathcal{L}_{\text{DINO}}(\tilde{\boldsymbol{v}}_i, \boldsymbol{v}_i)$, where $\tilde{\boldsymbol{v}}_i$ is the masked patch from the student model and $\boldsymbol{v}_i$ is the corresponding unmasked patch from the teacher model.

Slot attention (Locatello et al., 2020) is a variant of cross-attention that normalizes attention scores on the query side instead of the key side, introducing competition between queries and encouraging them to focus on different parts of the input. Our approach performs attentive pooling on patch embeddings according to their clustering assignments, sharing high-level intuition with slot attention if viewing the prototypes $\mathcal{C}$ as queries and patch embeddings $\boldsymbol{z}_{\theta,i} = f_\theta(\boldsymbol{x}_i)$ as keys. We thus follow tradition and call the pooled object features slots – prototypes adapted to image patches.

2.2 Motivation and Framework



(a) Clustering assignment of patch tokens. (b) Top-5 segments retrieved by the prototypes (by column).

Figure 1: **Comparison of concepts learned by iBOT and SlotMIM.** iBOT's prototypes can discover fine-grained patterns, and the quality improves if a smaller vocabulary is used (left). But these patterns are bottom-up and lack semantic meaning. In contrast, concepts with same tokens of SlotMIM are semantically coherent and more suitable for building instance discrimination pretext tasks (right).

High-level intuition. We decouple self-supervised learning on non-object-centric data into two subtasks: 1) learning to group image patches into objects (or stuff); and 2) learning to discriminate objects as previous works have done on object-centric data. The major challenge here is unsupervised object discovery, which we find could emerge from iBOT with a smaller number of prototypes.

Representation bottleneck induces objectness from iBOT. We first investigate the prototypes of iBOT, which is a set of embeddings $\mathcal{C} = \{\boldsymbol{c}_c\}_{c=1}^{C}$ that clusters image patches into C clusters and assigns each patch token a soft one-hot encoding $p_\theta(\boldsymbol{x}_i)$ identifying its clustering assignment. Conventionally, C is set to be 8192 to capture fine-grained patterns, which is good for learning representations. But in our case, the role of representation learning would be taken by another objective (contrastive learning between slots) and the prototypes are designated to focus on object discovery. We find a much smaller C, *e.g.*, 512 for COCO, would suit this goal better because it can build a very compact information bottleneck that forces the model to learn highly compositional concepts – objects. As shown in Fig. 1a, the clusters discovered by iBOT are very fine-grained (2nd row), and objectness emerges if a small vocabulary is used (3rd row). However, these patterns still lack semantic meaning and could split the same object into multiple parts. Also, it remains hard to match discovered objects between views as their semantics vary a lot despite having the same token (Fig. 1b, left). Both issues call for a set of semantic-level prototypes.

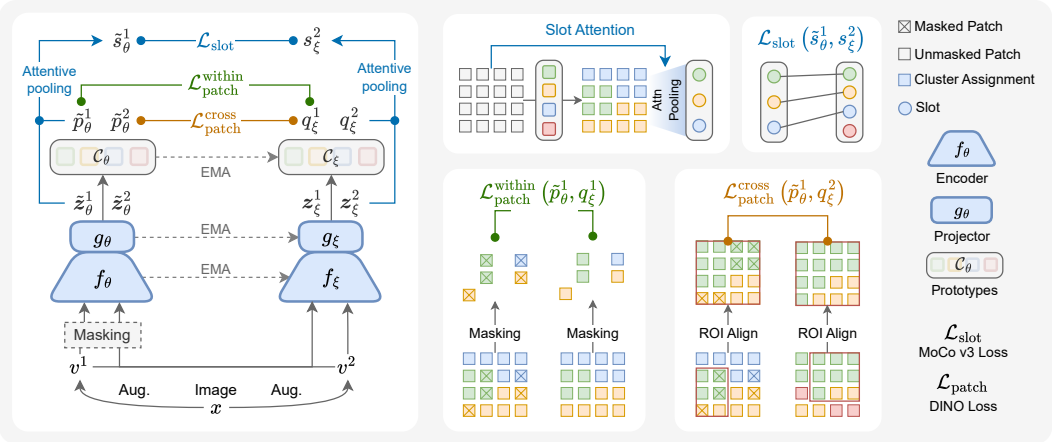


Figure 2: **Overview of SlotMIM.** We repurpose iBOT's within-view patch-level loss for object discovery, add a cross-view objective for semantic guidance, and build object-centric contrastive learning on top of object features (slots) grouped from patches with identical clustering assignments.

Cross-view consistency lifts object discovery to semantic-level. A key factor contributing to the lack of semantic meaning is that the iBOT loss $\mathcal{L}_{\text{iBOT}}$ is computed between patches within the same view. Consequently, there is no explicit guidance for learning invariant representations across different views of the same object or scene. We apply a simple yet effective fix: adding a cross-view consistency objective $\mathcal{L}_{\text{patch}}^{\text{cross}}$ that enforces patches undergone different photometric and geometric transformations to have the same token. To match patches between views, we adopt a SlotCon-style mechanism that crops & resizes the overlapping regions of two views (using ROIAlign). Formally, let $\boldsymbol{v}^1$ and $\boldsymbol{v}^2$ be two augmented views of the same image, and $\tilde{\boldsymbol{z}}_{\theta,i}^1 = f_\theta(\tilde{\boldsymbol{v}}_i^1)$ and $\boldsymbol{z}_{\xi,j}^2 = f_\xi(\boldsymbol{v}_j^2)$ be the corresponding patch embeddings. The cross-view consistency loss is defined as:

$$\mathcal{L}_{\text{patch}}^{\text{cross}}(\boldsymbol{v}^1, \boldsymbol{v}^2) = -\frac{1}{|\mathcal{P}|} \sum_{(i,j)\in\mathcal{P}} \sum_{c=1}^{C} \boldsymbol{q}_{\xi,i,c}^2 \log \tilde{\boldsymbol{p}}_{\theta,j,c}^1 \,, \tag{1}$$

where $\tilde{\boldsymbol{p}}_\theta^1 = \text{softmax}(\tilde{\boldsymbol{z}}_\theta^1 \cdot \mathcal{C}_\theta/\tau_s)$ and $\boldsymbol{q}_\xi^2 = \text{softmax}(\boldsymbol{z}_\xi^2 \cdot \mathcal{C}_\xi/\tau_t)$ are the cluster assignments of the student and teacher models respectively, τ_s and τ_t are temperature parameters of the student and teacher models respectively, and $\mathcal{P}$ is the set of matched patch pairs between views.

Object-level constrastive learning. Now that we have a set of object features that are aligned between views, we can apply a contrastive learning objective to perform object-centric learning. Not all slots are used. We only keep the slots that at least occupy one patch and we filter out the redundant ones by computing the following binary indicator: $\mathbb{1}_i = \exists_j$ such that $\text{argmax}_c(p_\theta(\boldsymbol{v}_j^1)_c) = i$. Those with the same tokens form positive pairs and others form negative pairs. We adopt a MoCo-style contrastive learning approach. Let $\boldsymbol{s}_{\theta,i}^1 = \sum_j p_\theta(\boldsymbol{v}_j^1)_i \boldsymbol{z}_{\theta,j}^1$ and $\boldsymbol{s}_{\xi,i}^2 = \sum_j q_\xi(\boldsymbol{v}_j^2)_i \boldsymbol{z}_{\xi,j}^2$ be the slots from the student and teacher models respectively. The contrastive loss is defined as:

$$\mathcal{L}_{\text{slot}}(\tilde{\boldsymbol{s}}_\theta^1, \boldsymbol{s}_\xi^2) = -\frac{1}{K} \sum_{i=1}^{C} \log \frac{\mathbb{1}_i^1 \mathbb{1}_i^2 \exp(h_\theta(\boldsymbol{s}_{\theta,i}^1) \cdot \boldsymbol{s}_{\xi,i}^2/\tau)}{\sum_{j=1}^{C} \mathbb{1}_i^1 \mathbb{1}_j^2 \exp(h_\theta(\boldsymbol{s}_{\theta,i}^1) \cdot \boldsymbol{s}_{\xi,j}^2/\tau)} \,, \tag{2}$$

where h_θ is a predictor MLP, $K = \sum_i \mathbb{1}_i^1 \mathbb{1}_i^2$ is the number of positive pairs and τ is a temperature parameter set to 0.2 following Chen et al. (2021). ℓ_2-normalization is applied to both slots and their predictions before computing the inner product. The final loss is a combination of these objectives:

$$\mathcal{L}_{\theta,\xi}(\tilde{\boldsymbol{v}}^1, \boldsymbol{v}^2) = \lambda_1 \mathcal{L}_{\text{patch}}^{\text{within}}(\tilde{\boldsymbol{v}}^1, \boldsymbol{v}^2) + \lambda_1 \mathcal{L}_{\text{patch}}^{\text{cross}}(\tilde{\boldsymbol{v}}^1, \boldsymbol{v}^2) + \lambda_2 \mathcal{L}_{\text{slot}}(\tilde{\boldsymbol{s}}_\theta^1, \boldsymbol{s}_\xi^2) \,, \tag{3}$$

where $\mathcal{L}_{\text{patch}}^{\text{within}}$ is exactly the same as $\mathcal{L}_{\text{iBOT}}$ and $\lambda_1 = 0.5$ and $\lambda_2 = 1$ are weighting coefficients. In practice the symmetrized objective $\mathcal{L}_{\theta,\xi}(\tilde{\boldsymbol{v}}^1, \boldsymbol{v}^2) + \mathcal{L}_{\theta,\xi}(\tilde{\boldsymbol{v}}^2, \boldsymbol{v}^1)$ is optimized.

Connection to previous work. As shown in Tab. 1, SlotMIM shares key ideas with previous self-supervised learning methods. MoCo v3 (Chen et al., 2021) and DINO (Caron et al., 2021) perform instance discrimination on image crops. SlotCon (Wen et al., 2022) performs contrastive learning between slots, but its objectness only receives high-level signals. This worked well for ResNet since the network architecture provided strong inductive bias for objectness. But when applied to ViT, MIM-like low-level signal is needed. Regarding iBOT, its patch-level and global self-distillation loss are built on the same set of prototypes, which however requires different levels of complexity. Our design allows the prototypes to focus on learning fine-grained patterns for patch-level loss, and semantic learning is achieved by other modules.

Table 1: Holistic comparison with previous methods.

Method	mask	$\mathcal{L}_{\text{patch}}^{\text{cross}}$	$\mathcal{L}_{\text{patch}}^{\text{within}}$	$\mathcal{L}_{\text{slot}}$	k-NN	ADE	Jacc	Loc
MoCo v3	✗	✗	✗	△	43.3	41.3	–	–
DINO	✗	✗	✗	○	**46.3**	40.5	–	–
SlotCon	✗	✓	✗	✓	42.9	47.1	40.1	59.6
iBOT	✓	✗	✓	○	45.3	44.5	–	–
SlotMIM	✓	✓	✓	✓	46.2	**49.1**	**43.9**	**62.5**

△: contrastive learning on image crops
○: self-distillation on image crops

3 Experiments

3.1 Setting

Table 2: **Overview of pre-training datasets.** We uniformly sample subsets of 241K[1] images from ImageNet, CC12M, and Ego4D. COCO+ is formed by combining `train` and `unlabeled` subsets of COCO. For Ego4D we first extract frames at 0.2 fps and then sample image subsets.

Pre-train Data	Source	#Image	#Obj/Img	#Class	Type	Video
INet-241K	ImageNet	241K	1.7	1000	OC	✗
COCO+	COCO	241K	7.3	80	SC	✗
CC-241K	CC12M	241K	–	–	Web	✗
Ego-241K	Ego4D	241K	–	–	Ego	✓

OC: Object-centric; SC: Scene-centric; Web: Web-crawled; Ego: Ego-centric

Dataset. We consider pre-training on four types of datasets, including object-centric ImageNet (Deng et al., 2009), scene-centric COCO (Lin et al., 2014), web-crawled CC12M (Changpinyo et al., 2021), and ego-centric Ego4D (Grauman et al., 2022). For the baseline setting, we uniformly sample 241K images from each dataset to form the training sets. See Tab. 2 for details. For larger-scale pre-training, we sample 1.28M images from the same sources, except for scene-centric data where we switch to Open Images (Kuznetsova et al., 2020).

Methods. We compare with a variety of ViT pre-training methods, including BEiT (Bao et al., 2022), SplitMask (El-Nouby et al., 2021), MAE (He et al., 2022), DINO (Caron et al., 2021), iBOT (Zhou et al., 2022), and DINOv2 (Oquab et al., 2024). We train with official code and suggested hyperparameters. For DINO and iBOT, training instability is observed when training on NOC data, and we tuned the teacher temperature if necessary for convergence.

Pre-training setting. We use ViT-B/16 (Dosovitskiy et al., 2021) as the backbone. At 241K data scale, all methods are trained for 800 epochs by default. At 1.28M data scale, we train for 400 epochs. The optimization hyperparameters follow Zhou et al. (2022).

Evaluation setting. We evaluate models on ImageNet-1K (Deng et al., 2009) and ADE20K (Zhou et al., 2017) following He et al. (2022). For ImageNet linear probing, we sweep between `[CLS]` token and average pooling and report best results of each model. For ImageNet fine-tuning, all models use the average-pooled token. Under both settings, we report top-1 validation accuracy of a single 224×224 center crop. ADE20K semantic segmentation experiments use UperNet (Xiao et al., 2018) and train for 160K iterations with batch size 16. Additionally, COCO object detection and instance segmentation is also considered to evaluate the transferability of pre-trained models. We follow the

[1] 1.28M subsets are also considered. For scene-centric data, we use Open Images dataset to scale up.

same setting in (Zhou et al., 2022) to train a Cascade Mask R-CNN (Cai & Vasconcelos, 2019) with $1\times$ schedule (12 epochs), and report box and mask AP.

Analytical metrics. We also introduce some numeric indicators to help analyze some properties of pre-trained models. This includes k-NN ImageNet classification ($k = 20$) following Caron et al. (2021), and object discovery metrics evaluated on Pascal VOC 2012 following Venkataramanan et al. (2024); Siméoni et al. (2021). Jaccard similarity measures the overlap between predicted mask P and the ground truth mask G as $J(P, G) = \frac{G \cap P}{G \cup P}$. We also compute CorLoc, which measures the percentage of correctly located boxes, where a predicted box is correct if it's IoU ≥ 0.5. Additionally, we maintain a running mean of the average number of active slots $\overline{K}$ in an image during training.

3.2 Results under Baseline Pre-training Budget

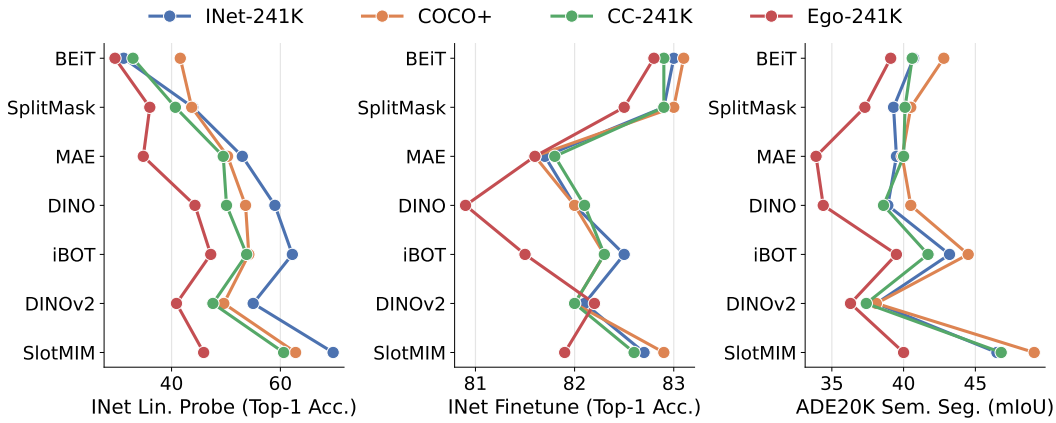


Figure 3: **Different models learn different levels of information from different datasets.** SlotMIM consistently outperforms prior arts whether pre-trained on object-centric data or not. Notably, when trained on COCO+, it transfers better than most ImageNet models despite the domain gap (middle). When evaluated on segmentation, the superiority of our method is even more pronounced (right).

We first evaluate models pre-trained on 241K-scale datasets, and show that NOC data can be good learning resources if used properly. The results are present in Fig. 3. Overall, SlotMIM achieves the best performance across classification and segmentation tasks, no matter learning from object-centric data or not. Below, we discuss some other interesting findings.

Features learned from NOC data can be linear separatable on ImageNet. From Fig. 3 (left), our models trained on COCO and CC achieve similarly good linear probing performance on ImageNet with best prior ImageNet-trained methods. As a clear contrast, all previous methods trained on NOC datasets (COCO, CC, and Ego4D) fall behind best ImageNet counterpart.

NOC data can be worth more than ImageNet for ImageNet. As shown in Fig. 3 (middle), under ImageNet fine-tuning setting, the top-3 methods (BEiT, SplitMask, and SlotMIM) have best performance when trained on COCO+ instead of ImageNet. For MAE and DINO, training on CC also transfers better than ImageNet. Note that this is uncommon given the domain gap between NOC pre-training data and OC downstream task, demonstrating that NOC data are information-rich learning resources.

NOC data is significantly beneficial for similar-domain downstream tasks. In Fig. 3 (right), we evaluate the models on ADE20K semantic segmentation. SlotMIM trained on COCO+ achieves the best performance, and our CC and ImageNet-trained models also surpass prior models by a large margin. This suggests that NOC data can be particularly useful for scene-understanding tasks.

Ego-centric data solely is not suitable for general-purpose models. In Fig. 3, we observe that models trained on Ego4D generally perform worse than those trained on other datasets. This is possibly due to video-based ego-centric data's redundancy and suggests that data diversity matters more for general-purpose pre-training. Still, as will be discussed in Sec. 3.5, ego-centric data can be effective for robot learning and SlotMIM learns the best representations from it.

3.3 Scaling Up Pre-training Schedule

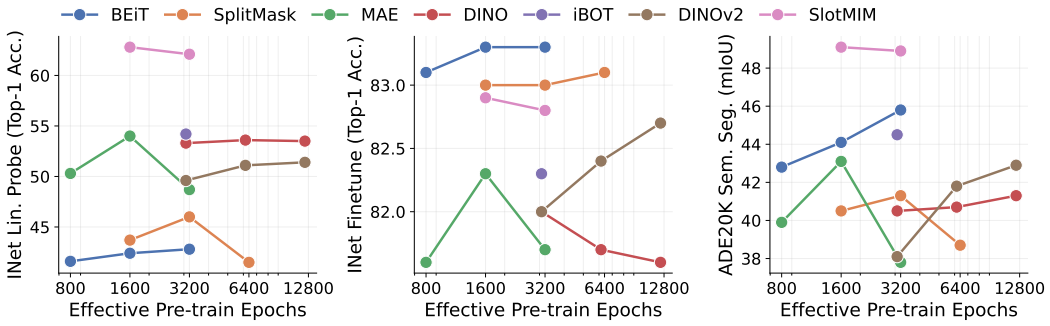


Figure 4: **A convergence check on COCO+ with longer training.** Existing methods either experience performance degradation or stagnate, or require significantly more epochs to reach better performance. SlotMIM achieves leading performance with shorter training (also without multi-crop).

SlotMIM is not only efficient in the need of data scale, but also in the need of training epochs. We take training on COCO+ as an example, and compare SlotMIM with other methods considering longer training schedules. For a fair comparison, we follow previous literature (Zhou et al., 2022) to calculate effective pre-training epochs for each method, which is 3.84× for methods using multi-crop (DINO, iBOT, and DINOv2), 2× for contrastive methods including SlotMIM, and 1× for non-contrastive methods (*e.g.*, BEiT and MAE). The results are shown in Fig. 4. We observe that SlotMIM achieves the best performance with the shortest training schedule, and other methods either require significantly more epochs to reach better performance or experience performance degradation or stagnation.

3.4 Scaling Up Pre-training Data

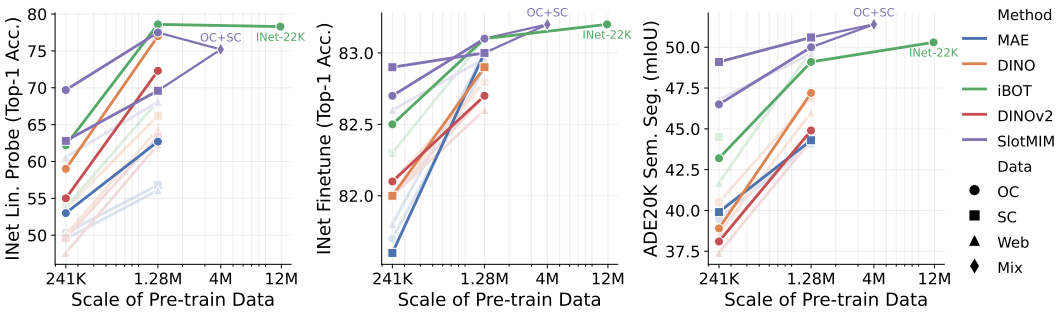


Figure 5: **Scaling laws on different data sources.** We scale up object-centric, scene-centric, and web-crawled data, and highlight the best (model, data) combinations. Our method learns strong and transferable representations with significant data efficiency and continues to improve with more data.

Superior data efficiency allows us to explore larger-scale pre-training data. In Fig. 5, we show that SlotMIM achieves strong performance with remarkable data efficiency.

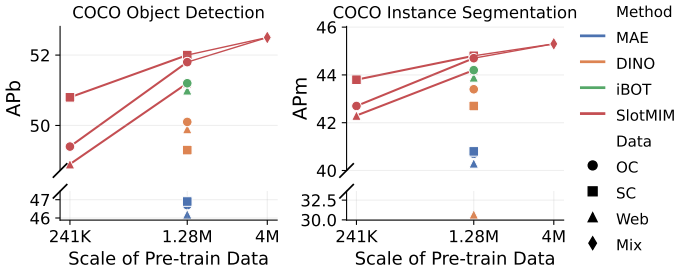
Comparable or better performance with small data scale. As shown in Figure 5, SlotMIM achieves comparable or superior performance to other methods using significantly less data. Our INet-241K model for ImageNet linear probing, and COCO+/INet-241K models for ImageNet fine-tuning and ADE20K semantic segmentation outperform or match most models trained on 1.28M ImageNet images across various tasks. This remarkable data efficiency demonstrates our approach's effectiveness in extracting rich, transferable features from limited data.

NOC pre-training rivals ImageNet pre-training for ImageNet. Interestingly, we observe that pre-training on NOC datasets like OpenImages-1.28M can lead to performance better than pre-training on ImageNet for the ImageNet classification task (fine-tuning setting). When scaled up to 4M scale, this trend becomes more pronounced. This aligns with the trend in Fig. 3 that NOC data can provide more information-rich features, which can be better-utilized by models like SlotMIM.

NOC data also possesses stronger scalability. We extend experiments to the 4M scale by combining INet-1.28M (Deng et al., 2009), COCO+ (Lin et al., 2014), OpenImages (Kuznetsova et al., 2020), Objects365 (Shao et al., 2019), and LVIS (Gupta et al., 2019b). Compared with previous efforts on scaling up with ImageNet-22K (12M images) (Russakovsky et al., 2015), the performance of our SlotMIM models continues to improve and surpasses them with 3× less data. This suggests that NOC data can be a more scalable learning resource.

3.4.1 Object Detection and Instance Segmentation

Figure 6: **Transfer learning experiments on COCO object detection and instance segmentation.** SlotMIM shows better data efficiency with both OC and NOC data, and the performance continually grows with more data, surpassing all prior SoTA models by a notable margin.



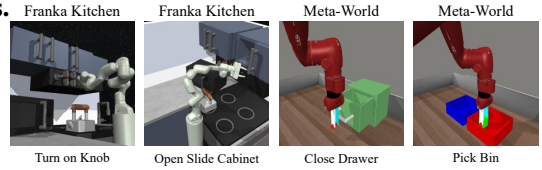
In Fig. 6, we also present an evaluation on COCO object detection and instance segmentation. The superiority of SlotMIM is clear and remains improving with increased data scale.

3.5 Pre-trained Vision Models for Motor Control

Previous sections showed that 1) NOC data offers rich, transferable features for image recognition and scene understanding tasks, and 2) its advantages are especially evident when there is strong alignment between pre-training data and target downstream tasks. In this section, we analyze the effects of OC/NOC data types (ego-centric and scene-centric) on robot manipulation benchmarks and the data efficiency of SlotMIM.

Table 3: **Overview of robot manipulation tasks.** Right: example tasks of each benchmark suite.

Bench. Suite	RGB	Proprio.	#Task	#Demo	#Seed
Franka Kitchen	✓	✗	5	25	3
Meta-World	✓	✗	8	25	3



Imitation learning setups. Following Hu et al. (2023), we compared our methods across two robot manipulation benchmarks using behavior cloning: Franka-kitchen (Gupta et al., 2019a) and Meta-world (Yu et al., 2019). We focus on efficient real-world learning with behavior cloning (BC) using a few human demonstrations per task in each benchmark suite. For each pre-trained vision model and task, we run 3 seeds of BC due to the result's high variability. Detailed setups for behavior cloning and example tasks are shown in Tab. 3. One-image observation for its comparable performance to stacks of images and higher computational efficiency. All tasks and environments use 224×224 RGB images without proprioceptive input. No image augmentations, such as random shifts, are applied. The policy training includes a few modifications: The policy network is trained for 20,000 steps, following R3M (Nair et al., 2023). We employ attentive pooling, as in V-Cond (Karamcheti et al., 2023), which is shown to be the better choice than the default `[CLS]` embedding head and provides better comparisons between pre-trained frozen visual representations.

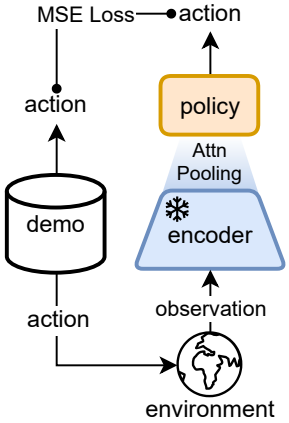


Figure 7: Behavior cloning with attentive probing.

Baselines. As shown in Fig. 8, MAE regime (blue line) including MVP (Radosavovic et al., 2023) and VC-1 (Majumdar et al., 2023) that leverages MAE (He et al., 2022) to pre-train the model across a massive collection of ego-centric videos (Grauman et al., 2022) and Internet data. V-Cond (Karamcheti et al., 2023) (purple point) further proposes language-driven representation

learning from human videos and associated captions. DINO (Caron et al., 2021) (orange line) is based on self-distillation and iBOT (Zhou et al., 2022) (green line) further combines MIM with self-distillation.

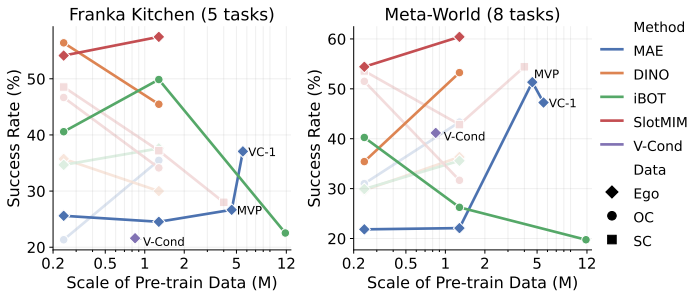


Figure 8: **Pre-training for robot manipulation tasks.** This evaluation considers three factors that influence manipulation success rates: data types (ego-centric ◆, object-centric ●, and scene-centric ■), pre-training methods, and data scale. Dark lines represent the best-performing data scaling for each pre-training method, while light lines indicate sub-optimal performance.

Fig. 8 examines the relationship between manipulation success rates and pre-training methods, comparing the trend of scaling dataset size across different data types: ego-centric ◆, object-centric ●, and scene-centric ■. Notably, increasing dataset size does not always improve performance across benchmarks, as also reported by VC-1(Majumdar et al., 2023).

Different scaling behaviors of OC/SC vs. ego-centric data. In object manipulation tasks as shown in Fig. 8, scaling scene-centric and object-centric data to the million level can lead to performance drops for methods like MAE, DINO, and iBOT. We hypothesize that self-supervised representation learning, including MIM, aims to learn invariance, where the feature extractor pulls images with similar visual content together in the embedding space, compressing the visual data. However, scaling up data may result in over-compression, causing performance drops in visuomotor control tasks.

By contrast, using ego-centric data for pre-training, MAE (blue line) and SlotMIM (red line) show positive data scaling effects. Unlike SC/OC data from vast Internet sources, ego-centric images are sampled from consecutive human videos that share contextual backgrounds or scenarios. The ego-centric data are among daily scenarios such as household, outdoor, workplace, and leisure *etc*. that are contextually similar to the robot manipulation scenarios (Grauman et al., 2022). Thus, invariance learning in ego-centric data tends to focus more on the differences within the same video or scenario, particularly in the foreground objects. This focus is critical for robot manipulation learning, as it requires effective interaction with these foreground objects.

SlotMIM is more data efficient in leveraging ego-centric data. Compared to general-purpose pre-trained models and state-of-the-art (SoTA) robot learning methods (e.g., MVP (Radosavovic et al., 2023) and VC-1 (Majumdar et al., 2023)), we demonstrate that SlotMIM (dark red ◆ line), pre-trained with just 241K data samples, can surpass prior methods that utilized over 1 million samples. When scaled to 1 million ego-centric data, it achieves the highest success rates compared to all other methods in the figure.

3.6 Ablation Study

Table 4: **Ablation study on effective modules.**

	mask	$\mathcal{L}_{\text{patch}}^{\text{cross}}$	$\mathcal{L}_{\text{patch}}^{\text{within}}$	$\mathcal{L}_{\text{slot}}$	k-NN	ADE	Jacc	Loc	$\overline{K}$
1	✗	✓	✗	✗	45.1	47.4	42.5	55.6	8.3
2	✓	✓	✗	✗	44.9	48.6	42.3	60.7	10.3
3	✓	✗	✓	✗	27.7	45.7	39.3	**65.5**	20.7
4	✓	✓	✗	✓	45.3	47.5	42.9	63.6	8.4
5	✓	✓	✓	✓	**46.2**	**49.1**	**43.9**	62.5	9.4

This section presents an ablation study on key SlotMIM design choices. Models are trained on COCO+ for 800 epochs. Tab. 4 demonstrates the impact of different modules, comparing k-NN ImageNet classification, ADE20K semantic segmentation, Jaccard similarity, and CorLoc for object discovery mask quality and localization recall. We report the average number of objects/stuff discovered per image. Results show Jaccard similarity correlates with representation quality, suggesting better object discovery improves representation learning. Introducing MIM enhances object localization and benefits segmentation tasks (rows 1 and 2). Cross-view consistency and slot contrastive losses contribute to improved object discovery (rows 2, 3, 5). The within-view loss can serve as a regularizer and improve representations (rows 4, 5).

Table 5: **Ablation studies on hyperparameters.** Default values are marked with a cyan background.

(a) **Number of prototypes**

C	k-NN	ADE	Jacc	Loc	$\overline{K}$
256	45.3	**49.1**	42.2	61.2	7.8
512	**46.2**	**49.1**	**43.9**	**62.5**	9.4
1024	45.6	48.4	42.8	62.6	10.8

(b) **Mask ratio (±0.2)**

	k-NN	ADE	Jacc	Loc	$\overline{K}$
0.3	**46.2**	**49.1**	43.9	62.5	9.4
0.4	45.8	48.6	45.0	62.6	8.1
0.5	44.3	48.2	**45.7**	**64.8**	7.1

(c) **Patch loss**

Type	k-NN	ADE	Jacc	Loc	$\overline{K}$
center	**46.2**	49.1	**43.9**	62.5	9.4
SH	45.1	**49.3**	40.8	**68.5**	15.2

In Tab. 5 we present ablations on some numeric design choices. Generally speaking, a smaller number of prototypes, a higher mask ratio, and the use of centering (Caron et al., 2021) instead of Sinkhorn-Knopp algorithm (Caron et al., 2020) encourage the network to discover more holistic concepts/objects, while the opposite discovers more fine-grained ones. Optimal representation is highly related to object discovery quality.

4 Related work

Self-supervised representation learning. Self-supervised representation learning aims to extract transferrable features from unlabeled data (Tian et al., 2020a; Caron et al., 2018; 2020; 2021; Asano et al., 2020; Chen et al., 2020). Two primary approaches have emerged: contrastive learning (Tian et al., 2020a; Chen et al., 2020; He et al., 2020), which learns by comparing positive and negative examples, and masked image modeling (He et al., 2022; Xie et al., 2022), which reconstructs masked regions of images. While these methods have shown success, they've primarily been tested on object-centric datasets like ImageNet-1K. Our study extends this by exploring self-supervised learning on large-scale non-object-centric datasets, demonstrating superior data efficiency compared to previous pre-training methods across various downstream applications. Additionally, we provide insights into the scalability and generalizability of these methods across diverse data types.

Learning on non-object centric data. Recent works have addressed the challenge of self-supervised learning on non-object centric data (Van Gansbeke et al., 2021; Oquab et al., 2024; Xie et al., 2021; Wang et al., 2021; Hénaff et al., 2021; 2022). These efforts include dense contrastive learning approaches (Wang et al., 2021; Xie et al., 2021), object-centric methods for dense prediction tasks (Hénaff et al., 2022), and slot-based contrastive learning frameworks (Wen et al., 2022). Additionally, some methods focus on learning from uncurated datasets (Caron et al., 2019; Tian et al., 2021; Bai et al., 2022). In our work, we decompose object representations to leverage established techniques that enable fine-grained pattern learning through patch-level target design, facilitating effective pre-training.

Scaling vision pre-training. Scaling vision pre-training to larger datasets and models has become a significant focus in recent years (Tian et al., 2021; Caron et al., 2019; Mu et al., 2022; Radford et al., 2021; Dehghani et al., 2023; Gadre et al., 2023; Schuhmann et al., 2022). The creation and use of massive datasets like LVD-142M (Oquab et al., 2024) and LAION-5B (Schuhmann et al., 2022) have also played a crucial role. Our method examines how non-object-centric datasets influence data scaling and the transferability of learned representations across downstream tasks, focusing on fine-grained data types: object-centric, scene-centric, ego-centric, and mixed.

5 Conclusion

This work revisits the use of non-object-centric (NOC) data for self-supervised visual representation learning. Our comprehensive study demonstrated that NOC data holds immense potential due to its rich information, which has been largely underutilized. To harness this potential, we formalized learning from NOC data into two sub-tasks: scene decomposition and object-centric representation learning. By repurposing and integrating established techniques to target these sub-tasks, we developed SlotMIM, a unified framework capable of effectively handling both NOC and object-centric data. Through extensive experiments across diverse datasets and downstream tasks, including robotics, we demonstrated the consistent superiority of our approach over existing methods. We hope our promising results open new avenues for scaling self-supervised learning using large volumes of NOC data, overcoming the limitations posed by conventional datasets in representation learning.

REFERENCES

Yuki M. Asano, Christian Rupprecht, and Andrea Vedaldi. Self-labelling via simultaneous clustering and representation learning. In *International Conference on Learning Representations*, 2020.

Yutong Bai, Xinlei Chen, Alexander Kirillov, Alan Yuille, and Alexander C. Berg. Point-level region contrast for object detection pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16061–16070, 2022.

Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. BEiT: BERT pre-training of image transformers. In *International Conference on Learning Representations*, 2022.

Zhaowei Cai and Nuno Vasconcelos. Cascade R-CNN: High quality object detection and instance segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(5):1483–1498, 2019.

Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *European Conference on Computer Vision*, pp. 139–156, 2018.

Mathilde Caron, Piotr Bojanowski, Julien Mairal, and Armand Joulin. Unsupervised pre-training of image features on non-curated data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2959–2968, 2019.

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *Advances in Neural Information Processing Systems*, pp. 9912–9924, 2020.

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 (ICCV)*, pp. 9650–9660, 2021.

Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12M: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3558–3568, 2021.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 1597–1607, 2020.

Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9640–9649, 2021.

Mostafa Dehghani, Josip Djolonga, Basil Mustafa, Piotr Padlewski, Jonathan Heek, Justin Gilmer, Andreas Peter Steiner, Mathilde Caron, Robert Geirhos, Ibrahim Alabdulmohsin, et al. Scaling vision transformers to 22 billion parameters. In *International Conference on Machine Learning*, pp. 7480–7512. PMLR, 2023.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255. IEEE, 2009.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

Alaaeldin El-Nouby, Gautier Izacard, Hugo Touvron, Ivan Laptev, Hervé Jegou, and Edouard Grave. Are large-scale datasets necessary for self-supervised pre-training? *arXiv preprint arXiv:2112.10740*, 2021.

Samir Y. Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, Eyal Orgad, Rahim Entezari, Giannis Daras, Sarah Pratt, Vivek Ramanujan, Yonatan Bitton, Kalyani Marathe, Stephen Mussmann, Richard Vencu, Mehdi Cherti, Ranjay Krishna, Pang Wei Koh, Olga Saukh, Alexander J. Ratner, Shuran Song, Hannaneh Hajishirzi, Ali Farhadi, Romain Beaumont, Sewoong Oh, Alex Dimakis, Jenia Jitsev, Yair Carmon, Vaishaal Shankar, and Ludwig Schmidt. DataComp: In search of the next generation of multimodal datasets. In *Advances in Neural Information Processing Systems*, pp. 27092–27112, 2023.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4D: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18995–19012, 2022.

Abhishek Gupta, Vikash Kumar, Corey Lynch, Sergey Levine, and Karol Hausman. Relay policy learning: Solving long-horizon tasks via imitation and reinforcement learning. *arXiv preprint arXiv:1910.11956*, 2019a.

Agrim Gupta, Piotr Dollar, and Ross Girshick. LVIS: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5356–5364, 2019b.

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*, pp. 9729–9738, 2020.

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*, pp. 16000–16009, 2022.

Olivier J. Hénaff, Skanda Koppula, Jean-Baptiste Alayrac, Aäron van den Oord, Oriol Vinyals, and João Carreira. Efficient visual pretraining with contrastive detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10086–10096, 2021.

Olivier J. Hénaff, Skanda Koppula, Evan Shelhamer, Daniel Zoran, Andrew Jaegle, Andrew Zisserman, João Carreira, and Relja Arandjelović. Object discovery and representation networks. In *European Conference on Computer Vision*, pp. 123–143, 2022.

Yingdong Hu, Renhao Wang, Li Erran Li, and Yang Gao. For pre-trained vision models in motor control, not all policy learning methods are created equal. In *International Conference on Machine Learning*, pp. 13628–13651, 2023.

Siddharth Karamcheti, Suraj Nair, Annie S. Chen, Thomas Kollar, Chelsea Finn, Dorsa Sadigh, and Percy Liang. Language-driven representation learning for robotics. In *Robotics: Science and Systems (RSS)*, 2023.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4015–4026, 2023.

Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. The Open Images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International journal of computer vision*, 128(7):1956–1981, 2020.

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*, pp. 740–755, 2014.

Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. Object-centric learning with slot attention. In *Advances in Neural Information Processing Systems*, pp. 11525–11538, 2020.

Arjun Majumdar, Karmesh Yadav, Sergio Arnaud, Jason Ma, Claire Chen, Sneha Silwal, Aryan Jain, Vincent-Pierre Berges, Tingfan Wu, Jay Vakil, et al. Where are we in the search for an artificial visual cortex for embodied intelligence? *Advances in Neural Information Processing Systems*, pp. 655–677, 2023.

Norman Mu, Alexander Kirillov, David Wagner, and Saining Xie. SLIP: Self-supervision meets language-image pre-training. In *European Conference on Computer Vision*, pp. 529–544, 2022.

Suraj Nair, Aravind Rajeswaran, Vikash Kumar, Chelsea Finn, and Abhinav Gupta. R3M: A universal visual representation for robot manipulation. In *Proceedings of The 6th Conference on Robot Learning*, pp. 892–909, 2023.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 8748–8763, 2021.

Ilija Radosavovic, Tete Xiao, Stephen James, Pieter Abbeel, Jitendra Malik, and Trevor Darrell. Real-world robot learning with masked visual pre-training. In *Conference on Robot Learning*, pp. 416–426, 2023.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5B: An open large-scale dataset for training next generation image-text models. In *Advances in Neural Information Processing Systems*, pp. 25278–25294, 2022.

Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8430–8439, 2019.

Oriane Siméoni, Gilles Puy, Huy V Vo, Simon Roburin, Spyros Gidaris, Andrei Bursuc, Patrick Pérez, Renaud Marlet, and Jean Ponce. Localizing objects with self-supervised transformers and no labels. *arXiv preprint arXiv:2109.14279*, 2021.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *European Conference on Computer Vision*, pp. 776–794, 2020a.

Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning? *Advances in Neural Information Processing Systems*, pp. 6827–6839, 2020b.

Yonglong Tian, Olivier J. Hénaff, and Aäron van den Oord. Divide and contrast: Self-supervised learning from uncurated data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10063–10074, 2021.

Wouter Van Gansbeke, Simon Vandenhende, Stamatios Georgoulis, and Luc V Gool. Revisiting contrastive methods for unsupervised learning of visual representations. In *Advances in Neural Information Processing Systems*, pp. 16238–16250, 2021.

Shashanka Venkataramanan, Mamshad Nayeem Rizve, João Carreira, Yuki M. Asano, and Yannis Avrithis. Is ImageNet worth 1 video? learning strong image encoders from 1 long unlabelled video. In *International Conference on Learning Representations*, 2024.

Xinlong Wang, Rufeng Zhang, Chunhua Shen, Tao Kong, and Lei Li. 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.

Chen Wei, Haoqi Fan, Saining Xie, Chao-Yuan Wu, Alan Yuille, and Christoph Feichtenhofer. Masked feature prediction for self-supervised visual pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14668–14678, 2022.

Xin Wen, Bingchen Zhao, Anlin Zheng, Xiangyu Zhang, and Xiaojuan Qi. Self-supervised visual representation learning with semantic grouping. In *Advances in Neural Information Processing Systems*, pp. 16423–16438, 2022.

Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 418–434, 2018.

Zhenda Xie, Yutong Lin, Zheng Zhang, Yue Cao, Stephen Lin, and Han Hu. Propagate yourself: Exploring pixel-level consistency for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16684–16693, 2021.

Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. SimMIM: A simple framework for masked image modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9653–9663, 2022.

Tianhe Yu, Deirdre Quillen, Zhanpeng He, Ryan Julian, Karol Hausman, Chelsea Finn, and Sergey Levine. Meta-world: A benchmark and evaluation for multi-task and meta reinforcement learning. In *Conference on Robot Learning (CoRL)*, 2019.

Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 633–641, 2017.

Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. Image BERT pre-training with online tokenizer. In *International Conference on Learning Representations*, 2022.

A Appendix

A.1 Implementation Details

A.1.1 Pre-training

Architecture. We use ViT-B/16 (Dosovitskiy et al., 2021) as our backbone. The projector g and predictor h are 3-layer MLPs with hidden dimension 4096 and output dimension 256.

Optimization. We use AdamW optimizer with a cosine learning rate schedule, peak learning rate of 1.5e-4, and weight decay of 0.05. The learning rate is linearly ramped up during the first 10 epochs to its base value scaled with the total batch size: $lr = lr_{\text{base}} \times \text{batch size}/256$. We train for 800 epochs on 241K-scale datasets and 400 epochs on 1.28M-scale datasets, with a batch size of 1024 distributed across 8 A100 GPUs. For experiments on 4M-scale datasets, we train 200 epochs.

Augmentation and masking. We use the same augmentation strategy as in iBOT (Zhou et al., 2022) except not using small local crops. The masking strategy follows (Zhou et al., 2022), with prediction ratio r uniformly sampled from range $[0.3 - 0.2, 0.3 + 0.2]$.

Hyperparameters. We set $\tau_s = 0.1$, $\tau_t = 0.07$. The number of prototypes is set to 512 for COCO and 1024 for other datasets.

A.1.2 EVALUATION

Linear probing and fine-tuning on ImageNet-1K. We follow (He et al., 2022) for details on ImageNet evaluations. For linear probing, we insert an extra BatchNorm layer without affine transformation between the features and the linear classifier. We train with batch size 4096, initial learning rate 0.1, and optimize using SGD for 90 epochs. We sweep between `[CLS]` token and average pooling and report the best results of pre-trained models. For fine-tuning, We train a linear classifier on frozen features for 100 epochs using SGD with momentum 0.9, batch size 1024, and initial learning rate 1e-3 with cosine decay. For both settings, accuracy is evaluated on a single 224×224 crop.

Semantic segmentation on ADE20K. We use UperNet Xiao et al. (2018) implemented in MMSegmentation following Zhou et al. (2022). Specifically, we fine-tune for 160k iterations with stochastic gradient descent, with a batch size of 16 and weight decay of 0.0005. The learning rate is 0.01 and decays following the poly schedule with power of 0.9 and min_lr of 0.0001.

Object detection and instance segmentation on COCO. COCO object detection and instance segmentation setting also follows Zhou et al. (2022), where the pre-trained model initialized a Cascade Mask R-CNN (Cai & Vasconcelos, 2019). The image scale is $[640, 800]$ during training and 800 at inference. We fine-tune all layers end-to-end on COCO Lin et al. (2014) `train2017` set with the standard $1\times$ schedule and report AP for boxes and masks on the `val2017` set.

Robot manipulation tasks. Following the setup of Hu et al. (2023), the policy network of behavior cloning includes a LayerNorm layer before the MLP. The policy training involves mini-batches of 128 samples, conducted over 20,000 steps with the Adam optimizer set to a learning rate of 0.0001. For each pre-trained vision model and task, we run 3 seeds of BC due to the result's high variability. One-image observation for its comparable performance to stacks of images and higher computational efficiency. All tasks and environments use 224×224 RGB images without proprioceptive input. No image augmentations, such as random shifts, are applied. We employ attentive pooling, as in V-Cond (Karamcheti et al., 2023), which is shown to be the better choice than the default `[CLS]` embedding head and provides better comparisons between pre-trained frozen visual representations.