Understanding Self-Supervised Pretraining with Part-Aware Representation Learning

Jie Zhu[1,2]*, **Jiyang Qi**[5*], **Mingyu Ding**[4,6*], **Xiaokang Chen**[3],
Ping Luo[6], **Xinggang Wang**[5], **Wenyu Liu**[5], **Leye Wang**[1,2], **Jingdong Wang**[7†]
zhujie@stu.pku.edu.cn, {jiyangqi, xgwang, liuwy}@hust.edu.cn, myding@berkeley.edu,
{pkucxk, leyewang}@pku.edu.cn, pluo@cs.hku.hk, wangjingdong@outlook.com
[1]*Key Lab of High Confidence Software Technologies (Peking University), Ministry of Education, China*
[2]*School of Computer Science, Peking University,* [3]*School of AI, Peking University,* [4]*UC Berkeley*
[5]*School of EIC, Huazhong University of Science & Technology,* [6]*University of Hong Kong,* [7]*Baidu*

Reviewed on OpenReview: *https://openreview.net/forum?id=HP7Qpui5YE*

Abstract

In this paper, we are interested in understanding self-supervised pretraining through studying the capability that self-supervised methods learn part-aware representations. The study is mainly motivated by that random views, used in contrastive learning, and random masked (visible) patches, used in masked image modeling, are often about object parts.

We explain that contrastive learning is a *part-to-whole* task: the projection layer hallucinates the whole object representation from the object part representation learned from the encoder, and that masked image modeling is a *part-to-part* task: the masked patches of the object are hallucinated from the visible patches. The explanation suggests that the self-supervised pretrained encoder leans toward understanding the object part. We empirically compare the off-the-shelf encoders pretrained with several representative methods on object-level recognition and part-level recognition. The results show that the fully-supervised model outperforms self-supervised models for object-level recognition, and most self-supervised contrastive learning and masked image modeling methods outperform the fully-supervised method for part-level recognition. It is observed that the combination of contrastive learning and masked image modeling further improves the performance.

1 Introduction

Self-supervised representation pretraining has been attracting a lot of research efforts recently. The goal is to train an encoder that maps an image to a representation from visual contents without the necessity of human annotation. Through this, it is expected that the encoder benefits the downstream tasks, e.g., segmentation and detection.

There are two main frameworks: contrastive learning[1] and masked image modeling. Contrastive learning aims to maximize the agreement of the embeddings of random augmented views from the same image. Masked image modeling partitions an image into masked patches and visible patches, and makes predictions for masked patches from visible patches. Figure 1 gives examples of random views for contrastive learning and masked and visible patches for masked image modeling.

We observe that a random view and a set of masked (visible) patches usually contain a portion of an object. It is also reported in self-supervised learning methods, e.g., DINO (Caron et al., 2021) and iBOT (Zhou et al.,

*Equal contribution.
†Corresponding author.
[1]In this paper, we use contrastive learning (CL) to refer to random-view based methods, e.g., SimCLR, MoCo, and BYOL.

(a) (b) (c) (d) (e)

Figure 1: (a) original image, (b-c) two random crops, and (d-e) masked and visible patches.

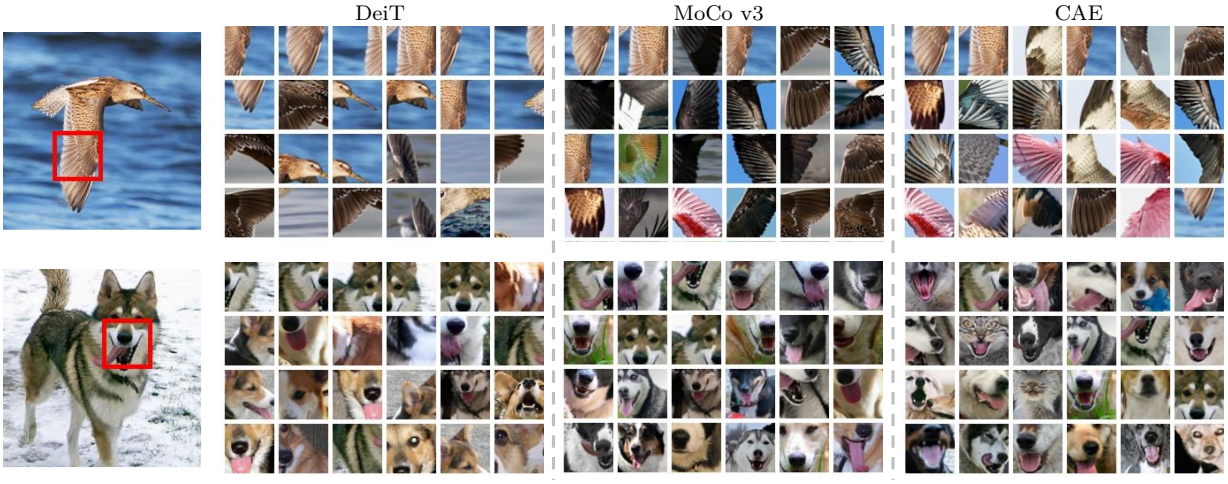

Figure 2: Top-24 patch retrieval results with three frozen encoders of DeiT, MoCo v3, and CAE, by taking the patch in the red box as the query. It can be seen that the retrieved results from CAE and MoCo v3 are about the object part (wing and dog mouth) and more precise than DeiT (about the whole object) implying that self-supervised pretraining methods, CAE and MoCo v3 are stronger at learning part-aware representations than the fully-supervised method DeiT. Details could be found in Sec. 3.

2021), that different attention heads in ViTs can attend to different semantic regions or parts of an object. In light of this, we attempt to understand self-supervised pretraining by studying the capability that the pretrained encoder learns part representations.

We present a *part-to-whole* explanation for typical contrastive learning methods (e.g., SimCLR (Chen et al., 2020), MoCo (Chen et al., 2021), and BYOL (Grill et al., 2020)): the embedding of the whole object is hallucinated from the embedding of the part of the object contained in the random crop through a projection layer. In this way, embeddings of random crops from the same image naturally agrees with each other. Masked image modeling is a *part-to-part* process: the embeddings of the masked patches of the object (a part of the object), are hallucinated from the visible patches (the other part of the object).

We empirically compare the supervised model DeiT (Touvron et al., 2020), which serves as an important baseline for analyzing SSL models, and typical self-supervised representation pretraining methods, including MoCo v3 (Chen et al., 2021), DINO (Caron et al., 2021), CAE (Chen et al., 2022a), MAE (He et al., 2021), BEiT (Bao et al., 2021), and iBOT (Zhou et al., 2021), on object-level recognition (image classification and object segmentation) and part-level recognition (patch retrieval, patch classification, and part segmentation). Figure 2 presents patch retrieval results using the encoders learned through CAE, MoCo v3, and DeiT, implying that the encoders pretrained by CAE and MoCo v3 are able to learn part-aware representations.

Through extensive studies and comparisons, we make the following observations. 1) DeiT outperforms contrastive learning and MIM methods except iBOT in object-level recognition tasks, which may benefit from its explicit object-level supervision. 2) In contrast, self-supervised methods learn better part-aware representations than DeiT. For example, while DeiT is superior to DINO and CAE by 0.4% and 2.3% on ADE20K object segmentation, DINO and CAE outperform DeiT by 1.6% and 1.1% on ADE20K part segmentation, respectively. 3) In contrastive learning, the encoder can learn part-aware information, while

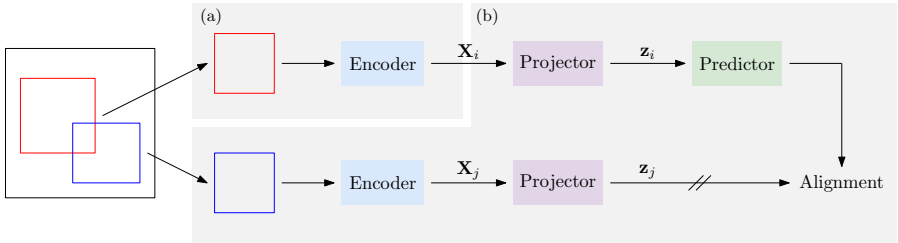

Figure 3: The pipeline of a typical contrastive learning approach. Two augmented views, red box and blue box, are generated from the original image. The augmented view in red is fed into the encoder and the projector, and then the predictor (which does not appear in earlier works like MoCo (Chen et al., 2021) and SimCLR (Chen et al., 2020)), and the view in blue is fed into the encoder and the projector. The two outputs are expected to be aligned. The gradient is stopped for the bottom stream.

the projected representation tends to be more about the whole object. The evidence could be found in part retrieval experiments on MoCo v3, DINO, and iBOT. 4) The MIM method CAE shows good potential in part-aware representation learning. Interestingly, the method combines contrastive learning and MIM is promising, e.g., iBOT learns better representations at both object and part levels.

This paper presents the following contributions:

- We study the capability of learning part-aware representations as a way of understanding self-supervised representation pertaining based on both qualitative analysis and empirical results.
- We explain masked image modeling as a part-to-part task and contrastive learning as a part-to-whole task, and speculate that self-supervised pretraining has the potential for learning part-aware representations.
- We empirically compare several pretrained models on object-level and part-level recognition tasks, showing interesting findings with supporting evidence of the capability of part-aware representation learning for self-supervised learning.

2 Related Work

Contrastive learning (CL). Contrastive pretraining has been an intense academic field in the CNN era. In this work, we use it to refer to methods for comparing random views (Caron et al., 2020; Chen et al., 2020; Zbontar et al., 2021; Xie et al., 2021b; Chen et al., 2021; Caron et al., 2021), including some instance discrimination work such as (Grill et al., 2020; Chen & He, 2021; Bardes et al., 2021). A representative work, SimCLR (Chen et al., 2020) learns representations through maximizing agreement between different views of the same image in the latent space. BYOL (Grill et al., 2020) uses two asymmetrical networks to bootstrap latent representation without negative samples involved during the interaction. As vision transformer (ViT) (Dosovitskiy et al., 2021) shows excellent performance via supervised learning, it is adopted subsequently in contrastive pertaining, and numerous outstanding works are proposed. For example, MoCo v3 (Chen et al., 2021) observes the hidden instability while training self-supervised ViT and solves it by using a fixed random patch projection. DINO (Caron et al., 2021) explores new properties derived from self-supervised ViT and accordingly designs a learning strategy interpreted as a form of self-distillation with no labels.

Masked image modeling (MIM). Masked image modeling is another self-supervised pretraining paradigm that attracts much attention recently. BEiT (Bao et al., 2021) follows masked language modeling in the natural language process (NLP) area and predicts tokens via mapping image patches by d-VAE (Ramesh et al., 2021). PeCo (Dong et al., 2021) boosts BEiT by taking into consideration more semantic information in visual tokens. MAE (He et al., 2021) learns rich hidden information by directly performing masked image reconstruction in RGB color space using ViT while SimMIM (Xie et al., 2021c) uses Swin-transformer (Liu et al., 2021). CAE (Chen et al., 2022a) adds a regressor between encoder and decoder, which is designed to align unmasked patches with masked ones, leading to a pure context encoder. Recently, a trend that combines MIM with siamese frameworks has surfaced and showed encouraging results including MST (Li

et al., 2021), SplitMask (El-Nouby et al., 2021), iBOT (Zhou et al., 2021), dBOT (Liu et al., 2022), and SIM (Tao et al., 2022).

Understanding self-supervised contrastive pretraining. The studies on understanding contrastive pretraining (Saunshi et al., 2022; Chen et al., 2022b; Zhong et al., 2022; Wei et al., 2022) mainly focus on random augmentations (views), contrastive loss function and its variants under the assumption that: the augmentations of inputs from the same class have significant overlap in the representation space, but there is little overlap for inputs from different classes. Our work is complementary to these studies. Inspired by the observation that random views usually contain a portion of an object, and methods (Caron et al., 2021; Zhou et al., 2021) show that different attention heads in ViTs can attend to different semantic regions of an object, we investigate what the encoder and the projector do in typical self-supervised contrastive pretraining. We speculate that the pretraining task is a part-to-whole problem, predicting the representation of the whole object through the projector from the representation (obtained from the encoder) of the part of an object. We use empirical results to verify our analysis.

Understanding self-supervised masked image modeling. The comparison of attention in different layers between the pretrained models from MIM and the supervised approach is conducted: MIM pretraining brings locality to the trained model with sufficient diversity on the attention heads (Xie et al., 2022a). Consistent with the analysis in NLP, empirical studies are conducted in Xie et al. (2022b) to verify that MIM benefits from larger models, more data, and longer training. CAE (Chen et al., 2022a) gives the comparison between contrastive and MIM and shows MIM cares about all patches and thus achieves better results for fine-tuning. Cao et al. (2022) provides a mathematical understanding of MIM. Kong & Zhang (2022) points out that the learned occlusion invariant feature contributes to the success of MIM. In this work, we speculate that masked image modeling is a *part-to-part* process: the embeddings of the masked part of the object are hallucinated from the visible part using the position information of the masked patches, leading to better part-aware representation than the supervised model DeiT (Touvron et al., 2020). More discussions can be found in Appendix A.8.

3 Understanding Contrastive Learning and Mask Image Modeling

In this section, we briefly review representative formulations of CL and MIM, and provide qualitative analysis of their part-aware explanations.

3.1 Contrastive Learning

Contrastive learning aims to learn the encoder through maximizing the agreement between differently augmented views of the same image in the representation space. An example pipeline is depicted in Figure 3. Given an image I, the augmentations, e.g., random cropping, random color distortion, and random Gaussian blur, are applied to generate a set of N augmented views, $\{\mathsf{V}_1, \mathsf{V}_2, \cdots, \mathsf{V}_N\}$. An augmented view V_n is fed into an encoder Encoder, generating the encoded representation $\mathbf{x}_n$, and followed by a projector, generating the projection $\mathbf{z}_n$. The basic goal is to maximize the agreement between the projections $\{\mathbf{z}_1, \mathbf{z}_2, \cdots, \mathbf{z}_N\}$, i.e., minimize the loss

$$\mathcal{L}_{\text{CPT}} = \sum\nolimits_{i=1}^{N} \sum\nolimits_{j=1}^{N} \ell\{\text{Projector}(\text{Encoder}(\mathsf{V}_i)), \text{Projector}(\text{Encoder}(\mathsf{V}_j))\}. \tag{1}$$

In the formulation with a contrastive loss, the agreement between the projections of random augmentations from different images is minimized.

Part-to-whole prediction explanation. Let us consider two crops randomly sampled from the original image (see the examples given in Figure 1(b-c)). The encoded representation of the first crop is expected to describe a part of the object dog; the encoded representation of the second crop is expected to describe another part of the object dog[2]. The two representations are related but different. Contrastive learning methods project the two encoded representations into two projected representations that are expected to

[2]In this paper, we mainly study the capability of learning representations about objects and parts, and leave the study of representations of the background as the future work.

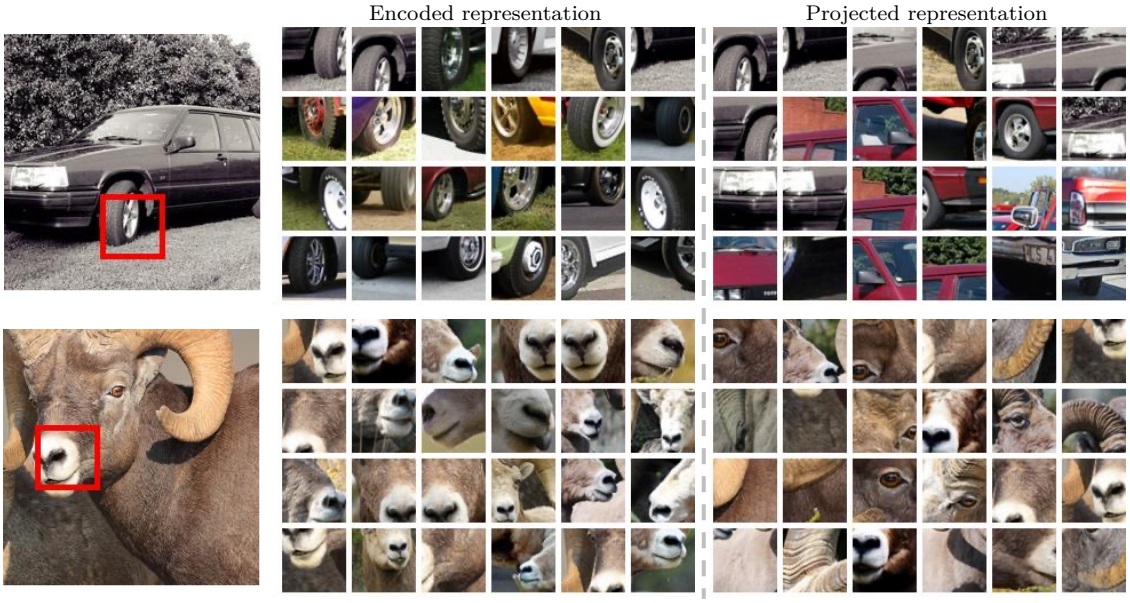

Figure 4: Illustration of patch search results using encoded representations and projections (pretrained with MoCo v3 as). Left: patch search results with encoded representations. Right: patch search results with projections. In each result, the small patch encircled by the red box is taken as the query. It can be seen that for encoded representations, the returned patches are about the same part, and for projections, the result patches are about the same object, verifying the *part-to-whole* hypothesis.

agree. We hypothesize that *the projection process maps the encoded part representation to the representation of the whole object*[3]. Through this way, the projected representations will agree to different views from the same image [4]. It is assumed that the part-to-whole projection is more reliable if the encoded representation is semantically richer and is able to describe the part information. The part-to-whole process suggests that the encoder pretrained by contrastive learning methods is potentially capable of learning part-aware representations.

Figure 4 provides patch search results of a representative contrastive learning method MoCo v3 (Chen et al., 2021) based on the encoded representations before and after the projections. The visualized patch retrieval results in Figure 4 as well as those in Figure 2 are obtained based on ImageNet (Deng et al., 2009) validation set. Concretely, we uniformly crop 49 patches sized 56×56 using a stride of 28 from each pre-processed 224×224 validation image in ImageNet. With all the cropped patches from the validation set, we select one patch as a query and find the top 24 patches with the highest cosine similarity with it.

One can see Figure 4 that the results through the encoded representations are mainly about the local part, and the results through the projections tend to include the other parts of the same object. In other words, the projections tend to be about the whole object. The search results verify the part-to-whole hypothesis. Similar observations are also shown in Chen et al. (2022b) where the embedding space tends to preserve more equivariance and locality, while the projection has more invariance.

3.2 Masked Image Modeling

Mask image modeling is the task of predicting some parts of an image from the remaining parts. An augmented view of an image is partitioned into patches, $\mathcal{R} = \{\mathsf{R}_1, \mathsf{R}_2, \ldots, \mathsf{R}_M\}$. The task is to predict a subset of patches $\mathcal{R}_m$, named masked patches, from the remaining patches $\mathcal{R}_v$, named visible patches. Considering contrastive learning that explicitly compares representations of random views, we take context autoencoder

[3]It is said that different parts have common causes in the external world (Becker & Hinton, 1992). Our hypothesis is that the common cause is the whole object.

[4]For methods that use local-global crops, e.g., DINO, contains a range of $0.05 - 0.4$ for local crops and $0.4 - 1$ for global crops, which in our hypothesis are expected to describe *smaller* and *larger* parts of the object, respectively.

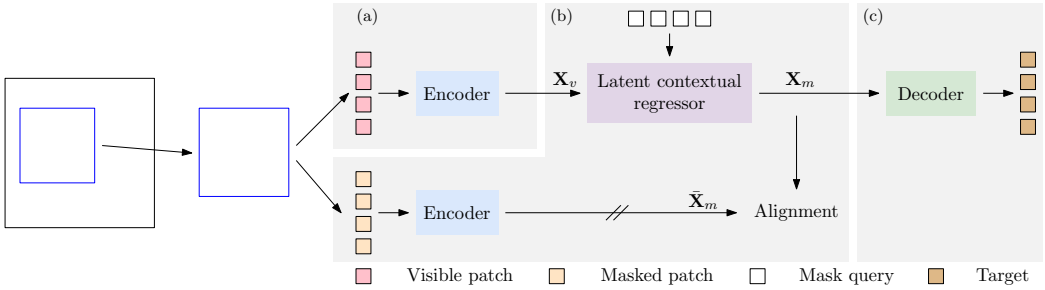

Figure 5: The pipeline of an MIM approach, context autoencoder (CAE). An augmented view (in blue) of the image is partitioned into visible and masked patches. The CAE approach feeds visible patches into the encoder and extracts their representations $\mathbf{X}_v$ and then completes the pretext task by predicting the representations $\mathbf{X}_m$ of the masked patches from the visible patches in the encoded representation space with latent contextual regressor and alignment constraint, and mapping predicted representations $\mathbf{X}_m$ of masked patches to the targets. The pretrained encoder in (a) is applied to downstream tasks by simply replacing the pretext task part (b, c) with the downstream task completion part.

(CAE) (Chen et al., 2022a) as an example that explicitly predicts the encoded representations of the masked patches from the encoded representations of the visible patches[5].

One goal of CAE (illustrated in Figure 5), which we call masked representation modeling (MRM), is to maximize the agreement between the predictions of the representations of masked patches (through a regressor) and the representation of masked patches computed from the encoder by minimizing the loss

$$\ell_{\text{MRM}}(\text{Regressor}(\text{Encoder}(\mathcal{R}_v)), \text{Encoder}(\mathcal{R}_m)). \tag{2}$$

Here, we do not include the positional embeddings of masked and visible patches for clarity. It is noted that MRM differs from contrastive learning: MRM does not compare multiple random views, but compares the regressed representations for masked patches and the encoded representations of masked patches. In addition, there is another loss for target prediction (reconstruction) for the masked patches, which is commonly used in masked image modeling (MIM) methods:

$$\ell_{\text{MIM}}(\text{Decoder}(\text{Regressor}(\text{Encoder}(\mathcal{R}_v))), \text{Target}(\mathcal{R}_m)), \tag{3}$$

where $\text{Target}(\mathcal{R}_m)$ is a function to map the masked patches to the targets, e.g., d-VAE (Ramesh et al., 2021) token used in CAE and BeiT (Bao et al., 2021), or normalized RGB values used in MAE (He et al., 2021).

Part-to-part prediction explanation. The masked image modeling approaches, including CAE, MAE, and BEiT, make use of the positions of masked patches for making predictions for masked patches from visible patches. The visible patches and masked patches often contain different parts of an object. In other words, MIM aims to predict the masked part of an object from the visible part. We name this a *part-to-part* process. There are two part-to-part tasks: one is to reconstruct the part targets from the visible part representations (MAE and CAE) or from the visible part raw pixels (BEiT), and the other one is to regress the masked part representations (CAE). The part-to-part process suggests that the encoder pretrained by MIM methods is potentially capable of learning part-aware representations. Figure 2 illustrates the capability with the patch retrieval results. More visualizations can be found in Appendix A.9.

4 Experiments

We study seven representative methods with the same ViT-B encoder, including a supervised method DeiT (Touvron et al., 2020) that serves as an important baseline for analyzing SSL models, contrastive learning methods MoCo v3 (Chen et al., 2021), DINO (Caron et al., 2021); masked image modeling (MIM) methods BEiT (Bao et al., 2021), MAE (He et al., 2021), and CAE (Chen et al., 2022a); and iBOT (Zhou

[5]Some MIM methods, such as BEiT (Bao et al., 2021) and MAE (Masked Autoencoder) (He et al., 2021) do not have an explicit process to predict the encoded representations of masked patches, instead, directly reconstruct the targets.

Table 1: Top-1 accuracy with linear probing, and attentive probing (Chen et al., 2022a), on the ImageNet classification benchmark (Deng et al., 2009).

Method	Linear	Attentive
Supervised Model:		
DeiT	81.8	81.8
Contrastive Learning:		
MoCo v3	76.2	77.0
DINO	77.3	77.8
Masked Image Modeling (MIM):		
BEiT	41.8	51.9
MAE	67.8	74.2
CAE	70.4	77.1
Contrastive Learning + MIM:		
iBOT	79.5	79.8

Table 2: Object retrieval results on CIFAR10 and the cropped object of COCO. The "Encoded" and "Projected" refer to the encoded and projected representations.

Methods	Object Retrieval (AP)			
	CIFAR10		COCO	
	Encoded	Projected	Encoded	Projected
Supervised Model:				
DeiT	71.3	–	71.2	–
Contrastive Learning:				
MoCo v3	40.6	56.0	47.7	55.9
DINO	31.3	61.5	38.0	57.3
Masked Image Modeling (MIM):				
BEiT	14.6	–	11.7	–
MAE	15.9	–	13.8	–
CAE	39.1	–	46.0	–
Contrastive Learning + MIM:				
iBOT	40.2	60.0	53.2	55.8

et al., 2021) that combines contrastive learning and MIM. We take the training epochs specified in each work to ensure that all compared models are properly trained and use publicly available checkpoints [6]: 300 for DeiT, 300 (600[7]) for MoCo v3, 400 (1600[7]) for DINO and iBOT, 800 for BEiT, and 1600 for MAE and CAE. Frozen encoders are used in all experiments to understand what these different representation pretraining methods learn. More details about models can be found in Appendix A.1.

4.1 Object-Level Recognition

We benchmark three widely-studied object-level recognition, i.e., image classification, object retrieval, and semantic segmentation to show the capability that the pretrained encoder learns object-level representations.

Image classification. We report the linear probing, and attentive probing results of the selected models on ImageNet (Deng et al., 2009). For attentive probing, we follow the protocol in CAE (Chen et al., 2022a) that append a cross-attention layer together with a batch normalization layer and a linear classifier.

We have the following observations from Table 1. ❶ The supervised model, DeiT performs better than self-supervised models at object-level recognition. ❷ The models that leverage contrastive learning, i.e., MoCo, DINO, and iBOT, show superior linear probing performance than MIM-based models, demonstrating they contain more object-aware high-level semantics. ❸ MIM-based models, e.g., CAE, show inferior results in linear probing while competitive results with contrastive-based methods in attentive probing. The reason might be that MIM is capable of attending to all the regions, including non-object regions in an image, thus needs a spatial feature selection step to attend to the object part, which is pointed out in Chen et al. (2022a). BEiT and MAE perform inferior in linear and attentive probes, implying that the two methods are less capable of learning semantics.

Object-level retrieval. We evaluate the object-level retrieval performance on two datasets, i.e., CIFAR10 (Krizhevsky et al., 2009) and COCO (Lin et al., 2014). For CIFAR10, we directly use each image (resized to 224×224) to perform the retrieval task. For COCO, we build the retrieval database by leveraging the bounding box (that is approximately square) in the annotation file to crop the objects and resizing to 224×224. The retrieval results are reported in Table 2. One can see that DeiT obtains the best performance over self-supervised methods in both datasets, showing stronger object-aware capability. Besides, we also notice that for part-to-whole SSL methods, e.g., MoCo v3, the projected feature has higher accuracy than the

[6]Considering that the training protocols of different models vary widely in most respects (e.g., batch size, augmentation, and ViT-related tricks to prevent training instability and crashes), it is unreasonable to use the identical setup. For example, DINO uses a centering and sharpening strategy to avoid collapse.

[7]The number of effective training epochs introduced in Zhou et al. (2021).

encoded one, possibly because the projection boosts the object-aware representations, which further supports our part-to-whole hypothesis.

Object-level semantic segmentation. We perform linear evaluation on ADE20K (Zhou et al., 2019) to show the object-level semantic capabilities of the pretrained models. A $4\times$ bilinear interpolation and a single 1×1 convolutional layer for pixel labeling are attached to the frozen encoder. The learning rate ($4e-4$), training iterations ($160k$), and batch size (16) among all the experiments maintain the same during training for fair comparisons. The input size is set to 512×512 following previous works (Bao et al., 2021; He et al., 2021; Chen et al., 2022a; Zhou et al., 2021).

Table 3: Linear evaluation of ADE20K (Zhou et al., 2019) object-level semantic segmentation (150 classes) using $4\times$ upsampling and a single 1×1 convolutional layer on frozen backbones.

Method	mIoU	mAcc	aAcc
Supervised Model:			
DeiT	34.9	44.2	75.4
Contrastive Learning:			
MoCo v3	34.7	43.9	75.9
DINO	34.5	43.5	76.1
Masked Image Modeling (MIM):			
BEiT	17.8	23.7	64.9
MAE	27.1	34.8	71.6
CAE	32.6	42.2	75.2
Contrastive Learning + MIM:			
iBOT	38.3	47.4	78.1

We can see from Table 3 that the supervised model DeiT outperforms all self-supervised models except iBOT, including contrastive learning and MIM methods on ADE20K object-level segmentation. This implies that these self-supervised models are not strong at object-level understanding, which is consistent with the observations for image classification. iBOT (Zhou et al., 2021), as a combination of contrastive learning and MIM, shows surprisingly better performance than the supervised model DeiT on ADE20K, implying the power of combining contrastive learning and masked image modeling for downstream tasks.

4.2 Part-Level Recognition

Self-supervised methods like iBOT (Zhou et al., 2021) and DINO (Caron et al., 2021) qualitatively show that different attention heads in ViTs can attend to different semantic regions of an object. We conduct the quantitative evaluation for part-aware representation obtained by pretrained models that is not well explored before, through three part-level recognition tasks, part retrieval, part classification, and part segmentation.

Table 4: Part retrieval and classification results on the cropped part patches of CUB-200-2011 and COCO. The “Encoded" and “Projected" refer to the encoded and projected representations. “Linear" and “Attentive" columns denote the linear probing and attentive probing accuracy, respectively.

Methods	Part Retrieval (AP)				Part Classification (Acc)			
	CUB-200-2011		COCO		CUB-200-2011		COCO	
	Encoded	Projected	Encoded	Projected	Linear	Attentive	Linear	Attentive
Supervised Model:								
DeiT	35.0	–	44.1	–	90.9	92.9	88.5	91.4
Contrastive Learning:								
MoCo v3	50.8	28.4	52.3	36.8	93.8	96.0	92.4	95.3
DINO	48.9	31.7	51.8	41.2	93.2	95.2	91.7	94.5
Masked Image Modeling (MIM):								
BEiT	27.9	–	35.3	–	55.4	86.5	69.3	86.5
MAE	28.5	–	37.1	–	86.9	92.8	88.0	93.9
CAE	58.0	–	57.0	–	89.5	95.8	91.1	95.5
Contrastive Learning + MIM:								
iBOT	49.3	31.2	59.2	41.5	93.8	95.8	92.1	95.1

Part retrieval. We conduct part retrieval experiments on two datasets, CUB-200-2011 (Wah et al., 2011) and COCO (Lin et al., 2014), which provide both the positions and corresponding categories of the keypoints. We build the part patch databases by cropping the patches where these keypoints are located at the center. We consider four and three keypoints from the two datasets, respectively. For each keypoint, we find the minimum L2 distance (d) from the distances between it and all the other keypoints in the same image, then

crop a $d \times d$ patch centered at this keypoint and resize it to 224×224 [8]. For each method, we directly use the pretrained encoders to extract features, without additional training processes, and take the better one from the class token or the average embedding of all patch tokens as the extracted representation. With each patch as the query patch, we attempt to retrieve patches belonging to the same category by calculating the cosine similarity between its representation and all the other patches' in the dataset and utilize the average precision (AP) as the retrieval metric. Finally, we average all the obtained AP scores (with all patches respectively taken as the query patch for retrieval) as the final retrieval score to evaluate the retrieval performance of the method.

The results are provided in Table 4. We have the following observations. ❶ Self-supervised models except BEiT and MAE outperform the supervised model DeiT, indicating the capability that contrastive learning and CAE learn part-aware representations. BEiT and MAE perform inferior, consistent to the observations in ImageNet classification in Table 1. ❷ iBOT performs the best, and the reason might be that the capability of learning part-aware representations is boosted by making use of both contrastive learning and masked image modeling.

We also report the part retrieval performance of the projected representations of contrastive learning methods in Table 4. The performance is much lower than the encoded representations. This provides an extra evidence for the part-to-whole hypothesis of contrastive learning: the projected representations are more about the whole object.

Finally, we visualize some patch retrieval results of the encoded representations on the ImageNet validation set in Figures 6 (More visualization is presented in Appendix A.9). It is observed that the retrieved patches of self-supervised methods are generally more about the semantics of the query part than that of DeiT. The results demonstrate that the encoded representations of DeiT focus more on object-level semantics, while the encoded representations of these self-supervised methods are more about part-level semantics. Among these methods, the retrieved patches of MAE have less semantic correlation but often share similar hues.

Part classification. We further conduct part classification experiments on the datasets used for part retrieval, requiring pretrained models to assign the patches into the corresponding category label. We consider two kinds of extra learnable layers, linear probing and attentive probing, for classification. For linear probing, we learn a supervised linear classification layer on the extracted class token of the frozen encoders. While for attentive probing, following Chen et al. (2022a), a cross attention module and a batch normalization layer without affine transformation are additionally inserted between the encoder and the linear classifier. And a new learnable class token is taken as the query of the cross attention module, to replace the original class token extracted by the frozen encoder. We use SGD optimizer with a learning rate of 0.4 and 0.04 for linear probing and attentive probing, respectively. For both linear probing and attentive probing, the models are trained for 90 epochs. And the momentum of SGD is set to 0.9, the weight decay is set to 0, and the batch size is set to 1024.

The results in Table 4 show that: ❶ While DeiT performs the best in the image classification task (see Table 1), for part classification, contrastive-based methods like MoCo v3, DINO, and iBOT outperform DeiT by more than 2% under both linear and attentive probing settings. ❷ Though MIM-based models CAE and MAE are inferior to DeiT in object-level classification (e.g., more than 10% and 4% lower in linear and attentive probing), they show competitive performance in linear probing and higher results than DeiT in attentive probing, demonstrating they learn better part-aware representations. ❸ BEiT is inferior to other works, and iBOT has good performance, implying that the probing quality of pretrained encoders is a good indicator for downstream performance.

Part segmentation. We perform part-level linear semantic segmentation to study the finer-grained part representation modeling capability of different pretraining paradigms on three widely used datasets: ADE20K-Part (Zhou et al., 2019) containing 209 parts from the ADE20K dataset (Zhou et al., 2019; Zhu et al., 2023), Pascal-Part (Chen et al., 2014) including 193 part categories, and LIP (Gong et al., 2017) consisting of 19 semantic human part labels. See Appendix A.2 for more dataset details. Similar to the object-level semantic segmentation experiments, linear evaluation is employed here. We maintain the same training

[8]We explain the approach of resizing patches to full image resolution instead of other methods to extract part representations in Appendix A.3

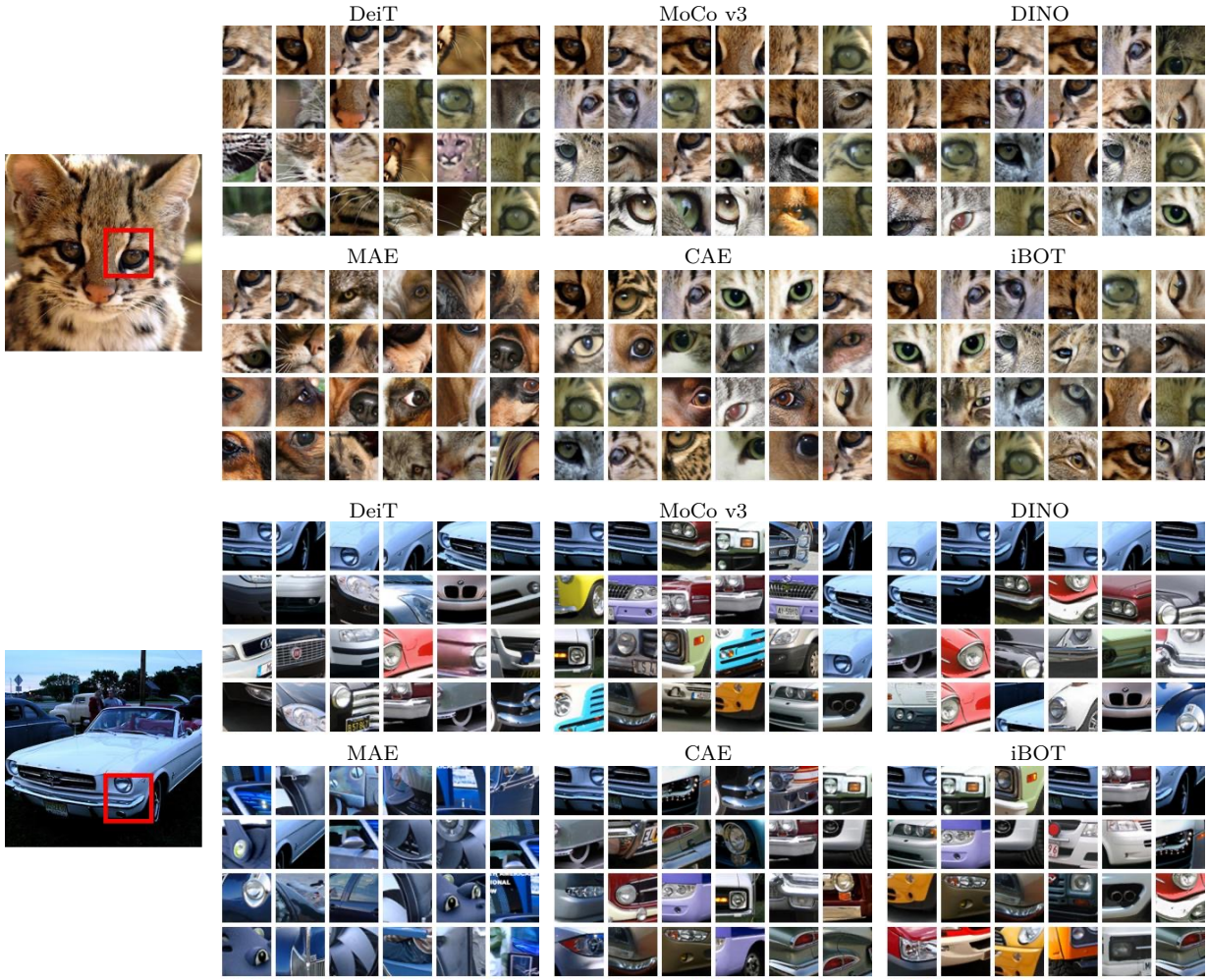

Figure 6: Patch retrieval comparisons of encoded representations on cropped patches from ImageNet.

Table 5: Part-level linear semantic segmentation results (%) on the ADE20K-Part, Pascal-Part, and LIP datasets.

Methods	ADE20K-Part 209 Part Classes			Pascal-Part 193 Part Classes			LIP 19 Part Classes		
	mIoU	mAcc	aAcc	mIoU	mAcc	aAcc	mIoU	mAcc	aAcc
Supervised Model:									
DeiT	27.3	34.7	69.2	27.4	36.1	65.8	41.4	52.6	73.5
Contrastive Learning:									
MoCo v3	27.1	34.7	70.1	27.1	35.8	66.0	41.9	53.0	74.5
DINO	28.9	36.8	70.3	27.8	36.5	66.4	41.0	51.9	74.0
Masked Image Modeling (MIM):									
BEiT	18.6	25.8	58.2	14.8	21.4	47.0	27.2	36.5	60.1
MAE	26.3	35.0	67.3	24.3	32.9	61.5	38.2	48.7	71.3
CAE	28.4	36.9	71.1	27.8	37.0	66.3	43.7	55.1	75.9
Contrastive Learning + MIM:									
iBOT	32.2	40.0	73.4	30.7	40.0	69.7	44.6	55.7	76.6

protocols including learning rate ($4e-4$), training iterations ($160k$), and batch size (16) for all methods for fair comparisons [9]. For ADE20K, the input size is set to 512×512 following previous works (Bao et al., 2021; He et al., 2021; Chen et al., 2022a; Zhou et al., 2021). For Pascal-Part, we adopt 480×480 as image input

[9]We also consider to do some hyperparameter tuning per method for more fair comparisons. See Appendix A.5

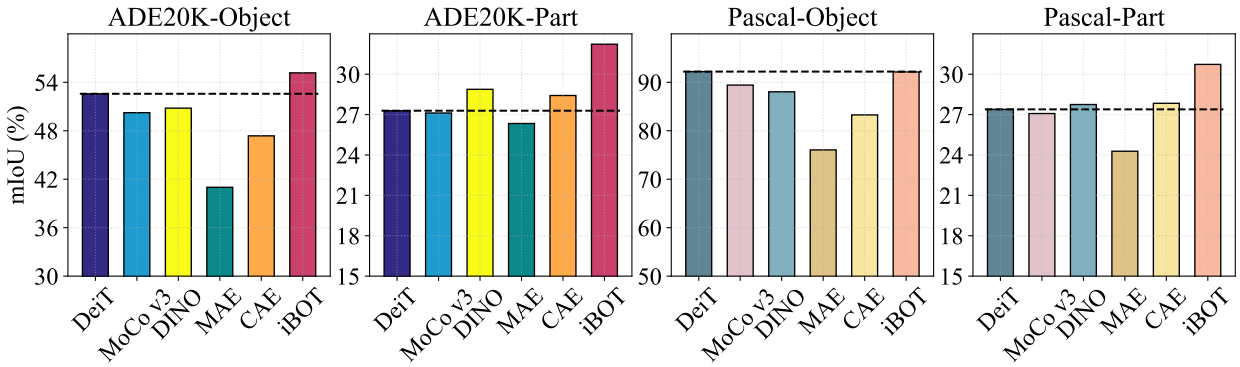

Figure 7: Comparisons between object-level and part-level semantic segmentation on ADE20K and Pascal-Part datasets. Though the supervised DeiT is superior over self-supervised models (i.e., MoCo v3, DINO, MAE, CAE) on object-level segmentation, it is generally inferior to self-supervised models on part segmentation, demonstrating self-supervised methods learn good part-aware representations. iBOT enjoys the benefits of contrastive learning and MIM. See Appendix A.4 for detailed results.

resolution following Contributors (2020). As for LIP, we use the same input size (320×320) proposed in LIP (Gong et al., 2017).

The results are reported in Table 5 with the following observations. ❶ Contrastive learning models, i.e., MoCo v3 and DINO, achieve competitive performance with the supervised model DeiT: DINO outperforms DeiT on ADE20K-Part and Pascal-Part, and MoCo v3 outperforms DeiT on LIP. ❷ The MIM model CAE, outperforms DeiT by large margins on all three datasets, e.g., 1.1% on ADE20K-Part and 2.3% on LIP, indicating CAE learns good part-aware representations. Similar to part retrieval, BEiT and MAE perform inferior in absolute metrics possibly due to pretraining quality in representation encoding [10]. Also, it is worth mentioning that the performance gap (1.0%) between DeiT and MAE on part-level segmentation is significantly narrowed compared to that (11.6%) of object-level segmentation as shown in Figure 7. This indicates that although MAE does not outperform DeiT on part-level tasks, it tends to learn more about parts than DeiT. Similar observations on BEiT can also be found in Appendix A.4. ❸ Compared with object-level segmentation results in Table 3, DeiT learns better object-level semantics by explicit supervision than both contrastive learning and MIM, however, it is generally inferior to self-supervised models on part segmentation. ❹ The model iBOT, which leverages both contrastive learning and MIM, outperforms all other works on three datasets, demonstrating its powerful capability in learning finer part-level semantics. Combining the two self-supervised learning techniques is thus a promising direction.

In summary, we show that self-supervised methods are potentially capable of learning part-aware representations. Among them, CAE is a representative MIM work, showing good performance by explicitly predicting the encoded representations of the masked patches in the encoding space; contrastive learning methods MoCo v3 and DINO outperform BEiT and MAE; and iBOT performs the best [11]. The observations are evidenced by three part-based segmentation benchmarks consistently.

4.3 Observation Summary between Object-Level and Part-Level Segmentation

We conduct both object-level and part-level linear semantic segmentation on different hierarchies of the same dataset. Considering that the 209 classes in ADE20K-Part are basically chosen from 59 object classes, we denote the 59-object dataset as ADE20K-Object. Similarly, Pascal-Object consists of 16 object categories, corresponding to the 193 part categories in Pascal-Part.

The results in Figure 7 show that: although self-supervised models (except iBOT) are inferior to the supervised DeiT on ADE20K-Object and Pascal-Object, they significantly narrow the performance gap on ADE20K-Part

[10] We discuss the reasons in detail and add human pose estimation experiments in Appendix A.6

[11] The iBOT authors verified that introducing MIM brings richer semantics to the patch tokens in Sec. 4.3.1 of their paper. This could be the reason for the superior performance of object-level and part-level tasks observed in iBOT where MIM and CL are combined.

and Pascal-Part and even outperform DeiT, demonstrating self-supervised methods tend to learn more about parts and potentially produce good part-aware representations. Similar observations could be found from the object classification in Table 1 and part classification in Table 4.

In comparison to contrastive learning, CAE shows a stronger capability of learning part-aware representations, and a weaker capability of learning object-level semantics. The superiority of iBOT, a combination of contrastive learning and masked image modeling, demonstrates that it enjoys the benefits of contrastive learning and masked image modeling. We have outlined several key takeaways in Appendix A.7 derived from our experiments and analysis, which we hope could inspire further works in self-supervised learning.

Additionally, although our main focus is not on in-the-wild datasets like COCO and OpenImages, we conduct a preliminary exploration on them. Following the same setting as in our paper, we conducted a quick experiment with a transformer-based Long-seq-MAE (Hu et al., 2022) that is pretrained on COCO. Our experiments show that its part retrieval (32.3% AP) on CUB-200 and LIP segmentation results (45.89% mIoU) outperform MAE (28.5% AP and 38.2% mIoU) pretrained on ImageNet1k. It also narrows the gap between DeiT (35.0% AP) in part retrieval and outperforms DeiT (41.4% mIoU) in part segmentation. These results indicate that pretraining on more in-the-wild datasets has the potential to learn better part representations as cropping can produce more parts of multiple objects in one image. We leave it as an interesting future direction.

5 Conclusion

We attempt to study the capability of learning part-aware representations of self-supervised representation pretraining methods. We provide speculations for contrastive learning and masked image modeling: part-to-whole and part-to-part, with empirical results justifying the speculations. Our study presents an aspect to understand what self-supervised representation pretraining methods learn.

Limitation and future work. The part-aware representation learning capability is one of the properties of self-supervised pretraining. There should be other characteristics that are worth exploring. The proposed approach has no ethical or societal issues on its own, except those inherited from computer vision.

Acknowledgments

This work is supported by National Science Foundation of China (NSFC) Grant No. 61972008.

References

Hangbo Bao, Li Dong, and Furu Wei. BEiT: BERT pre-training of image transformers. *arXiv:2106.08254*, 2021.

Adrien Bardes, Jean Ponce, and Yann LeCun. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. *arXiv preprint arXiv:2105.04906*, 2021.

Suzanna Becker and Geoffrey E. Hinton. A self-organizing neural network that discovers surfaces in random-dot stereograms. *Nature*, 355(6356):161–163, 1992.

Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Coco-stuff: Thing and stuff classes in context. In *CVPR*, pp. 1209–1218, 2018.

Shuhao Cao, Peng Xu, and David A Clifton. How to understand masked autoencoders. *arXiv preprint arXiv:2202.03670*, 2022.

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *arXiv preprint arXiv:2006.09882*, 2020.

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

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, volume 119 of *Proceedings of Machine Learning Research*, pp. 1597–1607. PMLR, 2020.

Xianjie Chen, Roozbeh Mottaghi, Xiaobai Liu, Sanja Fidler, Raquel Urtasun, and Alan Yuille. Detect what you can: Detecting and representing objects using holistic models and body parts. In *CVPR*, pp. 1971–1978, 2014.

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*, 2022a.

Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *CVPR*, pp. 15750–15758, 2021.

Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *ICCV*, pp. 9640–9649, 2021.

Yubei Chen, Adrien Bardes, Zengyi Li, and Yann LeCun. Intra-instance vicreg: Bag of self-supervised image patch embedding. *arXiv preprint arXiv:2206.08954*, 2022b.

MMSegmentation Contributors. MMSegmentation: Openmmlab semantic segmentation toolbox and benchmark. https://github.com/open-mmlab/mmsegmentation, 2020.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pp. 248–255. Ieee, 2009.

Mingyu Ding, Xiaochen Lian, Linjie Yang, Peng Wang, Xiaojie Jin, Zhiwu Lu, and Ping Luo. Hr-nas: Searching efficient high-resolution neural architectures with lightweight transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2982–2992, 2021.

Xiaoyi Dong, Jianmin Bao, Ting Zhang, Dongdong Chen, Weiming Zhang, Lu Yuan, Dong Chen, Fang Wen, and Nenghai Yu. Peco: Perceptual codebook for bert pre-training of vision transformers. *arXiv preprint arXiv:2111.12710*, 2021.

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 *ICLR*. OpenReview.net, 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.

M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2010 (VOC2010) Results. http://www.pascal-network.org/challenges/VOC/voc2010/workshop/index.html, 2010.

Ke Gong, Xiaodan Liang, Dongyu Zhang, Xiaohui Shen, and Liang Lin. Look into person: Self-supervised structure-sensitive learning and a new benchmark for human parsing. In *CVPR*, July 2017.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent: A new approach to self-supervised learning. *arXiv preprint arXiv:2006.07733*, 2020.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. *arXiv preprint arXiv:2111.06377*, 2021.

Ronghang Hu, Shoubhik Debnath, Saining Xie, and Xinlei Chen. Exploring long-sequence masked autoencoders. *arXiv preprint arXiv:2210.07224*, 2022.

Xiangwen Kong and Xiangyu Zhang. Understanding masked image modeling via learning occlusion invariant feature. *arXiv preprint arXiv:2208.04164*, 2022.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Zhaowen Li, Zhiyang Chen, Fan Yang, Wei Li, Yousong Zhu, Chaoyang Zhao, Rui Deng, Liwei Wu, Rui Zhao, Ming Tang, et al. Mst: Masked self-supervised transformer for visual representation. *Advances in Neural Information Processing Systems*, 34:13165–13176, 2021.

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 *ECCV*, pp. 740–755. Springer, 2014.

Xingbin Liu, Jinghao Zhou, Tao Kong, Xianming Lin, and Rongrong Ji. Exploring target representations for masked autoencoders. *arXiv preprint arXiv:2209.03917*, 2022.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, pp. 10012–10022, 2021.

Shlok Mishra, Anshul Shah, Ankan Bansal, Abhyuday Jagannatha, Abhishek Sharma, David Jacobs, and Dilip Krishnan. Object-aware cropping for self-supervised learning. *arXiv preprint arXiv:2112.00319*, 2021.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *ICML*, pp. 8821–8831. PMLR, 2021.

Nikunj Saunshi, Jordan T. Ash, Surbhi Goel, Dipendra Misra, Cyril Zhang, Sanjeev Arora, Sham M. Kakade, and Akshay Krishnamurthy. Understanding contrastive learning requires incorporating inductive biases. *CoRR*, abs/2202.14037, 2022. URL https://arxiv.org/abs/2202.14037.

Chenxin Tao, Xizhou Zhu, Gao Huang, Yu Qiao, Xiaogang Wang, and Jifeng Dai. Siamese image modeling for self-supervised vision representation learning. *arXiv preprint arXiv:2206.01204*, 2022.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. *arXiv preprint arXiv:2012.12877*, 2020.

Wouter Van Gansbeke, Simon Vandenhende, Stamatios Georgoulis, and Luc V Gool. Revisiting contrastive methods for unsupervised learning of visual representations. *Advances in Neural Information Processing Systems*, 34:16238–16250, 2021.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.

Xinlong Wang, Rufeng Zhang, Chunhua Shen, Tao Kong, and Lei Li. Dense contrastive learning for self-supervised visual pre-training. In *CVPR*, pp. 3024–3033, 2021.

Yixuan Wei, Han Hu, Zhenda Xie, Zheng Zhang, Yue Cao, Jianmin Bao, Dong Chen, and Baining Guo. Contrastive learning rivals masked image modeling in fine-tuning via feature distillation. *arXiv preprint arXiv:2205.14141*, 2022.

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*, pp. 8392–8401, 2021a.

Zhenda Xie, Yutong Lin, Zhuliang Yao, Zheng Zhang, Qi Dai, Yue Cao, and Han Hu. Self-supervised learning with swin transformers. *arXiv preprint arXiv:2105.04553*, 2021b.

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. *arXiv preprint arXiv:2111.09886*, 2021c.

Zhenda Xie, Zigang Geng, Jingcheng Hu, Zheng Zhang, Han Hu, and Yue Cao. Revealing the dark secrets of masked image modeling. *CoRR*, abs/2205.13543, 2022a. doi: 10.48550/arXiv.2205.13543. URL https://doi.org/10.48550/arXiv.2205.13543.

Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Yixuan Wei, Qi Dai, and Han Hu. On data scaling in masked image modeling. *CoRR*, abs/2206.04664, 2022b. doi: 10.48550/arXiv.2206.04664. URL https://doi.org/10.48550/arXiv.2206.04664.

Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. *arXiv preprint arXiv:2103.03230*, 2021.

Yuanyi Zhong, Haoran Tang, Junkun Chen, Jian Peng, and Yu-Xiong Wang. Is self-supervised learning more robust than supervised learning? *arXiv preprint arXiv:2206.05259*, 2022.

Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. *IJCV*, 127(3):302–321, 2019.

Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. Ibot: Image bert pre-training with online tokenizer. *arXiv preprint arXiv:2111.07832*, 2021.

Jie Zhu, Huabin Huang, Banghuai Li, and Leye Wang. E-crf: Embedded conditional random field for boundary-caused class weights confusion in semantic segmentation. In *The Eleventh International Conference on Learning Representations*, 2023.

A Appendix

A.1 Model Description

For all the models involved in the experiments including DeiT (Touvron et al., 2020), MoCo v3 (Chen et al., 2021), DINO (Caron et al., 2021), BEiT (Bao et al., 2021), MAE (He et al., 2021), CAE (Chen et al., 2022a), and iBOT (Zhou et al., 2021), we use their official code to implement the encoders. It is worth noticing that for DINO and iBOT, we choose the checkpoint of the teacher models as they have been reported to perform better than the student models in their papers (Caron et al., 2021; Zhou et al., 2021).

A.2 Datasets

ADE20K (Zhou et al., 2019) is one of the most challenging benchmarks, containing 150 fine-grained semantic concepts and a variety of scenes with 1,038 image-level labels. There are 20,210 images in the training set and 2,000 images in the validation set. We choose 59 out of total 150 semantic concepts that are concrete objects containing parts (Zhou et al., 2019), termed ADE20K-Object. We also select 209 part categories that emerge both in the training set and the validation set, called ADE20K-Part.

Pascal-Part (Chen et al., 2014) is a set of additional annotations for PASCAL VOC 2010 (Everingham et al., 2010), thereby holding the same statistics as those of PASCAL VOC 2010. It provides segmentation masks for each part of objects. Concretely, the dataset includes 20 object-level categories and 193 part-level categories. In our experiments, we remove 4 object categories that do not contain parts including boat, table, chair, and sofa.

LIP (Gong et al., 2017) is a large-scale benchmark for human parsing research, which includes 50,462 images with pixel-wise annotations on 19 semantic part labels. In detail, it includes 19,081 full-body images, 13,672 upper-body images, 403 lower-body images, 3,386 head-missed images, 2,778 back-view images and 21,028 images with occlusions. There are 30,462 images in the training set and 10,000 images in the validation set. The rest 10,000 images are served as the test set with missing labels for competition evaluation.

CUB-200-2011 (Wah et al., 2011) is a popular benchmark for fine-grained image classification, and also provides bounding box and part location annotations. It contains 11,788 images of 200 bird species and 15 part keypoint annotations per bird. In this work, we mainly leverage its part keypoint annotations. And only 4 part categories (right eye, right leg, left wing, and tail) are chosen to be considered in our experiments, to make sure that the selected keypoints are far enough away from each other and enough context information can be contained in the cropped patches. (We also tried using all keypoints and the conclusion is consistent.)

COCO (Caesar et al., 2018), as one of the most widely-used human pose estimation datasets, contains more than 200,000 images and 250,000 labeled person instances. Similar to CUB-200-2011 mentioned above, only 3 (nose, right wrist, and left ankle) of its 17 keypoint categories are considered in our experiments.

CIFAR10 (Krizhevsky et al., 2009) is a widely-used dataset that has $50,000$ training images and $10,000$ test images. CIFAR10 has 10 categories. The dimension for CIFAR10 images is $32 \times 32 \times 3$.

A.3 Part Resizing Approach

We explain the idea behind our part resizing approach below. There are three main reasons for using this approach:

• To maintain the shape of the position embedding the same as during pretraining, it is essential for the transformer to distinguish patches located at different positions. Rashly resizing position embedding, e.g., through interpolation, may introduce some noise and cause a mismatch with the parameters of other pretrained transformer components.

• Is there a domain gap between training data and the resized part? No may be the answer. As shown in Table 6, the crop operation during self-supervised pretraining, mostly adopts a scale range of $(0.08 - 1)$, followed by image resizing. Hence, the pretrained models may see these "upscaled" parts during training potentially. Therefore, there is no domain gap using our part resizing approach.

Table 6: The crop scale of different pretrained models.

Method	DeiT	MoCo v3	DINO	BEiT	MAE	CAE	iBOT
Crop Scale	0.08-1	0.08-1	0.4-1, 0.05-0.4	0.08-1	0.2-1	0.08-1	0.14-1, 0.05-0.4

• We also considered another manner: processing the whole image and then taking the individual patch representations. However, due to the presence of the attention module, different patch representations would interact globally. Therefore, the part representation would contain information from other parts and may lead to information leaks hence the confusion about part representations.

A.4 Detailed Results for Object-Level and Part-Level Segmentation

In this section, we provide detailed comparisons between object-level and part-level semantic segmentation in Table 7 and Table 8. Similar observations as in Figure 7 in the main paper are found: although the supervised DeiT is superior over self-supervised methods on ADE20K-Object and Pascal-Object except iBOT, it is generally inferior to self-supervised models on ADE20K-Part and Pascal-Part, demonstrating self-supervised methods can learn good part-aware representations. BEiT and MAE perform inferior, perhaps because the two methods do not have an explicit process to predict the encoded representations of masked patches, instead, directly reconstruct the targets.

Table 7: Linear semantic segmentation results on ADE20K-Object and ADE20K-Part.

Methods	Object Seg on ADE20K-Object 59 Object Classes			Part Seg on ADE20K-Part 209 Part Classes		
	mIoU	mAcc	aAcc	mIoU	mAcc	aAcc
Supervised Model:						
DeiT	52.6	62.9	83.8	27.3	34.7	69.2
Contrastive Learning:						
MoCo v3	50.2	60.4	83.6	27.1	34.7	70.1
DINO	50.8	60.8	83.9	28.9	36.8	70.3
Masked Image Modeling (MIM):						
BEiT	28.6	37.2	73.4	18.6	25.8	58.2
MAE	41.0	50.6	79.9	26.3	35.0	67.3
CAE	47.4	58.4	82.9	28.4	36.9	71.1
Contrastive Learning + MIM:						
iBOT	55.2	65.1	85.6	32.2	40.0	73.4

Table 8: Linear semantic segmentation results on Pascal-Object and Pascal-Part.

Methods	Object Seg on Pascal-Object 16 Object Classes			Part Seg on Pascal-Part 193 Part Classes		
	mIoU	mAcc	aAcc	mIoU	mAcc	aAcc
Supervised Model:						
DeiT	92.2	95.3	96.8	27.4	36.2	65.8
Contrastive Learning:						
MoCo v3	89.4	93.7	95.7	27.1	35.8	66.0
DINO	88.0	92.7	95.3	27.8	36.5	66.4
Masked Image Modeling (MIM):						
BEiT	56.4	69.0	76.8	14.8	21.4	47.0
MAE	76.1	84.6	89.5	24.3	32.9	61.5
CAE	83.3	89.7	93.2	27.8	37.0	66.3
Contrastive Learning + MIM:						
iBOT	92.1	95.3	97.1	30.7	40.0	69.7

Table 9: Different learning rates in LIP part segmentation.

Method	$1e-5$	$2e-4$	$3e-4$	$4e-4$	$5e-4$	$1e-3$
Supervised Model:						
DeiT	40.5	41.3	41.4	**41.4**	41.3	41.3
Contrastive Learning:						
MoCo v3	40.0	41.7	41.7	**41.9**	41.7	41.8
DINO	38.1	40.8	40.9	**41.0**	40.9	40.9
Masked Image Modeling (MIM):						
BEiT	23.1	27.2	27.4	27.2	27.1	**27.5**
MAE	34.7	38.5	38.6	38.2	38.7	**38.8**
CAE	42.8	44.1	44.2	43.7	**44.3**	44.3
Contrastive Learning + MIM:						
iBOT	41.5	44.4	44.5	44.6	44.5	**44.8**

Table 10: Linear human pose estimation results on the COCO dataset.

Method	AP	AP50	AR	AR50
Supervised Model:				
DeiT	6.8	27.5	14.7	44.3
Contrastive Learning:				
MoCo v3	12.6	42.1	21.61	56.3
DINO	16.6	49.0	25.9	62.1
Masked Image Modeling (MIM):				
MAE	**23.6**	**59.8**	**33.3**	**69.6**
CAE	17.5	50.1	28.9	64.0
Contrastive Learning + MIM:				
iBOT	15.5	48.1	26.0	62.8

A.5 Ablation on Learning Rate for Part Semantic Segmentation

We use batch size 16 and 160k iterations following previous methods He et al. (2021); Bao et al. (2021); Zhou et al. (2021). For optimizer, we adopt AdamW which is widely used for Vision Transformer (Dosovitskiy et al., 2021). As for learning rate, we conduct experiments on LIP part segmentation with different learning rates including $1e-5$, $2e-4$, $3e-4$, $4e-4$, $5e-4$, and $1e-3$ for all pretrained models. The results are presented in Table 9. The best results are bold. One can see that except for $1e-5$, other learning rates slightly influence the final performance. In terms of the overall performance of each model, MoCo v3, CAE, and iBOT show superior performance over supervised model DeiT, validating our findings in the main paper.

A.6 Why BEiT and MAE Perform Inferior?

In linear evaluation, the quality of output representations always depends on the pretraining quality of the encoder. A well-trained encoder usually produces representations of high quality. In our experiments, we find that BEiT only obtains 56.7 linear probe on ImageNet while other models achieve about 70 linear probe on ImageNet. **The large gap indicates that BEiT does not learn high-quality representations as other models, i.e., its pretraining quality is not good enough.** This could be the main reason that it performs worst on downstream tasks (both object-level and part-level tasks).

For MAE, we find that the colors of the retrieved patches in Figure 6 and Figure 8 in Appendix are quite similar. **Based on it, we speculate that its low-level color reconstruction target helps MAE learn better middle- or low-level features while providing limited high-level semantics.** To validate this point, we further conduct experiments on the human pose estimation task, which has been proven by previous work Ding et al. (2021) to be a task requiring middle-level features. Similar to linear probing, we train one linear layer over the features extracted by the frozen backbones to predict human keypoints. The results are shown in Table 10 and MAE outperforms all other counterparts by large margins. So we believe MAE tends to learn lower-level features and pay relatively less attention to the high-level semantics, which leads to its inferior performance on classification and segmentation tasks.

A.7 Takeaways

We have outlined several key takeaways derived from our experiments and analysis, which we believe could benefit further works in the self-supervised learning field.

- **1)** Supervised models care more about the whole objects, while self-supervised models have a stronger capability of learning part-aware representations. understanding self-supervised representation pretraining.

- **2)** Contrastive learning may learn higher-level semantics and be more semantically abundant than MIM. Combining MIM and CL is potentially capable of learning multi-level semantics.
- **3)** MIM methods, e.g., CAE, learn slightly better part-level features than contrastive learning methods. While the learned features of MIM methods, to some extent, depend on the reconstruction targets, e.g., the RGB (color) target used in MAE leads it to learn lower-level features, hence only sub-optimal results on high-level tasks.

Those takeaways are evidenced correspondingly as follows:

- **1)** In Figure 7, Table 7, and Table 8, from object-level to part-level, considerable performance improvements (relative to DeiT) are observed for all self-supervised models. Among these methods, DINO, CAE, and iBOT show larger improvements, demonstrating they can learn better features for part-aware segmentation. For example, DeiT (52.6% in mIoU) outperforms DINO (50.8%) and CAE (47.4%) by 1.8% and 5.2% in ADE20K object-level experiment. However, in ADE20K part-level experiment, the situation is completely opposite. DINO (28.9% in mIoU) outperforms DeiT (27.3%) by 1.6% and CAE (28.4%) outperforms DeiT by 1.1%. Similar results can be seen in segmentation experiments on the Pascal dataset, as well as classification results on COCO and CUB-200.
- **2)** As shown in the second row of Figure 2, the retrieved patches of MoCo v3 are all about dog mouths. While the retrieved patches of CAE contain some patches about the mouths of cats or foxes. Similarly, in Figure 8, the retrieved patches of CAE for the watch also include patches about the dial of the dial phone. Contrastive learning methods are more likely to retrieve patches with the same category while CAE sometimes retrieves some patches with similar textures. This observation will be more frequently found from the top-100 retrieved patches, based on which we show contrastive learning methods tend to learn better high-level semantics than MIM. And as shown in our paper, iBOT outperforms other methods by large margins in almost all tasks, including object-level and part-level tasks.
- **3)** ❶ On part retrieval, CAE outperforms contrastive learning methods by a large margin (about 8% and 5% on CUB-200 and COCO, respectively). On object-level linear probing or segmentation tasks, CAE performs clearly worse than contrastive learning methods. While this performance gap is significantly narrowed even closed on corresponding part-level tasks. ❷ the retrieved patches of MAE in Figure 6 and Figure 8 tend to share similar hues, and MAE performs much better than other methods on the human pose estimation task which requires middle-level features.

A.8 More Discussion about Related Work

Image crop is important in self-supervised learning and brings promising properties to self-supervised models: Van Gansbeke et al. (2021) shows that MoCo learns spatially structured representations when trained with a multi-crop strategy; By leveraging object-aware cropping, Mishra et al. (2021) encourages the self-supervised model to learn both object and scene level semantic representations from real-world uncurated datasets. Similarly, in our part-to-whole explanation, random crop on ImageNet generates views that describe parts of an object and are encoded to part representations and projected to whole for agreement. DenseCL (Wang et al., 2021) designs a dense self-supervised learning method that directly works at the level of pixels (or local features) by taking into account the correspondence between local features. DetCo (Xie et al., 2021a) achieves better performance trade-off on both classification and detection through multi-level supervision to intermediate representations and contrastive learning between global image and local patches.

A.9 More Examples for ImageNet Patch Retrieval Visualization

We visualize more patch retrieval results of the encoded representations on the ImageNet validation set in Figures 8.

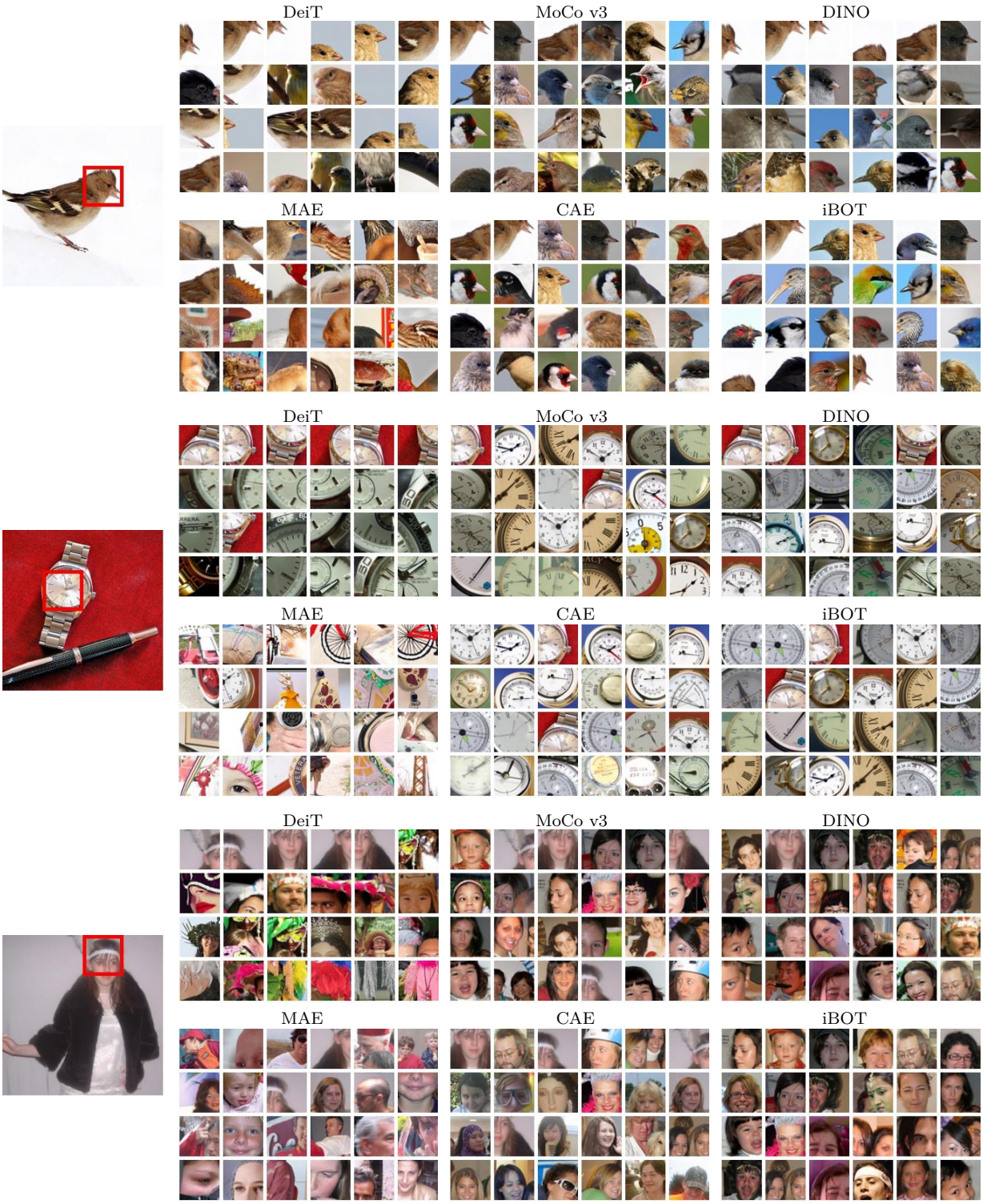

Figure 8: Patch retrieval comparisons of encoded representations on cropped patches from ImageNet.