Siamese Image Modeling for Self-Supervised Vision Representation Learning

Chenxin Tao[1*†], Xizhou Zhu[2,4*], Weijie Su[3*†], Gao Huang[1], Bin Li[3],
Jie Zhou[1], Yu Qiao[4], Xiaogang Wang[5], Jifeng Dai[1,4✉]
[1]Tsinghua University, [2]SenseTime Research, [3]University of Science and Technology of China,
[4]Shanghai Artificial Intelligence Laboratory, [5]The Chinese University of Hong Kong,
tcx20@mails.tsinghua.edu.cn, zhuwalter@sensetime.com,
jackroos@mail.ustc.edu.cn, {gaohuang,jzhou,daijifeng}@tsinghua.edu.cn,
binli@ustc.edu.cn, qiaoyu@pjlab.org.cn, xgwang@ee.cuhk.edu.hk

Abstract

Self-supervised learning (SSL) has delivered superior performance on a variety of downstream vision tasks. Two main-stream SSL frameworks have been proposed, i.e., Instance Discrimination (ID) and Masked Image Modeling (MIM). ID pulls together representations from different views of the same image, while avoiding feature collapse. It lacks spatial sensitivity, which requires modeling the local structure within each image. On the other hand, MIM reconstructs the original content given a masked image. It instead does not have good semantic alignment, which requires projecting semantically similar views into nearby representations. To address this dilemma, we observe that (1) semantic alignment can be achieved by matching different image views with strong augmentations; (2) spatial sensitivity can benefit from predicting dense representations with masked images. Driven by these analysis, we propose Siamese Image Modeling (SiameseIM), which predicts the dense representations of an augmented view, based on another masked view from the same image but with different augmentations. SiameseIM uses a Siamese network with two branches. The online branch encodes the first view, and predicts the second view's representation according to the relative positions between these two views. The target branch produces the target by encoding the second view. SiameseIM can surpass both ID and MIM on a wide range of downstream tasks, including ImageNet finetuning and linear probing, COCO and LVIS detection, and ADE20k semantic segmentation. The improvement is more significant in few-shot, long-tail and robustness-concerned scenarios. Code shall be released.

1. Introduction

Self-supervised learning (SSL) has been pursued in the vision domain for a long time [32]. It enables us to pre-train models without human-annotated labels, which makes it possible to exploit huge amounts of unlabeled data. SSL has provided competitive results against supervised pre-training baselines in various downstream tasks, including image classification [4, 23], object detection [35] and semantic segmentation [24].

To effectively train models in the SSL manner, researchers design the so-called "pretext tasks" to generate supervision signals. One of the most typical frameworks is *Instance Discrimination (ID)*, whose core idea is to pull together representations of different augmented views from the same image, and avoid representational collapse. Different variants of ID have been proposed, including contrastive learning [9, 24], asymmetric networks [12, 21], and feature decorrelation [5, 57]. A recent work [43] has shown the intrinsic consistency among these methods via their similar gradient structures. For ID methods, the representations of each image are well separated, thus inducing good linear separability. However, as shown in [35], for transfer learning on detection tasks with Vision Transformers [19], ID is not superior to supervised pre-training, and even lags behind random initialization given enough training time.

Recently, another SSL framework has gradually attracted more attention, namely *Masked Image Modeling (MIM)* [4, 23]. MIM methods train the model to reconstruct the original content from a masked image. Such practice can help to learn the rich local structures within an image, leading to excellent performance in dense prediction tasks such as object detection [35]. Nevertheless, MIM does not have good linear separability as ID, and usually performs poorly under the few-shot classification settings [1].

Both ID and MIM methods have their own strengths and

*Equal contribution. †This work is done when Chenxin Tao and Weijie Su are interns at Shanghai Artificial Intelligence Laboratory. ✉Corresponding author.

	ImageNet			COCO		ADE20k	LVIS		Robustness
	FT	LIN	$FT_{1\%}$	AP^b	AP^m	mIoU	AP^b_{rare}	AP^m_{rare}	avg score*
MoCo-v3 (ID method)	83.0	76.7	63.4	47.9	42.7	47.3	25.5	25.8	43.4
MAE (MIM method)	83.6	68.0	51.1	51.6	45.9	48.1	29.3	29.1	41.8
SiameseIM (ours)	84.1	78.0	65.1	52.1	46.2	51.1	30.9	30.1	47.9
Improve w.r.t. MoCo-v3	+1.1	+1.3	+1.7	+4.2	+3.5	+3.8	+5.4	+4.3	+4.5
Improve w.r.t. MAE	+0.5	+10.0	+14.0	+0.5	+0.3	+3.0	+1.6	+1.0	+6.1

Table 1. SiameseIM surpasses MoCo-v3 (ID method) and MAE (MIM method) on a wide range of downstream tasks. *The robustness average score is calculated by averaging top-1 acc of IN-A, IN-R, IN-S, and 1-mCE of IN-C. For detailed results, please refer to Section 4.2.

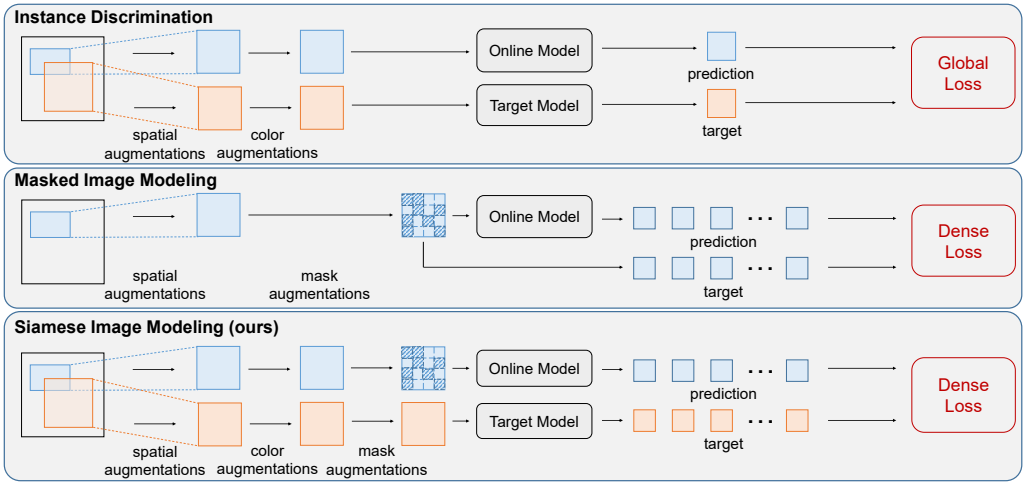

Figure 1. Comparisons among ID, MIM and SiameseIM. Matching different augmented views can help to learn semantic alignment, which is adopted by ID and SiameseIM. Predicting dense representations from masked images is beneficial to obtain spatial sensitivity, which is adopted by MIM and SiameseIM.

weaknesses. We argue that this dilemma is caused by neglecting the representation requirements of either semantic alignment or spatial sensitivity. Specifically, MIM operates within each image independently, regardless of the inter-image relationship. The representations of semantically similar images are not well aligned, which further results in poor linear probing and few-shot learning performances of MIM. On the other hand, ID only uses a global representation for the whole image, and thus fails to model the intra-image structure. The spatial sensitivity of features is therefore missing, and ID methods usually produce inferior results on dense prediction.

To overcome this dilemma, we observe the key factors for semantic alignment and spatial sensitivity: (1) semantic alignment requires that images with similar semantics are projected into nearby representations. This can be achieved by matching different augmented views from the same image. Strong augmentations are also beneficial because they provide more invariance to the model; (2) spatial sensitivity needs modeling the local structures within an image. Predicting dense representations from masked images thus helps, because it models the conditional distribution of image content within each image. These observations motivate us to predict the dense representations of an image from a masked view with different augmentations.

To this end, we propose Siamese Image Modeling (SiameseIM), which reconstructs the dense representations of an augmented view, based on another masked view from the same image but with different augmentations (see Fig. 1). It adopts a Siamese network with an online and a target branch. The online branch consists of an encoder that maps the first masked view into latent representations, and a decoder that reconstructs the representations of the second view according to the relative positions between these two views. The target branch only contains a momentum encoder that encodes the second view into the prediction target. The encoder is made up of a backbone and a projector. After the pre-training, we only use the online backbone for downstream tasks.

As shown in Tab. 1, SiameseIM is able to surpass both MIM and ID methods over a wide range of evaluation tasks, including full-data fine-tuning, few-shot learning and linear probing on ImageNet [16], object detection on COCO [36] and LVIS [22], semantic segmentation on ADE20k [58], as well as several robustness benchmarks [26–28, 47]. By gathering semantic alignment and spatial sensitivity in one model, SiameseIM can deliver superior results for all tasks. We also note that such improvements are more obvious on

ADE20k(∼3 points) and LVIS(∼1.6 point for rare classes) datasets. These long-tailed datasets demands semantic alignment and spatial sensitivity at the same time, and SiameseIM thus can deliver superior performance on them.

Our contributions can be summarized as follows:

- As a new form of SSL, SiameseIM is proposed to explore the possibilities of self-supervised pre-training. It displays for the first time that that using only a single dense loss is enough to learn semantic alignment and spatial sensitivity well at the same time;
- Compared with MIM methods, SiameseIM shows that reconstructing another view helps to obtain good semantic alignment. This also suggests that MIM framework can be used to reconstruct other targets with proper guidance, which opens a possible direction for MIM pretraining;
- Compared with ID methods, SiameseIM shows that dense supervision can be applied by matching the dense correspondence between two views strictly through their relative positions. We demonstrate dense supervision can bring a considerable improvement of spatial sensitivity;
- SiameseIM is able to surpass both MIM and ID methods over a wide range of tasks. SiameseIM obtains more improvements in few-shot, long-tail and robustness-concerned scenarios.

2. Related work

Instance Discrimination (ID). The core idea of instance discrimination is to pull together different augmented views of the same image and avoid representational collapse [45, 53]. In this way, models can learn to separate the representation of each image, leading to decent linear separability. There are three typical types of instance discrimination methods, while Siamese networks are always employed. *Contrastive Learning* methods [9, 11, 13, 24] push apart views from different images (negative samples) to avoid representational collapse. *Asymmetric Network* methods [12, 21] explore to get rid of negative samples with the help of an asymmetric network design. In these methods, a predictor network is only appended after one branch of the siamese network, and the other branch is detached from the gradient back-propagation. *Feature Decorrelation* methods [5, 29, 57] try to accomplish instance discrimination by reducing the redundancy among different feature dimensions. These different methods are then unified in UniGrad [43] by revealing that they share similar gradient structures. A recent work [44] has even successfully surpassed the performance of supervised learning with ResNets [25]. There have also been works on applying instance discrimination to Vision Transformers [7, 13], which demonstrate impressive performances.

The most common evaluation metric used for ID is linear probing, which trains a linear classifier on top of frozen representations. This metric concentrates on the linear separability of learned features. However, as [23] has pointed out, the dense prediction performance of ID on Vision Transformers is not superior to supervised pre-training, especially on object detection tasks.

Some previous ID works [33,37,49,52,55,56] have tried to introduce dense supervision to enhance local features. They employ different techniques to build dense correspondence between two views, including using Earth Mover's Distance [37], matching feature similarity [33, 49], finding nearest neighbors [55], using extra region proposal or flow modules [56]. However, these works either only focus on detection result, or rely on an extra global loss to improve linear probing result, which still requires to find a trade-off between semantic alignment and spatial sensitivity.

Our work display for the first time that only using a dense loss is enough to learn these two properties well at the same time. Unlike previous methods, we utilizes the relative positions between two views to strictly align the spatial correspondence. In doing so, SiameseIM outperforms the original ID methods by a large margin on a strong detection baseline.

Masked Image Modeling (MIM). Masked image modeling intends to reconstruct image content from a masked image, which is motivated by the masked language modeling in NLP [6, 17, 40, 41]. iGPT [8] first tries to reconstruct image pixels. ViT [19] has also tried to predict the mean color of the masked patch. However, these preliminary attempts are not competitive with their supervised counterparts. BEiT [4] reveals the power of MIM by predicting visual tokens from a pre-trained discrete VAE [42]. MAE [23] successfully performs pre-training via predicting raw pixels. It shows that the key point is to use a high masked ratio due to the high spatial redundancy. After that, different works [3,10,18,20,51] continue to push the limit by improving the quality of prediction targets. Some works [3,51] have shown that it is more effective to predict features rather than raw pixels for learning representations.

Unlike ID methods, MIM methods excel in transfer learning with full model fine-tuning with Vision Transformers, but lack good linearly-separated representations [23]. For example, given enough training epochs, BEiT [4] and MAE [23] can surpass other pre-training paradigms on detection tasks. However, under few-shot scenes, MIM methods are not as data-efficient as ID methods because of their poor linear separability [1].

Our work demonstrates that MIM can also produce the same adequate linear separable representations as ID. This is achieved by predicting the representation of another augmented view from the same image, rather than reconstructing the original view. Through reconstructing another view, SiameseIM can even surpass the linear probing performance of ID methods.

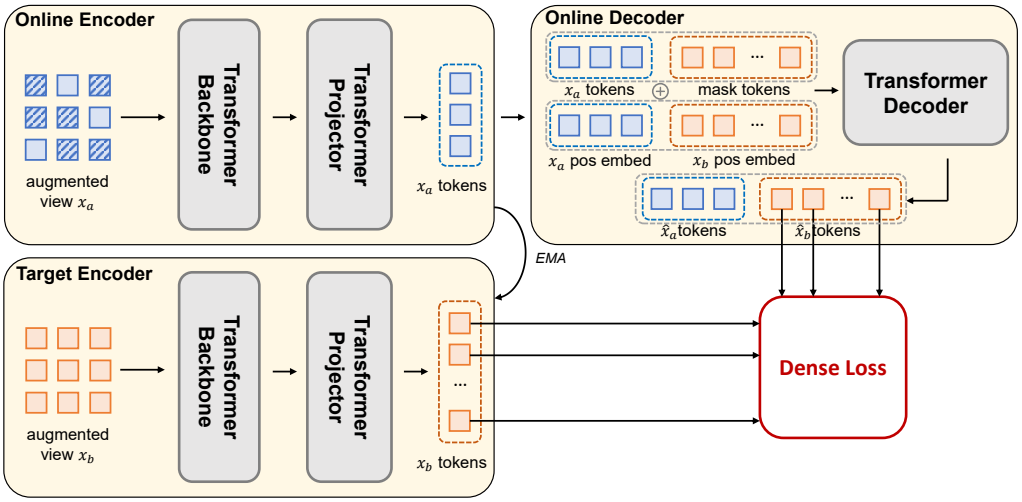

Figure 2. The overview of our Siamese Image Modeling (SiameseIM). Different augmented views are fed into the online and target branches. The online encoder operates on the visible patches of x_a. The online decoder accepts the x_a tokens as well as mask tokens that correspond to predicted x_b tokens. We use relative positions to inform the locations between x_a and x_b. The target encoder maps x_b to the target representations. We finally apply the dense loss on the dense representations.

Concurrent Work. Some recent works have also tried to combine ID and MIM methods [30, 39, 48, 59]. It was generally believed that ID and MIM provide two completely different pre-training supervisions. To benefit from both worlds, they simply combine these two supervisions. The most relevant to ours is iBoT [59], which combines two existing methods from ID and MIM, namely DINO [7] and BEiT [4]. While we find that as long as the local patches of different views are learned to be matched, the two tasks can be naturally unified. SiameseIM first enabled reconstructing dense features across different views. A single dense loss is designed to learn both semantic alignment and spatial sensitivity. This makes SiameseIM the first unified framework that naturally combines the best of MIM and ID methods.

3. Method

We depict our model of SiameseIM in Fig. 2. It takes two augmented views x_a and x_b of the same image as inputs. SiameseIM aims to predict the dense representations of x_b based on that of x_a. A Siamese network with an online and a target branch is used. The online branch is made up of an encoder that encodes the visible patches of x_a into a latent representation, and a decoder that predicts the representation of x_b according to the relative positions between x_a and x_b. The target branch only has a momentum encoder which takes x_b as input. The encoder consists of a backbone and a projector. After the pre-training, only the online backbone is used for downstream evaluation.

3.1. Augmented Inputs

ID methods adopt two different augmented views, while MIM methods utilize a single view. In our method, similar to ID methods, we feed two different views as the inputs to the online and target branches, respectively. As will be shown in Section 4.3, different views can significantly increase the linear probing result without harming the performance of object detection.

Apart from the number of views, there are also differences in the augmentations of previous methods (see Fig. 1). ID methods [9, 13, 24] tends to add stronger augmentations, which typically contain spatial and color augmentations. Whereas recently, MAE [23] reports that color augmentations are not beneficial for MIM pre-training. We find that color augmentations have different effects under different training settings. They can provide more invariance when used with different views, but such effect will vanish if paired with the same view (more analysis can be found in Section 4.3). As a result, we reserve both the spatial and color augmentations from ID methods [13].

Another difference is that MIM masks out some patches of the input image for reconstruction, which we refer as mask augmentation. With mask augmentation, the task of dense prediction can model the conditional distribution of image content within each image. The representations are trained to capture the local structure, and thus are endowed with spatial sensitivity. Therefore, we also apply mask augmentation to the view of the online branch.

3.2. Prediction Targets

There can be multiple choices for the prediction targets. For example, ID methods select to predict the features of different augmented views, while MIM methods are designed to predict pixels or features of the same view. We empirically find that feature prediction is superior for dif-

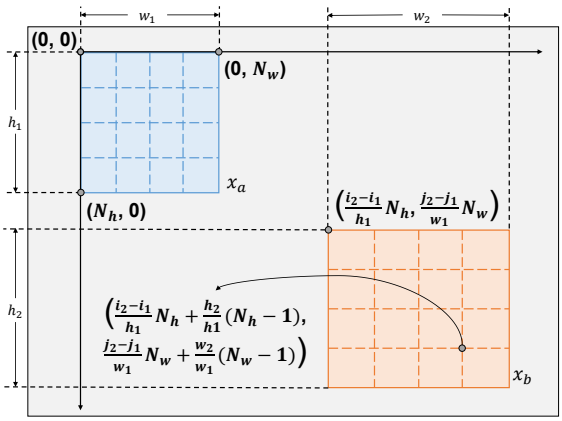

Figure 3. Positional embedding for online decoder. The positions are calculated with respect to the left-top origin of x_a.

ferent views (see Section 4.3). SiameseIM is thus designed to predict the features of another different augmented view from the same image. We shall describe how the prediction and target are calculated.

Online Branch will make the prediction. The online encoder first maps the masked view x_a into a latent representation $y_a \in \mathbb{R}^{N_v \times D}$, where N_v denotes the number of visible patches and D is the feature dimension. Following the practice in ID methods [9, 50], we append a projector after the backbone to form the encoder. The online decoder $g(\cdot)$ then combines y_a, mask tokens m, and their relative positions to calculate the prediction $y_b \in \mathbb{R}^{N \times D}$ as

$$y_b = g\Big(\text{Concat}\left(y_a + p_a,\ \{m + p_b^{(u,v)}\}_{u=1,v=1}^{N_h,N_w}\right)\Big). \tag{1}$$

Here, m indicates the learnable embedding of the mask token, which follows [23]. p_a is the position embeddings for x_a, and $p_b^{(u,v)}$ denotes the positional embedding for the patch of x_b at location (u, v), which will be introduced later. N_h and N_w denote the number of tokens along height and width dimensions in x_b (*e.g.*, $N_h = N_w = 14$), respectively. $N = N_h \times N_w$ is therefore the number of all tokens in x_b. Note that different from MIM methods, the mask tokens correspond to image patches from the different target view x_b instead of the input view x_a.

Target Branch is responsible for producing the target. The target encoder is an exponential moving average of the online encoder. It takes all tokens from x_b as input and outputs their latent representation $z_b \in \mathbb{R}^{N \times D}$. Note that we can also directly predict the raw pixels, where the target encoder is unnecessary. However, we find that feature prediction can give better performance when different views are used (see Section 4.3).

Positional Embedding for Online Decoder is necessarily required to inform the decoder of the corresponding locations of each patch in x_a and x_b. The decoder predicts the dense representations of x_b based on visible patches from x_a and their corresponding locations. For all input patches to the online decoder, including visible patches from x_a and mask tokens indicating x_b, their positional embeddings are calculated from the relative position with respect to the left-top origin of x_a. Fig. 3 shows the detailed process. Suppose the (left, top, height, width) positional properties of the two cropped views x_a and x_b in the original image are (i_1, j_1, h_1, w_1) and (i_2, j_2, h_2, w_2), respectively. The positions for x_a and mask tokens indicating x_b are

$$\begin{aligned} \tilde{p}_a^{(u,v)} &= (u-1, v-1), \\ \tilde{p}_b^{(u,v)} &= \left(\tfrac{h_2}{h_1}(u-1) + \tfrac{i_2-i_1}{h_1}N_h,\ \tfrac{w_2}{w_1}(v-1) + \tfrac{j_2-j_1}{w_1}N_w\right), \end{aligned} \tag{2}$$

where u and v are the location indexes along height and width dimensions. We further apply $\sin(\cdot)$ and $\cos(\cdot)$ operators to get the 2-D sin-cos positional embeddings, following the practice in MAE [23]. If the same view is used for input and target, $\tilde{p}_b^{(u,v)}$ will degenerate to $\tilde{p}_a^{(u,v)}$ as used in MAE. Instead, we choose to use different views because they are crucial for improving the linear separability of the representation (see Section 4.3). Moreover, to inform the scale variation between two different views, we add the relative scale changes as $s = \left(10 \log \frac{h_2}{h_1},\ \ 10 \log \frac{w_2}{w_1}\right)$, where $10\log(\cdot)$ is applied to make the numerical range similar to relative positions. Then, the final positional embeddings are calculated as

$$\begin{aligned} p_a^{(u,v)} &= \text{PE}(\tilde{p}_a^{(u,v)}), \\ p_b^{(u,v)} &= \text{Linear}\Big(\text{Concat}\left(\text{PE}(\tilde{p}_b^{(u,v)}),\ \text{PE}(s)\right)\Big), \end{aligned} \tag{3}$$

where PE is the sine-cosine positional encoding proposed by [46]. For x_b, we concatenate the relative positions and scale changes together, and use a linear layer to fit the dimension.

3.3. Loss Function

Loss functions guide the training direction, and thus shape the characteristic of the learned representations. ID methods usually adopt a loss over the globally averaged feature to separate representations among images, while MIM employs a dense loss on image patches to learn representations within each individual image. Interestingly, we find that a dense loss only is enough to train both semantic alignment and spatial sensitivity well. As a result, we only employ a dense loss for training SiameseIM.

Once the prediction y_b and target z_b have been calculated, we adopt a dense loss function for each predicted token. UniGrad [43] is employed because it is a unified loss of ID methods and is also memory-friendly. To apply UniGrad on the dense level, we treat each token representation

as an independent sample, *i.e.,* $y_b^i, i = 1, \ldots, N$. The corresponding positive sample is therefore z_b^i, and the negative samples consist of all the tokens from the target branch. The dense loss can then be computed according to[1]

$$L = \mathbb{E}_{\{y_b^i, z_b^i\}} \Big[- ||y_b^i - z_b^i||^2 + \lambda \sum_{u \in \mathcal{N}} (u^T y_b^i)^2 \Big], \quad (4)$$

where y_b^i comes from the online prediction, its target z_b^i is the positive sample, and all token representations from the target branch constitute the negative sample set $\mathcal{N}$. Note that most ID methods use the InfoNCE loss [45], which will require $\mathcal{O}(|\mathcal{N}|)$ memory to calculate the similarities. This is infeasible for the dense loss because of the vast number of negative sample patches. In contrast, UniGrad [43] only consumes $\mathcal{O}(D^2)$ memory by first calculating the covariance matrix of negative samples.

3.4. Discussion

As illustrated in Fig. 1, we would like to further emphasize the differences between SiameseIM and previous works from three perspectives:

(1) Compared with MIM methods, SiameseIM reveals that it's possible to reconstruct another augmented view rather than the same view. Such reconstruction can greatly enhance the semantic alignment of the model. We also show that strong augmentations can benefit this learning process;

(2) Compared with ID methods, SiameseIM shows that dense supervision can greatly improve spatial sensitivity, and the model can also learn semantic alignment well with a dense loss. By employing the relative positions between two views, SiameseIM is able to achieve strict spatial alignment and build dense correspondence without anbiguity. We also demonstrate a considerable boost on a strong detection baseline;

(3) For combining the best of MIM and ID methods, recent works have also made some attempts [30,39,48,59]. Nevertheless, their efforts do not jump out of the default setting of both frameworks, *i.e.,* they use different views only for ID, and apply MIM only on each view independently. These methods rely on both the global and dense loss as training objectives. In comparison, our work can naturally take the best of ID and MIM. SiameseIM enforces the similarity between different views from the dense level. Benefiting from our modeling, we can reveal the most important factors that influence the linear probing and object detection performances by gradually modifying ID or MIM methods to SiameseIM (see Section 4.3).

[1]Here we remove the L2 normalization in the original UniGrad formulation, as we empirically find that not applying normalization to the online prediction can improve the performance (see Section 4.3).

4. Experiments

4.1. Implementation Details

ViT-B/16 [19] is used as the backbone. Transformer encoder blocks [46] with BatchNorm [31] are adopted as the projector and decoder. Before calculating loss, if not specified, we follow MAE [23] to apply LayerNorm without affine parameters to target, and no normalization to prediction. We set $\lambda = 0.02$ in the loss function. During pre-training, we adopt the standard augmentation used in MoCo-v3 [13]. For masking strategy, if not specified, we follow BEiT [4] to use blockwise masking. We evaluate our model in various downstream tasks, including full data fine-tuning, few-shot learning and linear probing on ImageNet [16], object detection on COCO [36] and LVIS [22], semantic segmentation on ADE20k [58], as well as several robustness benchmarks [26–28,47]. Please refer to Appendix A for more implementation details.

4.2. Main Results

Image Classification. Tab. 2a shows the results of image classification tasks on ImageNet [16]. For full data fine-tuning, SiameseIM surpasses both MIM and ID methods. For linear probing, our work outperforms MAE [23] by 10 points, and MoCo-v3 [13] by 1.3 poins. When only 1% data is available, SiameseIM can outperform MoCo-v3 by 1.7 points and MAE by 14.0 points. This validates that SiameseIM has obtained good semantic alignment. Moreover, SiameseIM can already deliver comparable results with previous works with only 400 epochs' pretraining.

Common Object Detection. Tab. 2b reports the performance on COCO [36] detection. Compared to MoCo-v3, our method obtains 4.2 points improvement. SiameseIM is also better than MIM methods [4,23]. This comparison validates that our method gets good spatial sensitivity.

Semantic Segmentation. Tab. 2b also demonstrates segmentation results on ADE20k [58]. SiameseIM can surpass all pure ID and MIM methods over 3.0 points. Moreover, our method obtains 49.6 mIoU with only 400 epochs' pretraining, already on par with previous methods. Different from data-balanced ImageNet and COCO datasets, ADE20k contains classes that do not have enough labels. It thus demands both semantic alignment and spatial sensitivity of high quality. This shows the superiority of our method.

Long-tail Object Detection. Tab. 2c compares the results on LVIS [22] object detection. SiameseIM performs on par with MAE [23] on overall AP metric, and delivers 1.6 point gain on rare classes. Different from common object detection, long-tail object detection poses higher demand for semantic alignment because of rare classes. SiameseIM therefore displays larger improvement.

Robustness Evaluation. Tab. 2d shows the robustness eval-

Method	Epochs	ImageNet FT	ImageNet LIN	ImageNet $FT_{1\%}$
Supervised	300	81.8	-	-
DINO*	800†	82.8	78.2	-
iBOT*	1600†	84.0	79.5	-
DenseCL‡	400	82.2	69.7	49.9
MoCo-v3	600†	83.0	76.7	63.4
BEiT	800	83.2	-	-
MAE	400	83.1	62.5	-
MAE	1600	83.6	68.0	51.1
SiameseIM	400	83.7	76.8	61.8
SiameseIM	1600	**84.1** (+0.5)	**78.0** (+1.3)	**65.1** (+1.7)

(a) Image classification.

Method	Epochs	COCO AP^b	COCO AP^m	ADE20k mIoU
Supervised	300	47.9	42.9	47.4
DINO*	800†	50.1	43.4	46.8
iBOT*	1600†	51.2	44.2	50.0
DenseCL‡	400	46.6	41.6	44.5
MoCo-v3	600†	47.9	42.7	47.3
BEiT	800	49.8	44.4	47.1
MAE	400	50.6	45.1	45.0
MAE	1600	51.6	45.9	48.1
SiameseIM	400	50.7	44.9	49.6
SiameseIM	1600	**52.1** (+0.5)	**46.2** (+0.3)	**51.1** (+3.0)

(b) Common object detection and semantic segmentation.

Method	Epochs	AP^b	AP^b_{rare}	AP^m	AP^m_{rare}
Supervised	300	37.2	-	34.9	26.4
iBoT*	1600	36.9	29.1	34.6	28.9
DenseCL‡	400	33.8	25.1	32.1	24.6
MoCo-v3	600†	37.3	25.5	35.3	25.8
MAE	400	38.4	25.4	36.6	25.7
MAE	1600	40.1	29.3	**38.1**	29.1
SiameseIM	400	38.5	28.9	36.1	27.7
SiameseIM	1600	**40.5** (+0.4)	**30.9** (+1.6)	**38.1** (+0.0)	**30.1** (+1.0)

(c) Long-tail object detection on LVIS.

Method	Epochs	IN-A top-1	IN-R top-1	IN-Sketch top-1	IN-C 1-mCE
MSN*	1200†	37.5	50.0	36.3	53.4
iBoT*	1600†	42.4	50.9	36.9	55.5
DenseCL‡	400	30.8	43.8	29.9	48.1
MoCo-v3	600†	32.4	49.8	35.9	55.4
MAE	1600	35.9	48.3	34.5	48.3
SiameseIM	400	38.6	51.6	37.7	55.9
SiameseIM	1600	**43.8** (+7.9)	**52.5** (+2.7)	**38.3** (+2.4)	**57.1** (+1.7)

(d) Robustness evaluation.

Table 2. Results on downstream tasks. The numbers in gray cell are the main baselines. The bold numbers are the best results. "FT" denotes full data finetuning. "LIN" denotes linear probing. "$FT_{1\%}$" denotes 1% data finetuning. *These methods use multi-crop during pretraining. †These methods use a symmetric loss, so the effective number of epoch should be doubled. ‡We also apply DenseCL [49] to ViT-B and conduct the pre-train following MoCo-v3 [13].

uation on four datasets [26–28, 47]. Compared with MoCo-v3 [12], our method can bring an average of 4.5 gain. Compared with MAE [23], SiameseIM can leads by a large margin of an average of 6.1 points. The results suggest that SiameseIM helps to improve the robustness of representations both over ID and MIM methods.

4.3. Ablation Study

We carefully ablate the components of SiameseIM in this section to identify the most important factors for semantic alignment and spatial sensitivity. We focus on the linear probing and COCO detection results. We state the key observations as follows.

Predicting Pixels or Features. Tab. 3(ab) and (de) ablate what type of target to use. When the same view is used for input and target, we find that predicting raw pixels performs better than predicting features. On the contrary, it's superior to predict features if different views are used. We suspect that, for different views, predicting pixels presents a much more difficult pretext task than using the same view, whereas predicting features simplifies this reconstruction because the network can help to filter irrelevant details and extract semantic information.

Different Views. We demonstrate the effectiveness of different views by comparing Tab. 3(af). Here, we choose the best-performed setting for both the same view or different views. It's shown that different views significantly improve the linear probing performance by ∼11 points. This justifies our claim that matching different augmented views is the key to obtain semantic alignment.

Color Augmentations. Tab. 3(ac) and (ef) reports the effects of color augmentations. We observe different effects with the same view or different views. Color augmentations can help linear probing to obtain 3.5 points gain for different views, but this improvement vanishes with the same view. This coincides with the phenomena in SimCLR [9] and MAE [23]. We presume that if the same view is adopted, the color augmentations used for the target will be leaked to the model, which spoils the color variation.

BN/LN for Projector and Decoder. We also study how different normalizations will influence the model. The commonly used normalization in Transformer blocks is LayerNorm (LN) [2], while BatchNorm (BN) [31] proves to be important in ID methods [24]. We therefore try to replace LNs with BNs. Note that to preserve the vanilla ViT [19] backbone, we only conduct this replacement for the projector and decoder. Tab. 3(fg) displays that BN gives slightly better results on both linear probing and dense prediction.

	target type	different views	color aug	mask type	BN/LN in proj & dec	loss norm*	loss type	loss form	FT	LIN	AP^b	AP^m
single view with dense loss:												
MAE	pixel			random	LN	MAE-like	dense	L2	83.1	62.5	46.8	42.0
(a)	pixel			random	LN	MAE-like	dense	L2	82.8	62.3	47.3	42.5
(b)	feature			random	LN	MoCo-like	dense	UniGrad	81.0	48.7	43.5	39.2
(c)	pixel		✓	random	LN	MAE-like	dense	L2	82.0	59.9	46.3	41.8
multiple views with dense loss:												
(d)	pixel	✓		random	LN	MAE-like	dense	L2	78.7	46.2	38.1	34.8
(e)	feature	✓		random	LN	MoCo-like	dense	UniGrad	82.9	69.6	48.5	43.4
(f)	feature	✓	✓	random	LN	MoCo-like	dense	UniGrad	83.0	73.1	47.9	43.2
(g)	feature	✓	✓	random	BN	MoCo-like	dense	UniGrad	83.2	73.6	48.7	43.7
(h)	feature	✓	✓	blockwise	BN	MoCo-like	dense	UniGrad	83.5	74.7	50.0	44.5
(i)	feature	✓	✓	blockwise	BN	MAE-like	dense	UniGrad	83.7	76.8	49.8	44.2
(j)	feature	✓	✓	blockwise	BN	MAE-like	dense	L2	83.3	76.5	49.8	44.2
multiple views with global loss:												
(k)	feature	✓	✓	random	BN	MoCo-like	global	UniGrad	82.7	72.0	45.9	41.4
MoCo-v3 with mask	feature	✓	✓	random	BN	MoCo-like	global	UniGrad	82.8	72.2	45.0	40.5

Table 3. Ablations on SiameseIM. We focus on the performances of linear probing, object detection and instance segmentation tasks. All models are pre-trained for 400 epochs. The fine-tuning length is 90 epochs for linear probing and 25 epochs for COCO detection. The line with gray cells is our final setting.*MAE-like loss norm refers to apply LN without affine parameters to target and no normalization to prediction. MoCo-like loss norm refers to apply BN without affine parameters follow by l_2 normalization to both target and prediction.

Mask Type. Tab. 3(gh) compares two mask strategies. It shows that blockwise mask is beneficial for both downstream tasks. We think that it should be easier to reconstruct feature by just interpolating local features. Blockwise mask masks out continuous patches, which makes this hard and forces the model to capture long-range dependency.

Loss Normalization. Tab. 3(hi) ablates normalization in the loss function. MoCo-like normalization applies BN without affine parameters follow by l2 normalization to both target and prediction. MAE-like normalization only applies LN without affine parameters to the target. We find that MAE-like normalization performs better on linear probing and comparable on object detection campared with MoCo-like normalization, thus we adopt MAE-like loss normalization in our default setting.

Loss Form. We study the effect of different losses in Tab. 3(ij). It is shown that UniGrad loss produces slightly higher results than L2 loss.

Dense Supervision. Finally we study the role that the dense supervision plays in Tab. 3(gk). By adopting the dense loss, SiameseIM is able to get an improvement of 2.8 points on object detection and 2.3 points on instance segmentation. This comparison validates our observation that modeling dense representations from a masked image is beneficial for dense prediction tasks. We note that only using dense loss can also help to improve linear probing.

5. Conclusion

In Self-Supervised Learning (SSL), Instance Discrimination (ID) possesses good semantic alignment, while Masked Image Modeling (MIM) has decent spatial sensitivity. In this study, we propose a new SSL framework, namely SiameseIM, to show that it is possible to obtain two properties at the same time within a single dense loss. SiameseIM predicts the dense representations of an augmented view, based on another masked view from the same image. SiameseIM is able to outperform ID and MIM methods over a wide range of downstream tasks. We hope that SiameseIM can bring some insights and inspirations for self-supervised pre-training, and open new possibilities in this domain.

Limitations. The training of SiameseIM is less efficient compared to that of MAE. Using fewer tokens or smaller resolutions for the target branch may reduce the computation burden. Because this paper focuses on exploring the possibility of self-supervised pretraining, *i.e.,* combining linear separability and spatial sensitivity within a single loss, and revealing the connection between ID and MIM methods, we expect to propose a more efficient way to perform SiameseIM pretraining in future work.

Potential negative societal impacts. Our method has similar problems of the SSL paradigm. It requires huge computational resources to conduct large scale pretraining, which may consume a lot of electricity. Furthermore, it may possess biases in its digested data, and therefore should be used with caution.

Acknowledgments The work is partially supported by the National Natural Science Foundation of China under Grants No.U19B2044, No.61836011 and No.62022048. This work is also partially supported by the National Key R&D Program of China(NO.2022ZD0160100), and in part by Shanghai Committee of Science and Technology (Grant No. 21DZ1100100).

References

[1] Mahmoud Assran, Mathilde Caron, Ishan Misra, Piotr Bojanowski, Florian Bordes, Pascal Vincent, Armand Joulin, Michael Rabbat, and Nicolas Ballas. Masked siamese networks for label-efficient learning. *arXiv preprint arXiv:2204.07141*, 2022. 1, 3, 11

[2] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. 7

[3] Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. Data2vec: A general framework for self-supervised learning in speech, vision and language. *arXiv preprint arXiv:2202.03555*, 2022. 3

[4] Hangbo Bao, Li Dong, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021. 1, 3, 4, 6, 11, 12

[5] Adrien Bardes, Jean Ponce, and Yann LeCun. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. *arXiv preprint arXiv:2105.04906*, 2021. 1, 3

[6] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In *NeurIPS*, volume 33, pages 1877–1901, 2020. 3

[7] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, pages 9650–9660, 2021. 3, 4

[8] Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. Generative pretraining from pixels. In *ICML*, pages 1691–1703. PMLR, 2020. 3

[9] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, pages 1597–1607. PMLR, 2020. 1, 3, 4, 5, 7

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

[11] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 3

[12] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *CVPR*, pages 15750–15758, 2021. 1, 3, 7

[13] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *ICCV*, pages 9640–9649, 2021. 3, 4, 6, 7, 11

[14] ImageNet contributors. Imagenet terms of access. https://image-net.org/download, 2020. 13

[15] MMSegmentation Contributors. MMSegmentation: Openmmlab semantic segmentation toolbox and benchmark. https://github.com/open-mmlab/mmsegmentation, 2020. 12

[16] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255. Ieee, 2009. 2, 6

[17] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 3

[18] 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. 3

[19] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 1, 3, 6, 7, 11

[20] 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. 3

[21] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. In *NeurIPS*, volume 33, pages 21271–21284, 2020. 1, 3

[22] Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *CVPR*, pages 5356–5364, 2019. 2, 6, 12

[23] 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. 1, 3, 4, 5, 6, 7, 11

[24] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, pages 9729–9738, 2020. 1, 3, 4, 7

[25] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 3

[26] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *ICCV*, pages 8340–8349, 2021. 2, 6, 7, 13

[27] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019. 2, 6, 7, 13

[28] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *CVPR*, pages 15262–15271, 2021. 2, 6, 7, 13

[29] Tianyu Hua, Wenxiao Wang, Zihui Xue, Sucheng Ren, Yue Wang, and Hang Zhao. On feature decorrelation in self-supervised learning. In *ICCV*, pages 9598–9608, 2021. 3

[30] Zhicheng Huang, Xiaojie Jin, Chengze Lu, Qibin Hou, Ming-Ming Cheng, Dongmei Fu, Xiaohui Shen, and Jiashi

Feng. Contrastive masked autoencoders are stronger vision learners. *arXiv preprint arXiv:2207.13532*, 2022. 4, 6

[31] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, pages 448–456. PMLR, 2015. 6, 7, 11

[32] Longlong Jing and Yingli Tian. Self-supervised visual feature learning with deep neural networks: A survey. *TPAMI*, 43(11):4037–4058, 2020. 1

[33] Chunyuan Li, Jianwei Yang, Pengchuan Zhang, Mei Gao, Bin Xiao, Xiyang Dai, Lu Yuan, and Jianfeng Gao. Efficient self-supervised vision transformers for representation learning. *arXiv preprint arXiv:2106.09785*, 2021. 3

[34] Yanghao Li, Hanzi Mao, Ross Girshick, and Kaiming He. Exploring plain vision transformer backbones for object detection. *arXiv preprint arXiv:2203.16527*, 2022. 11, 12

[35] Yanghao Li, Saining Xie, Xinlei Chen, Piotr Dollar, Kaiming He, and Ross Girshick. Benchmarking detection transfer learning with vision transformers. *arXiv preprint arXiv:2111.11429*, 2021. 1

[36] 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*, pages 740–755. Springer, 2014. 2, 6, 11

[37] Songtao Liu, Zeming Li, and Jian Sun. Self-emd: Self-supervised object detection without imagenet. *arXiv preprint arXiv:2011.13677*, 2020. 3

[38] Julien Mairal. Cyanure: An open-source toolbox for empirical risk minimization for python, c++, and soon more. *arXiv preprint arXiv:1912.08165*, 2019. 11

[39] Shlok Mishra, Joshua Robinson, Huiwen Chang, David Jacobs, Aaron Sarna, Aaron Maschinot, and Dilip Krishnan. A simple, efficient and scalable contrastive masked autoencoder for learning visual representations. *arXiv preprint arXiv:2210.16870*, 2022. 4, 6

[40] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. 2018. 3

[41] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019. 3

[42] 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*, pages 8821–8831. PMLR, 2021. 3

[43] Chenxin Tao, Honghui Wang, Xizhou Zhu, Jiahua Dong, Shiji Song, Gao Huang, and Jifeng Dai. Exploring the equivalence of siamese self-supervised learning via a unified gradient framework. *arXiv preprint arXiv:2112.05141*, 2021. 1, 3, 5, 6

[44] Nenad Tomasev, Ioana Bica, Brian McWilliams, Lars Buesing, Razvan Pascanu, Charles Blundell, and Jovana Mitrovic. Pushing the limits of self-supervised resnets: Can we outperform supervised learning without labels on imagenet? *arXiv preprint arXiv:2201.05119*, 2022. 3

[45] Aaron Van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv e-prints*, pages arXiv–1807, 2018. 3, 6

[46] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, volume 30, 2017. 5, 6

[47] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. *NeurIPS*, 32, 2019. 2, 6, 7, 13

[48] Luya Wang, Feng Liang, Yangguang Li, Wanli Ouyang, Honggang Zhang, and Jing Shao. Repre: Improving self-supervised vision transformer with reconstructive pre-training. *arXiv preprint arXiv:2201.06857*, 2022. 4, 6

[49] Xinlong Wang, Rufeng Zhang, Chunhua Shen, Tao Kong, and Lei Li. Dense contrastive learning for self-supervised visual pre-training. In *CVPR*, pages 3024–3033, 2021. 3, 7

[50] Yizhou Wang, Shixiang Tang, Feng Zhu, Lei Bai, Rui Zhao, Donglian Qi, and Wanli Ouyang. Revisiting the transferability of supervised pretraining: an mlp perspective. *arXiv preprint arXiv:2112.00496*, 2021. 5

[51] Chen Wei, Haoqi Fan, Saining Xie, Chao-Yuan Wu, Alan Yuille, and Christoph Feichtenhofer. Masked feature prediction for self-supervised visual pre-training. *arXiv preprint arXiv:2112.09133*, 2021. 3

[52] Fangyun Wei, Yue Gao, Zhirong Wu, Han Hu, and Stephen Lin. Aligning pretraining for detection via object-level contrastive learning. *NeurIPS*, 34:22682–22694, 2021. 3

[53] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *CVPR*, pages 3733–3742, 2018. 3

[54] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *ECCV*, pages 418–434, 2018. 12

[55] 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 *CVPR*, pages 16684–16693, 2021. 3

[56] Yuwen Xiong, Mengye Ren, Wenyuan Zeng, and Raquel Urtasun. Self-supervised representation learning from flow equivariance. In *ICCV*, pages 10191–10200, 2021. 3

[57] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *ICML*, pages 12310–12320. PMLR, 2021. 1, 3

[58] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *CVPR*, pages 633–641, 2017. 2, 6

[59] 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. 4, 6