Self-Supervised Representation Learning from Arbitrary Scenarios

Zhaowen Li [1,2] Yousong Zhu [1✉] Zhiyang Chen [1,2] Zongxin Gao [7] Rui Zhao [6]
Chaoyang Zhao [1,5] Ming Tang [1] Jinqiao Wang [1,2,3,4,5✉]
Foundation Model Research Center, Institute of Automation, Chinese Academy of Science[1]
School of Artificial Intelligence, University of Chinese Academy of Sciences[2]
Peng Cheng Laboratory[3] Wuhan AI Research[4] Objecteye Inc.[5]
Qing Yuan Research Institute, Shanghai Jiao Tong University[6] Independent Researcher[7]
{zhaowen.li, yousong.zhu, jqwang}@nlpr.ia.ac.cn

Abstract

Current self-supervised methods can primarily be categorized into contrastive learning and masked image modeling. Extensive studies have demonstrated that combining these two approaches can achieve state-of-the-art performance. However, these methods essentially reinforce the global consistency of contrastive learning without taking into account the conflicts between these two approaches, which hinders their generalizability to arbitrary scenarios. In this paper, we theoretically prove that MAE serves as a patch-level contrastive learning, where each patch within an image is considered as a distinct category. This presents a significant conflict with global-level contrastive learning, which treats all patches in an image as an identical category. To address this conflict, this work abandons the non-generalizable global-level constraints and proposes explicit patch-level contrastive learning as a solution. Specifically, this work employs the encoder of MAE to generate dual-branch features, which then perform patch-level learning through a decoder. In contrast to global-level data augmentation in contrastive learning, our approach leverages patch-level feature augmentation to mitigate interference from global-level learning. Consequently, our approach can learn heterogeneous representations from a single image while avoiding the conflicts encountered by previous methods. Massive experiments affirm the potential of our method for learning from arbitrary scenarios.

1. Introduction

Currently, self-supervised learning (SSL) in computer vision [1, 2, 4–7, 18, 21, 22, 35, 59] is capable of adopting self-defined pseudo labels as supervision [28, 29] and holds the promise in leveraging large amounts of unlabeled data to build the foundational models. SSL methods focus on designing different pretext tasks, and there are two mainstream approaches for learning visual features in the community. One of the most promising directions among them is contrastive learning (CL) [2, 5–7, 18, 21, 25, 48]. It assumes that different views of a single image are of the same instance. Its objective is to learn a global-level feature representation that discriminates among images. CL leverages trainable networks to generate semantic pseudo labels [55], thereby enhancing the model's capability to extract semantic information. On the other hand, the masked image mod-

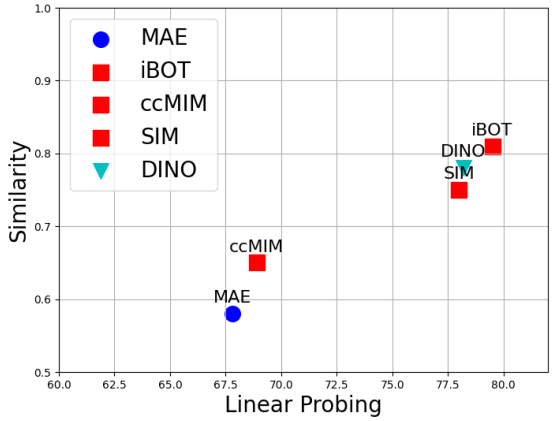

Figure 1. **Hybrid methods further enhance the global consistency of CL.** We measures representational similarity among tokens for some popular self-supervised pre-trained models. Among them, ccMIM [67] and SIM [54] employ MAE as the MIM baseline method while iBOT [69] adopts DINO [3] as the CL baseline method for hybrid approach. It indicates that hybrid methods, while increasing linear evaluation scores, also enhance the representational similarity among tokens. This suggests that hybrid methods fundamentally further strengthen the global semantic consistency of CL. All of the models adopt ViT-B.

Corresponding author: Yousong Zhu and Jinqiao Wang.

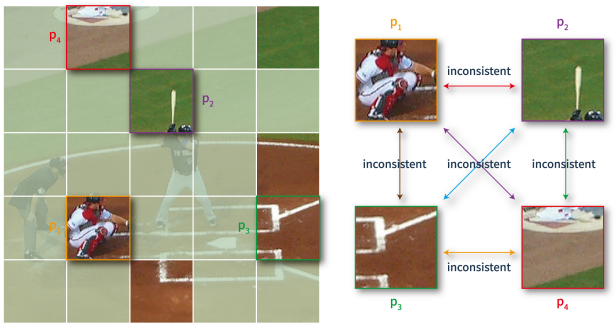

Figure 2. **Visualization of patches in an image from COCO dataset.** The image is partitioned into multiple patches, and different patches from the image may represent different semantics.

eling (MIM) [1, 4, 15, 22, 35, 63, 69] has became the focus in the community. MIM is a pretext task that involves masking some patches on the input image and predicting the original information of the masked patches based on their context. In these methods, BEiT [1] predicts discrete tokens created by the pre-trained model of VQVAE [56]. In MAE [22], the pseudo labels are the given pixels. Among these MIM methods, MAE shows impressive transfer performance on various dense prediction tasks.

The learning methodologies of CL and MIM are so distinct that a fundamental question arises: can their concurrent learning mutually enhance each other? In fact, extensive work [27, 32, 35, 44, 50, 52, 54, 64, 65, 67, 69] shows hybrid methods of CL and MIM achieve state-of-the-art (SOTA) performance on benchmark ImageNet [13] dataset. In our study, we choose some pre-trained models trained using popular self-supervised methods and measure their representational similarity for tokens as well as linear probing results, as depicted in Figure 1. Specifically, ccMIM [67] and SIM [54] employ MAE as the MIM baseline, while iBOT [69] adopts DINO [3] as the CL baseline for the hybrid approach. The figure illustrates that these hybrid methods have achieved better linear probing results compared to their respective baseline methods. However, they have further reinforced representational similarity for tokens, resulting in homogeneous token representations. This also implies that the hybrid methods essentially enhance the global consistency of CL, and the performance of SOTA methods heavily relies on the gains achieved through CL. Consequently, the success of these methods heavily relies on the single-centric-object data [37, 41] such as those in ImageNet dataset, and they cannot be generalized to non-iconic scenarios, making them unsuitable for training in arbitrary scenarios - regardless of whether they come from single-centric-object scenarios or non-iconic ones. Also, a recent study [34] indicates that the hybrid approach of MAE and CL exhibits inferior performance compared to the baseline in large-scale arbitrary scenarios. This goes against the primary promise of self-supervised learning, which aimed to build foundation models using a large amount of data for broader applicability.

Actually, the inability of SOTA methods to train in arbitrary scenarios can be attributed to the conflict between CL and MIM. In this paper, we first theoretically prove that MAE is a patch-level CL, treating each patch within an image as a separate category. This conflicts with global-level CL, where all patches from the image are treated as an identical category. As illustrated in Figure 2, a non-iconic image from the COCO [40] dataset is partitioned into multiple non-overlapping patches according to ViT [15], and each patch may represent a different semantics. MAE treats each object/patch in these patches as a distinct category, whereas CL considers all patches as an identical category. Therefore, when these two methods are combined for training on non-iconic data, significant conflicts and confusion arise, rendering the hybrid approach incapable of learning effective representations for non-iconic data and arbitrary scenarios.

To alleviate the problems, in this paper, we propose a self-supervised framework, called Arbitrary Self-supervised Learning (ASL), to learn representation from arbitrary scenarios. The framework leverages the principles of CL to enhance the patch-level semantic consistency of the model while avoiding the conflict between CL and MAE. Specifically, our approach involves designing a patch-level feature augmentation to effectively circumvent the limitation of global-level data augmentation in CL. According to the patch-level augmentation, ASL employs the encoder of MAE to generate dual-branch features, which then perform patch-level contrastive learning through a decoder module. We also retain the original MIM pretext task to ensure the rationality of the model. It is noteworthy that our approach is proposed for arbitrary scenarios, primarily experimenting on the non-iconic COCO dataset that closely represents natural scenes. Simultaneously, we conduct the pre-training experiments on both ImageNet and COCO. The results indicate that ASL is the sole method of attaining superior performance in arbitrary pre-training, further underscoring the generalizability and robustness of our approach.

Overall, we make the following contributions:

- We theoretically prove that MAE is a patch-level contrastive learning, which is in conflict with global contrastive learning, thereby rendering current hybird methods incapable of achieving generality in arbitrary scenarios pre-training.
- We propose a self-supervised framework, named Arbitrary Self-supervised Learning (ASL), to mitigate conflicts and allow the model to train in arbitrary scenarios.
- Extensive experiments demonstrate the effectiveness and transfer ability of our framework. Specifically, the models pre-trained with ASL achieve SOTA performance in arbitrary scenarios.

2. Related Work

It is known that the promise of self-supervised learning [1, 2, 4–7, 15, 18, 21, 22, 25, 30, 35, 37, 47, 48, 50, 51, 59, 63, 66, 69] in computer vision is to establish a visual foundation model by leveraging large dataset, just like ChatGPT [49] in natural language processing. These representations of the foundation model can be transferred to various downstream tasks [10, 12, 13, 19, 20, 33, 36, 38, 39, 53, 60].

2.1. Instance discrimination

The basic principle of contrastive learning/instance discrimination is that different views of an image are still the same category. As a significant representative of these approaches, MoCo [21] improves the training of contrastive learning methods by storing representations from a momentum encoder instead of the trained network. Then, SimCLR [5] shows that the memory bank can be entirely replaced with the elements from the same batch when the batch size is set large enough. Furthermore, BYOL [18] proposes an asymmetric structure and directly bootstraps the representations by attracting different features from the same instance and shows that contrastive learning without negative samples can also learn excellent visual representations. Also, the asymmetric structure is often directly adopted by subsequent work. Some approaches [23, 24, 58, 61] attempt to transfer global-level prior to pixel-level or region-level learning and acquire visual representations for dense tasks. Also, UniVIP [37] proposes a three-level pre-training task and learns versatile representations from non-iconic images. Recently, MoCo v3 [8] and DINO [3] replace CNNs with ViT [15] and achieve superior performance.

2.2. Masked image modeling

Motivated by MLM of BERT [14] in NLP, the pioneering work iGPT [4] operates on sequences of pixels and predicts unknown pixels. Benefiting from the proposal of Vision Transformer [11, 15, 42], MIM achieves performance comparable to CL. MST [35] is the first to introduce MIM into the siamese structure and proposes the attention-guided mask strategy. Then, iBOT [69] also adopts the siamese structure and obtains impressive performance. However, they have the prior/assumption of image semantic consistency. Besides, BEiT [1] proposes to predict discrete tokens for masked image modeling, yet its transfer performance mainly depends on the quality of the pre-trained model VQ-VAE [56]. Following the MIM design of ViT, SimMIM [63] predicts pixels of masked patches to perform the MIM task. Different from the above methods of performing MIM task in the encoder, MAE [22] designs a special encoder-decoder structure to make the decoder perform the MIM task and abandon the MIM-performing module in downstream visual tasks. Among these methods, MAE shows excellent performance in various downstream tasks.

2.3. Hybrid methods

An intuitive idea is to combine the above approaches [27, 32, 35, 44, 50, 52, 54, 64, 65, 67, 69] to achieve better performance. MST [35] firstly adopts MIM to a contrastive learning framework [3] and restores the pixels of patches, then iBOT [69] predicts the semantic tokens. Moreover, DINO v2 [50] achieves SOTA performance based on the iBOT through several engineering optimizations, with the most significant being data curation. It reduces the original dataset of 1.3 billion images to 142 million images, primarily using the ImageNet dataset as the main set of query images. However, it is not suitable for training in arbitrary scenarios. More methods [27, 32, 44, 52, 54, 64, 65, 67] are based on the MAE framework rather than the MST/iBOT architecture, as it conserves approximately 5 times the computational resources when compared to the MST/iBOT framework under the ViT-B setting. Additionally, as the model scales up, the proportion of saved computational resources increases further. However, current SOTA methods primarily benefit from contrastive learning, thus demonstrating SOTA performance on the benchmark single-centric-object ImageNet dataset. When extended to arbitrary or natural scenes, a pronounced conflict between CL and MAE arises.

3. Approach

3.1. Preliminary

3.1.1 Masked autoencoders

As noted in the prior work [22], for a dataset X without manual annotations, an image x can be divided into n regular non-overlapping patches. MAE first samples a subset of patches and masks the remaining ones, and then acquires the masked patches x_{m_i}, where $i = 1, ..., \eta$ and η is the set of possible masks of length, and visible patches x_{v_j}, where $j = 1, ..., \kappa$ and $\kappa = n - \eta$. The mask process of these patches follows a uniform distribution, named the random mask strategy. Moreover, MAE feds these visible patches into encoder (ViT) $f(\cdot)$, parameterized by θ, and obtains the encoded visible tokens, then puts these encoded visible tokens with $\mathbf{mask}$ tokens into decoder $g(\cdot)$ to perform the MIM task, parameterized by ξ, and obtains the predicted patches. Finally, MAE minimizes the mean squared error (MSE) between the reconstructed and original images in the pixel space. The loss of a single patch prediction is shown as Eq(1).

$$\mathcal{L}_{single}(i) = ||x_{m_i} - g(\xi; f(\theta; x_{v_1}, ..., x_{v_\kappa}, \mathbf{mask}))||^{\mathbf{2}}, \tag{1}$$

For MAE, the total loss function that trains the complete dataset is defined as Eq (2).

$$\mathcal{L}_{MAE} = \mathop{\mathbb{E}}_{\mathbf{x}\sim X} \mathop{\mathbb{E}}_{i\sim\eta} \mathcal{L}_{single}(i), \tag{2}$$

3.1.2 Contrastive learning

Similarly, for an image x in the dataset X, contrastive learning first generates two views under random data augmentation [5, 18], which are then fed into the network separately. Commonly, the network consists of the encoder and momentum encoder. Then, the two global features $\boldsymbol{z}_s$ and $\boldsymbol{z}_t$ can be obtained. Let $\text{sim}(\boldsymbol{z}_s, \boldsymbol{z}_t) = \boldsymbol{z}_s^\top \boldsymbol{z}_t / \|\boldsymbol{z}_s\| \|\boldsymbol{z}_t\|$ denotes the dot product between ℓ_2 normalized $\boldsymbol{z}_s$ and $\boldsymbol{z}_t$ ($i.e.$ cosine similarity). The InfoNCE loss function [5, 21, 48, 59] for a positive pair of examples is defined as Eq (3), where τ is a temperature coefficient. It pulls one prediction closer to the self-defined pseudo label while pushing other predictions in a mini-batch apart.

$$\mathcal{L}_{InfoNCE}(x) = -\log \frac{\exp(\text{sim}(\boldsymbol{z}_s, \boldsymbol{z}_t)/\tau)}{\sum_{k\neq s}^{2N}{}_{k=1} \exp(\text{sim}(\boldsymbol{z}_s, \boldsymbol{z}_k)/\tau)}, \quad (3)$$

Hence, the total loss function that trains the complete dataset is defined as Eq(4).

$$\mathcal{L}_{contras} = \mathop{\mathbb{E}}_{\mathbf{x}\sim X} \mathcal{L}_{InfoNCE}(x), \quad (4)$$

It is noted that the advantage of CL is that semantic learning and its self-defined pseudo label are acquired by the updatable and learnable network.

3.2. The analysis of MAE

According to [46], it is well known that minimizing MSE can be equivalent to maximum likelihood estimation. Therefore, a single patch prediction of MAE can be considered as the mean of a noisy prediction distribution, and the distribution can be modeled as a Gaussian distribution in the classic probabilistic interpretation [43] in Eq (5).

$$p(x_{m_i}|x_{inputs}; \theta, \xi) = \mathcal{N}(x_{m_i}; x_{p_i}, \sigma^2_{noise}\mathbf{I}), \quad (5)$$

x_{inputs} is the input tokens and equal to $\{x_{v_1}, ..., x_{v_\kappa}, \mathbf{mask}\}$, x_{p_i} is the single predicted patch, x_{m_i} is the pseudo label corresponding to this patch, and σ_{noise} is the scale of an independent and identically distributed error term $\epsilon \sim \mathcal{N}(0, \sigma^2_{noise}\mathbf{I})$. Meanwhile, MSE equals to the negative log likelihood (NLL) loss of the prediction distribution $p(x_{m_i}|x_{inputs}; \theta, \xi)$ [46]. The network trained by MSE predicts a single patch in fact learns to model $p(x_{m_i}|x_{inputs})$.

MAE shows impressive performance on various downstream tasks [22], it reconstructs the masked patches through learning from a large amount of data, and the pseudo labels are the masked image and have no interference from manual annotations. Hence, it is reasonable to consider the $p(x_{m_i})$ is uniform, and the MSE is equivalent to Eq(6), where C is a constant and equal to $p(x_{m_i})$, and the detailed derivation of this section in the Appendix.

$$\begin{aligned}\mathcal{L}_{single}(i) \cong &-\log \mathcal{N}(x_{m_i}; x_{p_i}, \sigma^2_{noise}\mathbf{I}) \\ &+ \log \int \mathcal{N}(x_{m_i}; x_{p_i}, \sigma^2_{noise}\mathbf{I}) \cdot p(x_{m_i}) dx_{m_i},\end{aligned} \quad (6)$$

For the second part of Eq(6), we utilize monte carlo method to solve $p(x_{m_i})$ and can obtain the Eq(7).

$$\begin{aligned}&\int \mathcal{N}(x_{m_i}; x_{p_i}, \sigma^2_{noise}\mathbf{I}) \cdot p(x_{m_i}) dx_{m_i} \\ &\approx \frac{1}{N} \sum_{b=1}^{N} \mathcal{N}(x_{m_{i_{(b)}}}; x_{p_i}, \sigma^2_{noise}\mathbf{I})),\end{aligned} \quad (7)$$

The monte carlo method treats all pseudo labels in a training batch as random samples from $p(x_{m_i})$. Hence, for pseudo labels in a training batch $B = \{x_{m_{i_{(1)}}}, x_{m_{i_{(2)}}}, ...x_{m_{i_{(N)}}}\}$, the loss is defined as Eq(8), where $\lambda = 2\sigma^2_{noise}$ is a temperature coefficient.

$$\mathcal{L}_{single}(i) \cong -\log \frac{\exp(-||x_{p_i} - x_{m_i}||^2/\lambda)}{\sum_{x'_{m_i}\in B} \exp(-||x_{p_i} - x'_{m_i}||^2/\lambda)}, \quad (8)$$

According to Eq (3) and Eq (8), it can be observed that the form of the two loss functions is similar. Both of the two loss functions pull together the prediction and self-defined pseudo label in the representation space, and push other predictions in a mini-batch apart. The MAE adopts the Euclidean distance while the contrastive learning measures the cosine similarity in representation space. Actually, MAE regards each masked patch as a category and performs patch-level contrastive learning with unlearnable pixels. It explicitly aligns the prediction with its corresponding masked patch and implicitly distinguishes other patches.

From the above description, when confronted with arbitrary scenes, MAE and CL will face a severe conflict issue.

3.3. Arbitrary self-supervised learning

Based on the previous analysis, it can be concluded that a global prior of CL is both unreasonable and conflicts with MIM when facing arbitrary scenarios. To enable learning from arbitrary scenarios, can there be a self-supervised method that harnesses the advantages of CL in semantic learning (learnable self-pseudo labels) and the heterogeneous representation learning of MIM, while avoiding global constraints and conflicts? The answer is affirmative. In this paper, we propose a new framework, named Arbitrary Self-supervised Learning (ASL), to learn from arbitrary scenarios while retaining their advantages.

Specifically, we adopt the popular MAE method as the basic MIM to ensure the initial heterogeneous representations. Then, inspired by CL using a dual-branch structure

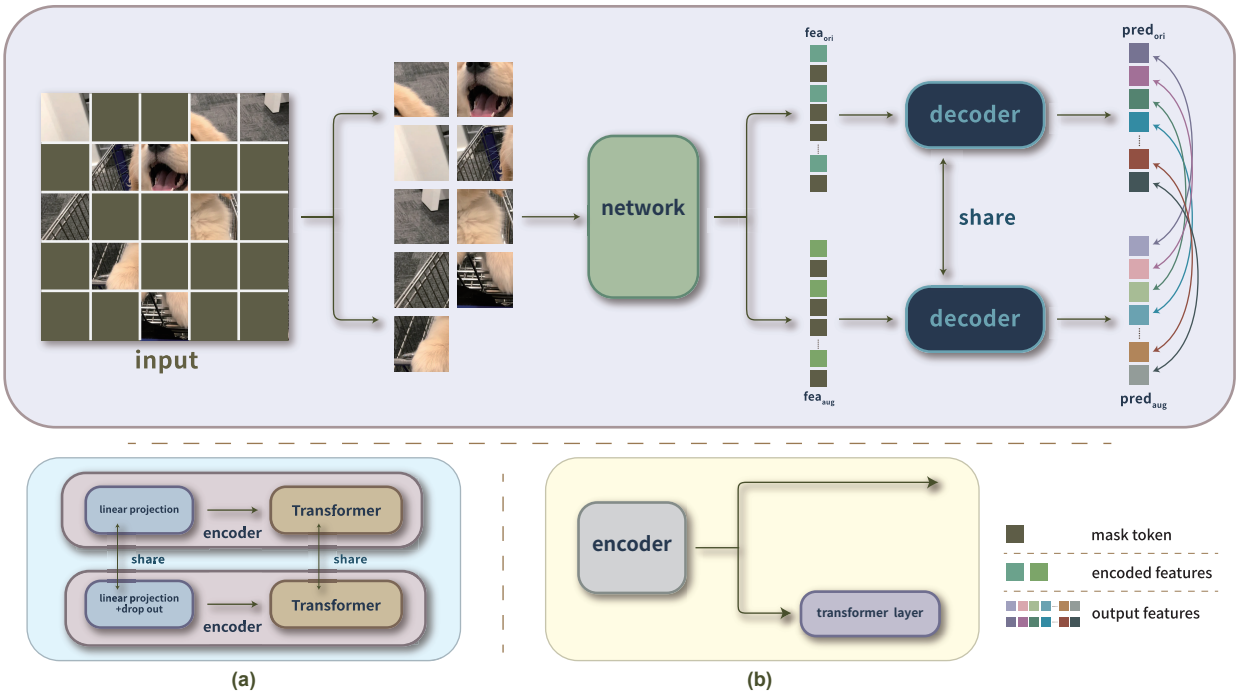

Figure 3. **The pipeline of our ASL.** The visible patches are fed into the network and the dual-branch features can be obtained by our designs. Then the encoded features with mask tokens are fed into the decoder and output the predictions. Finally, we pull together the predictions using L1 loss by one-to-one correspondence. We propose two designs in the network to obtain the dual-branch features: (a) Dropout in the linear embedding. (b) Asymmetric structure at the end of the encoder. In the figure, we adopt the *network* to represent (a) or (b) design. The **olive** patch represents the mask token.

to generate learnable self-pseudo labels, we propose two patch-level designs to generate dual-branch features based on the encoder of MAE while avoiding the global constraints. Subsequently, these dual-branch features are further processed through a decoder module, and CL is performed between the outputs patch by patch to for semantic learning. In contrast to traditional CL, our approach does not entail global data augmentation but instead focuses on patch-level feature augmentation. The specifics of our designs are outlined as follows:

Dropout in the linear embedding (DLE). According to previous work in NLP, SimCSE [17] has explored the use of dropout in word embedding for feature augmentation. In line with this, we also adopt dropout in the linear embedding of the encoder to perform patch-level augmentation and acquire two encoded features at the end of the encoder. The details of our method are described in Figure 3 (a).

Asymmetric structure at the end of the encoder (AEE). Some prior work [8, 18, 62] have demonstrated that introducing an asymmetric structure, which essentially is a nonlinear function, enhances the semantic features. With this in mind, we incorporate a Transformer layer [57] after the encoder in our approach to build an asymmetric structure for patch-level feature augmentation. This step produces dual-branch features: one is the feature output by the encoder, and the other is the feature that has passed through the Transformer layer. Figure 3 (b) illustrates the details of our method. The design is our default experimental setup.

The methodology described above pertains to the construction of feature augmentation. Our rationale for utilizing feature augmentation to construct dual-branch features is primarily based on two considerations: 1) Data augmentation in CL is global and not suitable for patch-level learning. 2) The asymmetric structure of CL inherently constitutes an implicit form of feature augmentation, and it is worth exploring whether its success in CL can be extrapolated to the broader domain of self-supervised learning.

After feature augmentation, we obtain the dual-branch features, fea_{ori} and fea_{aug}, then we feed them into the decoder module to obtain the corresponding predictions, $pred_{ori}$ and $pred_{aug}$, respectively. The patch-level CL between the outputs does not employ contrastive loss or cosine similarity loss but rather L1 loss.

The alignment between any pair is defined as Eq(9), where $\mathtt{sg}[\cdot]$ stands for stop gradient. It is noted that our alignment method does not include the alignment of $[\mathtt{cls}]$ token according to Section 3.2, which unleashes the potential of the method for training in arbitrary scenarios.

$$\mathcal{L}_s(i) = ||pred_{ori_i} - sg[pred_{aug_i}]|| + ||sg[pred_{ori_i}] - pred_{aug_i}||, \tag{9}$$

The loss function of the complete dataset is defined as Eq(10).

$$\mathcal{L}_{SEM} = \mathop{\mathbb{E}}_{\mathbf{x}\sim X} \mathop{\mathbb{E}}_{i\sim\eta} \mathcal{L}_s(i), \tag{10}$$

3.4. Objective function

The loss function of our ASL consists of the MAE loss and the semantic loss. Hence, the total loss of our ASL is formulated as Eq (11), and each loss coefficient is set to be 1 for equally weighted.

$$\mathcal{L}_{final} = \mathcal{L}_{MAE} + \mathcal{L}_{SEM}, \tag{11}$$

4. Experiments

4.1. Datasets and evaluation metrics

Pre-training datasets. In the pre-training stage, we select the four popular visual pre-training datasets, MS COCO [40], ImageNet [13], CIFAR10 [31], CIFAR100 [31], to investigate the versatility and transfer ability of the proposed method. For the COCO dataset, its `train2017` set contains ∼118k images. COCO contains more natural and diverse scenes and is a non-iconic image dataset. To further demonstrate the versatility of our algorithm for training in arbitrary scenarios, we conduct pre-training on both ImageNet and COCO datasets. Given the substantial difference in data volume between ImageNet and COCO, we select a subset of ImageNet, referred to as ImageNet-100, to maintain a balanced dataset size. Moreover, we pre-train on ImageNet-1K to demonstrate the effectiveness of ASL on curated data. Finally, CIFAR-10 and CIFAR-100 datasets are small-scale image datasets, and they all contain 60000 images with 32×32 size that belong to 10 and 100 categories, respectively. We pre-train models on the CIFAR-10 and CIFAR-100 datasets to demonstrate the effectiveness of ASL in the small-scale dataset.

Evaluation datasets. In this paper, ASL is evaluated by the linear probing and fine-tuning classification task on ImageNet, CIFAR-10, and CIFAR-100 datasets, which are popular image classification datasets. Additionally, for computer vision, object detection, and semantic segmentation are dense prediction tasks since training images of these tasks contain massive objects inside each image. The validation of dense prediction tasks can better reflect the transfer ability of self-supervised models. Therefore, we conduct extensive experiments on COCO [40] and ADE20k [68] datasets to verify the generalization and transfer ability of ASL. COCO is a popular object detection and instance segmentation dataset, which `train2017` contains about $118k$ images and test on the `val2017` contains $5k$ images. The challenging dataset contains human annotations for 80 classes. Moreover, ADE20K is also a challenging semantic segmentation dataset, and it contains about $25k$ training images and $2k$ validation images with 150 categories.

Method	Arch.	FLOPs	Epochs	LP	AP^b
Pre-training on COCO dataset:					
MAE [22]	ViT-B	1×	800	46.8%	49.3%
CMAE [27]	ViT-B	∼ 3×	800	46.6%	49.0%
ccMIM [67]	ViT-B	∼ 3.3×	800	46.7%	49.0%
SIM [54]	ViT-B	∼ 2.8×	800	46.0%	48.7%
iBOT [69]	ViT-B	∼ 6×	800	45.2%	48.6%
ASL	ViT-B	∼ 1.5×	800	**48.6%**	**50.3%**

Table 1. **Evaluation of pre-trained models on the COCO dataset.** We report Linear Probing (LP) top-1 accuracy on the ImageNet-1K `val` set and bbox mAP on COCO dataset.

Method	Arch.	Param.	Epochs	LP	FT
Pre-training on ImageNet-100 + COCO:					
MAE [22]	ViT-B	86	800	74.5%	90.9%
iBOT[69]	ViT-B	86	800	71.8%	88.5%
ccMIM [67]	ViT-B	86	800	72.3%	88.7%
ASL	ViT-B	86	800	79.6%	92.4%
ASL	ViT-B	86	4000	**85.9%**	**94.2%**

Table 2. **ImageNet-100 Top-1 accuracy of different methods under linear probing (LP) and fine-tuning (FT) setting to evaluate the performance of the model pre-trained on both ImageNet-100 and COCO.** We report top-1 accuracy on the ImageNet-100 `val` set.

The valuation metrics, pre-training settings, and the experimental settings of downstream tasks in the Appendix.

4.2. Pre-training on COCO dataset

In order to validate the adaptability of our approach for training in arbitrary scenarios, we initially conduct pre-training on the COCO dataset and subsequently conduct a fair comparison with other methods under the same experimental settings. In this comparison, we evaluate the ImageNet linear evaluation and the COCO detection results. Linear evaluation, which involves using frozen features, provides a more comprehensive assessment of a model's representation. Given that the pre-training experiments are conducted on the COCO dataset, using human-labeled data from COCO to test the learned representations is reasonable. As shown in Table 1, we observe that methods combining MAE with contrastive learning, such as CMAE, ccMIM, and SIM, exhibit weaker performance when pre-trained on the COCO dataset, even though they achieved SOTA performance on ImageNet. This observation suggests that existing SOTA methods are not well-suited for training in arbitrary scenarios. Furthermore, our approach demonstrates superior performance after pre-training on the COCO dataset, surpassing the performance of MAE. Generally, introducing a dual-branch structure similar to contrastive learning typically results in a significant increase in computational overhead, often exceeding threefold. Diverging from the emphasis on the encoder in hybrid methods, our AEE design prioritizes the decoder, significantly reduc-

Method	Pre-train Iterations	Dataset	Object detection			Instance segmentation		
			AP^b	AP^b_{50}	AP^b_{75}	AP^m	AP^m_{50}	AP^m_{75}
Supervised [9]	$\sim 94k$	ImageNet-1K	47.9%	-	-	42.9%	-	-
MoCo v3 [8]	$\sim 187k$	ImageNet-1K	47.9%	-	-	42.7%	-	-
BEiT [1]	$\sim 249k$	ImageNet-1K + DALLE	49.8%	-	-	44.4%	-	-
MAE [22]	$\sim 499k$	ImageNet-1K	50.4%	70.8%	55.7%	44.9%	68.3%	48.9%
CAE [9]	$\sim 499k$	ImageNet-1K	50.0%	70.9%	54.8%	44.0%	67.9%	47.6%
R-MAE [45]	$\sim 115k$	COCO	50.6%	-	-	45.0%	-	-
R-MAE [45]	$\sim 250k$	COCO	50.8%	-	-	45.2%	-	-
ccMIM [67]	$\sim 249k$	ImageNet-1K	50.3%	71.2%	55.0%	44.5%	68.4%	47.9%
ASL	$\sim 47k$	ImageNet-100 + COCO	50.7%	71.1%	55.6%	45.0%	68.3%	49.0%
ASL	$\sim 236k$	ImageNet-100 + COCO	**51.5%**	**71.9%**	**57.2%**	**45.7%**	**69.2%**	**50.0%**

Table 3. **Results of object detection and instance segmentation on COCO.** The architecture of various methods is ViT-B, and we report the bounding box AP and mask AP on COCO `val2017`.

Method	Dataset	Pre-train Iterations	ADE mIoU
Supervised [22]	ImageNet-1K	$\sim 93k$	47.4%
SplitMask [16]	ADE20K	$\sim 128k$	45.7%
MoCo v3 [8]	ImageNet-1K	$\sim 187k$	47.3%
BEiT [22]	ImageNet-1K+DALLE	$\sim 249k$	47.1%
R-MAE [45]	COCO	$\sim 115k$	46.8%
R-MAE [45]	COCO	$\sim 250k$	47.0%
ccMIM [67]	ImageNet-1K	$\sim 249k$	47.7%
MAE [22]	ImageNet-1K	$\sim 499k$	48.1%
ASL	ImageNet-100 + COCO	$\sim 236k$	**48.4%**

Table 4. **Results of semantic segmentation on ADE20K.** We report results measured by mIoU.

ing computational costs in comparison. Meanwhile, due to the fixed-size and lightweight design of the decoder, our design incurs decreasing computational overhead as the model scale increases. This highlights the potential of our method for pre-training in non-iconic scenarios.

4.3. Pre-training on both COCO and ImageNet

To further demonstrate the adaptability of ASL across diverse scenarios, we conduct the pre-training experiments on both ImageNet-100 and COCO datasets. We assess the performance of the model on benchmark tasks: image classification, object detection, and semantic segmentation. The ASL models are primarily trained with 800 epochs and 4000 epochs, with the number of training iterations for the 4000-epoch training similar to that of the ImageNet-1K iterations for the 800-epoch training.

4.3.1 Image classification

Since our training data comprises the ImageNet-100 dataset, we conducted testing for classification performance through both linear probing and fine-tuning on ImageNet-100. The experimental results are presented in Table 2. The results of pre-training observed from the table align with those of pre-training solely on COCO: ccMIM exhibits relatively weak performance in the context of the pre-training. This can be attributed to conflicts between CL and MAE when dealing with arbitrary scenarios. In contrast, our method successfully mitigates such conflicts and produces more robust representations in arbitrary scenarios.

4.3.2 Object detection and instance segmentation

To verify the transfer ability of ASL, we test it on object detection and instance segmentation with COCO [40] dataset. Here we report Box AP and mask AP on the validation set. According to the benchmark validation [22, 33], we choose the Mask R-CNN [20] as the test framework. Due to variations in the sizes of the pre-training datasets used, we opt to avoid epoch-based comparisons in this scenario. We convert epochs to a fixed batch size of 4096 iterations for a fair comparison. Table 3 shows the results of the learned representation by supervised method and different self-supervised methods. It shows our ASL, with approximately 236k iterations (4000 epochs), achieves impressive performance and outperforms 1600-epoch MAE, 1600-epoch CAE and 800-epoch ccMIM pre-trained on ImageNet. ASL surpasses various supervised and self-supervised models pre-trained on ImageNet and COCO, indicating its success in arbitrary scenarios.

4.3.3 Semantic segmentation

To further evaluate the generalization of our method, we conduct the semantic segmentation experiments on ADE20K dataset [68]. As shown in Table 4, ASL achieves the 48.4% mIoU. It also outperforms various supervised and self-supervised models pre-trained on ImageNet.

Overall, our ASL shows impressive performance on various mainstream visual benchmark tasks when pre-training on both COCO and ImageNet-100, which demonstrates

Method	LP	FT	AP^b	AP^m	mIoU
MAE	67.8%	83.6%	50.4%	44.9%	48.1%
ASL	69.2%	84.2%	51.3%	45.7%	49.2%

Table 5. **The ImageNet-1K pre-training results.** We report ImageNet classification, COCO detection, and instance segmentation, as well as ADE20k semantic segmentation results.

Method	CIFAR-10		CIFAR-100	
	Linear Prob.	*Finetuning*	*Linear Prob.*	*Finetuning*
Supervised [26]	-	91.3%	-	64.1%
DINO [3]	89.0%	94.4%	**65.8%**	76.3%
MAE [22]	87.3%	95.9%	54.0%	81.1%
ASL	**91.4%**	**96.5%**	60.3%	**82.1%**

Table 6. **Linear probing and fine-tuning classification results on small-scale datasets.** We adopt ViT-S as the base model.

ID	DLE	AEE	alignment	FLOPs	Top-1	AP^b
(a)	×	×	×	1×	46.8%	49.3%
(b)	✓	×	×	∼ 2×	46.5%	49.1%
(c)	✓	×	✓	∼ 2×	48.5%	**50.4%**
(d)	×	✓	×	∼ 1.5×	46.6%	49.2%
(e)	×	✓	✓	∼ 1.5×	**48.6%**	50.3%

Table 7. **Ablations for ASL: Effect of feature augmentation.** We conduct pre-training experiments on COCO dataset and report linear evaluation on ImageNet-1K and detection results on COCO.

ASL has the ability to learn general visual representations from arbitrary scenarios.

4.4. Pre-training on ImageNet-1K

Our method is designed for learning visual representations from uncurated arbitrary scenarios, and it can also yield stronger representations when applied to meticulously curated datasets. To demonstrate this point, we conduct experiments on the curated ImageNet dataset. As depicted in the experimental results shown in Table 5, ASL exhibits superior performance across various downstream tasks compared to MAE, highlighting the versatility of our approach.

4.5. Small-scale data pre-training

As shown in Table 6, the results of our method on these small datasets all show impressive performance. With the same training settings for all experiments, ASL consistently improves MAE by at least 4.1% for linear probing. However, MIM still lags behind CL in linear evaluation on CIFAR100 and its advantage lies in the fine-tuning superior performance. Our fine-tuning results also reveal it. Totally, these results suggest that ASL is effective and can learn more semantic features than MAE in small-scale datasets.

4.6. Ablation study

The effect of feature augmentation. We propose two designs to obtain a dual-branch structure by performing feature augmentation. As shown in Table 7, (a) is the results of MAE as the baseline when pre-trained on COCO. Table 7 (b) indicates the results of the DLE design are slightly inferior to the baseline. However, the results outperform the baseline MAE when aligning the predictions according to (c). Similarly, the results for (d) using AEE are not as favorable as MAE, while the results for (e) incorporating alignment operations outperform MAE. These experimental findings collectively indicate that creating learnable pseudo labels based on feature augmentation during the pre-training phase can improve the representation of pre-trained models. Due to the relatively comparable performance of (e) and (c), with (e) incurring significantly lower computational costs, we opt for (e) as the final experimental design.

The effect of alignment loss. We compare two loss functions: one commonly used in contrastive learning, which is cosine similarity loss, and the other is L1 loss. The following is a comparison between cosine similarity loss and L1 loss:

MAE	cosine similarity loss	L1 loss	[cls] align
46.8%	47.6%	48.6%	45.6%

From the above, experimental results indicate that the L1 loss facilitates the model in learning better representations compared to cosine similarity loss. Besides, we conduct an additional experiment involving alignment with the [cls] token on top of the L1 loss. The results indicate that introducing global contrastive learning during pre-training on the COCO dataset has a detrimental effect on performance.

5. Conclusions

This paper points out the inability of current hybrid SOTA methods to extend training to arbitrary scenarios due to the conflict between MAE and CL. It then theoretically identifies the root of the issue as MAE being patch-level CL, conflicting with global CL. To address this conflict, we introduce a novel self-supervised learning method ASL to learn visual representations from arbitrary scenarios. Specifically, this paper abandons the global-level consistency of CL and proposes explicit patch-level learning. ASL adopts patch-level feature augmentation to generate dual-branch features by the encoder of MAE, then these features perform patch-level learning through a decoder module. Finally, ASL can learn heterogeneous representations while avoiding the conflicts. Our method demonstrates robustness and versatility in multiple pre-training datasets and downstream visual tasks. We expect that our study will draw the attention of the community to large-scale pre-train from arbitrary scenarios and contribute to the development of visual foundation models.

Acknowledgement. This work was supported by National Key R&D Program of China under Grant No. 2021YFE0205700 and National Natural Science Foundation of China (No.62276260, No.62176254, No.62076235).

References

[1] Hangbo Bao, Li Dong, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021. 1, 2, 3, 7

[2] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *Advances in Neural Information Processing Systems*, pages 9912–9924, 2020. 1, 3

[3] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. *arXiv: Computer Vision and Pattern Recognition*, 2021. 1, 2, 3, 8

[4] Mark Chen, Alec Radford, Rewon Child, Jeffrey K Wu, Heewoo Jun, David Luan, and Ilya Sutskever. Generative pretraining from pixels. In *International Conference on Machine Learning*, pages 1691–1703, 2020. 1, 2, 3

[5] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020. 1, 3, 4

[6] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey E. Hinton. Big self-supervised models are strong semi-supervised learners. In *Advances in Neural Information Processing Systems*, pages 22243–22255, 2020.

[7] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 1, 3

[8] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. *arXiv preprint arXiv:2104.02057*, 2021. 3, 5, 7

[9] 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. 7

[10] Zhiyang Chen, Yousong Zhu, Zhaowen Li, Fan Yang, Wei Li, Haixin Wang, Chaoyang Zhao, Liwei Wu, Rui Zhao, Jinqiao Wang, et al. Obj2seq: Formatting objects as sequences with class prompt for visual tasks. *Advances in Neural Information Processing Systems*, 35:2494–2506, 2022. 3

[11] Zhiyang Chen, Yousong Zhu, Fan Yang, Zhaowen Li, Chaoyang Zhao, Jinqiao Wang, and Ming Tang. The devil is in details: Delving into lite ffn design for vision transformers. *Available at SSRN 4299967*, 2023. 3

[12] Zhiyang Chen, Yousong Zhu, Yufei Zhan, Zhaowen Li, Chaoyang Zhao, Jinqiao Wang, and Ming Tang. Mitigating hallucination in visual language models with visual supervision. *arXiv preprint arXiv:2311.16479*, 2023. 3

[13] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 2, 3, 6

[14] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina N. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, 2018. 3

[15] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. 2, 3

[16] 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. 7

[17] Tianyu Gao, Xingcheng Yao, and Danqi Chen. Simcse: Simple contrastive learning of sentence embeddings. *arXiv preprint arXiv:2104.08821*, 2021. 5

[18] 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, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent: A new approach to self-supervised learning. In *Advances in Neural Information Processing Systems*, pages 21271–21284, 2020. 1, 3, 4, 5

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 3

[20] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *The IEEE International Conference on Computer Vision*, 2017. 3, 7

[21] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. *arXiv preprint arXiv:1911.05722*, 2019. 1, 3, 4

[22] 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, 2, 3, 4, 6, 7, 8

[23] Olivier J Hénaff, Skanda Koppula, Jean-Baptiste Alayrac, Aaron Van den Oord, Oriol Vinyals, and Joao Carreira. Efficient visual pretraining with contrastive detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10086–10096, 2021. 3

[24] Olivier J Hénaff, Skanda Koppula, Evan Shelhamer, Daniel Zoran, Andrew Jaegle, Andrew Zisserman, João Carreira, and Relja Arandjelović. Object discovery and representation networks. In *European Conference on Computer Vision*, pages 123–143. Springer, 2022. 3

[25] R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. *International Conference on Learning Representations*, 2019. 1, 3

[26] Tianyu Hua, Yonglong Tian, Sucheng Ren, Hang Zhao, and Leonid Sigal. Self-supervision through random seg-

ments with autoregressive coding (randsac). *arXiv preprint arXiv:2203.12054*, 2022. 8

[27] 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. 2, 3, 6

[28] Ashish Jaiswal, Ashwin Ramesh Babu, Mohammad Zaki Zadeh, Debapriya Banerjee, and Fillia Makedon. A survey on contrastive self-supervised learning. *Technologies*, 9(1): 2, 2020. 1

[29] Longlong Jing and Yingli Tian. Self-supervised visual feature learning with deep neural networks: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 43(11):4037–4058, 2020. 1

[30] Nikos Komodakis and Spyros Gidaris. Unsupervised representation learning by predicting image rotations. In *International Conference on Learning Representations*, 2018. 3

[31] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 6

[32] Youngwan Lee, Jeffrey Willette, Jonghee Kim, Juho Lee, and Sung Ju Hwang. Exploring the role of mean teachers in self-supervised masked auto-encoders. *arXiv preprint arXiv:2210.02077*, 2022. 2, 3

[33] 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. 3, 7

[34] Yanghao Li, Haoqi Fan, Ronghang Hu, Christoph Feichtenhofer, and Kaiming He. Scaling language-image pre-training via masking. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 23390–23400, 2023. 2

[35] Zhaowen Li, Zhiyang Chen, Fan Yang, Wei Li, Yousong Zhu, Chaoyang Zhao, Rui Deng, Liwei Wu, Rui Zhao, Ming Tang, and Jinqiao Wang. Mst: Masked self-supervised transformer for visual representation. In *Advances in Neural Information Processing Systems*, 2021. 1, 2, 3

[36] Zhaowen Li, Xu Zhao, Chaoyang Zhao, Ming Tang, and Jinqiao Wang. Transfering low-frequency features for domain adaptation. In *2022 IEEE International Conference on Multimedia and Expo (ICME)*, pages 01–06. IEEE, 2022. 3

[37] Zhaowen Li, Yousong Zhu, Fan Yang, Wei Li, Chaoyang Zhao, Yingying Chen, Zhiyang Chen, Jiahao Xie, Liwei Wu, Rui Zhao, et al. Univip: A unified framework for self-supervised visual pre-training. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 14627–14636, 2022. 2, 3

[38] Zhaowen Li, Xu Zhao, Peigeng Ding, Zongxing Gao, Yuting Yang, Ming Tang, and Jinqiao Wang. Freconv: Frequency branch-and-integration convolutional networks. In *2023 IEEE International Conference on Multimedia and Expo (ICME)*, pages 258–263. IEEE, 2023. 3

[39] Zhaowen Li, Yousong Zhu, Zhiyang Chen, Wei Li, Chaoyang Zhao, Liwei Wu, Rui Zhao, Ming Tang, and Jinqiao Wang. Efficient masked autoencoders with self-consistency. *arXiv preprint arXiv:2302.14431*, 2023. 3

[40] 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 *Proceedings of the European Conference on Computer Vision*, 2014. 2, 6, 7

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

[42] Yang Liu, Yao Zhang, Yixin Wang, Feng Hou, Jin Yuan, Jiang Tian, Yang Zhang, Zhongchao Shi, Jianping Fan, and Zhiqiang He. A survey of visual transformers. *IEEE Transactions on Neural Networks and Learning Systems*, 2023. 3

[43] Peter McCullagh and John A Nelder. *Generalized linear models*. Routledge, 2019. 4

[44] 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. 2, 3

[45] Duy-Kien Nguyen, Vaibhav Aggarwal, Yanghao Li, Martin R Oswald, Alexander Kirillov, Cees GM Snoek, and Xinlei Chen. R-mae: Regions meet masked autoencoders. *arXiv preprint arXiv:2306.05411*, 2023. 7

[46] David A Nix and Andreas S Weigend. Estimating the mean and variance of the target probability distribution. In *Proceedings of 1994 ieee international conference on neural networks (ICNN'94)*, pages 55–60. IEEE, 1994. 4

[47] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *Proceedings of the European Conference on Computer Vision*, 2016. 3

[48] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 1, 3, 4

[49] OpenAI. Introducing chatgpt, 2022. 3

[50] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 2, 3

[51] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016. 3

[52] Hao Quan, Xingyu Li, Weixing Chen, Mingchen Zou, Ruijie Yang, Tingting Zheng, Ruiqun Qi, Xinghua Gao, and Xiaoyu Cui. Global contrast masked autoencoders are powerful pathological representation learners. *arXiv preprint arXiv:2205.09048*, 2022. 2, 3

[53] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*, 2015. 3

[54] Chenxin Tao, Xizhou Zhu, Gao Huang, Yu Qiao, Xiaogang Wang, and Jifeng Dai. Siamese image modeling for self-supervised vision representation learning. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2023. 1, 2, 3, 6

[55] Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning? *Advances in Neural Information Processing Systems*, 33:6827–6839, 2020. 1

[56] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017. 2, 3

[57] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 5998–6008, 2017. 5

[58] Xinlong Wang, Rufeng Zhang, Chunhua Shen, Tao Kong, and Lei Li. Dense contrastive learning for self-supervised visual pre-training. *arXiv preprint arXiv:2011.09157*, 2020. 3

[59] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 1, 3, 4

[60] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *Proceedings of the European Conference on Computer Vision*, pages 418–434, 2018. 3

[61] Zhenda Xie, Yutong Lin, Zheng Zhang, Yue Cao, Stephen Lin, and Han Hu. Propagate yourself: Exploring pixel-level consistency for unsupervised visual representation learning. *arXiv preprint arXiv:2011.10043*, 2020. 3

[62] 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*, 2021. 5

[63] 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*, 2021. 2, 3

[64] Kun Yi, Yixiao Ge, Xiaotong Li, Shusheng Yang, Dian Li, Jianping Wu, Ying Shan, and Xiaohu Qie. Masked image modeling with denoising contrast. *arXiv preprint arXiv:2205.09616*, 2022. 2, 3

[65] Qi Zhang, Yifei Wang, and Yisen Wang. How mask matters: Towards theoretical understandings of masked autoencoders. *arXiv preprint arXiv:2210.08344*, 2022. 2, 3

[66] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *Proceedings of the European Conference on Computer Vision*, 2016. 3

[67] Shaofeng Zhang, Feng Zhu, Rui Zhao, and Junchi Yan. Contextual image masking modeling via synergized contrasting without view augmentation for faster and better visual pretraining. In *The Eleventh International Conference on Learning Representations*. 1, 2, 3, 6, 7

[68] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. *International Journal of Computer Vision*, 127(3):302–321, 2019. 6, 7

[69] 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. 1, 2, 3, 6