Information-density Masking Strategy for Masked Image Modeling

He Zhu*†, Yang Chen*, Guyue Hu‡, and Shan Yu*†§
*Brainnetome Center, National Laboratory of Pattern Recognition (NLPR),
Institute of Automation, Chinese Academy of Sciences(CASIA)
†School of Future Technology, University of Chinese Academy of Sciences(UCAS)
‡School of Computer Science and Engineering, Nanyang Technological University
§School of Artificial Intelligence, University of Chinese Academy of Sciences(UCAS)
{he.zhu, shan.yu}@nlpr.ia.ac.cn, yang.chen@ia.ac.cn, guyue.hu@ntu.edu.sg

***Abstract*—Recent representation learning approaches mainly fall into two paradigms: contrastive learning (CL) and masked image modeling (MIM). Combining these two methods may boost the performance, but its learning process still heavily depends on the random masking strategy. We conjecture that the random masking may hinder learning the comprehensive relationship between concept and visual patches. To overcome these limitations, we propose an information-density masking (IDM) strategy for general visual transformers. Specifically, the IDM mask out the visual patches according to their activation values of attention maps. To obtain the attention maps before the reconstruction, a self-supervised training framework CAMAE is further proposed. In addition, in order to reduce the redundancy among different attention maps, we introduce a pattern-learning balance (PLB) sampling to adaptively adjust the learning progress in different attention spaces. Extensive experiments indicate that our method efficiently retains more comprehensive visual characteristics and achieves state-of-the-art performance.**

***Index Terms*—Masked image modeling, contrastive learning, unsupervised learning**

I. Introduction

In recent years, contrastive learning (CL) and masked image modeling (MIM) have attracted wide interest in the field of self-supervised learning (SSL). CL-based methods [1]–[3] distinguish whether two input images come from the same instance, while MIM methods [4] reconstruct the masked image from visible patches that capture semantic features. Both methods could enable the visual encoder to learn rich and holistic representations. However, Hinton et al. [1] pointed out that the CL may lead the encoder to remove the low-level transformation-dependent characteristics, which can be harmful to generalized representation learning. Meanwhile, previous works [4] found that the MIM requires much more training epochs, and it only captures limited high-level semantics within local details. How to learn comprehensive visual representation is still an open question.

Previous works find that combining CL and MIM improves the quality of learned representations. iBOT [5] proposes a method of self-distillation learning by combining CL and MIM, but this method is hard to capture pixel-level visual characteristics and needs many training epochs. CMAE [6] designs a multi-task framework to coordinate the learning of pixel-level and conceptual representations, but the training efficiency problem is still unsolved.

We conjecture that the random masking strategy limits the efficiency of learning the relationship between concept and visual patches: the random masking strategy may contain many semantic-irrelevant patches, thus the training process needs more sampling attempts to capture useful characteristics. Moreover, previously proposed masking approaches like grid or block masks both fail to achieve effective semantic-wise masking [4]. Importantly, DINO [7] identifies that self-supervised visual transformer (ViT) enables attention maps to locate the semantic regions of images. Inspired by this finding, here we propose an information-density masking (IDM) strategy to produce effective mask patterns as shown in Figure 1. To provide the attention maps for IDM, we suggest the Contrastive Attentional Masked Auto-encoder (CAMAE) framework. Specifically, IDM samples an attention map from CAMAE with multiple attention heads in each step, and it sorts the visual patches based on their activation values of the sampled attention map. In such a manner, IDM masks the most informative 75% patches for MIM training. Finally, we find that the captured total patterns have strong redundancy though each attention space only captures a specific mode, suggesting that the learning progress among patterns could be imbalanced. To overcome this issue, we design the pattern-learning balance (PLB) sampling to balance the learning among different attention spaces. In summary, the contribution of this work includes:

- We propose an information-density masking (IDM) strategy, which produces informative masks for MIM through attention information. The IDM significantly reduces the training epochs needed for MIM training (from 1600 epochs to 400 epochs).
- We discover that the visual patterns captured by different attention maps have significant overlaps that induce unbalanced mask modeling between foreground and background. Thus we design a pattern-learning balance

This work was funded by the Beijing Advanced Discipline Fund, CAS Project for Young Scientists in Basic Research YSBR-041, the International Partnership Program of CAS 173211KYSB2020002, and the National Natural Science Foundation of China 11905291.

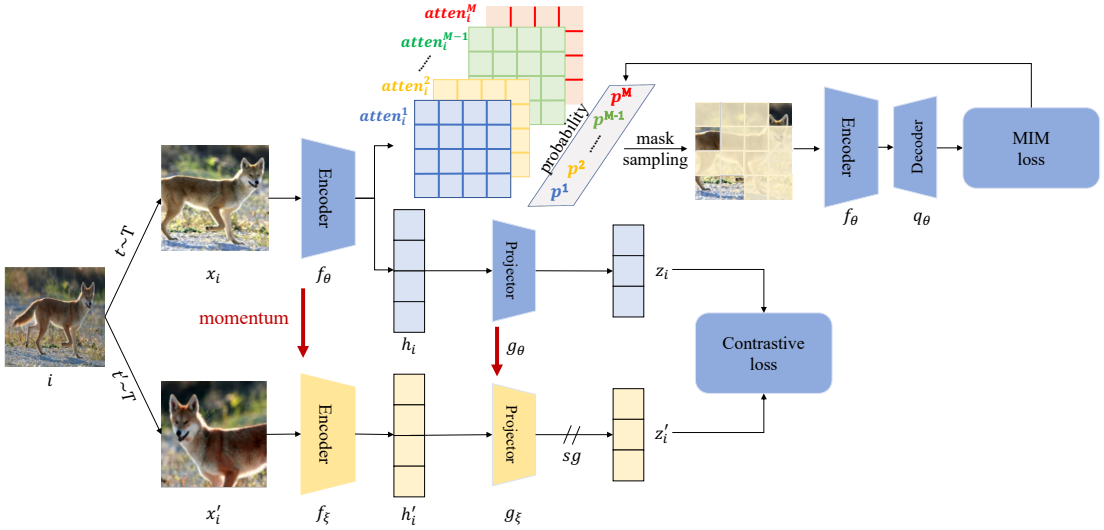


Fig. 1: The pipeline of the proposed CAMAE framework. $T_i/T_i^{'}$ are the random data augmentation operators sampled from the same family of augmentations. $*_\theta$ are the parameters of the encoder network, and $*_\xi$ represent the parameters are momentum updated by $*_\theta$ and sg means stop-gradient. $\{atten_i^1, .., atten_i^M\}$ means the set of multiple attention maps (number=M) of i-th images in the last block of ViT. $[p^1, ..., p^M]$ is an M-dimension vector representing sampling probabilities distribution. The "MIM loss" is used to update the sampling probabilities according to our pattern-learning balanced strategy. Noted that, f_θ is the same network that appears twice in the pipeline.

(PLB) strategy to re-balance the visual patterns during the learning progress. The PLB eventually improves the linear performance by 3% than the random sampling strategy.

- We propose a new self-supervised learning framework that combines the above strategies to learn comprehensive visual structures for representation learning, which significantly improves performance on the downstream task (e.g. linear classification on the ImageNet-100 has a gain of 6.0% over MoCo v3).

II. Related Work

Contrastive learning. The CL compares two augmented views of the same input to capture semantic-invariant features, such as object-related characteristics. Some works [1], [2], [8]–[10] attract the positive (similar) pairs and repulse negative (different) pairs to learn the representations. Some works [7], [11], [12] also provide a self-distillation framework, which only matches positive samples to learn the representations. Moreover, Zbontar et al. [13] proposes to minimize the redundancy between the vector components of positive pairs' vectors could also achieve good representation learning. However, Hinton et al. [1] has pointed out that CL-based methods neglect some essential information such as colour or object orientation.

Masked image modeling. Since masked language modeling in the NLP field achieves great success, several pioneering works [14], [15] propose Masked Image modeling (MIM). iGPT [14] reconstructs the masked down-sampled images to train the visual encoder, but the down-sample operation loses visual details. BEiT [16] tokenizes the visual patches via a discrete VAE to learn representations by predicting masked tokens, but the discrete operation still loses information. SimMIM [17] proposes to reconstruct the masked image area in the original pixel space, but the encoding needs to input the masked visual token. In contrast, MAE [4] only inputs unmasked visual tokens to the encoding, which greatly reduces the computation. Moreover, AttMask [18] proposes to produce the mask guided by attention maps, but it uses the average activation of different attention maps, which ignores different visual patterns. CMAE [6] combines CL and MAE as multi-task framework to boost the performance. Importantly, all the above MIM methods are inefficient, which require abundant training epochs to learn effective representations.

III. Method

As shown in Figure 1, the CL and MIM in our framework coordinate to supplement each other to learn different levels of visual structure. However, the MIM requires abundant training epochs induced by the random masking strategy. To further improve the efficiency of the MIM training, we find that the attention activation in ViT is beneficial to produce a valuable mask pattern. The details of each component will be introduced in the following.

A. Contrastive learning

First of all, we obtain the two augmented views x_i and $x_i^{'}$ of the input image i. Consistent with the CL framework [10], here we apply an encoder f_θ and a momentum encoder f_ξ to extract their hidden output h_i and $h_i^{'}$. The projectors g_θ, g_ξ

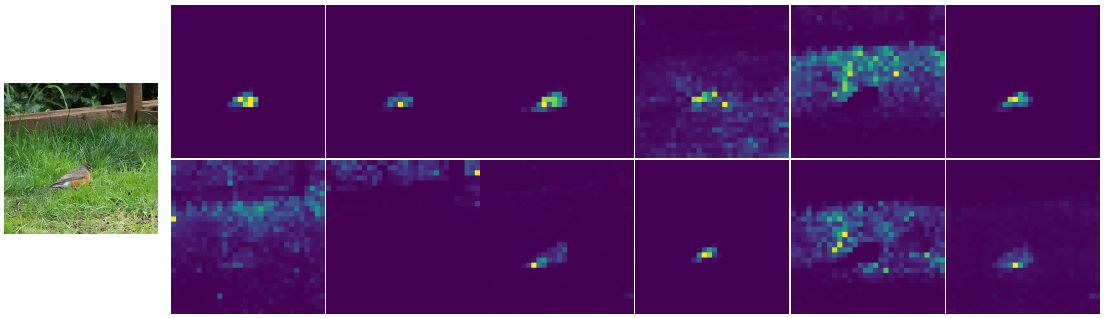


Fig. 2: Visualization of different attention maps of the last block of ViT-B/16 trained by our CAMAE, which contains various visual patterns and could be used to produce effective masks for MIM training.

further cast the hidden outputs to the visual representations z_i and $z_i^{'}$, and the contrastive objective loss is:

$$\mathcal{L}_c(x_i) = -\log \frac{\exp(z_i \cdot z_i'/\tau)}{\exp(z_i \cdot z_i'/\tau) + \sum_{j \neq i} \exp(z_i \cdot z_j'/\tau)}, \quad (1)$$

where z_j' is the representation of other samples in the same batch produced by the momentum projector.

B. Masked image modeling

The masked images are trained by the MIM task upon the same encoder f_θ with an additional decoder module q_θ, and the masked reconstruct loss $\mathcal{L}^m$ of m-th attention space is:

$$\mathcal{L}^m = \sum_i \mathcal{L}^m(x_i) = \sum_h mask_{i,h}^m * (x_{i,h} - \tilde{x}_{i,h})^2, \quad (2)$$

where $mask_{i,h}^m$ is the binary matrix representing whether h-th patch of i-th image should be masked, and the meaning of m will be introduced later. $x_{i,h}$ means the pixel input of the i-th image's h-th patch, $\tilde{x}_{i,h}$ means the reconstructed pixel output corresponding to the i-th image's h-th patch.

The reconstruction loss is calculated by Equation 2, which equals the mean squared error between the reconstructed pixel output and the original pixel input. We note that $mask_{i,h}^m$ will filter the unmasked pixels, and this loss will only calculate the reconstruction error based on masked pixels.

C. Information-density masking strategy

Then, to improve the training efficiency of the MIM, we propose an information-density mask (IDM) strategy to produce semantic-related masks rather than random erasing. In general visual transformer networks, the last block has multiple self-attention maps $\{atten_i^1, .., atten_i^M\}$ as shown in Figure 1. The values of the attention map (patch size $\times$ patch size) represent the masked importance of corresponding visual patches. IDM sorts the visual patches according to the activation values of the attention map and produces the semantic-related mask as shown in Figure 3. If the value is bigger, it is more important to the representations, so our IDM masks out the top 75% values visual patches to train a MIM task.

Equation 3 indicates the relationship between mask and attention maps:

$$mask_{i,h}^m = I(h, atten_i^m), \quad (3)$$

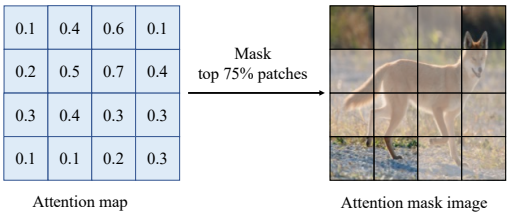


Fig. 3: Schematic illustration of the IDM mask. The real patch size is 16.

where the $I(\cdot)$ represents the IDM module in the original paper, $atten_i^m$ is the i-th image's m-th attention head activation value, and h indicates the index of individual image patches. Specifically, if the h-th image patch activation value over the top 25% total patches of $atten_i^m$, then this patch should be masked.

However, the encoder has multiple attention maps output, and IDM needs to decide which one should be selected to produce the mask. Random sampling is a good choice, but we will provide a more efficient approach in the next section.

D. Pattern-learning balanced sampling

Although previous work [7] proves the attention maps are rich in high-level features, we further discover that different attention maps contain various visual patterns. As shown in Figure 2, the visual modes are highly semantic-related. However, the captured semantic patterns are heavily imbalanced because some visual modes are redundant. Therefore, the sampling strategy should balance the learning progress between different patterns.

Therefore, we propose a patterns-learning balance sample strategy (PLB) to bridge the learning gap. PLB samples well-trained attention patterns with low probability, otherwise samples with high probability. PLB uses the reconstruction loss to represent the learning progress of each attention space. Specifically, PLB will gather statistics of the average loss value during each epoch, it saves the reconstruct loss $\mathcal{L}^m$ related to the m-th attention map. After each training epoch, PLB will update the probability by the saved loss value according to

Equation 4:

$$p^m \leftarrow (1-\lambda) * p^m + \lambda * \frac{exp(\mathcal{L}^m/\epsilon)}{\sum_k^M exp(\mathcal{L}^k/\epsilon)}, \tag{4}$$

where M is the number of attention maps, and λ, ϵ are hyper-parameters (λ=0.99,ϵ=0.2)

At each step, PLB will sample the index m from P to select the attention map to produce mask governed by Equation 5 and 3. During initialization, all attention spaces have the same probability.

$$m \sim P(p^1, ..., p^M). \tag{5}$$

In conclusion, the total training loss is:

$$L = \sum_i (L_c(x_i) + \sum_m L^m(x_i)). \tag{6}$$

IV. Experiments

A. Implementation details

For a fair comparison, our pre-training and linear classification experiments are conducted on the ImageNet-100/1K dataset, following the same protocol as [4].

Architecture. We use the vision transformers (ViT) [15] and ViT-B/16 as the backbone f. For ViTs, /16 denotes its patch size is 16. We pre-train and fine-tune the Transformers with 224-size images, so the total number of patch tokens is 196. The projection head h is a 3-layer MLPs following [10].

Pre-training. We pre-train the ViT-B on the eight V100 GPUs with a batch size of 4096 for 400 epochs. We use Xavier uniform [19] to initialize all Transformer blocks. We use the AdamW optimizer [20] in the pre-training. We use the linear lr scaling rule: lr = base lr × batchsize / 256. The mask ratio is 75%. The τ of contrastive loss is 0.2. The dimension of hidden output of the encoder is 768. The dimension of the output of the projector is 256. The self-distillation setting follows [7] and contrastive learning setting is the same as [10]. The MIM experiments are similar to [4].

Downstream task. (1)Linear probing and fine-tuning: We use a linear classifier on the frozen representations to evaluate the quality of pre-trained features. For this process, we use the same regularization strategies as [4] and set weight decay as zero. Our fine-tuning follows the common practice of supervised ViT training. (2)Semi-supervised learning: We split 1 percent and 10 percent training dataset according to [1], and we follow the experimental setting with [13]. (3)Detection/segmentation [1] are based on the open-source code to evaluate the pre-trained encoder. The image scale is in [640, 800] pixels during training and 800 at inference.

B. Effectiveness of IDM

Previous studies of CL [7] prove that self-supervised features of ViT contain various semantic visual characteristics. In this paper, we investigate whether CL-pretrained ViTs could provide effective masks when applying the IDM method.

[1]https://github.com/bytedance/ibot

TABLE I: The evaluation of MAE when using different IDM methods on the ImageNet-100 dataset. The best results are in **bold**.

Method	IDM	Epoch	Acc.
MAE [4]	-	400	61.3
MAE	DINO [7]	800	33.8
MAE	MoCo [10]	800	35.7
MAE	End-to-end	400	62.5
CAMAE	-	400	83.0
CAMAE	End-to-end	400	**83.2**

The results of Table I show that the IDM cannot use well-trained attention maps, only the end-to-end IDM strategy could work well.

C. Effectiveness of PLB

In Table II we compare different masking strategies: (1) Random: randomly masking the visual patches (2) Attention: randomly sampling attention maps to produce masks. (3) PLB: PLB sampling attention maps to produce masks.

TABLE II: The evaluation of the PLB. Experiments on the linear classification of ImageNet-100. The "Mask" means different mask strategies. "Acc." indicates the top-1 accuracy of linear classification. The best results are in **bold**.

Method	Mask	Epoch	Acc.
CAMAE	Random	400	83.0
	IDM	400	83.2
	IDM+PLB	400	**86.5**

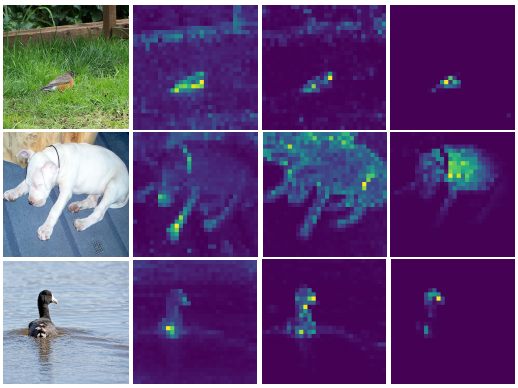
The results show that the PLB strategy significantly improves the performance and efficiency of pre-training with jointing the contrastive pretext task. However, we find the PLB is not helpful for MAE framework. Because the PLB tends to sample the mask patterns that contributed more to the reconstruction loss with high probabilities. We found that the PLB tends to capture noisy attention maps with large reconstruct loss in the MAE framework without CL. Thus CL is necessary for PLB.

D. Visualization

We visualize the object-related attention map produced by different SSL methods as shown in Figure 4. In the first image, there is a bird standing on the grass ground, in the previous CL methods, such as iBOT or MAE, the object-related attention maps always contain the background grass, and the attention map of our method mainly focuses on the bird itself. This result indicates that the attention maps captured by CAMAE decouple the background and object information much better when compared to other methods.

E. Linear evaluation

To verify the effectiveness of our method, we use linear classification to evaluate the quality of learned representations.



(a) Image (b) iBOT (c) MAE (d) Our

Fig. 4: Visualization of attention map compared with different MIM methods.

TABLE III: Pre-experiments on ImageNet-100/1K linear classification. Models based on ViT-B/16. "Lin." indicates the top-1 accuracy of linear evaluation of classification task with ImageNet-100/1K. "Fin." means the fine-tune results. "hours" indicates the number of hours required. The best results are in **bold**.

Method	Epoch	IN-100			IN-1K	
		hours	Lin.	Fin.	Lin.	Fin.
MoCo v3 [10]	300	14	80.2	86.8	76.7	83.2
DINO [7]	300	24	81.8	88.1	78.2	83.6
MAE [4]	800	32	74.4	85.7	72.7	83.6
iBOT [5]	800	60	83.3	89.8	79.4	84.0
AttMask [18]	100	-	-	-	76.1	-
CMAE [6]	800	-	-	-	-	84.4
CAMAE	100	8	81.2	87.6	77.2	81.5
CAMAE	400	29	83.2	89.9	79.3	84.1
CAMAE + PLB	400	29	**86.5**	**90.9**	**80.2**	**84.7**

In Table III, the CAMAE outperforms other self-supervised methods and achieves a new state-of-the-art performance.

Moreover, training epochs and hours in Table III represent the efficiency of pre-training, and we can see that CAMAE does not require too many training epochs, achieving efficient representation learning. In addition, DINO [7] and iBOT [5] use multi-crop technology to boost performance, but it slows down the training process. The experiment results indicate that our proposal significantly improves linear classification performance with fewer training epochs compared to other MIM methods.

F. Detection/segmentation

Here we evaluate the ImageNet-100 pre-trained ViT-B/16 on the detection/segmentation task as shown in Table IV. The results indicate that our approach learns general representations and can be well-transferred to fine-grain visual tasks.

TABLE IV: Instance segmentation and object detection results on COCO. * denotes reproduced results. The best results are in **bold**.

Method	AP^{bb}	AP^{bb}_{50}	AP^{bb}_{75}	AP^{mk}	AP^{mk}_{50}	AP^{mk}_{75}
MoCo v3*	38.1	56.8	41.1	33.6	54.0	35.6
CAMAE	38.0	56.6	40.7	33.5	53.7	35.3
CAMAE+PLB	**38.8**	**57.1**	**41.5**	**34.0**	**54.7**	**36.2**

TABLE V: The top-1 semi-supervised classification accuracy of the ImageNet-100 using 1% and 10% training examples. * denotes reproduced results. The best results are in **bold**.

Method	1% Label	10% Label
MoCo v3* [10]	55.9	77.0
DINO* [7]	57.2	79.6
iBOT* [5]	57.6	80.8
CAMAE	58.4	81.1
CAMAE+PLB	**62.5**	**82.1**

G. Semi-supervised learning

We use 1% and 10% subset of the ImageNet-100 to test the semi-supervised learning performance as shown in Table V.

The results show that our proposal consistently improves the performance of the semi-supervised learning task.

V. Discussion and Conclusions

In this paper, we propose a new masking strategy named IDM to address the training inefficiency of the MIM task. We demonstrate the effectiveness of producing the mask with attention maps, which significantly improves the training efficiency and boosts the downstream performance. To promote consistent pattern learning, we design a pattern-learning balanced sampling strategy. Experimental results show that our framework decouples the object and background in the attention maps much better than previous methods. Thus, our method achieves considerable and consistent gains in downstream performance over the state-of-the-art methods. In summary, the proposed IDM is a versatile, transferable, and low-cost approach to improve representation learning.

References

[1] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton, "A simple framework for contrastive learning of visual representations," in *ICML 2020*, 2020, vol. 119 of *Proceedings of Machine Learning Research*, pp. 1597–1607.

[2] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick, "Momentum contrast for unsupervised visual representation learning," in *CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, 2020, pp. 9726–9735.

[3] Xu Luo, Yuxuan Chen, Liangjian Wen, Lili Pan, and Zenglin Xu, "Boosting few-shot classification with view-learnable contrastive learning," in *2021 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2021, pp. 1–6.

[4] 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.

[5] 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.

[6] 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.

[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*, 2021, pp. 9650–9660.

[8] Xinlei Chen, Haoqi Fan, Ross B. Girshick, and Kaiming He, "Improved baselines with momentum contrastive learning," *CoRR*, vol. abs/2003.04297, 2020.

[9] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey E. Hinton, "Big self-supervised models are strong semi-supervised learners," in *NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

[10] Xinlei Chen, Saining Xie, and Kaiming He, "An empirical study of training self-supervised vision transformers," in *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, 2021, pp. 9620–9629.

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

[12] Xinlei Chen and Kaiming He, "Exploring simple siamese representation learning," in *CVPR 2021, virtual, June 19-25, 2021*, 2021, pp. 15750–15758.

[13] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny, "Barlow twins: Self-supervised learning via redundancy reduction," in *ICML 2021, 18-24 July 2021, Virtual Event*, Marina Meila and Tong Zhang, Eds., 2021, vol. 139 of *Proceedings of Machine Learning Research*, pp. 12310–12320.

[14] Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever, "Generative pretraining from pixels," in *International Conference on Machine Learning*. PMLR, 2020, pp. 1691–1703.

[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 *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*, 2021.

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

[17] 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.

[18] Ioannis Kakogeorgiou, Spyros Gidaris, Bill Psomas, Yannis Avrithis, Andrei Bursuc, Konstantinos Karantzalos, and Nikos Komodakis, "What to hide from your students: Attention-guided masked image modeling," in *ECCV 2022*. Springer, 2022, pp. 300–318.

[19] Y. Bengio and X. Glorot, "Understanding the difficulty of training deep feed forward neural networks," *Proc. AISTATS, 2010*, 2010.

[20] Ilya Loshchilov and Frank Hutter, "Decoupled weight decay regularization," in *International Conference on Learning Representations*, 2018.