Retaining Diverse Information in Contrastive Learning through Multiple Projectors

He Zhu, Shan Yu

***Abstract*—*Contrastive Learning (CL)* achieves great success in learning visual representations by comparing two augmented views of the same images. However, this very design removes transformation-dependent visual information from the pre-training, which leads to incomplete representations and is harmful for downstream tasks. It's still an open question to retain such information in the CL pre-training process. In this paper, we propose a *Multi-Projector Contrastive Learning (MPCL)* to address this issue, which produces multi-view contrastive candidates to retain more comprehensive visual characteristics. In addition, we introduce a contrast regularization to construct multiple projectors as different as possible, thereby facilitating the diversity of preserved information. Finally, to promote a consistent learning process for multi-projector, we design a projector training balance strategy to adjust the learning preference of different projectors. MPCL can be applied to various CL frameworks to effectively protect visual characteristics. Experimental results show that the method performs well on subsequent tasks such as linear and semi-supervised image classification, object detection, and semantic segmentation. Importantly, the visual transformer trained by MPCL improves 2% absolute points of linear evaluation beyond the MoCo-v3 on the ImageNet-100 dataset.**

***Index Terms*—multiple projectors contrastive learning, projector contrastive regularization, projector mining**

I. Introduction

CONTRASTIVE learning [1], [2], [3], [4] has achieved great success in the field of self-supervised learning. The common motivation behind CL frameworks is the InfoMax principle [5], which guides the network to maximize common features between two transformed views of the same images. Although CL can perform better than supervised learning in various downstream tasks, there are intrinsic limitations of this framework that have not been addressed adequately. For instance, Hinton et al. [1] point out that optimizing the contrastive loss may lead the network to remove the basic visual characteristics, which can be harmful for learning generalizable representations of images. How to preserve these characteristics better in CL remains unclear.

Previous studies have proposed several approaches to address this issue. For example, [6] brings up the alignment loss to preserve the relations of image pairs but it inevitably impairs the uniformity of embedding spaces. After that, [7] proposes a hardness-aware loss that enhances alignment while protecting uniformity, but it only retains the pair-related features rather than the comprehensive characteristics and has achieved only limited success in empirical evaluation. Meanwhile, [1] and [8] identify the projector, which is widely used in CL to compress the representations extracted from the encoder network to a lower-dimensional projection, can retain visual information. However, performance on downstream tasks hardly changes with different choices of a single projector [1]. We conjecture that visual features with complex structures may be difficult to be fully retained by a single projector. Therefore, to address this issue, here we propose a multi-projector design to retain more comprehensive characteristics.

To verify our hypothesis, a toy experiment of MPCL is conducted based on ImageNet-100. MPCL has multiple projectors which share the same backbone, as shown in Figure.1. The evaluation of this pre-trained model is carried out according to previously published alignment loss [6] and linear classification. The alignment loss measures the distance of positive pairs. Importantly, Table I shows that the combination of projectors with the different structures has a higher similarity of positive pairs which indicates more underlying visual relationships are retained in the learned representations, and it yields better downstream performance than the multiple projectors with the same structure or the single projector.

TABLE I
METRICS OF PRE-TRAINED MODELS WITH DIFFERENT SETTINGS. MODELS ARE PRE-TRAINED ON IMAGENET-100 BASED ON RESNET-50 (R50). IN THIS EXPERIMENT, "SINGLE" MEANS SINGLE PROJECTOR, AND "MULTIPLE" REPRESENTS USING FOUR PROJECTORS. "BASELINE" CORRESPONDS TO THE ORIGINAL MOCO V2 FRAMEWORK. THE "SAME"/"DIFFERENT" INDICATES THAT THE HIDDEN DIMENSION OF EACH PROJECTOR IS THE SAME/DIFFERENT. THE VALUE OF $1-\mathcal{L}_{Ali}$ MEANS THE SIMILARITY OF POSITIVE PAIRS. THE VALUE OF $-\mathcal{L}_{Uni}$ IS A UNIFORMITY METRIC AS [6]. "TOP-1" INDICATES TOP-1 ACCURACY OF LINEAR EVALUATION OF DOWNSTREAM CLASSIFICATION TASK WITH IMAGENET-100.

Method	Projector	$1-\mathcal{L}_{Ali}$	$-\mathcal{L}_{Uni}$	Top-1
Baseline	Single	0.66	2.00	77.54
Same	Multiple	0.66	2.00	77.32
Different	Multiple	0.79	2.03	**78.12**

In addition, a combination of projectors with the same structure does not outperform a single one, which indicates

This work was supported in part by the National Key Research and Development Program of China under Grant 2017YFA0105203, the Strategic Priority Research Program of the Chinese Academy of Sciences (CAS) under Grant XDB32040200, the International Partnership Program of CAS under Grant 173211KYSB20200021, and the Beijing Academy of Artificial Intelligence (BAAI, to S.Y.)

He Zhu is with the Brainnetome Center, National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences(CASIA), Beijing. School of Future Technology, University of Chinese Academy of Sciences(UCAS), Beijing. (e-mail: he.zhu@nlpr.ia.ac.cn)

Shan Yu is with Brainnetome Center, National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences(CASIA), Beijing. School of Future Technology, University of Chinese Academy of Sciences(UCAS), Beijing. CAS Center for Excellence in Brain Science and Intelligence Technology(CEBSIT),Beijing. (e-mail: shan.yu@nlpr.ia.ac.cn)

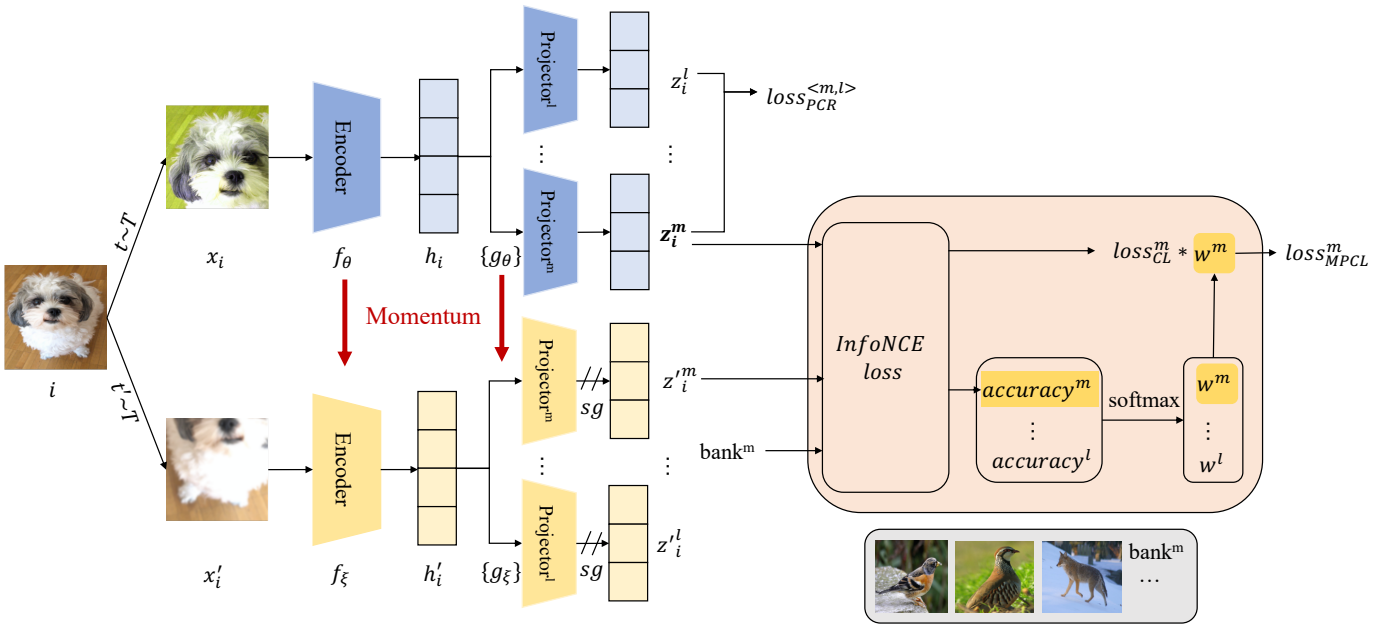

Fig. 1. The pipeline of the proposed multi-projector contrastive learning framework. $T_i/T_i^{'}$ are the random data augmentation operators sampled from the same family of augmentations. m,l indexes the different projector numbers. $*_\theta$ are the parameters of the network, and parameters $*_\xi$ are momentum updated by $*_\theta$. sg means stop-gradient. Each projector has its contrastive loss and accuracy, the loss weight of different projectors is updated according to the average accuracy after the last epoch training, this re-weight contrastive loss L^m_{MPCL} is shown in the figure. Projector contrastive regularization(PCR) $L^{m,l}_{PCR}$ are calculated pairwise to reduce the similarity among arbitrary two projectors. The $bank^m$ stores the m-th projector negative samples. z^m_i (red colour) is the representation produced by the m-th head, the PCR loss and MPCL loss are calculated respectively as shown in the figure.

that a naive design of multiple projectors cannot automatically retain diverse information. Given this, we propose a new regularization to facilitate projector diversity. Moreover, we design a strategy to balance the training processes of different projectors.

Our main contributions are thus summarized as follows:

- We hypothesized that a single projector is difficult to retain full visual information and propose a multi-projector contrastive learning framework to better retain complex visual feature structure.
- We propose a contrastive regularization term to facilitate the diversity of projectors to preserve more information.
- We design a training strategy to enable the network automatically adjusts its learning preference of different projectors, achieving a balanced learning process for all projectors.

II. Related Work

Contrastive learning. Chen et al. [1] and He et al. [9] pioneered the original CL framework almost simultaneously, which has become very popular recently, with the main idea of learning useful representations through attracting positive pairs and repulsing negative pairs. Chen et al. [4], [2] improved the baseline by applying powerful data augmentation and fine-tuning projector settings. Furthermore, to reduce the CL extension module, previous studies proposed that the network only attracts positive pairs to achieve representation learning [10], [11], [3], [12]. These subsequent variants of the CL framework have greatly simplified its use, with the importance of the projector consistently verified. Although careful handling of projector settings is suggested, no previous study has looked into using multiple projectors to retain more comprehensive information.

Understanding CL. To understand the mechanism underlying the strength of CL, lots of studies explore the characteristics of contrastive features. Recently, lots of experimental studies [7], [13], [14], [15], [6], [16], [17], [18] find that similar representation pairs may come from highly different images, which means that the system learned to ignore such differences and thereby impairing the quality of learned representations. Several plain strategies are provided to reduce the influence of the wrong contrastive candidates through redesigning data sampling methods or the loss functions, such as negative sampling [13], [14], data augmentation [15], mixing samples [19], supervision [2] or distribution loss function [6]. But these studies do not recognize the underlying mechanisms causing the loss of useful visual characteristics.

Multiple heads representation learning. Multiple heads representation learning is usually used for multi-task learning [20], such as classification and detection heads in the detector [21], [22]. In addition, multi-head modules are treated as an attention mechanism to generate redundancy latent sub-spaces to capture adequate relationships [23], [24], [25], which significantly benefits natural language processing. The primary concepts of previous applications involve collaborative representation learning [26]. Wu et al. [27] and Gu et al. [28] propose that decoupling output for different tasks benefits backbone representation learning, but they just

split the output without any other constraints. To the best of our knowledge, few studies have explored the advantages of encouraging diversity in multiple heads representation learning.

III. METHODS

A. MPCL approach

In this section, we introduce a multiple-projector contrastive learning framework:

Given an unlabeled training set $X = \{x_1, x_2, \ldots, x_N\}$, for the framework with M projectors in MPCL, the CL loss of the m-th projector is:

$$\mathcal{L}^m(x_i) = -\log \frac{\exp(s^m_{i,i}/\tau)}{\exp(s^m_{i,i}/\tau) + \sum\limits_{k\neq i} \exp(s^m_{i,k}/\tau)} \tag{1}$$

where $s^m_{i,j} = g^m_\theta(f_\theta(x_i))^T g^m_\xi(f_\xi(x_j))$. $f_\theta(\cdot), f_\xi(\cdot)$ is the shared feature/momentum extractor that maps the images from pixel space to embedding space [1]; $g^m_\theta(\cdot), g^m_\xi(\cdot)$ is corresponding to the m-th projector cascaded behind the feature/momentum extractor and τ is a temperature hyper-parameter.

B. Projector contrastive regularization (PCR)

Multiple projectors with different structures cannot guarantee to maximize the diversity of retained information, and different structures are cumbersome for implementation. Thus, to address this issue, we propose the projector contrastive regularization (PCR) approach: for each representation, its projection from one projector needs to be different from the outputs on other projectors as Eq.2.

$$\mathcal{L}^{B,m}_{PCR} = \sum_b^B \frac{1}{M-1} \sum_{l\neq m}^M \frac{\langle z^m_b, z^l_b \rangle_d}{\|z^m_b\|_2 \left\|z^l_b\right\|_2} \tag{2}$$

where m, l are the indices of the projectors, z^m, z^l are the outputs of the different projectors, B is the batch size, b is the index of samples in the batch, and d is the index of the vector components of the output.

Note, $Z_{m,l} = \frac{\langle z^m_b, z^l_b \rangle_d}{\|z^m\|_2 \|z^l\|_2}$ is the cross-correlation matrix computed between the outputs of the two different projectors.

The PCR loss maximizes the variability of the representations learned by different projectors. It relies on batch statistics to measure this variability.

C. Projector training balance (PTB)

Multiple projectors have variable training progress due to different learning difficulties that they face. To balance the training process, the network needs to adjust different learning preferences, which means more attention should be paid to the projector with poor contrastive performance.

A new projector mining strategy enables the network to adjust its learning preference. Specifically, each projector has its loss and accuracy of the contrastive task, the network could assign the weight of different projectors according to their last epoch's accuracy by:

$$w^m = \frac{exp((1-acc^m)/\epsilon)}{\sum_k^M exp((1-acc^k)/\epsilon)} \tag{3}$$

where acc^m represents the one epoch mean accuracy of m-th projector's contrastive task, ϵ is a hyper-parameter ($\epsilon = 0.5$).

Therefore, the total loss is :

$$\mathcal{L} = \frac{1}{M} \sum_n^N \sum_m^M (\sum_i^{b_n} w^m \mathcal{L}^m(x_i) + \lambda \mathcal{L}^{b_n,m}_{PCR}) \tag{4}$$

where M is the total projector number, N is the total number of training batches, b_n is n-th batch data, and λ is a hyper-parameter (λ=0.01), and m indexes the projector numbers.

IV. EXPERIMENTS

A. Implementation details

For the ImageNet-100/ImageNet experiments, our method is based on the official MoCo code. We use eight 3080-ti GPUs for training and the batch size is 128/256.

Detection/segmentation[1] and semi-supervised classification[2] 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. Our method and MoCo use the same hyper-parameters as the ImageNet-supervised counterpart (i.e., we do not perform any method-specific tuning).

B. PCR effectiveness

In this section, we provide the experimental evidence to verify the effectiveness of contrastive regularization (based on the ImageNet-100 linear classification experiments), as shown in Table II. To this end, several widely used techniques are compared, including initialization, and dropout.

TABLE II
DIFFERENT METHODS AIM AT PROMOTING THE DIVERSITY OF DIFFERENT PROJECTORS. EVALUATION OF IMAGENET-100 LINEAR CLASSIFICATION. MODELS BASED ON RESNET-50 WITH FOUR PROJECTORS. 'MIX INIT.' INDICATES WHETHER THE MULTIPLE PROJECTORS USED MIXED INITIALIZATION. 'DROPOUT' INDICATES PROJECTORS WITH DROPOUT. 'ORTH.' MEANS WEIGHTS OF MPHS WERE CONSTRAINED TO BE ORTHOGONAL. "TOP-1" INDICATES THE TOP-1 ACCURACY OF LINEAR EVALUATION OF DOWNSTREAM CLASSIFICATION TASK WITH IMAGENET-100. THE BEST RESULTS ARE IN **BOLD**.

Framework	Method	Arch.	Epoch	Top-1
MoCo v2	-	R50	200	77.50
MPH	-	R50	200	77.32
MPH	Dropout	R50	200	76.31
MPH	Mix Init.	R50	200	77.90
MPH	Orth.	R50	200	77.70
MPH	PCR	R50	200	**78.70**

In the 'mix init.' experiments, different projectors use different initialization methods to build diverse projectors, i.e., namely Kaiming [29] and Xavier[30]. In the 'dropout' experiments, dropout is introduced to the projectors [31] to generate diverse features.

The results show that the PCR method retains more diverse features and achieves better representation learning.

[1]https://github.com/facebookresearch/moco
[2]https://github.com/facebookresearch/barlowtwins

C. Calculation costs

The use of MPHs introduces additional computational costs during the pre-training process. Here we evaluate the computational cost of the multi-projector and optimize the projector number settings. In the evaluation of projector numbers on linear classification as shown in Table III. Additional projectors benefit the performance, which indicates diverse projectors can retain more comprehensive information. However, we find that too many projectors can impair the performance although PTB improves, which is probably due to increased difficulty in training a highly heterogeneous. Thus, in the following experiments, four projectors are used as the standard setting.

TABLE III
PRE-EXPERIMENTS ON IMAGENET-100 LINEAR CLASSIFICATION. MODELS WERE BASED ON RESNET-50. "NUM." IS THE NUMBER OF PROJECTORS. "TOP-1" INDICATES THE TOP-1 ACCURACY. "+PTB" MEANS TRAINING WITH PTB. THE BEST RESULTS ARE IN **BOLD**.

Method	Num.	Params.	Epoch	Top-1	+PTB
MPH + PCR	1	27.97	200	77.50	-
	2	28.23	200	77.52	77.73
	4	28.75	200	**78.70**	**79.18**
	8	29.80	200	77.83	78.82

D. Downstream tasks

1) Linear classification: Table IV demonstrates the same behaviors of the linear classification as shown in the toy experiment. Previous solutions achieve little success in linear probing, but our approach shows significant improvement.

TABLE IV
LINEAR CLASSIFICATION EVALUATION ON IMAGENET. TOP-1 CENTER-CROP ACCURACY OF FULLY CONNECTED CLASSIFIERS FOR IMAGENET IS REPORTED. * DENOTES REPRODUCED RESULTS. THE BEST RESULTS ARE IN **BOLD**.

Method	Epochs	Top 1
MoCo v2* [4]	200	67.50
Align* [6]	200	67.69
Hard Sampling* [7]	200	67.55
MoCHi*[19]	200	67.60
ContrastiveCrop* [18]	200	67.80
MPH + PCR	200	68.22
MPH + PCR + PTB	200	**68.54**

2) Semi-supervised classification: We fine-tune the pre-trained ResNet-50 on a subset of the ImageNet dataset using our method. We use 1% and 10% subsets according to SimCLR [1]. Table V shows that our approach outperforms competing methods in the semi-supervised learning task.

3) Detection and segmentation: Following previous research [9], we use Mask R-CNN [33] with a C4 backbone, with batch normalization tuned and synchronized across GPUs. Table VI shows the object detection and semantic segmentation results for the COCO dataset [32], which indicates that our proposal also has a better transferability on various visual downstream tasks.

4) Self-supervised Learning Frameworks: To evaluate the generalization performance of the MPCL, here we adapt it to Barlow Twins (ResNet) [12] and MoCo v3 [25] (Visual Transformer [34]). As shown in Table VII, the improvement of PCR is less affected by the batch size, and the results show that our method improves the performance of the advanced CL frameworks and consistently has a positive effect on various networks.

TABLE V
SEMI-SUPERVISED LEARNING ON IMAGENET USING 1% AND 10% TRAINING EXAMPLES. * DENOTES REPRODUCED RESULTS. THE BEST RESULTS ARE IN **BOLD**.

Method	1% Label		10% Label	
	Top-1	Top5	Top-1	Top5
Supervised [32]	25.40	48.40	56.40	80.40
MoCo v2* [4]	43.25	72.68	63.52	86.21
MoCHi* [19]	43.42	72.02	63.45	86.12
MPH+PCR	43.84	73.39	64.43	87.10
MPH+PCR+PTB	**44.12**	**73.61**	**64.75**	**87.30**

TABLE VI
INSTANCE SEGMENTATION AND OBJECT DETECTION RESULTS ON COCO WITH THE ×1 TRAINING SCHEDULE AND A C4 BACKBONE. * DENOTES REPRODUCED RESULTS. THE BEST RESULTS ARE IN **BOLD**.

Method	AP^{mk}	AP^{mk}_{50}	AP^{mk}_{75}	AP^{bb}	AP^{bb}_{50}	AP^{bb}_{75}
MoCo v2* [4]	34.2	55.4	36.2	39.0	58.6	41.9
MoCHi*[19]	34.4	55.6	36.7	39.2	58.8	42.4
CCrop* [18]	34.5	55.5	36.4	39.2	58.8	42.2
MPH+PCR	34.2	55.9	36.6	39.6	59.3	42.9
MPH+PCR+PTB	**34.6**	**56.2**	**36.9**	**40.0**	**59.5**	**43.1**

TABLE VII
EXPERIMENTS OF IMAGENET-100 LINEAR CLASSIFICATION. MODELS WERE BASED ON RESNET-50/VIT-BASE. "SAME" AND "DIFF." INDICATE THAT THE HIDDEN DIMENSION OF MPHS IS SAME/DIFFERENT, RESPECTIVELY. "TOP-1" INDICATES THE TOP-1 ACCURACY OF LINEAR EVALUATION OF DOWNSTREAM CLASSIFICATION TASK WITH IMAGENET-100. THE BEST RESULTS ARE IN **BOLD**.

Method	Arch.	Epochs	Batch	Top 1
Barlow Twins[12]	Res50	300	1024	79.86
MPH + PCR	Res50	300	1024	80.36
MPH + PCR + PBT	Res50	300	1024	80.58
iBOT [35]	ViT-B	300	512	81.30
MAE [36]	ViT-B	300	2048	74.40
MoCo v3 [25]	ViT-B	300	2048	79.60
MoCo v3 [25]	ViT-B	300	4096	80.08
MPH(Same)	ViT-B	300	4096	79.74
MPH(Diff.)	ViT-B	300	4096	80.81
MPH+PCR	ViT-B	300	4096	81.68
MPH+PCR+PTB	ViT-B	300	2048	81.68
MPH+PCR+PTB	ViT-B	300	4096	**82.02**

V. DISCUSSION AND CONCLUSIONS

In this paper, we propose a multi-projector contrastive learning approach to address the information loss problem in CL. For the first time, we demonstrate that projectors of different structures can jointly retain more features of the image, and explore the encouragement of diverse preservation of projectors of the same structure by PCR. Last but not least, to promote a consistent learning process, we devise a projector training balance strategy. Experimental results show that MPCL minimizes information loss by preserving distinct features and provides considerable gains over state-of-the-art methods, which consistently positively impact transfer learning performance. MPCL is versatile, transferable, and low-cost among other approaches to improve representation learning.

REFERENCES

[1] T. Chen, S. Kornblith, M. Norouzi, G. E. Hinton, A simple framework for contrastive learning of visual representations, in: Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event, Vol. 119 of Proceedings of Machine Learning Research, pp. 1597–1607.

[2] T. Chen, S. Kornblith, K. Swersky, M. Norouzi, G. E. Hinton, Big self-supervised models are strong semi-supervised learners, in: H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, H. Lin (Eds.), Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual, 2020.

[3] X. Chen, K. He, Exploring simple siamese representation learning, in: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021, pp. 15750–15758.

[4] X. Chen, H. Fan, R. B. Girshick, K. He, Improved baselines with momentum contrastive learning, CoRR abs/2003.04297.

[5] R. Linsker, Self-organization in a perceptual network, Computer 21 (3) (1988) 105–117.

[6] T. Wang, P. Isola, Understanding contrastive representation learning through alignment and uniformity on the hypersphere, in: Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event, Vol. 119 of Proceedings of Machine Learning Research, pp. 9929–9939.

[7] F. Wang, H. Liu, Understanding the behaviour of contrastive loss, in: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021, pp. 2495–2504.

[8] Y. Wang, S. Tang, F. Zhu, L. Bai, R. Zhao, D. Qi, W. Ouyang, Revisiting the transferability of supervised pretraining: an mlp perspective, arXiv preprint arXiv:2112.00496.

[9] K. He, H. Fan, Y. Wu, S. Xie, R. B. Girshick, Momentum contrast for unsupervised visual representation learning, in: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020, pp. 9726–9735.

[10] J. Grill, F. Strub, F. Altché, C. Tallec, P. H. Richemond, E. Buchatskaya, C. Doersch, B. Á. Pires, Z. Guo, M. G. Azar, B. Piot, K. Kavukcuoglu, R. Munos, M. Valko, Bootstrap your own latent - A new approach to self-supervised learning, in: H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, H. Lin (Eds.), Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual, 2020.

[11] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, A. Joulin, Unsupervised learning of visual features by contrasting cluster assignments, in: H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, H. Lin (Eds.), Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual, 2020.

[12] J. Zbontar, L. Jing, I. Misra, Y. LeCun, S. Deny, Barlow twins: Self-supervised learning via redundancy reduction, in: M. Meila, T. Zhang (Eds.), Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event, Vol. 139 of Proceedings of Machine Learning Research, pp. 12310–12320.

[13] J. D. Robinson, C. Chuang, S. Sra, S. Jegelka, Contrastive learning with hard negative samples, in: 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021.

[14] C. Chuang, J. Robinson, Y. Lin, A. Torralba, S. Jegelka, Debiased contrastive learning, in: H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, H. Lin (Eds.), Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual, 2020.

[15] S. Purushwalkam, A. Gupta, Demystifying contrastive self-supervised learning: Invariances, augmentations and dataset biases, in: H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, H. Lin (Eds.), Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual, 2020.

[16] T. Xiao, X. Wang, A. A. Efros, T. Darrell, What should not be contrastive in contrastive learning, in: 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021.

[17] F. Wang, H. Liu, D. Guo, F. Sun, Unsupervised representation learning by invariance propagation, in: H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, H. Lin (Eds.), Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual, 2020.

[18] X. Peng, K. Wang, Z. Zhu, Y. You, Crafting better contrastive views for siamese representation learning, CoRR abs/2202.03278.

[19] Y. Kalantidis, M. B. Sariyildiz, N. Pion, P. Weinzaepfel, D. Larlus, Hard negative mixing for contrastive learning, in: H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, H. Lin (Eds.), Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual, 2020.

[20] O. Sener, V. Koltun, Multi-task learning as multi-objective optimization, in: S. Bengio, H. M. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, R. Garnett (Eds.), Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada, 2018, pp. 525–536.

[21] S. Ren, K. He, R. B. Girshick, J. Sun, Faster R-CNN: towards real-time object detection with region proposal networks, IEEE Trans. Pattern Anal. Mach. Intell. 39 (6) (2017) 1137–1149. doi:10.1109/TPAMI.2016.2577031.
URL https://doi.org/10.1109/TPAMI.2016.2577031

[22] Z. Cai, N. Vasconcelos, Cascade R-CNN: delving into high quality object detection, in: 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018, pp. 6154–6162.

[23] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, in: I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA, 2017, pp. 5998–6008.

[24] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, in: J. Burstein, C. Doran, T. Solorio (Eds.), Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers), pp. 4171–4186.

[25] X. Chen, S. Xie, K. 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, pp. 9620–9629.

[26] M. K. Ebrahimpour, G. Qian, A. Beach, Multi-head deep metric learning using global and local representations, in: IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2022, Waikoloa, HI, USA, January 3-8, 2022, pp. 1340–1349.

[27] Y. Wu, Y. Chen, L. Yuan, Z. Liu, L. Wang, H. Li, Y. Fu, Rethinking classification and localization for object detection, in: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020, pp. 10183–10192.

[28] Z. Ge, S. Liu, F. Wang, Z. Li, J. Sun, YOLOX: exceeding YOLO series in 2021, CoRR abs/2107.08430.

[29] K. He, X. Zhang, S. Ren, J. Sun, Delving deep into rectifiers: Surpassing human-level performance on imagenet classification, in: 2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015, pp. 1026–1034.

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

[31] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, Dropout: a simple way to prevent neural networks from overfitting, J. Mach. Learn. Res. 15 (1) (2014) 1929–1958.
URL http://dl.acm.org/citation.cfm?id=2670313

[32] L. Beyer, X. Zhai, A. Oliver, A. Kolesnikov, S4L: self-supervised semi-supervised learning, in: 2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019, pp. 1476–1485.

[33] K. He, G. Gkioxari, P. Dollár, R. B. Girshick, Mask R-CNN, in: IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017, pp. 2980–2988.

[34] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. 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.

[35] J. Zhou, C. Wei, H. Wang, W. Shen, C. Xie, A. Yuille, T. Kong, ibot: Image bert pre-training with online tokenizer, arXiv preprint arXiv:2111.07832.

[36] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, R. Girshick, Masked autoencoders are scalable vision learners, arXiv preprint arXiv:2111.06377.