Supplementary Material for Self-Supervised Representation Learning with Cross-Context Learning between Global and Hypercolumn Features

Zheng Gao Chen Feng Ioannis Patras

Queen Mary University of London, Mile End Road, London, E1 4NS
{z.gao, chen.feng, i.patras}@qmul.ac.uk

1. Discussion with works related to intermediate features

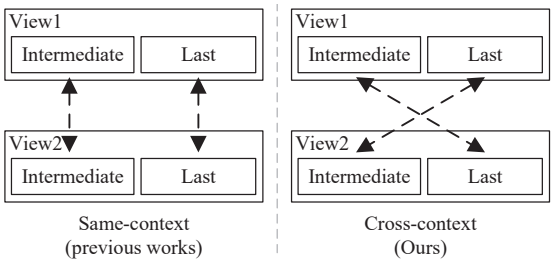

Figure 1. Comparison of our method with previous works.

While some existing works have already explored multi-level self-supervisions (intermediate features) in self-supervised learning (SSL) [6, 12, 16, 17], our method is fundamentally different from these works in the following aspects:

- **Our goal is to present a novel framework to alleviate the "class collision problem" in contrastive learning [2, 13]**, which is different from the goal of previous works where they aim to apply self-supervised objective over intermediate levels.
- **Previous works commonly adopt a straightforward way to extend the self-supervised objective over the last global features to multi-level learning on the intermediate features** by inducing both intermediate features and global features to the self-supervised objective simultaneously [6, 16, 17]. In practice, most of them use the intermediate features from the teacher to supervise the corresponding features from the student (same-context in Fig. 1), which is the application of knowledge distillation in SSL. **In contrast, we propose a cross-layer learning strategy where intermediate features and global features are used as each other's supervisory signal (cross-contest in Fig. 1)**. The superiority of cross-context over same-context is shown in ablation (Tab. 5 in the main paper). **Moreover, we outperform OBoW [6], which also adopts the same-context strategy (Tab. 1, 2 in the main paper, Tab. 2 in the supplementary material)**.
- Another work [12] encourages the intermediate representations to learn from the last layer via the contrastive loss, which is still different from our cross-context (cross-layer) learning. Besides, our objective measures the instance relations with cross-entropy loss to alleviate the "class collision problem" while work [12] fails to do so as it is still based on contrastive objective.
- **Therefore, compared with current works, we have a different goal and to achieve that goal we adopt a different way of leveraging intermediate features for producing better supervisory signal**.

2. Additional experiment results

Table 1. **Results of IN-1K linear classification using hypercolumn**. **hyper** is the result using hypercolumn as the input to the linear classifier.

Method	IN-1K Acc.
MoCo-v2	67.5
ReSSL	69.3
CGH	70.5
CGH (hyper)	**70.8**

2.1. IN-1K classification using hypercolumn

We investigate the effectiveness of hypercolumn by using it for linear classification in Tab. 1. CGH with hypercolumn outperforms its counterpart that directly uses representation vectors after global average pooling for classi-

Table 2. **Transfer learning on COCO object detection and instance segmentation using ResNet-50 pre-trained on IN-1K**. We report the bounding-box AP (AP^{bb}) for object detection and mask AP (AP^{mk}) for instance segmentation. †: our reproduction using the official codes. $*$: results cited from [5].

Method	Epochs	COCO Det.			COCO Instance Seg.		
		AP^{bb}	AP^{bb}_{50}	AP^{bb}_{75}	AP^{mk}	AP^{mk}_{50}	AP^{mk}_{75}
Asymmetric loss.							
MoCo-v2 [4]	200	38.8	58.0	42.0	34.0	55.2	36.3
OBoW [6]†	200	38.6	58.0	41.8	33.8	54.8	36.2
ReSSL [18]†	200	38.3	57.7	41.3	33.4	54.7	35.3
CGH	200	**39.0**	**58.8**	**42.2**	**34.2**	**55.3**	**36.5**
Symmetric loss. 2× FLOPS							
SimCLR [3]$*$	200	37.9	57.7	40.9	33.3	54.6	35.3
SwAV [1]$*$	200	37.6	57.6	40.3	33.1	54.2	35.1
SimSiam [5]$*$	200	37.9	57.5	40.9	33.2	54.2	35.2
BYOL [7]$*$	200	37.9	57.8	40.9	33.2	54.3	35.0
LEWEL [11]	200	38.5	58.9	41.2	33.7	55.5	35.5
Multi-crop							
CGH (Multi)	200	**39.3**	**59.3**	**42.7**	**34.4**	**55.9**	**36.6**

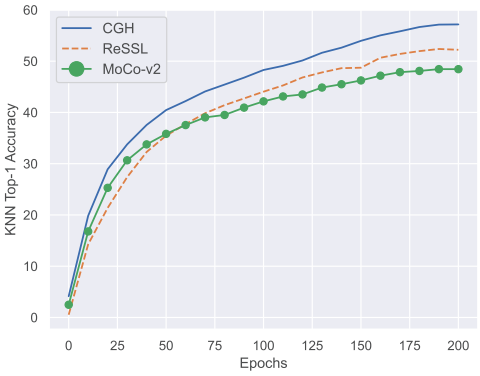

Figure 2. KNN validation accuracy with respect to epochs during pre-training.

fication, which indicates the hypercolumn provides better supervisory signal.

Table 3. Comparison of pre-training running time relative to supervised training.

Method	Time/ Epoch	Linear Acc.	VOC 07+12 Det.
Supervised	1.00	76.5	81.3
MoCo-v2 [4]	1.62	67.5	82.4
ReSSL [18]	1.62	69.3	82.2
BYOL [7]	2.90	70.6	81.4
CGH	2.01	70.5	82.6

2.2. Visualization of training progress

Following [15], we present the KNN classification accuracy with respect to epoch number in Fig. 2, which is a useful metric to monitor the training progress. The KNN classifier is evaluated on the validation set of IN-1K. The KNN accuracy plot shows that the proposed method achieves a steady and consistent improvement. Note that in Fig. 2 we perform the KNN classification using the embedding from the MLP head as in [15]. However, in KNN evaluation, we build the KNN classifier on top of the global average pooling layer of ResNet by following [8].

2.3. COCO object detection and instance segmentation

For COCO object detection and instance segmentation, we fine-tune the Mask R-CNN [9] with ResNet-50-C4 backbone using the model pre-trained on IN-1K. Following [4, 18], we adopt the 1x schedule used in the detetron2 [14], which fine-tunes the model for $90,000$ iterations. The results on COCO are reported in Tab. 2. CGH outperforms ReSSL on all tasks, which demonstrates the effectiveness of the learned representations. Moreover, our method achieves better performance than 2x backprop methods like SimCLR, SwAV, SimSiam and BYOL and competitive results with SOTA methods like MoCo-v2 [4] and LEWEL [11].

3. Visualization of feature representations

We use t-SNE [10] to visualize the learned representation on the training set of Tiny-ImageNet. The first 20 classes of

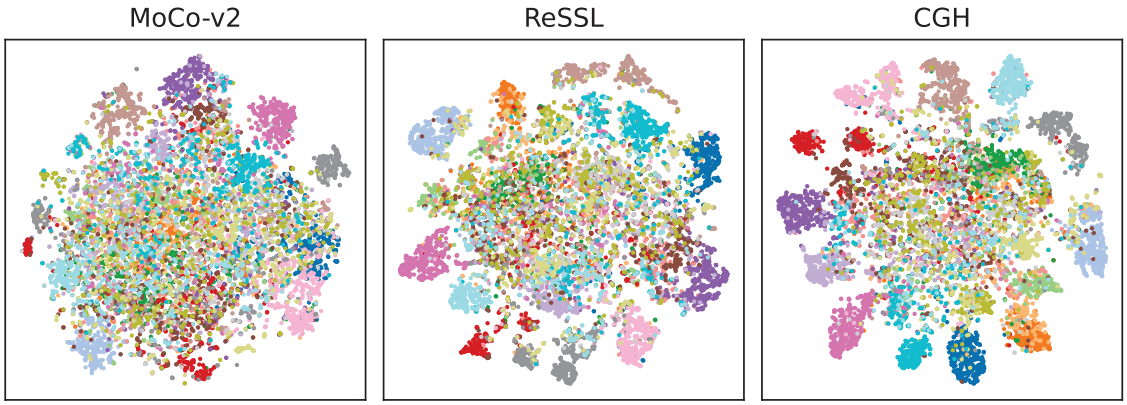

Figure 3. The t-SNE visualization on the training set of Tiny-ImageNet for the first 20 classes. The classes are represented by different colours.

Tiny-ImageNet are selected for the visualization. We report the comparison results of three methods, i.e., MoCo-v2, ReSSL and the proposed CGH in Fig. 3. As shown in Fig. 3, the proposed CGH has better class separation than MoCo-v2 and ReSSL. The t-SNE visualization results demonstrate that the proposed CGH can produce more discriminative representations, which benefit the performance on various downstream tasks.

4. Training cost analysis

In this section, we compare the training cost of our method with the baselines in Tab. 3. For all methods, we perform the pre-training on IN-1K for 200 epochs with ResNet-50 backbone using 2 NVIDIA A100 GPUs. We measure the time consumption relative to supervised IN-1K training ("Supervised") based on the running time of one training epoch ("Time/Epoch"). Note that BYOL uses a batch size of 4096 to achieve the reported performance while we report the training cost using a batch size of 256 due to limited GPU memory. The results show that CGH outperforms ReSSL by 1.2% and 0.4% on IN-1K linear classification and PASCAL VOC object detection with reasonable cost increase (2.01 vs. 1.62). Moreover, compared with 2x backprop methods like BYOL, the proposed method achieves 1.2% improvement on detection and similar performance on classification (70.5 vs. 70.6) with much less training cost (2.01 vs. 2.90).

5. Negative societal impact

Generally self-supervised learning needs to pre-train with multiple GPUs for a long time to achieve competitive results with supervised learning. Our method also has such limitation. However, our method has better performance than SOTA self-supervised learning methods with similar (or shorter) training time, *e.g.*, our CGH (1x backprop) achieves compatible performance with BYOL (2x backprop method with longer training time) on classification and object detection (Tab. 3).

References

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

[2] Mayee Chen, Daniel Y Fu, Avanika Narayan, Michael Zhang, Zhao Song, Kayvon Fatahalian, and Christopher Re. Perfectly balanced: Improving transfer and robustness of supervised contrastive learning. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 3090–3122. PMLR, 17–23 Jul 2022. 1

[3] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607. PMLR, 13–18 Jul 2020. 2

[4] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 2

[5] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15745–15753, 2021. 2

[6] Spyros Gidaris, Andrei Bursuc, Gilles Puy, Nikos Komodakis, Matthieu Cord, and Patrick Pérez. Obow: Online bag-of-visual-words generation for self-supervised learning.

In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6826–6836, 2021. 1, 2

[7] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, Bilal Piot, koray kavukcuoglu, Remi Munos, and Michal Valko. Bootstrap your own latent - a new approach to self-supervised learning. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 21271–21284. Curran Associates, Inc., 2020. 2

[8] Yuanfan Guo, Minghao Xu, Jiawen Li, Bingbing Ni, Xuanyu Zhu, Zhenbang Sun, and Yi Xu. Hcsc: Hierarchical contrastive selective coding. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9696–9705, 2022. 2

[9] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2980–2988, 2017. 2

[10] Geoffrey E Hinton and Sam Roweis. Stochastic neighbor embedding. In S. Becker, S. Thrun, and K. Obermayer, editors, *Advances in Neural Information Processing Systems*, volume 15. MIT Press, 2002. 2

[11] Lang Huang, Shan You, Mingkai Zheng, Fei Wang, Chen Qian, and Toshihiko Yamasaki. Learning where to learn in cross-view self-supervised learning. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14431–14440, 2022. 2

[12] Jiho Jang, Seonhoon Kim, Kiyoon Yoo, Chaerin Kong, Jangho Kim, and Nojun Kwak. Self-distilled self-supervised representation learning. In *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2828–2838, 2023. 1

[13] Junnan Li, Pan Zhou, Caiming Xiong, and Steven Hoi. Prototypical contrastive learning of unsupervised representations. In *International Conference on Learning Representations*, 2021. 1

[14] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019. 2

[15] Zhirong Wu, Yuanjun Xiong, Stella X. Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3733–3742, 2018. 2

[16] Haohang Xu, Xiaopeng Zhang, Hao Li, Lingxi Xie, Wenrui Dai, Hongkai Xiong, and Qi Tian. Seed the views: Hierarchical semantic alignment for contrastive representation learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):3753–3767, 2023. 1

[17] Ryota Yoshihashi, Shuhei Nishimura, Dai Yonebayashi, Yuya Otsuka, Tomohiro Tanaka, and Takashi Miyazaki. Ladder siamese network: a method and insights for multi-level self-supervised learning. *arXiv preprint arXiv:2211.13844*, 2022. 1

[18] Mingkai Zheng, Shan You, Fei Wang, Chen Qian, Changshui Zhang, Xiaogang Wang, and Chang Xu. Ressl: Relational self-supervised learning with weak augmentation. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 2543–2555. Curran Associates, Inc., 2021. 2