A Appendix

Optionally include extra information (complete proofs, additional experiments and plots) in the appendix. This section will often be part of the supplemental material.

A.1 Proof of proposition 1

Proposition 1. There exists a negative-positive coupling (NPC) multiplier $q_{B,i}^{(1)}$ in the gradient of $L_i^{(1)}$:

$$\begin{cases} -\nabla_{\mathbf{z}_i^{(1)}} L_i^{(1)} = \frac{q_{B,i}^{(1)}}{\tau} \left[\mathbf{z}_i^{(2)} - \sum_{l \in \{1,2\}, j \in [\![1,N]\!], j \neq i} \frac{\exp \langle \mathbf{z}_i^{(1)}, \mathbf{z}_j^{(l)} \rangle / \tau}{\sum_{q \in \{1,2\}, j \in [\![1,N]\!], j \neq i} \exp(\langle \mathbf{z}_i^{(1)}, \mathbf{z}_j^{(q)} \rangle / \tau)} \cdot \mathbf{z}_j^{(l)} \right] \\ -\nabla_{\mathbf{z}_i^{(2)}} L_i^{(1)} = \frac{q_{B,i}^{(1)}}{\tau} \cdot \mathbf{z}_i^{(1)} \\ -\nabla_{\mathbf{z}_j^{(l)}} L_i^{(1)} = -\frac{q_{B,i}^{(1)}}{\tau} \frac{\exp \langle \mathbf{z}_i^{(1)}, \mathbf{z}_j^{(l)} \rangle / \tau}{\sum_{q \in \{1,2\}, j \in [\![1,N]\!], j \neq i} \exp(\langle \mathbf{z}_i^{(1)}, \mathbf{z}_j^{(q)} \rangle / \tau)} \cdot \mathbf{z}_i^{(1)} \end{cases}$$

where the NPC multiplier $q_{B,i}^{(1)}$ is:

$$q_{B,i}^{(1)} = 1 - \frac{\exp(\langle \mathbf{z}_i^{(1)}, \mathbf{z}_i^{(2)} \rangle / \tau)}{\sum_{q \in \{1,2\}, j \in [\![1,N]\!], j \neq i} \exp(\langle \mathbf{z}_i^{(1)}, \mathbf{z}_j^{(q)} \rangle / \tau)}$$

Due to the symmetry, a similar NPC multiplier $q_{B,i}^{(k)}$ exists in the gradient of $L_i^{(k)}, k \in \{1,2\}, i \in [\![1,N]\!]$.

Proof.

$$\begin{aligned} -\nabla_{\mathbf{z}_i^{(1)}} L_i^{(1)} &= \frac{\mathbf{z}_i}{\tau} - \frac{1}{Y} \cdot \exp(\langle \mathbf{z}_i^{(1)}, \mathbf{z}_i^{(2)} \rangle / \tau) \cdot \frac{\mathbf{z}_i^{(2)}}{\tau} - \frac{1}{Y} \cdot \sum_{q \in \{1,2\}, j \in [\![1,N]\!], j \neq i} \exp(\langle \mathbf{z}_i^{(1)}, \mathbf{z}_j^{(q)} \rangle / \tau) \frac{\mathbf{z}_j^{(q)}}{\tau} \\ &= (1 - \frac{1}{Y} \cdot \exp(\langle \mathbf{z}_i^{(1)}, \mathbf{z}_i^{(2)} \rangle / \tau)) \frac{\mathbf{z}_i^{(2)}}{\tau} - \frac{1}{Y} \cdot \sum_{q \in \{1,2\}, j \in [\![1,N]\!], j \neq i} \exp(\langle \mathbf{z}_i^{(1)}, \mathbf{z}_j^{(q)} \rangle / \tau) \frac{\mathbf{z}_j^{(q)}}{\tau} \\ &= \frac{1}{\tau} (1 - \frac{1}{Y} \cdot \exp(\langle \mathbf{z}_i^{(1)}, \mathbf{z}_i^{(2)} \rangle / \tau)) \left[\mathbf{z}_i^{(2)} - \sum_{q \in \{1,2\}, j \in [\![1,N]\!], j \neq i} \frac{\exp(\langle \mathbf{z}_i^{(1)}, \mathbf{z}_j^{(q)} \rangle / \tau)}{U} \cdot \mathbf{z}_j^{(q)} \right] \\ &= \frac{q_{B,i}^{(1)}}{\tau} \left[\mathbf{z}_i^{(2)} - \sum_{q \in \{1,2\}, j \in [\![1,N]\!], j \neq i} \frac{\exp(\langle \mathbf{z}_i^{(1)}, \mathbf{z}_j^{(q)} \rangle / \tau)}{U} \cdot \mathbf{z}_j^{(q)} \right] \end{aligned}$$

where $Y = \exp(\langle \mathbf{z}_i^{(1)}, \mathbf{z}_i^{(2)} \rangle / \tau) + \sum_{q \in \{1,2\}, j \in [\![1,N]\!], j \neq i} \exp(\langle \mathbf{z}_i^{(1)}, \mathbf{z}_j^{(q)} \rangle / \tau)$, $U = \sum_{q \in \{1,2\}, j \in [\![1,N]\!], j \neq i} \exp(\langle \mathbf{z}_i^{(1)}, \mathbf{z}_j^{(q)} \rangle / \tau)$.

$$\begin{aligned} -\nabla_{\mathbf{z}_i^{(2)}} L_i^{(1)} &= \frac{1}{\tau} \mathbf{z}_i^{(1)} - \frac{1}{Y} \exp(\langle \mathbf{z}_i^{(1)}, \mathbf{z}_i^{(2)} \rangle / \tau) \cdot \frac{\mathbf{z}_i^{(1)}}{\tau} \\ &= \frac{q_{B,i}^{(1)}}{\tau} \cdot \mathbf{z}_i^{(1)} \end{aligned}$$

$$\begin{aligned} -\nabla_{\mathbf{z}_j^{(l)}} L_i^{(1)} &= \frac{1}{Y} \exp(\langle \mathbf{z}_i^{(1)}, \mathbf{z}_j^{(q)} \rangle / \tau) \cdot \frac{\mathbf{z}_i^{(1)}}{\tau} \\ &= \frac{q_{B,i}^{(1)}}{\tau} \cdot \frac{\exp(\langle \mathbf{z}_i^{(1)}, \mathbf{z}_j^{(q)} \rangle / \tau)}{U} \mathbf{z}_i^{(1)} \end{aligned}$$

□

A.2 Proof of proposition 2

Proposition 2. Removing the positive pair from the denominator of Equation 2 leads to a decoupled contrastive learning loss. If we remove the NPC multiplier $q_{B,i}^{(k)}$ from Equation 2, we reach a decoupled contrastive learning loss $L_{DC} = \sum_{k \in \{1,2\}, i \in [\![1,N]\!]} L_{DC,i}^{(k)}$, where $L_{DC,i}^{(k)}$ is:

$$
\begin{aligned}
L_{DC,i}^{(k)} &= -\log \frac{\exp(\langle \mathbf{z}_i^{(1)}, \mathbf{z}_i^{(2)} \rangle / \tau)}{\cancel{\exp(\langle \mathbf{z}_i^{(1)}, \mathbf{z}_i^{(2)} \rangle / \tau)} + \sum_{l \in \{1,2\}, j \in [\![1,N]\!], j \neq i} \exp(\langle \mathbf{z}_i^{(k)}, \mathbf{z}_j^{(l)} \rangle / \tau)} \\
&= -\langle \mathbf{z}_i^{(1)}, \mathbf{z}_i^{(2)} \rangle / \tau + \log \sum_{l \in \{1,2\}, j \in [\![1,N]\!], j \neq i} \exp(\langle \mathbf{z}_i^{(k)}, \mathbf{z}_j^{(l)} \rangle / \tau)
\end{aligned}
$$

Proof. By removing the positive term the denominator of Equation 4, we can repeat the procedure in the proof of Proposition 1 and see that the coupling term disappears.

□

A.3 Linear classification on ImageNet-1K

Top-1 accuracies of linear evaluation in Table 5 shows that, we compare with the state-of-the-art SSL approaches on ImageNet-1K. For fairness, we list the batch size and learning epoch of each individual approach, which are shown in the original paper. During pre-training, our DCL is based on a ResNet-50 backbone, with two views with size 224 × 224. Without relatively huge batch sizes or other pre-training schemes, i.e., momentum encoder, clustering, and prediction head, our DCL relies on its simplicity to reach competitive performance. We report both 200-epoch and 400-epoch versions of our DCL. It achieves 69.5% under the batch size of 256 and 400-epoch pre-training, which is better than SimCLR [8] in their optimal case, i.e., batch size of 4096, and 1000-epoch. Note that SwAV [26], BYOL [15], SimCLR [8], and PIRL [27] need huge batch size of 4096, and SwAV [17] further applies multi-cropping as generating extra views to reach optimal performance.

Table 5: ImageNet-1K top-1 accuracies (%) of linear classifiers trained on representations of different SSL methods.

Method	Architecture	Param. (M)	Batch size	Epochs	Top-1 (%)
Relative-Loc. [28]	ResNet-50	24	256	200	49.3
Rotation-Pred. [3]	ResNet-50	24	256	200	55.0
DeepCluster [26]	ResNet-50	24	256	200	57.7
NPID [4]	ResNet-50	24	256	200	56.5
Local Agg. [29]	ResNet-50	24	256	200	58.8
MoCo [7]	ResNet-50	24	256	200	60.6
SimCLR [8]	ResNet-50	28	256	200	61.8
CMC [6]	ResNet-50$_{L+ab}$	47	256	280	64.1
MoCo v2 [25]	ResNet-50	28	256	200	67.5
SwAV [17]	ResNet-50	28	4096	200	69.1
SimSiam [16]	ResNet-50	28	256	200	70.0
InfoMin [30]	ResNet-50	28	256	200	70.1
BYOL [15]	ResNet-50	28	4096	200	70.6
DCL	ResNet-50	28	256	200	67.8
PIRL [27]	ResNet-50	24	256	800	63.6
SimCLR [8]	ResNet-50	28	4096	1000	69.3
MoCo v2 [25]	ResNet-50	28	256	800	71.1
SwAV [17]	ResNet-50	28	4096	400	70.7
SimSiam [16]	ResNet-50	28	256	800	71.3
DCL	ResNet-50	28	256	400	69.5

A.4 Implementation details

DCL augmentations. We follow the settings of SimCLR [8] to set up the data augmentations. We use $RandomResizedCrop$ with scale in [0.08, 1.0] and follow by $RandomHorizontalFlip$. Then, $ColorJittering$ with strength in [0.8, 0.8, 0.8, 0.2] with probability of 0.8, and $RandomGrayscale$ with probability of 0.2. $GaussianBlur$ includes Gaussian kernel with standard deviation in [0.1, 2.0].

Linear evaluation. Following the open-sourced project, OpenSelfSup [23], we first train the linear classifier with batch size 256 for 100 epochs. We use the SGD optimizer with momentum = 0.9, and weight decay = 0. The base lr is set to 30.0 and decay by 0.1 at epoch [60, 80]. We further demonstrate the linear evaluation protocol of SimSiam [16], which raises the batch size to 4096 for 90 epochs. The optimizer is switched to LARS optimizer with base $lr = 1.2$ and cosine decay schedule. The momentum and weight decay are remained unchanged. We found the second one slightly improves the performance.

References

[1] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European Conference on Computer Vision*, pages 69–84. Springer, 2016.

[2] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *European conference on computer vision*, pages 649–666. Springer, 2016.

[3] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. *arXiv preprint arXiv:1803.07728*, 2018.

[4] 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*, pages 3733–3742, 2018.

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

[6] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. *arXiv preprint arXiv:1906.05849*, 2019.

[7] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020.

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

[9] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[10] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

[11] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2006), 17-22 June 2006, New York, NY, USA*, pages 1735–1742. IEEE Computer Society, 2006.

[12] Mang Ye, Xu Zhang, Pong C Yuen, and Shih-Fu Chang. Unsupervised embedding learning via invariant and spreading instance feature. In *Proceedings of the IEEE Conference on computer vision and pattern recognition*, pages 6210–6219, 2019.

[13] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *CoRR*, abs/2004.11362, 2020.

[14] Yang You, Igor Gitman, and Boris Ginsburg. Large batch training of convolutional networks. *arXiv preprint arXiv:1708.03888*, 2017.

[15] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Ávila Pires, Zhaohan 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 Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

[16] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. *CoRR*, abs/2011.10566, 2020.

[17] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *arXiv preprint arXiv:2006.09882*, 2020.

[18] Alexey Dosovitskiy, Philipp Fischer, Jost Tobias Springenberg, Martin Riedmiller, and Thomas Brox. Discriminative unsupervised feature learning with exemplar convolutional neural networks. *IEEE transactions on pattern analysis and machine intelligence*, 38(9):1734–1747, 2015.

[19] 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.

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

[21] Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 215–223. JMLR Workshop and Conference Proceedings, 2011.

[22] 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.

[23] Xiaohang Zhan, Jiahao Xie, Ziwei Liu, Dahua Lin, and Chen Change Loy. OpenSelfSup: Open mmlab self-supervised learning toolbox and benchmark. 2020.

[24] Xudong Wang, Ziwei Liu, and X Yu Stella. Unsupervised feature learning by cross-level instance-group discrimination.

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

[26] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 132–149, 2018.

[27] Ishan Misra and Laurens van der Maaten. Self-supervised learning of pretext-invariant representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6707–6717, 2020.

[28] Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised visual representation learning by context prediction. In *Proceedings of the IEEE international conference on computer vision*, pages 1422–1430, 2015.

[29] Chengxu Zhuang, Alex Lin Zhai, and Daniel Yamins. Local aggregation for unsupervised learning of visual embeddings. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6002–6012, 2019.

[30] Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning. *arXiv preprint arXiv:2005.10243*, 2020.