Exploring Data Augmentations on Self-/Semi-/Fully- Supervised Pre-trained Models

Anonymous Author(s)
Affiliation
Address
`email`

Abstract

Data augmentation has become a standard component of vision pre-trained models to capture the invariance between augmented views. In practice, augmentation techniques that mask regions of a sample with zero/mean values or patches from other samples are commonly employed in pre-trained models with self-/semi-/fully-supervised contrastive losses. However, the underlying mechanism behind the effectiveness of these augmentation techniques remains poorly explored. To investigate the problems, we conduct an empirical study to quantify how data augmentation affects performance. Concretely, we apply 4 types of data augmentations termed with `Random Erasing`, `CutOut`, `CutMix` and `MixUp` to a series of self-/semi-/fully- supervised pre-trained models. We report their performance on vision tasks such as image classification, object detection, instance segmentation, and semantic segmentation. We then explicitly evaluate the invariance and diversity of the feature embedding. We observe that: 1) Masking regions of the images decreases the invariance of the learned feature embedding while providing a more considerable diversity. 2) Manual annotations do not change the invariance or diversity of the learned feature embedding. 3) The `MixUp` approach improves the diversity significantly, with only a marginal decrease in terms of the invariance.

1 Introduction

Recently, self-/semi-/fully- supervised contrastive learning has achieved promising performance in learning meaningful representations during pre-training. Besides, the pre-trained models are successfully transferred to many downstream tasks, such as image classification, object detection, and instance segmentation. Terminologically, self-supervised contrastive learning refers to the pre-training without any labels introduced. While we term it as the semi-/fully- supervised contrastive learning when providing partial/all ground truths labels.

In the pure self-supervised configurations, data augmentations act as an essential component of self-supervised contrastive learning [1, 2, 3]. The algorithms are optimized to minimize the distance between different augmented views from the same sample (*a.k.a.* the anchor), while pushing views from different samples (the contrastive ones) away from the anchor. On the other hand, previous studies [1] show that with a limited amount of labels introduced, semi-supervised contrastive learning achieves better performance in related downstream tasks. Furthermore, fully-supervised contrastive learning with all ground truths further boosts the performance [4].

In practice, augmentation techniques that mask regions of a sample with zero/mean values or patches from other samples are commonly employed in semi-/fully- supervised (non-contrastive) learning. However, this family of augmentation techniques is not often applied in contrastive configurations, and the underlying mechanism behind the effectiveness of these augmentation techniques remains poorly explored. In this study, we implement 4 types of data augmentations termed with `Random`

Erasing, CutOut, CutMix and MixUp to a series of self-/semi-/fully- supervised pre-trained models. We then conduct a numerical study to quantify how data augmentation affects performance.

To this end, we clarify the terms *invariance* and *diversity* and provide the methods to calculate them explicitly. We then evaluate the invariance and diversity of the feature embedding of numerous pre-trained models. We demonstrate that *invariance* and *diversity* are closely related to the downstream tasks. Besides, we observe that: 1) Masking regions of the images decreases the invariance of the learned feature embedding while providing a more considerable diversity. 2) Manual annotations do not change the invariance or diversity of the learned feature embeddings. 3) The `MixUp` approach improves the diversity significantly, with only a marginal decrease in terms of the invariance.

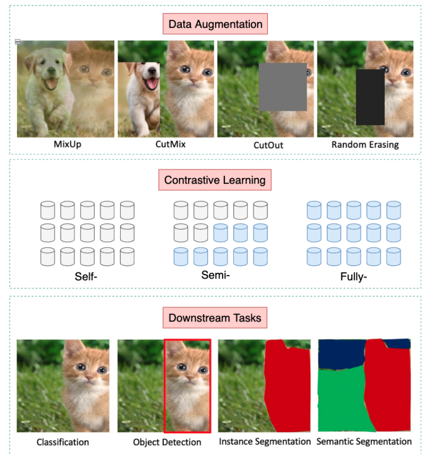

Figure 1: Illustration of our empirical study on four data augmentations (MixUp, CutMix, CutOut, Random Erasing), three pre-training types(self-, semi-, fully-supervised), and four downstream tasks (classification, object detection, instance segmentation, semantic segmentation).

Overall, the main contributions of this work can be summarized as follows:

- We conduct a comprehensive empirical study by quantifying how data augmentation affects the self-/semi-/fully- supervised contrastive learning frameworks.
- We provide an approach to measure the quality of the augmented view by explicitly examining the invariance and diversity metrics for self-/semi-/fully- supervised pre-trained models.
- Extensive experiments on various downstream benchmarks demonstrate that invariance and diversity are important metrics for the contrastive learning frameworks. Data augmentations that provide better invariance and diversity result in better performance in downstream tasks.

2 Methodology

In this work, we conduct an empirical study to quantify the effect of data augmentation techniques on the self-/semi-/fully- supervised contrastive learning frameworks. First, we begin with the formal problem setup for this empirical study. Then, we introduce self-/semi-/fully- supervised InfoNCE loss for comparisons. Finally, we propose two metrics, invariance, and diversity, to measure the quality of the augmented views between the anchor.

Notations. Given a pre-training set of N sample/label pairs, $\mathcal{N} = \{\boldsymbol{x}_i, \boldsymbol{y}_i\}_{k=1,\cdots,N}$. Under the commonly-used contrastive learning setting [3, 2], we generate two views $\boldsymbol{q}_i, \boldsymbol{k}_i$ for each sample $\boldsymbol{x}_i$. A set of negative samples for each sample $\boldsymbol{x}_i$ is $\mathcal{M}(i) = \{\boldsymbol{k}_m\}_{m=1,2,\cdots,M}$ and M is the number of negative samples.

2.1 Preliminaries: Self- & Fully-Supervised Contrastive Loss

Under the self-supervised contrastive learning framework, the main objective for each sample $\boldsymbol{x}_i$ is to maximize the similarity between the query $\boldsymbol{q}_i$. The corresponding augmented view $\boldsymbol{k}_i$, while minimizing the similarity between the query $\boldsymbol{q}_i$ and the negative sample $\boldsymbol{k}_m$. Thus, the overall objective $\mathcal{L}^{\text{self}}$ is formulated as:

$$\mathcal{L}^{\text{self}} = \sum_{i\in\mathcal{I}} \mathcal{L}_i^{\text{self}} = -\sum_{i\in\mathcal{I}} \log \frac{\kappa}{\kappa + \sum_{m\in\mathcal{M}(i)} \exp(\boldsymbol{q}_i \cdot \boldsymbol{k}_m/\tau)} \tag{1}$$

where κ is the positive similarity term, $\exp(\mathbf{q}_i \cdot \mathbf{k}_i/\tau)$, and $\mathcal{M}(i)$ denote the set of negative samples. τ is a temperature parameter.

By introducing all ground truths in the pre-training stage, we generate a new set $\mathcal{M}'(i)$ of negative samples, where the labels of negative samples are different from that of the anchor. Then, we define the fully-supervised objective $\mathcal{L}^{\text{full}}$ with the new negative set $\mathcal{M}'(i)$ as:

$$\mathcal{L}^{\text{full}} = \sum_{i \in \mathcal{I}} \mathcal{L}_i^{\text{full}} = \sum_{i \in \mathcal{I}} -\log \frac{\kappa}{\kappa + \sum_{m \in \mathcal{M}'(i)} \exp(\boldsymbol{q}_i \cdot \boldsymbol{k}_m / \tau)} \tag{2}$$

where $\mathcal{M}'(i) = \{\boldsymbol{k}_m | \boldsymbol{y}_m \neq \boldsymbol{y}_i\}$, and other settings are the same as in Eq. 1.

2.2 Semi-Supervised Contrastive Loss

In practice, it is unrealistic to acquire all labels from a large-scale pre-training set. Instead, obtaining partial annotations is operable. In this way, we split the original set $\mathcal{N}$ into two subsets, labelled set $\mathcal{D}$ and unlabelled set $\mathcal{U}$. Given the sample $\boldsymbol{x}_i$ in the labelled set $\mathcal{D}$, we maintain a negative samples queue $\mathcal{M}_d(i)$ and a label queue $\mathcal{Y}_d(i)$. In the meanwhile, we keep a negative samples queue $\mathcal{M}_u(i)$ for each sample in the unlabelled set $\mathcal{U}$. Then, we apply the fully-supervised contrastive loss $\mathcal{L}_i^{\text{full}}$ to the labelled set $\mathcal{D}$ and the self-supervised contrastive loss $\mathcal{L}_i^{\text{self}}$ to the unlabelled set $\mathcal{U}$. Therefore, the overall objective of semi-supervised contrastive loss is defined as:

$$\begin{aligned}\mathcal{L}^{\text{semi}} = \sum_{i \in \mathcal{D}} \mathcal{L}_i^{\text{full}} + \sum_{i \in \mathcal{U}} \mathcal{L}_i^{\text{self}} &= \sum_{i \in \mathcal{D}} -\log \frac{\kappa}{\kappa + \sum_{m \in \mathcal{M}_d(i)} \exp(\boldsymbol{q}_i \cdot \boldsymbol{k}_m / \tau)} \\ &- \sum_{i \in \mathcal{U}} \log \frac{\kappa}{\kappa + \sum_{m \in \mathcal{M}_u(i)} \exp(\boldsymbol{q}_i \cdot \boldsymbol{k}_m / \tau)}\end{aligned} \tag{3}$$

where $\mathcal{M}_d(i), \mathcal{M}_u(i)$ denotes the negative samples queue for the labelled set $\mathcal{D}$ and the unlabelled set $\mathcal{U}$. $\mathcal{M}_d(i) = \{\boldsymbol{k}_m | \boldsymbol{y}_d(i) \neq \boldsymbol{y}_i\}$. Other terms are the same as in Eq. 1 and 2.

2.3 Invariance

In order to measure the invariance between the augmented views $\boldsymbol{q}_i$ and the anchor $\boldsymbol{x}_i$, we propose a metric to calculate the normalized similarity invariance of the views in terms of the embedding space. Specifically, we take a set $\mathcal{V}_i$ of views, $\mathcal{V}_i = \{\boldsymbol{q}_i^v, v = 1, \cdots, V\}$, by applying data augmentations to the original sample $\boldsymbol{x}_i$. Then we calculate the normalized embeddings similarity between the augmented views $\boldsymbol{q}_i^v$ and the raw sample $\boldsymbol{x}_i$. Thus, we formulate the invariance metric of augmented views as:

$$\mathcal{L}_{inv} = \frac{1}{NV} \sum_{i=1}^{N} \sum_{v=1}^{V} \frac{\mathcal{S}(\boldsymbol{q}_i^v, \boldsymbol{x}_i)}{\mathcal{S}(\boldsymbol{x}_i, \boldsymbol{x}_i)} \tag{4}$$

where $\mathcal{S}(\boldsymbol{x}_i, \boldsymbol{q}_i^v)$ denotes the dot product metric for calculating the distance between $\boldsymbol{q}_i^v$ and $\boldsymbol{x}_i$. Note that $\mathcal{L}_{inv}$ achieves the maximum value 1 when $\boldsymbol{q}_i^v = \boldsymbol{x}_i$. This means that the augmented views have the maximum invariance from the anchor.

2.4 Diversity

In order to measure the quality of the augmented view in a comprehensive manner, we also propose to qualify the diversity of the augmented views. Specifically, we introduce a metric named *diversity* to measure how different the augmented views in the set $\mathcal{V}_i$ are. Based on the dot product distance metric $\mathcal{S}$, we define the diversity between two augmented views $\boldsymbol{q}_i^v$ and $\boldsymbol{q}_i^w$ as:

$$\mathcal{L}_{div} = \frac{1}{NV(V-1)} \sum_{i=1}^{N} \sum_{v=1}^{V} \sum_{w \neq v}^{V} \exp\left(\frac{\mathcal{S}(\boldsymbol{q}_i^v, \boldsymbol{q}_i^w)}{\sigma}\right) \tag{5}$$

where $\mathcal{S}(\boldsymbol{q}_i^v, \boldsymbol{q}_i^w)$ denotes the dot product distance metric between $\boldsymbol{q}_i^v$ and $\boldsymbol{q}_i^w$. σ is a scale parameter. In this way, we simultaneously maximize the diversity and invariance of the augmented views together to acquire views with best quality for self-/semi-/fully- supervised contrastive learning.

Table 1: Comparisons of linear classification evaluation on ImageNet-100 via applying four data augmentations to MoCo v2, where models are trained on frozen features from pre-trained encoders. Bold and underlined numbers denote the first and second place.

Method	Arch.	Param.(M)	Batch	Epochs	Top-1(%)	Top-5(%)	$\mathcal{L}_{inv}$	$\mathcal{L}_{div}$
MoCo v2 [3]	ResNet-50	24	256	200	81.65	95.77	**0.72**	0.23
MoCo v2 + Random Erasing	ResNet-50	24	256	200	81.04	95.27	0.59	0.42
MoCo v2 + CutOut	ResNet-50	24	256	200	82.64	95.84	0.67	0.36
MoCo v2 + CutMix	ResNet-50	24	256	200	83.51	96.51	0.61	**0.53**
MoCo v2 + MixUp	ResNet-50	24	256	200	84.08	96.79	0.69	0.45
MoCo v2 + 10% label	ResNet-50	24	256	200	82.26	95.80	**0.72**	0.23
MoCo v2 + 30% label	ResNet-50	24	256	200	82.55	95.83	**0.72**	0.23
MoCo v2 + 50% label	ResNet-50	24	256	200	83.21	96.36	**0.72**	0.23
MoCo v2 + 70% label	ResNet-50	24	256	200	83.75	96.62	**0.72**	0.23
MoCo v2 + 100% label	ResNet-50	24	256	200	84.93	97.18	**0.72**	0.23
MoCo v2 + MixUp + 50% label	ResNet-50	24	256	200	85.59	97.43	0.69	0.45
MoCo v2 + MixUp + 100% label	ResNet-50	24	256	200	**87.86**	**98.15**	0.69	0.45

3 Experiments

In this part, we conduct extensive experiments by transferring our model to four main downstream tasks, including linear classification, object detection, instance segmentation and semantic segmentation. In the meanwhile, we introduce $\mathcal{L}_{inv}$ and $\mathcal{L}_{div}$ to quantify how data augmentation affects the self-/semi-/fully-supervised pre-trained models. We give a comprehensive analysis on the effect of data augmentation and supervision during pre-training on various downstream tasks.

Linear Classification. Table 1 reports the top-1 and top-5 accuracy for linear classification on ImageNet-100 benchmark by applying four data augmentations to MoCo v2, where models are trained on frozen features from the pre-trained models. We can observe that MoCo v2+MixUp achieves better performance than other three data augmentations, including Random Erasing, CutOut, and CutMix. This is because the augmented views generated from MixUp have larger invariance between themselves and the anchor image. Meanwhile, with the increase of the number of given labels, we can observe an obvious performance gain in terms of both top-1 and top-5 accuracies, although our augmented views are not changed. This demonstrates the effectiveness of semi-/fully-supervised learning in learning more meaningful features for classification. Adding MixUp to the fully-supervised learning boosts the top-1 and top-5 accuracies to 87.86% and 98.15%. In terms of the invariance and diversity between augmented views, adding MixUp to the original MoCo v2 achieves the largest invariance score $\mathcal{L}_{inv}$ with best linear classification performance compared to other data augmentation techniques. In the meanwhile, all data augmentation techniques indeed increase the diversity score $\mathcal{L}_{div}$ while achieving better results than the baseline, which demonstrates the importance of measuring the quality of the augmented view by the proposed metrics. Furthermore, adding semi-supervised samples to MoCo v2 do not change the invariance and diversity scores as only augmented views are evaluated during training.

We compare data augmentation based semi-/fully-supervised models and other self-supervised methods for the linear classification evaluation on ImageNet-1K, as shown in Table 2 in Appendix. Applying MixUp to MoCo v2 increases the top-accuracy from 67.5% to 68.4%, which shows the effectiveness of additional data augmentations on the views generated by the baselines. With the increase of the number of given labels during pre-training, the linear classification accuracy consistently increases. Particularly, MoCo v2+MixUp+100% label achieves the best top-1 accuracy in terms of linear classification. Please see more experimental details and results in Appendix.

4 Conclusion

In this work, we perform a comprehensive empirical study to quantify how the self-/semi-/fully-supervised pre-trained models are affected by different data augmentation techniques. An approach is introduced to measure the quality of the augmented view by explicitly examining the invariance and diversity metrics for self-/semi-/fully- supervised pre-trained models. We also conduct extensive experiments on various downstream benchmarks, which demonstrate that invariance and diversity are important metrics for contrastive learning frameworks. Data augmentations that provide better invariance and diversity result in better performance in downstream tasks.

References

[1] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of International Conference on Machine Learning (ICML)*, 2020. 1, 8, 9, 10

[2] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9729–9738, 2020. 1, 2, 8, 9, 10

[3] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 1, 2, 4, 8, 9, 10

[4] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 1, 8

[5] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 34, pages 13001–13008, 2020. 8

[6] Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017. 8

[7] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *Proceedings of International Conference on Learning Representations*, 2018. 8

[8] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6023–6032, 2019. 8

[9] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *Proceedings of European Conference on Computer Vision (ECCV)*, 2016. 8

[10] Shin'ya Yamaguchi, Sekitoshi Kanai, Tetsuya Shioda, and Shoichiro Takeda. Multiple pretext-task for self-supervised learning via mixing multiple image transformations. *arXiv preprint arXiv:1912.11603*, 2019. 8

[11] Deepak Pathak, Philipp Krähenbühl, Jeff Donahue, Trevor Darrell, and Alexei A. Efros. Context encoders: Feature learning by inpainting. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2536–2544, 2016. 8

[12] Ishan Misra and Laurens van der Maaten. Self-supervised learning of pretext-invariant representations. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 8

[13] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *Proceedings of European Conference on Computer Vision (ECCV)*, 2016. 8

[14] Tete Xiao, Xiaolong Wang, Alexei A Efros, and Trevor Darrell. What should not be contrastive in contrastive learning. In *Proceedings of International Conference on Learning Representations (ICLR)*, 2021. 8

[15] Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning? In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, pages 6827–6839, 2020. 8

[16] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *Proceedings of European Conference on Computer Vision (ECCV)*, 2020. 8, 9

[17] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey Hinton. Big self-supervised models are strong semi-supervised learners. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 8

[18] 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 *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 8, 9, 10

[19] Yue Cao, Zhenda Xie, Bin Liu, Yutong Lin, Zheng Zhang, and Han Hu. Parametric instance classification for unsupervised visual feature learning. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, pages 15614–15624, 2020. 8, 9

[20] Hu Qianjiang, Wang Xiao, Hu Wei, and Qi Guo-Jun. AdCo: adversarial contrast for efficient learning of unsupervised representations from self-trained negative adversaries. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1074–1083, 2021. 8, 9

[21] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 8, 9, 10

[22] 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. 8, 9, 10

[23] Junnan Li, Pan Zhou, Caiming Xiong, and Steven Hoi. Prototypical contrastive learning of unsupervised representations. In *Proceedings of International Conference on Learning Representations (ICLR)*, 2021. 8, 9, 10

[24] Xudong Wang, Ziwei Liu, and Stella X Yu. CLD: unsupervised feature learning by cross-level instance-group discrimination. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 8

[25] Shentong Mo, Zhun Sun, and Chao Li. Siamese prototypical contrastive learning. In *BMVC*, 2021. 8

[26] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. *arXiv preprint arXiv:1906.05849*, 2019. 8

[27] 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 (NeurIPS)*, 2015. 9

[28] Mark Everingham, Luc Van Gool, Christopher K. Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, pages 303–338, 2010. 9

[29] Tsung yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 740–755, 2014. 9

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

[31] Agrim Gupta, Piotr Dollár, and Ross B. Girshick. LVIS: A dataset for large vocabulary instance segmentation. *arXiv preprint arXiv:1908.03195*, 2019. 9

[32] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3213–3223, 2016. 9

[33] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3431–3440, 2015. 9

[34] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5122–5130, 2017. 9

[35] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. *International Journal of Computer Vision (IJCV)*, 127:302–321, 2018. 9

[36] 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 (CVPR)*, pages 9729–9738, 2020. 9

[37] 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 (CVPR)*, 2018. 9

[38] Chengxu Zhuang, Alex Lin Zhai, and Daniel Yamins. Local aggregation for unsupervised learning of visual embeddings. In *Proceedings of the International Conference on Computer Vision (ICCV)*, pages 6002–6012, 2019. 9

[39] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 9

[40] Yannis Kalantidis, Mert Bulent Sariyildiz, Noe Pion, Philippe Weinzaepfel, and Diane Larlus. Hard negative mixing for contrastive learning. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 9, 10

[41] Yuwen Xiong, Mengye Ren, and Raquel Urtasun. LoCo: local contrastive representation learning. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, pages 11142–11153, 2020. 9

[42] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021. 11

Appendix

A Related Work

Data Augmentation. In the vision community, a branch of data augmentation methods [5, 6, 7, 8] have achieved promising performance in image related tasks, such as image classification and object detection. Typically, Random Erasing [5] selected a rectangle region in an image and erased its pixels with random values to reduce over-fitting and increase the robustness of trained model to occlusion. CutOut [6] randomly masked square regions of training images and tried to capture less prominent features for classification. MixUp [7] applied a convex combination of pairs of examples and their labels to improve the generalization of neural network architectures. CutMix [8] cut patches and pasted them from training images with mixed ground truth labels to train strong classifiers with localizable features. Recently, pretext tasks [9, 10, 11, 12, 13, 14] have been proven to be effective in self-supervised learning for meaningful visual representations. Researchers explore various pretext tasks to improve the quality of pre-trained representations, which includes colorization [9, 10], context autoencoders [11], spatial jigsaw puzzles [12, 13] and discriminate orientation [14].

However, a comprehensive recipe for data augmentations used in self-supervised learning is unexplored. In this work, we conduct an empirical study to exploit four main data augmentations over self-supervised methods on commonly-used benchmarks in terms of various downstream tasks. We further introduce invariance and diversity to quantify how data augmentation affects the performance of self-supervised pre-trained models.

Self-Supervised Learning. In the self-supervised literature, researchers aim to exploit the internal characteristics of data and leverage pretext tasks to train a model. Recently, an unsupervised framework that learns effective views with data augmentation was proposed by Tian *et al.* [15] to reduce the mutual information between views. CMC [16] introduced a multi-view contrastive learning framework with any number of views to learn view-agnostic representations. Another pretext task of solving jigsaw puzzles was developed in PIRL [12] to improve the semantic quality of learned image representations, achieving better object detection results than supervised pre-training.

In the past years, contrastive learning has shown its effectiveness in self-supervised learning, where various instance-wise contrastive learning frameworks [1, 17, 18, 2, 3, 4, 19, 20, 21] and prototype-level contrastive methods [22, 23, 24, 25] were proposed. The general idea of the instance-wise contrastive learning is to close the distance of the embedding of different views from the same instance while pushing embeddings of views from different instances away. One common way is to use a large batch size to accumulate positive and negative pairs in the same batch. For instance, Chen *et al.* [1] proposed a simple framework with a learnable nonlinear projection head and a large batch size to improve the quality of the pre-trained representations. To make best use of a large amount of unlabelled data, they present a bigger unsupervised pre-training network and introduce distillation with unlabeled data in SimCLR v2 [17] to improve the performance in downstream tasks. The dynamic dictionary was used with a moving-averaged encoder in MoCo series [3, 2] to build a dynamic dictionary to to update negative instances in a queue of large size.

Nevertheless, how to leverage labels in the momentum queue based pre-training is unexplored, especially their impacts on various downstream tasks, such as image classification, object detection, and semantic segmentation. This motivates us to comprehensively explore the effect of self-/semi-/full supervision on pre-trained models that are transferred to the aforementioned tasks. In the meanwhile, we quantify the effect of data augmentation on self-/semi-/fully- supervised contrastive learning frameworks.

B Pre-training Datasets & Settings

Following previous methods [2, 3, 26, 16], we use two popular benchmarks, *ImageNet-100* [16] and *ImageNet-1K*. The ImageNet-100 pre-trained model is evaluated on linear classification, and the ImageNet-1K model is transferred to various downstream tasks, including linear classification, object detection, instance segmentation and semantic segmentation.

For self-supervised pre-training on ImageNet-100 and ImageNet-1K, we closely follow the original MoCo v2 implementation [3]. SGD is used as our optimizer, where we apply a weight decay of 0.0001, a momentum of 0.9, and a batch size of 256. Our model is trained for 200 epochs with a

Table 2: Comparisons of linear classification evaluation on ImageNet-1K, where all results are trained under the same architecture. Parameters are of the feature extractor [36]. Views denote the number of images fed into the encoder in one iteration under batch size 1.

Method	Arch.	Param.(M)	Batch	Epochs	Views	Top-1 (%)
NPID [37]	ResNet-50	24	256	200	2x224	58.5
LocalAgg [38]	ResNet-50	24	128	200	2x224	58.8
MoCo [2]	ResNet-50	24	256	200	2x224	60.6
SimCLR [1]	ResNet-50	24	256	200	2x224	61.9
CPC v2 [39]	ResNet-50	24	512	200	2x224	63.8
CMC [16]	ResNet-50	47	128	240	2x224	66.2
MoCo v2 [3]	ResNet-50	24	256	200	2x224	67.5
PCL v2 [23]	ResNet-50	24	512	200	2x224	67.6
PIC [19]	ResNet-50	24	512	200	2x224	67.6
MoCHi [40]	ResNet-50	24	512	200	2x224	68.0
AdCo [20]	ResNet-50	24	256	200	2x224	68.6
SwAV [22]	ResNet-50	24	4096	200	2x224	69.1
LoCo [41]	ResNet-50	24	4096	800	2x224	69.5
BYOL [18]	ResNet-50	24	4096	200	4x224	70.6
SimSiam [21]	ResNet-50	24	256	200	4x224	70.0
MoCo v2 + MixUp	ResNet-50	24	256	200	2x224	68.4
MoCo v2 + MixUp + 50% label	ResNet-50	24	256	200	2x224	69.3
MoCo v2 + MixUp + 100% label	ResNet-50	24	256	200	2x224	**71.2**

initial learning rate of 0.03. The learning rate is then decayed by a factor of 10 at 120 and 160 epochs. For semi-/fully supervised pre-training, we use the same setting except that some or all labels are provided for maintaining the negative queue with labels.

C Transferring Datasets & Settings

Linear Classification. We evaluate linear classification on *ImageNet-100.* and *ImageNet-1K.* dataset, where a linear classifier is trained on frozen features from pre-trained weights. We report top-1,top-5 accuracy for ImageNet-100, and top-1 accuracy for ImageNet-1K.

Object Detection. For a fair comparison with previous work [2, 3], we fine-tune a Faster R-CNN detector [27] with C4-backbone end-to-end on the *PASCAL VOC* [28] 07+12 trainval set and evaluate on the VOC 07 test set. For *MS-COCO* [29] benchmark, we use the same hyper-parameters in MoCo [2], and fine-tune a Mask R-CNN [30] with C4 backbone on the train2017 set with 2x schedule and evaluate on val2017 set. The COCO box metrics (AP, AP_{50}, AP_{75}) are reported on both datasets.

Instance Segmentation. In terms of instance segmentation, we evaluate our pre-trained models on three popular benchmarks, including *MS-COCO* [29], *LVIS v1.0* [31], and *Cityscapes* [32]. For MS-COCO, we follow the same setting as the Mask R-CNN [30] used in the object detection task, where the COCO mask metrics (AP^m, AP^m_{50}, AP^m_{75}) are reported. For LVIS, we fine-tune an FCN model [33] on train set for 80k iterations and test on val set. We use the commonly-used metrics, AP, AP_c, AP_f, and AP_r for evaluation. For Cityscapes, an FCN model [33] is fine-tuned end-to-end on train_fine set for 40k iterations and test on val set, where AP^m and AP^m_{50} are reported for comparison.

Semantic Segmentation. We use *Cityscapes* [32] and *ADE20K* [34, 35] to evaluate semantic segmentation. For both benchmarks, we fine-tune an FCN model [33] on the train set for 40k iterations and test on the val set. Following previous work [2], we report two metrics (mIoU, mIoU_{sup}) for Cityscapes and four metrics (mIoU, fwIoU, mACC, pACC) for ADE20K to have a comprehensive comparison.

D Additional Experiments

Object Detection. We transfer various self-supervised pre-trained models to PASCAL VOC for object detection, and report the comparison results of AP, AP_{50}, and AP_{75} in Table 3a. As can be seen, adding MixUp to the pre-training with the highest invariance achieves the best results compared to other data augmentations. This further shows the importance of learning the invariance during pre-training for object detection on PASCAL VOC. We further evaluate our models pre-trained by

Table 3: Comparison results of object detection and instance segmentation on PASCAL VOC & COCO. Bold and underline denote the first and second place.

Method	AP	AP_{50}	AP_{75}
Random Initialization	32.80	59.00	31.60
Supervised	54.20	81.60	59.80
SimCLR [1]	51.50	79.40	55.60
BOYL [18]	51.90	81.00	56.50
SwAV [22]	55.40	81.50	61.40
MoCo [2]	55.90	81.50	62.60
MoCov2 [3]	57.00	82.40	63.60
SimSiam [21]	57.00	82.40	63.70
MoCov2 + Random Erasing	56.39	81.79	62.92
MoCov2 + CutOut	57.49	82.83	63.06
MoCov2 + CutMix	57.22	82.91	63.95
MoCov2 + MixUp	**57.61**	**82.96**	**64.30**

(a) PASCAL VOC.

Method	AP^b	AP^b_{50}	AP^b_{75}	AP^m	AP^m_{50}	AP^m_{75}
Random Initialization	32.80	50.90	35.30	29.90	47.90	32.00
Supervised	39.70	59.50	43.30	35.90	56.60	38.60
SwAV [22]	37.60	57.60	40.30	33.10	54.20	35.10
SimSiam [21]	39.20	59.30	42.10	34.40	56.00	36.70
MoCo [2]	40.70	60.50	44.10	35.40	57.30	37.60
MoCHi [40]	39.40	59.00	42.70	34.50	55.70	36.70
MoCov2 [3]	39.80	59.80	43.60	**36.10**	56.90	38.70
PCL [23]	41.00	60.80	44.20	35.60	57.40	37.80
MoCov2 + Random Erasing	40.14	60.25	43.82	35.35	57.13	37.75
MoCov2 + CutOut	40.84	60.73	44.25	35.72	57.41	**38.76**
MoCov2 + CutMix	40.75	60.67	44.12	35.53	57.23	38.24
MoCov2 + MixUp	**41.07**	**60.96**	**44.50**	36.05	**57.69**	38.37

(b) COCO.

Table 4: Comparison results of instance and semantic segmentation. Bold and underline denote the first and second place.

Method	AP	AP_c	AP_f	AP_r
MoCov2 [3]	17.08	8.16	15.35	22.94
+ Random Erasing	16.92	8.03	15.12	22.85
+ CutOut	17.19	8.18	15.36	23.06
+ CutMix	17.11	8.16	15.35	22.96
+ MixUp	**17.33**	**8.22**	**15.42**	**23.09**

(a) LVIS.

Method	AP^m	AP^m_{50}	mIoU	$mIoU_{sup}$
MoCov2 [3]	22.57	48.19	55.48	79.72
+ Random Erasing	22.51	48.15	55.35	79.63
+ CutOut	22.76	48.25	55.80	79.90
+ CutMix	22.55	48.19	55.45	79.67
+ MixUp	**22.83**	**48.28**	**55.92**	**79.96**

(b) Cityscapes.

Method	mIoU	fwIoU	mACC	pACC
MoCov2 [3]	20.62	54.68	27.15	69.59
+ Random Erasing	20.51	54.61	27.07	69.52
+ CutOut	20.76	54.72	27.19	69.61
+ CutMix	20.67	54.69	27.09	69.55
+ MixUp	**20.93**	**54.80**	**27.26**	**69.65**

(c) ADE20K.

various data augmentations on MS-COCO for a comprehensive comparison. The experimental results are reported in Table 3b. MoCo v2 + MixUp consistently achieves the best performance in terms of all metrics (AP^b, AP^b_{50}, AP^b_{75}), which further demonstrates the effectiveness of MixUp in learning a larger invariance between the augmented views and the anchor image.

Instance Segmentation. The comparison results of instance segmentation on MS-COCO are reported in Table 3b. We can observe that MoCov2 + CutOut achieves the best AP^m_{75} compared to other data augmentations. This is because MoCov2 + CutOut has the lowest diversity $\mathcal{L}_{div}$ between augmented views, demonstrating the importance of reducing the diversity of augmented views to improve the performance of instance segmentation. In Table 4a, we report the comparison results of instance segmentation by fine-tuning our pre-trained models on LVIS v1.0 benchmark. MoCo v2 + MixUp outperforms MoCo v2 + CutOut by a small margin since they achieves comparable diversity score $\mathcal{L}_{div}$ between augmented views, as we reported in Table 1. Moreover, MoCo v2 + Random Erasing achieves the worst performance in terms of all metrics. This shows the importance of keeping invariant features during pre-training while increasing the diversity of augmented views. We compare the results of instance segmentation on Cityscapes in Table 4b. We can observe a similar trend as LVIS v1.0 dataset, where MoCo v2 + MixUp performs the best while MoCo v2 + Random Erasing performs the worst, which further demonstrates the importance of learning the invariances from augmented views during pre-training and increasing the diversity of augmented views at the same time.

Semantic Segmentation. Table 4b shows the comparison results of semantic segmentation fine-tuned on Cityscapes dataset. MoCov2 + MixUp and MoCov2 + CutOut achieve comparable performance in terms of both metrics. This shows the effectiveness of learning the invariance and diversity together from augmented views during pre-training. With the smallest invariance score $\mathcal{L}_{inv}$, MoCov2 + Random Erasing performs worse than other data augmentations. In Table 4c, we report the comparison results of semantic segmentation fine-tuned on ADE20K dataset. We can make similar observations as the Cityscapes dataset. Compared to other data augmentations, MoCo v2 + Random Erasing achieves the worst results while MoCov2 + MixUp achieves the best performance. This further demonstrates the effectiveness of MixUp in keeping the invariance and increasing the diversity at the pre-training stage.

Table 5: Ablation Studies on augmented views and batch size, where top-1, top-5 accuracy, $\mathcal{L}_{inv}$, and $\mathcal{L}_{div}$ are reported on ImageNet-100.

# of views (V)	Top-1 (%)	Top-5 (%)	$\mathcal{L}_{inv}$	$\mathcal{L}_{div}$
2	**84.08**	**96.79**	**0.69**	0.45
3	82.37	95.81	0.58	0.53
4	81.55	95.68	0.51	**0.59**

(a) Augmented Views.

batch size (N)	Top-1 (%)	Top-5 (%)	$\mathcal{L}_{inv}$	$\mathcal{L}_{div}$
32	82.13	95.65	**0.75**	0.52
64	82.78	95.91	0.73	0.49
128	83.27	96.38	0.72	0.47
256	**84.08**	**96.79**	0.69	**0.45**
512	83.49	96.52	0.61	0.57
1024	82.92	96.23	0.58	0.63

(b) Batch Size.

E Additional Analysis

In this part, we explore the effect of the number of augmented views V and batch size N on the invariance and diversity. All experiments for ablation studies are conducted with MoCo v2 + MixUp on ImageNet-100 dataset.

Number of augmented views. In order to explore how the number of augmented V views affects the invariance and diversity, we set the value of V to 2, 3, and 4. The experimental results are reported in Table 5a. As can be seen, when V is set to 2, we achieve the best top-1 and top-5 accuracies with the largest invariance score $\mathcal{L}_{inv}$ and the smallest diversity score $\mathcal{L}_{div}$. With the increase in the number of augmented views, the performance of our model decreases a lot, which demonstrates the importance of selecting the right augmented views for contrastive learning.

Batch size. In order to demonstrate the effect of batch size on the final performance of invariance and diversity. Specifically, we set the number of batch size N to 32, 64, 128, 256, 512, 1024, and report the comparison results in Table 5b. When the batch size is set to 256, our model achieves the best performance in terms of the top-1 and top-5 accuracy. In the meanwhile, with the decrease in the batch size, both the invariance and diversity score increases, resulting in performance degradation.

F Limitation

The crucial limitation of this work is the scale of the datasets and backbones. Due to limited computational resources, the majority of the experiments are carried out on the ImageNet-100 dataset using the ResNet-50. Therefore we are unsure about the availability of the conclusions on much larger datasets and backbones. For instance, we do not perform experiments on costful transformer-based frameworks, such as DINO [42]. Nevertheless, we consider the results should generalize to other situations. On the other hand, we cannot enumerate all types of data augmentations that mask out information about the image. In recent studies, the patch-wise CutOut is shown effective in self-supervised algorithms such as masked image modeling. While in this work, we focus on the contrastive learning algorithm, the analysis of other data augmentations will be conducted in future works.

G Broader Impact.

The empirical results of our study benefit self-/semi-/fully- supervised pre-trained frameworks in the literature. Moreover, the analysis of the invariance and diversity terms helps in designing the appropriate data augmentation for the downstream tasks.