Curriculum SSL: An Efficient And Effective Data Augmentation For Contrastive Learning

Hongchao Fang [1]

Abstract

Contrastive learning has been one of the most successful approach for self-supervised learning and is gaining more attentions ever since. Recent works showed the importance of crafting good positive pairs, especially for contrastive methods that only leverage on positive pairs. Most previous works implement the standard data augmentation with random cropping and did not take into account the congruence within positive pairs. While some works tried to tackle this problem at the cost of computational overhead. In this work, we propose **MICO(Mixup Contrastive learning)**, an efficient yet effective method that produces better positive image pairs without hindering the training speed. Specifically, we use MixUp as a stronger data augmentation method and leverage curriculum learning to overcome the difficulty of additional perturbations on the input image. Empirical results show that our method achieved consistent improvements on multiple contrastive learning frameworks and 3% on cifar100 datasets.

1. Introduction

Advances in contrastive learning show the promising aspect of self-supervised learning (SSL) for its capability of taming enormous amount of unlabeled data. Recent contrastive methods (Grill et al., 2020; He et al., 2019; Chen et al., 2020b; Chen & He, 2020; Chen et al., 2020a; Caron et al., 2020) have achieved amazing performances on wide range of computer vision tasks, including object detection and instance segmentation, that are comparable or even superior than its supervised counterpart. The success of contrastive learning can be largely attributed to the advantage of Siamese structure (Bromley et al., 1993; Chen & He, 2020) where multiple inputs go through a weight-sharing network and the goal is to minimize distance between inputs in the feature space. Contrastive methods differ in structures based on presence of negative pairs, the presence of positive pair, however, is universal across line of works. Therefore, it is of necessity to design good positive pairs. Current contrastive methods generate positive pairs by applying data augmentations. Standard practice includes random crop, horizontal flip, color jitter and Gaussian noise (Wu et al., 2018; He et al., 2019). More complicated data augmentations (Peng et al., 2022; Selvaraju et al., 2020) have also been proposed at the cost of additional computational overhead. In order to produce more information-rich positive pairs, the contrastive learning will introduce stronger data augmentation, but it also adds noise and subject to performance degradation(Tian et al., 2020). One way to avoid this performance degradation is to introduce curriculum learning (Bengio et al., 2009).

Inspired by human learning, curriculum learning(CL) principles on solving tasks with increasing difficulties. Under the CL strategy, the model will be first solving easy tasks, and gradually transitions to solving harder tasks later in the training stage. This strategy has achieved great success in supervised computer vision tasks (Gong et al., 2016) as well as in other domains such as NLP, robotics, and speech(Tay et al., 2019). The empirical success of CL establishes CL as a general solution suitable for multifarious problems. Specifically, CL works well under noisy condition (Wu et al., 2020) which inspired us to explore the usage of curriculum learning in self-supervised settings.

In this paper, we propose MICO, a simple yet effective method for contrastive learning with strong data augmentation. Compared to the traditional contrastive learning method and MixUp augmentation, we applied curriculum learning for the lambda in MixUp to form different levels of tasks and achieved better performance on cifar10/cifar100.

Our main contribution can be summarized as

- Define difficulty as data augmentation strength of curriculum learning, and introduce stepwise scheduler for curriculum learning.
- Use MixUp as a special data augmentation and show the effectiveness of curriculum learning in multiple datasets
- Combined MixUp augmentation and contrastive learning to form a new framework with higher performance.

2. Related works

In this section, we will introduce contrastive learning and curriculum learning relevant to our work.

2.1. Contrastive Learning

Contrastive learning aims to pull similar samples closer and dissimilar samples far away in the visual representation by solving various pretext tasks (Jaiswal et al., 2020). Early works explore unlabeled training by using similarity metric based loss function (Chopra et al., 2005). FaceNet first proposes triplet loss with positive and negative samples, and generalized to include multiple negative pairs (Schroff et al., 2015).

SimCLR proposes an end-to-end learning paradigm for contrastive learning with large batchsize and a projection head(Chen et al., 2020a). InstDisc avoids large batchsize by using memory bank to save the feature representation of negative samples (Wu et al., 2018). MoCo replaces computationally expensive memory bank with a shared-parameter momentum encoder to generate a queue of negative samples with smooth updates to ensure consistency (He et al., 2019). More recent works have shown the possibility of learning with only positive samples. BYOL uses online network to predict the output of the target network with exponential moving average(EMA) (Grill et al., 2020). By using the stop gradient operation and EMA of the target network, the author claimed the collapse of training in absence of negative samples can be avoided. On top of that, Simsiam explores the core element that prevents collapsing by conducting ablation study on the removal of the EMA and the stop gradient operation, and deemed the stop gradient operation as the critical component that prevents the collapsing.

2.2. Curriculum Learning

Inspired by how human learns, curriculum learning mimics the curricula system in schooling system as human learns material more efficiently in an organized order. This order is canonically from easy to hard, as human learns the concept faster with easier tasks and consolidates the learned concept on harder tasks. The pioneering work on this direction dates back to 1990s where Elmar proposed network should train from limited memory to mature stage (1993). Vanilla CL was proposed as an easy-to-hard learning paradigm that reorders training samples with a priori rule that discriminate easy and difficult samples (Bengio et al., 2009). Self-paced learning(SPL) does not rely on prior rule. Instead, SPL computes the order with respect to the performance of the model (Kumar et al., 2010). SPCL defines the order of training samples by combining predefined rules and learning based metrics on matrix factorization tasks (Jiang et al., 2015). Curriculum dropout conceptually utilizes the easy-to-hard paradigm by gradually increasing the dropout probability with time scheduling and shows superior results against overfitting (Morerio et al., 2017).

Empirically, CL has proven its effectiveness in multiple domains as a general training paradigm. SPCN demonstrates under supervised setting, CL improves baseline on image classfication (Li & Gong, 2017). Some works show that facial recognition tasks also benefit from CL (Lin et al., 2018; Huang et al., 2020).

Most contrastive methods require strong data augmentation to create diversified positive pairs in order to learn good visual representations. AutoAugment learns the best data augmentation setting as a RL problem and tries to find the combination that has the highest accuracy on evaluation set (Cubuk et al., 2018b). Build upon AutoAugment, RandAugment reduces the data augmentation search space by introducing an augmentation magnitude parameter (Cubuk et al., 2019). Cutout randomly masks out a square area of the input image (DeVries & Taylor, 2017). CutMix creates new data by conducting regional cut-and-paste mixture of two images (Yun et al., 2019). Conditioned on coefficient lambda, Mixup conducts pixel-wise mixture on global image (Zhang et al., 2017). Although most of the data augmentation methods were proposed for supervised setting, works have shown the feasibility of their implications on contrastive methods. MixCo explores pairwise mixup between positive and negative samples (Kim et al., 2020). ContrastiveCrop designed a semantic-aware cropping strategy that contains target object and reduces MI comparing to random crop (Peng et al., 2022).

3. Method

In this section we propose a generic curriculum learning strategy on SSL. We will provide an overview on how to apply curriculum learning onto SSL in Section 3.1. We will provide specific CL setup for hard negative sampling in Section 3.3 and for data augmentation in Section 3.4.

3.1. Overview

Curriculum learning generally relies on 1) *difficulty measurer* which defines the ”difficulty” of the task and 2) *training scheduler* which regulates how difficulty is changed throughout the training (Bengio et al., 2009; Wang et al., 2022). We empirically investigate how to apply curriculum learning onto self-supervised learning tasks with simple yet effective definition of difficulty under the setting of hard negative sampling (Robinson et al., 2020) and data augmentation strength. We further propose Curriculum-Mixup in Section 3.5 as an extension of showing the power of curriculum learning as a continuation method that helps to ease the difficult optimization problem.

3.2. CL on Rotation Prediction

RotNet uses image rotation prediction as the pretext tasks for unsupervised pretraining (Gidaris et al., 2018). Formally, this problem is setup as

$$F(X^{y*}|\theta) = \{F^y(X^{y*}|\theta)\}_{y=1}^{K} \tag{1}$$

where $y \in \{0, 90, 180, 270\}$ is the rotation degree X^y is the rotated image with rotation degree y. The goal is to predict the rotation degree given the transformed image from the original untransformed image. Intuitively, the larger the rotation angle, the more difficult to predict.

However, this global level of image rotation does not attribute to the local importance of the input image. In particular, when the target object takes only a small portion of the image, the rotation of the entire image will most likely unable to capture the rotation invariant information. We propose crop-and-rotate based on this assumption. Figure 1 shows a general pipeline of the crop-and-rotate prediction method. We follow ContrastiveCrop by using semantic-aware localization to bound the center object. Specifically, we use heatmap to define the bounding box and update the bounding boxes 4 times during the entire pretraining stage to minimize the incurred computation cost. For each image, we only rotate the part that is within the bounding box.

Although predicting rotation angles requires model to learn rotation invariant information, the rotated image does not contain the unrotated information (except when $y = 0$). It is reasonable to assume that when both rotated information and unrotated information are both passed to the model, this contrast will helps the model to learn a better visual representation. Under this assumption, we propose rotate-and-mix. Figure 2 shows the general pipeline of the proposed method.

While all three proposed rotation prediction method intuitively make sense, it is possible that the additional cropping and mixing will cause the model not converging due to the difficulty and noise added to the training process. To tackle the additional hardness, we propose a curriculum learning strategy on y where image is gradually rotated from 0 degree to 270 degree with training scheduler. Empirically, curriculum learning has shown to be useful under the limited training time and noisy training setting (Wu et al., 2020). We follow the basic discrete training scheduler where the rotation angle is changed based on training progress.

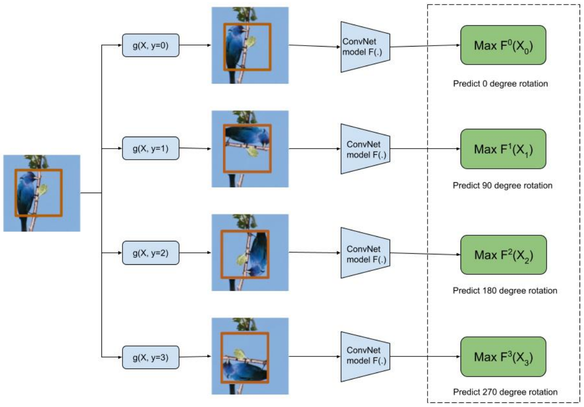

Figure 1. Crop-and-rotate general pipeline.

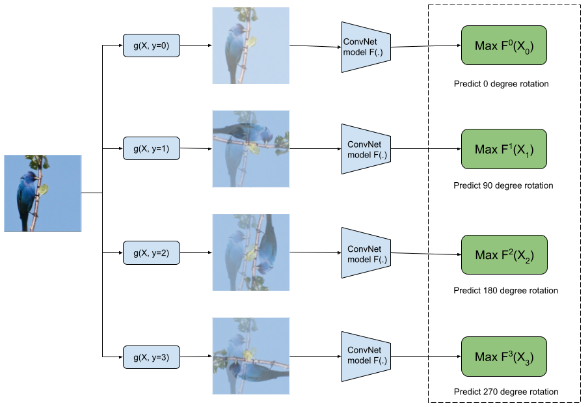

Figure 2. rotate-and-mix general pipeline.

3.3. CL on Hard Negative Sampling

By up-weighting negative distribution over points with different label $p_x^-(x^-)$, Robinson et al. (2020) proposed a sampling probability

$$q_\beta(x^-) \propto e^{\beta f(x)^\top f(x^-)} p(x^-) \tag{2}$$

where $\beta \geq 0$ is the concentration parameter that controls the degree of similarity between negative samples x^- and the anchor image. Larger β value means higher similarity between the anchor image and the negative sample, and thus, hardness of the task. This exact hardness adjustable sampling strategy forms a trivial yet self-contained definition of the difficulty for the curriculum learning.

Denote $S(\beta, i)$ as training scheduler where β is the concentration parameter described above and i is the i_{th} step during training, a naive CL of sampling probability is $q_\beta^i(x^-) \propto e^{S(\beta,i) f(x)^\top f(x^-)} p(x^-)$. We adopt a discrete scheduler inspired by (Spitkovsky et al., 2010) for its simplicity and effectiveness. Section **??** contains our explorations on different training schedulers.

While HCL (Robinson et al., 2020) focuses on mining the hard negative samples (images from similar but different classes), conceptually it works on hard positive samples as well. Analogous to hard negative sampling, the hard positive sampling follows

$$q_\beta(x^+) \propto e^{(k-\beta)f(x)^\top f(x^+)} p(x^+) \tag{3}$$

where $\beta \geq 0$ is the concentration parameter that controls the degree of similarity between negative samples x^+ and the anchor image. Note that equation 2 and equation 3 share same β and thus follow same training rate scheduler $S(\beta, i)$. k is some predefined constant such that $k - \beta$ is decreasing as β increases. Under these settings, we propose a unified hard contrastive sampling technique.

While HCL samples hard negative images from the batch with a controlling parameter β, due to the nature of unsupervised learning, in extreme circumstances large β could result false negative sampling. Since the ground truth label of a particular sample is irretrievable during the training stage, the value of β is forced to be restrained in a safe interval. Therefore, the power of the hard negative sampling would not be fully harnessed. An alternative approach would be first use hard negative sampling with a smaller β, and then use data augmentation to further increase the hardness. One of the method is to use adversarial training to augment the negative samples. Specifically, we use pgd-attack (Madry et al., 2017) as a data augmentation method to generate harder negative images with the following equations:

$$x^{t+1} = \prod_{x+S} (x^t + \alpha \mathrm{sgn}(\nabla_x L(\theta, x, y))) \tag{4}$$

Where α is the step size and S is the number of iteration. It can be seen as a minimax optimization problem. After hard negative sampling, we maximize the loss of the noise ϵ and minimize the loss of the negative sample with ϵ.

3.4. CL on Data Augmentation

Since contrastive learning depends on contrastive difference between two separately augmented data views, the stronger the data augmentation strength, the more dissimilar two views are, and hence the harder the task. Therefore, it is intuitive to define the difficulty as the strength of the data augmentation in curriculum setting. Under standard data augmentation settings (random cropping, color jitter, grayscale), we apply curriculum learning on color distortion methods (color jitter, grayscale) since they have the most significant impact on contrastive learning performance (Chen et al., 2020a). We adopt the same training scheduler as in Section 3.3 as we empirically find it is favorable under data augmentation as well.

While the vanilla data augmentation works empirically, handcrafted values maybe suboptimal. Automatic data augmentation method, AutoAugment for example, applies reinforcement learning strategy to find an optimal policy that maximize the performance of the model on validation dataset (Cubuk et al., 2018a). Follow AutoAugment, we train 5 policies where each policy consists two data augmentation methods. While the vanilla AutoAugment learns the optimal augmentation methods, the magnitude, however, is learned simultaneously without smoothing techniques. Due to inter-batch statistical difference, directly learned magnitude values might change drastically in the worst scenario and cause the model to converge in a suboptimal solution. To overcome this problem, we proposed to add an auxiliary module that restrict and guide the update value of the magnitude. Figure 3 shows a general structure of the proposed strategy. The training scheduler model is updated alongside with the policy by the controller and works as an additional normalization to ensure the smoothness of the changing the policy.

The training scheduler module can be implemented with self-paced learning strategy guided by the validation loss. Formally, we are trying to minimize

$$\min_{\mathbf{w},\mathbf{v}\in[01]^n} \mathbb{E}(\mathbf{w}, \mathbf{v}, \lambda, \Psi) = \sum_{i=1}^{n} L(y_i, g(\mathbf{x}_i, \mathbf{w})) + f(\mathbf{v}; \lambda) \tag{5}$$

where w is the model parameter, $L(y_i, g(\mathbf{x}_i, \mathbf{w}))$ is the loss function, λ controls the pace and v denotes the weight variables reflecting the policy's importance (Jiang et al., 2015). $f(\mathbf{v}; \lambda)$ is the pacing function that controls the magnitude of S

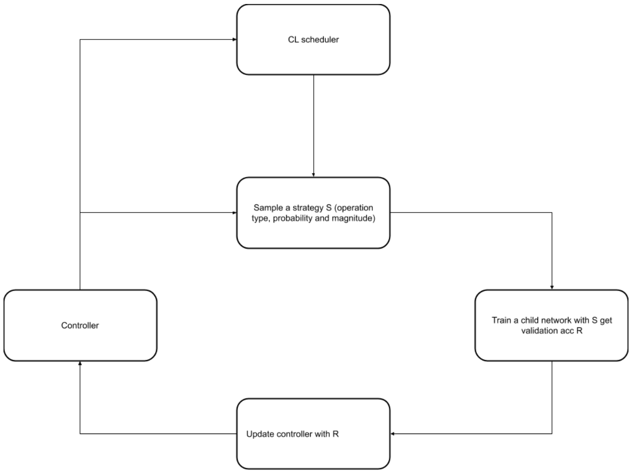
CL scheduler
Sample a strategy S (operation type, probability and magnitude)
Controller
Train a child network with S get validation acc R
Update controller with R

Figure 3. normalized autoaugment general pipeline.

3.5. Curriculum Mixup

As a special type of data augmentation, Mixup uses global mixture to generate new data point consisting the original two input images. Formally, given raw input vectors x_i, x_j with corresponding one-hot label y_i, y_j, Mixup is defined as

$$\tilde{x} = \lambda x_i + (1 - \lambda)x_j$$

$$\tilde{y} = \lambda y_i + (1 - \lambda)y_j$$

where $\lambda \in [0, 1]$ which encourage the model to generalize well on unseen dataset by generating these virtual data points.

Under self-supervised setting, Mixup is only conducted on the image and not on the labels, therefore we are only interested in $\tilde{x} = \lambda x_i + (1 - \lambda)x_j$ where $\lambda = \text{Beta}(\alpha, \alpha)$ and α is a hyperparameter. Let $f(x)$ be the model $g(x)$ produces intermediate representation given input x and $f(x) = h(g(x))$. Formally, this can be defined as a modified InfoNCE loss:

$$L_{MixUp} = -\sum_{i=1}^{n} \log \frac{\exp(f(x_i) \cdot f(\lambda x_i + (1 - \lambda)x'_i)/\tau)}{\sum_{j=0}^{K} \exp(f(x_i) \cdot f(x_j)/\tau)} \tag{6}$$

We consider Mixup as a bridge between the strongly augmented positive pairs. Although strong data augmentation resulting in reduced MI positive pairs, additional noises incurred by augmentations makes the contrastive model hard to learn good visual representations (Tian et al., 2020). The vanilla Mixup has a fixed mixing proportion between two images and does not provide enough information during final training stage. Our experiments in Table 1 show that the vanilla Mixup on SimSiam does not significantly improve the baseline.

Curriculum learning has been showed empirically that it can help model's convergence in noisy setting (Wu et al., 2020). To help contrastive model to converge in noisy setting incurred by strong data augmentation, we use CL with Mixup throughout the training stage. Similar to progressive CL, we define the difficulty as the magnitude of mixup parameter λ. A large λ means two images are more similar, thus a easier task. A small λ means two images are more dissimilar, thus a more difficult task. We use a stepwise CL scheduler for its empirical superior performance and low computational overhead.

Since we are doing the image mixture solely on positive pairs, our method is applicable to all contrastive methods that consider positive pairs. Algorithm 1 shows how to apply our strategy on SimSiam framework.

4. Experiment

In this section, we empirically investigate how does our method perform with popular contrastive learning methods and common benchmarking datasets. We will first introduce datasets and baseline approachesin Section 4.2

Algorithm 1 Curriculum Mixup in SimSiam

```
# f: backbone + projection mlp
# h: prediction mlp
# S: training scheduler
for epoch in total_epochs:
    alpha = S(epoch, total_epochs)
    for x in loader:
       λ = Beta(alpha, alpha)
       x1, x2 = aug(x), aug(x)
       mixed_x2 = λx1 + (1-λ)x2
       z1, z2 = f(x1), f(mixed_x2)
       p1, p2 = h(z1), h(z2)
       L = D(p1, z2)/2 + D(p2, z1)/2
       L.backward()
       update(f, h)
def D(p, z):
   z = z.detach()
   p = normalize(p, dim=1)
   z = normalize(z, dim=1)
   return -(p*z).sum(dim=1).mean()
```

4.1. Baseline Approach

We present our implementation details in Section 4.4.Section 4.5 contains linear evaluation results on baseline and our methods. We present an ablation study in Section 4.6 denoting the influence of mixup crops with different levels of curriculum learning.

4.2. Datasets

We assess our method on standard objection recognition datasets.**CIFAR-10/100**(Krizhevsky & Hinton, 2009) consists 60000 color images of size 32 × 32 pixels. CIFAR-10 has 10 classes with 5000 training images and 1000 testing images per class. CIFAR-100 has 100 classes with 500 training images and 100 testing images per class.

4.3. Baseline Approach

We use wide range of state-of-the-art contrastive learning frameworks to show the generalizibility of our method. These frameworks include SimSiam(Chen & He, 2020) and BYOL(Grill et al., 2020) that only leverage positive pairs, and MOCO(He et al., 2019; Chen et al., 2020b) and SimCLR(Chen et al., 2020a) which use both positive and negative samples.

4.4. Implementations

Since our goal is to provide insights on how to apply curriculum learning strategy onto self-supervised tasks, we follow the hyperparameters and settings of the baseline if not otherwise explicitly stated. Specifically, we use ResNet-18 as the backbone for CIFAR-10/100 experiments. For each contrastive learning method, we explore the performance for three frames: Baseline, MixUp, and C-Mixup(mixup with

Table 1. Downstream linear classification results for different contrastive learning methods on small datasets.

Method	CIFAR-10			CIFAR-100		
	BASELINE	MIXUP	C-MIXUP	BASELINE	MIXUP	C-MIXUP
SimCLR	90.38	91.38	**91.55**	59.86	63.88	**63.98**
MOCO	**90.86**	88.99	90.67	**62.20**	61.48	61.98
BYOL	93.45	**93.76**	93.58	69.62	**70.13**	69.87
SimSiam	90.83	91.03	**92.66**	66.1	66.05	**69.39**

curriculum learning). For the Mixup framework, we use 0.5 as our mixup lambda, and for C-mixup, we use 4-step lambda(0.2, 0.4, 0.6, 0.8) for a single augmented picture.

4.5. Linear Classification

Table1 shows the effectiveness of our Mixup and C-Mixup framework on different contrastive learning methods. Generally, C-mixup and mixup can improve methods with limited negative samples like SimSiam, BYOL, and SimCLR. Among them, BYOL benefits more from stable mixup lambda, and SimSiam and SimCLR gain more from increasing lambda.

4.6. Ablation Study

For the ablation study, we investigated the influence of schedule for curriculum learning based on MOCO. We experiment with four kinds of curriculum strategies, 2-step(0.33, 0.66), 4-step(0.2, 0.4, 0.6, 0.8), Linear(0 1) and reverse cosine(0.8, 0.6, 0.4, 0.2). We assess the four strategies on the CIFAR-10 dataset and use two types of augmentations: Standard single augmentation and double augmentation.

In Table 2, for a tiny epoch, the 2-step strategy performs best on standard augmentation, and the 4-step strategy works best on double augmentation. However, for significant epochs, the 4-step strategy performs best in all augmentations in Table 3.

5. Conclusion

In this study, we introduced MICO, a novel approach that seamlessly integrates Mixup data augmentation with contrastive learning frameworks. Our method, focused on enhancing the generation of positive image pairs, demonstrates a significant improvement in learning efficiency and performance across various datasets and models.

Our experiments validate that both Mixup and C-Mixup frameworks notably enhance the performance of contrastive learning methods, particularly in scenarios with limited negative samples. We observed distinct advantages in applying stable Mixup lambda in BYOL, while SimSiam and SimCLR showed more substantial improvements with an increasing lambda schedule. These findings underscore the adaptability and effectiveness of our approach across different contrastive learning frameworks.

Overall, MICO represents a significant step forward in the field of self-supervised learning. By leveraging Mixup as a stronger data augmentation method and combining it with curriculum learning, we have demonstrated a method that not only improves the quality of positive pairs in contrastive learning but also maintains computational efficiency. Our approach sets a new benchmark for future research in self-supervised and contrastive learning methods, paving the way for more robust and efficient learning algorithms.

Future work will focus on exploring the potential of MICO in broader applications, including more diverse datasets and real-world scenarios, to further establish its versatility and effectiveness in various learning environments.

References

Bengio, Y., Louradour, J., Collobert, R., and Weston, J. Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pp. 41–48, New York, NY, USA, 2009. Association for Computing Machinery. ISBN 9781605585161. doi: 10.1145/1553374.1553380. URL https://doi.org/10.1145/1553374.1553380.

Bromley, J., Guyon, I., LeCun, Y., Säckinger, E., and Shah, R. Signature verification using a "siamese" time delay neural network. In *Proceedings of the 6th International Conference on Neural Information Processing Systems*, NIPS'93, pp. 737–744, San Francisco, CA, USA, 1993. Morgan Kaufmann Publishers Inc.

Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., and Joulin, A. Unsupervised learning of visual features by contrasting cluster assignments, 2020. URL https://arxiv.org/abs/2006.09882.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. A simple framework for contrastive learning of visual representations, 2020a. URL https://arxiv.org/abs/2002.05709.

Chen, X. and He, K. Exploring simple siamese represen-

Table 2. Curriculum Learning on direct data augmentation cifar10 on MOCO v1

Method	Epoch	augmentation strength	scheduler	Accuracy
MOCO v1	200	Standard	baseline	84.86
			2 step	86.33
			4 step	85.09
			linear	83.8
			reverse cosine	83.43
		Double	baseline	84.4
			4 step	86.68
			linear	84.96
			reverse cosine	84.02

Table 3. Curriculum Learning on direct data augmentation cifar10 on MOCO v1

Method	Epoch	augmentation strength	scheduler	Accuracy
MOCO v1	800	Standard	baseline	89.69
			4 step	90.02
		one and half	baseline	89.42
			4 step	90.39
		Double	baseline	89.06
			4 step	90.54
		Triple	baseline	88.27
			4 step	90.07

tation learning, 2020. URL https://arxiv.org/abs/2011.10566.

Chen, X., Fan, H., Girshick, R., and He, K. Improved baselines with momentum contrastive learning, 2020b. URL https://arxiv.org/abs/2003.04297.

Chopra, S., Hadsell, R., and LeCun, Y. Learning a similarity metric discriminatively, with application to face verification. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pp. 539–546 vol. 1, 2005. doi: 10.1109/CVPR.2005.202.

Cubuk, E. D., Zoph, B., Mane, D., Vasudevan, V., and Le, Q. V. Autoaugment: Learning augmentation policies from data, 2018a. URL https://arxiv.org/abs/1805.09501.

Cubuk, E. D., Zoph, B., Mane, D., Vasudevan, V., and Le, Q. V. Autoaugment: Learning augmentation policies from data, 2018b. URL https://arxiv.org/abs/1805.09501.

Cubuk, E. D., Zoph, B., Shlens, J., and Le, Q. V. Randaugment: Practical automated data augmentation with a reduced search space, 2019. URL https://arxiv.org/abs/1909.13719.

DeVries, T. and Taylor, G. W. Improved regularization of convolutional neural networks with cutout, 2017. URL https://arxiv.org/abs/1708.04552.

Elman, J. L. Learning and development in neural networks: the importance of starting small. *Cognition*, 48(1):71–99, 1993. ISSN 0010-0277. doi: https://doi.org/10.1016/0010-0277(93)90058-4. URL https://www.sciencedirect.com/science/article/pii/0010027793900584.

Gidaris, S., Singh, P., and Komodakis, N. Unsupervised representation learning by predicting image rotations, 2018. URL https://arxiv.org/abs/1803.07728.

Gong, C., Tao, D., Maybank, S. J., Liu, W., Kang, G., and Yang, J. Multi-modal curriculum learning for semi-supervised image classification. *IEEE Transactions on Image Processing*, 25(7):3249–3260, 2016. doi: 10.1109/TIP.2016.2563981.

Grill, J.-B., Strub, F., Altché, F., Tallec, C., Richemond, P. H., Buchatskaya, E., Doersch, C., Pires, B. A., Guo, Z. D., Azar, M. G., Piot, B., Kavukcuoglu, K., Munos, R., and Valko, M. Bootstrap your own latent: A new approach to self-supervised learning, 2020. URL https://arxiv.org/abs/2006.07733.

He, K., Fan, H., Wu, Y., Xie, S., and Girshick, R. Momentum contrast for unsupervised visual representation learning, 2019. URL https://arxiv.org/abs/1911.05722.

Huang, Y., Wang, Y., Tai, Y., Liu, X., Shen, P., Li, S., Li, J., and Huang, F. Curricularface: Adaptive curriculum learning loss for deep face recognition, 2020. URL https://arxiv.org/abs/2004.00288.

Jaiswal, A., Babu, A. R., Zadeh, M. Z., Banerjee, D., and Makedon, F. A survey on contrastive self-supervised learning, 2020. URL https://arxiv.org/abs/2011.00362.

Jiang, L., Meng, D., Zhao, Q., Shan, S., and Hauptmann, A. G. Self-paced curriculum learning. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, AAAI'15, pp. 2694–2700. AAAI Press, 2015. ISBN 0262511290.

Kim, S., Lee, G., Bae, S., and Yun, S.-Y. Mixco: Mix-up contrastive learning for visual representation, 2020. URL https://arxiv.org/abs/2010.06300.

Krizhevsky, A. and Hinton, G. Learning multiple layers of features from tiny images. *Master's thesis, Department of Computer Science, University of Toronto*, 2009.

Kumar, M., Packer, B., and Koller, D. Self-paced learning for latent variable models. In Lafferty, J., Williams, C., Shawe-Taylor, J., Zemel, R., and Culotta, A. (eds.), *Advances in Neural Information Processing Systems*, volume 23. Curran Associates, Inc., 2010. URL https://proceedings.neurips.cc/paper/2010/file/e57c6b956a6521b28495f2886ca0977a-Paper.pdf.

Li, H. and Gong, M. Self-paced convolutional neural networks. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pp. 2110–2116, 2017. doi: 10.24963/ijcai.2017/293. URL https://doi.org/10.24963/ijcai.2017/293.

Lin, L., Wang, K., Meng, D., Zuo, W., and Zhang, L. Active self-paced learning for cost-effective and progressive face identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(1):7–19, 2018. doi: 10.1109/TPAMI.2017.2652459.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks, 2017. URL https://arxiv.org/abs/1706.06083.

Morerio, P., Cavazza, J., Volpi, R., Vidal, R., and Murino, V. Curriculum dropout, 2017. URL https://arxiv.org/abs/1703.06229.

Peng, X., Wang, K., Zhu, Z., and You, Y. Crafting better contrastive views for siamese representation learning. *arXiv preprint arXiv:2202.03278*, 2022.

Robinson, J., Chuang, C.-Y., Sra, S., and Jegelka, S. Contrastive learning with hard negative samples, 2020. URL https://arxiv.org/abs/2010.04592.

Schroff, F., Kalenichenko, D., and Philbin, J. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

Selvaraju, R. R., Desai, K., Johnson, J., and Naik, N. Casting your model: Learning to localize improves self-supervised representations, 2020. URL https://arxiv.org/abs/2012.04630.

Spitkovsky, V. I., Alshawi, H., and Jurafsky, D. From baby steps to leapfrog: How "less is more" in unsupervised dependency parsing. pp. 751–759. The Association for Computational Linguistics, 2010. URL http://dblp.uni-trier.de/db/conf/naacl/naacl2010.html#SpitkovskyAJ10.

Tay, Y., Wang, S., Tuan, L. A., Fu, J., Phan, M. C., Yuan, X., Rao, J., Hui, S. C., and Zhang, A. Simple and effective curriculum pointer-generator networks for reading comprehension over long narratives, 2019. URL https://arxiv.org/abs/1905.10847.

Tian, Y., Sun, C., Poole, B., Krishnan, D., Schmid, C., and Isola, P. What makes for good views for contrastive learning? *arXiv preprint arXiv:2005.10243*, 2020.

Wang, X., Chen, Y., and Zhu, W. A survey on curriculum learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9):4555–4576, 2022. doi: 10.1109/TPAMI.2021.3069908.

Wu, X., Dyer, E., and Neyshabur, B. When do curricula work?, 2020. URL https://arxiv.org/abs/2012.03107.

Wu, Z., Xiong, Y., Yu, S., and Lin, D. Unsupervised feature learning via non-parametric instance-level discrimination, 2018. URL https://arxiv.org/abs/1805.01978.

Yun, S., Han, D., Oh, S. J., Chun, S., Choe, J., and Yoo, Y. Cutmix: Regularization strategy to train strong classifiers with localizable features, 2019. URL https://arxiv.org/abs/1905.04899.

Zhang, H., Cisse, M., Dauphin, Y. N., and Lopez-Paz, D. mixup: Beyond empirical risk minimization, 2017. URL https://arxiv.org/abs/1710.09412.