Center-wise Local Image Mixture For Contrastive Representation Learning

Hao Li[1]
hao_li_96@163.com
Xiaopeng Zhang[2]
zxphistory@gmail.com
Hongkai Xiong[1]
xionghongkai@sjtu.edu.cn

[1] Institute of M.I.N
Shanghai Jiao Tong University
Shanghai, China

[2] Huawei, Inc.
Shanghai, China

Abstract

Contrastive learning based on instance discrimination trains model to discriminate different transformations of the anchor sample from other samples, which does not consider the semantic similarity among samples. This paper proposes a new kind of contrastive learning method, named CLIM, which uses positives from other samples in the dataset. This is achieved by searching local similar samples of the anchor, and selecting samples that are closer to the corresponding cluster center, which we denote as center-wise local image selection. The selected samples are instantiated via an data mixture strategy, which performs as a smoothing regularization. As a result, CLIM encourages both local similarity and global aggregation in a robust way, which we find is beneficial for feature representation. Besides, we introduce *multi-resolution* augmentation, which enables the representation to be scale invariant. We reach 75.5% top-1 accuracy with linear evaluation over ResNet-50, and 59.3% top-1 accuracy when fine-tuned with only 1% labels.

1 Introduction

Recently, self-supervised learning has attracted more attention due to its free of human labels. In self-supervised learning, the network aims at exploring the intrinsic distributions of images via a series of predefined pretext tasks [6, 16]. Among them, contrastive learning pulls closer together the positive pairs composed by different transformations of the same image, *e.g.,* cropping, color distortion, *etc.*, and has been demonstrated to be able to generate features that are comparable with those produced by supervised pretraining [4]. However, contrasting two images that are *de facto* similar in semantic space is not optimal for general representations. It is intuitive to pull semantically similar images for better transferability. DeepCluster [1] and Local Aggregation [28] relax the extreme instance discrimination task via discriminating groups of images instead of an individual image. However, due to the lack of labels, it is inevitable that the positive pairs contain noisy samples, which limits the performance.

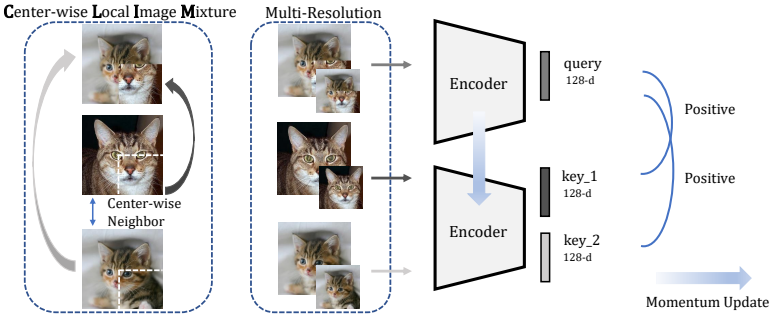

Figure 1: An illustration of the proposed **CLIM** and **Multi-resolution** data augmentations.

In this paper, we target at expanding instance discrimination by exploring local similarities and facilitate transitivity among images. Towards this goal, we need to solve two issues: i) how to select similar images as positive pairs, and ii) how to incorporate these positive pairs, which inevitably contain noisy assignments, into contrastive learning. This paper proposes a new kind of contrastive learning method, named *Center-wise Local Image Mixture*, to tackle the above two issues in a robust way. CLIM consists of two core elements, *i.e.,* a center-wise positive sample selection, as well as a data mixing operation. For center-wise sample selection, we search for nearest neighbors of an image, and only retain similar samples that are closer to the corresponding cluster center. As a result, an image is pulled towards the corresponding center without breaking the local similarity. Considering that the selected positive samples inevitably contain noisy samples, we apply data mixing augmentation as a regularization strategy to avoid predictions with high confidence on selected positive samples. In this way, similar samples are pulled together in a smoother and robust way, which we find is beneficial for general representation.

Furthermore, we propose *multi-resolution* augmentation, which aims at contrasting the same image (patch) at different resolutions explicitly, to enable the representation to be scale invariant. We argue that although previous operations such as crop and resize introduce multi-resolution implicitly, they do not compare the same patch at different resolutions directly. As comparisons, multi-resolution incorporates scale invariance into contrastive learning, and significantly boosts the performance even based on a strong baseline.

We evaluate the feature representation on several self-supervised learning benchmarks. On ImageNet linear evaluation protocol, we achieve 75.5% top-1 accuracy with a standard ResNet-50, and achieve 59.3% top-1 accuracy when finetuned with only 1% labels. We also validate its transferring ability on several downstream tasks, and consistently outperform the fully supervised counterparts.

2 Related Work

Unsupervised Representation Learning. Unsupervised learning aims at exploring the intrinsic distribution of data samples via constructing a series of pretext tasks without human labels. These pretext tasks take many forms and vary in utilizing different properties of images. Among them, one family of methods takes advantage of the spatial properties of images, typical pretext tasks include predicting the relative spatial positions of patches [16], or inferring the missing parts of images by colorization [27], or rotation prediction [6]. Recent

progress in self-supervised learning mainly benefits from instance discrimination, which regards each image (and augmentations of itself) as one class for contrastive learning. The motivation behind these works is the InfoMax principle, which aims at maximizing mutual information [21, 23] across different augmentations of the same image [3, 9, 21].

Contrastive learning does not consider the relationship between different samples. Some related works [8, 14] propose to use different mining strategies to select positive samples. CoCLR [8] proposes to use rgb flow and optical flow to select positive samples alternately and interactively. In [14], the authors propose MIL-NCE to learn a joint embedding space where semantically related videos and texts are close and far away otherwise. Both these two methods use other modalities to facilitate the samples selection. Our proposed method, CLIM selects the positive samples from the nearest neighbors with a direction provided by cluster center, which helps to pull together semantically similar images.

Data Augmentation. Instance discrimination relies on data augmentations, *e.g.,* random cropping, color jittering, horizontal flipping, to define a large set of vicinities for each image. As has been demonstrated in [3, 22], the effectiveness of instance discrimination methods strongly relies on the type of augmentations, hoping that the network holds invariance in the local vicinities of each sample.

Mixing different images is widely used as a data augmentation to help alleviate overfitting in training deep networks. In particular, Mixup [26] combines two samples linearly on pixel level, where the target of the synthetic image was a linear combination of one-hot labels. Following Mixup, there are a few variants [12, 25]. Cutmix [25] cuts out a patch from image, pastes it on another image, and mix their labels according to the area proportion. Attribute Mix [12] mixes the attentive regions to generate new samples.

In contrastive learning, Un-mix [20] uses data mixing augmentation to expand data space, which is totally different from CLIM. Without the restriction of labels, the improvements Un-mix brings are limited. CLIM uses data mixing method to regularise the relationship between positive sample pairs. The data mixing operations are only performed on positive sample pairs.

3 Method

In this section, we start by reviewing contrastive learning for unsupervised representation learning. Then we elaborate our proposed CLIM method, which targets at pulling similar samples via center-wise similar sample selection, followed by a data mixing regularization. We also present multi-resolution augmentation that further improves the performance.

3.1 Contrastive Learning

Contrastive learning targets at training an encoder to map positive pairs to similar representations while pushing away the negative samples in the embedding space. Given unlabeled training set $\boldsymbol{X} = \{x_1, x_2, ..., x_n\}$. Instance-wise contrastive learning aims to learn an encoder f_q that maps the samples $\boldsymbol{X}$ to embedding space $\boldsymbol{V} = \{v_1, v_2, ..., v_n\}$ by optimizing a contrastive loss. Take the Noise Contrastive Estimator (NCE) [17] as an example, the contrastive

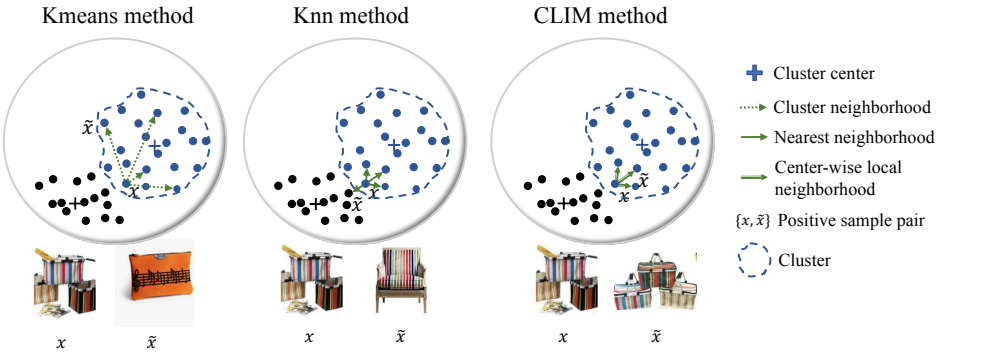

Figure 2: Comparison of three positive sample selection strategies, *i.e.,* K-means, Knn, and the proposed center-wise local sample selection.

loss is defined as:

$$\mathcal{L}_{nce}(x_i, x_i') = -\log \frac{\exp(\frac{f_q(x_i)\cdot f_k(x_i')}{\tau})}{\exp(\frac{f_q(x_i)\cdot f_k(x_i')}{\tau}) + \sum_{j=1}^{N} \exp(\frac{f_q(x_i)\cdot f_k(x_j')}{\tau})} \tag{1}$$

where τ is the temperature parameter, and x_i' and x_j' denote the positive and negative samples of x_i, respectively. The encoder f_k can be shared [2, 3] or momentum update of encoder f_q [9]. N denotes the number of negative samples sampled.

3.2 CLIM: Center-wise Local Image Mixture

In contrastive learning, each sample as well as its transformations are treated as a separate class, while all other samples are regarded as negative examples. In principle, semantically similar samples should be endowed with similar feature representation in the embedding space, while current contrastive methods does not consider the semantic similarities among different samples. To solve this issue, we propose a new contrastive learning method, termed as CLIM, pulls together samples that are semantically similar in an efficient and robust way. The proposed CLIM consists of two elements, *i.e.,* center-wise local similar sample selection, and a data mixing regularization, which would be described in details in the following.

Center-wise Local Positive Sample Selection.

To facilitate transitivity among samples, we propose a positive sample selection strategy that considers both local similarity and global aggregation to expand the neighborhood space of current anchor sample. This is achieved by searching similar samples within a cluster that the anchor sample belongs to, and only retaining samples that are closer to the corresponding cluster center. We denote it as center-wise local selection as these samples are picked out towards the cluster center among the local neighborhood of an image. In this way, similar samples are progressively pulled to the predefined cluster centers, while do not break the local similarity.

Specifically, given a set of unlabeled images $\boldsymbol{X} = \{x_1, x_2, ..., x_n\}$ and the corresponding embedding $\boldsymbol{V} = \{v_1, v_2, ..., v_n\}$ with encoder f_θ, where $v_i = f_\theta(x_i)$. We cluster the representations $\boldsymbol{V}$ using a standard K-means algorithm, and obtain m centers $\mathbf{C} = \{c_1, c_2, ..., c_m\}$.

Given an anchor x_i with its assigned cluster $c(x_i) \in \boldsymbol{C}$, denote the sample set that belongs to $c(x_i)$ as $\boldsymbol{\Omega}_1 = \{x|c(x) = c(x_i)\}$. We search the k nearest neighbors of x_i over the entire space with L2 distance, obtaining sample set $\boldsymbol{\Omega}_2 = \{x_{i1}, ..., x_{ik}\}$. The positive samples are selected based on the following rule:

$$\boldsymbol{\Omega}_p = \{x|d(f_\theta(x), v_{c(x_i)}) \leq d(f_\theta(x_i), v_{c(x_i)}), x \in \boldsymbol{\Omega}_1 \cap \boldsymbol{\Omega}_2\} \tag{2}$$

where $d(\cdot,\cdot)$ denotes the L2 distance of two samples, and $v_{c(x_i)}$ denotes the feature representation of the corresponding cluster center, respectively. In this way, the samples are aggregated towards the predefined clusters, and meanwhile maintaining the local similarity.

Our method combines the advantages of cluster and nearest neighbor methods. An illustration comparing the three methods is shown in Fig. 2. Cluster-based method regards all samples that belong to the same center as positive pairs, which breaks the local similarity among samples especially when the anchor is around the boundary. While nearest neighbor-based method independently pulling samples of an anchor, and does not encourage the well-clustered goal. As a result, the embedding space is not highly concentrated among multiple similar anchors. As comparisons, by center-wise sample selection, similar samples are progressively pulled to the predefined center as well as considering the local similarity. In Fig. 2, some visualizations of positive sample pairs are provided. We used MoCo pretrained model to infer on the validation set and do K-means on the output features. Almost 48% samples in this cluster are from the same category. The remaining 52% of the samples, especially those on the cluster boundaries, have no similarities either in layout or category information. We also use KNN to search Top-K neighbors of every sample on the output features space. It was found that that the nearest neighborhoods of the anchor sample have at least one of these two characteristics: sharing the same category or sharing the similar layout and color information with the anchor sample. CLIM integrates the advantages of KNN and K-means, and selects positive samples sharing similar category, layout and color information with the anchor sample. In the experimental section, we would demonstrate its advantages over simply cluster and knn in feature representation.

Data Mixing Regularization. Once we obtain the positive samples of an anchor, one direct way is to treat these samples similar as the augmented ones for contrastive learning. However, similarity computation in high dimensional space is complex, which will bring the issue of uncertainty and trust-worthiness. To solve this issue, we make use of data mixture strategy, which aims at mixing patches from two different images as augmented samples for contrasting. Data mixing is widely used in supervised learning as data augmentation. Here data mixing is used as a regularization strategy to avoid high confidence for the selected positive samples.

We mix selected positive sample with the anchor sample, and mixed samples obtained can be used to constitute positive pairs $(x_{mix}, \tilde{x}_i)$ and (x_{mix}, x_i). In this way, the selected sample $\tilde{x}_i$ and anchor x_i can be pulled together in a soft manner. Specifically, we conduct data mixing as Cutmix [25], which can be described as:

$$x_{mix} = \mathbf{M} \odot x_i + (\mathbf{1} - \mathbf{M}) \odot \tilde{x}_i \tag{3}$$

where $\mathbf{M} \in \{0,1\}^{W \times H}$ denotes a binary mask indicating the mixed rectangle region of an image, *i.e.,* where to cutout the region in x_i and replaced with a randomly selected patch from $\tilde{x}_i$, and W, H denotes the wide and height of an image, respectively. $\mathbf{1}$ is a binary mask filled with ones, and $\odot$ is the element-wise multiplication operation. For mask $\mathbf{M}$ generation, we

follow the setting in [25]. For the mixed sample x_{mix}, the positive sample can be either x_i or $\tilde{x}_i$, and we reformulate the contrastive learning as combing two NCE loss:

$$\mathcal{L}_{mix}(x_i, \tilde{x}_i) = \lambda \cdot \mathcal{L}_{nce}(x_{mix}, x_i) + (1-\lambda) \cdot \mathcal{L}_{nce}(x_{mix}, \tilde{x}_i) \tag{4}$$

where the combination ratio λ is sampled from beta distribution $\text{Beta}(\alpha, \alpha)$ with parameter α. The advantages are twofold: first, mixed samples help to expand the neighborhood space of current anchor sample for better representation; second, minimizing the two terms simultaneously can help to maximize the mutual information between x_i and $\tilde{x}_i$ in a soft manner and perform as smoothing regularization on the prediction for selected positive samples.

3.3 Multi-resolution Data Augmentation

Data augmentation plays a key role in contrastive learning, crop augmentation is one of the most effective way [3]. In a typical crop augmentation, a sample x with size $H \times W$ is randomly cropped with ratio σ, and resized to $K_{train} \times K_{train}$ as augmented samples, where $K_{train} \times K_{train}$ denotes the input resolution for model training. Hence the scaling factor w.r.t. sample x can be described as:

$$s = \frac{1}{\sigma} \cdot \frac{K_{train}}{\sqrt{H \times W}}. \tag{5}$$

For crop augmentation, the parameter K_{train} is fixed, and the crop ratio σ is randomly selected among positive pairs. As a result, different crop augmentations usually contain different contents, which can be regarded as modeling occlusion invariance to some extent, where each crop sees one view of an image. In this section, we propose a simple but effective data augmentation strategy, named multi-resolution augmentation, which enables the representation to be scale invariant of an example. The highlight is that it is better for contrasting positive pairs with the same content but different resolutions. Specifically, for each positive we keep the crop ratio σ fixed, and adjust K_{train} to different resolutions for contrastive loss. An illustration is shown in Fig. 1 .Using multi-resolution, the objective function can be generalized as:

$$\mathcal{L}_{mr} = \sum_{r,r' \in \{r_1,\ldots,r_n\}} \mathcal{L}_{mix}(x_i^r, \tilde{x}_i^{r'}), \tag{6}$$

where $\{r_1, \ldots, r_n\}$ indicates the resolution set. In this way, the encoder would be encouraged to discriminate the positive samples with different resolutions from a series of negative keys, which will maximize the mutual information between inputs with different resolutions and discard redundant information brought by resolutions.

Comparing with Multi-crop Augmentation. There exist recent works that aim at improving crop augmentations, including multi-crop [2] and jigsaw-crop [25]. However, both methods target at reducing crop ratio σ in Eq.5 and resolution K_{train} simultaneously to bridge different parts of an object, and do not explicitly model scale invariance. As comparisons, our proposed multi-resolution strategy fixes the crop ratio to explicitly model scale invariance. In the experimental section, we would validate the discrepancy of the two augmentation strategies.

Table 1: Top-1 accuracies under linear evaluation on ImageNet, using ResNet-50 as encoder

Method	Accuracy (%)
Supervised	76.5
Colorization [27]	39.6
Jigsaw [16]	45.7
NPID [23]	54.0
LA [28]	58.8
MoCo [9]	60.6
SeLa [24]	61.5
PIRL [15]	63.6
CPCv2 [10]	63.8
PCL [13]	65.9
SimCLR [3]	70.0
MoCo v2 [5]	71.1
SimCLRv2 [4]	71.7
InfoMin [22]	73.0
BYOL [7]	74.3
SwAV [2]	75.3
CLIM	**75.5**

Table 2: Semi-supervised learning with few shot ImageNet labels, using ResNet-50 as encoder (averaged by 5 trials)

Method	Top-1 / Top-5			
	1% labels		10% labels	
Supervised	25.4	48.4	56.4	56.4
PIRL	30.7	57.2	60.4	83.8
SimCLR	48.3	75.5	65.6	87.8
MoCo v2	52.4	78.4	65.3	86.6
BYOL	53.2	78.4	68.8	89.0
SwAV	53.9	78.5	**70.2**	**89.9**
SimCLRv2	57.9	**82.5**	68.4	89.2
CLIM	**59.3**	81.6	70.0	89.3

Table 3: Transfer learning on VOC object detection (averaged by 5 trials)

Method	Accuracy (%)	
	AP_{50}	AP_{75}
Supervised	81.4	58.8
MoCo v2	82.5	64.0
SwAV	82.6	-
CLIM	**82.8**	**64.5**

4 Experimental Results

In this section, we assess our pretrained feature representation on several unsupervised benchmarks. We evaluate it on ImageNet under linear evaluation and semi-supervised settings. Then we transfer the learned features to different downstream tasks. We also analyze the performance of our representation with detailed ablation studies. For brief expression, except for the ablation study, we denote our method as CLIM, which includes two kinds of data augmentations.

4.1 Linear evaluation on ImageNet

The feature representation is trained based on ImageNet 2012 [19], using a standard ResNet-50 structure as backbone. We follow the setting in MoCo v2 [5], and the training details are listed in Appendix. We first evaluate our features by training a linear classifier on top of the frozen representation, following a common protocol in [9, 21]. For linear classifier, the learning rate is initialized as 30 and decayed by 0.1 after 60, 80 epochs, respectively. Table.1 shows the top-1 accuracies with center crop evaluation. Our method achieves an accuracy of 75.5%, surpassing MoCo v2 baseline (71.1%) by 4.4%, and nearly approaching the supervised learning baseline (76.5%).

Table 4: Transfer learning on COCO detection and instance segmentation (5 trials)

Method	Mask R-CNN,R50-FPN,Det						Mask R-CNN,R50-FPN,InsSeg					
	1× schedule			2× schedule			1× schedule			2× schedule		
	AP^{bb}	AP^{bb}_{50}	AP^{bb}_{75}	AP^{bb}	AP^{bb}_{50}	AP^{bb}_{75}	AP^{mk}	AP^{mk}_{50}	AP^{mk}_{75}	AP^{mk}	AP^{mk}_{50}	AP^{mk}_{75}
Supervised	38.9	59.6	42.0	40.6	61.3	44.4	35.4	56.5	38.1	36.8	58.1	39.5
MoCo v2	39.2	59.9	42.7	41.5	62.2	45.3	35.7	56.8	38.1	37.5	59.1	40.1
CLIM	**39.5**	**60.0**	**43.3**	**41.8**	**62.3**	**45.7**	**35.8**	**57.0**	**38.6**	**37.7**	**59.4**	**40.5**

Table 5: Transfer learning on LVIS long-tailed instance segmentation (5 trials)

Method	Object Det			Instance Seg		
	AP^{bb}	AP^{bb}_{50}	AP^{bb}_{75}	AP^{bb}	AP^{mk}_{50}	AP^{mk}_{75}
Supervised	24.1	39.4	25.0	24.2	37.8	25.1
MoCo v2	25.1	40.4	26.1	25.3	38.4	27.0
CLIM	**25.5**	**41.2**	**26.7**	**25.6**	**39.5**	**27.5**

4.2 Semi-supervised training on ImageNet

We also evaluate our method by fine-tuning the pretrained model with a small subset of labels, following the semi-supervised settings in [2, 3, 7, 11]. For fair comparisons, we use the same fixed 1% and 10% splits of training data as in [3], and fine-tune all layers using SGD optimizer with momentum of 0.9, and learning rate of 0.0001 for backbone, 10 for the newly initialized fc layer. The fine-tune epochs is set as 60, and the learning rate is decayed by 0.1 after every 20 epochs. During training, only random cropping and flipping data augmentations are used for fair comparison. The results are reported in Table.2. CLIM achieves 59.3% top-1 accuracy with only 1% labels, and 70.0% with 10% labels. The performance gains are larger with 1% labels, *e.g.,* 6.1% higher than BYOL, and 5.4% better than SwAV, which demonstrates that the proposed feature representation is mainly suitable for extremely few shot learning. Note that SimCLR v2 makes use of other tricks like more MLP layers for better performance, while our method simply adds one fc layer, and still achieves better performance under both settings.

4.3 Downstream tasks

We also evaluate our feature representation on several downstream tasks, including object detection and instance segmentation, to evaluate the transferability of the learned features. For fair comparison, all experiments follow MoCo settings.

PASCAL VOC Object Detection. Following the evaluation protocol in [9], we use Faster R-CNN [18] with R50-C4 as backbone. We fine-tune all layers on the trainval set of VOC07+12 for 2× schedule and evaluate on the test set of VOC2007. We report the performances under the metric of AP50 and AP75. As shown in Table 3, on PASCAL VOC, CLIM achieves 82.8% and 64.5% mAP under AP50 and AP75 metric, which is 1.4 points and 5.7 points higher than the fully supervised counterparts, and is slightly better than the results of MoCo v2.

COCO Object Detection and Instance Segmentation. We also evaluate the representation learned on a large scale COCO dataset. Following [9], we choose Mask R-CNN with

FPN as backbone, and fine-tune all the layers on the train set and evaluate on the val set of COCO2017. In Table.4, we report results under both $1\times$ and $2\times$ schedules. We show that CLIM consistently outperforms the supervised pretrained model and MoCo v2. Under 2X schedule, we achieve 41.8% and 37.7% detection and segmentation accuracies, respectively, which is 1.2 points and 1.1 points better than the supervised couterparts, and also slightly better than the highly optimized MoCo v2.

LVIS Long Tailed Instance Segmentation. Different from VOC and COCO where the number of training samples is comparable, LVIS is a long-tailed dataset, which contains more than 1200 categories, among them some categories only have less than ten instances. The main challenge is to learn accurate few shot models for classes among the tail of the class distribution, for which little data is available. We evaluate our features on this long-tailed dataset to validate how the unsupervised representation boosts the performance. Similarly, we fine-tune the model (Mask R-CNN, R50-FPN) on the train set and evaluate on the val set of Lvis v0.5. Table.5 shows the result under $2\times$ schedule. CLIM outperforms the supervised pretrained model by a large margin and slightly better than MoCo v2. We claim that it is mainly to the proposed data mixing data augmentation, which is able to learn generalized representations even with extremely few labeled data.

4.4 Ablation Study

In this section, we present ablation studies to better understand how each component affects the performance. Unless specified, we train the model for 200 epochs over the ImageNet-1000 and report the top-1 classification accuracy under linear evaluation protocol.

Positive Sample Selection. We first analyze the advantages of our proposed center-wise local sample selection strategy. The compared sample selection alternatives include:

- Random selection: Randomly select a sample from all unlabeled data.
- KNN selection: Use k-nearest neighbors to build the correlation map among samples, and randomly select a sample from the Top-k ($k = 10$) nearest neighbors as positive sample.
- K-means selection: Use k-means clustering algorithm to obtain k cluster centers, and randomly select a sample from the corresponding cluster as positive sample.
- KNN $\cap$ K-means selection: Use K-means clustering algorithm to obtain k cluster centers, and randomly select nearest neighbor within the cluster as positive sample.

The results are shown in the second column of Table.6. In order the inspect the influence of sample selection, we do not conduct cutmix augmentation, and these positive samples are simply pulled via a standard contrastive loss. Noted that, we test many hyper-parameter settings for each sample selection strategy, and choose the best settings for them. It can be shown that comparing with the MoCo baseline, both KNN and cluster-based sample selection boost the performance, while our proposed center-wise selection surpasses these two methods by 1% and 1.3%, respectively.

Data Mixing Augmentation. Data mixing helps to expand the neighborhood space of the target sample, and acts as smoothing regularization for the prediction. As shown in the third column of Table.6, cutmix augmentation consistently improve the performance, comparing with directly pulling similar samples in contrastive loss, and achieve 70.1% accuracy with only 200 training epochs. Notably, with randomly selected positive samples, cutmix operation even obtains 67.1% accuracy, slightly lower than the MoCo baseline, while significantly better than no mixing with only 62.3% accuracy. This can be attributed to the smoothing regularization of cutmix, which is able to alleviate the effect of noisy samples and update model in a more robust way.

Table 6: Impact of different sample selection

Strategy	Accuracy (%)	
	no mixing	+cutmix
MoCo v2	67.5	-
Random	62.3	67.1
KNN	68.3	69.5
K-means	68.0	69.2
KNN ∩ K-means	68.5	69.6
Center-wise	**69.3**	**70.1**

Table 7: Impact of different multiple resolutions

Method	Resolution	Accuracy (%)
Multi-Crop	2×224 + 2×96	69.7
Multi-Reso	$r, r' \in \{224, 96\}$	70.4
	$r, r' \in \{224, 128\}$	71.7
	$r, r' \in \{224, 160\}$	**72.3**
	$r, r' \in \{224, 224\}$	71.4

Multiple Resolution. Based on CLIM, we further add multi-resolution data augmentation to validate its effectiveness. The results of introducing different resolutions are shown in Table.7. Using multiple resolutions setting with $r, r' \in \{224, 160\}$, our method achieves an accuracy of 72.3% with only 200 epochs, which surpasses the baseline of MoCo by 4.8%, and even much better than the results of MoCo with 800 epochs (71.1%). We also compare our multi-resolution augmentation with multi-crop augmentation proposed in [2]. $2 \times 224 + 2 \times 96$ denotes using two 224×224 crops with crop-scale $\sigma \sim U(0.2, 1.0)$ and two 96×96 crops with $\sigma \sim U(0.05, 0.14)$, referring to [2]. We find that multi-crop slightly deteriorates the performance of CLIM (70.1% versus 69.7%), partially because data mixing behaves like image cropping augmentation, and shares similarity with multi-crop strategy.

More Ablation Studies. We provide more ablation studies for reference in the Appendix. Detailed comparisons include 1) Number of clusters m and k in KNN, 2) Hyper-parameters α in Cutmix , 3) Different choices of mixing methods and 4) Ablation study on longer training schedule.

5 Conclusion

In this work, we proposed CLIM data augmentation. Center-wise positive sample selection considers both local similarity and global aggregation property. In such way, similar samples are progressively aggregated to a series of predefined clusters, while not breaking the local similarity. Data mixing augmentation acts as a smoothing regularization for contrastive loss between neighborhood space. Furthermore, we present a simple but effective multi-resolution augmentation, which explicitly model scale invariance to further improve the representation. Experiments evaluated on several unsupervised benchmarks demonstrate the effectiveness of our method.

References

[1] 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.

[2] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Ar-

mand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *arXiv preprint arXiv:2006.09882*, 2020.

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

[4] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey Hinton. Big self-supervised models are strong semi-supervised learners. *arXiv preprint arXiv:2006.10029*, 2020.

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

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

[7] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent: A new approach to self-supervised learning. *arXiv preprint arXiv:2006.07733*, 2020.

[8] Tengda Han, Weidi Xie, and Andrew Zisserman. Self-supervised co-training for video representation learning. *arXiv preprint arXiv:2010.09709*, 2020.

[9] 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.

[10] Olivier J Hénaff, Aravind Srinivas, Jeffrey De Fauw, Ali Razavi, Carl Doersch, SM Eslami, and Aaron van den Oord. Data-efficient image recognition with contrastive predictive coding. *arXiv preprint arXiv:1905.09272*, 2019.

[11] Simon Kornblith, Jonathon Shlens, and Quoc V Le. Do better imagenet models transfer better? In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2661–2671, 2019.

[12] Hao Li, Xiaopeng Zhang, Qi Tian, and Hongkai Xiong. Attribute mix: semantic data augmentation for fine grained recognition. In *2020 IEEE International Conference on Visual Communications and Image Processing (VCIP)*, pages 243–246. IEEE, 2020.

[13] Junnan Li, Pan Zhou, Caiming Xiong, Richard Socher, and Steven CH Hoi. Prototypical contrastive learning of unsupervised representations. *arXiv preprint arXiv:2005.04966*, 2020.

[14] Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. End-to-end learning of visual representations from uncurated instructional videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9879–9889, 2020.

[15] 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.

[16] 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.

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

[18] 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*, pages 91–99, 2015.

[19] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115 (3):211–252, 2015.

[20] Zhiqiang Shen, Zechun Liu, Zhuang Liu, Marios Savvides, Trevor Darrell, and Eric Xing. Un-mix: Rethinking image mixtures for unsupervised visual representation learning. *arXiv preprint arXiv:2003.05438*, 2020.

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

[22] 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.

[23] 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.

[24] Asano YM., Rupprecht C., and Vedaldi A. Self-labelling via simultaneous clustering and representation learning. In *International Conference on Learning Representations (ICLR)*, 2020.

[25] 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 International Conference on Computer Vision*, pages 6023–6032, 2019.

[26] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.

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

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