PIECING AND CHIPPING: AN EFFECTIVE SOLUTION FOR THE INFORMATION-ERASING VIEW GENERATION IN SELF-SUPERVISED LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

In self-supervised learning frameworks, deep networks are optimized to align different views of an instance that contains the similar visual semantic information. The views are generated by conducting series of data augmentation to the anchor samples. Although the data augmentation operations are often designed to be aggressive and extensive to lower the mutual information between views, the family of Information-Erasing data augmentation that masks out region of images is barely considered. In this work, we propose the Piecing and Chipping enhanced Erasing Augmentation (PCEA) approach to making the self-supervised learning algorithms benefit from the effectiveness of Information-Erasing data augmentation. Specifically, we design a pipeline to generate mutually weakly related transformed views using random erasing and build corresponding loss terms to take advantage of these views. Extensive experiments demonstrate the effectiveness of our method. Particularly, applying our PCEA to MoCo v2 improves the baseline by 12.84%, 3.3% in terms of linear classification on ImageNet-100 and ImageNet-1K.

1 INTRODUCTION

The deep convolutional neural networks (CNNs) (Krizhevsky et al., 2012) have a great success in computer vision tasks, and in recent years, self-supervised learning (Oord et al., 2018; Chen et al., 2020b;c; He et al., 2020; Li et al., 2021; Zbontar et al., 2021; Grill et al., 2020; Chuang et al., 2020; Hu et al., 2020; Kim et al., 2020; Zhu et al., 2020; Caron et al., 2020; Xiao et al., 2020; Kalantidis et al., 2020) also achieve a great success and gained attentions because of its ability of reducing the labor cost on large-scale dataset annotation. Self-supervised learning aims at learning some forms of image representations by figuring out a pattern that can explain the image reasonably. The learned pattern can be used in downstream tasks, such as image classification, object detection, segmentation and etc. The self-supervised learning can be achieved majorly in two different styles: contrastive (Chen et al., 2020b;c; He et al., 2020; Chuang et al., 2020) and non-contrastive (Li et al., 2021; Zbontar et al., 2021; Grill et al., 2020) (though the detailed taxonomy of self-supervised learning is the topic of this study). The key component of both styles is the generation of views of the anchor sample.

The term "view" in self-supervised learning is roughly grounded as "augmented or transformed samples that maintain semantically similar information to the anchor sample". In the computer vision tasks, the generation of views is accomplished a series of domain transformation operations, *e.g.* `ColorJitter`, `RandomGrayscale`, `GaussianBlur`. Former literature has examined the influence of the adaptation of different types of transformation. In these works, the composition of transformation operations is considered as the crucial part for learning good representations (Chen et al., 2020b). And the proper approach to reduce the mutual information between views while keeping task-relevant information intact (Tian et al., 2020).

One family of data augmentation that is commonly employed in computer vision tasks is "information-erasing". By which, we refer to the methods that mask small regions of an image, such that the information concerning the objects in the image is erased (DeVries & Taylor, 2017; Yun et al., 2019; French et al., 2020; Singh & Lee, 2017; Chen et al., 2020a). However, this family of data augmentation is barely seen in self-supervised learning algorithms. While in (Chen et al., 2020b), the researchers also denote the Cutout (DeVries & Taylor, 2017) as an unflavored augmentation method to generate

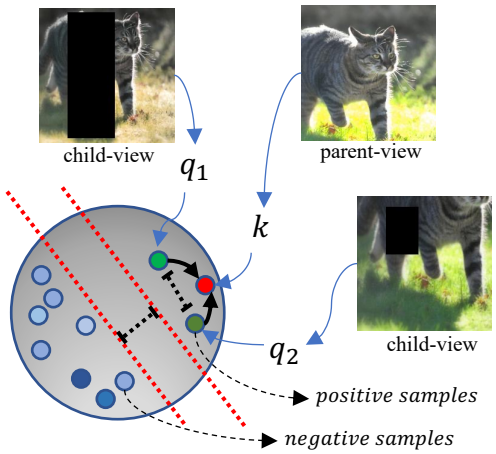

Figure 1: Semantic representation with fine-grained instance discrimination method.

views. We conjecture the primary reason for the inferior performance of the information-erasing family in the self-supervised learning algorithms is the inconsistency in preserving the task-relevant information. At the same time, information-erasing methods do not contribute to the reduction of the mutual information between views and anchor images in the non-masked regions. As a consequence, the generated views could be valueless for feature extractors to learning semantically meaningful representations.

In this work, we tackle the aforementioned drawbacks of inconsistency and mutual information reduction. We build an approach with the simple random erasing method to provide stable views with high qualities to improve the performance of self-supervised learning algorithms. We refer to our approach as Piecing and Chipping enhanced Erasing Augmentation (PCEA), which is built upon four motivations:

1. Multiple instances of erasing augmented images are generated and pieced, and we chip the larger image irregularly such that views would be weakly related by acquiring peripheral patches from other views;
2. We resize the irregularly chipped views without preserving the aspect ratio to reduce mutual information in the non-mask regions;
3. We feed more than one view (two in this work) to the "positive pair" loss head of the self-supervise algorithms to lessen the inconsistency brought by random selection of masked regions.
4. Considering the above approach for the view generation, we also regularize the predicted similarity between these views. Thus, we could largely prevent the non-task-relevant information from being memorized.

In simple terms, we spawn weakly related child-views that are similar to their parent-view while being considerably different from each other. The overall approach is shown in Figure 1 and Figure 2. The dark and light blue spots denote the negatives samples from different images. The red (k) and green ones (q_1 and q_2) indicate the positive pairs. The proposed method aims to enlarge the margin between the blue/non-blue spots and the distance between the spots in green color (q_1 and q_2). For positives in red and green, the margin between the red (k) and each green spot (q_1 or q_2) is narrowed.

In our experimental analysis, we firstly compare the effectiveness of the proposed approach with other information-erasing family data augmentation. We keep the comparison fair by offering multiple child-views for all the augmentation methods as demonstrating in Figure 1. We show that the piecing-and-chipping-based random erasing augmentation out-performs other well-designed augmentation methods by a large margin. We also conduct experiments compared with other state-of-the-art self-supervised learning algorithms. Specifically, we employ MoCo v2 (Chen et al., 2020d) as our backbone and modify its view generation codes with the proposed PCEA. We then achieve a

competitive performance on the linear-probe classification task using the ImageNet-1K datasets. Overall, the **main contributions** of this work can be summarized as follows:

- We propose Piecing and Chipping enhanced Erasing Augmentation (PCEA), a novel data augmentation approach for the view generation in self-supervised learning algorithms.
- The proposed PCEA data augmentation approach also offers a novel method of utilizing multiply child-views. The method not only reduces the inconsistency in the view generation process but also regularizes the utilization of non-task-relevant information during the self-supervised learning progress.
- We conduct extensive experiments to demonstrate the effectiveness of our method. To the best of our knowledge, this is the first successful attempt in involving the Information-Erasing family data augmentation in self-supervised learning algorithms.

2 Related Work

2.1 self-supervised learning

A wide range of self-supervised learning algorithms has been proposed to improve the quality of learned representations. Recent self-supervised learning algorithms can be divided into two categories: *Non-contrastive* ones that employ positive pairs of sample; *Contrastive* ones that employ negative pairs of samples. Here the terms positive/negative do not strictly refer to pairs of sample with similar/different semantic information, but pairs of views generated from the same or different anchor samples. In the family of non-contrastive self-supervised learning, BYOL (Grill et al., 2020) achieves an outstanding performance, which relies on two neural networks to represent the visual semantic information; the online and target network interact and learn from each other. SimSiam architecture (Chen & He, 2020) aims at enlarging the similarity between the two augmented views of one image with a shared encoder network. On the other hand, typical contrastive self-supervised learning applies multi-layer perceptions and stop-gradient tricks in case of collapsing (Chen et al., 2020b). To reduce the memory cost of large amount of negative samples, MoCo (He et al., 2020) proposes a momentum memory bank to record negative samples of previous steps. SWAV (Caron et al., 2020) is an online algorithm, which improves the contrastive method without the pairwise comparison. An online clustering loss is constructed, and a multi-crop strategy is introduced to increase the number of views without the extra computational overhead. In this study, we employ both of the self-supervised learning algorithm families to verify the effectiveness and efficiency of our proposed method.

2.2 Data augmentation in self-supervised learning

Data augmentation in vanilla computer vision tasks helps to improve performance by increasing the amount of training data. Specifically, in practical implementation, this technology helps the model find the indistinguishable features in the image, that can reduce the over-fitting of the model like a regularizer. However, in the scenarios of self-supervised learning, the data augmentation plays a much different role. In the SimCLR paper (Chen et al., 2020b), the author carefully examine the effects of different data augmentation *w.r.t.* the downstream classification tasks. In their conclusion, the Gaussian blur for the input images and a stronger color distortion act as critical roles in obtaining an effective predicted result. SimCLR has experimentally demonstrated that the ImageNet linear classification accuracy at Top-1 is increased from 59.6% to 63.2% by stronger color distortion strength. This conclusion is further confirmed in Chen et al. (2020c), which shows that the accuracy of MoCo v1 with extra blur augmentation is increased by 2.8% to 63.4%. Furthermore, Tian et al. (2020) argues the proper data augmentation should reduce the mutual information between views while keeping task-relevant information, and develops the more aggressive `info-min` data augmentation approach. However, we consider the regular induced data augmentations are still limited in the desire of fully using the semantic information of visual representation in self-supervised learning. In this work, we focus on the family of data augmentation that masks out semantic information straightforwardly.

2.3 Information Erasing Data augmentation

In paper (Noroozi & Favaro, 2016), a puzzle-based data augmentation method is developed with an unsupervised visual representation manner, which builds a CNN to solve **Jigsaw puzzles** as a pretext

task for enhancing classification and detection performance. In paper (DeVries & Taylor, 2017), a method named **"CutOut"** is designed for the objective classification task, which randomly masks square regions of training images and tries to find out less prominent features. These two methods can be regarded as early explorations of advanced data augmentation for object classification and detection-related tasks. With their convenience and efficiency, these two methods reached the highest level of computer vision-related tasks at that time and profoundly influenced other methods. However, this type of single splicing and deletion of images or image parts also limits the performance of the models.

In the object localization area, a weakly supervised framework named **"Hide-and-Seek"** is proposed in the paper (Singh & Lee, 2017), which randomly hides patches of the images and enhances the model. In this method, not only the most discriminative part of the image can be identified, but other parts with weak discriminative can also be identified. Through the overall organization of each part in the image, the discriminative performance of the model is improved. Another method, **MixUp** is designed in paper (Zhang et al., 2017), which aims to provide an image data augmentation idea with a convex combination of the training data. With the state-of-art performance in several tasks such as ImageNet2021, CIFAR-10, and CIFAR-100, the method Mixup inspires a potential clue for unsupervised, semi-supervised, and reinforcement learning. Different from the traditional regional dropout or patch removal methods, a **CutMix** data augmentation method is proposed in paper (Yun et al., 2019), which cuts patches and pasted them among training images with ground truth labels to enhance the reliability and stability of the model. These three methods provide new ideas for data augmentation, and the methods based on them also archived the highest level at that time. However, these methods still do not completely get rid of the relatively inflexible processing methods for images or patches, such as the proportion of the original image and the shape of the patches, which also restricts the performance of the model.

Based on these studies above, a regional dropout strategy is designed as **GridMask** in paper (Chen et al., 2020a), which provides a controllable method to delete patches of a training image. Compared with previous methods, this structured information dropping method is more effective and avoids random information dropping. At the same time, to overcome the shortages of squared patches in previous studies, a Gaussian filter-based data augmentation method **"Milking CowMask"** is proposed in paper (French et al., 2020). This method provides more flexibly shaped masks according to turnable parameters in Gaussian filter with fewer correlations and reaches a new state-of-art performance in related tasks. However, these methods discussed above focus on increasing the discriminability of samples in the entire dataset, which results in limited performance in the self-supervised learning cases.

3 METHODOLOGY

3.1 PCEA: PIECING AND CHIPPING ENHANCED ERASING AUGMENTATION

In this section, we first introduce how the views are generated in the proposed Piecing and Chipping enhanced Erasing Augmentation (PCEA) method. The overall approach is depicted in Figure 2.

We refer 2 pipelines of image transformation operations as T_1 and T_2. T_2 is an ordinary adopted data augmentation method used in state-of-the-art self-supervised learning algorithms (in this paper, we employ the data augmentation strategy in MoCov2 (Chen et al., 2020d)). T_1 is based on T_2, with an additional masking operation (in this paper, we employ random erasing). The PCEA method is described as follows:

- `Step 1:` For each image $x \in \mathbb{R}^{w \times h \times c1}$, we generate views $x\_v_{1,2,3,4}$ and x_k using T_1 and T_2, respectively. The $x\_v_{1,2,3,4}$ are denoted as "child-views", while x_k is denoted as the "parent-view".
- `Step 2:` We piece the 4 different child-views $x\_v_{1,2,3,4}$ (224×224) to obtain a larger image (448×448). This newly generated image is considered as an alternative image to obtain more positive samples with more substantial semantic information.

[1]For the rest of the paper, we let the $w, h = 224$ and omit the channel notation c, for sake of good readability.

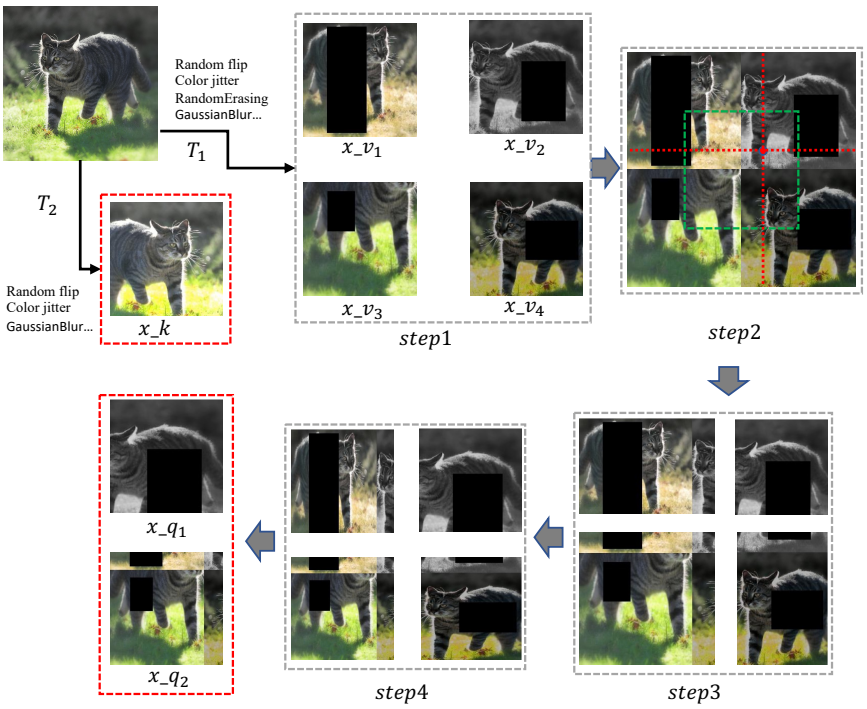

Figure 2: The process of the proposed Piecing and Chipping enhanced Erasing Augmentation.

- `Step 3`: We locate a candidate region (green rectangle in the figure) at the centroid of the newly generated image. [2] We then (uniformly) randomly select a segmentation point in the candidate region, and chip the image vertically and horizontally. Thus we obtain a new set of child-views $x\_q_{1,2,3,4}$.
- `Step 4`: The set of new child-views are resized to their original size (224×224), without preserving the aspect ration. We finally select 2 child-views as the new positive pairs of their parent-view x_k.

3.2 Similarity Regularization loss

Although x_{q_1} and x_{q_2} can still be roughly judged as identical by human beings, the randomness of in choosing the erasing and the change of aspect ratio develop a considerable margin between the semantic information of visual representation in x_{q_1} and x_{q_2}. Meanwhile, the ordinary `InfoNCE` loss aligns both child-views to their parent view. To prevent the deep model implicitly aligns the child-views, we put an additional Similarity Regularization (SimReg) loss term to attain explicit discrimination between them. This loss term is implemented with a simple cosine similarity between the embedded representations of the child-views. According to this, the loss between q_1 and q_2 is defined as (1).

$$\mathcal{L}_{\text{SimReg}} = \frac{q_1 \cdot q_2}{\max(\|q_1\|_2, \|q_2\|_2)} \tag{1}$$

In the experimental analysis, we find that the loss term is insensitive to the loss weight (so-called λ in many literature) empirically. Therefore, we leave the loss weight hyper-parameter to be 1.0 in all experimental configurations.

[2] Here we set the size of candidate region to be same as the 'child-views' (224×224), more details are discussed in the ablation study.

4 EXPERIMENTS

4.1 DATASETS & EXPERIMENTAL CONFIGURATIONS

Datasets. In this work, we conduct experiments on ImageNet ILSVRC-2021 dataset (Deng et al., 2009) with 1.28 million images in 1000 categories (**ImageNet-1K**) and a subset of images in 100 categories (**ImageNet-100**), which have been widely utilized as benchmark datasets (Tian et al., 2019; He et al., 2020; Grill et al., 2020; Hu et al., 2020). We also construct a more difficult subset of the original ImageNet-1K dataset, named Small-ImageNet-1000 (**S-ImageNet-1K**). S-ImageNet-1K only selects 10 percent of the images from each of the categories, which aims at reducing the richness of visual representations while maintaining the same representation distribution as the original ImageNet-1K. The evaluation is carried out by training a linear probe for the classification task, while keeping the weights of feature extractor frozen. For ImageNet-1K and ImageNet-100, we employ the commonly adopted classification accuracy as the evaluation metric. For S-ImageNet-1K, we employ the average correct classification rate among all 1000 categories proposed in Le & Yang (2015) as our evaluation metric.

In addition, we also employ the widely acknowledged MsCOCO (Lin et al., 2014) to verify the proposed PCEA with the object detection task. We fine-tune the ImageNet-1K pre-trained backbone models using the `train2017` split, and perform evaluation on the `val2017` split.

Configurations. We employ the vanilla ResNet-50 (He et al., 2016) equipped with an global average pooling on its head as our backbone architecture. We employ the projection head with only one linear layer for encoding f_θ and f_ε. The feature dimensions of the output of ResNet-50 pooling layer and the embedding vector are 2048 and 128, respectively. For other hyper-parameters, we keep the same configuration as in MoCo v2 (Chen et al., 2020d) and SimSiam (Chen & He, 2021). In the MoCo v2 algorithm, the augmented views x_q_1 and x_q_2 are fed into the encoder network f_θ with back-propagation. Meanwhile, x_k is represented as $k = f_\varepsilon(x\_k)$ without back-propagation, where $f_\varepsilon(\cdot)$ denotes the momentum encoder. In the SimSiam algorithm, we simply average the similarity between multiply child-views and the parent views. For the detection task, we adopt the commonly used Faster-RCNN with ResNet-50 as the baseline architecture.

Training. During training, a mini-batch size of 256 is used in 8 GPUs (Tesla V100 16G), and the initial learning rate is defined as 0.03. SGD (Loshchilov & Hutter, 2016) is used as the optimizer, the weight decay and the momentum update parameter is defined as 0.0001 and 0.9. 200/100 epochs are trained with a cosine learning rate decay for MoCo v2 and SimSiam, respectively. The number of negative samples in momentum queue and the sliding queue are 65536 and 32768, respectively. The temperature is set as 0.2.

4.2 EXPERIMENTAL RESULTS

ImageNet-100. Following previous work (Chen et al., 2020d; Chen & He, 2021), we evaluate nine data augmentation methods on MoCo v2 (Chen et al., 2020d) and SimSiam (Chen & He, 2021), where linear classifier are trained on frozen features from these methods. The comparison results are reported in Table 1. As can be seen, applying PCEA to MoCo v2 with negative samples involved achieves the best performance against baselines using other data augmentation methods. Particularly, our PCEA outperforms the vanilla baseline by 12.84% and 3.85% in terms of top-1 and top-5 accuracy. This demonstrates the effectiveness of our PCEA in learning discriminative representations by treating the positive and negative samples separately. We can also observe that SimSiam (Chen & He, 2021) with our PCEA achieves superior performance on the ImageNet-100 dataset against previous data augmentations, which further validates the generalizability of our PCEA to existing contrastive self-supervised methods.

ImageNet-1K. Furthermore, we compare our PCEA with existing state-of-the-art self-supervised methods under the linear classification setting in Table 2. From the results, we can observe that our PCEA outperforms MoCo v2, the vanilla baseline by a large margin, *i.e.,* 3.3% in terms of top-1 accuracy. Meanwhile, we also achieve competitive results with previous methods in terms of top-1 and top-5 accuracy, which further demonstrates the advantage of our PCEA over baselines under the same linear classification setting.

Table 1: Top-1/top-5 accuracy for linear classification on ImageNet-100 via applying nine data augmentations to MoCo v2 and SimSiam, where models are trained on frozen features from different methods. Bold and underline numbers denote the first and second place.

Method	Data Aug.	Param.(M)	Batch	Epochs	Top-1(%)	Top-5(%)
MoCo v2	-	28	256	200	81.65	95.77
	Jigsaw Puzzles	28	256	200	78.41	94.69
	CutOut	28	256	200	82.64	95.84
	Hide-and-Seek	28	256	200	82.72	95.87
	MixUp	28	256	200	84.08	96.79
	CutMix	28	256	200	83.51	96.51
	Random Erasing	28	256	200	81.04	95.27
	Grid Mask	28	256	200	80.35	94.42
	Milking CowMask	28	256	200	<u>85.32</u>	<u>97.35</u>
	PCEA (ours)	28	256	200	**94.49**	**99.62**
SimSiam	-	28	256	100	72.32	91.35
	Jigsaw Puzzles	28	256	100	68.15	90.13
	CutOut	28	256	100	72.26	91.17
	Hide-and-Seek	28	256	100	74.38	91.87
	MixUp	28	256	100	72.16	91.03
	CutMix	28	256	100	73.55	92.06
	Random Erasing	28	256	100	73.42	92.01
	Grid Mask	28	256	100	70.47	90.75
	Milking CowMask	28	256	100	<u>75.26</u>	<u>93.22</u>
	PCEA (ours)	28	256	100	**84.25**	**96.71**

Table 2: Comparisons between PCEA and other methods under the linear classification evaluation. For fair comparison, all results are trained under the same architecture on ImageNet-1K training set and validation set. Parameters are of the feature extractor He et al. (2020). 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 (%)	Top-5 (%)
InstDisc	ResNet-50	24	256	200	2x224	58.5	-
LocalAgg	ResNet-50	24	128	200	2x224	58.8	-
MoCo	ResNet-50	24	256	200	2x224	60.6	-
MoCo v2	ResNet-50	24	256	200	2x224	67.5	-
CMC	ResNet-50	47	128	240	2x224	66.2	87.0
SimCLR	ResNet-50	24	256	200	2x224	61.9	-
PCL v2	ResNet-50	24	512	200	2x224	67.6	-
CPC v2	ResNet-50	24	512	200	2x224	63.8	85.3
PIC	ResNet-50	24	512	200	2x224	67.6	-
MoCHi	ResNet-50	24	512	200	2x224	68.0	-
AdCo	ResNet-50	24	256	200	2x224	68.6	-
SwAV	ResNet-50	24	4096	200	2x224	69.1	-
BYOL	ResNet-50	24	256	200	4x224	70.6	-
SimSiam	ResNet-50	24	256	200	4x224	70.0	-
MoCo v2 + PCEA (ours)	ResNet-50	24	256	200	3x224	**70.8**	**90.2**

Table 3: Top-1 accuracy for linear classification on S-ImageNet-1K, where models are trained on frozen features from different methods. Bold numbers denote the first place.

Method	**Param.(M)**	**Batch**	**Epochs**	**Top-1(%)**
MoCo v2	28	256	200	42.3
SwAV	28	256	200	53.6
BYOL	28	256	200	54.1
MoCo v2 + PCEA (ours)	28	256	200	**57.3**

S-ImageNet-1K. Table 3 reports the comparison results of linear classification on our S-ImageNet-1K dataset, a smaller dataset with the same distribution of the original ImageNet-1K, but lacks richness in visual representations. The proposed PCEA with MoCo v2 out-performs its baseline algorithm with a large margin (15.0%) in terms of top-1 accuracy. This superior performance validates the effectiveness and efficiency of PCEA in difficult configurations.

MsCOCO. Table 4 reports the detection performance (mAP) on the Ms COCO datasets. The proposed PCEA with MoCo v2 achieve the best result compared to state-of-the-art self-supervisely pre-trained backbones. Specifically, it outperforms its baseline MoCo v2 by 1.1% and supervisely trained model by 4.3%.

Table 4: The object detection result on MsCOCO 2017 datasets. all the models use Faster RCNN-ResNet-50 and are finetuned using the 1x schedule.

Pre-trained Model	**#Epochs**	**mAP**
Supervised	200	35.2
Supervised	300	38.2
MoCo-v2	300	39.3
SwAV	300	38.4
Barlow Twins	300	39.2
MoCo v2 + PCEA (ours)	200	39.5

5 Ablation Study

In this section, we conduct extensive ablation studies to explore how each step of our PCEA and the size of candidate region affect the final performance of our approach. Unless specified, we perform the experiments on ImageNet-100 dataset.

5.1 Ablation on Each Step of PCEA

Table 5: Ablation study on each step of PCEA on ImageNet-100 dataset.

Step 1	Step 2 & Step 3	Step 4	Top-1 (%)	Top-5 (%)
✗	✗	✗	81.65	95.77
✓	✗	✗	87.15	97.87
✓	✓	✗	90.76	98.21
✓	✓	✓	**94.49**	**99.62**

In order to explore the effect of each step of our PCEA on the final performance, we ablate each step and report the experimental results in Table 5. The methods with different steps are analysed, which describe the effectiveness of each steps in the Mosaic process. The top-1 accuracy on ImageNet-100 with the same data augmentation processing as MoCo v2 (Chen et al., 2020d) is 81.65% with two inputs(k and q). After that, the result benefits an promotion with two inputs as x_q_1 and x_q_2 in step 1, which increases the performance by 5.5%. As for the combination of step 2 and step3, we

Table 6: Ablation study on the size of candidate region on ImageNet-100 dataset.

Size	Top-1 (%)	Top-5 (%)
28*28	88.07	98.17
56*56	88.65	98.33
112*112	89.41	98.57
224*224	**94.49**	**99.62**
336*336	88.80	98.49
448*448	88.56	98.44

Table 7: The ablation experiments on different child-views and SimReg loss.

Datasets	# Child-Views	SimReg loss	Top-1	Top-5
S-ImageNet-1K	1	✗	42.3	64.7
S-ImageNet-1K	2	✗	53.4	76.4
S-ImageNet-1K	3	✗	48.3	71.4
S-ImageNet-1K	2	✓	**57.3**	**80.2**
ImageNet-1K	2	✗	70.3	90.1
ImageNet-1K	2	✓	70.8	90.3

use padding and random crop to modify the output images instead of resizing the splitted images into 224*224, which achieves a higher accuracy as 90.76%. Adding step 4 to previous three steps boosts the top-1 and top-5 accuracy to 94.49% and 99.62%, which indeed validates the rationality of interpolation in our PCEA to capture the fine-grained instance features.

5.2 Ablation on the Size of Candidate Region

To analyze how the size of the candidate region affect the final performance of our PCEA, we vary the size from 28, 56, 112, 224, 336, 448. The comparison results are reported in Table 6. As can be seen, our PCEA with the size of 224×224 achieves the best performance compared to other size settings. With the increase of the size of the candidate region, the performance of our PCEA degrades a lot, which could be caused by more background information introduced in the selected region. In the meanwhile, when the size of the candidate region is decreased to 112×112, our PCEA performs worse than the best result in terms of top-1 and top-5 accuracy. This further shows the importance of choosing the right size of the candidate region to learn more discriminative representations during pre-training.

5.3 Ablation on number of views in loss terms

We modify the number of child-views participated in the self-supervise learning loss (`InfoNCE` for MoCo, `CosSim` for SimSiam). The loss terms are duplicated and averaged according to the number child-views. We also conduct experiments on the effects of the SimReg loss term. Table 7 reports these results on both S-ImageNet-1K and ImageNet-1K. It can be seen that, two child-views achieves the best performance among different configurations. On the other hand, the SimReg loss functions overwhelmingly in the difficult S-ImageNet-1K dataset.

6 Conclusion

In this work, we propose Piecing and Chipping enhanced Erasing Augmentation (PCEA), a novel approach to employ information-erasing family of data augmentation methods in self-supervised learning scenarios. We exploit eight existing information-erasing data augmentation over previous methods on commonly-used benchmark datasets. We also equip the PCEA on 2 popular self-supervised learning baseline algorithm. Both results prove that the effectiveness and efficiency of the proposed PCEA approach. We believe the involvement of information-erasing family of data augmentation has a border impact on further developing of self-supervised learning algorithm.

REFERENCES

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.

Pengguang Chen, Shu Liu, Hengshuang Zhao, and Jiaya Jia. Gridmask data augmentation. *arXiv preprint arXiv:2001.04086*, 2020a.

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)*, 2020b.

Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. *arXiv preprint arXiv:2011.10566*, 2020.

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

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

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

Ching-Yao Chuang, Joshua Robinson, Yen-Chen Lin, Antonio Torralba, and Stefanie Jegelka. Debiased contrastive learning. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

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

Geoff French, Avital Oliver, and Tim Salimans. Milking cowmask for semi-supervised image classification. *arXiv preprint arXiv:2003.12022*, 2020.

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.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

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*, pp. 9729–9738, 2020.

Qianjiang Hu, Xiao Wang, Wei Hu, and Guo-Jun Qi. Adco: Adversarial contrast for efficient learning of unsupervised representations from self-trained negative adversaries. *arXiv preprint arXiv:2011.08435*, 2020.

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.

Sungnyun Kim, Gihun Lee, Sangmin Bae, and Se-Young Yun. Mixco: Mix-up contrastive learning for visual representation. In *Neural Information Processing Systems (NeurIPS) Workshop on Self-Supervised Learning*, 2020.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105, 2012.

Ya Le and Xuan Yang. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7):3, 2015.

Yazhe Li, Roman Pogodin, Danica J Sutherland, and Arthur Gretton. Self-supervised learning with kernel dependence maximization. *arXiv preprint arXiv:2106.08320*, 2021.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016.

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

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

Krishna Kumar Singh and Yong Jae Lee. Hide-and-seek: Forcing a network to be meticulous for weakly-supervised object and action localization. In *2017 IEEE international conference on computer vision (ICCV)*, pp. 3544–3553. IEEE, 2017.

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

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.

Tete Xiao, Xiaolong Wang, Alexei A Efros, and Trevor Darrell. What should not be contrastive in contrastive learning. *arXiv preprint arXiv:2008.05659*, 2020.

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*, pp. 6023–6032, 2019.

Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. *arXiv preprint arXiv:2103.03230*, 2021.

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

Benjin Zhu, Junqiang Huang, Zeming Li, Xiangyu Zhang, and Jian Sun. Eqco: Equivalent rules for self-supervised contrastive learning. *arXiv preprint arXiv:2010.01929*, 2020.