CRAFTING BETTER CONTRASTIVE VIEWS FOR SIAMESE REPRESENTATION LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent self-supervised contrastive learning methods greatly benefit from the Siamese structure that aims at minimizing distances between positive pairs. For high performance Siamese representation learning, one of the keys is to design good contrastive pairs. Most previous works simply apply random sampling to make different crops of the same image, which overlooks the semantic information that may degrade the quality of views. In this work, we propose *ContrastiveCrop*, which could effectively generate better crops for Siamese representation learning. Firstly, a semantic-aware object localization strategy is proposed within the training process in a fully unsupervised manner. This guides us to generate contrastive views which could avoid most false positives (*i.e.* object *v.s.* background). Moreover, we empirically find that views with similar appearances are trivial for the Siamese model training. Thus, a center-suppressed sampling is further designed to enlarge the variance of crops. Remarkably, our method takes a careful consideration of positive pairs for contrastive learning with negligible extra training overhead. As a plug-and-play and framework-agnostic module, *ContrastiveCrop* consistently improves SimCLR, MoCo, BYOL, SimSiam by $0.4\% \sim 2.0\%$ classification accuracy on CIFAR-10, CIFAR-100, Tiny ImageNet and STL-10. Superior results are also achieved on downstream detection and segmentation tasks when pre-trained on ImageNet-1K.

1 INTRODUCTION

Self-supervised learning (SSL) has attracted much attention in the computer vision community due to its potential of exploiting large amount of unlabeled data. As a mainstream approach in SSL, contrastive learning has achieved higher performance on several downstream tasks (*e.g.*, object detection, segmentation and pose estimation (Ren et al. (2015), He et al. (2017), Güler et al. (2018), Everingham et al. (2010), Lin et al. (2014))) than its supervised counterpart. Such promising results can be largely attributed to the Siamese structure, which is commonly applied in state-of-the-art unsupervised methods, including SimCLR (Chen et al. (2020a)), MoCo V1 & V2 (He et al. (2020), Chen et al. (2020b)), BYOL (Grill et al. (2020)) and SimSiam (Chen & He (2021)). Typically, the Siamese structure takes two augmented views from an image as input, and minimizes their distance in the embedding space. With proper views selected, Siamese networks demonstrate a strong capability to learn generic visual features (Tian et al. (2020)).

One of the key issues of contrastive learning is to design positives selection. Some works generate different positive views by strong data augmentation, such as color distortion and jigsaw transformation (Tian et al. (2020), Chen et al. (2021)). Another work (Shen et al. (2020)) applies mixture in an unsupervised manner to produce positive pairs of multiple samples. Additionally, different from data augmentation, Zhu et al. (2021) creates hard positives with transformation at the feature level. Despite different techniques, these works commonly apply *RandomCrop* to sample multiple views of an image, and further make the views more diverse. As a basic sampling method, *RandomCrop* enables all individual crops to be selected equiprobably. However, it fails to look at the semantic information of paired views, which helps to learn better representations more efficiently and accurately. As shown in Fig.1(a), random crops are prone to miss the object when no prior of object (*e.g.* scale and location) is given. Optimizing the distance between object and background in the embedding space would mislead the learning of representations. Besides, Fig.1(c) indicates that random crops cannot always carry sufficient variances of an object. Such views with large similarity are trivial for learning discriminative models.

In this paper, we propose *ContrastiveCrop*, aiming to craft better contrastive pairs for Siamese representation learning. False positives indicate that a better sampling strategy for contrastive learning should consider the content information of the image. Hereby, we propose a semantic-aware localization scheme. The module serves as a guidance to select crops, avoiding most false positives, as shown in Fig. 1(b). Moreover, we propose a center-suppressed sampling strategy to tackle trivial positive pairs with large similarity. Fig. 1(d) shows that our crops are more likely to cover different parts of the object. The semantic-aware localization and center-suppressed sampling scheme can be gracefully combined to generate better crops for contrastive learning.

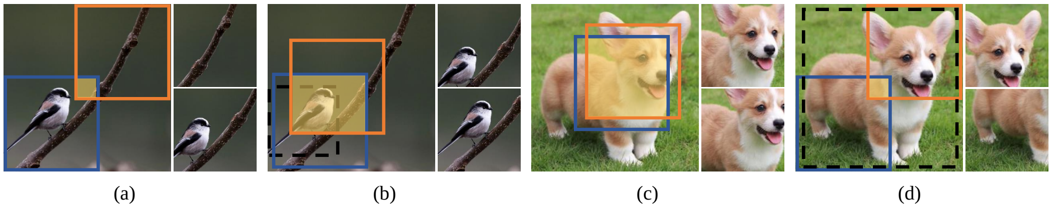


Figure 1: The motivation of our proposed *ContrastiveCrop*. (a) and (c) are generated by typical *RandomCrop*, while (b) and (d) are crops from our method. We address the false positive problem (object *v.s.* background) shown in (a) by localizing the object and restricting the crop center within the bounding box (the black dashed box) in (b). Moreover, we enlarge the variance of crops in (d) by keeping them away from the center, which avoids the close appearance as shown in (c).

The proposed *ContrastiveCrop* considers both semantic infomation and maintaining large variance when making pairs. As a plug-and-play method, it can be easily applied into the Siamese structure. More importantly, our approach is agnostic to contrastive frameworks, regardless using negative samples or not. With negligible training overhead, our strategy consistently improves SimCLR, MoCo, BYOL, SimSiam by $0.4\% \sim 2.0\%$ classification accuracy on CIFAR-10, CIFAR-100, Tiny ImageNet and STL-10. Superior results are also achieved on downstream detection and segmentation tasks when pre-trained on ImageNet-1K.

The main contributions of this paper can be summarized as:

- To the best of our knowledge, this is the first work to investigate the problem of commonly used *RandomCrop* in contrastive learning. We propose *ContrastiveCrop* that is customized to generate better views for this task.
- In *ContrastiveCrop*, the semantic-aware localization is adopted to avoid most false positives and the center-suppressed sampling strategy is applied to reduce trivial positive pairs.
- *ContrastiveCrop* consistently outperforms *RandomCrop* with popular contrastive methods on a variety of datasets, showing its effectiveness and generality for Siamese representation learning.

2 Related Work

In this section, we introduce contrastive learning and positives selection related to this work.

2.1 Contrastive Learning

The core idea of contrastive learning is to pull positive pairs closer while pushing negatives apart in the embedding space. This methodology has shown great promise in learning visual representations without annotation (Bachman et al. (2019), Hénaff et al. (2019), Wu et al. (2018), Misra & Maaten (2020), Oord et al. (2018), Ye et al. (2019)). More recently, contrastive methods based on the Siamese structure achieve remarkable performance on downstream tasks (Chen et al. (2020a), He et al. (2020), Chen et al. (2020b), Grill et al. (2020), Chen & He (2021), Xie et al. (2021a), Wang et al. (2021), Xie et al. (2021b), Dwibedi et al. (2021)), some of which even surpass supervised models. The milestone work is SimCLR (Chen et al. (2020a)), which presents a simple framework for contrastive visual representation learning. It significantly improves the quality of learned representations with a non-linear transformation head. Another famous work is MoCo (He et al. (2020)), which uses a memory bank to store large number of negative samples and smoothly updates it with momentum for better consistency. Methods that learn useful representations without negative samples are also proposed. BYOL (Grill et al. (2020)) trains an online network to predict the output of the target network, with the latter slowly updated with momentum. The authors hypothesize that the additional projector to the online network and the momentum encoder are important to avoid collapsed solutions without negative samples. SimSiam (Chen & He (2021)) further explores simple Siamese networks that can learn meaningful representations without negative sample pairs, large batches and momentum encoders. The role of stop-gradient is emphasized in preventing collapsing.

2.2 Positives Selection

One of the key issues in contrastive learning is the design of positives selection. An intuitive approach to generating positive pairs is to create different views of a sample. Tian et al. (2020) propose an *InfoMin principle* to catch a sweet point of mutual information between views, and accordingly generate positive pairs with its *InfoMin Augmentation*. Different from data augmentation, Zhu et al. (2021) create hard positives by repelling paired representations in the feature space. Additionally, Dwibedi et al. (2021) apply a support set to search nearest neighbors of positives and use them as cross-sample positives. These works commonly apply *RandomCrop* as the basic sampling method to generate views. We find that it may not be the optimal solution for contrastive learning. Therefore, we propose *ContrastiveCrop* that is tailored to make better views for contrastive learning.

3 ContrastiveCrop for Siamese Representation Learning

In this section, we introduce *ContrastiveCrop* for Siamese representation learning. Firstly, We briefly review *RandomCrop* as preliminary knowledge. Then, we describe semantic-aware localization and center-suppressed sampling as two submodules of our *ContrastiveCrop*. Finally, favorable properties of our method are further discussed for better understanding.

3.1 Preliminary: *RandomCrop*

RandomCrop, an efficient data augmentation method, has been widely used in both supervised learning and self-supervised learning (SSL). Here, we briefly review this technique, using API in Pytorch[1] as an example. Given an image, we first determine the scale s and aspect ratio r of the crop from a pre-defined range. Then, the height and width of the crop can be obtained with s and r. Finally, the location of the crop is randomly selected, as long as the whole crop lies within the image. The procedure of *RandomCrop* can be formulated as

$$(x, y, h, w) = \mathbb{R}_{crop}(s, r, I), \tag{1}$$

where $\mathbb{R}_{crop}(\cdot, \cdot, \cdot)$ is the random sampling function that returns a quaternion (x, y, h, w) representing the crop. We denote I as the input image, (x, y) as the coordinate of the crop center, and (h, w) as the height and width of crop. Usually, the scale s and aspect ratio r of crops are set flexibly, so that crops of variant sizes could be made.

In principle, *RandomCrop* enables all individual crops to be selected and could provide diverse views of a sample. However, it performs sampling equiprobably, ignoring the semantic information of images. As shown in Fig. 1(a), *RandomCrop* is prone to generate false positives when the scale of object is small. Given objects with variant scales in contrastive learning, *RandomCrop* would inevitably generate false positives due to lack of the consideration of semantic information. Similarly, optimizing the false positives in Fig. 3 may mislead the learning of representation. Therefore, designing a semantic-aware sampling strategy for crops is crucial and vital for Siamese representation learning.

[1] https://pytorch.org/vision/stable/transforms.html

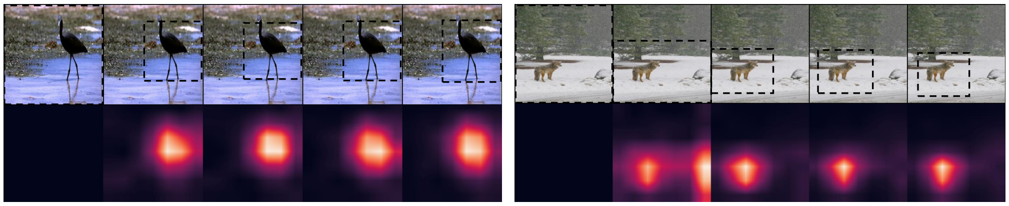


Figure 2: The training dynamic of localization is shown from left to right in each subfigure. We initialize the localization box as the whole image, and update it at a regular interval using the latest heatmap. Note that our goal is not to derive precise localization, but to guide generation of crops by finding the object of interest.

3.2 Semantic-aware Localization

To tackle the issue of poor content understanding in *RandomCrop*, we design a semantic-aware localization module that can effectively reduce false positives in an unsupervised manner. To better study the process of feature learning in Siamese networks, we visualize the heatmaps generated at different training stages (*e.g.*, 0th, 20th, 40th, 60th, 80th epoch) in Fig. 2. Note that we derive the heatmap by summing the features of last convolutional layer across the channel dimension and normalizing it to [0, 1]. There are several inspirations from visualization: 1) The Siamese representation learning framework is capable of capturing the location of the object, which can be leveraged to guide the generation of better crops; 2) Heatmaps can roughly indicate the object, but may need some warm-up at early stages.

Based on above analyses, we propose to detect the object during the training process using the information in heatmaps. Specifically, *RandomCrop* is applied at early stage of training to collect semantic information of the whole image. Then, we apply an indicator function to obtain the bounding box of object B from heatmaps, which can be written as,

$$B = L(\mathbb{1}[M > k]), \tag{2}$$

where M represents heatmap, k is the threshold of activations, $\mathbb{1}$ is the indicator function and L calculates the rectangular closure of activated positions. After obtaining the bounding box B, the semantic crops could be generated as follows,

$$[\dot{x}, \dot{y}, \dot{h}, \dot{w}] = \mathbb{R}_{crop}(s, r, B), \tag{3}$$

where the definitions of $\dot{x}$, $\dot{y}$, $\dot{h}$, $\dot{w}$, s, r, and $\mathbb{R}_{crop}$ are similar to Eq. 1. Considering the coarse localization, we enlarge the operable region by only constraining center of crops within B. Note that the bounding box is progressively updated at a regular interval during the training process. We make comparison between *RandomCrop* and *RandomCrop + Semantic-aware Localization* in Fig. 3. One can find that the false positives reduce dramatically when the proposed module is applied.

3.3 Center-suppressed Sampling

The semantic-aware localization scheme serves as a useful guidance to reduce false positive cases, but increases the probability of making trivial pairs due to the smaller operable region. In this subsection, we introduce the center-suppressed sampling that aims to tackle this dilemma. The idea is to reduce the probability of crops gathering around center. Specifically, we adopt the symmetric beta distribution $\beta(\alpha, \alpha)$, which allows to easily control its shape with different α. As the goal is to enlarge the variance, we set $\alpha < 1$ which gives a lower probability near center and higher at other

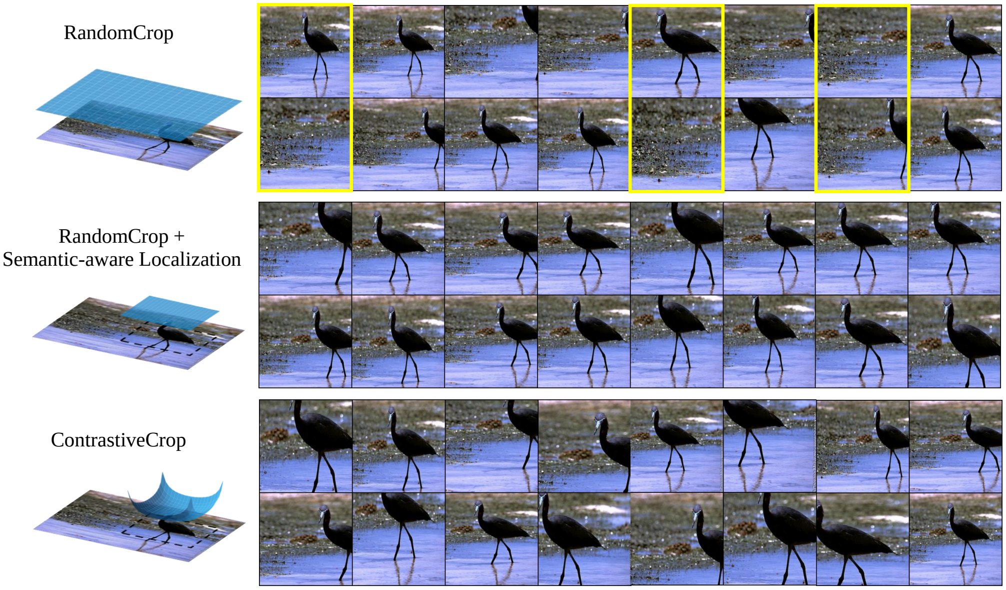


Figure 3: Visualization of *RandomCrop*, *RandomCrop* + *Semantic-aware Localization* and our *ContrastiveCrop*. We show the sampling distributions and operable regions for three settings on the left, and correspondent sampled pairs on the right. Pairs made by *RandomCrop* include several false positives that totally miss the object (marked in yellow box). Using *RandomCrop* with *Semantic-aware Localization* reduces false positives, but introduces easy positive pairs that share large similarity. Last, our *ContrastiveCrop* could reduce false positive pairs while increasing variance at the same time.

positions. In this way, crops are more likely to be scattered within the operable region and large overlap could be significantly avoided.

Algorithm 1: *ContrastiveCrop* for Siamese Representation Learning

```
Input: Image $I$, Crop Scale $s$, Crop Ratio $r$, Threshold of Activations $k$, Parameter of $\beta$ Distribution $\alpha$
    $h = \sqrt{s \cdot r}$    // Height of the crop.
    $w = \sqrt{s/r}$    // Width of the crop.
    $F = Forward(I)$    // Get the features of last layer by forwarding the image.
    $M = Normalize(F)$    // Get the heatmap by normalizing features across channel dimension.
    $B = L(\mathbb{1}[M > k])$    // Get the bounding box by Eq. 2.
    $x = B_{x0} + (B_{x1} - B_{x0}) \cdot u, u \sim \beta(\alpha, \alpha)$ // Sample coordinate $x$ of crop center from $\beta$ distribution
    $y = B_{y0} + (B_{y1} - B_{y0}) \cdot v, v \sim \beta(\alpha, \alpha)$ // Sample coordinate $y$ of crop center from $\beta$ distribution
Output: Crop $C = (x, y, h, w)$
```

Combining center-suppressed sampling with semantic-aware localization, we can finally formulate our *ContrastiveCrop* as

$$(\dot{x}, \dot{y}, \dot{h}, \dot{w}) = \mathbb{C}_{crop}(s, r, B), \tag{4}$$

where $\mathbb{C}_{crop}$ denotes sampling function that uses a center-suppressed distribution and B is the same bounding box as in Eq. 3. Note that β distribution with $\alpha < 1$ is not the only choice for sampling. Other distributions with similar shape (*e.g.*, quadratic function) could also achieve the same goal. The effect of our *ContrastiveCrop* is visualized in Fig. 3. Comparing to *RandomCrop*, our method significantly reduces false positive pairs due to the semantic-aware localization. Meanwhile, it introduces larger variance within a pair by applying the center-suppressed distribution. We show the pipeline for *ContrastiveCrop* in Algorithm 1. The whole module is agnostic to other transformations and can be easily integrated into general contrastive learning frameworks.

3.4 Discussion

To better understand the behavior of *ContrativeCrop*, we discuss several properties that may contribute to its effectiveness. We first investigate the relation between semantic information and variance. We calculate the class scores of single crop that may indicate richness of categorized semantic information. Distance between representations of positive pairs is also obtained to represent variance. Both the class score and distance are statistical results from a standard supervised model (*i.e.*, ResNet-50) with large number of trials. Their relation is shown in Fig. 4. *ContrastiveCrop* conveys more semantic information than *RandomCrop* at the same level of variance, showing the effectiveness of semantic-aware localization. Furthermore, with equal semantic information, *ContrastiveCrop* achieves larger variance than *RandomCrop*, which can be owed to center-suppressed sampling.

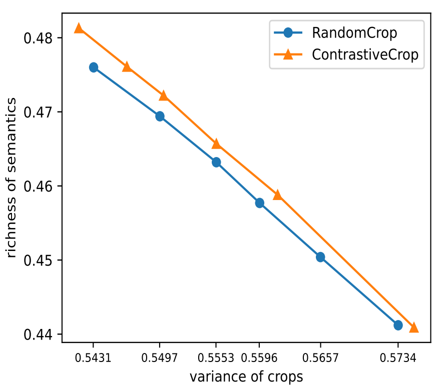


Figure 4: The relation between richness of semantics and variance of crops. Both scores are statistical results from supervisedly trained ResNet-50 with large number of trials. Our *ContrastiveCrop* conveys more semantic information than *RandomCrop* at the same level of variance, and yields larger variance with equal semantic information.

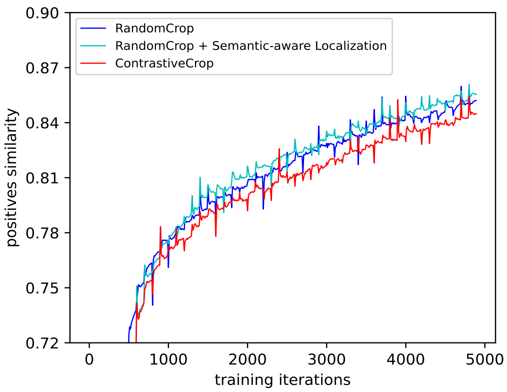


Figure 5: Similarity of positive pairs in training. Smaller positives similarity indicates harder positive samples which may enhance the training (Zhu et al. (2021)). Take *RandomCrop* as baseline, adding only localization results in slightly larger similarity. Our *ContrastiveCrop* combines both semantic-aware localization and center-suppressed sampling, which effectively reduces similarity of positives.

We further visualize the similarity of positive pairs in the training process. As shown in Fig. 5, adding only semantic-aware localization to *RandomCrop* slightly increases similarity, as localization restrains crops in a smaller operable region. Our *ContrastiveCrop* incorporates center-suppressed sampling, showing smaller positives similarity than the other two. This indicates positive pairs sampled by ours are harder ones, which could help learning more view-invariant features as suggested in FT (Zhu et al. (2021)). However, different from FT that reduces positives similarity in the feature space, we directly sample harder crops from raw data, while taking a careful consideration of semantic information.

4 Experiments

In this section, we conduct extensive experiments with popular contrastive methods on a variety of datasets, to demonstrate the effectiveness and generality of our method. We first introduce the datasets and contrastive methods in Sec. 4.1. Sec. 4.2 describes the implementation details. We then evaluate our method with the common linear evaluation protocol in Sec. 4.3. Results of ablation experiments are shown in Sec. 4.4. Finally, Sec. 4.5 presents transfer performance on object detection and segmentation tasks.

4.1 Datasets & Baseline Approaches

We perform evaluation of our method with state-of-the-art unsupervised contrastive methods, on a wide range of datasets. The datasets include **CIFAR-10/CIAFR-100** (Krizhevsky et al. (2009)), **Tiny ImageNet**, **STL-10** (Coates et al. (2011)) and **ImageNet** (Russakovsky et al. (2015)). Generally, these datasets are built for object recognition and the images contain iconic view of objects. The baseline contrastive methods include SimCLR, MoCo V1 & V2, BYOL and SimSiam.

4.2 Implementation Details

Our *ContrastiveCrop* aims to make better views for contrastive learning, which is agnostic to unsupervised learning frameworks. We strictly keep the same training setting when making comparison. Larger gains could be expected with further hyper-parameter tuning, which is not the focus of this work.

For small datasets (*i.e.*, CIFAR-10/100, Tiny ImageNet and STL-10), we use the same training setup in *all* experiments. At the pre-train stage, we train ResNet-18 (He et al. (2016)) for 500 epochs with a batch size of 512 and a cosine-annealed learning rate of 0.5. In our method, we set $k = 0.1$ for the threshold of activations and $\alpha = 0.1$ for sampling. Localization boxes are updated at a frequency of every 100 epochs, which adds negligible extra training overhead. The linear classifier is trained for 100 epochs with initial learning rate of 10.0 multiplied by 0.1 at 60th and 80th epochs.

For experiments on ImageNet, we adopt ResNet-50 as the base model. Pre-train settings of MoCo and SimSiam exactly follow their original works. We reproduce SimCLR with a smaller batch size of 512 and cosine-annealed learning rate of 0.05. We set the $k = 0.1$, $\alpha = 0.6$ for all experiments with 100 epochs and 200 epochs. *ContrastiveCrop* starts from the 20th epoch, after which boxes are updated every 10 epochs. We adopt the same setting as (He et al. (2020)) for training the linear classifier for all baseline methods.

All the experiments are conducted on an 8-GPU server. We use SGD optimizer with momentum of 0.9, weight decay of 10^{-4} and 0 for pre-train and linear evaluation respectively.

Method	CIFAR-10		CIFAR-100		Tiny ImageNet		STL-10	
	R-Crop	*C-Crop*	*R-Crop*	*C-Crop*	*R-Crop*	*C-Crop*	*R-Crop*	*C-Crop*
SimCLR (Chen et al. (2020a))	89.63	**90.08**	60.30	**61.91**	45.19	**46.21**	88.95	**89.53**
MoCo (He et al. (2020))	86.73	**88.78**	56.10	**57.65**	47.09	**47.98**	89.17	**89.81**
BYOL (Grill et al. (2020))	91.96	**92.54**	63.75	**64.62**	46.08	**47.23**	91.84	**92.42**
SimSiam (Chen & He (2021))	90.96	**91.48**	64.79	**65.82**	43.03	**44.54**	89.39	**89.83**

Table 1: Linear classification results for different contrastive methods and datasets. *R-Crop* and *C-Crop* mean *RandomCrop* and *ContrastiveCrop* respectively. We adopt ResNet-18 as the base model and reproduce all the methods with a unified training setup as described in Sec. 4.2.

4.3 Linear Classification

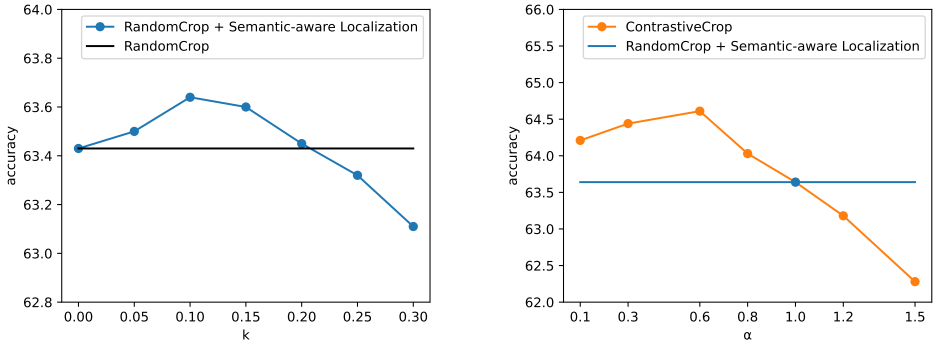
In this section, we verify our method with linear classification following the common protocol. We freeze pre-trained weights of the encoder and train a supervised linear classifier on top of it. Top-1 classification accuracy results on the validation set are reported.

Results on CIFAR-10/100, Tiny ImageNet and STL-10. Our results on these small datasets are shown in Tab. 1. With the same training setup for *all* experiments, *ContrastiveCrop* consistently improves baseline methods by at least 0.4%. Results show that the proposed method is generic and does not require heavy parameter tuning. The localization boxes are updated 5 times in the whole training of 500 epochs, adding negligible training overhead.

Results on ImageNet. The results of ImageNet are two-part: 1) standard ImageNet-1K (IN-1K), which is used for pre-training. 2) IN-200, which consists of 200 random classes of IN-1K and is used for ablation experiments. As shown in Tab. 2, our method outperforms *RandomCrop* with

SimCLR, MoCo V1, MoCo V2, SimSiam on IN-1K by 0.25%, 1.09%, 0.49% and 0.33% respectively. A larger improvement is seen on IN-200. The consistent gain over baseline methods shows the effectiveness and generality of *ContrastiveCrop* for contrastive methods.

4.4 Ablation Studies



(a) Accuracy *v.s.* k

(b) Accuracy *v.s.* α

Figure 6: Ablation results on IN-200 w.r.t. k and α. Subfigure (a) compares *RandomCrop + Semantic-aware Localization* (blue plot) with the *RandomCrop* baseline (black plot). In subfigure (b), we fix the best $k = 0.1$ for localization (blue plot) and compare it with *ContrastiveCrop* to study the influence of different α.

In ablation studies, we investigate the semantic-aware localization module and center-suppressed sampling independently. We also study the effect of *ContrastiveCrop* when it is combined with different transformations. We conduct experiments with ResNet-50 and report the linear classification results on IN-200.

Semantic-aware Localization. In this work, the unsupervised semantic-aware localization serves as a guidance to make crops. We study the influence of k that determines the size of the localization box, and compare it to *RandomCrop* that does not use localization (*i.e.*, $k = 0$). Experimental results are shown in Fig. 6(a). One can find that the localization box can slightly outperforms *RandomCrop* baseline (black plot) when k is kept at a low level from 0.05 to 0.2. This shows the effectiveness of largely removing false positives. However, as k increases over 0.2, the performance starts to fall quickly. We argue that smaller bounding box dramatically reduces variance of views, making it trivial to learn discriminative features.

Center-suppressed Sampling. We use β distribution for the center-suppressed sampling, which allows to control its variance with different α. Here we investigate the impact of different variance by iterating over multiple α. Results are shown in Fig. 6 (b) with $k = 0.1$ for localization. When $\alpha < 1$, our *ContrastiveCrop* consistently outperforms *RandomCrop* with localization, showing the effect of center-suppressed sampling. We also study $\alpha > 1$ that has a smaller variance than uniform distribution (*i.e.*, $\alpha = 1$). A drop in accuracy is observed with $\alpha > 1$. This indicates that larger variance of crops is required for better contrast.

***ContrasitveCrop* with Other Transformations.** To further compare the effect of *ContrastiveCrop* and *RandomCrop*, we study other image transformations used in MoCo V2 (Chen et al. (2020b)), including *Flip*, *ColorJitter*, *Grayscale* and *Blur*. The ablation results are shown in Tab. 3. In case all other transformations are removed, *ContrastiveCrop* is 0.4% higher than *RandomCrop*, which is a direct evidence of our superiority. Moreover, with only one extra transformation, *ContrastiveCrop* outperforms *RandomCrop* by $0.3\% \sim 0.8\%$. The largest gap of 1.2% is achieved when all of the transformations are incorporated. These results show that our *ContrastiveCrop* is compatible and orthogonal to other transformations.

Method	IN-200		IN-1K	
	R-Crop	*C-Crop*	*R-Crop*	*C-Crop*
SimCLR	62.14	**63.08**	61.60	**61.85**
MoCo V1	64.52	**65.80**	57.25	**58.34**
MoCo V2	63.43	**64.61**	64.40	**64.89**
SimSiam	62.89	**63.54**	65.62	**65.95**

Table 2: Linear classification results (100 epochs) on IN-200 and IN-1K. We use exactly the same training setup for comparison of a method.

Flip	*ColorJitter + Grayscale*	*Blur*	*R-Crop*	*C-Crop*
✓	✓	✓	63.4	**64.6**
✓			50.4	**50.9**
	✓		60.6	**61.4**
		✓	44.9	**45.2**
			45.5	**45.9**

Table 3: Ablation of other transformations used in MoCo V2. We combine *ColorJitter* and *Grayscale* as one color transformation. The results are from ResNet-50 pre-trained on IN-200 for 100 epochs.

pre-train	IN-1K	VOC detection			COCO instance seg.			COCO detection		
	Top-1	AP	AP_{50}	AP_{75}	AP^{mk}	AP^{mk}_{50}	AP^{mk}_{75}	AP^{bb}	AP^{bb}_{50}	AP^{bb}_{75}
random init	-	33.8	60.2	33.1	29.3	46.9	30.8	26.4	44.0	27.8
supervised	76.1	53.5	81.3	58.8	33.3	54.7	35.2	38.2	58.2	41.2
infomin	70.1	57.6	82.7	64.6	34.1	55.2	36.3	39.0	58.5	42.0
MoCoV1 (He et al. (2020))	60.6	55.9	81.5	62.6	33.6	54.8	35.6	38.5	58.3	41.6
MoCoV1 + *ContrastiveCrop*	**61.1**	**56.3**	**82.1**	62.6	**33.9**	**55.2**	**36.1**	**38.6**	58.2	41.6

Table 4: Fine-tuning results on PASCAL VOC detection and COCO detection and instance segmentation. All models are pre-trained for 200 epochs on ImageNet-1K. On VOC, the training and evaluation sets are `trainval2007+2012` and `test2007`, on COCO are the `train2017` and `val2017`. All models are fine-tuned for 24K iterations on VOC and 90K on COCO.

4.5 Downstream Tasks

In this section, we measure the transferability of our method on the object detection and instance segmentation task. Following previous works (He et al. (2020), Zhu et al. (2021), we pre-train ResNet-50 on IN-1K for 200 epochs. For downstream tasks, we use PASCAL VOC (Everingham et al. (2010)) and COCO (Lin et al. (2014)) as our benchmarks and we adopt the same setups as in MoCo's detectron2 codebase[2]. All layers of pre-trained models are fine-tuned end-to-end at target datasets.

PASVAL VOC Object Detection. The detector is Faster R-CNN (Ren et al. (2015)) with a backbone of R50-C4 (He et al. (2017)). We fine-tune the model on the `trainval2007+2012` split and evaluate on the VOC `test2007`. The results are present in Tab. 4. Compared with MoCo V1 baseline, our method achieves improvement of +0.4AP and +0.5AP_{50}.

COCO Object Detection/Segmentation. The model is Mask R-CNN (He et al. (2017)) with R50-C4 backbone. We fine-tune 90K iterations on the `train2017` set and evaluate on `val2017`. As shown in Tab. 4, the proposed *ContrastiveCrop* achieved superior performance in most metrics.

5 Conclusion

In this work, we propose *ContrastiveCrop*, that is tailored to make better contrastive views for Siamese representation learning. *ContrastiveCrop* adopts semantic-aware localization to avoid most false positives and applies the center-suppressed sampling to reduce trivial positive pairs. We innovatively take semantic information into account when transforming a sample, and thoroughly investigate the suitable variance for contrastive learning. We have shown the effectiveness and generality of our method through extensive experiments with state-of-the-art contrastive methods including SimCLR, MoCo, BYOL and SimSiam. Finally, we hope this work could inspire future research in designing positives selection to further exploit the potential of contrastive learning.

[2] https://github.com/facebookresearch/detectron2

REFERENCES

Philip Bachman, R Devon Hjelm, and William Buchwalter. Learning representations by maximizing mutual information across views. In *NeurIPS*, pp. 15509–15519, 2019.

Pengguang Chen, Shu Liu, and Jiaya Jia. Jigsaw clustering for unsupervised visual representation learning. In *CVPR*, pp. 11526–11535, 2021.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, pp. 1597–1607, 2020a.

Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. *CVPR*, 2021.

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

Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *AISTATS*, pp. 215–223, 2011.

Debidatta Dwibedi, Yusuf Aytar, Jonathan Tompson, Pierre Sermanet, and Andrew Zisserman. With a little help from my friends: Nearest-neighbor contrastive learning of visual representations. *arXiv preprint arXiv:2104.14548*, 2021.

Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 88(2):303–338, 2010.

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.

Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. Densepose: Dense human pose estimation in the wild. In *CVPR*, pp. 7297–7306, 2018.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, pp. 2961–2969, 2017.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, pp. 9729–9738, 2020.

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.

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

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 *ECCV*, pp. 740–755. Springer, 2014.

Ishan Misra and Laurens van der Maaten. Self-supervised learning of pretext-invariant representations. In *CVPR*, pp. 6707–6717, 2020.

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

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NeurIPS*, 2015.

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. *IJCV*, 2015.

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

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.

Xinlong Wang, Rufeng Zhang, Chunhua Shen, Tao Kong, and Lei Li. Dense contrastive learning for self-supervised visual pre-training. *CVPR*, 2021.

Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *CVPR*, 2018.

Enze Xie, Jian Ding, Wenhai Wang, Xiaohang Zhan, Hang Xu, Zhenguo Li, and Ping Luo. Detco: Unsupervised contrastive learning for object detection. *arXiv preprint arXiv:2102.04803*, 2021a.

Zhenda Xie, Yutong Lin, Zheng Zhang, Yue Cao, Stephen Lin, and Han Hu. Propagate yourself: Exploring pixel-level consistency for unsupervised visual representation learning. *CVPR*, 2021b.

Mang Ye, Xu Zhang, Pong C Yuen, and Shih-Fu Chang. Unsupervised embedding learning via invariant and spreading instance feature. In *CVPR*, pp. 6210–6219, 2019.

Rui Zhu, Bingchen Zhao, Jingen Liu, Zhenglong Sun, and Chang Wen Chen. Improving contrastive learning by visualizing feature transformation. *arXiv preprint arXiv:2108.02982*, 2021.