Crafting Better Contrastive Views for Siamese Representation Learning

Xiangyu Peng[1*] Kai Wang[1*] Zheng Zhu[2] Mang Wang[3] Yang You[1†]
[1]National University of Singapore [2]Tsinghua University [3]Alibaba Group
{xiangyupeng, kai.wang, youy}@comp.nus.edu.sg
zhengzhu@ieee.org wangmang.wm@alibaba-inc.com
Code: https://github.com/xyupeng/ContrastiveCrop

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* vs. *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 [16, 18, 21, 27, 32]) 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 [5], MoCo [20], BYOL [17] and SimSiam [8]. 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 [37].

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 distor-

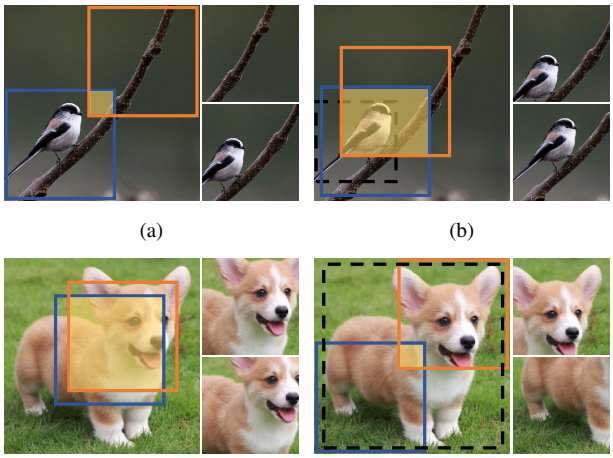


Figure 1. The motivation of our proposed *ContrastiveCrop*. (a) and (c) are generated by typical *RandomCrop*, while (b) and (d) are crops from our method. We address the false positive problem (object *vs*. background) shown in (a) by localizing the object and restricting the crop center within the 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).

*Equal contribution.
†Corresponding author.

tion and jigsaw transformation [4, 37]. Another work [34] applies mixture [48, 49] in an unsupervised manner to produce positive pairs that incorporate multiple samples. Additionally, different from data augmentation, [50] 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. 1a, 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. 1c 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 of an image. Hereby, we propose a semantic-aware localization scheme, which serves as a guidance to select crops and avoid most false positives, as shown in Fig. 1b. Moreover, we propose a center-suppressed sampling strategy to tackle trivial positive pairs with large similarity. Fig. 1d 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.

The proposed *ContrastiveCrop* considers both semantic information 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 are 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 works

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 [2, 23, 29, 30, 36, 43, 47]. More recently, contrastive methods based on the Siamese structure achieve remarkable performance on downstream tasks [5, 7, 8, 15, 17, 20, 40, 45, 46], some of which even surpass supervised models.

The milestone work is SimCLR [5], 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 [20], 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 [17] 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 [8] 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. In addition to framework design, theoretical analyses and empirical studies have also been proposed to better understand the behavior and properties of contrastive learning [1, 3, 6, 9, 24, 31, 35, 39, 39, 41, 44, 52].

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 using data augmentation. Most SSL works apply data augmentation pipelines that are directly adapted from those in supervised learning [12, 13, 19, 26, 48, 49]. Chen *et al*. [5] comprehensively study the effect of a range of data transformations, and find out the composition made of random cropping and random color distortion can lead to better performance. Tian *et al*. [37] propose an *InfoMin principle* to catch a sweet point of mutual information between

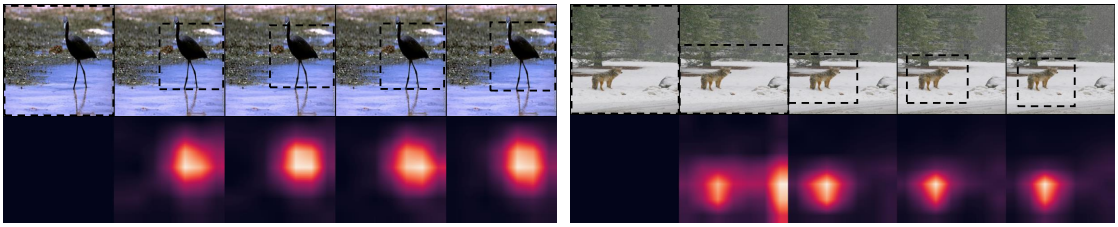


Figure 2. The training dynamic of localization is shown from left to right in each subfigure. We initialize the localization box as the whole image, and update it at a regular interval using the latest heatmap. Note that our goal is not to derive precise localization, but to guide generation of crops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

views, and accordingly generate positive pairs with its *InfoMin Augmentation*. A close work to this paper is [33], which also uses unsupervised saliency maps as a constraint of crops, but crops are still randomly sampled. All these works commonly apply *RandomCrop* as the basic sampling method to generate input views, which we find may not be the optimal solution for contrastive learning. [28] take object-scene relation into account when making crops, but require additional object proposal algorithms. In this work, we propose *ContrastiveCrop* that is tailored to create better positives views for contrastive learning, without the need of external functions.

3. Method

In this section, we introduce *ContrastiveCrop* for Siamese representation learning. Firstly, we briefly review *RandomCrop* as the 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, 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 I, we first determine the scale s and aspect ratio r of the crop from a pre-defined range (*e.g.*, $s \in [0.2, 1.0]$ and $r \in [3/4, 4/3]$) Then, the height and width of the crop can be obtained with s and r. After that, the location of the crop is randomly selected on the image plane, 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, thus could provide diverse views of a sample. However, it performs sampling equiprobably (*i.e.*, each single view is sampled with the same probability), which ignores the semantic information of images. As shown in Fig. 1a, *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. As a result, optimizing the false positives in Fig. 3 may mislead the learning of good representations. Therefore, designing a semantic-aware sampling strategy for crops is crucial and vital for Siamese representation learning.

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.*, 0*th*, 20*th*, 40*th*, 60*th*, 80*th* 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 locate 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

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

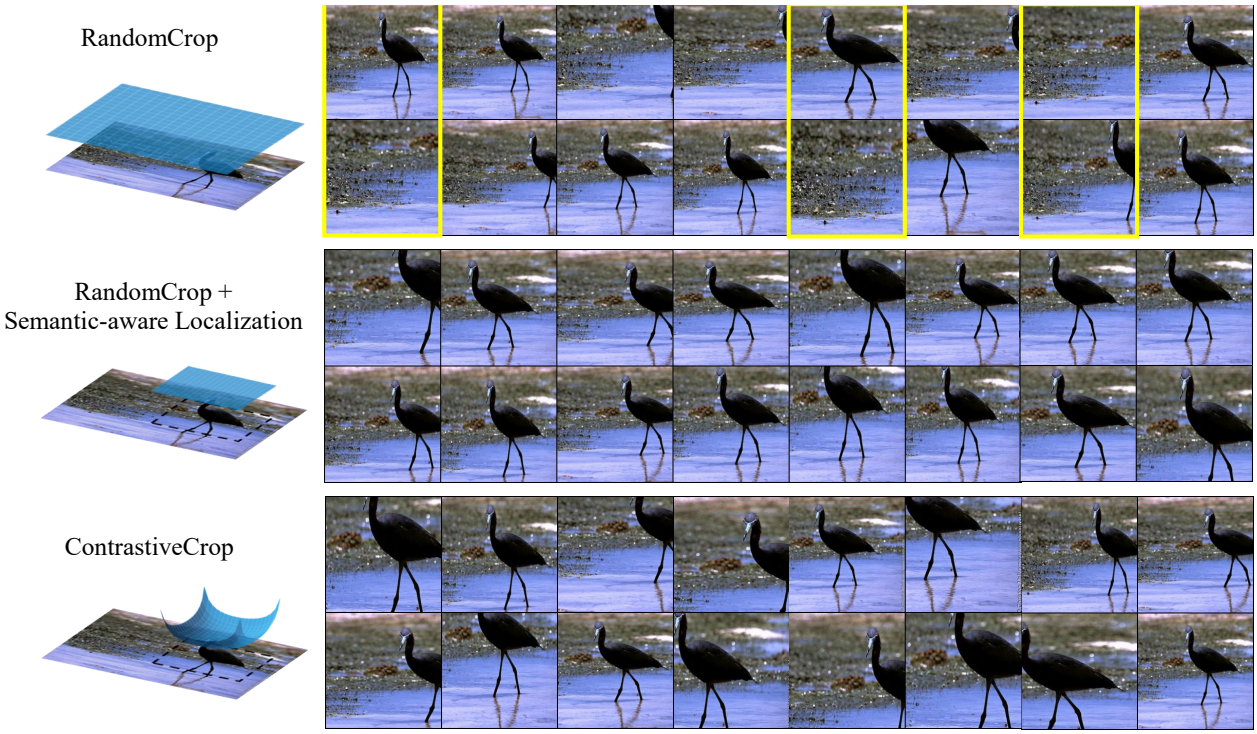


Figure 3. Visualization of *RandomCrop*, *RandomCrop + Semantic-aware Localization* and our *ContrastiveCrop*. We show the sampling distributions and operable regions for three settings on the left, and correspondent sampled pairs on the right. Pairs made by *RandomCrop* include several false positives 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.

can be written as,

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

where M represents heatmap, $k \in [0, 1]$ 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 probable coarse localization, we enlarge the operable region by only constraining center of crops within B. This also reduces the potential negative effect of resolution discrepancy at training and inference stages [38].

At the training stage, the bounding box is progressively updated at a regular interval to leverage the latest features learned by the model. Note that our goal is not to derive precise localization, but to guide generation of crops by finding the object of interest. The scale of the bounding box is controlled by the threshold parameter $k \in [0, 1]$. Generally, a larger k leads to a small box and would limit the diversity of crops to be made. A smaller k, however, may include much unrelated background texture and is not sufficient for finding the object. We study the effect of different threshold k in Sec. 4.4. We empirically find that the proposed localization module is not sensitive to this parameter and could improve over baseline within a wide range of k.

Finally, we show the sampling effect of semantic-aware localization in Fig. 3. Compared with *RandomCrop*, one can find that the false positive pairs reduce dramatically when the proposed module is applied. This provides evidence that self-supervised neural networks trained without annotations are capable of recognizing the object of interest as well as its location. In this way, additional region proposals or ground truth bounding boxes are no longer necessary for views generation [10, 51].

3.3. Center-suppressed Sampling

The semantic-aware localization scheme provides useful guidance to reduce false positive cases, but increases the probability of close appearance pairs due to the smaller operable region. In this subsection, we introduce the center-suppressed sampling that aims to tackle this dilemma.

Algorithm 1 *ContrastiveCrop* for Siamese Representation Learning

Input: Image I, Crop Scale s, Crop Ratio r, Threshold of Activations k, Parameter of β Distribution α.

$h = \sqrt{s \cdot r}$ ▷ Height of the crop

$w = \sqrt{s/r}$ ▷ Width of the crop

$F = Forward(I)$ ▷ Features of last layer

$M = Normalize(F)$ ▷ Heatmap after normalizing

$B = L(\mathbb{1}[M > k])$ ▷ Bounding box by Eq. 2.

$x = B_{x0} + (B_{x1} - B_{x0}) \cdot u, u \sim \beta(\alpha, \alpha)$

$y = B_{y0} + (B_{y1} - B_{y0}) \cdot v, v \sim \beta(\alpha, \alpha)$

▷ Sample crop center x and y from β distribution

Output: Crop $C = (x, y, h, w)$

The main idea is to reduce the probability of crops gathering around center by pushing them apart. Specifically, we adopt the beta distribution $\beta(\alpha, \alpha)$ with two identical parameters α, which shows a symmetric function. In this way, we could easily control the shape of the distribution with different α. As the goal is to enlarge the variance of crops, we set $\alpha < 1$ which gives us a U-shaped distribution (*i.e.*, with lower probability near center and greater one at other positions). In this way, crops are more likely to be scattered to near border lines of the operable region, and cases of much overlap could be largely avoided.

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 applies a center-suppressed distribution, and B is the same bounding box as in Eq. 3. Note that the shape of beta distribution is determined by the parameter α and affects the variance of crops. We study the impact of different α in in Sec. 4.4, including $\alpha > 1$ that gives a inverted U shape.

The effect of our *ContrastiveCrop* is visualized in Fig. 3. Compared with *RandomCrop*, our method could significantly reduce false positive pairs due to the semantic-aware localization. Meanwhile, it introduces larger variance within a positive 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 positives similarity. We take the class score of a crop as an indicator of richness of categorized semantic information. The similarity of positive pairs is calculated in the latent space as the cosine similarity between positive representations. Both the class score and similarity are average results of a large number of cropping trials from a standard ResNet-50 [22] trained with ImageNet [14] labels. Their relation is shown in Fig. 4. One can find that *ContrastiveCrop* conveys more semantic information than *RandomCrop* at the same level of variance, showing the

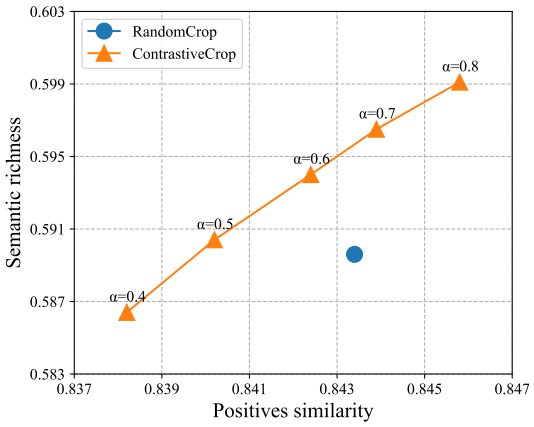


Figure 4. The relation between semantic richness and positives similarity. Dots are obtained by varying α, and scores of each dot are calculated by averaging results of a large number of cropping trials. Compared with *RandomCrop*, our *ContrastiveCrop* conveys more semantic information at the same level of similarity (vertical), and yields less similar positive pairs under equal semantic information (horizontal).

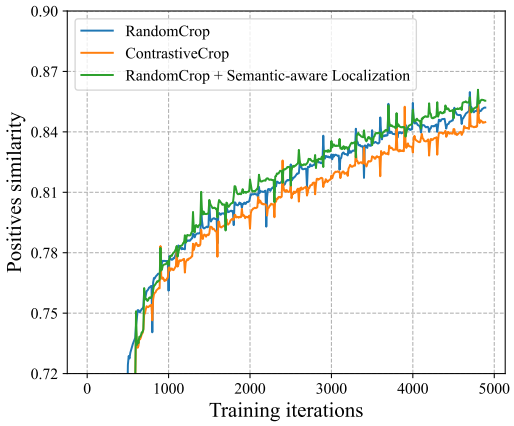


Figure 5. Similarity of positive pairs in training. Smaller positives similarity indicates harder positive samples which may enhance representation learning [50]. Taking *RandomCrop* as baseline, adding only localization results in slightly larger similarity. Our *ContrastiveCrop* combines both semantic-aware localization and center-suppressed sampling, which effectively reduces similarity of positives.

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 [5]	89.63	**90.08**	60.30	**61.91**	45.19	**46.21**	88.95	**89.53**
MoCo [20]	86.73	**88.78**	56.10	**57.65**	47.09	**47.98**	89.17	**89.81**
BYOL [17]	91.96	**92.54**	63.75	**64.62**	46.08	**47.23**	91.84	**92.42**
SimSiam [8]	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.

effectiveness of semantic-aware localization. Furthermore, with equal semantic information, *ContrastiveCrop* achieves larger variance than *RandomCrop*, which can be owed to center-suppressed sampling.

We further visualize the similarity of positive pairs in the training process in Fig. 5. As shown in the figure, adding only semantic-aware localization to *RandomCrop* slightly increases similarity, as localization restrains crops in a smaller operable region. Our *ContrastiveCrop* further incorporates center-suppressed sampling, showing smaller positives similarity than the other two. This indicates positive pairs sampled by *ContrastiveCrop* are harder ones, which are helpful in learning more view-invariant features as suggested in FT [50]. 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 downstream 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** [25], **Tiny ImageNet**, **STL-10** [11] and **ImageNet** [14]. Generally, these datasets are built for object recognition and the images contain iconic view of objects. The baseline contrastive methods include SimCLR [5], MoCo V1 & V2 [7, 20], BYOL [17] and SimSiam [8].

4.2. Implementation Details

Our *ContrastiveCrop* aims to make better views for contrastive learning, which is agnostic to self-supervised learning frameworks and their related training components, such as backbone networks, losses, optimizers, *etc*. Thus, 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-training stage, we train ResNet-18 [22] for 500 epochs with a batch size of 512 and a cosine-annealed learning rate of 0.5. The linear classifier is trained for 100 epochs with initial learning rate of 10.0 multiplied by 0.1 at the 60*th* and 80*th* epochs.

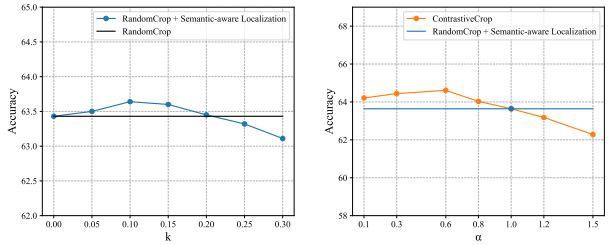
For experiments on ImageNet, we adopt ResNet-50 as the base model. Pre-training settings of MoCo V1 & V2 and SimSiam exactly follow their original works. We reproduce SimCLR with a smaller batch size of 512 and cosine-annealed learning rate of 0.05. For linear evaluation, we adopt the same setting as in [20] for all baseline methods.

For our method, we set $k = 0.1$ for the threshold of activations and $\alpha = 0.6$ for sampling. Localization boxes are updated at a frequency of 20% (*i.e*., 4 updates in total, except the last epoch), which adds negligible extra training overhead; *RandomCrop* is applied before the first update to collect global information, as described in Sec. 3.2. All the experiments are conducted with 8 GPUs. We use SGD optimizer with a momentum of 0.9 and a weight decay of 10^{-4} and 0 for pre-training and linear evaluation respectively.

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, *Con-*



(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* (the black plot) with *RandomCrop* + *Semantic-aware Localization* (the blue plot). In subfigure (b), we fix the best $k = 0.1$ for localization (the blue plot) and compare it with *ContrastiveCrop* to study the influence of different α.

Method	Arch.	Epoch	IN-200 Top-1	IN-1K Top-1
SimCLR	R50	100	62.14	61.60
SimCLR + Ours	R50	100	**63.08**	**61.85**
MoCo V1	R50	100	64.52	57.25
MoCo V1 + Ours	R50	100	**65.80**	**58.34**
MoCo V2	R50	100	63.43	64.40
MoCo V2 + Ours	R50	100	**64.61**	**64.89**
SimSiam	R50	100	62.89	65.62
SimSiam + Ours	R50	100	**63.54**	**65.95**

Table 2. Comparison of *RandomCrop* and our *ContrastiveCrop* with linear classification results on IN-200 and IN-1K. Models are pre-trained for 100 epochs, with the same training setup within a method for fair comparison.

trastiveCrop 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 at a frequency of 20% in the training process (*i.e.*, 4 times in total, except the last epoch), 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

In ablation studies, we investigate the semantic-aware localization module and center-suppressed sampling respectively. We also study the effect of *ContrastiveCrop* when it is combined with different transformations. We conduct experiments with MoCo V2 and ResNet-50, and report the linear classification results on IN-200.

Semantic-aware Localization. In our method, the unsupervised semantic-aware localization serves as a guidance to make crops. We study the influence of k that determines the scale of the localization box, with a larger k leading to a smaller box. We also make comparison with *RandomCrop* that does not use localization (*i.e.*, $k = 0$). Experimental results are shown in Fig. 6a. One can find that using localization box outperforms *RandomCrop* baseline (the black plot) within a range from 0.05 to 0.2. This shows the effectiveness of largely removing false positives. However, as k increases over 0.25, the performance starts to fall quickly. We suggest the reason is smaller bounding boxes dramatically reduce variance of views, making it trivial to learn discriminative features.

Freq.	0%	10%	20%	30%	50%
Acc. (%)	63.43	64.40	64.61	64.40	64.11

Table 3. Linear classification accuracy w.r.t different update frequencies of localization boxes. *RandomCrop* is applied before the first update.

We also study the effect of update frequency of localization boxes in Tab. 3. It shows that only one update in the middle of training (*i.e.*, 50%) could outperform *RandomCrop* baseline (*i.e.*, 0%) with a non-trivial margin. A larger improvement is seen in a range of $10\% \sim 30\%$ where there are more updates. These results show that our method could work well for different update frequencies.

Center-suppressed Sampling. In this work, 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. 6b 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 their combinations with other image transformations. Here, we choose the transformations used in

Pre-train	IN-1K Top-1	VOC detection AP	VOC detection AP_{50}	VOC detection AP_{75}	COCO instance seg. AP^{mk}	COCO instance seg. AP^{mk}_{50}	COCO instance seg. AP^{mk}_{75}	COCO detection AP^{bb}	COCO detection AP^{bb}_{50}	COCO detection 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 [37]	70.1	57.6	82.7	64.6	34.1	55.2	36.3	39.0	58.5	42.0
MoCoV1 [20]	60.6	55.9	81.5	62.6	33.6	54.8	35.6	38.5	58.3	41.6
MoCoV1 + *ContrastiveCrop*	**63.0**	**56.1**	**81.7**	**63.0**	**33.9**	**55.2**	**36.1**	**38.8**	**58.5**	**41.9**

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.

MoCo V2 [7], including *Flip*, *ColorJitter*, *Grayscale* and *Blur*. The ablation results are shown in Tab. 5. In case all other transformations are removed, *ContrastiveCrop* is 0.4% higher than *RandomCrop*, which is a direct evidence of its 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, which indicates that the potential of *ContrastiveCrop* can be larger exploited with further color transformations. Additionally, these results show that our *ContrastiveCrop* is compatible and orthogonal to other transformations.

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 5. Ablation of other transformations used in MoCo V2. We combine *ColorJitter* and *Grayscale* as a single color transformation. *R-Crop* and *C-Crop* denote *RandomCrop* and *ContrastiveCrop*, respectively. The results are from ResNet-50 pre-trained on IN-200 for 100 epochs.

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 [20, 50], we pre-train ResNet-50 on IN-1K for 200 epochs. For downstream tasks, we use PASCAL VOC [16] and COCO [27] as our benchmarks and we adopt the same setups as in MoCo's detectron2 codebase [42]. All layers of pre-trained models are fine-tuned end-to-end at target datasets.

PASCAL VOC Object Detection. Following [20], we use Faster R-CNN [32] with a backbone of R50-C4 [21] as the detector. 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 a consistent improvement of +0.2AP, +0.2AP_{50} and +0.4AP_{75}.

COCO Object Detection/Instance Segmentation. The model for both detection and segmentation is Mask R-CNN [21] 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* achieves superior performance in all 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 of positives designing, considering its significant role in contrastive learning.

Acknowledgements. This research is supported by the National Research Foundation, Singapore under its AI Singapore Programme (AISG Award No: AISG2-PhD-2021-08-008). We thank Google TFRC for supporting us to get access to the Cloud TPUs. We thank CSCS (Swiss National Supercomputing Centre) for supporting us to get access to the Piz Daint supercomputer. We thank TACC (Texas Advanced Computing Center) for supporting us to get access to the Longhorn supercomputer and the Frontera supercomputer. We thank LuxProvide (Luxembourg national supercomputer HPC organization) for supporting us to get access to the MeluXina supercomputer.

References

[1] Sanjeev Arora, Hrishikesh Khandeparkar, Mikhail Khodak, Orestis Plevrakis, and Nikunj Saunshi. A theoretical analysis of contrastive unsupervised representation learning. *arXiv preprint arXiv:1902.09229*, 2019. 2

[2] Philip Bachman, R Devon Hjelm, and William Buchwalter. Learning representations by maximizing mutual information across views. In *NeurIPS*, 2019. 2

[3] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. *arXiv preprint arXiv:2104.14294*, 2021. 2

[4] Pengguang Chen, Shu Liu, and Jiaya Jia. Jigsaw clustering for unsupervised visual representation learning. In *CVPR*, 2021. 2

[5] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020. 1, 2, 6

[6] Ting Chen, Calvin Luo, and Lala Li. Intriguing properties of contrastive losses. *arXiv preprint arXiv:2011.02803*, 2020. 2

[7] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 2, 6, 8

[8] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. *CVPR*, 2021. 1, 2, 6

[9] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. *arXiv preprint arXiv:2104.02057*, 2021. 2

[10] Ming-Ming Cheng, Ziming Zhang, Wen-Yan Lin, and Philip Torr. Bing: Binarized normed gradients for objectness estimation at 300fps. In *CVPR*, 2014. 4

[11] Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *AISTATS*, 2011. 6

[12] Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. Autoaugment: Learning augmentation policies from data. *arXiv preprint arXiv:1805.09501*, 2018. 2

[13] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *CVPR*, 2020. 2

[14] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 5, 6

[15] 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. 2

[16] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 2010. 1, 8

[17] 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. 1, 2, 6

[18] Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. Densepose: Dense human pose estimation in the wild. In *CVPR*, 2018. 1

[19] Ryuichiro Hataya, Jan Zdenek, Kazuki Yoshizoe, and Hideki Nakayama. Faster autoaugment: Learning augmentation strategies using backpropagation. In *ECCV*, 2020. 2

[20] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020. 1, 2, 6, 8

[21] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, 2017. 1, 8

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 5, 6

[23] 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. 2

[24] Tianyu Hua, Wenxiao Wang, Zihui Xue, Sucheng Ren, Yue Wang, and Hang Zhao. On feature decorrelation in self-supervised learning. In *ICCV*, 2021. 2

[25] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. Technical report, Citeseer, 2009. 6

[26] Sungbin Lim, Ildoo Kim, Taesup Kim, Chiheon Kim, and Sungwoong Kim. Fast autoaugment. *NeurIPS*, 2019. 2

[27] 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*, 2014. 1, 8

[28] Shlok Mishra, Anshul Shah, Ankan Bansal, Abhyuday Jagannatha, Abhishek Sharma, David Jacobs, and Dilip Krishnan. Object-aware cropping for self-supervised learning. *arXiv preprint arXiv:2112.00319*, 2021. 3

[29] Ishan Misra and Laurens van der Maaten. Self-supervised learning of pretext-invariant representations. In *CVPR*, 2020. 2

[30] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 2

[31] Senthil Purushwalkam and Abhinav Gupta. Demystifying contrastive self-supervised learning: Invariances, augmentations and dataset biases. *arXiv preprint arXiv:2007.13916*, 2020. 2

[32] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NeurIPS*, 2015. 1, 8

[33] Ramprasaath R. Selvaraju, Karan Desai, Justin Johnson, and Nikhil Naik. Casting your model: Learning to localize improves self-supervised representations. In *CVPR*, pages 11058–11067, June 2021. 3

[34] 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. 2

[35] Yuandong Tian, Xinlei Chen, and Surya Ganguli. Understanding self-supervised learning dynamics without contrastive pairs. *arXiv preprint arXiv:2102.06810*, 2021. 2

[36] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. *ECCV*, 2019. 2

[37] 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. 1, 2, 8

[38] Hugo Touvron, Andrea Vedaldi, Matthijs Douze, and Hervé Jégou. Fixing the train-test resolution discrepancy. *arXiv preprint arXiv:1906.06423*, 2019. 4

[39] Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *ICML*, 2020. 2

[40] Xinlong Wang, Rufeng Zhang, Chunhua Shen, Tao Kong, and Lei Li. Dense contrastive learning for self-supervised visual pre-training. *CVPR*, 2021. 2

[41] Mike Wu, Chengxu Zhuang, Milan Mosse, Daniel Yamins, and Noah Goodman. On mutual information in contrastive learning for visual representations. *arXiv preprint arXiv:2005.13149*, 2020. 2

[42] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019. 8

[43] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *CVPR*, 2018. 2

[44] Tete Xiao, Xiaolong Wang, Alexei A Efros, and Trevor Darrell. What should not be contrastive in contrastive learning. *ICLR*, 2021. 2

[45] 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*, 2021. 2

[46] 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*, 2021. 2

[47] Mang Ye, Xu Zhang, Pong C Yuen, and Shih-Fu Chang. Unsupervised embedding learning via invariant and spreading instance feature. In *CVPR*, 2019. 2

[48] 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 *ICCV*, 2019. 2

[49] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *ICLR*, 2018. 2

[50] 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. 2, 5, 6, 8

[51] C Lawrence Zitnick and Piotr Dollár. Edge boxes: Locating object proposals from edges. In *ECCV*, 2014. 4

[52] Barret Zoph, Golnaz Ghiasi, Tsung-Yi Lin, Yin Cui, Hanxiao Liu, Ekin D Cubuk, and Quoc V Le. Rethinking pre-training and self-training. *arXiv preprint arXiv:2006.06882*, 2020. 2