A Detailed steps of the JointCrop

For a clearer understanding of JointCrop, we provide the detailed steps in Alg. 1.

Algorithm 1: The steps of J-Crop(β) different from RandomCrop.

Hyperparameter: the area range $[s_{\min}, s_{\max}]$, the β controls the difficulty
Output: Areas of a pair of positive views s_1 and s_2
/* In RandomCrop, s_1 and s_2 are independently and identically distributed and their ratio is likely to be close to 1 */
$s_1 \sim \mathcal{U}[s_{\min}, s_{\max}]$;
$s_2 \sim \mathcal{U}[s_{\min}, s_{\max}]$;
return s_1, s_2;
/* JointCrop controls the joint distribution of s_1 and s_2 by controlling the area ratio $s_r = s_2/s_1$ */
$s_b = \log \frac{s_{\max}}{s_{\min}}$;
if $\beta > 0$ **then**
  $\log s_r \sim \mathcal{N}_T(0, \frac{1}{\beta} s_b, -s_b, s_b)$;
else if $\beta = 0$ **then**
  $\log s_r \sim \mathcal{U}[-s_b, s_b]$;
else if $\beta < 0$ **then**
  $\log s_r \sim \mathcal{N}_T(0, -\frac{1}{\beta} s_b, -s_b, s_b)$;
  if $\log s_r < 0$ **then**
    $\log s_r = -s_b - \log s_r$;
  else
    $\log s_r = s_b - \log s_r$;
  end
end
$s_1 \sim \mathcal{U}\left[\max\left(s_{\min}, \frac{s_{\min}}{s_r}\right), \min\left(\frac{s_{\max}}{s_r}, s_{\max}\right)\right]$;
$s_2 = s_1 \times s_r$;
return s_1, s_2;
/* With s_1 and s_2 known, we can use the same steps as RandomCrop to compute h_1, w_1, i_1, j_1 and then get a pair of positive views */

B Code of Our Method

Different code frameworks are employed for various methods, as illustrated in Tab. 9. The official code is directly cloned, and modifications are made to the data augmentation components of these codes. For instance, in replacing RandomCrop with JointCrop, it is necessary to overload the RandomCrop class 'torchvision.transforms.RandomResizedCrop' and amalgamate the cropping process for a pair of positive samples into a singular class or function to achieve a 'joint' operation. Only this segment of the code requires modification, while the remainder of the code is maintained as per the baseline code utilized.

The code for pre-training with JointCrop and performing linear fine-tuning on MoCo v3 has been included in the supplementary material. Detailed instructions for setting up the environment and executing the code are provided in the README. We used different codebases for various baselines, as outlined in Tab. 9.

Method	Code Link
MoCo v1	https://github.com/facebookresearch/moco
MoCo v2	https://github.com/facebookresearch/moco
MoCo v3	https://github.com/facebookresearch/moco-v3
SimSiam	https://github.com/open-mmlab/mmselfsup
BYOL	https://github.com/open-mmlab/mmselfsup
SimCLR	https://github.com/open-mmlab/mmselfsup
Dino	https://github.com/facebookresearch/dino

Table 9: The code base we use for various methods.

C Limitation and Future Work

The limitations of this paper and the directions for our future work are summarized as follows.

C.1 Limitation

- Our approach necessitates pre-training from scratch, consuming significant energy, time, and computational resources. Nonetheless, we intend to make the pre-trained weights publicly accessible, enabling others to fine-tune their downstream tasks accordingly.
- Instance Discrimination (ID) and Masked Image Modeling (MIM) represent prominent paradigms in self-supervised learning (SSL). Our methodology is exclusively adaptable to ID (contrastive learning) and is not applicable to MIM (generative SSL, such as MAE).

C.2 Future work

- We aim to achieve outcomes from training over extended epochs. However, such an endeavor demands considerable time and computational resources. The most extensive pre-training demonstrated herein is the outcome of 300 epochs of MoCo v3 pre-training.
- We plan to investigate the integration of various JointAugmentation techniques to generate challenging pairs of positive samples, thereby enhancing the model's feature representation capabilities.

D Analysis for JointCrop and MultiCrop

Multi-Crop enables positive pairs to have both a global view and a local view by manually specifying the area size during cropping and using more than two views. For example, a typical setup includes two global views with an area ranging from 40% to 100% of the original image and four local views with an area between 5% and 40% of the image. The key distinction between Multi-Crop and our JointCrop is that: (1) In Multi-Crop, the data augmentations for positive pairs are independent, whereas in JointCrop, they are not. (2)

JointCrop incurs no additional computational cost, whereas Multi-Crop takes approximately 1.35 times longer than the baseline. (3) The underlying concepts of our JointCrop can be extended to other augmentations, such as JointBlur and JointColor, among others. In contrast, Multi-Crop lacks this kind of extensibility. (4) JointCrop can be further integrated with Multi-Crop to enhance its capabilities. We tested the combination of JointCrop and Multi-Crop. Specifically, we applied JointCrop to both the global and local pairs of Dino (Caron et al. 2021) with the same Multi-Crop configuration of 2x160+4x96, and then trained and fine-tuned on ImageNet-1K. As shown in Table 10, our JointCrop can be effectively combined with Multi-Crop to further enhance the performance of contrastive learning.

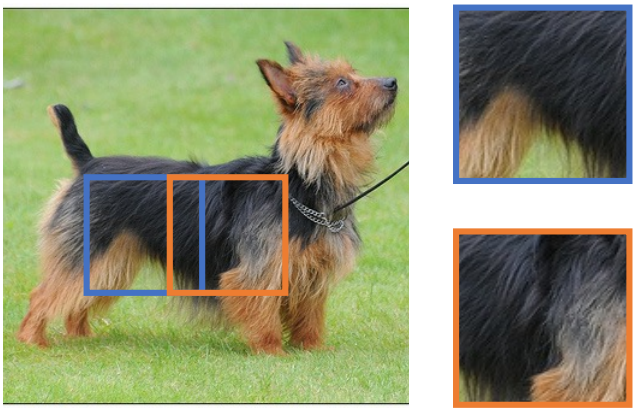
Model	Multi-Crop	Multi-Crop+JointCrop
Dino	66.36	66.55 (global) 66.58 (local)

Table 10: Results of the Combination of Multi-Crop and JointCrop. As these baseline were not available in their original papers, we have reproduced them for our analysis.

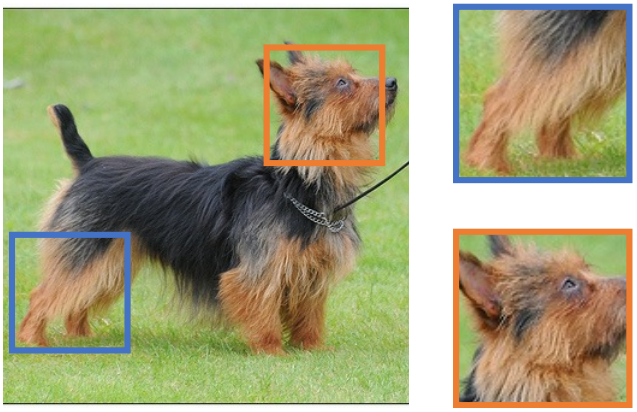
E Analysis for Distance between Positive Pairs in JointCrop

In JointCrop, we control h_1, w_1, and h_2, w_2 by controlling the area ratios s_1 and s_2. However, Crop requires sample locations (i_1, j_1) and (i_2, j_2) in addition to sample widths and heights. The Euclidean distance $d = \sqrt{(i_1 - i_2)^2 + (j_1 - j_2)^2}$ of the positional coordinates between a positive pair also affects its difficulty. A pair of positive samples that are close together may have similar features, *e.g.*, the two parts of a dog's body shown in Fig. 6a. This pair is itself very similar and simple for CL. Whereas a pair of positive samples that are farther away may have dissimilar features but have the same semantic information (they are on one image, after all), *e.g.*, the dog's head and dog's leg in Fig. 6b. This pair is more challenging and may help models learn better.

Can we add control over distance in JointCrop to get more challenging samples? This is certainly possible, but let's first note that JointCrop actually implicitly controls the distance already. In Sec. 3.1 we have shown that $i \sim \mathcal{U}[0, \mathrm{W} - w]$ and $j \sim \mathcal{U}[0, \mathrm{H} - h]$. Compared to RandomCrop, JointCrop directly controls the area ratio s_2/s_1 and indirectly changes the distribution of s_1. This indirectly changes the distributions of h_1 and w_1 and then affects the distributions of i_1 and j_1. Similarly, the distributions of i_2 and j_2 are indirectly controlled. Therefore, JointCrop will also affect distance d. Direct theoretical derivation of distances is very complex. We repeated the experiment three times, each time taking 100,000 positive pairs from J-Crop(β) and measuring the distance between them, as illustrated in Fig. 7. A smaller β does increase the Euclidean distance between pairs of positive samples.



(a) Close views may be similar.



(b) Distant views may be challenging.

Figure 6: Distance between views may affect difficulty.

F Analysis for the Combination of JointCrop and JointBlur

In this paper we propose JointCrop and JointBlur, and both JointCrop and JointBlur can improve the linear evaluation accuracy over baselines under multiple datasets and on multiple CL methods. However, what if we use them together, *e.g.*, the combination of JointCrop and JointBlur?

In fact, these two augmentation methods are in conflict with each other. JointCrop encourages the ratio of the area of a positive pair of samples to be farther away from 1, that is, with a higher probability of obtaining a "large view" and a "small view". The "large" and "small" here do not refer to the size of the images, as they are all resized to the same size (default to 224×224). They refer to the area of the views in the original image. While, JointBlur makes the GaussianBlur kernel of a pair of positive samples more different, that is, with a higher probability of obtaining a "fuzzy view" and a "clear view". If we use the two methods at the same time, we may obtain overly challenging samples, *e.g.*, pairs of positive samples with a "small fuzzy view" and a "large clear view", which are too difficult to learn good feature representations.

Examples of such a case are given in Figure 8. Figure 8a shows an image from ImageNet-1K, and Figures 8b and 8c

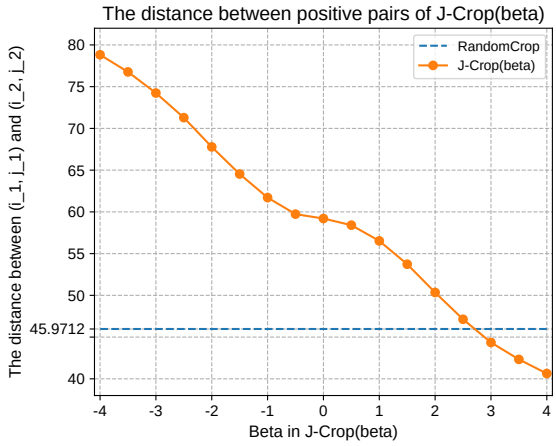


Figure 7: The distance between pairs also affects difficulty. JointCrop not only controls the area ratios directly, but also indirectly controls the distance between positive pairs.

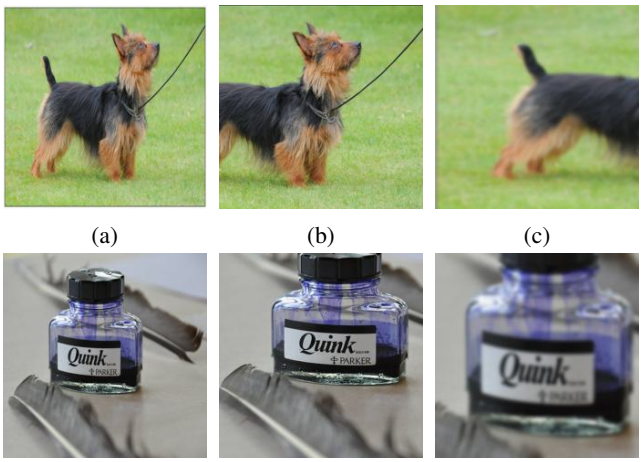


Figure 8: Examples of pairs of positive samples with a "small fuzzy view" (c and f) and a "large clear view" (b and e). The subfigures (a) and (d) are the original images.

show a pair of positive samples obtained using JointCrop and JointBlur. Figure 8b has a larger area in the original image but uses a weaker GaussianBlur, while Figure 8c has a small area in the original image. Figure 8c is already not as high-resolution as Figure 8b when it is resized to the same size. In addition, with a stronger GaussianBlur, Figures 8b and 8c are too difficult to learn a good feature representation. Figures 8d to 8f are also similar examples.

While the simultaneous application of JointCrop and JointBlur might produce samples that are excessively challenging, employing them in sequence can be beneficial. For instance, one can apply either JointCrop or JointBlur with a certain probability when generating positive samples. Our experiments, conducted using MoCo v1, demonstrate that employing J-Crop(0) or J-Blur(0) individually enhances the baseline accuracy from 57.25% to 60.87% and 60.58%, respectively. Furthermore, applying either of these methods with a probability of 0.5 yields an improved result of 61.24%.

G Combine JointCrop and ContrastiveCrop

Our JointCrop method explicitly manages the area ratio of two cropping regions $\frac{h_2 \cdot w_2}{h_1 \cdot w_1}$ and implicitly influences the distance d between positive pairs, in contrast to ContrastiveCrop (Peng et al. 2022), which directly controls the cropped regions i_1, j_1 and i_2, j_2. ContrastiveCrop is bifurcated into two segments.

- *Semantic-aware Localization* restricts the cropping area through a heatmap to preclude object absence, thereby increasing the likelihood of samples appearing proximal to, particularly at the center of, the heatmap.
- *Center-suppressed Sampling* enforces a β-distribution for coordinates i_1, j_1 and i_2, j_2 to diversify positive pairs that are overly analogous.

The integration of JointCrop and Center-suppressed Sampling might yield superior outcomes, as both methods engender more challenging views. However, *Semantic-aware Localization* by constricting the cropping area, offers less challenging views. The results of MoCo v1 on ImageNet-1K are presented in Tab. 11. Employing our JointCrop independently surpasses the MoCo v1 baseline and ContrastiveCrop, while the amalgamation of our JointCrop with *Center-suppressed Sampling* (a component of ContrastiveCrop) yields enhanced results (as indicated by bolded number in Tab. 11). Furthermore, JointCrop is also amenable to integration with additional techniques.

Method	ImageNet-1K Top-1 Accuracy
MoCo v1 Baseline	57.25
MoCo v1 + ContrastiveCrop	58.34
MoCo v1 + JointCrop(0)	60.87
JointCrop(0) + *Center-suppressed Sampling*	**61.33**

Table 11: The combination of our proposed JointCrop and ContrastiveCrop on ImageNet-1K. The training setup is the same as MoCo v1 and ContrastiveCrop.

H Derivation of Probability Density Functions for Area Ratios in RandomCrop

Assume that $\widetilde{s}_1, \widetilde{s}_2 \sim \mathcal{U}[s_{min}, s_{max}]$, where $\mathcal{U}$ denotes a uniform distribution. Consequently, the probability density functions $g(y)$ for $\widetilde{s}_1$ and $\widetilde{s}_2$ are identical.

$$g(y) = \frac{1}{s_{max} - s_{min}} \tag{4}$$

Let the random variable $X = \dfrac{\widetilde{s}_1}{\widetilde{s}_2}$ be defined, and for $s_{min} \leq x \leq 1$, the cumulative density function $F(x)$ is specified.

$$\begin{aligned} F(x) = p(X \leq x) &= \int_{s_{min}/x}^{s_{max}} \int_{s_{min}}^{\widetilde{s}_2 y} g^2(y)\, \mathrm{d}\widetilde{s}_1 \mathrm{d}\widetilde{s}_2 \\ &= \frac{1}{(s_{max} - s_{min})^2} \int_{s_{min}/x}^{s_{max}} (\widetilde{s}_2 y - s_{min})\, \mathrm{d}\widetilde{s}_2 \\ &= \frac{1}{(s_{max} - s_{min})^2} \left[\frac{x}{2} \widetilde{s}_2^2 - s_{min} \widetilde{s}_2 \right]_{s_{min}/x}^{s_{max}} \\ &= \frac{1}{(s_{max} - s_{min})^2} \left[\frac{s_{max}^2}{2} x - s_{max} s_{min} + \frac{s_{min}^2}{2x} \right] \end{aligned} \tag{5}$$

For $\frac{s_{min}}{s_{max}} < x < 1$, the probability density function of Y is derived from Eq. (5).

$$\begin{aligned} f(x) &= \frac{\mathrm{d}}{\mathrm{d}x} F(x) \\ &= \frac{1}{(s_{max} - s_{min})^2} \left[\frac{s_{max}^2}{2} - \frac{s_{min}^2}{2x^2} \right] \\ &= \frac{s_{max}^2 x^2 - s_{min}^2}{2x^2 (s_{max} - s_{min})^2} \end{aligned} \tag{6}$$

Similarly, $f(x)$ can be determined for $1 < x \leq \frac{s_{max}}{s_{min}}$.

$$f(x) = \frac{s_{max}^2 - s_{min}^2 x^2}{2x^2 (s_{max} - s_{min})^2} \tag{7}$$

In summary, the probability density function $f(x)$ for $X = \widetilde{s}_1 / \widetilde{s}_2$ is presented in Eq. (8).

$$f(x) = \begin{cases} \dfrac{s_{max}^2 x^2 - s_{min}^2}{2x^2 (s_{max} - s_{min})^2}, & \dfrac{s_{min}}{s_{max}} \leq x \leq 1, \\ \dfrac{s_{max}^2 - s_{min}^2 x^2}{2x^2 (s_{max} - s_{min})^2}, & 1 < x \leq \dfrac{s_{max}}{s_{min}}. \end{cases} \tag{8}$$

In RandomCrop, typical settings are $s_{min} = 0.2$ and $s_{max} = 1$. By substituting these values into Eq. (8), the probability density map of $d(R(I_j; \widetilde{s}_1), R(I_j; \widetilde{s}_2))$ is obtained, as shown in Eq. (10).

$$a \sim F(x) = \begin{cases} \dfrac{25}{32} x - \dfrac{5}{16} + \dfrac{1}{32x}, & 0.2 \leq x \leq 1, \\ -\dfrac{1}{32} x + \dfrac{21}{16} - \dfrac{25}{32x}, & 1 < x \leq 5. \end{cases} \tag{9}$$

$$\widetilde{s}_r = \frac{\widetilde{s}_1}{\widetilde{s}_2} \sim f(x) = \begin{cases} \dfrac{25}{32} - \dfrac{1}{32x^2}, & 0.2 \leq x \leq 1, \\ -\dfrac{1}{32} + \dfrac{25}{32x^2}, & 1 < x \leq 5. \end{cases} \tag{10}$$