A. Details on Temporal Interval Sampling

Here we describe how to sample the temporal interval $t \in [0, T]$ from a given distribution $P(t)$. Suppose $P(t)$ is a power function:

$$P(t) = at^b + c, \tag{2}$$

where a, b and c are constants. We adopt the technique of inverse transform sampling [64] by first calculating the cumulative distribution function (CDF) $F(t)$ of $P(t)$ as:

$$F(t) = \int_{-\infty}^{t} P(x)\, dx = \frac{a}{b+1} t^{b+1} + ct, \tag{3}$$

where $t \in [0, T]$. To sample a temporal interval t, we then generate a random variable $v \sim U(0, 1)$ from a standard uniform distribution and calculate $t = F^{-1}(v)$. Notice that it is difficult to directly compute the closed-form solution of the inverse function of $F(\cdot)$. Considering the facts that the temporal interval t is an integer representing the number of frames between the start frames of two clips and $F(\cdot)$ is monotonically increasing, we use a simple binary search method in Algorithm 2 to find t. The algorithm is demonstrated below and the complexity is $\mathcal{O}(\log T)$.

Algorithm 2: Temporal Interval Sampling

Input: random variable $v \sim U(0, 1)$, CDF function $F(\cdot)$
  $upper\_bound = T$
  $lower\_bound = 0$
  while $upper\_bound - lower\_bound > 1$ **do**
    $t = \text{int}((upper\_bound + lower\_bound)/2)$
    if $F(t) > v$ **do**
      $upper\_bound = t$
    else do
      $lower\_bound = t$
  end while
Output: temporal interval $t \approx F^{-1}(v)$

B. Additional Results

B.1. Semi-Supervised Learning on Kinetics-600

We also conduct semi-supervised learning on K600. Similar to K400, we sample 1% and 10% videos from each class in the training set, forming two balanced subsets, respectively. The evaluation set remains the same. As in Table 10, CVRL shows strong performance especially when there is only 1% labeled data.

B.2. Comparison with RandAugment

We are interested in the performance of strong spatial augmentations that are widely used in supervised learning.

Method	Backbone	K600 Top-1 Acc. (Δ *vs.* Sup.)	
		1% label	10% label
Supervised	R3D-50	4.3	45.3
SimCLR infla.	R3D-50	16.9 (12.6↑)	51.4 (6.1↑)
ImageNet infla.	R3D-50	19.7 (15.4↑)	48.3 (3.0↑)
CVRL	R3D-50	**36.7 (32.4↑)**	**56.1 (10.8↑)**

Table 10. **Semi-supervised learning results on Kinetics-600.**

Augmentation method	Accuracy (%)	
	top-1	top-5
RandAugment w/ temporal consistency	54.2	77.9
Proposed	**63.8**	**85.2**

Table 11. **Performance of different spatial augmentations in pre-training (200 epochs).** Our proposed augmentation method outperforms RandAugment with temporal consistency.

We experiment with RandAugment [12] to randomly select 2 operators from a pool of 14. We conduct experiments with 200 epochs pre-training on Kinetics-400 [38]. For linear evaluation, RandAugment with temporal consistency achieves 54.2% top-1 accuracy as shown in Table 11, which is lower than our temporally consistent spatial augmentation presented in Algorithm 1, implying that strong augmentations optimized for supervised image recognition do not necessarily perform as well in the self-supervised video representation learning.

C. Illustrations

C.1. Pre-Training and Linear Evaluation

More detailed pre-training statistics on Kinetics-400 [38] are illustrated in Figure 5. We display four metrics: (1) contrastive loss, (2) regularization loss, (3) entropy and (4) pre-training accuracy. The total loss is the sum of contrastive loss and regularization loss. We also provide linear evaluation statistics in Figure 6, where all models are pre-trained on Kinetics-400 for 800 epochs corresponding to Figure 5.

C.2. Temporally Consistent Spatial Augmentation

We illustrate the proposed temporally consistent spatial augmentation method in Figure 7. Given an original video clip (top row), simply applying spatial augmentations to each frame independently would break the motion cues across frames (middle row). The proposed temporally consistent spatial augmentation (bottom row) would augment the spatial domain of the video clip while maintaining their natural temporal motion changes.

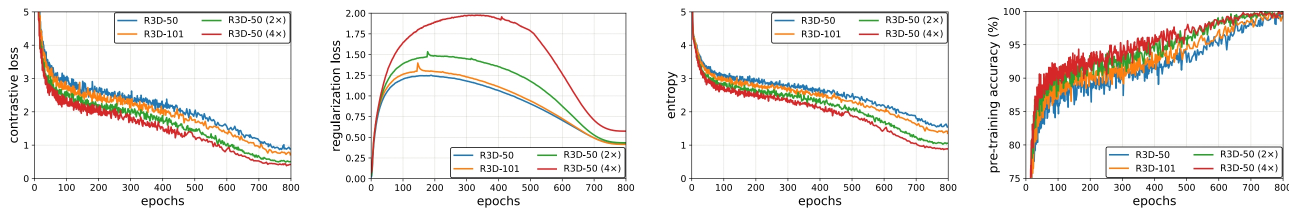

Figure 5. **Model pre-training statistics:** contrastive loss, regularization loss, entropy and pre-training accuracy on Kinetics-400.

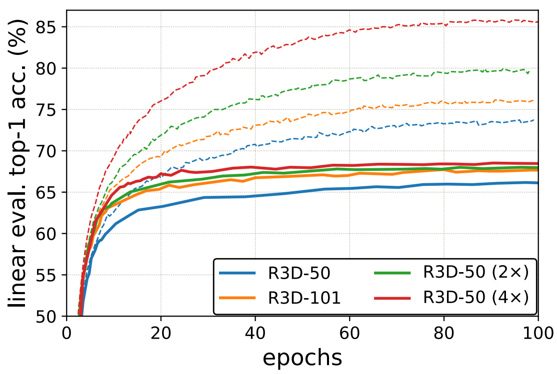

Figure 6. **Linear evaluation training (dashed-line) and evaluation (solid-line) top-1 accuracy** on Kinetics-400.

Figure 7. **Illustration of temporally consistent spatial augmentation.** The middle row indicates frame-level spatial augmentations without temporal consistency which would be detrimental to the video representation learning.