SELF-SUPERVISED VIDEO PRETRAINING YIELDS STRONG IMAGE REPRESENTATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Videos contain far more information than still images, and hold the potential for learning rich representations of the visual world. Yet, pretraining on image datasets has remained the dominant paradigm for learning representations that capture spatial information and previous attempts at video pretraining have fallen short on image understanding tasks. In this work we revisit self-supervised learning of image representations from the dynamic evolution of video frames. To that end, we propose a dataset curation procedure that addresses the domain mismatch between video and image datasets, and develop a contrastive learning framework which handles the complex transformations present in natural videos. This simple paradigm for distilling knowledge from videos to image representations, called VITO, performs surprisingly well on a variety of image-based transfer learning tasks. For the first time, our video-pretrained model closes the gap with ImageNet pretraining on semantic segmentation on PASCAL and ADE20k and object detection on COCO and LVIS, raising the possibility of video-pretraining becoming the new default for learning image representations.

1 INTRODUCTION

Pretraining on large image datasets has been the dominant paradigm for learning representations that understand the visual world (Krizhevsky et al., 2012; He et al., 2016). In particular, self-supervised methods, which learn representations that are invariant to specific image transformations, have proven very powerful, surpassing supervised pretraining on a variety of downstream tasks (He et al., 2020; Hénaff et al., 2019; Chen et al., 2020; Caron et al., 2021). Although the synthetic augmentations used in these transformations capture important image priors such as scale-, color-, and translation-invariance, they pale in comparison to the complex changes in pose, viewpoint, and motion that arise in natural videos. Furthermore, while learning invariant representations has shown to be a powerful proxy for downstream tasks like image classification, intelligent behavior requires also learning about the shapes, sizes, orientations, and the relationships objects can appear in, which videos are well suited to inform.

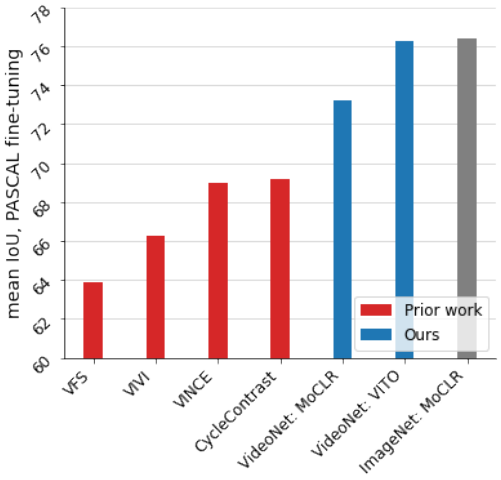

Figure 1: Closing the gap between video- and image-pretraining. Prior work learning image representations from video have lagged behind ImageNet pretraining. We propose a method for data curation, VideoNet, and self-supervised learning objective, VITO, which together close this gap.

While self-supervised video pretraining has seen a variety of recent successful applications in *video* representation learning (Toering et al., 2022; Dave et al., 2022; Feichtenhofer et al., 2022; Ni et al., 2022), it has typically lagged behind ImageNet pretraining when learning *image* representations (Gordon et al., 2020; Wu & Wang, 2021). Furthermore, the specifics of video-representation archi-

tectures make their comparison with image-based architectures difficult, obfuscating the role of the underlying data and learning paradigm in the quality of the resulting representations.

In this work we perform a systematic comparison of image- and video-based learning of image representations. Starting from a strong, self-supervised contrastive baseline, we find the spatial content of standard video datasets to have a detrimental effect on the quality of the resulting representations, as measured by their performance on canonical scene understanding tasks. We therefore introduce a straightforward video curation procedure—VideoNet—which aligns their class distribution with that of ImageNet, and which partially redresses the imbalance between image and video learning. Additionally, we propose three simple modifications to the standard contrastive paradigm to account for the particularities of video data: less aggressive crop augmentation, multi-scale attention pooling, and enriching view generation with natural temporal deformations. Together, these improvements yield large gains on semantic segmentation on PASCAL and ADE20K and object detection on COCO and LVIS, closing the gap between image- and video-based representation learning for the first time. This gives a new life to the promise of video pretraining serving as a general purpose means of learning visual representations.

2 Related work

Video-based pretraining. Many prior works have considered the problem of self-supervised representation learning for capturing spatio-temporal invariances. These span a wide range of approaches, beginning with traditional methods that leveraged temporal coherence, optical flow, and object tracking (Wiskott & Sejnowski, 2002; Hurri & Hyvärinen, 2003; Agrawal et al., 2015; Wang & Gupta, 2015; Pathak et al., 2017; Goroshin et al., 2015). More recently, there have been many successful examples of approaches that leverage contrastive learning, masked autoencoding, and other self-supervised pretext tasks (Sermanet et al., 2018; Recasens et al., 2021; Qian et al., 2021; Dave et al., 2022; Dorkenwald et al., 2022; Feichtenhofer et al., 2021; 2022) to learn strong video representations. However, most of these methods require specialized video architectures and transfer to video-based tasks (action recognition, motion segmentation, object tracking, etc.) to measure the quality of the learned representations. However, the natural deformations like the different viewpoints of an object created due to motion are powerful learning signals that should allow for learning better image representations as well. Recent works (Gordon et al., 2020; Alayrac et al., 2020; Wu & Wang, 2021; Tschannen et al., 2020; Xu & Wang, 2021) have made attempts in this direction, using either "deflated" video models or frame-based contrastive methods.

The most similar prior works to ours are Gordon et al. (2020), Xu & Wang (2021), and Wu & Wang (2021) which utilize contrastive learning to learn frame-level representations from videos. However, our method differs in its curation of video datasets and its ability to handle temporal deformations in the contrastive learning framework. Although these works report gains on video-centric tasks such as object tracking and video segmentation, in the context of canonical scene understanding tasks used to evaluate image representations, we find that these methods underperform state-of-the-art ImageNet-pretrained models.

Contrastive learning for fine-grained scene understanding. In this work, we specifically focus on evaluations that require real-world scene understanding, namely semantic segmentation and object detection. Self-supervised learning has greatly benefited fine-grained scene understanding tasks, and there has been significant progress using dense contrastive losses that chose positive pairs for local features by spatial proximity and/or feature affinity across two views (Xie et al., 2021; Bai et al., 2022; O Pinheiro et al., 2020; Wang et al., 2021). However, as described in Sharma et al. (2022), dense contrastive losses fail when ground truth correspondences cannot be easily obtained across views, as is the case when including natural temporal augmentations. Sharma et al. (2022) propose to resolve this by performing tracking of local features via optical flow, but this is cumbersome and relies on brittle, hand-engineered features that can be costly to compute. Therefore, in this work, we revert to the standard global contrastive loss formulation, and show that in the right setting, this can be sufficient for capturing complex invariances without the need for dense losses or explicit correspondence matching.

3 METHODS

In our experiments we pretrain image representations using image or video datasets, then transfer them to a range of downstream tasks that test their spatial understanding.

3.1 PRETRAINING VISUAL REPRESENTATIONS

We start by describing our self-supervised baseline for learning representations from images or individual video frames, **MoCLR**, before adding the simple modifications which together make our method for distilling **vi**deos in**to** image representations, which we call **VITO**.

A strong contrastive baseline. MoCLR (Tian et al., 2021) is a simple but powerful way of learning image representations from still images. Given an image $\boldsymbol{x}$ (or a frame from a video), we generate a small number of "views" via random cropping, resizing, and color jittering. Each view $\boldsymbol{v}^l$ is encoded with a feature extractor f_θ into a spatial map of hidden vectors $\boldsymbol{h}_\theta^l = f_\theta(\boldsymbol{v}^l)$ where θ are the parameters of the *online* network being optimized. Following Chen et al. (2020), these hidden vectors are average pooled into a single vector $\hat{\boldsymbol{h}}_\theta^l$ then transformed with a two-layer MLP g_θ, yielding non-linear projections $\boldsymbol{z}_\theta^l = g_\theta(\hat{\boldsymbol{h}}_\theta^l)$ which we rescale such that their euclidean norm is equal to $1/\sqrt{\tau}$, where $\tau = 0.1$.

We wish to enforce invariance of these features across views. In theory, one could regress one projection $\boldsymbol{z}_\theta^i$ onto its target $\boldsymbol{z}_\theta^j$, however it is helpful to stabilize these targets by encoding them instead with specific *target* networks g_ξ and f_ξ, whose parameters ξ vary more slowly, as shown by Grill et al. (2020). We enforce this invariance using the standard contrastive loss of van den Oord et al. (2018)

$$\mathcal{L}^{ij}(\theta;\xi) = -\log \frac{\exp(\boldsymbol{z}_\theta^i \cdot \boldsymbol{z}_\xi^j)}{\exp(\boldsymbol{z}_\theta^i \cdot \boldsymbol{z}_\xi^j) + \sum_n \exp(\boldsymbol{z}_\theta^i \cdot \boldsymbol{z}_\xi^n)}, \tag{1}$$

where $\{\boldsymbol{z}_\xi^n\}_n$ are *negative* features from other images in the batch. We generalize this loss to multiple views by evaluating it for all pairs $\mathcal{L}(\theta;\xi) = \sum_{i \neq j} \mathcal{L}^{ij}(\theta;\xi)$. We update the online network with the loss gradients, and the target network with an exponential moving average of the online network:

$$\theta \leftarrow \text{optimizer}(\theta, \nabla_\theta \mathcal{L}(\theta;\xi), \lambda_\theta) \tag{2}$$

$$\xi \leftarrow (1-\lambda_\xi)\xi + \lambda_\xi \theta \tag{3}$$

By combining the contrastive loss from SimCLR (Chen et al., 2020) and the momentum architecture of MoCo (He et al., 2020) and BYOL (Grill et al., 2020), MoCLR (similarly to MoCo v3 (Chen et al., 2021), which it is akin to) benefits from the best of each approach and been shown to yield state-of-the-art performance of a variety of downstream tasks.

Adapting synthetic augmentations for video frames. The synthetic augmentation pipeline that has become ubiquitous in the self-supervised learning literature has worked extremely well for pretraining models on ImageNet. However, video frames (or even uncurated image data) are typically more complex than ImageNet images. Specifically, uncurated video frames generally have more variable viewpoints, and a larger field-of-view that can cover multiple (not necessarily centered) objects in complex scenes. The standard random resized crop (RRC) operation, which has been found to be essential in self-supervised training methods like SimCLR and BYOL (Chen et al., 2020; Grill et al., 2020), is in fact a very large scale transformation, where the smallest crops can cover only 8% of the original image. While this can enable learning strong invariances across views when the image is relatively homogeneous in content (e.g. featuring a single object), for video frames, this can result in views that have very different semantic content (e.g. entirely different objects), dampening the selectivity of the representation for different object classes.

As a result, we suggest and empirically validate that larger crop sizes (e.g. increasing the minimum crop size to 40%) are beneficial when applying contrastive learning to video frames, even though this results in smaller synthetic scale transformations.

Learning from temporal augmentations. When learning from still images, we apply the random pre-processing pipeline of BYOL (Grill et al., 2020), which includes random cropping, flipping, blurring, and point-wise color transformations (see appendix for the the detail of their distributions):

$$\boldsymbol{v}^1 \sim \mathcal{A}_1(\boldsymbol{x}) \qquad \boldsymbol{v}^2 \sim \mathcal{A}_2(\boldsymbol{x}). \tag{4}$$

When learning from videos, we sample frames according to a distribution $\mathcal{T}$ and transform each frame using the same pipeline as above:

$$\boldsymbol{v}^1 \sim \mathcal{A}_1(\boldsymbol{x}_1) \qquad \boldsymbol{v}^2 \sim \mathcal{A}_2(\boldsymbol{x}_2) \qquad \boldsymbol{x}_1, \boldsymbol{x}_2 \sim \mathcal{T}(\{\boldsymbol{x}_t\}_{t=1,\ldots,T}) \tag{5}$$

Recent works have also suggested similar methodologies for learning from temporal augmentations. In Gordon et al. (2020), $\mathcal{T}$ samples pairs with a fixed time delay. Xu & Wang (2021) choose a distribution that involves uniform sampling over independent chunks of a given video clip. In this work, we propose a simpler approach where we sample from a uniform distribution over the entire video segment of length $T = 2.56s$. In Figure A.1, we show that these sampling schemes induce very different distributions of absolute time differences between pairs of frames. Our marginal sampling scheme is arguably the most natural as the mode of the distribution is at 0, meaning that it is not biased to over-represent any specific time difference (similarly to the random-resized crop operation in space). While not a huge effect, we find that when evaluated on multiple downstream tasks, this temporal sampling method outperforms the other methods (Figure A.1).

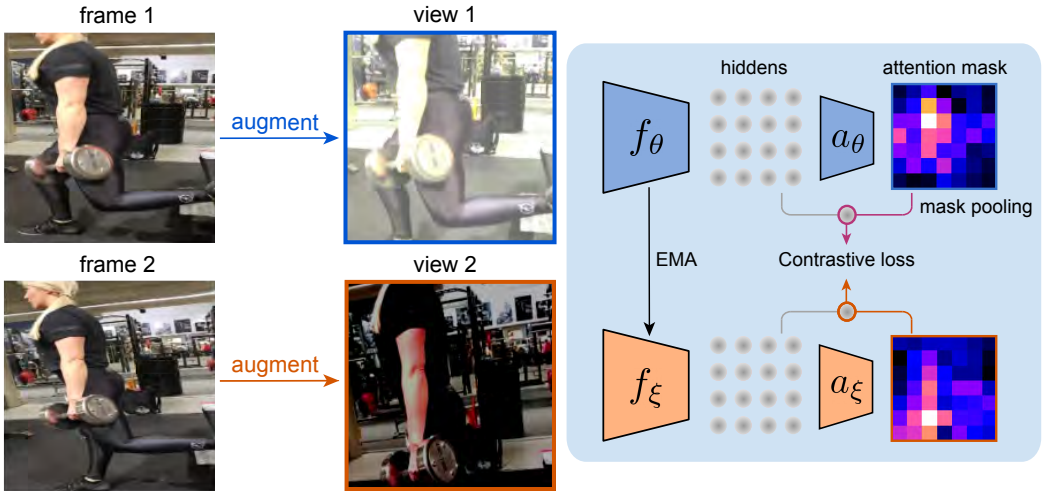

Figure 2: Learning to attend to stable video content. The network takes as input two augmented frames from a video. The attention module takes as input the resulting feature map from one view and produces a mask that attends to features that are likely to be predictable or uniquely identifiable in the other, temporally-displaced view. The attention-gated features are then pooled accordingly. Both the feature extractor and attention module are trained to satisfy the contrastive objective.

Multi-scale contrastive attention pooling. Typical contrastive frameworks use a simple global average pooling of hidden vectors to obtain a single feature for each input that can then be passed through the projector network g_θ. This global average pooling (GAP) aggregates content across the whole image, which can enable higher-level semantic tasks like image classification. However, global pooling makes it hard to localize features in space. As a result, for fine-grained tasks like semantic segmentation, it has been shown that using local (dense) contrastive losses can lead to significant improvements (Wang et al., 2021; Xie et al., 2021; Bai et al., 2022; Hénaff et al., 2021).

Most local contrastive methods require establishing local correspondences across the two views of the input image such that the contrastive loss can be applied to locally pooled features that correspond to similar image content. However, as shown in Sharma et al. (2022), when temporal augmentations are introduced, local correspondences cannot be established without some form of object or point tracking. These methods can be quite cumbersome and involve tuning many dataset-dependent hyperparameters, so in this work, we propose a more general, adaptive method for learning (at multiple scales) what features should be attended to in order to solve the contrastive learning problem across temporally displaced views.

As shown in Figure 2, given a view $\boldsymbol{v}^l$ the feature extractor outputs a spatial map of feature vectors $\boldsymbol{h}_\theta^{l,s} \in \mathcal{R}^{h \times w \times c}$ at a given scale s, where different scales correspond to the outputs of different blocks of a ResNet for example. We introduce a 2-layer attention MLP a_θ^s per scale which outputs an input-dependent mask $\boldsymbol{m}^{l,s} = a_\theta(\boldsymbol{h}_\theta^{l,s})$ that we use to spatially weight and pool hidden vectors:

$$\hat{\boldsymbol{h}}_\theta^{l,s} = \frac{1}{\sum_{i,j} \boldsymbol{m}^{l,s}[i,j]} \sum_{i,j} \boldsymbol{m}^{l,s}[i,j] \; \boldsymbol{h}_\theta^{l,s}[i,j]. \tag{6}$$

Given the attention-pooled features from multiple scales, we concatenate them before transforming them with the two-layer MLP projector: $\boldsymbol{z}_\theta^l = g_\theta(\hat{\boldsymbol{h}}_\theta^l)$ where $\hat{\boldsymbol{h}}_\theta^l = [\hat{\boldsymbol{h}}_\theta^{l,s},\ s \in 1...S]$.

This framework allows the projector function to utilize a multi-scale, adaptively localized representation to solve the contrastive learning problem. This is especially important given the much large dynamic range of scales at which objects can appear in videos as opposed to single-object ImageNet images. This method is related to that of Jetley et al. (2018), which applied a variant of spatial attention pooling in the context of supervised image classification, and more loosely to attention-based backbones which have shown great success in self-supervised learning (Caron et al., 2021). Note however that, rather than requiring specialized network operations, our multi-scale attention pooling module can be a lightweight addition to standard convolutional architectures. In our experiments, we find that for the canonical ResNet-50 architecture, attending over the outputs of the last two ResNet blocks (i.e. $S = 2$) is optimal given our evaluations.

3.2 Evaluating visual representations

Classification has traditionally been the default means of evaluating image representations. Classifying single objects however does not require many of the defining features of real-world scene understanding. Semantic segmentation and object detection provide more relevant tests as they require that a representation understand fine and coarse object boundaries, shapes, sizes, and viewpoints. We therefore evaluate on two semantic segmentation datasets, PASCAL VOC (Everingham et al., 2015) and ADE20K (Zhou et al., 2017), which respectively test object-level understanding and complex scene understanding. We also evaluate on the well-known COCO object detection dataset (Lin et al., 2014) as well as the more challenging long-tailed LVIS dataset (Gupta et al., 2019).

Semantic segmentation on PASCAL and ADE20K. We evaluate ResNet models by attaching a fully-convolutional network (FCN, Long et al. (2015)) and fine-tuning end-to-end, following He et al. (2020). We fine-tune for 45 and 60 epochs on PASCAL and ADE20K respectively, and report the mean intersection over union (mIoU) averaged across 5 runs.

Object detection on COCO and LVIS. We evaluate pretrained ResNet's using the FCOS* architecture, following Hénaff et al. (2022). FCOS* is the implementation of a single-stage detector based on FCOS (Tian et al., 2019), and improved with the collection of techniques from Wu et al. (2020), Zhang et al. (2020), and Feng et al. (2021), full details can be found in Hénaff et al. (2022). The pretrained network is used to initialize the backbone of the FCOS* model, which is then fine-tuned for 30 epochs. We report bounding-box AP averaged across 5 runs.

3.3 Addressing dataset domain mismatch

We begin investigating the potential for video learning with standard datasets including Kinetics, AudioSet, and YouTube-8M. Yet prior work has shown that even self-supervised methods are sensitive to the pretraining distribution. We therefore hypothesized that video pretraining might benefit from a data distribution that is more aligned with the statistics of standard image datasets.

As a test of this hypothesis, we developed a simple data curation pipeline (which we will refer to as *VideoNet*) to filter online videos such that our training data more closely matches the distribution of ImageNet categories. For each of the 1,000 ImageNet categories, we retrieved 5,000 video clips whose title included the category's name or a synonym. We then filtered these videos by applying an image classifier to verify that the videos contained the intended object category. For this we ran a pretrained ResNet-50 ImageNet classifier on the first 100 frames of each video and discarded videos for which the query category was not equal to the ResNet's top-1 prediction for any of the frames.

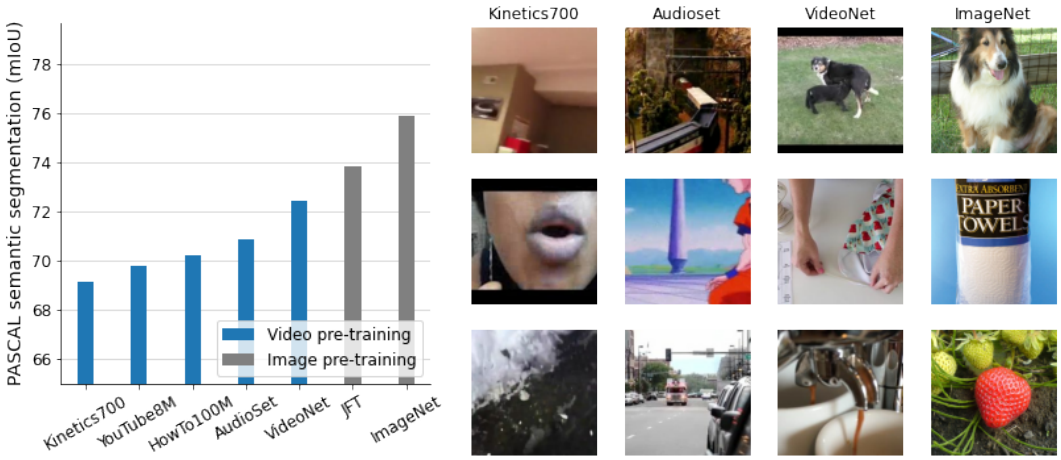

Figure 3: **Impact of pretraining data's spatial content on representation quality.** All models learn from single frames sampled from image datasets (grey bars) or individual videos (blue bars). For all models we use a ResNet-50 backbone and 2-view MoCLR pretraining for 300 epochs.

We additionally discarded videos of less than $10s$ in length. This procedure resulted in a dataset of 1,180,042 videos in total.

We note that while the VideoNet procedure is close in conceptualization to the method used to create the R2V2 dataset proposed by Gordon et al. (2020), it differs in a few ways. First, we utilize full video clips that allow us to uniformly sample frames at any time point rather than the fixed sampling of frames that are $5s$ apart in R2V2. This coarse temporal sampling reduces the total number of frames a model can learn from, but also limits the resolution of temporal augmentations used in the contrastive framework: displacements of $4s$ are more likely to have changes in semantic content than continuously sampled frames from a small interval. Second, by using the ImageNet classifier to filter videos, we can reduce the mismatch with the ImageNet class distribution that can arise from incorrect tagging and noisy labeling of online videos. This is somewhat verified by the fact that only 24% of the retrieved videos met our filtering criteria.

4 Results

4.1 Effect of pretraining data

To demonstrate the effect of the pretraining data distribution on transfer performance, we first pretrain baseline 2-view MoCLR ResNet-50 models on a variety of image and video datasets, where we initially treat video datasets as collections of individual frames. We train each model for 300 ImageNet-equivalent epochs, referred to hereafter as "epochs" (i.e. 1 epoch = learning from 1.28M examples, irrespective of the dataset), such that each model benefits from the same amount of computation. Figure 3 (left) shows their transfer performance on PASCAL semantic segmentation.

Training on standard datasets. As expected, ImageNet pretraining works very well, but pretraining on standard video datasets results in a substantial drop in performance. This performance gap between video and image pretraining can be attributed to a combination of increased complexity and field-of-view of video frames and domain mismatch between the dataset categories (Figure 3, right). This is also evident with uncurated image datasets (like JFT) which have heavy-tailed class distributions, consistent with Tian et al. (2021).

Training with VideoNet curation. We find that applying the same baseline pretraining to frames from our curated video dataset performs better than existing large-scale video datasets like Audioset, but still underperforms image pretraining on JFT and ImageNet (Figure 3).

This result demonstrates the importance of aligning the semantic content between video and image datasets. As a result, we choose this filtered video dataset as our primary pretraining dataset for the rest of this work, with the goal of closing the gap with ImageNet pretraining performance.

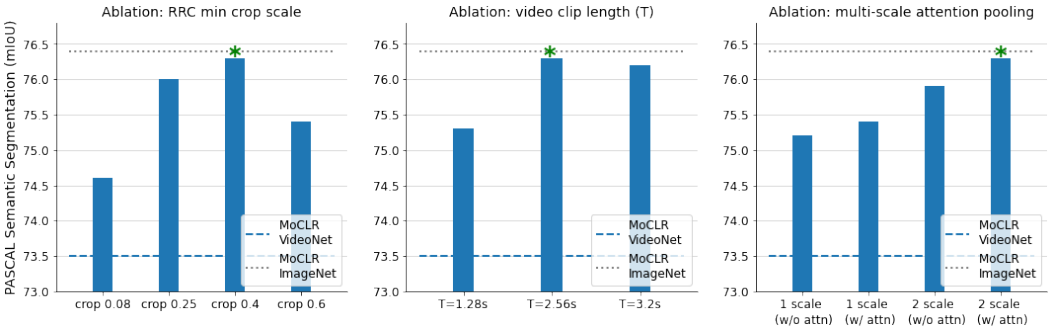

Figure 4: Effects of crop scale, temporal augmentations, and multi-scale attention pooling on representation quality. All ablations are performed relative to VITO's configuration (denoted by a green *) which uses 2-scale attention pooling, less aggressive random resized crop (RRC) scale (0.4 as opposed to 0.08 as is typically used), and temporal augmentations uniformly sampled in a window of length $T = 2.56s$. We also compare to our baseline MoCLR model trained on single frames, either from ImageNet (indicated by dotted gray line) or VideoNet (indicated by the dashed blue line). We evaluate all models by transferring to PASCAL semantic segmentation.

4.2 Leveraging video data with VITO

We now turn to our proposed method for distilling videos into image representations, **VITO**. We first investigate the choices in our pretraining paradigm by varying each of the three main components of our model in isolation: adaptation of spatial crop size, temporal augmentations, and multi-scale attention pooling. For the following ablations, we train VITO and its variants for 300 epochs on VideoNet, using 3 views, then transfer to PASCAL semantic segmentation. Similarly, we pretrain MoCLR baselines on ImageNet and VideoNet for 300 epochs using 3 views.

Adapting spatial augmentations. We validate in Figure 4 (left) our hypothesis that increasing the minimum crop-scale in the RRC operation during training leads to models that generalize better to fine-grained tasks like semantic segmentation. Specifically, we find that a minimum crop scale of 0.4 (as opposed to the traditional 0.08) results in the best transfer performance (+1.7% mIoU).

Temporal augmentations. As described in section 3.1, for each training example, we sample 3 views using marginal sampling of each frame from the video clip of length $T = 2.56$ seconds. This length determines the distribution (and accordingly the mean) time difference between any pair of frames. As a result, the total length impacts the time-scale over which the contrastive model learns invariances. We verify our choice by varying the total length of clips. While going to longer time-scales $T = 3.2s$ does not hurt performance much, we find a significant improvement over using shorter clips (e.g. $T = 1.28s$, +1.0% mIoU; Figure 4, center). This suggests that invariance to the rich temporal deformations present in longer video clips is indeed a beneficial criterion for learning fine-grained spatial representations. Note however that with an optimum interval of $2.56s$, more coarsely sampled video datasets may be of limited utility.

Multi-scale attention pooling. We breakdown the proposed multi-scale contrastive attention pooling into components to study the effect of both the multi-scale aspect separately from the introduction of attention pooling. While we find only modest gains from adding attention pooling to a single-scale version of the model (+0.2% mIoU), we find that the 2-scale (without attention pooling) model significantly improves over the single scale model (+0.6% mIoU). Interestingly, however, we find that the combination of the 2-scale model with attention pooling has a synergistic effect leading to a +1% mIoU improvement over the single-scale attention model.

Combined, the three modifications VITO makes to the contrastive framework result in a 2.8% mIoU improvement over MoCLR pretrained on VideoNet, closing the gap with ImageNet pretraining when transferring to PASCAL semantic segmentation. In the next sections, we seek to understand the mechanism underlying this improvement, and validate it on other downstream tasks.

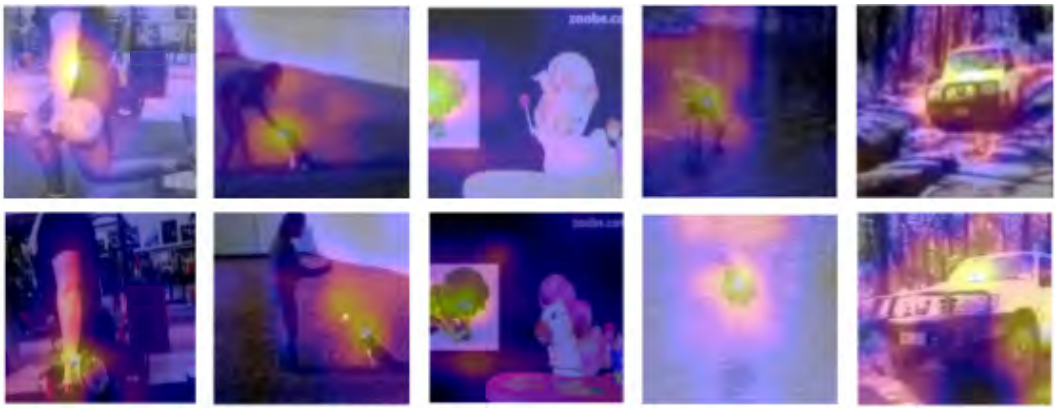

Figure 5: Example augmented frames with overlaid (resized) learned attention masks. Attention is computed from the output of the final block of the VITO trained ResNet-50. Crucially, the attention masks are computed independently, such that the attention module can only use spatial cues.

4.3 SEMANTIC BINDING WITH CONTRASTIVE ATTENTION POOLING

The ablation study demonstrated that multi-scale attention improves the performance of VITO in semantic segmentation. To probe why this may be, we visualize and interpret the learned attention masks (Figure 5). For simplicity, we only visualize the masks from the coarsest scale (output feature map), but the interpretation naturally extends to the multi-scale version as these masks are learned with independent attention modules.

Because the attention masks are not computed jointly across each view, for a given video frame, the attention module must marginalize over the training data to make a statistical prediction—what should be attended to in the first view in order to minimize the contrastive loss across possible second views. Specifically, the attention must focus on content that is most likely to be stable across time while still being discriminative (or unique) relative to other frames from other videos. Different examples appear to trade-off these criteria differently, yet systematically. For example, in the third column of Figure 5 even though the animated characters on the right side of both frames may be discriminative content, the attention module has learned to focus on the static picture on the left as it is the content that is most likely to be stable across time. For this pair of frames the prediction is correct—the attention disregards content that is changing too abruptly—despite not having access to motion cues. On the other hand, the example in the fourth column demonstrates a scenario where the model has attended to stable, but primarily discriminative content (the bird) rather than the background, which is also very stable but most likely less unique relative to other videos.

Even beyond the ability to localize stable, yet discriminative content, it seems that our method also enables "semantic binding" of visually different, but semantically related features. This can be seen in the first pair of frames, as the model has learned to associate an arm or elbow (in the first frame) with the dumbbell (in the second frame), demonstrating an understanding that these two semantically related concepts co-occur and thus are predictive of one another given the right embedding.

Binding co-occuring features appears as an intuitive explanation for why these representations would perform well on semantic segmentation. It is particularly interesting that training end-to-end with a standard contrastive loss can produce complex behavior reminiscent of the DINO approach (Caron et al., 2021) even though we use a single, two-layer MLP attention module, as opposed to large-scale transformer architectures which use attention throughout the network.

4.4 COMPARISONS TO PRIOR WORK

Having shown that VITO is potentially learning novel and powerful visual representations that transfer well to PASCAL semantic segmentation, we now present in Table 1 the transfer performance of VITO against many recent strong image and video pretraining baselines on the full set of semantic segmentation and object detection evaluations.

Table 1: VITO closes the gap between video- and image-pretraining of ResNet-50. For external models, we finetune publicly available checkpoints.

			Semantic segmentation		Object detection	
Pretraining	Dataset	Epochs	PASCAL	ADE20K	COCO	LVIS
Random Init			53.0	27.9	39.0	21.1
Methods pretraining on video datasets						
VFS (Xu & Wang, 2021)	K400	100	63.9	31.4	41.6	23.2
VIVI (Tschannen et al., 2020)	YouTube8M	192	65.8	34.2	41.3	23.2
VINCE (Gordon et al., 2020)	R2V2	200	69.0	35.7	42.4	24.4
CycleContrast (Wu & Wang, 2021)	R2V2	200	69.2	35.6	42.8	24.5
MMV (Alayrac et al., 2020)	AS + HT	1600	70.6	32.5	41.3	24.2
VITO	VideoNet	200	**75.5**	**39.2**	**43.6**	**25.6**
MoCLR	VideoNet	300	73.5	38.0	43.1	24.8
VITO	AudioSet	300	73.6	38.5	43.2	25.0
VITO	VideoNet	300	**76.3**	**39.4**	**44.0**	**25.7**
Methods pretraining on ImageNet						
Supervised	ImageNet	200	71.3	33.5	44.2	25.2
BYOL (Grill et al., 2020)	ImageNet	300	76.1	38.8	43.7	25.5
MoCLR (Tian et al., 2021)	ImageNet	300	76.4	39.2	43.9	25.8
DINO (Caron et al., 2021)	ImageNet	300	76.1	39.0	44.3	26.4

Comparison to prior video-pretraining. Given a similar computational budget as prior works (200 epochs and 3 views) VITO delivers large gains over all prior methods. For example, VITO improves over VIVI (Tschannen et al., 2020) by 10%/5%/2%/2%, highlighting the importance of data curation and our contrastive formulation. VITO improves over VINCE (Gordon et al., 2020) by 7%/4/1%/1%, highlighting the importance of fine-grained temporal augmentations. Finally, VITO improves even over MMV (Alayrac et al., 2020) by 5%/7%/2%/2%, despite their use of large-scale text supervision, highlighting the relevance of video-only learning. We revisit the importance of data curation by applying VITO to AudioSet, and find that despite a drop in performance relative to VideoNet-pretraining, it continues to outperform all prior video learning methods.

Comparison to ImageNet pretraining. Finally, we compare our VITO-pretrained VideoNet model to a host of state-of-the-art ImageNet-pretrained methods. Surprisingly, we find VITO to be competitive with the best of such methods, outperforming BYOL and DINO on PASCAL, matching BYOL and MoCLR on COCO and LVIS, and surpassing all methods on ADE20K. VITO largely surpasses supervised ImageNet pretraining on 3 downstream evaluations. This is, to the best of our knowledge, the first example of video pretraining competing with ImageNet pretraining on such tasks.

5 Discussion

We propose VITO, a simple method for distilling videos into image representations. The key features of our method include improved dataset curation, adapting standard synthetic augmentations to video frames, and using attention-guided contrastive learning. Taken together, these components have helped us achieve, to our knowledge, the first results of video pretraining closing the gap with ImageNet pretraining on canonical scene understanding benchmarks.

We believe this work can be a foundation for future video pretraining efforts, as our approach is powerful, yet simple and extensible in almost every aspect. For example, because we base our learning paradigm on standard architectures and contrastive learning methods, it is easy to extend or adapt our approach to leverage continuing advancements in image-based contrastive, and more generally, self-supervised learning objectives. Additionally, while we have shown the benefits of a simple attention module for learning from video data, there are great opportunities to extend our approach to take advantage of more powerful attentional architectures. In sum, despite the many successes in video representation learning, our results suggest that there is a great untapped potential for video pretraining to become the default paradigm for learning image representations.

6 REPRODUCIBILITY STATEMENT

We will release our pretrained models along with the code needed to implement the VITO model architecture. Pretraining and evaluation details about architectures, optimization, and hyperparameters have all been detailed in the appendix. The VideoNet procedure for curating video datasets can be reproduced with standard ImageNet classifiers and publicly available online videos.

REFERENCES

Pulkit Agrawal, Joao Carreira, and Jitendra Malik. Learning to see by moving. In *ICCV*, 2015.

Jean-Baptiste Alayrac, Adria Recasens, Rosalia Schneider, Relja Arandjelović, Jason Ramapuram, Jeffrey De Fauw, Lucas Smaira, Sander Dieleman, and Andrew Zisserman. Self-supervised multimodal versatile networks. *Advances in Neural Information Processing Systems*, 33:25–37, 2020.

Yutong Bai, Xinlei Chen, Alexander Kirillov, Alan Yuille, and Alexander C Berg. Point-level region contrast for object detection pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16061–16070, 2022.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9650–9660, 2021.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PMLR, 2020.

Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9640–9649, 2021.

Ishan Dave, Rohit Gupta, Mamshad Nayeem Rizve, and Mubarak Shah. Tclr: Temporal contrastive learning for video representation. *Computer Vision and Image Understanding*, 219:103406, 2022.

Michael Dorkenwald, Fanyi Xiao, Biagio Brattoli, Joseph Tighe, and Davide Modolo. Scvrl: Shuffled contrastive video representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4132–4141, 2022.

Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *International journal of computer vision*, 111(1):98–136, 2015.

Christoph Feichtenhofer, Haoqi Fan, Bo Xiong, Ross Girshick, and Kaiming He. A large-scale study on unsupervised spatiotemporal representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3299–3309, 2021.

Christoph Feichtenhofer, Haoqi Fan, Yanghao Li, and Kaiming He. Masked autoencoders as spatiotemporal learners. *arXiv preprint arXiv:2205.09113*, 2022.

Chengjian Feng, Yujie Zhong, Yu Gao, Matthew R Scott, and Weilin Huang. TOOD: Task-aligned one-stage object detection. In *Int. Conf. Comput. Vis.*, 2021.

Daniel Gordon, Kiana Ehsani, Dieter Fox, and Ali Farhadi. Watching the world go by: Representation learning from unlabeled videos. *arXiv preprint arXiv:2003.07990*, 2020.

Ross Goroshin, Joan Bruna, Jonathan Tompson, David Eigen, and Yann LeCun. Unsupervised learning of spatiotemporally coherent metrics. In *Proceedings of the IEEE international conference on computer vision*, pp. 4086–4093, 2015.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in Neural Information Processing Systems*, 33, 2020.

Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5356–5364, 2019.

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

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9729–9738, 2020.

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.

Olivier J Hénaff, Skanda Koppula, Jean-Baptiste Alayrac, Aaron van den Oord, Oriol Vinyals, and João Carreira. Efficient visual pretraining with contrastive detection. In *ICCV*, 2021.

Olivier J Hénaff, Skanda Koppula, Evan Shelhamer, Daniel Zoran, Andrew Jaegle, Andrew Zisserman, João Carreira, and Relja Arandjelović. Object discovery and representation networks. *arXiv preprint arXiv:2203.08777*, 2022.

Jarmo Hurri and Aapo Hyvärinen. Simple-cell-like receptive fields maximize temporal coherence in natural video. *Neural Computation*, 15(3):663–691, 2003.

Saumya Jetley, Nicholas A Lord, Namhoon Lee, and Philip HS Torr. Learn to pay attention. *arXiv preprint arXiv:1804.02391*, 2018.

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

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 *Eur. Conf. Comput. Vis.*, pp. 740–755. Springer, 2014.

Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3431–3440, 2015.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *Int. Conf. Learn. Represent.*, 2019.

Jingcheng Ni, Nan Zhou, Jie Qin, Qian Wu, Junqi Liu, Boxun Li, and Di Huang. Motion sensitive contrastive learning for self-supervised video representation. *arXiv preprint arXiv:2208.06105*, 2022.

Pedro O O Pinheiro, Amjad Almahairi, Ryan Benmalek, Florian Golemo, and Aaron C Courville. Unsupervised learning of dense visual representations. *Advances in Neural Information Processing Systems*, 33:4489–4500, 2020.

Deepak Pathak, Ross Girshick, Piotr Dollár, Trevor Darrell, and Bharath Hariharan. Learning features by watching objects move. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2701–2710, 2017.

Rui Qian, Tianjian Meng, Boqing Gong, Ming-Hsuan Yang, Huisheng Wang, Serge Belongie, and Yin Cui. Spatiotemporal contrastive video representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6964–6974, 2021.

Adrià Recasens, Pauline Luc, Jean-Baptiste Alayrac, Luyu Wang, Ross Hemsley, Florian Strub, Corentin Tallec, Mateusz Malinowski, Viorica Patraucean, Florent Altché, Michal Valko, Jean-Bastien Grill, Aäron van den Oord, and Andrew Zisserman. Broaden your views for self-supervised video learning. In *Int. Conf. Comput. Vis.*, 2021.

Pierre Sermanet, Corey Lynch, Yevgen Chebotar, Jasmine Hsu, Eric Jang, Stefan Schaal, Sergey Levine, and Google Brain. Time-contrastive networks: Self-supervised learning from video. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1134–1141. IEEE, 2018.

Yash Sharma, Yi Zhu, Chris Russell, and Thomas Brox. Pixel-level correspondence for self-supervised learning from video. *arXiv preprint arXiv:2207.03866*, 2022.

Yonglong Tian, Olivier J Henaff, and Aäron van den Oord. Divide and contrast: Self-supervised learning from uncurated data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10063–10074, 2021.

Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. FCOS: Fully convolutional one-stage object detection. In *Int. Conf. Comput. Vis.*, 2019.

Martine Toering, Ioannis Gatopoulos, Maarten Stol, and Vincent Tao Hu. Self-supervised video representation learning with cross-stream prototypical contrasting. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 108–118, 2022.

Michael Tschannen, Josip Djolonga, Marvin Ritter, Aravindh Mahendran, Neil Houlsby, Sylvain Gelly, and Mario Lucic. Self-supervised learning of video-induced visual invariances. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13806–13815, 2020.

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

Xiaolong Wang and Abhinav Gupta. Unsupervised learning of visual representations using videos. In *Proceedings of the IEEE international conference on computer vision*, pp. 2794–2802, 2015.

Xinlong Wang, Rufeng Zhang, Chunhua Shen, Tao Kong, and Lei Li. Dense contrastive learning for self-supervised visual pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3024–3033, 2021.

Laurenz Wiskott and Terrence J Sejnowski. Slow feature analysis: Unsupervised learning of invariances. *Neural computation*, 14(4):715–770, 2002.

Haiping Wu and Xiaolong Wang. Contrastive learning of image representations with cross-video cycle-consistency. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10149–10159, 2021.

Shengkai Wu, Xiaoping Li, and Xinggang Wang. IoU-aware single-stage object detector for accurate localization. *Image and Vision Computing*, 2020.

Zhenda Xie, Yutong Lin, Zheng Zhang, Yue Cao, Stephen Lin, and Han Hu. Propagate yourself: Exploring pixel-level consistency for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16684–16693, 2021.

Jiarui Xu and Xiaolong Wang. Rethinking self-supervised correspondence learning: A video frame-level similarity perspective. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10075–10085, 2021.

Yang You, Igor Gitman, and Boris Ginsburg. Large batch training of convolutional networks. *arXiv preprint arXiv:1708.03888*, 2017.

Shifeng Zhang, Cheng Chi, Yongqiang Yao, Zhen Lei, and Stan Z. Li. Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020.

Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 633–641, 2017.

A APPENDIX

A.1 IMPLEMENTATION: DATA PRE-PROCESSING

Self-supervised pretraining. Each frame is randomly augmented by composing the following operations, each applied with a given probability:

1. random cropping: a random patch of the image is selected, whose area is uniformly sampled in $[s \cdot \mathcal{A}, \mathcal{A}]$, where $\mathcal{A}$ is the area of the original image, and whose aspect ratio is logarithmically sampled in $[3/4, 4/3]$. s is a scale hyper-parameter set to 0.08 when learning from ImageNet, and 0.4 when learning from videos. Regardless, the patch is then resized to 224 $\times$224 pixels using bicubic interpolation;
2. horizontal flipping;
3. color jittering: the brightness, contrast, saturation and hue are shifted by a uniformly distributed offset;
4. color dropping: the RGB image is replaced by its grey-scale values;
5. gaussian blurring with a 23$\times$23 square kernel and a standard deviation uniformly sampled from $[0.1, 2.0]$;
6. solarization: a point-wise color transformation $x \mapsto x \cdot 1_{x<0.5} + (1-x) \cdot 1_{x \geq 0.5}$ with pixels x in $[0, 1]$.

The augmented frames $\boldsymbol{v}^1$ and $\boldsymbol{v}^2$ result from augmentations sampled from distributions $\mathcal{A}_1$ and $\mathcal{A}_2$ respectively. These distributions apply the primitives described above with different probabilities, and different magnitudes. The following table specifies these parameters for the BYOL framework Grill et al. (2020), which we adopt without modification.

Parameter	$\mathcal{A}_1$	$\mathcal{A}_2$
Random crop probability	1.0	
Flip probability	0.5	
Color jittering probability	0.8	
Color dropping probability	0.2	
Brightness adjustment max	0.4	
Contrast adjustment max	0.4	
Saturation adjustment max	0.2	
Hue adjustment max	0.1	
Gaussian blurring probability	1.0	0.1
Solarization probability	0.0	0.2

Transfer to COCO and LVIS. The target resolution is 800$\times$1024. During testing, an image is resized by a factor s while preserving the aspect ratio, such that it is tightly contained inside the target resolution, and then padded. When fine-tuning, the image is rescaled by a factor of $u \cdot s$ where u is uniformly sampled in $[0.8, 1.25]$, and is then cropped or padded to the target resolution.

Transfer to PASCAL and ADE20K. During training, images are randomly flipped and scaled by a factor in $[0.5, 2.0]$. Training and testing are performed with 512$\times$512-resolution images. When fine-tuning on ADE20K, we aditionally use photometric transformations from the mmseg[1] codebase.

A.2 IMPLEMENTATION: OPTIMIZATION

Self-supervised pretraining. We pretrain ResNet-50 using the LARS optimizer (You et al., 2017) with a batch size of 4096 split across 128 Cloud TPU v3 workers. We adopt the optimization details of BYOL, scaling the learning rate linearly with the batch size and decaying it according to a cosine schedule. The base learning rate is 0.3 and the weight decay is 10^{-6}.

Transfer to PASCAL and ADE20K with Fully Convolutional Networks. We fine-tune for 45 epochs on PASCAL or 60 epochs on ADE20K with stochastic gradient descent, with a batch size of

[1] `https://github.com/open-mmlab/mmsegmentation`

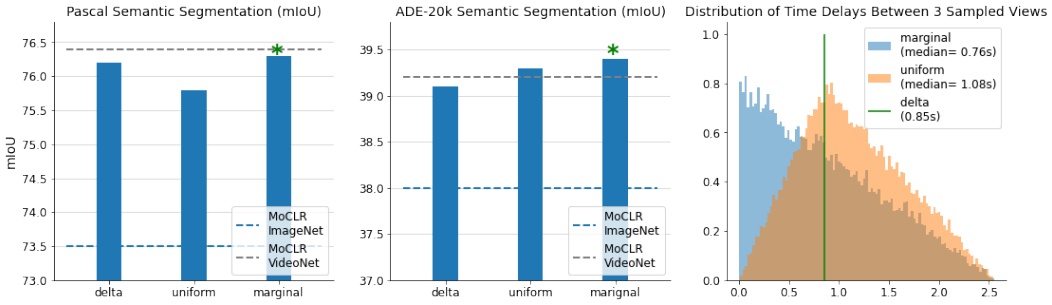

Figure A.1: Ablating different temporal sampling schemes. *Delta* refers to fixed time sampling beteween frames as in Gordon et al. (2020). *Uniform* refers to chunking time into non-overlapping blocks and uniformly sampling within each chunk as in Xu & Wang (2021). *Marginal* sampling (ours) refers to simple uniform sampling from the full video clip of length $T = 2.56s$. First two panels show that marginal sampling is best overall across transfer to PASCAL and ADE20K. Third panel shows the distribution of absolute time-differences between any two pairs of frames under each sampling scheme (assuming 3 views are sampled per clip).

16 and weight decay of 0.005. The learning rate is initially set to 0.04 and decayed exponentially with a factor of 0.9^n where n is the iteration number. When fine-tuning external models, we sweep over the base learning rate and weight decay and report performance given the optimal configuration.

Transfer to COCO and LVIS with FCOS*. The network is fine-tuned for 30 epochs with batch size 128 split across 16 workers, with AdamW (Loshchilov & Hutter, 2019), weight decay 10^{-4} and base learning rate of 10^{-3}. The learning rate rises linearly for $\frac{1}{4}$ of an epoch, and is dropped twice by a factor of 10, after $\frac{2}{3}$ and $\frac{8}{9}$ of the total training time.