
Long-Short Temporal Contrastive Learning of Video Transformers

Jue Wang¹

Gedas Bertasius¹

Du Tran¹

Lorenzo Torresani^{1,2}

¹ Facebook AI ² Dartmouth College

Abstract

Video transformers have recently emerged as a competitive alternative to 3D CNNs for video understanding. However, due to their large number of parameters and reduced inductive biases, these models require supervised pretraining on large-scale image datasets to achieve top performance. In this paper, we empirically demonstrate that self-supervised pretraining of video transformers on video-only datasets can lead to action recognition results that are on par or better than those obtained with supervised pretraining on large-scale image datasets, even massive ones such as ImageNet-21K. Since transformer-based models are effective at capturing dependencies over extended temporal spans, we propose a simple learning procedure that forces the model to match a long-term view to a short-term view of the same video. Our approach, named Long-Short Temporal Contrastive Learning (LSTCL), enables video transformers to learn an effective clip-level representation by predicting temporal context captured from a longer temporal extent. To demonstrate the generality of our findings, we implement and validate our approach under three different self-supervised contrastive learning frameworks (MoCo v3, BYOL, SimSiam) using two distinct video-transformer architectures, including an improved variant of the Swin Transformer augmented with space-time attention. We conduct a thorough ablation study and show that LSTCL achieves competitive performance on multiple video benchmarks and represents a convincing alternative to supervised image-based pretraining.

1 Introduction

Since the introduction of AlexNet [34], deep convolutional neural networks (CNNs) have emerged as the prominent model in numerous computer vision tasks [29, 20, 48, 21, 57, 56]. More recently, the transformer architecture has received much attention due to its impressive performance in the field of natural language processing (NLP) [14]. While CNNs rely on the local operation of convolution, the building block of transformers is self-attention [53] which is particularly effective at modelling long-range dependencies. In the image domain, the Vision Transformer (ViT) [15] was proposed as a pure convolution-free architecture which uses self-attention between non-overlapping patches in all layers of the model. ViT was shown to be competitive with state-of-the-art CNNs on the task of image categorization. In the last few months, several adaptations of ViT to video have been proposed [5, 2, 3]. In order to capture salient temporal information from the video, these works typically extend the self-attention mechanism to operate along the time axis in addition to within each frame. Since video transformers have large numbers of parameters and fewer inductive biases compared to CNN architectures, they require large-scale pretraining on supervised image datasets, such as ImageNet-21K [46] or JFT [3], in order to achieve top performance.

Self-supervised learning has been demonstrated to be an effective solution to eliminate the need for large-scale supervised pretraining of transformers both in NLP [14] as well as in image-analysis [51, 8]. In this work, we show that self-supervised video pretraining provides an effective way for training

video transformers without any external data. Specifically, we introduce Long-Short Temporal Contrastive Learning (LSTCL), a contrastive formulation that maximizes representation similarity between a long video clip (say, 8 seconds long) and a much shorter clip (say, 2 seconds long) where both clips are sampled from the same video. We argue that by training the short-clip representation to match the long-clip representation, the model is forced to extrapolate from a short extent the contextual information exhibited in the longer temporal span. As the long clip includes temporal segments not included in the short clip, this self-supervised strategy trains the model to anticipate the future and to predict the past from a small temporal window in order to match the representation extracted from the long clip. We believe that this is a good pretext for video representation learning, as it can be accomplished only by a successful understanding and recognition of the structure and correlation of atomic actions in a long video. Furthermore, such framework is particularly suitable for video transformers as they have been recently shown to effectively capture long-term temporal cues [5]. In this work we demonstrate that these long-term temporal cues can be effectively encoded into a short-range clip-level representation leading to a substantial improvement in video classification performance.

To demonstrate the generality of our findings, we experiment with two different video transformer architectures whose code is publicly available. The first is TimeSformer [5], which reduces the computational cost of self-attention over the 3D video volume by means of a space-time factorization. The second architecture is the Swin transformer [35], which we further extend into a 3D version, dubbed Space-Time Swin transformer, that computes hierarchical spatiotemporal self-attention by using 3D shifting windows. We show that our unsupervised LSTCL pretraining scheme allows both of these video transformers to outperform their respective counterparts pretrained with full supervision on the large-scale ImageNet-21K dataset.

In summary, the contributions of this paper can be summarized as follows:

- We introduce Long-Short Temporal Contrastive Learning (LSTCL), which trains the clip-level representation to effectively encode temporal context from the longer video.
- We demonstrate that in the context of video transformer models, our proposed LSTCL pretraining provides an effective alternative to large-scale supervised pretraining on images.
- We propose a Space-Time Swin transformer for spatiotemporal feature learning, and show that it achieves strong results on multiple action recognition benchmarks.

2 Related work

Self-supervised Learning in Images. Early attempts at self-supervised visual representation learning used a variety of pretext tasks, such as image rotation prediction [33], auto-encoder learning [54, 43, 49], or solving jigsaw puzzles [40]. In comparison, recent approaches in self-supervised learning leverage contrastive learning [28, 10, 11, 13, 8, 45, 27]. The idea is to generate two views of the same image through data augmentation and then minimize the distance of their representations while, optionally, maximizing the distance to other images [10, 28]. One disadvantage of contrastive learning is that it requires a large number of negative examples, which implies a large batch size [10] or the use of a memory bank [28]. To tackle the high computational cost of such contrastive approaches, several recent methods proposed to eliminate the reliance on negative samples [12, 13, 25, 7].

Self-supervised Learning in Videos. Early methods for self-supervised video representation learning focused on predictive spatiotemporal ordering tasks [22, 31, 1, 50, 59, 39]. Just like in the image domain, more recent approaches [26, 19, 27, 44] adopt contrastive learning objectives. Our method also falls into the latter category of contrastive approaches. In comparison to prior contrastive video methods, we propose a contrastive formulation where a positive pair is generated from a short clip and a long clip, both of which are sampled from the same video. This pushes our model to learn a short clip-level representation that captures global video-level context.

The approach that is most closely related to our own is the BraVe system, described in concurrent work [45]. BraVe shares the same underlying idea of training a model to match a long (broad) view to a short (narrow) view of the same video. However, our work differs in several aspects. First of all, our main focus is to leverage self-supervised learning as a means to train video transformers without labeled image data while BraVe is applied to 3D CNNs. Video transformers are emerging as a competitive alternative to 3D CNNs. However, as discussed earlier, they suffer from the limitation

of requiring image-based supervised pretraining. Thus, we believe that this is an important and timely problem to address. Additionally, we note that our LSTCL is a lot simpler than BraVe: while our model uses shared parameters, a single projection network, and a single prediction network, BraVe requires separate backbones, separate projection networks, and separate prediction networks for the two views in order to achieve the best performance; furthermore, while LSTCL can be applied with any traditional contrastive loss (as demonstrated by our experiments with MoCo v3, BYOL, and SimSiam), BraVe uses a combination of two specific regression objectives (broad-to-narrow and narrow-to-broad) and employs distinct augmentation strategies for the two views. Despite the bare-bone simplicity of our learning formulation, we demonstrate that it delivers impressive results, elevating the accuracy of video transformers to the state-of-the-art on challenging action classification benchmarks *without* the need of any supervised image-level pretraining.

Transformers in Vision. Transformer-based models [53, 14] currently define the state-of-the-art for the majority of natural language processing (NLP) tasks. Similarly, there have also been several attempts to adopt transformer-based architectures for vision problems. Initially, these attempts focused on architectures mixing convolution with self-attention [58, 60, 6, 30, 61]. The recent introduction of Vision Transformer (ViT) [15] has demonstrated that it is possible to achieve competitive image classification results with a convolution-free architecture. To increase the data-efficiency aspect of the original ViT, Touvron et al. [51] proposed a training recipe based on distillation. Lastly, the recently introduced Swin transformer [35] significantly reduces the number of parameters and the cost of ViT by employing local rather than global self-attention.

ViT models were also adapted to the video domain by introducing different forms of spatiotemporal self-attention [5, 3, 2]. However, due to their large number of parameters, these models typically require large amounts of training data, which typically comes in the form of a large scale labeled dataset (such as ImageNet or JFT). To address this issue, Fan et al. [16] introduced a multi-scale vision transformer (MViT), which uses a much smaller number of parameters and can be successfully trained from scratch. Instead of reducing the model capacity, as done in MViT, we show that it is possible to train large-capacity video transformer models without any external data by means of our proposed LSTCL self-supervised learning framework.

3 Video Transformers

Several recent attempts have been made to extend ViT to the video domain [5, 3, 16, 2]. Most video transformers share common principles, which we review below. We then discuss specific designs differentiating the video transformers considered in our experiments.

3.1 Overview

Video decomposition. Given an input video clip $X \in \mathbb{R}^{H \times W \times C \times T}$ that has T frames of $H \times W$ pixels, each frame is decomposed into N non-overlapping patches, where each patch is of size $P \times P$ and $N = (HW)/P^2$. The original video clip is then represented as a sequence of vectorized patches, with $\mathbf{p}_{(i,t)} \in \mathbb{R}^{P^2 \cdot C}$ denoting the vector corresponding to the i -th patch in frame t , with $i = 1, 2, \dots, N$ and $t = 1, 2, \dots, T$.

Linear and positional embeddings. Each patch $\mathbf{p}_{(i,t)}$ is linearly embedded into a feature vector $\mathbf{z}_{(i,t)}^0 \in \mathbb{R}^D$ obtained by means of a learnable matrix $W \in \mathbb{R}^{D \times (P^2 \cdot C)}$ and a learnable vector $\mathbf{e}_{(i,t)} \in \mathbb{R}^D$ representing the spatial-temporal positional embedding: $\mathbf{z}_{(i,t)}^0 = W\mathbf{p}_{(i,t)} + \mathbf{e}_{(i,t)}$.

Multi-headed attention. The multi-headed self-attention block (MHA) is the key component of the transformer, which is carried out by means of a query-key-value computation for each patch, and it is interleaved with layer normalization [4] (LN) and a multilayer perceptron (MLP) within each block ℓ . Thus, the intermediate representation $\tilde{\mathbf{z}}^\ell$ for a patch in block ℓ is obtained from its features in the previous block, as:

$$\tilde{\mathbf{z}}^\ell = \text{MHA}(\text{LN}(\mathbf{z}^{\ell-1})) + \mathbf{z}^{\ell-1}, \quad \mathbf{z}^\ell = \text{MLP}(\text{LN}(\tilde{\mathbf{z}}^\ell)) + \tilde{\mathbf{z}}^\ell. \quad (1)$$

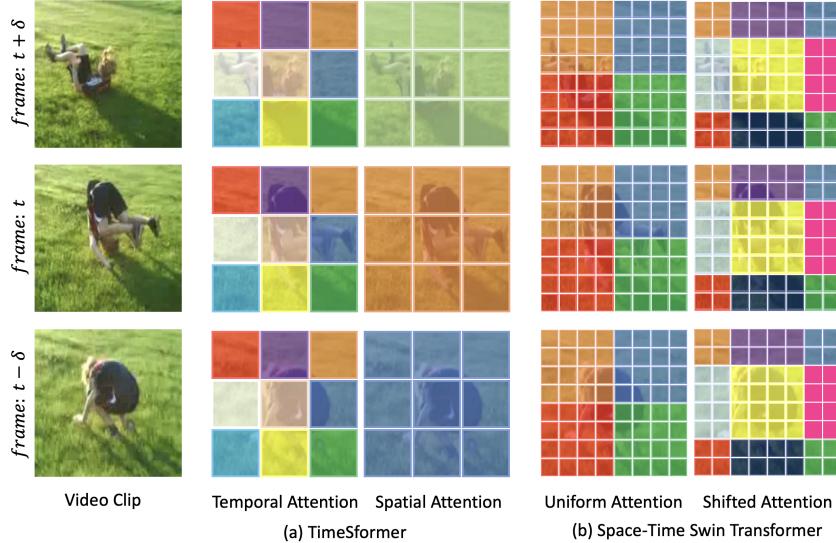


Figure 1: An illustration of the self-attention mechanisms in TimeSformer and Space-Time (ST) Swin Transformer. Each column in the figure depicts a different self-attention block. Patches that have the same color are compared during the self-attention computation.

Classification. As in BERT [14], a classification token $p_{(0,0)}$ is added at the beginning of the input sequence. In the last layer of the network, a linear layer with softmax activation function is attached to the classification token in order to output the final classification probabilities.

3.2 TimeSformer

TimeSformer [5] is a direct extension of ViT [15] to the video domain. It uses two independent multi-head attention blocks for spatial and temporal self-attention. As shown in Figure 1, the spatial self-attention compares the query patch only to image patches appearing in the same frame. Conversely, the temporal self-attention compares the query patch to the image patches in the same spatial location but from the other frames. The decomposition over space and time dramatically reduces the cost of self-attention compared to a dense comparison over all pairs of patches of the video. Thus, the feature representation is computed as:

$$\begin{aligned} z_t^\ell &= \text{MHA}_{\text{Time}}(\text{LN}(z^{\ell-1})) + z^{\ell-1} \\ z_s^\ell &= \text{MHA}_{\text{Space}}(\text{LN}(z_t^\ell)) + z_t^\ell \\ z^\ell &= \text{MLP}(\text{LN}(z_s^\ell)) + z_s^\ell \end{aligned} \tag{2}$$

3.3 Space-Time Swin Transformer

Compared to ViT, the Swin Transformer [35] applies self-attention locally. The features are learned hierarchically by aggregating information from local neighborhoods of patches in each layer. Here we adapt the original Swin transformer, which was introduced for still-images, to video. We name this new variant Space-Time Swin Transformer (ST Swin). Instead of considering 2D neighborhoods of image patches for self-attention computation, ST Swin uses local 3D space-time volumes. Specifically, as proposed in the original paper [35], ST Swin uses two distinct self-attention mechanisms: uniform partition and shifted partition. In our case, both of these self-attention schemes are adapted to video by considering the temporal dimension in the local patch neighborhoods. As shown in Figure 1, the uniform partition splits the entire clip into 4 non-overlapping 3D sections, with each section sharing the same partition index. Spatiotemporal self-attention is then computed between image patches that have the same partition index. Similarly, shifted partition generates multiple non-overlapping 3D sections at different scales, and spatiotemporal patches within each section are compared for

self-attention computation. The uniform partition and the shifted partition are stacked to form two successive attention blocks, which implement cross-window connections further increasing the model capacity. Thus, the complete transformation carried out in each layer ℓ of the ST Swin transformer can be summarized as follows::

$$\begin{aligned}\mathbf{z}_u^\ell &= \text{MHA}_{\text{Uniform}}(\text{LN}(\mathbf{z}^{\ell-1})) + \mathbf{z}^{\ell-1} \\ \mathbf{z}^\ell &= \text{MLP}(\text{LN}(\mathbf{z}_u^\ell)) + \mathbf{z}_u^\ell \\ \mathbf{z}_s^{\ell+1} &= \text{MHA}_{\text{Shift}}(\text{LN}(\mathbf{z}^\ell)) + \mathbf{z}^\ell \\ \mathbf{z}^{\ell+1} &= \text{MLP}(\text{LN}(\mathbf{z}_s^{\ell+1})) + \mathbf{z}_s^{\ell+1}\end{aligned}\tag{3}$$

We adopt the 3D relative positional embedding and the patch merging strategy used in the original Swin [35]. However, we only merge image patches along the spatial axis while maintaining fixed temporal resolution through the layers.

4 Long-Short Temporal Contrastive Learning

Overview. Video transformers have been recently shown to be particularly effective at long-range temporal modeling [5]. Our aim is to design a contrastive learning framework that exploits this characteristic. Our proposed Long-Short Temporal Contrastive Learning (LSTCL) framework takes as input a pair of clips sampled from the same video—a long clip and a short clip. The procedure trains the video transformer to match the representation of the short clip to the representation of the long clip. This forces the model to predict the future and the past from a small temporal window, which is beneficial for capturing the general structure of the video. Below we describe specific details related to our LSTCL.

Given a batch B of *unlabeled* training videos, we randomly sample a short clip and a long clip from each of them. While both the long and the short clip include a total of T frames, we use largely different sampling temporal strides τ_S and τ_L with $\tau_S < \tau_L$ in order for the long clip to cover a much longer temporal extent than the short clip. The sets of short and long clips in the batch B are denoted as $X_S = \{x_S^1, x_S^2, \dots x_S^B\}$ and $X_L = \{x_L^1, x_L^2, \dots x_L^B\}$, respectively, where x_S^i and x_L^i represent the short clip and the long clip sampled from the i -th example in the batch. The set of short clips is processed by an encoder f_q to yield a set of “query” examples $Q = \{q^1, q^2, \dots q^B\}$ where $q^i = f_q(x_S^i) \in \mathbb{R}^D$. The set of long clips is processed by a separate encoder f_k to produce “key” examples $K = \{k^1, k^2, \dots k^B\}$. We optimize the encoders to yield similar query-key representations for pairs consisting of a long clip and a short clip taken from the same video, and dissimilar representations for cases where the long clip and the short clip are sampled from different videos. This is achieved by adopting an InfoNCE [41] loss on the sets Q and K :

$$\mathcal{L}_{NCE} = \sum_i -\log \frac{\exp(q^i \top k^i / \rho)}{\exp(q^i \top k^i / \rho) + \sum_{j \neq i} \exp(q^i \top k^j / \rho)}\tag{4}$$

where ρ is a temperature hyperparameter that controls the sharpness of the output distribution. As commonly done in contrastive learning approaches [13, 12, 25, 7], we symmetrize the loss function. In our case this is achieved by adding to the loss term above a dual term obtained by reversing the role of the long and the short clips, i.e., by computing queries from long clips $q^i = f_q(x_L^i)$ and keys from short clips $k^i = f_k(x_S^i)$. The encoder f_q consists of a video transformer backbone, a MLP projection head and an additional prediction MLP head. The purpose of the prediction layer is to transform the representation of the query clip to match the key. The encoder f_k consists of a video transformer backbone and a MLP projection head. Our experiments present results obtained with different contrastive learning optimizations to update the parameters of f_q and f_k . In the case of our default optimization based on MoCo v3, the parameters of f_q are updated by minimizing \mathcal{L}_{NCE} via backpropagation, while the parameters of f_k are updated as a moving average of the parameters of f_q . We refer the reader to our appendix for the details of the optimizations based on the other contrastive learning frameworks considered in our experiments—BYOL and SimSiam.

Clip Sampling Strategy. Since we want our model to be able to extrapolate the context observed in the entire video from the brief extent of the short clip, we propose to sample the long and the short clip *at random* and *independently* from each video. By doing so, the learning cannot leverage any synchrony between the two clips and because the temporal offset will be random for every

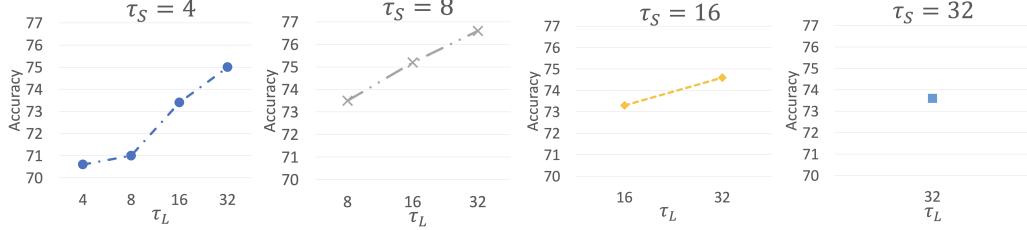


Figure 2: We study how the temporal extents of the short and the long view (controlled by τ_S and τ_L) in LSTCL affect the video-level accuracy on Kinetics-400. We can see that, for each choice of τ_S , accuracy monotonically increases as the long stride τ_L is made larger. The best result is obtained for $\tau_S = 8$ and $\tau_L = 32$, corresponding to a long view that is 4 times longer than the short view.

pair of long-short samples, the optimization will force the short clip representation to encode as much as possible of the context exhibited over the entire video. To demonstrate the value of random independent sampling, in our ablation study we contrast this strategy (named “Random Independent”) to two alternative schemes. The first, named “Random Included,” consists in sampling the short clip at random but so as to fall completely within the temporal extent spanned by the long clip (which is sampled first at random). The second, named “Random Disjoint,” samples the two clips at random but it enforces the constraint that they cannot overlap at all, i.e., they are completely disjoint. We refer the reader to our experiments which validate our hypothesis that random independent sampling is indeed the superior strategy for long-short temporal contrastive learning of video transformers.

Implementation Details. We implement LSTCL under three different and popular contrastive learning frameworks: BYOL [25], MoCoV3 [13], and SimSiam [12]. For training we adopt the video data augmentations described in [19] using clips of size $224 \times 224 \times 8$ sampled from the video. We experiment with two video transformer architectures: TimeSformer with divided Space-Time attention [5] and our adaptation of the Swin-B model [35] to video (Space-Time Swin). We use the AdamW [38] optimizer, which is commonly used for training vision transformer models [5, 8, 13, 3, 2, 15, 51]. In our default set-up, we train LSTCL for 200 epochs on the 240K videos of Kinetics-400 [32] using linear warm-up [23] during the first 40 epochs. We apply a cosine decay schedule [37] after the warm-up and the learning rate is set to $lr \times \text{BatchSize}/256$. We adopt the base learning rate and weight decay from [13]. Our experiments are run on 64 V100 GPUs with a distributed training set-up in Pytorch [42]. The training of 200 epochs takes about 3 days.

5 Experiments

We evaluate our proposed LSTCL on several popular action recognition benchmarks such as Kinetics-400 [32], Kinetics-600 datasets [9], and Something-Something-V2 [24] (SSv2). Our experimental setup is as follows. First, we perform self-supervised LSTCL pretraining on Kinetics-400 with clips of $T = 8$ frames but using *distinct* temporal sampling strides for the short view and the long view, so that the two views effectively span temporal extents of different lengths in seconds. Afterwards, we finetune the LSTCL-pretrained model for 200 epochs in a fully supervised fashion on each of these three datasets. During inference, we sample uniformly 5 clip with center cropping from each video and average the sample-level predictions to perform video-level classification.

5.1 Ablation Studies

Importance of the Temporal Extent. We first ablate the choice of τ_S and τ_L for self-supervised training, while keeping the finetuning temporal stride fixed to the value $\tau = 8$ (i.e., sampling a frame every 8 from the video starting from a random frame). Figure 2 shows how different combinations of τ_S and τ_L affect the final video-level accuracy on Kinetics-400. For ease of interpretation we split the visualization of results over 4 distinct plots, representing 4 different values of τ_S : $\tau_S \in \{4, 8, 16, 32\}$. Each plot shows how the final video-level accuracy varies for different temporal stride values τ_L of the long clip where $\tau_L \geq \tau_S$ and τ_S is kept fixed. There are two important observations we can make from these results. The first is that, for each choice of τ_S , the larger the gap between the two strides (i.e., the larger the value of $\tau_L - \tau_S$), the higher is the accuracy. This can be seen in the first three plots where the accuracy curve monotonically increases as τ_L is made larger starting from the initial

Table 1: We analyze the potential benefits of randomly sampling either τ_S and/or τ_L (for the short and the long clips, respectively). Accuracy is measured for video-level classification on Kinetics-400 after pretraining with our LSTCL system using MoCo v3. The best result is still achieved for fixed values of $\tau_S = 8$ and $\tau_L = 32$.

τ_S	τ_L	Accuracy
4	{8,16,32}	73.9
{4,8,16}	32	74.8
8	{8, 16,32}	75.5
{8,16,32}	32	75.9
{4,8,16,32}	{4,8,16,32}	75.6
{8,16,32}	{8,16,32}	76.0
8	32	76.6

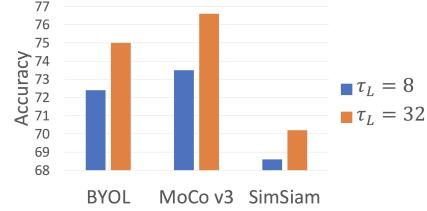


Figure 3: Kinetics-400 accuracy achieved by pretraining with LSTCL using different self-supervised learning frameworks, with a fixed stride for the short clip ($\tau_S = 8$) and two possible stride values for the long clip ($\tau_L \in \{8, 32\}$). All three methods benefit from using views of different lengths (corresponding to $\tau_L = 32$, instead of $\tau_L = \tau_S = 8$).

value of $\tau_L = \tau_S$. This validates the importance of contrasting views of different temporal lengths during self-supervised pretraining. The second finding we can draw from these results is that our model performs best when $\tau_S = 8$ and $\tau_L = 32$. This result makes intuitive sense as a short clip sampled with $\tau_S = 8$ is temporally long enough to allow to predict the context of the long clip; at the same time it is short enough to allow the method to use a long view that is significantly longer (up to 4 times longer than the short view). Conversely, choosing a larger value of τ_S (i.e., 16 or 32) reduces the maximum possible gap $\tau_L - \tau_S$ between the two views, while choosing a smaller value of τ_S (i.e., 4) would cause the contrastive learning between the two views to be overly difficult due to the excessive brevity of the short clip.

In Table 1, we include additional performance points corresponding to settings where τ_S and/or τ_L are sampled randomly for each training video clip. Specifically, the first row in the table shows the performance of our system when $\tau_S = 4$ and τ_L is sampled randomly from $\{8, 16, 32\}$; the second row represents the opposite setting where τ_L is kept fixed ($\tau_L = 32$) and τ_S is randomly sampled from $\{4, 8, 16\}$; the third and fourth row shows the similar setting with the first and second rows but exclude $\tau_S = 4$; the fifth and the sixth rows shows a configuration where both temporal strides are randomly chosen for each training video clip, subject to $\tau_S \leq \tau_L$. As before, we keep the finetuning temporal stride fixed to the value $\tau = 8$. The results in Table 1 clearly show that adding randomness in the choice of the temporal extents for the long and short clips does not produce improved performance. The best performance is still achieved when $\tau_S = 8$ and $\tau_L = 32$ (shown in the last row). Thus, we adopt this setup for all subsequent experiments.

Now we turn to analyze the impact of the *finetuning* temporal stride τ on video-level accuracy. The two plots in Figure 4 show how the accuracy on Kinetics-400 varies as we change the value of τ (on the horizontal axis) for two different choices of τ_S ($\tau_S = 4$ in the left plot, and $\tau_S = 8$ in the right plot). The different curves in each plot correspond to different choices of τ_L . Overall we see that setting the finetuning stride to $\tau = 8$ tends to produce the best results across all possible choices of τ_S and τ_L . Again, this makes sense, since $\tau = 8$ makes the 5 inference clips short enough not to overlap so that they provide complementary information for the video-level classification. At the same time, $\tau = 8$ implies an inference clip long enough to yield good classification on its own.

Experimenting with Different Contrastive Learning Frameworks. Next, we investigate the effects of different contrastive learning frameworks in our LSTCL system. Specifically, we experiment with three recent approaches, namely BYOL, MoCo v3, and SimSiam. From the results in Figure 3 we can notice that a larger temporal stride τ_L for the long view leads to better accuracy for all three of these frameworks. Specifically, setting $\tau_L = 32$ leads to the following performance gains compared to the setting where $\tau_L = \tau_S = 8$: +2.6% for BYOL, +3.1% for MoCo v3, and +1.6% for SimSiam. The lower absolute performance of SimSiam can be explained by the lack of the momentum-encoder, which we observed to be important when training video transformer models with LSTCL. Thus, based on these result, for all subsequent experiments, we adopt MoCo v3 as our base learning framework.

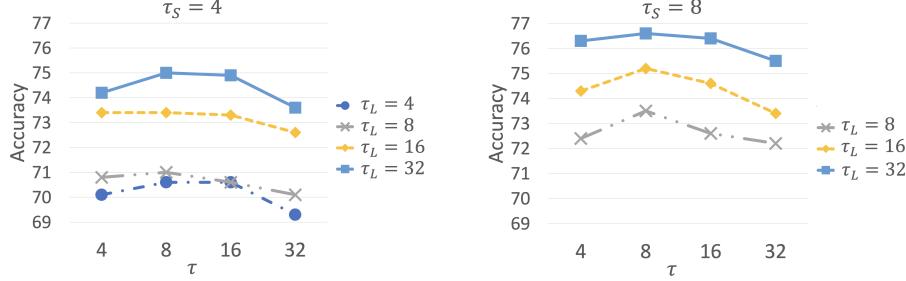


Figure 4: The plots show video-level accuracy on Kinetics-400 for different values of the temporal sampling stride τ used for supervised finetuning and test-time inference.

Clip Sampling Strategy in LSTCL. In Table 2, we study the effect of different clip sampling strategies. These results indicate that random independent sampling where the short and long clips works best in our setting. Intuitively, this makes sense as it forces our model to extrapolate to different video views.

Table 2: Comparison of different clip sampling strategies for LSTCL on Kinetics-400. In these experiments we use $\tau_S = 8$ and $\tau_L = 32$ for LSTCL and $\tau = 8$ for finetuning.

Sampling Method	Accuracy
Random Disjoint	72.6
Random Included	76.2
Random Independent	76.6

Table 3: Comparing self-supervised pretraining using LSTCL to training from scratch and supervised pretraining on ImageNet-1K. The results show video classification accuracy on Kinetics-400 for three video transformer architectures: TimeSformer, Swin, and Space-Time (ST) Swin.

Model	Scratch	ImageNet-1K	LSTCL
TimeSformer	60.4	76.2	76.6
Swin	66.2	73.3	73.5
ST Swin	71.1	76.0	79.8

Video Transformers. In Table 3 we compare the performance of three distinct video transformer architectures: TimeSformer, Swin, and Space-Time (ST) Swin. We train each of these models under three different scenarios on Kinetics-400: 1) from scratch (without using any pretraining), 2) using supervised pretraining on the large-scale ImageNet-1K dataset, and 3) using our self-supervised LSTCL pretraining. We can see that among these three training strategies, our LSTCL pretraining provides the highest accuracy, outperforming the models that use large-scale supervised ImageNet-1K pretraining for all three architectures.

5.2 Comparison to the State-of-the-Art

For our final experiments, we adopt the Space-Time Swin transformer as it performed best in our ablation studies. For this comparison to the state-of-the-art, we also train models using clips of $T = 16$ frames during both pretraining with LSTCL and supervised finetuning. Even in this case we set the temporal stride to $\tau_S = 4$ for short clips and $\tau_L = 16$ for long clips.

Kinetics-400 & Kinetics-600. In Table 4, we report results on Kinetics-400, listing for each method the clip size, the accuracy, the inference cost (in TFLOPs), and the number of parameters. We group methods on the basis of the input clip size, since models trained on longer clips or higher resolution frames tend to yield higher accuracy. The first two groups include models operating on clips of the same size as those used by our system (8×224^2 and 16×224^2). It can be seen that the ST Swin model pretrained with LSTCL achieves the highest accuracy among all previous methods that use the same input clip size and that do not make use of additional data. Furthermore, compared to prior video transformer models that are pretrained with full supervision on large-scaled labeled datasets (the bottom part of the table), our method still achieves competitive results and actually often yields better accuracy. Lastly, note that compared to training our ST Swin model from scratch, LSTCL pretraining leads to a significant 8.7% boost on the Kinetics-400 dataset.

Table 5 shows a comparison with the state-of-the-art on the Kinetics-600 dataset. Even here we see that ST Swin pretrained with LSTCL achieves the best accuracy within the two groups of models

Table 4: Comparison to the state-of-the-art on the Kinetics-400 dataset. Among methods using the same clip sizes as our models and no additional data (first two groups in the table), ST Swin networks pre-trained with LSTCL achieve the highest reported accuracy. ST Swin models trained with LSTCL are competitive with respect to networks that use longer or higher-resolution clips (third group) or that leverage additional labeled data for supervised pre-training (bottom group).

Method	Clip Size	Additional Data (# Samples)	Top-1	Top-5	TFLOPs	Params
SlowFast [18]	8 x 224 ²	-	77.9	93.2	3.0	59.9M
TimeSformer-scratch	8 x 224 ²	-	60.4	76.7	0.59	121.4M
ST Swin from scratch	8 x 224 ²	-	71.1	85.2	0.60	88.0M
ST Swin w/ LSTCL	8 x 224 ²	-	79.8	94.0	0.60	88.0M
CorrNet-101 [55]	16 x 224 ²	-	79.2	-	7.0	-
SlowFast-NL [18]	16 x 224 ²	-	79.8	93.9	7.0	59.9M
MViT-B [16]	16 x 224 ²	-	78.4	93.5	0.36	36.6M
ST Swin w/ LSTCL	16 x 224 ²	-	81.2	95.0	1.80	88.0M
X3D-XL [17]	16 x 312 ²	-	79.1	93.9	1.45	11.0M
ip-CSN-152 [52]	32 x 224 ²	-	77.8	92.8	3.3	32.8M
MViT-B [16]	64 x 224 ²	-	81.2	95.1	4.09	36.6M
TimeSformer [5]	8 x 224 ²	ImageNet-21K (14M)	78.0	93.7	0.59	121.4M
STAM [47]	16 x 224 ²	ImageNet-21K (14M)	79.3	-	0.27	96.0M
TEINet [36]	16 x 224 ²	ImageNet-1K (1.2M)	76.2	92.5	1.8	-
ViViT-H [3]	16 x 224 ²	JFT (300M)	84.8	95.8	118.8	632.0M
ViViT-L [3]	16 x 320 ²	ImageNet-21K (14M)	81.6	94.7	47.9	310.0M
VATT-B [2]	32 x 320 ²	AudioSet + HowTo100M (3.2M)	79.6	94.9	9.08	88.0M
VATT-L [2]	32 x 320 ²	AudioSet + HowTo100M (3.2M)	82.1	95.5	29.8	88.0M
TimeSformer-L [5]	96 x 224 ²	ImageNet-21K (14M)	80.7	94.7	7.14	121.4M

using the same clip sizes as our networks. Furthermore, LSTCL produces a gain of 7.3% compared to learning from scratch.

Table 5: Video-level accuracy on Kinetics-600. The ST Swin model trained with LSTCL achieves results comparable with the state-of-the-art *without* using additional data or labels.

Method	Clip Size	Additional Data (# Samples)	Top-1	Top-5
SlowFast [18]	8 x 224 ²	-	80.4	94.8
ST Swin from scratch	8 x 224 ²	-	74.7	92.2
ST Swin w/ LSTCL	8 x 224 ²	-	82.0	95.5
SlowFast [18]	16 x 224 ²	-	81.8	95.1
MViT-B [16]	16 x 224 ²	-	82.1	95.7
ST Swin w/ LSTCL	16 x 224 ²	-	83.2	96.4
X3D-XL [17]	16 x 312 ²	-	81.9	95.9
MViT-B [16]	32 x 224 ²	-	83.4	96.3
TimeSformer [5]	8 x 224 ²	ImageNet-21K (14M)	79.1	94.4
ViViT-H [3]	16 x 224 ²	JFT (300M)	85.8	96.5
ViViT-L [3]	16 x 320 ²	ImageNet-21K (14M)	83.0	95.7
VATT-B [2]	32 x 320 ²	AudioSet + HowTo100M (3.2M)	80.5	95.5
VATT-L [2]	32 x 320 ²	AudioSet + HowTo100M (3.2M)	83.6	96.6
TimeSformer-L [5]	96 x 224 ²	ImageNet-21K (14M)	82.2	95.6

Table 6: Video-level classification accuracy on Something-Something-V2. Our ST Swin models pre-trained *without* labels using LSTCL yield results on par with the state-of-the-art.

Method	Clip Size	Additional Data (# Samples)	Pretraining	Top-1	Top-5
TimeSformer [5]	8 x 224 ²	ImageNet-21K (14M)	supervised	59.5	-
ResNet50 [19]	8 x 224 ²	K400 (240K)	unsupervised	55.8	-
ST Swin from scratch	8 x 224 ²	-	-	38.4	65.5
ST Swin w/ LSTCL	8 x 224 ²	K400 (240K)	unsupervised	64.8	89.4
TEINet [36]	16 x 224 ²	ImageNet-1K (1.2M)	supervised	64.7	-
ViViT-L [3]	16 x 224 ²	ImageNet-21K (14M)	supervised	65.4	89.8
MViT-B [16]	16 x 224 ²	K400 (240K)	supervised	64.7	89.2
ST Swin w/ LSTCL	16 x 224 ²	K400 (240K)	unsupervised	66.3	90.4
TimeSformer-L [5]	96 x 224 ²	ImageNet-21K (14M)	supervised	62.4	-
MViT-B [16]	64 x 224 ²	K400 (240K)	supervised	67.7	90.9

Something-Something-V2. In Table 6 we report the performance on the Something-Something-V2 dataset. Most previously proposed methods leverage *supervised* large-scale pre-training on external datasets in order to achieve strong performance on this benchmark, since the dataset is relatively small. The results in the table highlight that our ST Swin models pre-trained *without* labels on

Kinetics-400 using LSTCL achieves higher accuracy than methods that leverage pretraining on larger datasets and using manually labeled data. Moreover, our LSTCL pretraining yields a gain 26.4% over the same model trained from scratch. This remarkable improvement is due to the fact that the Something-Something-V2 dataset requires thorough temporal reasoning for good accuracy. Our LSTCL method trains the clip representation to predict the temporal context from the entire video and thus yields strong benefits on this benchmark.

6 Conclusion

This paper introduces Long-Short Temporal Contrastive Learning (LSTCL), an unsupervised pre-training scheme for video transformers. By contrasting representations obtained from a long view and a short view of each video, it forces the model to encode context from the whole video into the features of short clips. We demonstrate our LSTCL under three different contrastive frameworks and two video transformer architectures including a new variant, Space-Time Swin transformer. LSTCL eliminates the need for large-scaled supervised image pretraining in video transformers. Through extensive experiments, we show that unsupervised pretraining with LSTCL leads to similar or better video classification accuracy compared to pretraining with full supervision on ImageNet-21K and it achieves competitive results on three different video classification benchmarks.

A Self-Supervised Learning Frameworks

In the main paper we presented experimental results obtained by implementing our proposed LSTCL procedure under three popular self-supervised contrastive learning frameworks: MoCo v3 [13], BYOL [25], and SimSiam [12]. Since we used the framework of MoCo v3 to present our method (section 4), here we provide a short description of the other two frameworks—BYOL and SimSiam—and of how we use them in our method.

BYOL. BYOL is a self-supervised learning framework consisting of a momentum-encoder (f_{θ_m}), an online encoder (f_θ) and a predictor MLP (g_{θ_p}) which is connected to the online encoder. The momentum-encoder is the moving average of the online encoder, which is controlled by a momentum parameter m . There is no back-propagation into the momentum-encoder parameters θ_m . The momentum update can be written as:

$$\theta_m = m\theta_m + (1 - m)\theta \quad (5)$$

Differently from MoCo, BYOL uses only positive pairs of examples. It minimizes the negative cosine similarity over all positive pairs. In our setting, the set of short clips is processed by the online encoder f_θ to yield a set of “query” examples $Q = \{q^1, q^2, \dots, q^B\}$ where $q^i = f_\theta(x_S^i) \in \mathbb{R}^D$. The set of long clips is processed by the momentum encoder f_{θ_m} to produce “key” examples $K = \{k^1, k^2, \dots, k^B\}$. Then, the LSTCL method with BYOL loss minimizes the following objective:

$$\mathcal{L}_{BYOL} = \sum_i [2 - 2 \cdot \frac{g_{\theta_p}(q^i)^\top k^i}{\|g_{\theta_p}(q^i)\|_2 \cdot \|k^i\|_2}] \quad (6)$$

We symmetrize the objective by adding to the loss term above a dual term obtained by reversing the role of the long and the short clips, i.e., by computing queries from long clips with the online encoder ($q^i = f_\theta(x_L^i)$) and keys from the short clips with the momentum encoder ($k^i = f_{\theta_m}(x_S^i)$).

SimSiam. SimSiam can be viewed as the BYOL method without momentum-encoder and momentum update. Thus, in our setting the queries and the keys are computed from the short view and the long view, respectively, but using the same encoder: $q^i = f_\theta(x_S^i)$, $k^i = f_\theta(x_L^i)$. During training SimSiam applies the stop-gradient operation on the key view. Thus, LSTCL with SimSiam loss minimizes the objective:

$$\mathcal{L}_{SimSiam} = \sum_i [2 - 2 \cdot \frac{g_{\theta_p}(q^i)^\top SG(k^i)}{\|g_{\theta_p}(q^i)\|_2 \cdot \|SG(k^i)\|_2}] \quad (7)$$

where $SG()$ denotes the stop-gradient operation. Even here we symmetrize the objective by adding a dual term with reversed roles for the long and the short clip.

References

- [1] Pulkit Agrawal, Joao Carreira, and Jitendra Malik. Learning to see by moving. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, December 2015.
- [2] Hassan Akbari, Linagzhe Yuan, Rui Qian, Wei-Hong Chuang, Shih-Fu Chang, Yin Cui, and Boqing Gong. Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text. *arXiv preprint arXiv:2104.11178*, 2021.
- [3] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. *arXiv preprint arXiv:2103.15691*, 2021.
- [4] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.
- [5] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? *arXiv preprint arXiv:2102.05095*, 2021.
- [6] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pages 213–229. Springer, 2020.
- [7] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *arXiv preprint arXiv:2006.09882*, 2020.
- [8] 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.
- [9] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017.
- [10] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.
- [11] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020.
- [12] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. *arXiv preprint arXiv:2011.10566*, 2020.
- [13] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised visual transformers. *arXiv preprint arXiv:2104.02057*, 2021.
- [14] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.
- [15] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.
- [16] Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. Multiscale vision transformers. *arXiv preprint arXiv:2104.11227*, 2021.
- [17] Christoph Feichtenhofer. X3d: Expanding architectures for efficient video recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 203–213, 2020.
- [18] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6202–6211, 2019.
- [19] Christoph Feichtenhofer, Haoqi Fan, Bo Xiong, Ross Girshick, and Kaiming He. A large-scale study on unsupervised spatiotemporal representation learning. *arXiv preprint arXiv:2010.09709*, 2021.
- [20] Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep learning*, volume 1. MIT press Cambridge, 2016.
- [21] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *arXiv preprint arXiv:1406.2661*, 2014.
- [22] 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 (ICCV)*, December 2015.
- [23] Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large minibatch sgd: Training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*, 2017.
- [24] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. The " something something" video database for learning and evaluating visual common sense. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5842–5850, 2017.
- [25] 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.
- [26] Tengda Han, Weidi Xie, and Andrew Zisserman. Video representation learning by dense predictive coding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019.

- [27] Tengda Han, Weidi Xie, and Andrew Zisserman. Self-supervised co-training for video representation learning. *arXiv preprint arXiv:2010.09709*, 2020.
- [28] 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*, pages 9729–9738, 2020.
- [29] 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*, pages 770–778, 2016.
- [30] Zilong Huang, Xinggang Wang, Lichao Huang, Chang Huang, Yunchao Wei, and Wenyu Liu. Ccnet: Criss-cross attention for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 603–612, 2019.
- [31] Phillip Isola, Daniel Zoran, Dilip Krishnan, and Edward H. Adelson. Learning visual groups from co-occurrences in space and time. *CoRR*, abs/1511.06811, 2015.
- [32] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.
- [33] Nikos Komodakis and Spyros Gidaris. Unsupervised representation learning by predicting image rotations. In *International Conference on Learning Representations (ICLR)*, 2018.
- [34] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105, 2012.
- [35] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. *arXiv preprint arXiv:2103.14030*, 2021.
- [36] Zhaoyang Liu, Donghao Luo, Yabiao Wang, Limin Wang, Ying Tai, Chengjie Wang, Jilin Li, Feiyue Huang, and Tong Lu. Teinet: Towards an efficient architecture for video recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11669–11676, 2020.
- [37] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016.
- [38] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.
- [39] Ishan Misra, C. Lawrence Zitnick, and Martial Hebert. Shuffle and learn: Unsupervised learning using temporal order verification. In *Proceedings of (ECCV) European Conference on Computer Vision*, pages 527 – 544, October 2016.
- [40] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European conference on computer vision*, pages 69–84. Springer, 2016.
- [41] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.
- [42] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *arXiv preprint arXiv:1912.01703*, 2019.
- [43] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2536–2544, 2016.
- [44] Rui Qian, Tianjian Meng, Boqing Gong, Ming-Hsuan Yang, Huisheng Wang, Serge J. Belongie, and Yin Cui. Spatiotemporal contrastive video representation learning. *CoRR*, abs/2008.03800, 2020.
- [45] Adrià Recasens, Pauline Luc, Jean-Baptiste Alayrac, Luyu Wang, Florian Strub, Corentin Tallec, Mateusz Malinowski, Viorica Patrascu, Florent Altché, Michal Valko, et al. Broaden your views for self-supervised video learning. *arXiv preprint arXiv:2103.16559*, 2021.
- [46] Tal Ridnik, Emanuel Ben-Baruch, Asaf Noy, and Lihi Zelnik-Manor. Imagenet-21k pretraining for the masses. *arXiv preprint arXiv:2104.10972*, 2021.
- [47] Gilad Sharir, Asaf Noy, and Lihi Zelnik-Manor. An image is worth 16x16 words, what is a video worth? *arXiv preprint arXiv:2103.13915*, 2021.
- [48] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. *arXiv preprint arXiv:1406.2199*, 2014.
- [49] Jingkuan Song, Hanwang Zhang, Xiangpeng Li, Lianli Gao, Meng Wang, and Richang Hong. Self-supervised video hashing with hierarchical binary auto-encoder. *IEEE Transactions on Image Processing*, 27(7):3210–3221, 2018.
- [50] Nitish Srivastava, Elman Mansimov, and Ruslan Salakhudinov. Unsupervised learning of video representations using lstms. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 843–852, Lille, France, 07–09 Jul 2015. PMLR.
- [51] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. *arXiv preprint arXiv:2012.12877*, 2020.
- [52] Du Tran, Heng Wang, Lorenzo Torresani, and Matt Feiszli. Video classification with channel-separated convolutional networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5552–5561, 2019.
- [53] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017.

- [54] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*, pages 1096–1103, 2008.
- [55] Heng Wang, Du Tran, Lorenzo Torresani, and Matt Feiszli. Video modeling with correlation networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 352–361, 2020.
- [56] Jue Wang and Anoop Cherian. Learning discriminative video representations using adversarial perturbations. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 685–701, 2018.
- [57] Jue Wang and Anoop Cherian. Gods: Generalized one-class discriminative subspaces for anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8201–8211, 2019.
- [58] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7794–7803, 2018.
- [59] Xiaolong Wang and Abhinav Gupta. Unsupervised learning of visual representations using videos. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 2794–2802, 2015.
- [60] Yuqing Wang, Zhaoliang Xu, Xinlong Wang, Chunhua Shen, Baoshan Cheng, Hao Shen, and Huaxia Xia. End-to-end video instance segmentation with transformers. *arXiv preprint arXiv:2011.14503*, 2020.
- [61] Li Zhang, Dan Xu, Anurag Arnab, and Philip HS Torr. Dynamic graph message passing networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3726–3735, 2020.