Time-Equivariant Contrastive Video Representation Learning

Simon Jenni Hailin Jin
Adobe Research
{jenni,hljin}@adobe.com

Abstract

We introduce a novel self-supervised contrastive learning method to learn representations from unlabelled videos. Existing approaches ignore the specifics of input distortions, e.g., by learning invariance to temporal transformations. Instead, we argue that video representation should preserve video dynamics and reflect temporal manipulations of the input. Therefore, we exploit novel constraints to build representations that are equivariant to temporal transformations and better capture video dynamics. In our method, relative temporal transformations between augmented clips of a video are encoded in a vector and contrasted with other transformation vectors. To support temporal equivariance learning, we additionally propose the self-supervised classification of two clips of a video into 1. overlapping 2. ordered, or 3. unordered. Our experiments show that time-equivariant representations achieve state-of-the-art results in video retrieval and action recognition benchmarks on UCF101, HMDB51, and Diving48.

1. Introduction

A general video representation should accurately capture both scene appearance and dynamics to perform well on various video understanding tasks, *e.g.*, action recognition or video retrieval. While large annotated video datasets [75, 1] advanced the state-of-the-art in video understanding, such annotations are costly. Furthermore, supervised learning from sparse action labels can introduce biases towards appearance and neglect features related to motion when actions are recognizable from static frames [42, 53]. We consider self-supervised learning (SSL) [18] as a possible solution, enabling both the training on abundant unlabelled videos and biasing learning towards video dynamics.

In SSL, learning tasks are created for which supervision does not require human labor. On images, state-of-the-art SSL approaches based on contrastive learning are approaching or exceeding supervised pre-training. Fundamentally, the task in these methods is to discriminate different training examples while simultaneously learning invariance to a set of input transformations, *e.g.*, of the spatial domain via random cropping and horizontal flipping. How should the additional temporal domain in videos be used for SSL?

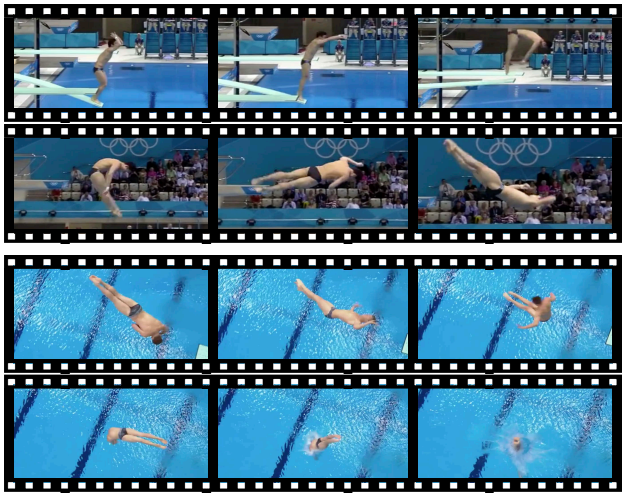

Figure 1. **The importance of motion for video understanding.** We show four clips from two videos of the Diving48 dataset [42]. Different classes in this dataset are defined primarily by different motion patterns. How should these examples be used for contrastive representation learning? Besides recognizing the clips as distinct, our method exploits that the relative temporal shifts between the top two clips and the bottom two clips are identical to learn a temporally equivariant representation. This promotes detailed learning of motion patterns.

A straightforward approach would be to learn invariance to temporal input transformations. Since such invariance could encourage the representation to disregard video dynamics, prior works proposed learning distinctiveness instead [51, 15], *i.e.*, treating different temporal crops from the same video as distinct examples. In contrast, rather than merely learning to recognize two temporal augmentations as different, our main idea is to learn precisely *how* they differ. We, therefore, propose to exploit additional free supervision in the form of the known temporal transformations. Besides learning that two clips show different moments in time, the model will also learn the clips' temporal order and even the temporal shift between them. This

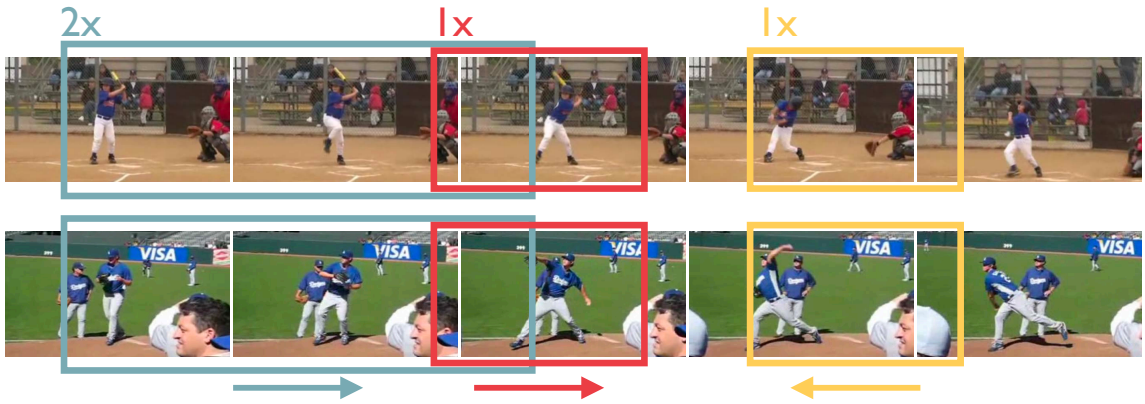

Figure 2. **Illustration of temporal transformations and the proposed learning task.** The colored rectangles illustrate three different temporally augmented clips extracted from two videos. The playback speed is indicated above the box (2× speed resulting in double the temporal extend with the same clip size) and the playback direction with an arrow below the box. The goal of our contrastive equivariance model is to recognize identical relative transformations, *e.g.*, recognize that the temporal transformations between blue and yellow clips in both videos are identical. To support the equivariance learning, we introduce a new auxiliary task of classifying two clips into 1. non-overlapping with correct order (blue & yellow), 2. overlapping (blue & red), 3. non-overlapping with incorrect order (yellow & red).

should improve the learning of dynamics and increase the models' temporal reasoning capabilities, which is crucial in situations where motion is the primary discriminating feature (*e.g.*, see Figure 1). The recognition of input transformations requires that the model represents input changes in a predictable way, *e.g.*, when the learned representation is equivariant to the transformation.

In this paper, we propose a contrastive approach for learning representations that exhibit such equivariance to temporal input transformations. In our method, we encode the relative temporal transformation between two input clips in a feature-vector and contrast it with other, distinct, relative transformation vectors in the mini-batch. Therefore, a positive pair for contrastive equivariance learning results from applying the same relative transformation to two different examples. This framework is very flexible, as it allows us to encode the desired equivariance in a set of input transformations. Aside from standard video augmentations (*e.g.*, random spatial and temporal cropping), we also explore equivariance to reverse playback and playback at higher speeds.

It turns out that learning temporal equivariance is considerably more difficult than learning spatial equivariance (which we also study in experiments). However, temporal equivariance and the resulting preservation of motion features lead to much-improved transfer performance. To increase the network's sensitivity to motion patterns and improve the temporal equivariance learning, we propose as a novel pretext task the three-way classification of two clips into 1. non-overlapping with correct temporal order, 2. overlapping, 3. non-overlapping with an incorrect order. Additionally, we also pose as auxiliary tasks the classification of the playback speed and playback direction as proposed in prior works [63, 5]. All these auxiliary tasks align with the temporal equivariance objective, the optimization of which they support. Figure 2 illustrates the considered transformations and the proposed learning tasks.

Contributions. To summarize, we make the following contributions: 1) We introduce a novel contrastive learning approach to learn representations equivariant to a set of input transformations; 2) We study how equivariance to temporal or spatial transformations affects the quality of the learned features; 3) We introduce the pretext task of clip overlap/order prediction to support the temporal equivariance learning; 4) We show that time-equivariant representations achieve state-of-the-art transfer learning performance in several action recognition and video retrieval benchmarks on UCF101 [54], HMDB51 [38], and Diving48 [42].

2. Related Work

Contrastive Learning Methods. Current state-of-the-art methods in unsupervised image representation learning are primarily based on contrastive learning. The origin of these methods can be traced to instance discrimination tasks [19]. Wu *et al.* [65] proposed a non-parametric formulation of this task based on a noise-contrastive estimation that allowed scaling instance discrimination to a large number of instances. [10] proposed several improvements regarding architecture designs, stronger augmentations, and instance discrimination among large mini-batches. Some methods rely on a queue of past negatives of momentum-encoded examples [29] to lessen the need for large batches. Other methods remove the need for explicit negatives altogether [24, 11]. Several recent works proposed contrastive approaches beyond instance-level discrimination,

i.e., they learn a grouping or clustering of examples using the contrastive framework [9, 62]. Fundamental to these approaches is the learning of transformation invariance. [66] showed that it could be beneficial also to learn distinctiveness to image transformations depending on the downstream tasks. In contrast, we propose a contrastive approach to learn equivariance to a set of input transformations and demonstrated its benefits on video representation learning.

Contrastive Video Representation Learning. Several works have explored the use of contrastive learning on videos. Some propose an extension to videos by adding spatially consistent temporal cropping to the set of augmentations [52]. Others propose to treat temporally augmented clips as distinct instances [15, 51]. Some works explored the combination of contrastive learning with other SSL tasks via multi-task learning [4, 59, 56]. Multi-modal contrastive learning on video has also shown promising results. [27] propose a joint contrastive training between RGB and optical flow. Other works rely on weak supervision in the form of text [46] or exploit the audio accompanying the video [2, 3]. Related methods extract different "views" of the data from different intermediate layers of the network [17, 69]. Our method focuses only on the raw RGB data and goes beyond temporal distinctiveness by learning equivariance. SSL pretext tasks support this equivariance learning.

Temporal Self-Supervision. Self-supervised learning comprises pretext tasks where supervision does not involve human labor. This approach proved effective on images where pretext task include predicting the location of image patches [18, 48], predicting color from grayscale images [72, 73, 39], inpainting image patches [50], clustering [8, 74], or predicting image transformations [22, 32, 33]. Several methods exploit video for self-supervised image representation learning, *e.g.*, via tracking [61, 49, 23]. Besides adapting image SSL tasks to videos [35, 25, 55, 16], several works proposed using temporal self-supervision in videos, *e.g.*, in the arrangement of video frames [47, 6, 21, 40, 40], the temporal arrangement of video clips [68, 36], the arrow of time [63] or the video playback speed [20, 5, 70, 34].Our method leverages such SSL tasks to guide and improve temporal equivariance learning.

Learning Equivariant Representations. Translational equivariance is one of the defining features of CNN architectures [41]. Several works have proposed network designs that endow models with additional equivariances, *e.g.*, capsule networks [30, 31], group equivariant convolutional networks [14], or harmonic networks [64]. Equivariance has also been explored for representation learning on images, *e.g.* by considering a discretized transformation space and predicting multiples of 90° rotations [22], or regressing the parameters of a relative transformation between two input images [71]. Ours is a more general learning approach that is not limited to discrete or parametrized transformations.

3. Model

Our approach combines a temporal equivariance objective (Sec. 3.1 and Fig. 3) with an instance discrimination objective (Sec. 3.3 and Fig.4). A set of auxiliary SSL tasks which are aligned with the equivariance objective, guide the network towards learning temporal features (Sec 3.2).

Let $\mathcal{D} = \{x_1, x_2, \dots, x_N\}$ be a set of unlabelled training videos. Further, let F denote the neural network we want to learn and let $F(x_i) \in \mathbb{R}^D$ be the learned representation of x_i. Our goal is to learn network weights such that $F(x_i)$ is equivariant to a set of temporal transformations $\mathcal{T}$ and invariant to another set $\mathcal{S}$ consisting of spatial augmentations and color jittering. Concretely, for $\sigma \sim \mathcal{S}$ we desire $\forall x \in \mathcal{D} : F_\phi(\sigma(x)) \approx F(x)$ and for $\tau \sim \mathcal{T}$ we want $\forall x \in \mathcal{D} : F_\phi(\tau(x)) \approx \bar{\tau}(F(x))$. The function $\bar{\tau} : \mathbb{R}^D \mapsto \mathbb{R}^D$ represents a transformation in feature space that corresponds to the input transformation τ. We can find an example of equivariance in the feature maps of CNNs: Spatial shifts in the input are reflected in corresponding shifts in the feature map. In contrast, we do not fix the behavior of $\bar{\tau}$. The key property we are interested in is that each $\tau \sim \mathcal{T}$ is recognizable from the pair $\big(F(x), F(\tau(x))\big)$ irrespective of the example $x \in \mathcal{D}$.

3.1. Temporal Equivariance Learning

Different approaches exist to obtain equivariance to a set of transformations $\mathcal{T}$, depending on the nature of $\mathcal{T}$. Before introducing our approach, we will discuss some existing examples from the literature.

When the set is finite, *i.e.*, $\mathcal{T} = \{\tau_1, \dots, \tau_k\}$ and when it is also possible to identify τ_i from a single example $\tau_i(x)$, then standard k-way classification is an option. Examples of this setting can be found in the classification of playback direction [63] and the recognition of playback speeds [20, 5, 70] or other distinct temporal transformations [34, 47].

If the set $\mathcal{T}$ is finite but the τ_i are not identifiable from $\tau_i(x)$ alone, then we have to consider the pair $\big(x, \tau_i(x)\big)$ to classify the relative transformation between x and $\tau_i(x)$. Tasks concerning the temporal ordering of clips [68, 36] somewhat resemble this case (although they typically consider more than two clips).

The more general case is when the set $\mathcal{T}$ is infinite. If all $\tau_i \in \mathcal{T}$ are parametrizable, *i.e.*, τ_i can be described by some parameters $\theta \in \mathbb{R}^d$, then the regression of θ from $\big(x, \tau_i(x)\big)$ can be used to learn equivariance. An example of this can be found on images [71] where τ_i are projective transformations. Even when the parametrization of complex relative transformations is possible, this approach is not practical: Various classification tasks for discrete parameters (*e.g.*, playback direction) have to be balanced with regression tasks for continuous parameters.

We aim to learn equivariance in an even more general case where we can access data containing at minimum two

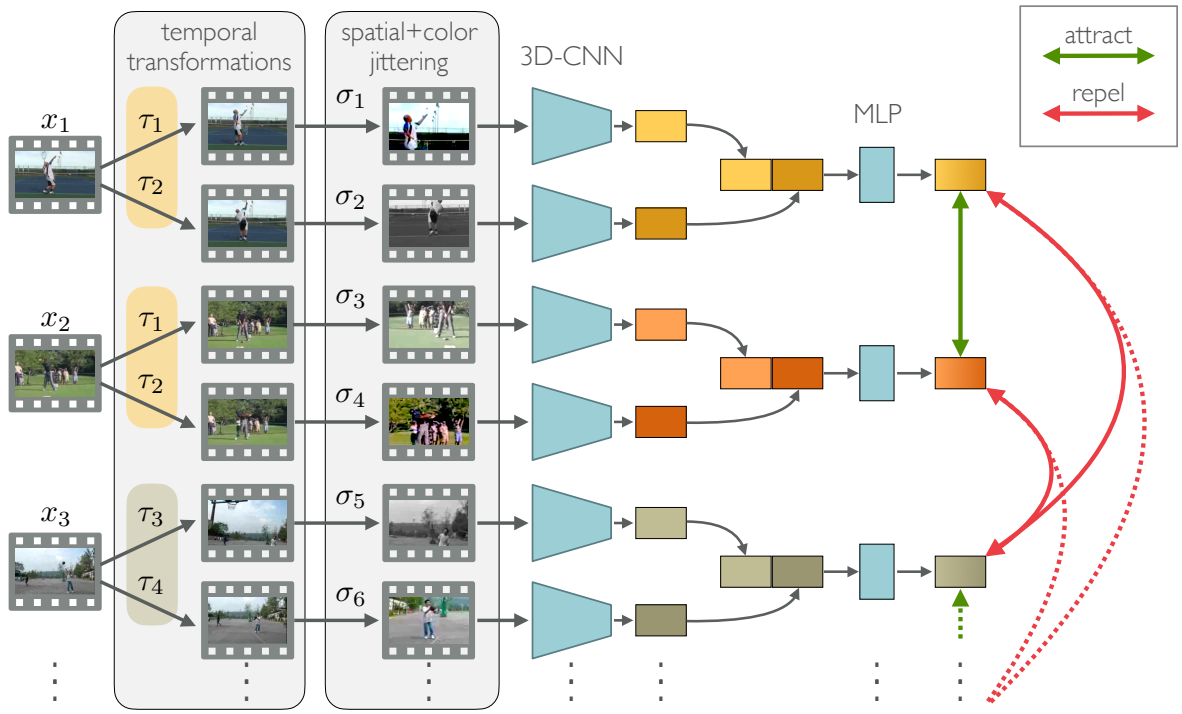

Figure 3. **Overview of the time-equivariant contrastive learning model.** On the left, we show example training videos that are passed through two sets of input transformations. First, we apply temporal transformations (*e.g.*, temporal cropping), which define the set of desired equivariances. We apply the same temporal transformations to two training examples in the batch to enforce the equivariance constraint. Samples then pass through further spatial and color augmentations, which represent desired invariances. All the examples are then encoded in the feature space of some shared 3D-CNN. The feature vectors of the two augmentations for a given video are then concatenated and fed through an MLP that encodes the relative temporal transformation. These vectors are then attracted to the vectors encoding the same transformation and contrasted with other vectors in the batch representing distinct transformations.

pairs with identical relative transformations. That is, the data contains pairs $\big(\tau_p(x_i), \tau_q(x_i)\big)$ and $\big(\tau_p(x_j), \tau_q(x_j)\big)$. This approach does not require a parametrization and also includes the case of non-parametric transformations.

Equivariance via Transformation Discrimination. In what follows, we will omit the application of $\sigma \sim \mathcal{S}$ to each example x_i to keep the notation uncluttered. A different σ is applied implicitly to each occurrence of an example x_i, therefore encouraging invariance to $\mathcal{S}$ in the process. We represent the relative transformation between $\tau_p(x_i)$ and $\tau_q(x_i)$ by $\psi_i^{pq} = \psi\big([F(\tau_p(x_i))^\top, F(\tau_q(x_i))^\top]\big) \in \mathbb{R}^D$, where $\psi(\cdot)$ denotes a multi-layer perceptron, taking as input the concatenation of the feature vectors of $\tau_p(x_i)$ and $\tau_q(x_i)$. We want ψ_i^{pq} to be similar to ψ_j^{pq} for $i \neq j$, and dissimilar to ψ_j^{rs} for $p, q \neq r, s$. This can be achieved using the contrastive learning framework, by otpimizing

$$\mathcal{L}_{\text{equi}} = -\mathbb{E}\left[\log \frac{d(\psi_i^{pq}, \psi_j^{pq})}{d(\psi_i^{pq}, \psi_j^{pq}) + \sum_{rs \neq pq} d(\psi_i^{pq}, \psi_k^{rs})}\right], \tag{1}$$

where

$$d(\mathbf{x}, \mathbf{y}) := \exp\left(\frac{1}{\lambda} \frac{\mathbf{x} \cdot \texttt{stopgrad}(\mathbf{y})}{\|\mathbf{x}\|_2 \|\texttt{stopgrad}(\mathbf{y})\|_2}\right) \tag{2}$$

is a measure of similarity between two feature vectors, with $\lambda = 0.1$ being a temperature parameter and $\texttt{stopgrad}(\cdot)$ indicating that gradients are not back-propagated through the argument (this is similar to [11]). Note that, in practice, the summation in the denominator of Eq.1 is performed over other examples in the same training mini-batch. The training on the transformation discrimination objective is illustrated in Figure 3.

3.2. Auxillary Temporal SSL Objectives

In practice, we find that learning equivariance to temporal transformations $\mathcal{T}$ by optimizing Eq.1 alone is difficult. Initial network parameters do not seem to capture temporal features well, and the optimization stays stuck in a bad region of the parameter space. To alleviate this issue, we exploit SSL tasks that use auxiliary signals that come for free with the transformations τ and are aligned with the equivariance objective. The purpose of these tasks is to steer the network F towards capturing temporal features related to video dynamics. This, in turn, helps the network in optimizing the equivariance objective in Eq.1. We explore three types of auxiliary SSL tasks, 1. classification

of the playback-speed, 2. classification of the playback direction, and 3. classification of the overlap or order of the two compared clips. We describe each of these tasks in detail below.

Speed Classification. In this case, the temporal transformations contain changes in the playback speed. We consider up to four different speed classes corresponding to a $1\times$, $2\times$, $4\times$, and $8\times$ increase of the original playback speed. We train a separate non-linear classifier on the feature representation $F(x)$ to classify these speed types. This task has already been proposed and explored in several prior works [20, 5, 70, 34].

Direction Classification. In this case, the temporal transformations can result in videos being played backward. As in the speed classification, this transformation is predictable from a single transformed example. Thus, to predict this transformation, we train a binary classifier on top of the learned video representation $F(x)$. This self-supervised task was proposed in [63].

Overlap-/Order Classification. Finally, we propose a self-supervised classification task concerning relative time-shifts between two clips, *i.e.*, this task concerns a pair $(\tau_p(x), \tau_q(x))$ with τ_i containing temporal shifts. We discretize the event space into three distinct classes: 1. $\tau_p(x)$ comes entirely before $\tau_q(x)$, 2. $\tau_p(x)$ comes entirely after $\tau_q(x)$, and 3. $\tau_p(x)$ and $\tau_q(x)$ temporally overlap. We solve this task by feeding the concatenated feature vectors, *i.e.*, $[F(\tau_p(x))^\top, F(\tau_q(x))^\top]$ to a non-linear classifier.

Note that accurately solving the temporal equivariance task in Eq.1 subsumes all these auxiliary SSL tasks. Figure 2 illustrates the temporal transformations we consider, the different auxiliary tasks, and how the equivariance objective relates to them.

3.3. Instance Discrimination Objective

Besides learning equivariance to certain input transformations $\mathcal{T}$, we also want the learned video representation $F(x)$ to be sensitive to the actual input scene changing. We achieve this by adding a standard instance contrastive learning objective [10]. Let $\tilde{x}_i^p = \sigma_p \circ \tau_p(x_i)$ denote an augmented training example with some randomly sampled $\sigma_p \sim \mathcal{S}$ and $\tau_p \sim \mathcal{T}$. To obtain instance discriminative features, we encode these augmented examples in a vector $\phi_i^p = \phi(F(\tilde{x}_i^p)) \in \mathbb{R}^D$, where ϕ represents a multi-layer perceptron. We can now formualte the instance discrimination objective as

$$\mathcal{L}_{\text{inst}} = -\mathbb{E}\left[\log \frac{d(\phi_i^p, \phi_i^q)}{d(\phi_i^p, \phi_i^q) + \sum_{j \neq i} d(\phi_i^p, \phi_j^r)}\right], \quad (3)$$

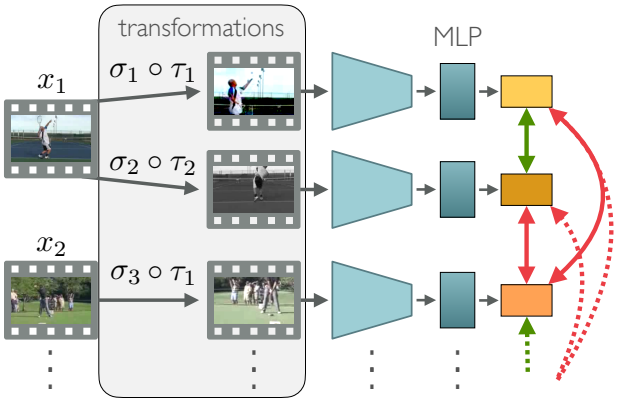

Figure 4. **Instance discrimination task.** We illustrate the task of instance discrimination which we perform on the same batch of examples used for the equivariance objective (compare to Fig.3). Augmentations of the same example are attracted, and other examples are repelled. Note that the MLPs in the two cases are different.

where $d(\cdot, \cdot)$ is again defined as in Eq.2. The training on the instance discrimination objective is illustrated in Figure 4. As shown in the Figure, we perform the two objectives in Equation 1 and 3 on the same examples in the mini-batch. However, the two MLPs ψ and ϕ learn different non-linear projections of $F(x)$ that capture changes in the input transformations and input instance, respectively.

3.4. Implementation Details

Networks were trained using the AdamW optimizer [44] with default parameters and a weight decay of 10^{-4}. During pre-training, the learning rate is first linearly increased to $3 \cdot 10^{-4}$ (10^{-4} for fine-tuning) and then decayed with a cosine annealing schedule [43]. In addition to the temporal input transformations, we also use standard data-augmentations used in contrastive methods, *e.g.*, horizontal flipping, color-jittering, and random cropping. We use a set of different backbone architectures (3D-ResNets[28], R(2+1)D [57], S3D-G [67]) and always consider as learned representation $F(x)$ the output of the global pooling layer. The MLPs $\phi(\cdot)$ and $\psi(\cdot)$ each have two hidden layers and preserve the original feature dimension. The various classification heads for the auxiliary SSL tasks each have a single hidden layer. Input clips contain 16 frames at a resolution of 128×128 if not specified otherwise, and we skip three frames during transfer learning (corresponding to the $4\times$ speed encountered during pre-training). We perform inference by averaging predictions of multiple temporal and spatial crops following prior works [27, 26]. Features for linear probes and nearest-neighbor retrieval are similarly obtained by averaging features from multiple crops and standardizing the result based on training set statistics.

Table 1. **Ablation experiments.** We illustrate the influence of different input equivariances (a)-(d), the different self-supervised learning objectives (e)-(k), and the composition of the auxiliary SSL objective (l)-(o) on the feature performance. We report action recognition accuracy using linear classifiers, full fine-tuning, and nearest-neighbor classification.

		UCF101			HMDB51		
	Ablation	Linear	Finetune	1-NN	Linear	Finetune	1-NN
(a)	no equivariance	54.2	73.4	48.7	25.2	45.4	17.8
(b)	spatial only	62.8	75.5	45.6	38.2	48.9	19.3
(c)	spatial + temporal	70.7	80.8	53.0	44.1	57.6	27.5
(d)	temporal only	**74.1**	**83.7**	**62.1**	**47.5**	**60.8**	**31.5**
(e)	$\mathcal{L}_{inst}$	54.2	73.4	48.7	25.2	45.4	17.8
(f)	$\mathcal{L}_{equi}$	36.2	74.5	21.2	21.2	52.2	10.7
(g)	$\mathcal{L}_{aux}$	67.5	**84.2**	49.8	41.7	59.6	26.6
(h)	$\mathcal{L}_{inst} + \mathcal{L}_{equi}$	63.4	77.0	52.1	37.4	48.8	21.4
(i)	$\mathcal{L}_{inst} + \mathcal{L}_{aux}$	71.8	83.0	59.9	44.9	58.8	30.6
(j)	$\mathcal{L}_{equi} + \mathcal{L}_{aux}$	67.1	83.3	48.0	42.8	**61.0**	27.0
(k)	$\mathcal{L}_{inst} + \mathcal{L}_{equi} + \mathcal{L}_{aux}$	**74.1**	83.7	**62.1**	**47.5**	60.8	**31.5**
(l)	`aux = speed`	69.0	82.3	56.6	42.0	58.3	27.4
(m)	`aux = speed + order`	70.1	81.6	57.9	44.3	58.6	29.0
(n)	`aux = speed + rev`	71.1	83.4	60.0	45.6	59.7	27.5
(o)	`aux = speed + rev + order`	**74.1**	**83.7**	**62.1**	**47.5**	**60.8**	**31.5**

4. Experiments

Datasets. We consider four datasets in our experiments. Kinetics-400 [75] contains around 240K training videos and is used for unsupervised pre-training, *i.e.*, we do not use the action labels. UCF101 [54] and HMDB51 [38] are smaller datasets with human action labels and are used to evaluate the learned representation in transfer to action recognition via fine-tuning and as fixed feature extractors for video retrieval. We also use UCF101 training split 1 for unsupervised pre-training in ablation and retrieval experiments. Finally, we evaluate our method on Diving48 [42], a dataset in which different action classes differ primarily in their long-range motion patterns rather than static frame appearance.

4.1. Ablation Experiments

We perform ablation experiments to investigate the effect of learning equivariance to different types of input transformations and the influence of the various self-supervised learning objectives. We pre-train a 3D-ResNet-18 network for 400 epochs on UCF101 training split 1 with a batch size of 192. The learned representations are then evaluated on UCF101 and HMDB51 via linear SVM classifiers, full finetuning, and nearest neighbors on the action recognition task. Results of the following ablations are summarized in Table 1:

(a)-(d) Equivariance vs. invariance: We compare models trained without equivariance (a), equivariance to only spatial transformations (b), spatial and temporal equivariance (c), and only temporal equivariance (d). These experiments thus change the composition of $\mathcal{T}$ and $\mathcal{S}$, *e.g.*, in (b) $\mathcal{T}$ consists of only spatial augmentations. The model with temporal equivariance and spatial invariance performs best in all metrics. Interestingly, we observe that spatial equivariance leads to better performance in most cases compared to models without any equivariance training. Note that we used consistent temporal augmentations for spatial equivariance learning to remove possible ambiguities.

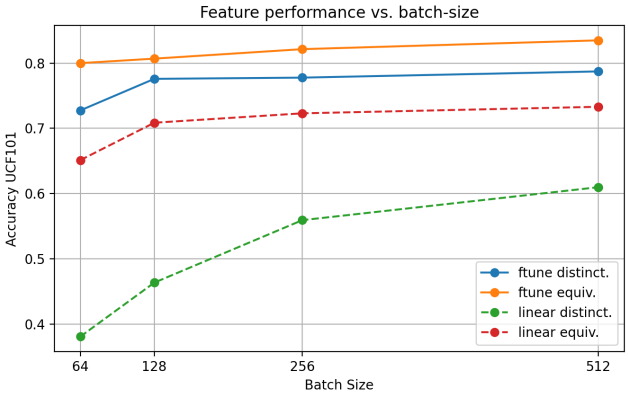

Figure 5. **Equivariance vs. distinctiveness.** We show how contrastive equivariance learning scales with the number of negative examples (*i.e.*, batch size) and compare to a model merely contrasting between temporal augmentations (*i.e.*, learning distinctiveness).

(e)-(k) Training objectives: We investigate how feature performance is influenced by the different training objectives concerning temporal equivariance $\mathcal{L}_{equi}$, instance dis-

Table 2. **Comparison to prior work on self-supervised video representation learning.** We report action recognition accuracy after fine-tuning to UCF101 and HMDB51. We indicate the pre-training dataset, input resolution, number of input frames, network architecture, and pre-training data modality (V=RGB, F=optical-flow, A=audio, T=text).

Method	Dataset	Res.	Frames	Network	Mod.	UCF101	HMDB51
VCOP [68]	UCF101	112	16	R(2+1)D	V	72.4	30.9
PRP [70]	UCF101	112	16	R(2+1)D	V	72.1	35.0
Var. PSP [12]	UCF101	112	16	R(2+1)D	V	74.8	36.8
Temp.-Trans. [34]	UCF101	112	16	R(2+1)D	V	81.6	46.4
Pace Pred. [59]	Kinetics-400	112	16	R(2+1)D	V	77.1	36.6
VideoDIM [17]	Kinetics-400	128	32	R(2+1)D	V	79.7	49.2
TCLR [15]	Kinetics-400	112	16	R(2+1)D	V	84.3	54.2
CBT [55]	Kinetics-600	112	16	S3D	V	79.5	44.6
SpeedNet [5]	Kinetics-400	224	64	S3D-G	V	81.1	48.8
3D ST-puzzle [36]	Kinetics-400	224	16	R3D-18	V	65.8	33.7
MemDPC [26]	Kinetics-400	224	40	R3D-34	V	78.1	41.2
CVRL [52]	Kinetics-400	224	32	R3D-50	V	92.1	65.4
STS [58]	Kinetics-400	224	64	S3D-G	V+F	89.0	62.0
CoCLR [27]	Kinetics-400	128	32	S3D	V+F	87.9	54.6
AVTS [37]	Kinetics-400	224	25	MC3	V+A	85.8	56.9
XDC [3]	Kinetics-400	224	8	R(2+1)D	V+A	84.2	47.1
GDT [51]	Kinetics-400	112	32	R(2+1)D	V+A	89.3	60.0
MIL-NCE [46]	HowTo100M	224	32	S3D	V+T	91.3	61.0
Ours	Kinetics-400	128	16	R3D-18	V	87.1	63.6
Ours	Kinetics-400	112	16	R(2+1)D	V	88.2	62.2
Ours	Kinetics-400	128	32	S3D-G	V	86.9	63.5

crimination $\mathcal{L}_{inst}$, and auxiliary SLL tasks $\mathcal{L}_{aux}$. Overall, we observe that the combination of all the objectives performs best, especially regarding fixed-feature evaluation via linear and nearest neighbor classification. Interestingly, the cases (g) and (j) training only for temporal equivariance perform best in finetuning, suggesting that representations with a strong initial bias towards motion generalize better with few labeled examples. Noteworthy is also the importance of $\mathcal{L}_{aux}$ in combination with $\mathcal{L}_{equi}$. Indeed, we find that networks are unable to optimize $\mathcal{L}_{equi}$ well without the guiding training signals from $\mathcal{L}_{aux}$.

(e)-(k) Auxillary SSL tasks: We demonstrate the effect of the different auxiliary SSL tasks, *i.e.*, speed prediction, direction prediction, and the proposed clip-order/overlap prediction. Each of these tasks corresponds to different temporal transformation types (speed changes, playback direction, and temporal shifts), all of which are beneficial and improve the equivariance objective's optimization.

Finally, in Figure 5 we show how our models' performance scales with the number of negative samples available for contrastive learning. We compare this to a model that is trained to be distinctive, but not equivariant, to temporal transformations. Besides superior performance, we observe that the equivariance model achieves good performance even with small batch sizes. This is of practical value given the large memory footprint of 3D-CNNs.

4.2. Comparison to Prior Works

Action recognition on UCF101 and HMDB51. The most common evaluation of self-supervised video representations is via complete finetuning for action recognition on UCF101 and HMDB51. We compare to prior works using three common backbone architectures. Networks are pre-trained on Kinetics-400 for 200 epochs. The batch size is set to 512 for 3D-ResNet-18, 256 for S3D-G, and 192 for R(2+1)D (adjusting for those networks' different memory footprints). Finetuning is performed for 100 epochs on UCF101 and 200 epochs on HMDB51 using a batch-size of 32. Our results and a comparison to prior works can be found in Table 2. A fair comparison to prior works is difficult since network architectures, pre-training datasets, data modalities, and evaluation protocols can vary significantly. We indicate input resolution, the number of frames, network architecture, and pre-training data modalities, all of which influence performance, in the table. When possible, we report numbers that correspond most closely to our setting, *i.e.*, we report results using only RGB videos at test time when available.

Our method achieves very competitive performance - es-

Table 3. **Video Retrieval Performance on UCF101 and HMDB51.** We report recall at k (R@k) for k-NN based video retrieval. Query videos are taken from test split 1 and retrievals computed on train split 1 of UCF101 and HMDB, respectively. * indicates Kinetics pre-training.

		UCF101				HMDB51			
Method	**Network**	**R@1**	**R@5**	**R@10**	**R@20**	**R@1**	**R@5**	**R@10**	**R@20**
OPN [40]	AlexNet	19.9	28.7	34.0	40.6	-	-	-	-
Büchler *et al*. [7]	AlexNet	25.7	36.2	42.2	49.2	-	-	-	-
STS [58]	C3D	30.1	49.6	58.8	67.6	13.9	33.3	44.7	59.5
Pace Pred. [59]	C3D	31.9	49.7	59.2	68.9	12.5	32.2	45.4	61.0
PRP [70]	R3D-18	22.8	38.5	46.7	55.2	-	-	-	-
VCOP [68]	R3D-18	14.1	30.3	40.4	51.1	7.6	22.9	34.4	48.0
VCP [45]	R3D-18	18.6	33.6	42.5	53.5	7.6	24.4	36.6	53.6
Var. PSP [12]	R3D-18	24.6	41.9	51.3	62.7	10.3	26.6	38.8	51.6
PCL [56]	R3D-18	40.5	59.4	68.9	77.4	16.8	38.4	53.4	68.9
MemDPC [26]	R3D-18	20.2	40.4	52.4	64.7	7.7	25.7	40.6	57.7
Temp.-Trans. [34]*	R3D-18	26.1	48.5	59.1	69.6	-	-	-	-
SpeedNet [5]*	S3D-G	13.0	28.1	37.5	49.5	-	-	-	-
CoCLR [27]	S3D	53.3	69.4	76.6	82.0	23.2	43.2	53.5	65.5
TCLR [15]	R(2+1)D	56.9	72.2	79.0	84.6	24.1	45.8	58.3	75.3
GDT [51]*	R(2+1)D	57.4	73.4	80.8	88.1	25.4	51.4	63.9	75.0
Ours	R3D-18	63.6	79.0	84.8	89.9	**32.2**	**60.3**	**71.6**	**81.5**
Ours	R(2+1)D	**64.3**	**80.9**	**86.4**	**90.6**	29.5	55.8	68.0	78.2

pecially on HMDB51 - with all tested architectures and outperforms other RGB-only approaches, except for [52] which is based on a considerably larger architecture with much higher computational requirements.

Video Retrieval on UCF101 and HMDB51. We also compare to prior works on the video retrieval task in Table 3. Nearest-neighbors are computed using cosine similarity, and we use networks pre-trained on UCF101 in this evaluation, following the majority of prior work. We observe strong performance with our method and achieve state-of-the-art results on both datasets. This is consistent with the drastic improvements resulting from combining temporal equivariance and instance discrimination learning, as observed in the ablations (see Table 1). The results also show the benefit of learning temporal equivariance compared to models that merely learn to distinguish between temporal transformations [15, 51].

Evaluation on Diving48. Finally, we evaluate our method on Diving48 [42], which is a challenging dataset as it requires capturing long-term motion patterns to recognize the classes accurately. Furthermore, unlike in other benchmarks, appearance does not correlate strongly with the label. We pre-train an R(2+1)D on the Diving48 training set and follow the most common evaluation using 16 consecutive input frames (which corresponds to $1\times$ playback in our method). For completeness, we also report results using $2\times$ playback (*i.e*., covering 32 input frames) and results on updated action labels (**V2**). The strong inductive motion bias in our method leads to state-of-the-art results.

Table 4. **Comparison on Diving48.** We report classification accuracy using old (**V1**) and updated (**V2**) diving labels. We indicate playback speed for our results. Prior works all use $1\times$ speed.

	Accuracy	
Method	**V1**	**V2**
Random Init. ($1\times$)	18.8	50.7
Random Init. ($2\times$)	26.6	64.3
RESOUND-C3D [42]	16.4	-
TSN [60]	16.8	-
Debiased R3D-18 [13]	20.5	-
TCRL [15]	22.9	-
Ours ($1\times$)	29.9	71.2
Ours ($2\times$)	34.9	76.2

5. Conclusions

We introduced a novel self-supervised learning approach to learn video representations that are equivariant to temporal input transformations. This method is motivated by the importance of dynamics for video understanding and the desire to reflect variations in the input motion patterns. Our method extends the contrastive learning framework with equivariance constraints on relative temporal transformations between augmented samples. Experiments demonstrate that representations with temporal equivariance achieve state-of-the-art performance on classic vision tasks such as action recognition or video retrieval.

References

[1] Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. Youtube-8m: A large-scale video classification benchmark. *arXiv preprint arXiv:1609.08675*, 2016. 1

[2] Triantafyllos Afouras, Andrew Owens, Joon Son Chung, and Andrew Zisserman. Self-supervised learning of audio-visual objects from video. *arXiv preprint arXiv:2008.04237*, 2020. 3

[3] Humam Alwassel, Dhruv Mahajan, Bruno Korbar, Lorenzo Torresani, Bernard Ghanem, and Du Tran. Self-supervised learning by cross-modal audio-video clustering. *arXiv preprint arXiv:1911.12667*, 2019. 3, 7

[4] Yutong Bai, Haoqi Fan, Ishan Misra, Ganesh Venkatesh, Yongyi Lu, Yuyin Zhou, Qihang Yu, Vikas Chandra, and Alan Yuille. Can temporal information help with contrastive self-supervised learning? *arXiv preprint arXiv:2011.13046*, 2020. 3

[5] Sagie Benaim, Ariel Ephrat, Oran Lang, Inbar Mosseri, William T Freeman, Michael Rubinstein, Michal Irani, and Tali Dekel. Speednet: Learning the speediness in videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9922–9931, 2020. 2, 3, 5, 7, 8

[6] Biagio Brattoli, Uta Büchler, Anna-Sophia Wahl, Martin E Schwab, and Björn Ommer. Lstm self-supervision for detailed behavior analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, 2017. 3

[7] Uta Büchler, Biagio Brattoli, and Björn Ommer. Improving spatiotemporal self-supervision by deep reinforcement learning. *arXiv preprint arXiv:1807.11293*, 2018. 8

[8] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 132–149, 2018. 3

[9] 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. 3

[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. 2, 5

[11] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. *arXiv preprint arXiv:2011.10566*, 2020. 2, 4

[12] Hyeon Cho, Taehoon Kim, Hyung Jin Chang, and Wonjun Hwang. Self-supervised spatio-temporal representation learning using variable playback speed prediction. *arXiv preprint arXiv:2003.02692*, 2020. 7, 8

[13] Jinwoo Choi, Chen Gao, Joseph CE Messou, and Jia-Bin Huang. Why can't i dance in the mall? learning to mitigate scene bias in action recognition. *arXiv preprint arXiv:1912.05534*, 2019. 8

[14] Taco Cohen and Max Welling. Group equivariant convolutional networks. In *International conference on machine learning*, pages 2990–2999. PMLR, 2016. 3

[15] Ishan Dave, Rohit Gupta, Mamshad Nayeem Rizve, and Mubarak Shah. Tclr: Temporal contrastive learning for video representation. *arXiv preprint arXiv:2101.07974*, 2021. 1, 3, 7, 8

[16] 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. 3

[17] R Devon et al. Representation learning with video deep infomax. *arXiv preprint arXiv:2007.13278*, 2020. 3, 7

[18] Carl Doersch, Abhinav Gupta, and Alexei A. Efros. Unsupervised visual representation learning by context prediction. *ICCV*, 2015. 1, 3

[19] Alexey Dosovitskiy, Philipp Fischer, Jost Tobias Springenberg, Martin Riedmiller, and Thomas Brox. Discriminative unsupervised feature learning with exemplar convolutional neural networks. *IEEE transactions on pattern analysis and machine intelligence*, 38(9):1734–1747, 2015. 2

[20] Dave Epstein, Boyuan Chen, and Carl Vondrick. Oops! predicting unintentional action in video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 919–929, 2020. 3, 5

[21] Basura Fernando, Hakan Bilen, Efstratios Gavves, and Stephen Gould. Self-supervised video representation learning with odd-one-out networks. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, pages 5729–5738. IEEE, 2017. 3

[22] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. In *International Conference on Learning Representations*, 2018. 3

[23] 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. 3

[24] 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. 2

[25] Tengda Han, Weidi Xie, and Andrew Zisserman. Video representation learning by dense predictive coding. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 0–0, 2019. 3

[26] Tengda Han, Weidi Xie, and Andrew Zisserman. Memory-augmented dense predictive coding for video representation learning. *arXiv preprint arXiv:2008.01065*, 2020. 5, 7, 8

[27] Tengda Han, Weidi Xie, and Andrew Zisserman. Self-supervised co-training for video representation learning. *arXiv preprint arXiv:2010.09709*, 2020. 3, 5, 7, 8

[28] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6546–6555, 2018. 5

[29] 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. 2

[30] Geoffrey E Hinton, Alex Krizhevsky, and Sida D Wang. Transforming auto-encoders. In *International Conference on Artificial Neural Networks*, pages 44–51. Springer, 2011. 3

[31] Geoffrey E Hinton, Sara Sabour, and Nicholas Frosst. Matrix capsules with em routing. In *International conference on learning representations*, 2018. 3

[32] S. Jenni and P. Favaro. Self-supervised feature learning by learning to spot artifacts. In *CVPR*, 2018. 3

[33] Simon Jenni, Hailin Jin, and Paolo Favaro. Steering self-supervised feature learning beyond local pixel statistics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6408–6417, 2020. 3

[34] Simon Jenni, Givi Meishvili, and Paolo Favaro. Video representation learning by recognizing temporal transformations. *arXiv preprint arXiv:2007.10730*, 2020. 3, 5, 7, 8

[35] Longlong Jing, Xiaodong Yang, Jingen Liu, and Yingli Tian. Self-supervised spatiotemporal feature learning via video rotation prediction. *arXiv preprint arXiv:1811.11387*, 2018. 3

[36] Dahun Kim, Donghyeon Cho, and In So Kweon. Self-supervised video representation learning with space-time cubic puzzles. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8545–8552, 2019. 3, 7

[37] Bruno Korbar, Du Tran, and Lorenzo Torresani. Cooperative learning of audio and video models from self-supervised synchronization. In *Advances in Neural Information Processing Systems*, pages 7763–7774, 2018. 7

[38] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. HMDB: a large video database for human motion recognition. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2011. 2, 6

[39] Gustav Larsson, Michael Maire, and Gregory Shakhnarovich. Colorization as a proxy task for visual understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6874–6883, 2017. 3

[40] Hsin-Ying Lee, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. Unsupervised representation learning by sorting sequences. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 667–676, 2017. 3, 8

[41] Karel Lenc and Andrea Vedaldi. Understanding image representations by measuring their equivariance and equivalence. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 991–999, 2015. 3

[42] Yingwei Li, Yi Li, and Nuno Vasconcelos. Resound: Towards action recognition without representation bias. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 513–528, 2018. 1, 2, 6, 8

[43] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016. 5

[44] Ilya Loshchilov and Frank Hutter. Fixing weight decay regularization in adam. *arXiv preprint arXiv:1711.05101*, 2017. 5

[45] Dezhao Luo, Chang Liu, Yu Zhou, Dongbao Yang, Can Ma, Qixiang Ye, and Weiping Wang. Video cloze procedure for self-supervised spatio-temporal learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11701–11708, 2020. 8

[46] Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. End-to-end learning of visual representations from uncurated instructional videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9879–9889, 2020. 3, 7

[47] Ishan Misra, C Lawrence Zitnick, and Martial Hebert. Shuffle and learn: unsupervised learning using temporal order verification. In *European Conference on Computer Vision*, pages 527–544. Springer, 2016. 3

[48] 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. 3

[49] Andrew Owens, Jiajun Wu, Josh H McDermott, William T Freeman, and Antonio Torralba. Ambient sound provides supervision for visual learning. In *European Conference on Computer Vision*, pages 801–816. Springer, 2016. 3

[50] Deepak Pathak, Philipp Krähenbühl, Jeff Donahue, Trevor Darrell, and Alexei Efros. Context encoders: Feature learning by inpainting. In *CVPR*, 2016. 3

[51] Mandela Patrick, Yuki M Asano, Ruth Fong, João F Henriques, Geoffrey Zweig, and Andrea Vedaldi. Multi-modal self-supervision from generalized data transformations. *arXiv preprint arXiv:2003.04298*, 2020. 1, 3, 7, 8

[52] Rui Qian, Tianjian Meng, Boqing Gong, Ming-Hsuan Yang, Huisheng Wang, Serge Belongie, and Yin Cui. Spatiotemporal contrastive video representation learning. *arXiv preprint arXiv:2008.03800*, 2020. 3, 7, 8

[53] Konrad Schindler and Luc Van Gool. Action snippets: How many frames does human action recognition require? In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2008. 1

[54] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 2, 6

[55] Chen Sun, Fabien Baradel, Kevin Murphy, and Cordelia Schmid. Contrastive bidirectional transformer for temporal representation learning. *arXiv preprint arXiv:1906.05743*, 2019. 3, 7

[56] Li Tao, Xueting Wang, and Toshihiko Yamasaki. Self-supervised video representation using pretext-contrastive learning. *arXiv preprint arXiv:2010.15464*, 2020. 3, 8

[57] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6450–6459, 2018. 5

[58] Jiangliu Wang, Jianbo Jiao, Linchao Bao, Shengfeng He, Wei Liu, and Yun-hui Liu. Self-supervised video representation learning by uncovering spatio-temporal statistics. *arXiv preprint arXiv:2008.13426*, 2020. 7, 8

[59] Jiangliu Wang, Jianbo Jiao, and Yun-Hui Liu. Self-supervised video representation learning by pace prediction. In *European Conference on Computer Vision*, pages 504–521. Springer, 2020. 3, 7, 8

[60] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *European conference on computer vision*, pages 20–36. Springer, 2016. 8

[61] Xiaolong Wang and Abhinav Gupta. Unsupervised learning of visual representations using videos. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2794–2802, 2015. 3

[62] Xudong Wang, Ziwei Liu, and Stella X Yu. Unsupervised feature learning by cross-level discrimination between instances and groups. *arXiv preprint arXiv:2008.03813*, 2020. 3

[63] Donglai Wei, Joseph Lim, Andrew Zisserman, and William T Freeman. Learning and using the arrow of time. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8052–8060, 2018. 2, 3, 5

[64] Daniel E Worrall, Stephan J Garbin, Daniyar Turmukhambetov, and Gabriel J Brostow. Harmonic networks: Deep translation and rotation equivariance. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5028–5037, 2017. 3

[65] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3733–3742, 2018. 2

[66] Tete Xiao, Xiaolong Wang, Alexei A Efros, and Trevor Darrell. What should not be contrastive in contrastive learning. *arXiv preprint arXiv:2008.05659*, 2020. 3

[67] Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 305–321, 2018. 5

[68] Dejing Xu, Jun Xiao, Zhou Zhao, Jian Shao, Di Xie, and Yueting Zhuang. Self-supervised spatiotemporal learning via video clip order prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10334–10343, 2019. 3, 7, 8

[69] Fei Xue, Hongbing Ji, Wenbo Zhang, and Yi Cao. Self-supervised video representation learning by maximizing mutual information. *Signal Processing: Image Communication*, 88:115967, 2020. 3

[70] Yuan Yao, Chang Liu, Dezhao Luo, Yu Zhou, and Qixiang Ye. Video playback rate perception for self-supervised spatio-temporal representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6548–6557, 2020. 3, 5, 7, 8

[71] Liheng Zhang, Guo-Jun Qi, Liqiang Wang, and Jiebo Luo. Aet vs. aed: Unsupervised representation learning by auto-encoding transformations rather than data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2547–2555, 2019. 3

[72] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *European Conference on Computer Vision*, pages 649–666. Springer, 2016. 3

[73] Richard Zhang, Phillip Isola, and Alexei A Efros. Split-brain autoencoders: Unsupervised learning by cross-channel prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1058–1067, 2017. 3

[74] Chengxu Zhuang, Alex Lin Zhai, and Daniel Yamins. Local aggregation for unsupervised learning of visual embeddings. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6002–6012, 2019. 3

[75] Andrew Zisserman, Joao Carreira, Karen Simonyan, Will Kay, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, et al. The kinetics human action video dataset. *ArXiv*, 2017. 1, 6