TransRank: Self-supervised Video Representation Learning
via Ranking-based Transformation Recognition

Haodong Duan[1]   Nanxuan Zhao[1,3,4] ✉   Kai Chen[2,5]   Dahua Lin[1,2,3]
[1]The Chinese University of HongKong   [2]Shanghai AI Laboratory
[3]Centre of Perceptual and Interactive Intelligence   [4]University of Bath   [5]SenseTime Research

Abstract

*Recognizing transformation types applied to a video clip (RecogTrans) is a long-established paradigm for self-supervised video representation learning, which achieves much inferior performance compared to instance discrimination approaches (InstDisc) in recent works. However, based on a thorough comparison of representative Recog-Trans and InstDisc methods, we observe the great potential of RecogTrans on both semantic-related and temporal-related downstream tasks. Based on hard-label classification, existing RecogTrans approaches suffer from noisy supervision signals in pre-training. To mitigate this problem, we developed **TransRank**, a unified framework for recognizing **Trans**formations in a **Rank**ing formulation. TransRank provides accurate supervision signals by recognizing transformations relatively, consistently outperforming the classification-based formulation. Meanwhile, the unified framework can be instantiated with an arbitrary set of temporal or spatial transformations, demonstrating good generality. With a ranking-based formulation and several empirical practices, we achieve competitive performance on video retrieval and action recognition. Under the same setting, TransRank surpasses the previous state-of-the-art method [28] by **6.4%** on UCF101 and **8.3%** on HMDB51 for action recognition (Top1 Acc); improves video retrieval on UCF101 by **20.4%** (R@1). The promising results validate that* RecogTrans *is still a worth exploring paradigm for video self-supervised learning. Codes will be released at* `https://github.com/kennymckormick/TransRank`.

1. Introduction

Effective video representation is of crucial importance for various video understanding tasks, including action recognition [8, 19, 43], temporal localization [5, 67, 68], and video retrieval [62, 70]. To ensure the quality, models pretrained on large-scale video recognition datasets [8, 15, 43]
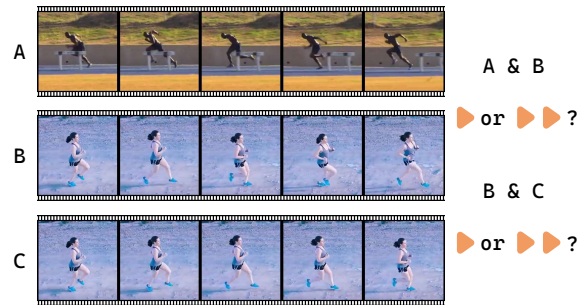
✉Corresponding Author.

Figure 1. **Quiz Time!** Two temporal transformations (Normal Speed $1\times$ and Sped Up $2\times$) are applied to 3 clips. In pair (A, B), (B, C), two clips are played with $1\times$, $2\times$ speed, respectively. Can you find which clip is played with $2\times$ speed in (A, B)? What about the pair (B, C)? Answers are in the footnote[1] on the next page.

have been widely adopted as the initial training point. However, labeling such large video datasets is notoriously costly and time-consuming, limiting the growth rate of labeled video datasets and the evolution of supervised video representation. Considering the infinite supply and the exorbitant annotating cost, learning video representation with self-supervision [16, 41, 49] has drawn increasing attention.

Relying on pretext tasks, video self-supervised learning (video SSL) can obtain good representation without human annotation. The learned representation is then transferred to benefit a series of downstream tasks via finetuning. Most pretext tasks fall into two broad categories: recognizing transformation types [3, 28, 29] (*RecogTrans*) and instance discrimination [9, 46, 57] (*InstDisc*). RecogTrans pretext tasks aim to classify the transformation applied to video clips. The applied transformation can be either spatial (rotation [29], resizing [36]) or temporal (different playback rates [3, 28, 59]). Recently, following the success of contrastive learning in the image domain [10, 20, 24], InstDisc-based pretext tasks gradually become the dominant approach for video self-supervised learning, significantly outperforming RecogTrans-based ones on video downstream tasks, including recognition and retrieval [32, 50].

Rather than directly continuing this trendy research direction, the popularity of InstDisc-based approaches raises

several questions that we care about. First, is there any difference between the representations learned via RecogTrans and InstDisc? Is InstDisc-based representation generally more powerful, or do two representations focus on different aspects, having their own merits? Besides, what is the primary cause of the inferior performance of RecogTrans? Is it because of general limitations of this framework or because of lacking good practices?

To answer these questions, we first conduct a comprehensive study with thorough comparisons on representative methods of RecogTrans and InstDisc. We find that the representation learned via temporal RecogTrans (*RecogTrans-T*) has some unique properties, distinguishing itself from representations learned via InstDisc and spatial RecogTrans (*RecogTrans-S*). The representation learned via RecogTrans-T does not include large amounts of semantic cues, leading to relatively poor performance in downstream settings that directly evaluate the learned representations on semantic-related tasks, like video retrieval and linear evaluation. Meanwhile, RecogTrans-T shows an impressive capability of temporal modeling and does well across different temporal-related tasks. Furthermore, when evaluated on action recognition with a finetuning setting, RecogTrans-T can surpass all alternative SSL methods with a proper finetuning strategy. These findings demonstrate the great potential of RecogTrans-T approaches and motivate us to further explore this direction.

RecogTrans suffers from noisy supervision signals caused by the ignorance of video intrinsic properties. To this end, we develop **TransRank**: a unified framework for recognizing **Trans**formations in a **Rank**ing formulation. The core of our TransRank framework is to consider RecogTrans tasks in a relative manner due to the different intrinsic speeds of videos (Figure 1). For example, 'boxing' and 'running' are much faster than 'tai-chi' and 'walking' in human perception. Even for the same action, the intrinsic speeds can vary a lot when performed by different people (in Figure 1, the runner in clip-A runs much faster than the runner in clip-B). Compared to the classification-based formulation (**TransCls**), TransRank adopts a more accurate and distinct supervision signal, thus outperforming TransCls consistently across different settings. Moreover, the ranking formulation is not detrimental to the universality. TransRank can be instantiated with an arbitrary set of temporal (or spatial) transformations. We further conduct an extensive ablation study on choices of the transformation set, good practices during pre-training and finetuning, and evaluate learned representations on diverse downstream tasks. Competitive performance on downstream tasks, including video retrieval and action recognition, can be obtained with

the ranking-based framework and several good practices.

In summary, we make the following contributions:

1) We revisit several SSL approaches based on RecogTrans and InstDisc, demonstrating the *great potential* of RecogTrans in video self-supervised learning.

2) We develop a new framework called TransRank, which provides more accurate supervision signals than RecogTrans based on hard-label classification, and can be applied to various temporal and spatial pretext tasks.

3) With TransRank and several good practices, we improve the RecogTrans-based video SSL to the next level. Under the same setting, TransRank outperforms a previous state-of-the-art work [28] by **6.4%** on UCF101 and **8.3%** on HMDB51. We achieve decent recognition results (**90.7%**, **64.2%** on UCF101, HMDB51) with a simple R(2+1)D-18 backbone and visual-only inputs. Encouraging results validate that RecogTrans is still worth exploring.

2. Related Work

Self-supervised learning is a long-standing problem attracting a lot of research works in different directions, such as image [6, 14, 18, 44, 45], video [3, 21, 28, 30, 57], and cross-modal [1, 2, 48]. We briefly introduce the most related works on self-supervised video representation learning in this section. The mainstream methods can be categorized into *RecogTrans* and *InstDisc*.

RecogTrans approaches. RecogTrans aims to train a model to recognize the applied transformation to data. Specifically, for video representation learning, the transformations can be either spatial [29, 30] or temporal [3, 28, 34, 61, 65]. Spatial RecogTrans extends existing image pretext tasks (*i.e.*, Jigsaw [44], rotation prediction [18]) to handle video inputs. For temporal RecogTrans, one pretext task is the order prediction: frames [34, 42] or clips [61] from a video are shuffled, and a model is learned to predict the order. Another well-known pretext task is playback style prediction. AoT [59] first proposes to predict the arrow of time in videos, which can be regarded as a binary classification of forward or rewind. SpeedNet [3] and PRP [65] learn a model for playback rate perception. Besides different playing speeds, RTT [28] further introduces more temporal transformations (*e.g.*, periodic, random) to enrich the transformation set for RecogTrans-T. Despite the diversified temporal transformations adopted, these works all conduct RecogTrans in a hard-label classification manner, without considering the intrinsic speeds of different videos. In contrast, our work considers the transformation recognition task in a relative manner, which provides clean and distinct supervision signals in pre-training.

InstDisc approaches. Learning models to perform instance discrimination is another paradigm in self-supervised learning. Following the great success of InstDisc approaches in image representation learning [4, 7, 10–12, 20,

Table 1. **Preliminary results on semantic-related tasks.**

(a) **Video retrieval results on UCF101 and HMDB51.**

Method	UCF101			HMDB51		
	R@1	R@5	R@10	R@1	R@5	R@10
3D-RotNet [29]	40.8	56.8	65.0	17.5	39.1	53.1
SpeedNet [3]	24.4	39.1	48.2	12.4	30.9	43.5
SimSiam [12]	39.0	53.1	60.6	17.1	37.3	48.6
MoCo [24]	**50.1**	**63.8**	**71.9**	**21.8**	**43.9**	**57.0**

(b) **Video recognition results on UCF101 and HMDB51.**

Method	UCF101		HMDB51	
	Linear	Finetune	Linear	Finetune
3D-RotNet [29]	51.6	77.9	23.1	47.7
SpeedNet [3]	31.6	**81.9**	18.4	**51.6**
SimSiam [12]	35.4	81.2	27.9	49.5
MoCo [24]	**63.4**	79.4	**35.3**	46.2

24, 54, 66], InstDisc-based video representation learning has attracted wide attention in recent days. [57] first proposes to combine pace prediction and contrastive learning for video representation learning. Following works improve vanilla contrastive learning in different ways: VTHCL [63] proposes hierarchical contrastive learning to learn visual tempo consistency; VideoMoCo [46] adopts adversarial learning as a temporal data augmentation; CoCLR [22] proposes to learn strong representations via co-training two modalities: RGB and Flow. Meanwhile, another line of works [9, 25, 27, 52] aims to combine contrastive learning and RecogTrans to learn better representations. Recent works [16, 49] show that pure InstDisc approaches can achieve much superior performance on downstream tasks to RecogTrans approaches, with a strong backbone and a large-scale training set. Based on thorough comparisons of representative InstDisc and TransRecog methods, we find that representations learned by two method families focus on different aspects: representations learned with InstDisc are **not** generally more powerful. Besides, with good practices, RecogTrans can also achieve competitive downstream performance on semantic tasks like action recognition.

Other approaches. Besides two main categories, there still exist various approaches that exploit spatiotemporal information in different ways, for example, by future prediction [21, 55], co-occurrence [26], or temporal coherence [33, 56, 58]. Besides, videos are a rich source of multiple modalities. One can also exploit the rich supervision signals from other modalities, including text [1, 35, 40, 51], audio [2, 31, 48], and optical flow [17, 39].

3. Preliminary Comparisons of *InstDisc* and *RecogTrans*

We first conduct pilot experiments to have a preliminary understanding of representations learned by two self-supervised paradigms: *InstDisc* and *RecogTrans*. We select two representative methods from each paradigm: SpeedNet [3] (temporal), 3D-RotNet [29] (spatial) for Recog-Trans; MoCo [24], SimSiam [12] for InstDisc. We use
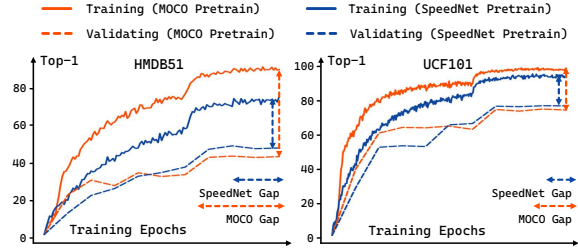

Figure 2. **The train-validation accuracy gaps on downstream tasks.** Initializing with weights pre-trained on RecogTrans-T (i.e., SpeedNet) is less vulnerable to overfitting.

Table 2. **Preliminary results on temporal-related tasks.**

	Motion	Sync-U	Order-U	Sync-H	Order-H	Avg. Rank
Random Guess	20.0	14.7	50.0	14.7	50.0	-
3D-RotNet [29]	69.1	45.8	62.6	43.3	60.3	3.4
SpeedNet [3]	**73.3**	46.5	**82.9**	**46.8**	**80.6**	**1.2**
MoCo [24]	68.0	43.4	69.5	40.6	67.5	3.6
SimSiam [12]	69.6	**48.9**	75.1	45.1	71.4	1.8

R3D-18 for all methods and train them on MiniKinetics [60] for 200 epochs with strong augmentations in Sec 4.4[2]. In the transferring stage, we evaluate the learned representations on various downstream tasks, which can be categorized as semantic-related tasks and temporal-related tasks. We briefly introduce these downstream tasks in this section and leave the detailed setting of each task to the appendix.

Semantic-related Tasks. Semantic-related tasks are the major downstream tasks used in previous literature, focusing on understanding the semantics of a video clip [22, 28, 57]. There exist three main settings in total: 1). *Nearest-Neighbor Evaluation*, which tests if a query class is present in the top-k retrievals; 2). *Linear Evaluation*, which trains a linear classifier based on the learned representation to classify a novel dataset. 3). *Finetuning*, which finetunes the whole network initialized with the learned weights to train a new classifier. In real-world scenarios, finetuning performance is of *greater importance* since better performance on downstream tasks is usually obtained via finetuning.

Temporal-related Tasks. Temporal-related tasks focus on recognizing temporal patterns of video clips, which is of crucial importance in video understanding [37, 64, 69]. We evaluate the temporal modeling capability of learned representations with three temporal-related tasks: 1). *Motion Type Prediction (Motion)* [47], which categorizes the motion in a video into five pre-defined types (*e.g.*, linear and oscillatory); 2). *Synchronization (Sync)* [28], which predicts the temporal pattern of two overlapping video clips (*e.g.*, clip-A is ahead of clip-B with $1/2$ overlapping); 3). *Temporal Order Prediction (Order)* [28], which determines the temporal order of two non-overlapping clips. We demonstrate three tasks in detail in the appendix.

We evaluate four SSL methods on all the semantic-

---

[2]We also perform a basic hyper-parameter search to ensure a suitable setting for each algorithm, *i.e.*, our MoCo implementation surpasses VideoMoco [46] with the same backbone (R3D-18).

related and temporal-related tasks for a comprehensive study. We use the official split-1 of UCF101 [50] and HMDB51 [32] in all downstream tasks[3].

Table 1 shows the evaluation results of different representations under three settings for semantic-related tasks. The performance of video retrieval and linear classification is directly correlated with the quality of the learned representation. MoCo, thanks to its contrastive pre-training, achieves the best results under the first two settings, while SpeedNet becomes the underdog due to the lack of semantic cues in the learned representation. However, once fine-tuned on downstream datasets, SpeedNet quickly adapts to the action recognition task, outperforming all other methods on UCF101 and HMDB51. It's also worth noting that the weight pre-trained on SpeedNet is less vulnerable to overfitting when finetuned on downstream recognition tasks (Figure 2). For temporal-related tasks (Table 2), SpeedNet achieves the best results across 4 of 5 evaluation settings, demonstrating its great temporal modeling capability.

So far, we can issue several questions raised before: 1) The evaluation on diversified downstream tasks demonstrates that representations learned via RecogTrans or InstDisc have their unique benefits: no one is generally more powerful across all evaluation protocols. 2) There is a good chance that the inferior performance reported in previous RecogTrans works is due to lacking good practices. The representation learned by the reimplemented SpeedNet is more powerful compared to the original paper [3], with less training data (MiniKinetics *vs.* K400) and an inferior backbone (R3D-18 *vs.* S3D-G). The improvement is due to the better practices we adopted, like strong augmentations or large finetuning learning rates.

Preliminary experiments have shown good representation learning capability and great potential of RecogTrans-T. However, SpeedNet, as a possible instantiation, still has plenty of room for improvement. First, in SpeedNet, clips are directly categorized as Normal Speed or Sped Up without considering the intrinsic speeds of different videos. Besides, with only two temporal transformations, it is highly likely that SpeedNet did not exploit the full capacity of RecogTrans-T. In this work, we develop a unified framework **TransRank** to further realize the potential. TransRank can achieve outstanding transfer learning performance on downstream tasks, providing substantial evidence for the potential of RecogTrans.

4. TransRank: Ranking-based RecogTrans

TransRank well improves two drawbacks mentioned above. With a ranking-based formulation, TransRank considers the intrinsic properties (like speeds) of videos and the per-video basis difficulty of distinguishing different trans-

formations (*e.g.*, it can be easy, hard, or impossible to distinguish forward and rewind for different videos), thus providing cleaner supervision signals. Meanwhile, it can be instantiated with an arbitrary set of transformations, thus retaining the generality.

4.1. Learning Framework

TransRank recognizes transformations relatively (*e.g.*, a video played with $2\times$ speed is faster than the same video played with $1\times$ speed), and supports various temporal (or spatial) transformations. A deterministic version of TransRank is illustrated in Figure 3, which samples three clips from each video and applies three different transformations to each of them. More generally, TransRank can be instantiated with M different temporal transformations $\{T_1, ..., T_M\}$. In the training stage, it samples N different clips $\{c_1, ..., c_N\}$ from a video, and apply a temporal transformation $T_{t_i}, t_i \in \{1, ..., M\}$ deterministically or stochastically to each c_i. For each clip c_i, TransRank predicts a score vector $S_i = [s_{i1}, ..., s_{iM}]$, s_{ij} denotes the confidence score of the clip c_i being transformed by T_j. On top of the score matrix $S \in \mathbb{R}^{N \times M}$, TransRank adopts a margin ranking loss to learn the transformation recognition task:

$$\mathcal{L}_{\mathcal{T}\text{-}\mathcal{MR}} = \frac{1}{\sum\limits_{i,j} \mathbf{1}_{t_i \neq t_j}(i,j)} \sum\limits_{i,j,t_i \neq t_j} \max(0, s_{jt_i} - s_{it_i} + m), \text{ (1)}$$

where m indicates the margin, which is a hyper-parameter and $\mathbf{1}_{t_i \neq t_j}$ is an indicator that equals to 1 if $t_i \neq t_j$, otherwise 0. The intuition is, for two clips c_i, c_j sampled from a same video with T_{t_i}, T_{t_j} applied, respectively, the score s_{it_i} should be larger than s_{jt_i}, and vice versa. Meanwhile, for a classification formulation, the cross-entropy loss is used:

$$\mathcal{L}_{\mathcal{T}\text{-}\mathcal{CE}} = \frac{1}{N} \sum\limits_i -log(exp(s_{it_i}) / \sum\limits_{j=1}^{M} exp(s_{ij})). \quad \text{(2)}$$

Intuitively, representations learned via TransRank are of better temporal modeling capability: the supervision signal is less noisy with the intrinsic speeds of videos considered. For a better illustration, we conduct a toy experiment and show the results in Figure 4. We train TransRank and TransCls separately with two temporal transformations: Normal Speed ($T_1 : 1\times$) and Sped Up ($T_2 : 2\times$) on MiniKinetics[4]. Then for each testing clip c_i, two scores s_{i1}, s_{i2} will be predicted per model. To verify whether the model is aware of the temporal variance, during testing, we also send clips with two unseen transformations $0.5\times$ and $4\times$ into models for predictions. The underlying idea is that the predicted scores should reflect the scales of transformations adequately. Specifically, we define $s_{i2} - s_{i1}$ as the *speediness score*, and it should be larger with a more drastic transformation. With ranking-based training, TransRank can generalize to unseen transformations $0.5\times$ and $4\times$, predicting

---

[3]Except *Motion*: only HMDB51 is annotated with motion types.

[4]For MiniKinetics videos, 4 transformations $0.5\times$, $1\times$, $2\times$, $4\times$ sample 1 frame per 1, 2, 4, 8 frames to form a clip, respectively.
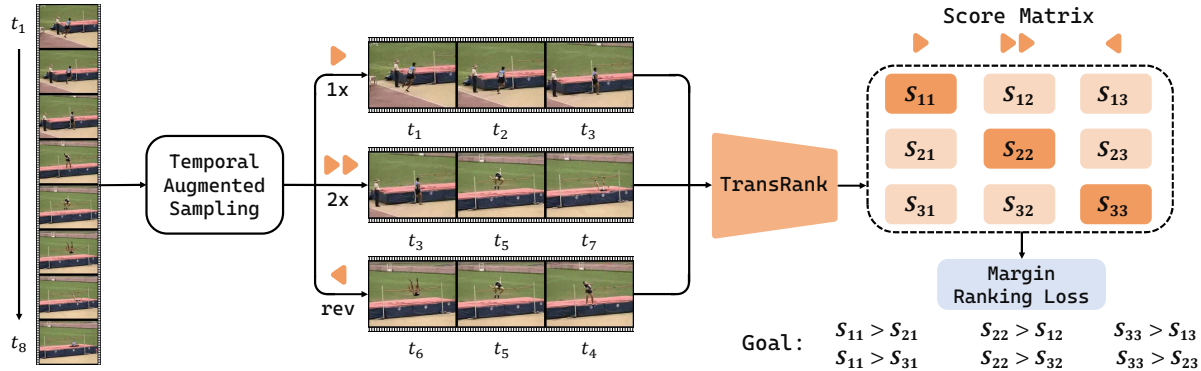
Figure 3. **The TransRank Framework.** In this example, three clips are sampled from a video, with three different temporal transformations: $1\times$, $2\times$, and rev applied, respectively. For each clip, TransRank predicts three scores, corresponding to three transformations. A margin ranking loss is applied to the 3×3 score matrix.

the speediness score accurately. Instead, the new temporal transformations cannot be handled by TransCls.

4.2. On Building the Temporal Transformation Set

TransRank can be instantiated with an arbitrary set of temporal transformations. However, not every transformation will be helpful in representation learning. We briefly introduce the temporal transformations we used and discuss the characteristic that a 'good' transformation needs to own.

A major family of transformations used in previous works [3, 9, 28, 57] is speed-related ones. Different speed-related transformations artificially change the frame rate when sampling clips (l consecutive frames form a clip) from a video. We consider three speed-related transformations $\{1\times, 2\times, 4\times\}$ with a base frame interval 2: for transformation $n\times$, the index sequence of frames in an l-frame clip is $\rho + [0, 2n, 2 \cdot 2n, ..., (l-1) \cdot 2n]$, where ρ is a random offset.

Inspired by the AoT [59] task, we also use another temporal transformation rev in TransRank, which performs frame sampling reversely. rev can also be instantiated with different speeds, namely rev-$n\times$: the index sequence of frames in a rev-$n\times$ clip is $\rho + [(l-1) \cdot 2n, (l-2) \cdot 2n, ..., 2n, 0]$. We abbreviate rev-$1\times$ as rev.

It's worth noting that **not all** transformations are useful in TransRank. One principle is, the model needs to analyze as many frames as possible to recognize a transformation. If a transformation can be recognized by watching a small proportion of frames, the model may take the shortcut and fail to learn a good representation for the whole clip. We provide two examples of such 'bad' transformations, namely shuffle (frame indices are randomly permuted) and palindrome ($\rho + [0, 2 \cdot 2n, ..., (l-2) \cdot 2n, (l-1) \cdot 2n, ..., 2n]$), for which critical cues for recognizing the transformation can be obtained by watching a small part of the clip. In experiments, we find recognizing such transformations is much easier, while adding them to the transformation set leads to performance degradation on downstream tasks.
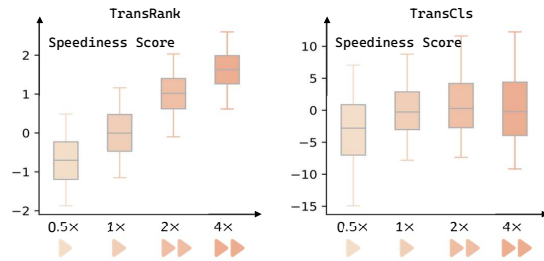

Figure 4. **TransRank vs. TransCls: The Speediness Distribution on MiniKinetics validation dataset.** A stronger transformation (e.g., $4\times$) should have a larger speediness score. TransRank can successfully reflect this, while TransCls failed to achieve it. [5]

4.3. The Generality of TransRank

TransRank is a general framework, not restricted to temporal transformations. We take two spatial transformations as examples to show: 1). How to apply TransRank to recognize spatial transformations? 2). Which kind of spatial transformation can benefit from the ranking formulation?

Estimate Aspect Ratio. RandomResizedCrop is often adopted for data augmentation, which crops a region of a random size and aspect ratio from the original clip and re-sizes it to a target size. Inspired by V3S [36], we propose to exploit the original aspect ratio of a cropped clip as a supervision signal for video SSL. However, videos may have different intrinsic aspect ratios, just like they have different intrinsic speeds. Formulating the aspect ratio estimation as a classification or regression problem is unreasonable, while TransRank can be a suitable framework. For N clips $\{c_1, ..., c_N\}$ sampled from a video with original aspect ratios $\{r_1, ..., r_N\}$, we estimate a ratio score r' for each clip (lower $r' \to$ smaller r). A margin ranking loss is used to ensure the ranking of r' to be the same as r:

$$\mathcal{L}_{\mathcal{S}\text{-}\mathcal{MR}} = \frac{2}{N \cdot (N-1)} \sum_{i,j,r_i < r_j} \left[\max(0, r_i' - r_j' + m) \right]. \quad (3)$$

---

is 0, and the average of $2\times$ clips is 1. Whiskers denote 5%, 95% quantiles; boxes denote 25%, 75% quantiles; middle lines in boxes denote the median. Please find class-specific visualization in the appendix.

[5]Speediness scores are normalized so that the average score of $1\times$ clips

Estimate Rotation. Estimating rotations for images or videos is a well-known pretext task in SSL [18, 29]. Each image or video clip is randomly rotated with four different degrees (0°, 90°, 180°, 270°), and a network is trained to recognize the applied rotation. This task can also be jointly trained with temporal TransRank. However, the ranking formulation does not lead to additional benefits over TransCls, since the rotation category is a more distinct supervision signal compared to the playback rate.

Final Loss. By jointly optimizing the temporal TransRank objective $\mathcal{L}_{\mathcal{T}\text{-}\mathcal{MR}}$ and the spatial TransRank objectives $\mathcal{L}_{\mathcal{S}\text{-}\mathcal{MR}_i}$, the final loss is:

$$\mathcal{L} = \mathcal{L}_{\mathcal{T}\text{-}\mathcal{MR}} + \sum_i \lambda_i \mathcal{L}_{\mathcal{S}\text{-}\mathcal{MR}_i}, \qquad (4)$$

λ_i is the loss weight for the i_{th} spatial pretext task.

4.4. Data Augmentations

To consider an extreme case: if two clips ($a : 1\times, b : 2\times$) are sampled from a video with no data augmentation, they can be easily distinguished with no knowledge required. One may just find the 3_{rd} frame of a and the 2_{nd} frame of b are the same (and the 1_{st} frame of a and b are the same). When sampling multiple clips from a video, we perform strong spatiotemporal augmentations to prevent TransRank from taking shortcuts and secure a useful representation.

Temporal Augmentations. Two temporal augmentations are applied when sampling clips from a video: 1) When selecting indices for a clip, to introduce variability in the time domain, we add random jitter to the frame interval. We multiply it by a random factor sampled from a uniform distribution $[0.8, 1.2]$ (*e.g.*, $1\times: 0.8-1.2\times, 2\times: 1.6-2.4\times$, still distinguishable). 2) Multiple clips are uniformly sampled from the entire video, with different clip offsets.

Spatial Augmentations. We apply *clip-wise* strong spatial augmentations to multiple clips sampled from the same video. Take RandomResizedCrop as an example: clips in a video are cropped with different random regions. Other spatial augmentations we used include RandomGrayScale and RandomColorJitter. For joint training with rotation estimation, we further apply RandomRotate, which randomly rotates a video clip with four degrees.

5. Experiments

5.1. Implementation Details.

Network Architectures and Inputs. We mainly consider two backbones: R3D-18 [23] and R(2+1)D-18 [53], used frequently in previous video SSL literature. N clips from a video are fed into the backbones, followed by global average pooling, to extract N 512-d feature vectors. Based on extracted features, a temporal head is adopted to predict M-d confidence scores for M temporal transformations. For optional spatial pretext tasks, one spatial head

Table 3. **The effect of different temporal transformations on downstream action recognition tasks.**

	TransRank (Ours)		TransCls	
T-Trans Set	UCF101	HMDB51	UCF101	HMDB51
$\{1\times, 2\times\}$	81.3	51.4	80.5	50.8
$\{1\times, 2\times, \text{rev}\}$	84.1	54.5	83.3	52.7
$\{1\times, 2\times, \text{rev, palindrome}\}$	82.9	54.7	82.4	54.0
$\{1\times, 2\times, \text{rev, shuffle}\}$	83.8	53.6	83.0	53.6
$\{1\times, 2\times, \text{rev}, 4\times\}$	85.1	56.3	84.2	54.3
$\{1\times, 2\times, \text{rev, rev-}2\times\}$	85.3	54.9	84.9	53.6

is used to predict scores for each task. The temporal and spatial heads can be instantiated with fully connected layers or multi-layer perceptrons. Our implementation is based on the action recognition codebase MMAction2 [13]. More details on pre-training and finetuning are in the appendix.

Datasets. We use four datasets in experiments: 1). *MiniKinetics* [60] is a subset of the widely used action recognition dataset Kinetics400. It consists of 200 action classes and 76K training videos; each lasts around 10 seconds. We use the training split of MiniKinetics as the pre-training dataset and *K200* to denote MiniKinetics in experiments. 2). *SthV1* [19] is a large collection of labeled video clips showing humans performing pre-defined basic actions. It consists of 86K training videos and 11K validation videos covering 174 action categories. Each clip lasts around 2 to 6 seconds. We treat SthV1 as another pre-training dataset and also adopt it as a downstream recognition task in some experiments using K200 for pre-training. 3). *UCF101* [50] & *HMDB51* [32] are two relatively small action recognition datasets, consisting of 13K videos of 101 classes and 7K videos of 51 classes, respectively. Both datasets are divided into three official train/test splits. We use the split-1 for ablation study and report the average performance over three splits for comparison with the state-of-the-art.

5.2. Ablation Study

By default, we initialize TransRank-T with temporal-transformation (T-Trans) set $\{1\times, 2\times, \text{rev}\}$ and R3D-18, adopt K200 as the pre-training dataset. All heads are instantiated with fully-connected layers unless specified.

The effect of different T-Trans sets. To evaluate the advantage of the ranking formulation, we first compare the downstream performance of TransRank and TransCls instantiated with the same T-Trans set. For both frameworks, we start with the T-Trans set $\{1\times, 2\times\}$ (same as Speed-Net), and then add another major transformation rev and two auxiliary transformations $4\times, \text{rev-}2\times$. We show results in Table 3. The downstream recognition accuracies increase monotonically under both frameworks as the T-Trans set gets larger, validating the importance of having diverse transformations for RecogTrans-T to learn good representations. Besides, TransRank outperforms TransCls consistently across different temporal transformation sets, demonstrating the effectiveness of the TransRank framework.

Table 4. **The effect of spatial and temporal augmentations.**

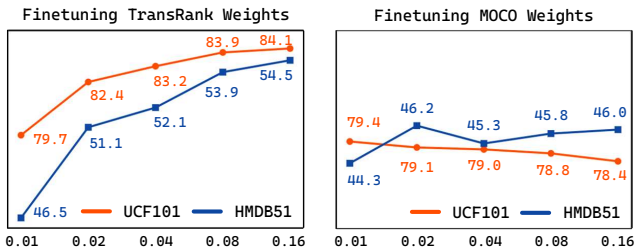| Experimental Setup		Downstream Tasks	
Spatial Augs.	Temporal Augs.	HMDB51	UCF101
✗	✗	50.4	80.2
✗	✓	50.8	80.8
✓	✗	54.2	83.8
✓	✓	**54.5**	**84.1**

Figure 5. **The effect of LRs on finetuning SSL models.** The finetuning LR ranges from 0.01 to 0.16 with a batch size of 128. Finetuning with a large LR is essential for TransRank to work well (x-axis: LR, y-axis: Top1 Acc).

We further add two transformations: palindrome and shuffle to the T-Trans set $\{1\times, 2\times, \text{rev}\}$, to validate the design principle of transformation (Sec 4.2). We observe performance drops in Table 3 and also find that TransRank can easily recognize these two transformations (with Top-1 accuracy > 99.5%) during pre-training. The findings are consistent with our assumption in Sec 4.2 that the model may take the shortcut and fail to learn a good representation as it can achieve the task by watching a small portion of frames.

Good practices for pre-training & finetuning. We adopt several good practices for pre-training and finetuning of TransRank, including strong augmentations (spatial & temporal) and large finetuning learning rates. We first ablate the effect of strong augmentations. Table 4 demonstrates that both spatial and temporal augmentations contribute to learning good representations for transferring, while the improvement from strong spatial augmentations is more significant. Besides, finetuning with a large learning rate (LR) is critical for TransRank to succeed on downstream recognition tasks. We finetune TransRank and MoCo with a wide range of the initial LR. Figure 5 shows that for TransRank, the downstream recognition performance may drastically deteriorate with a small initial LR. The Top-1 accuracy drops 4.4% and 8% on UCF101 and HMDB51, respectively, when using 0.01 as the initial LR. For the InstDisc approach MoCo, the downstream performance is much less vulnerable to changes of LR.

TransRank + spatial transformations. TransRank is a general framework and can be further extended to spatial transformations. We formulate aspect ratio estimation and rotation estimation to fit the TransRank framework and jointly train the spatial tasks with TransRank-T. The loss weight λ for each spatial task is 0.5. We compare TransRank and TransCls for each spatial task in Table 5, eval-

Table 5. **Extending TransRank on spatial transformations.** We jointly train TransRank-T with two spatial tasks. For both tasks, our ranking formulation achieves better or comparable performance than vanilla classification formulation.

Method	UCF101	HMDB51
TransRank-T	84.1	54.5
+ Aspect (Cls)	84.1	53.5
+ Aspect (Ours)	**84.8**	**56.6**
+ Rotation (Cls)	85.2	**57.1**
+ Rotation (Ours)	**85.3**	56.8

Table 6. **Improve the feature quality with MLP heads.** The feature quality can be significantly improved by replacing the fc head with a 2-layer MLP (TransRank-ST includes two spatial tasks).

Method	w. MLP	Order	Sync	R@1	R@5	R@10
TransRank-T	✗	89.0	52.1	22.0	44.1	54.4
TransRank-T	✓	**92.0**	**57.4**	**32.5**	**54.0**	**65.0**
TransRank-ST	✗	82.3	53.7	39.2	59.1	70.0
TransRank-ST	✓	**83.3**	**58.4**	**51.1**	**70.4**	**78.3**

uating their finetuning performance on action recognition. For aspect ratio estimation, TransRank outperforms TransCls by a large margin. Since each video may have its own aspect ratio, it is unreasonable to classify with a hard label. Thus, there is no performance gain after adding this task on TransCls. Since the supervision signal is distinct and less ambiguous for rotation estimation, both formulations lead to good finetuning performance.

Improve feature quality with MLP heads. With finetuning, TransRank achieves competitive results on downstream tasks. However, the experiments show that RecogTrans approaches are of inferior performance under settings directly evaluating the learned features, *i.e.*, video retrieval. We ascribe the deficiency to the head design in previous works [3, 28]: using a single fully-connected (FC) layer to predict the confidence for transformations. We replace the FC head with a 2-layer MLP (128 hidden channels). Table 6 shows that the simple modification drastically improves the feature quality. On UCF101, the Top-1 recall increases by 10% for both TransRank-T and TransRank-ST. Besides, the improvements of two temporal-related tasks *Sync* and *Order* are also significant. We also evaluate the pre-trained models under the finetuning setting, but have not observed any significant difference (FC head *vs.* MLP head).

5.3. Comparison to the state-of-the-art

We compare to prior works on self-supervised video representation learning in Table 7. We report Top-1 accuracies (average of 3 folds) on UCF101 and HMDB51 for both RGB and RGBDiff (the differences of RGB frames, cheap to obtain) modalities. Since the backbones and pre-training settings used in prior works vary, we mainly compare results using commonly used network architectures (*i.e.*, R3D-18, R(2+1)D-18). We adopt TransRank-ST with the T-Trans set $\{1\times, 2\times, \text{rev}, \text{rev-}2\times\}$ and two spatial tasks (the loss

Table 7. **Comparisons with other video self-supervised learning methods on UCF101 and HMDB51.** Methods with * include InstDisc in pre-training. Methods with # use heavy backbones. Abbreviations are used for each modality: R → RGB, RD → RGBDiff, F → Flow, A → Audio.

Method	Backbone	Input Size	Modality	Pre-train Data	UCF101	HMDB51
SpeedNet# [3]	S3D-G	64×224^2	R	K400 (28d)	81.1	48.8
CoCLR*# [22]	S3D	32×128^2	R	K400 (28d)	87.9	54.6
RSPNet*# [9]	S3D-G	16×224^2	R	K400 (28d)	89.9	59.6
ASCNet*# [25]	S3D-G	64×224^2	R	K400 (28d)	90.8	60.5
CoCLR*# [22]	S3D	32×128^2	R + F	K400 (28d)	90.6	62.9
CVRL*# [49]	R3D-50	32×224^2	R	K400 (28d)	92.2	66.7
ST-Puzzle [30]	R3D-18	16×80^2	R	K400 (28d)	65.8	33.7
3D-RotNet [29]	R3D-18	16×112^2	R	K600 (45d)	66.0	37.1
DPC [21]	R3D-18	25×128^2	R	K400 (28d)	68.2	34.5
RSPNet* [9]	R3D-18	16×112^2	R	K400 (28d)	74.3	41.8
3D-RotNet [29]	R3D-18	16×112^2	R + RD	K600 (45d)	76.7	47.0
RTT [28]	R3D-18	16×112^2	R	K600 (45d)	79.3	49.8
TransRank-ST (Ours)	R3D-18	16×112^2	R	K200 (9d)	85.7	58.1
TransRank-ST (Ours)	R3D-18	16×112^2	RD	K200 (9d)	88.9	61.4
TransRank-ST (Ours)	R3D-18	16×112^2	R + RD	UCF101 (1d)	88.5	63.0
TransRank-ST (Ours)	R3D-18	16×112^2	R + RD	SthV1 (4d)	89.0	61.6
TransRank-ST (Ours)	R3D-18	16×112^2	R + RD	K200 (9d)	89.6	63.5
VCOP [61]	R(2+1)D-18	16×112^2	R	UCF101 (1d)	72.4	30.9
PacePred* [57]	R(2+1)D-18	16×112^2	R	K400 (28d)	77.1	36.6
VideoMoCo* [46]	R(2+1)D-18	32×112^2	R	K400 (28d)	78.7	49.2
V3S [36]	R(2+1)D-18	16×112^2	R	K400 (28d)	79.2	40.4
RSPNet* [9]	R(2+1)D-18	16×112^2	R	K400 (28d)	81.1	44.6
RTT [28]	R(2+1)D-18	16×112^2	R	UCF101 (1d)	81.6	46.4
XDC* [2]	R(2+1)D-18	32×224^2	R + A	K400 (28d)	86.8	52.6
TransRank-ST (Ours)	R(2+1)D-18	16×112^2	R	K200 (9d)	87.8	60.1
TransRank-ST (Ours)	R(2+1)D-18	16×112^2	D	K200 (9d)	89.7	63.0
TransRank-ST (Ours)	R(2+1)D-18	16×112^2	R + RD	K200 (9d)	90.7	64.2

Table 8. **Video retrieval performance on the split 1 of UCF101 and HMDB51.** Three regions correspond to other RecogTrans approaches, TransRank-ST (**Ours**), and InstDisc approaches.

Method	Backbone	Pre-train Data	UCF101			HMDB51		
			R@1	R@5	R@10	R@1	R@5	R@10
SpeedNet [3]	S3D-G	K400	13.0	28.1	37.5	-	-	-
VCP [38]	R3D-18	UCF101	18.6	33.6	42.5	7.6	24.4	33.6
PRP [65]	R3D-18	UCF101	22.8	38.5	46.7	8.2	25.8	38.5
RTT [28]	R3D-18	K600	26.1	48.5	59.1	-	-	-
V3S [36]	R3D-18	UCF101	28.3	43.7	51.3	10.8	30.6	42.3
PacePred* [57]	R3D-18	UCF101	23.8	38.1	46.4	9.6	26.9	41.1
RSPNet* [9]	R3D-18	K400	41.1	59.4	68.4	-	-	-
CoCLR* [22]	S3D	UCF101	53.3	69.4	76.6	23.2	43.2	53.5
CoCLR* [22]	S3D	K400	46.3	62.8	69.5	20.6	43.0	54.0
TransRank-ST	R3D-18	UCF101	46.5	63.7	72.8	19.4	45.4	59.1
TransRank-ST	R3D-18	K200	**54.0**	**71.8**	**79.6**	**25.5**	**52.3**	**65.8**

Table 9. **Comparisons with supervised pre-training.**

Pre-training	Modality	UCF101	HMDB51	SthV1
Supervised	RGB	81.4	54.8	42.4
Supervised	Diff	88.7	58.0	42.0
Supervised	2Stream	88.9	59.9	44.9
TransRank-ST (**Ours**)	RGB	**85.7**	**58.1**	**43.1**
TransRank-ST (**Ours**)	Diff	**88.9**	**61.4**	**43.2**
TransRank-ST (**Ours**)	2Stream	**89.6**	**63.5**	**45.8**

Comparisons to supervised pre-training. In Table 9, we compare TransRank-ST with supervised pre-training on 3 datasets: UCF101, HMDB51, SthV1, each with 3 settings: RGB, Diff, 2Stream. TransRank pre-train outperforms supervised pre-train across all downstream settings.

weight λ is 0.5 for each task). We train TransRank-ST on all datasets for 200 epochs.

Action Recognition. TransRank-ST achieves competitive performance with all backbones over both datasets. With a standard R3D-18 backbone, TransRank-ST outperforms the state-of-the-art RecogTrans approach RTT [28] by 6.4% and 8.3% on UCF101 and HMDB51, with much fewer training data (K200 *vs*. K600). With RGB modality and the cheap RGBDiff modality combined, TransRank-ST with R(2+1)D-18 backbone achieves 90.7% and 64.2% Top-1 accuracies on UCF101 and HMDB51. TransRank outperforms the state-of-the-art RecogTrans methods by a large margin, and is comparable with state-of-the-art InstDisc approaches (*i.e*., CoCLR) which adopts a more advanced backbone and a much longer input.

Video Retrieval. We perform another quantitative evaluation of the learned representation via nearest-neighbor (NN) retrieval. We adopt TransRank-ST (R3D-18) trained on RGB inputs as the feature extractor. For each video, we extract and average features of 10 clips to obtain one single 512-d feature vector. Cosine similarity is used as the metric to determine the nearest neighbors. For each clip in the testing split, we query clips in the training split and get N nearest neighbors ($N = 1, 5, 10$). A query is correctly classified if N-nearest neighbors contain at least one video of the same class as the query. We report the mean accuracy for different N (R@N) and compare it to prior works in Table 8. TransRank-ST significantly outperforms all RecogTrans approaches, and achieves comparable performance to state-of-the-art InstDisc approaches like CoCLR [22].

6. Discussion

To conclude, we demonstrate the *great potential* of *RecogTrans*-based video self-supervised learning by introducing a unified framework named TransRank. We have shown its effectiveness through extensive experiments on ablation studies and comparisons with state-of-the-art methods. Given the initial success on marrying RecogTrans with InstDisc [9, 27, 57], how to use TransRank to further boost this research line is also worth exploring. We will release our code and pre-train models to facilitate future research.

Broader Impact. Self-supervised learning is a data-hungry task, consuming expensive computational resources, though we have mitigated the effort and expense of collecting annotation. Since we have verified our model in multiple aspects and downstream tasks, we hope our released code and models can serve as a solid baseline for RecogTrans methods and deliver good initializations to benefit downstream tasks. Besides, data-driven methods often bring the risk of learning biases and preserve them in downstream tasks. We encourage users to carefully consider the consequences of the biases when adopting our model.

References

[1] Jean-Baptiste Alayrac, Adria Recasens, Rosalia Schneider, Relja Arandjelovic, Jason Ramapuram, Jeffrey De Fauw, Lucas Smaira, Sander Dieleman, and Andrew Zisserman. Self-supervised multimodal versatile networks. *NeurIPS*, 2(6):7, 2020. 2, 3

[2] Humam Alwassel, Dhruv Mahajan, Bruno Korbar, Lorenzo Torresani, Bernard Ghanem, and Du Tran. Self-supervised learning by cross-modal audio-video clustering. *arXiv:1911.12667*, 2019. 2, 3, 8

[3] 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 *CVPR*, pages 9922–9931, 2020. 1, 2, 3, 4, 5, 7, 8

[4] Andrew Brown, Weidi Xie, Vicky Kalogeiton, and Andrew Zisserman. Smooth-ap: Smoothing the path towards large-scale image retrieval. In *European Conference on Computer Vision*, pages 677–694. Springer, 2020. 2

[5] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *CVPR*, pages 961–970, 2015. 1

[6] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *ECCV*, pages 132–149, 2018. 2

[7] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *arXiv:2006.09882*, 2020. 2

[8] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, pages 6299–6308, 2017. 1

[9] Peihao Chen, Deng Huang, Dongliang He, Xiang Long, Runhao Zeng, Shilei Wen, Mingkui Tan, and Chuang Gan. Rspnet: Relative speed perception for unsupervised video representation learning. In *AAAI*, volume 1, 2021. 1, 3, 5, 8

[10] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, pages 1597–1607. PMLR, 2020. 1, 2

[11] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey Hinton. Big self-supervised models are strong semi-supervised learners. *arXiv:2006.10029*, 2020. 2

[12] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *CVPR*, pages 15750–15758, 2021. 2, 3

[13] MMAction2 Contributors. Openmmlab's next generation video understanding toolbox and benchmark. https://github.com/open-mmlab/mmaction2, 2020. 6

[14] Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised visual representation learning by context prediction. In *ICCV*, pages 1422–1430, 2015. 2

[15] Haodong Duan, Yue Zhao, Yuanjun Xiong, Wentao Liu, and Dahua Lin. Omni-sourced webly-supervised learning for video recognition. In *ECCV*, pages 670–688. Springer, 2020. 1

[16] Christoph Feichtenhofer, Haoqi Fan, Bo Xiong, Ross Girshick, and Kaiming He. A large-scale study on unsupervised spatiotemporal representation learning. In *CVPR*, pages 3299–3309, 2021. 1, 3

[17] Chuang Gan, Boqing Gong, Kun Liu, Hao Su, and Leonidas J Guibas. Geometry guided convolutional neural networks for self-supervised video representation learning. In *CVPR*, pages 5589–5597, 2018. 3

[18] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. *arXiv:1803.07728*, 2018. 2, 6

[19] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzyńska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, Florian Hoppe, Christian Thurau, Ingo Bax, and Roland Memisevic. The "something something" video database for learning and evaluating visual common sense, 2017. 1, 6

[20] 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:2006.07733*, 2020. 1, 2

[21] Tengda Han, Weidi Xie, and Andrew Zisserman. Video representation learning by dense predictive coding. In *CVPRW*, pages 0–0, 2019. 2, 3, 8

[22] Tengda Han, Weidi Xie, and Andrew Zisserman. Self-supervised co-training for video representation learning. In *Neurips*, 2020. 3, 8

[23] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In *CVPR*, pages 6546–6555, 2018. 6

[24] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, pages 9729–9738, 2020. 1, 2, 3

[25] Deng Huang, Wenhao Wu, Weiwen Hu, Xu Liu, Dongliang He, Zhihua Wu, Xiangmiao Wu, Mingkui Tan, and Errui Ding. Ascnet: Self-supervised video representation learning with appearance-speed consistency. *arXiv:2106.02342*, 2021. 3, 8

[26] Phillip Isola, Daniel Zoran, Dilip Krishnan, and Edward H Adelson. Learning visual groups from co-occurrences in space and time. *arXiv:1511.06811*, 2015. 3

[27] Simon Jenni and Hailin Jin. Time-equivariant contrastive video representation learning. In *ICCV*, pages 9970–9980, 2021. 3, 8

[28] Simon Jenni, Givi Meishvili, and Paolo Favaro. Video representation learning by recognizing temporal transformations. In *ECCV*, pages 425–442. Springer, 2020. 1, 2, 3, 5, 7, 8

[29] Longlong Jing, Xiaodong Yang, Jingen Liu, and Yingli Tian. Self-supervised spatiotemporal feature learning via video rotation prediction. *arXiv:1811.11387*, 2018. 1, 2, 3, 6, 8

[30] Dahun Kim, Donghyeon Cho, and In So Kweon. Self-supervised video representation learning with space-time cubic puzzles. In *AAAI*, volume 33, pages 8545–8552, 2019. 2, 8

[31] Bruno Korbar, Du Tran, and Lorenzo Torresani. Cooperative learning of audio and video models from self-supervised synchronization. *arXiv:1807.00230*, 2018. 3

[32] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. Hmdb: a large video database for human motion recognition. In *ICCV*, pages 2556–2563. IEEE, 2011. 1, 4, 6

[33] Zihang Lai, Erika Lu, and Weidi Xie. Mast: A memory-augmented self-supervised tracker. In *CVPR*, pages 6479–6488, 2020. 3

[34] Hsin-Ying Lee, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. Unsupervised representation learning by sorting sequences. In *ICCV*, pages 667–676, 2017. 2

[35] Tianhao Li and Limin Wang. Learning spatiotemporal features via video and text pair discrimination. *arXiv:2001.05691*, 2020. 3

[36] Wei Li, Dezhao Luo, Bo Fang, Yu Zhou, and Weiping Wang. Video 3d sampling for self-supervised representation learning. *arXiv:2107.03578*, 2021. 1, 5, 8

[37] Ji Lin, Chuang Gan, and Song Han. Tsm: Temporal shift module for efficient video understanding. In *ICCV*, pages 7083–7093, 2019. 3

[38] 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 *AAAI*, volume 34, pages 11701–11708, 2020. 8

[39] Zelun Luo, Boya Peng, De-An Huang, Alexandre Alahi, and Li Fei-Fei. Unsupervised learning of long-term motion dynamics for videos. In *CVPR*, pages 2203–2212, 2017. 3

[40] 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 *CVPR*, pages 9879–9889, 2020. 3

[41] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *ICCV*, pages 2630–2640, 2019. 1

[42] Ishan Misra, C Lawrence Zitnick, and Martial Hebert. Shuffle and learn: unsupervised learning using temporal order verification. In *ECCV*, pages 527–544. Springer, 2016. 2

[43] Mathew Monfort, Alex Andonian, Bolei Zhou, Kandan Ramakrishnan, Sarah Adel Bargal, Yan Yan, Lisa Brown, Quanfu Fan, Dan Gutfreund, Carl Vondrick, et al. Moments in time dataset: one million videos for event understanding. *TPAMI*, 2019. 1

[44] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *ECCV*, pages 69–84. Springer, 2016. 2

[45] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv:1807.03748*, 2018. 2

[46] Tian Pan, Yibing Song, Tianyu Yang, Wenhao Jiang, and Wei Liu. Videomoco: Contrastive video representation learning with temporally adversarial examples. In *CVPR*, pages 11205–11214, 2021. 1, 3, 8

[47] Rishubh Parihar, Gaurav Ramola, Ranajit Saha, Ravi Kini, Aniket Rege, and Sudha Velusamy. Spatio-temporal video representation learning for ai based video playback style prediction. *arXiv:2110.01015*, 2021. 3

[48] Mandela Patrick, Yuki M Asano, Polina Kuznetsova, Ruth Fong, João F Henriques, Geoffrey Zweig, and Andrea Vedaldi. On compositions of transformations in contrastive self-supervised learning. In *ICCV*, pages 9577–9587, 2021. 2, 3

[49] Rui Qian, Tianjian Meng, Boqing Gong, Ming-Hsuan Yang, Huisheng Wang, Serge Belongie, and Yin Cui. Spatiotemporal contrastive video representation learning, 2021. 1, 3, 8

[50] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv:1212.0402*, 2012. 1, 4, 6

[51] Chen Sun, Fabien Baradel, Kevin Murphy, and Cordelia Schmid. Learning video representations using contrastive bidirectional transformer. *arXiv:1906.05743*, 2019. 3

[52] Li Tao, Xueting Wang, and Toshihiko Yamasaki. Selfsupervised video representation using pretext-contrastive learning. *arXiv:2010.15464*, 2, 2020. 3

[53] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *CVPR*, pages 6450–6459, 2018. 6

[54] Ali Varamesh, Ali Diba, Tinne Tuytelaars, and Luc Van Gool. Self-supervised ranking for representation learning. *arXiv preprint arXiv:2010.07258*, 2020. 2

[55] Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Anticipating visual representations from unlabeled video. In *CVPR*, pages 98–106, 2016. 3

[56] Guangting Wang, Yizhou Zhou, Chong Luo, Wenxuan Xie, Wenjun Zeng, and Zhiwei Xiong. Unsupervised visual representation learning by tracking patches in video. In *CVPR*, pages 2563–2572, 2021. 3

[57] Jiangliu Wang, Jianbo Jiao, and Yun-Hui Liu. Self-supervised video representation learning by pace prediction. In *ECCV*, pages 504–521. Springer, 2020. 1, 2, 3, 5, 8

[58] Xiaolong Wang, Allan Jabri, and Alexei A Efros. Learning correspondence from the cycle-consistency of time. In *CVPR*, pages 2566–2576, 2019. 3

[59] Donglai Wei, Joseph J Lim, Andrew Zisserman, and William T Freeman. Learning and using the arrow of time. In *CVPR*, pages 8052–8060, 2018. 1, 2, 5

[60] Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. Rethinking spatiotemporal feature learning for video understanding. *arXiv:1712.04851*, 1(2):5, 2017. 3, 6

[61] Dejing Xu, Jun Xiao, Zhou Zhao, Jian Shao, Di Xie, and Yueting Zhuang. Self-supervised spatiotemporal learning via video clip order prediction. In *CVPR*, pages 10334–10343, 2019. 2, 8

[62] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *CVPR*, pages 5288–5296, 2016. 1

[63] Ceyuan Yang, Yinghao Xu, Bo Dai, and Bolei Zhou. Video representation learning with visual tempo consistency. *arXiv:2006.15489*, 2020. 3

[64] Ceyuan Yang, Yinghao Xu, Jianping Shi, Bo Dai, and Bolei Zhou. Temporal pyramid network for action recognition. In *CVPR*, pages 591–600, 2020. 3

[65] Yuan Yao, Chang Liu, Dezhao Luo, Yu Zhou, and Qixiang Ye. Video playback rate perception for self-supervised spatio-temporal representation learning. In *CVPR*, pages 6548–6557, 2020. 2, 8

[66] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. *arXiv:2103.03230*, 2021. 2

[67] Hang Zhao, Antonio Torralba, Lorenzo Torresani, and Zhicheng Yan. Hacs: Human action clips and segments dataset for recognition and temporal localization. In *ICCV*, pages 8668–8678, 2019. 1

[68] Yue Zhao, Yuanjun Xiong, Limin Wang, Zhirong Wu, Xiaoou Tang, and Dahua Lin. Temporal action detection with structured segment networks. In *ICCV*, pages 2914–2923, 2017. 1

[69] Bolei Zhou, Alex Andonian, Aude Oliva, and Antonio Torralba. Temporal relational reasoning in videos. In *ECCV*, pages 803–818, 2018. 3

[70] Luowei Zhou, Chenliang Xu, and Jason J Corso. Towards automatic learning of procedures from web instructional videos. In *AAAI*, 2018. 1