Enhancing Self-supervised Video Representation Learning via Multi-level Feature Optimization

Rui Qian[1], Yuxi Li[1,2], Huabin Liu[1], John See[3], Shuangrui Ding[1], Xian Liu[4], Dian Li[5], Weiyao Lin[1]*

[1]Shanghai Jiao Tong Univeristy, [2]Tencent Youtu Lab, [3]Heriot-Watt University
[4]Zhejiang University, [5]Tencent PCG

{qrui9911,huabinliu,dsr1212,wylin}@sjtu.edu.cn
{yukiyxli,goodli}@tencent.com,j.see@hw.ac.uk,alvinliu@zju.edu.cn

Abstract

The crux of self-supervised video representation learning is to build general features from unlabeled videos. However, most recent works have mainly focused on high-level semantics and neglected lower-level representations and their temporal relationship which are crucial for general video understanding. To address these challenges, this paper proposes a multi-level feature optimization framework to improve the generalization and temporal modeling ability of learned video representations. Concretely, high-level features obtained from naive and prototypical contrastive learning are utilized to build distribution graphs, guiding the process of low-level and mid-level feature learning. We also devise a simple temporal modeling module from multi-level features to enhance motion pattern learning. Experiments demonstrate that multi-level feature optimization with the graph constraint and temporal modeling can greatly improve the representation ability in video understanding. Code is available here.

1. Introduction

Video representation learning has been a fundamental problem in computer vision to solve a series of video analysis tasks, *e.g.*, action recognition and detection [11, 72, 7, 18, 39, 79], video retrieval [40, 45], video caption [60, 48], and etc. To address this problem, some large-scale human annotated datasets, *e.g.*, Kinetics [11], ActivityNet [7], YouTube-8M [1], are developed to facilitate video understanding in specific downstream tasks. However, human labeling on videos is expensive, and fully-supervised methods fail to leverage massive unlabeled video data. Therefore, it is significant to develop unsupervised video representation learning without resorting to manual labeling.
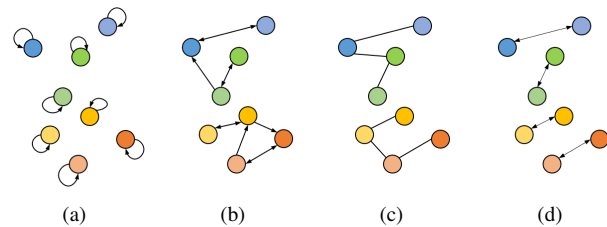

Figure 1. Graph presentation of four conditions. The nodes denote different samples, the edges present sample-wise relationships, and different colors indicate different characteristics, *e.g.*, appearance, motion and semantic. Fig. 1(a) one hot label in InfoNCE loss, we use self-loop to present only the instance with its augmented view are regarded as positive. Fig. 1(b) instance-wise similarity distribution, measured by cosine similarity in embedding space[1]. We use arrows to show the samples with similarity above threshold. Fig. 1(c) semantic-wise distribution, we connect samples of the same category. Fig. 1(d) comprehensive distribution formulated by the intersection of the former two. Note that we omit self-loops in the last three for concise presentation.

To achieve this goal, early works designed various pretext tasks to uncover effective supervision from video sequences [6, 46, 33, 31, 74, 63]. Recently, contrastive learning has shown to be powerful in image representation learning [28, 47, 55, 12, 26, 77]. It encourages augmentation invariant representations by leveraging instance discrimination to attract augmented samples of the same instance and repel those of different instances. Later, beyond naive instance discrimination, inter-image relationships and semantic structures are proved helpful for learning high-quality representations [36, 67]. To expand this pipeline to video domain, diverse spatiotemporal augmentation techniques are proposed to construct contrastive pairs and enhance motion modeling [17, 50, 64, 75, 14]. Some works used contrastive learning to form temporal cycle or make future prediction to boost dense spatiotemporal feature mod-

---

[1]The instance-wise similarity distribution is asymmetric due to the normalization, it is a directed graph.

*Corresponding author. Email: wylin@sjtu.edu.cn

eling [30, 23].

However, there are obvious limitations in these works. Firstly, previous works only explore either instance-wise or semantic-wise distribution [17, 50, 36], lacking a comprehensive perspective over both sides. Secondly, less effort has been placed on low-level features than high-level representations, while the former is proven critical for knowledge transfer [80]. Third, directly performing temporal augmentations, *e.g.*, shuffle and reverse, at input level instead of feature level could impair feature learning [4].

To address these challenges, we propose a novel framework that explicitly optimizes features from a *unified multi-level view* to achieve more general representations. The representations from different levels of deep neural networks show different generalization and abstraction properties. Specifically, it is of common view that high-level features are more representative towards instances or semantics but less feasible towards cross-task transfer. In contrast, low-level features are transfer-friendly but lack structural information over samples, and are particularly sensitive to temporal statistics.

This consideration is particularly meaningful from different perspectives. In a high-level sense, we optimize the deep representation from two aspects: 1) instance discrimination with conventional InfoNCE loss; 2) semantic structure modeling with a prototypical branch. In this way, the high-level representations can procure structural relationship among samples by formulating both instance- and semantic-wise relationships into distribution graphs as depicted in Fig. 1. In a low-level sense, these distribution graphs serve as reliable cues to aggregate samples that share similar semantics and instance characteristics (*e.g.*, appearance, motion) for better optimization in multiple shallower feature spaces. Through this, low-level features are imposed with high-level relation knowledge while keeping good cross-task generalization ability.

Since low-level representation is sensitive to input temporal sequences, we replace the previous data-level temporal augmentation methods with a multi-level solution to enhance the temporal modeling of the pretrained representation. First, we apply temporal augmentation on multi-level features to construct contrastive pairs that have different motion patterns with the objective designed to distinguish the augmented samples and original ones. Second, a retrieval task is proposed to match the features in short and long time spans based on their semantic consistency. Compared with previous data-level solutions, our method avoids forcing the backbone model to adapt to unnatural sequences which corrupts spatiotemporal statistics. Experimental results reveal that our proposed simple temporal modeling is more general and suits different network backbones, while the conventional augmentation technique is somewhat limited to two-pathway networks like SlowFast [16].

In brief, our contributions can be summarized as:

- We propose a multi-level feature optimization framework for unsupervised video representation learning. Both instance- and semantic-wise knowledge learned from high-level features are leveraged to form a more reliable self-supervisory signal, which is employed to optimize low-level feature distributions thereby enhancing transferability.

- We develop a simple but effective temporal modeling module with a multi-level augmentation scheme for more robust temporal analysis.

- Our method achieves state-of-the-art performance on two downstream tasks, action recognition and video retrieval, across two datasets, UCF-101 and HMDB-51. Ablation studies demonstrate the efficacy of multi-level feature optimization as well as the new temporal modeling strategy.

2. Related Work

2.1. Contrastive Representation Learning

Contrastive learning aims to discriminate instances by attracting the positive pairs and repelling the negative pairs [20, 19, 70]. A line of works have adopted this approach for self-supervised representation learning [28, 47, 55, 12, 26, 77]. But there exits one main drawback of the one-hot labels in InfoNCE loss, *i.e.* it only regards the augmentation of the query as positive, and considers all other samples as equally negative. To address this problem, [68, 15] employed the similarity distribution in the embedding space to guide contrastive learning in another view. Further, [69, 71, 67, 21, 36] demonstrated that the semantic-wise relationships between different samples could improve the high-level representation. To better extract the latent semantics in unlabelled data, [10, 2, 3, 51] leveraged Sinkhorn-Knopp algorithm [13] to generate uniformly distributed clusters as pseudo labels for pretraining. However, [80] demonstrated only utilizing instance discrimination or semantic label is not the optimal solution to establish transferable representations. Therefore, we propose to jointly consider the instance- and semantic-wise similarity distribution to form a reliable self-supervision signal, which simultaneously maintains the learned instance-wise unique information and filters out hard negatives.

2.2. Multi-level Feature Analysis

The features of different layers in the deep neural network tend to possess different attributes, *e.g.*, lower-level features contain more information of object shapes and are more transferable, while higher-level features contain

more texture cues and are more specific to certain semantics [29, 83, 78, 80]. [80] demonstrated that it is the low-level and mid-level features that majorly transfer from pre-trained networks to downstream tasks. However, most existing works on self-supervised representation learning only focus on high-level features. Though [73] attempted to optimize intermediate feature vectors but did not establish relationships between different levels. While in our work, we use joint constraint of instance- and semantic-wise distributions inferred from high-level features to explicitly optimize low-level and mid-level representations, which significantly facilitates pretrained knowledge transfer.

2.3. Self-supervised Video Representation Learning

In self-supervised video representation learning, a line of works designed various pretext tasks, *e.g.*, temporal ordering [46, 74, 75], spatiotemporal puzzles [33, 63], colorization [59], playback speed prediction [31, 6] and temporal cycle-consistency [66, 30, 37]. Some works proposed to predict future frames from the given sequence to learn feature embeddings [58, 57, 43, 5]. Recently, inspired by the success of contrastive learning in static image, a line of works expanded contrastive learning pipeline to video domain [17, 50, 44, 64, 41]. Typically, [22, 23] employed InfoNCE loss for dense future prediction, [34, 24] performed instance discrimination across different domains to boost video representation. Though contrastive self-supervised learning contributes to better representation, the temporal information in videos is not well leveraged. [4] revealed that directly applying temporal augmentations on input sequences even impairs the performance since these unnatural sequences could corrupt spatiotemporal statistics. To tackle this problem, [62, 61] disentangled static appearance and dynamic motion information but required complex training procedures. In contrast, we propose a simple yet effective operation to apply temporal augmentations on extracted multi-level features. In this way, we manage to embed the temporal characteristics to the video backbone without enforcing the network to adapt to unnatural sequences.

3. Method

In this section, we introduce our proposed multi-level feature optimization framework as shown in Fig. 2. Concretely, we first present our instance and semantic discrimination on high-level representations. Next, we develop the instance- and semantic-wise distribution graph to generate reliable constraint for multi-level feature optimization. Then, we propose a simple temporal modeling approach to improve temporal discrimination at different time scales.

3.1. Beyond Instance Discrimination

Recent contrastive learning methods based on instance discrimination have shown superior performance on self-supervised representation learning, but the one-hot labels in InfoNCE loss neglect the relationship between different samples. Specifically, as shown in Eq. 1,

$$\mathcal{L}_{NCE} = -log \frac{h(\mathbf{q}, \mathbf{q}')}{h(\mathbf{q}, \mathbf{q}') + \sum_{i=1}^{N} h(\mathbf{q}, \mathbf{k}_i)}, \quad (1)$$

where $h(\mathbf{u}, \mathbf{v}) = \exp(\mathbf{u}^T\mathbf{v}/(\tau||\mathbf{u}||_2||\mathbf{v}||_2))$ with temperature τ, given query $\mathbf{q}$ with its augmentation $\mathbf{q}'$, and a negative key list $\{\mathbf{k}_1, \mathbf{k}_2, ..., \mathbf{k}_N\}$, the InfoNCE loss only regards the augmentation of the query as positive and takes all other samples as equally negative. However, considering that existing contrastive self-supervised learning pipelines mostly require large negative pools, there exist some negative samples that may share similar characteristics, *e.g.*, appearance, motion or category, with the query. Under this circumstance, better instance discrimination would even lead to performance drop in downstream tasks [56]. To this end, besides instance-wise discrimination, we explicitly develop another branch on the projected high-level feature vectors for inter-sample relationship modeling.

Mathematically, we denote the projected high-level feature vector of the i-th sample and a-th augmentation view as $\mathbf{z}_i^a \in \mathbb{R}^C$, where C is the channel dimension. The instance discrimination learning objective can be formulated as

$$\mathcal{L}_{ins} = -\sum_{i=1}^{N}\sum_{a=1}^{2} log \frac{h(\mathbf{z}_i^1, \mathbf{z}_i^2)}{\sum_{j=1}^{N} h(\mathbf{z}_i^a, \mathbf{z}_j^*)}, \quad (2)$$

$$h(\mathbf{z}_i^a, \mathbf{z}_j^*) = \begin{cases} h(\mathbf{z}_i^1, \mathbf{z}_i^2) & \text{if } i = j \\ h(\mathbf{z}_i^a, \mathbf{z}_j^1) + h(\mathbf{z}_i^a, \mathbf{z}_j^2) & \text{if } i \neq j \end{cases} \quad (3)$$

where two augmentation views are adopted, and N is the number of samples within a batch. For inter-sample relationship modeling, we draw motivation from parametric classification approaches [8, 36] by defining a learnable matrix $\mathbf{P} \in \mathbb{R}^{C \times K}$ as prototypes to serve as pseudo category centers, where K is the number of prototypes[1]. We perform matrix multiplication between $\mathbf{z}_i^a$ and the prototypes $\mathbf{P}$ followed with softmax regression to produce the semantic-wise distribution $\mathbf{p}_i^a \in \mathbb{R}^K$. In the absence of category annotations, it is intuitive to encourage $\mathbf{p}_i$ of different augmentations to be consistent, but it lacks the discrimination between different semantics, which can lead to feature space collapse [9]. Inspired by [10, 2, 3] (where clustering is regarded as an optimal transport problem), we employ Sinkhorn-Knopp algorithm [13] to transform a set of distributions $\{\mathbf{p}_1^a, \mathbf{p}_2^a, ..., \mathbf{p}_N^a\}$ into soft targets $\{\mathbf{s}_1^a, \mathbf{s}_2^a, ..., \mathbf{s}_N^a\}$, where $\mathbf{s}_i^a \in \mathbb{R}^K$, is uniformly distributed at category level, indicating that there are around $\frac{N}{K}$ samples per category. In this way, the generated soft targets explicitly discriminate samples of different semantic groups and avoids trivial so-

---

[1] K is not required to be consistent with the number of semantic classes in the training set, it can be set to a comparatively large number as shown in experiments.
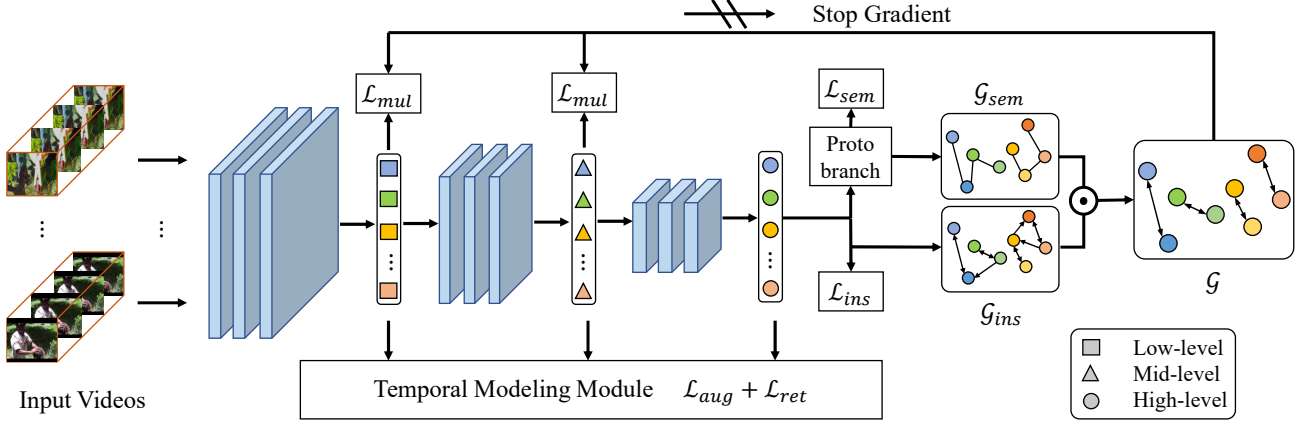
Figure 2. An overview of the multi-level feature optimization framework. We perform instance and semantic discrimination on high-level representations and infer two similarity distribution graphs $\mathcal{G}_{ins}$ and $\mathcal{G}_{sem}$, which are combined into $\mathcal{G}$, a reliable self-supervisory signal to guide low-level and mid-level representation learning. Note that we stop the gradient from back-propagating to the inferred distribution. To exploit multi-level features of different resolutions, we propose a temporal modeling strategy to enhance motion pattern discrimination.

lutions. Therefore, we optimize the model by minimizing cross-entropy between the soft targets and probability distributions of different augmentations as in Eq 4:

$$\mathcal{L}_{sem} = -\sum_{i=1}^{N}\sum_{k=1}^{K} \mathbf{s}_i^1(k)log\mathbf{p}_i^2(k) + \mathbf{s}_i^2(k)log\mathbf{p}_i^1(k), \quad (4)$$

where two augmentation views are adopted. Considering that K could be larger than batch size, we design a queue to store the semantic-wise distributions from previous batches to ensure equal partition into K prototypes, but using only those from the current batch for gradient back-propagation. Different from previous methods [26, 10, 67], we store the inferred distributions in the queue, which would generally change slower than feature vectors in the training phase. Therefore, our method could work with small batch sizes without requiring a slow-progressing momentum encoder.

Finally, we jointly leverage $\mathcal{L}_{ins}$ and $\mathcal{L}_{sem}$ to form the self-supervisory objective for high-level representations:

$$\mathcal{L}_{high} = \mathcal{L}_{ins} + \mathcal{L}_{sem}. \quad (5)$$

This enables the network to simultaneously discriminate the instances of different characteristics and uncover potential instance groups that share similar semantics.

3.2. Graph Constraint for Multi-level Features

The instance- and semantic-wise constraints lead to effective high-level representations, but it is worth noting that it is the lower-level features that mainly transfer from the pretrained network to downstream tasks [80]. Therefore, it is crucial to also pay attention to lower-level representations. However, directly applying either instance or semantic discrimination to intermediate layers does not bring improvement [73, 80], hence it remains a challenge to impose reasonable guidance on these features. Since we could infer instance- and semantic-wise distributions from high-

level features as mentioned in Section 3.1, it is intuitive to produce an ideal self-supervisory signal by taking these two distributions into consideration.

Particularly, we denote the instance-wise similarity distribution as a directed graph $\mathcal{G}_{ins}$, and semantic-wise distribution as an undirected graph $\mathcal{G}_{sem}$. Both two graphs consist of N nodes representing N different samples within a batch, and $N \times N$ edges indicating the relationship between each sample. The detailed formulation of the edges $\mathcal{E}$ is:

$$\mathcal{E}_{ins}(i,j) = \begin{cases} \mathcal{W}(i,j) & \text{if } \frac{\mathcal{W}(i,j)}{\sum_{j=1}^{N}\mathcal{W}(i,j)} \geq \eta \\ 0 & \text{if } \frac{\mathcal{W}(i,j)}{\sum_{j=1}^{N}\mathcal{W}(i,j)} < \eta \end{cases}, \quad (6)$$

$$\mathcal{E}_{sem}(i,j) = \begin{cases} 1 & \text{if } \mathrm{argmax}(\mathbf{s}_i^*) = \mathrm{argmax}(\mathbf{s}_j^*) \\ 0 & \text{if } \mathrm{argmax}(\mathbf{s}_i^*) \neq \mathrm{argmax}(\mathbf{s}_j^*) \end{cases}, \quad (7)$$

$$s.t. \quad \mathcal{W}(i,j) = \begin{cases} h(\mathbf{z}_i^1, \mathbf{z}_i^2) & \text{if } i = j \\ \bar{h}(\mathbf{z}_i^*, \mathbf{z}_j^*) & \text{if } i \neq j \end{cases}, \quad (8)$$

where two augmentation views are adopted as in Section 3.1, and η is a threshold hyper-parameter, $\bar{h}(\mathbf{z}_i^*, \mathbf{z}_j^*) = \frac{1}{4}\sum_{m=1}^{2}\sum_{n=1}^{2}h(\mathbf{z}_i^m, \mathbf{z}_j^n)$, $\mathbf{s}_i^* = \mathbf{s}_i^1 + \mathbf{s}_i^2$. In this way, $\mathcal{E}_{ins}$ indicates the inferred instance-wise similarity distribution, which respects inter-sample relationship and is more realistic data distribution than the one-hot encoding. Meanwhile, to filter out hard negatives that share high similarity in $\mathcal{E}_{ins}$, we employ $\mathcal{E}_{sem}$ to truncate the edges between nodes of different pseudo categories. Under this circumstance, we manage to comprehensively utilize unique instance-wise information and high-level semantics to generate reliable self-supervision for low-level and mid-level features. Mathematically, we jointly leverage $\mathcal{G}_{ins}$ and $\mathcal{G}_{sem}$ to form the combined graph $\mathcal{G}$, whose edge weights $\mathcal{E}$ serve as the final

soft targets:

$$\mathcal{E}(i,j) = \frac{\mathcal{E}_{ins}(i,j)\mathcal{E}_{sem}(i,j)}{\sum_{k=1}^{N} \mathcal{E}_{ins}(i,k)\mathcal{E}_{sem}(i,k)}. \quad (9)$$

We then calculate cross entropy between $\mathcal{E}$ and inferred similarity distribution to optimize lower-level features, *i.e.*,

$$\mathcal{L}_{mul} = -\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{a=1}^{2} \mathcal{E}(i,j)log\frac{h(\mathbf{z}_{\mathbf{r}_i}^a, \mathbf{z}_{\mathbf{r}_j}^*)}{\sum_{j=1}^{N} h(\mathbf{z}_{\mathbf{r}_i}^a, \mathbf{z}_{\mathbf{r}_j}^*)}, \quad (10)$$

where r indicates the feature level (low-level or mid-level) and $\mathbf{z_r}$ is the projected feature vectors of the r-th level. With this learning objective, we obtain more robust and representative lower-level features to facilitate knowledge transfer.

3.3. Temporal Modeling

Under the proposed multi-level representation optimization framework, it is intuitive to utilize the temporal information at diverse time scales to enhance motion pattern modeling since the features at different layers possess different temporal characteristics.

Motivated by previous works in video action recognition [65, 81, 49], achieving robust temporal modeling entails two aspects: 1) Semantic discrimination between different motion patterns; 2) Semantic consistency under different temporal views. Therefore, we devise two learning objectives to accomplish this.

First, for motion pattern discrimination, we use general temporal transformations, *e.g.*, temporal shuffle and reverse, to augment samples of various motion patterns. However, since the backbone is learned from scratch, directly applying augmentations on the input data will force the network to adapt to unnatural sequences. We develop a simple yet effective operation to perform temporal augmentation on multi-level features $\mathbf{f_r}$, and then leverage a lightweight motion excitation module [38] to extract motion enhanced feature representations. Temporal transformations that result in semantically inconsistent motion patterns can be regarded as a negative pair of the original sample and the InfoNCE loss is used to discriminate these augmented pairs, *i.e.*,

$$\mathcal{L}_{aug} = -\sum_{i=1}^{N}\sum_{a=1}^{2} log\frac{h(\text{ME}(\mathbf{f}_{\mathbf{r}_i}^1), \text{ME}(\mathbf{f}_{\mathbf{r}_i}^2))}{h(\text{ME}(\mathbf{f}_{\mathbf{r}_i}^1), \text{ME}(\mathbf{f}_{\mathbf{r}_i}^2)) + \mathbf{neg}_i^a}, \quad (11)$$

$$s.t. \quad \mathbf{neg}_i^a = \sum_{k=1}^{A} h(\text{ME}(\mathbf{f}_{\mathbf{r}_i}^a), \text{ME}(\text{Aug}_k(\mathbf{f}_{\mathbf{r}_i}^a))), \quad (12)$$

where ME is implemented by the Motion Excitation module in [38] followed with spatiotemporal average pooling and a two-layer multi-layer perception (MLP), Aug_k indicates k-th temporal augmentation operation. In this way, we embed the ability to discriminate motion patterns into the backbone network. Second, to boost the consistency under different temporal views, we propose to match the

feature of a specific timestamp from sequences of different lengths. Concretely, for a short sequence v_s covering timestamp $[t_1, t_2]$ and a long sequence v_l covering $[t_3, t_4]$, where $t_3 < t_1 < t_2 < t_4$, we aim to retrieve the feature at each timestamp of v_s in the feature set of v_l. Similarly, we also formulate it as a contrastive learning problem, where the feature of corresponding timestamp in v_l serves as the positive key, while others serve as negatives, *i.e.*,

$$\mathcal{L}_{ret} = -\sum_{t_q \in [t_1, t_2]} log\frac{h(v_s(t_q), v_l(t_q))}{\sum_{t_k \in [t_3, t_4]} h(v_s(t_q), v_l(t_k))}. \quad (13)$$

By leveraging $\mathcal{L}_{aug}$ and $\mathcal{L}_{ret}$, we achieve motion pattern discrimination as well as temporally consistent understanding of different views. Moreover, both learning objectives are implemented on multi-level features with diverse resolutions, leading to more robust temporal modeling.

4. Experiment

4.1. Dataset and Evaluation

We use three popular video action recognition datasets, Kinetics-400 [11], UCF-101 [53] and HMDB-51 [35]. For self-supervised pretraining, we use the training set of UCF-101 or Kinetics-400 for fair comparisons. For the downstream tasks, following [6, 23, 34], we use split 1 of UCF-101 and HMDB-51 for evaluation.

4.2. Implementation Details

Self-supervised Pretraining. We use R3D-18 [25] or S3D [72] as the backbone network. For temporal augmentation, we use temporal shuffle and reverse as two typical transformations. For the definition of contrastive pairs, we regard clips from the same video as positive pairs, and those of different videos as negative. Specifically, we randomly sample 32 RGB frames within a video, and uniformly split them into two 16-frame clips with resolution 112×112 to form positive pairs. For the proposed timestamp retrieval, we regard 16-frame clips as short sequences and the 32-frame clips as long sequences. For multi-level feature optimization, we formulate it as a two-stage procedure. In the first few epochs, we only use Eq. 5 to optimize high-level features until they could generate reliable soft targets in Eq. 9. Then, we jointly use Eq. 5 and Eq. 10 for multi-level feature learning. The specific definition of multi-level features is listed in the Supplementary Material. We use batch size of 256, and set default number of prototypes to 1000 with queue length 1024. In total, we train for 100 epochs on Kinetics-400, and 300 epochs on UCF-101 using ADAM with an initial learning rate of 10^{-3} and weight decay of 10^{-5}. The learning rate is decayed by 10 at 70 epochs for Kinetics-400, and 200 epochs for UCF-101.
Action Recognition. For action recognition, we initialize the backbone with pretrained model parameters except

Method	Backbone	Dataset	Res	Freeze	UCF	HMDB
CBT [54]	S3D	K600	112	✓	54.0	29.5
CCL [34]	R3D-18	K400	112	✓	52.1	27.8
MemDPC† [23]	R2D3D-34	K400	224	✓	54.1	30.5
TaCo [4]	R3D	K400	-	✓	59.6	26.7
Ours	S3D	K400	112	✓	61.1	31.7
Ours	R3D-18	K400	112	✓	**63.2**	**33.4**
Order [74]	R(2+1)D	UCF	112	✗	72.4	30.9
VCP [42]	R3D	UCF	112	✗	66.0	31.5
STS [63]	R3D-18	UCF	112	✗	**77.8**	40.7
PRP [76]	R(2+1)D	UCF	112	✗	72.1	35.0
Ours	S3D	UCF	112	✗	74.3	37.2
Ours	R3D-18	UCF	112	✗	76.2	**41.1**
RotNet [32]	R3D-18	K400	112	✗	62.9	33.7
CBT [54]	S3D	K600	112	✗	79.5	44.6
TempTrans [31]	R3D-18	K400	112	✗	79.3	49.8
Pace [64]	R(2+1)D	K400	112	✗	77.1	36.6
ST-Puzzle [33]	R3D-18	K400	224	✗	63.9	33.7
SpeedNet [6]	S3D-G	K400	224	✗	**81.1**	46.8
MemDPC [23]	R2D3D-34	K400	224	✗	78.1	41.2
DSM [61]	R3D-34	K400	224	✗	78.2	**52.8**
Ours	S3D	K400	112	✗	76.5	42.3
Ours	R3D-18	K400	112	✗	79.1	47.6

Table 1. Comparison results for action recognition task. We show of settings of the backbone used, pretraining dataset, resolution for fair comparison. Freeze (tick) indicates linear probe, while no freeze (cross) indicates the finetune mode. † means using two-stream networks, *i.e.*, RGB and optical flow.

Method	Multi-level	UCF	HMDB
w/o	✗	58.1	28.8
One-hot	✓	57.8	28.5
Instance	✓	59.4	30.1
Semantic	✓	54.5	26.3
Combined	✗	60.4	32.3
Combined	✓	**63.2**	**33.4**

Table 2. Ablation study on multi-level feature optimization. Results are based on R3D-18.

Method	Backbone	UCF	HMDB
w/o TM	SlowFast	57.8	30.3
w/ CTM	SlowFast	**65.4**	**34.8**
w/ NTM	SlowFast	63.9	34.1
w/o TM	R3D-18	55.9	28.1
w/ CTM	R3D-18	43.2	21.1
w/ NTM	R3D-18	**63.2**	**33.4**
w/o TM	S3D	53.8	27.2
w/ CTM	S3D	41.1	19.8
w/ NTM	S3D	**61.1**	**31.7**

Table 3. Ablation study on temporal modeling. Results are shown on three backbones: one two-pathway network SlowFast, two single-pathway networks (R3D-18, S3D). TM: temporal modeling, CTM: convention temporal modeling approach, NTM: our temporal modeling strategy.

the last fully-connected layer. There are two settings for this task: 1) Finetune the whole network in a fully supervised manner (denoted as *finetune*); 2) Only train the linear classifier (denoted as *linear probe*). For evaluation, following [74, 64], we uniformly sample 10 clips for each video, then center crop and resize them to 112×112. The final prediction of each video is the average softmax probabilities of each clip. Performance is measured by Top-1 accuracy.

Video Retrieval. For video retrieval, we directly use the pretrained model as a feature extractor without finetuning. Following [74, 42], we select videos in test set as query, and aim to retrieve k-nearest neighbors in training set. We employ the cosine similarity in feature space to measure the similarity, and use Top-k recall R@k for evaluation.

4.3. Evaluation on Downstream Tasks

Action Recognition. In this subsection, we compare our method with recent state-of-the-art self-supervised video representation learning approaches on video action recognition. In Table 1, we report the Top-1 accuracy of two settings, *i.e.*, linear probe and finetune. We exclude models with much deeper backbones or multi-modal data from our comparison. Note that [27] reports that with various training settings in finetune mode, even training from scratch

could reach performances comparable to that using pretrained models. Therefore, the linear probe can more consistently compare the learned representations.

Under the linear probe setting, our method obtains the best results on both datasets. Specifically, our method with S3D and R3D-18 backbones outperform contrastive learning based approaches, CBT [54] and CCL [34], respectively, by a large margin. Even when compared with MemDPC [23] which leverages two stream information (RGB and flow), with larger resolution, our method still shows significant advantages. Additionally, our method also outperforms TaCo [4], an approach that carefully designs various temporal pretext tasks, demonstrating our model's ability to represent the temporal aspect generally.

Under the end-to-end finetune setting, for models pretrained on UCF-101, our method can outperform approaches that used simple temporal order or playback rate as their pretext task, and is comparable to STS [63] which designed a complex learning scheme to characterize appearance and motion statistics. This demonstrates that our method is capable of robust spatiotemporal modeling. For models pretrained on Kinetics dataset, our method is comparable to recent state-of-the-art approaches even with smaller resolution; this is evident of the generalization of

Method	Backbone	Dataset	UCF-101				HMDB-51			
			R@1	R@5	R@10	R@20	R@1	R@5	R@10	R@20
SpeedNet [6]	S3D-G	Kinetics-400	13.0	28.1	37.5	49.5	-	-	-	-
VCP [42]	R3D	UCF-101	18.6	33.6	42.5	53.3	7.6	24.4	36.3	53.6
Pace [64]	R3D-18	UCF-101	23.8	38.1	46.4	56.6	9.6	26.9	41.1	56.1
MemDPC [23]	R2D3D-34	UCF-101	20.2	40.4	52.4	64.7	7.7	25.7	40.6	57.7
PRP [76]	R3D	UCF-101	22.8	38.5	46.7	55.2	8.2	25.8	38.5	53.3
DSM [61]	I3D	UCF-101	17.4	35.2	45.3	57.8	7.6	23.3	36.5	52.5
STS [63]	R3D-18	UCF-101	38.3	**59.9**	68.9	77.2	18.0	37.2	50.7	**64.8**
Ours	R3D-18	UCF-101	**39.6**	57.6	**69.2**	**78.0**	**18.8**	**39.2**	**51.0**	63.7
Ours	R3D-18	Kinetics-400	41.5	60.6	71.2	80.1	20.7	40.8	55.2	68.3

Table 4. Comparison results for video retrieval task. We report R@k (k=1,5,10,20) on UCF-101 and HMDB-51 datasets.

learned representations. Note that due to limited computational resources, we only report results with resolution 112 and training epochs 100. According to [50, 63], further improvement is expected when using resolution 224 and more epochs for self-supervised pretraining.

Video Retrieval. Besides the video action recognition task, we also report the video retrieval performance. Table 4 shows the quantitative results on UCF-101 and HMDB-51. Our method pretrained on UCF-101 is superior to other approaches over two datasets. Our method is significantly better than those using temporal cues to design pretext tasks. Though DSM [61] and STS [63] designed elaborate operations to build static appearance and dynamic motion statistics, our higher performance indicates good transferability of knowledge to the downstream task, hence showing the efficacy of our multi-level feature optimization. Further improvement can be observed when Kinetics-400 is utilized.

4.4. Ablation Study

Here, we present ablation studies on key modules in the framework as well as some crucial experiment settings. We report the results on action recognition under linear probe setting to evaluate the learned video representations.

Multi-level Optimization. In this work, we use the graph constraint in Eq. 9 to guide lower-level feature learning. We compare it with using different constraints for lower-level features: one-hot labels in Eq. 1, only instance-wise distribution in Eq. 6, only semantic-wise distribution in Eq. 7, and no constraint *i.e.*, only $\mathcal{L}_{high}$ for high-level features. Besides, we also compare with using combined graph constraint on high-level features as an extra loss term. Results on R3D-18 for these different settings are shown in Table 2. We regard the method that only optimizes high-level representations without any constraints as baseline. When using one-hot labels (in InfoNCE loss) as self-supervision for lower-level features, the poorer-than-baseline performance can be explained by gradient competition [52, 73]. Using $\mathcal{E}_{ins}$ improves the results, while $\mathcal{E}_{sem}$ only appears to badly corrupt the learned representations. This is because $\mathcal{E}_{ins}$ is a soft probability distribution, the learning process is similar to distilling knowledge from high-level representations.

On the contrary, $\mathcal{E}_{sem}$ is a hard 0-1 distribution and enforces invariance between samples of the same inferred category. Our method employing both constraints shows significant improvement as combining both distributions can yield reliable self-supervision. We also observe that introducing graph constraint on high-level representations does bring improvement but still less effective than our full pipeline. This shows that the multi-level feature optimization produces more transferable representations.

Temporal Modeling. We compare our temporal modeling approach with the conventional temporal augmentation technique, which shuffles or reverses input video clips to construct contrastive pairs, on three backbones (R3D-18, S3D, and two-pathway network SlowFast). Table 3 shows that the conventional approach improves the action recognition performance of SlowFast network, but not the performance of R3D and S3D. This is because the model is trained from scratch, it needs to learn robust spatiotemporal statistics from the input data. For R3D and S3D that use a single 3D convolution pathway to learn 3D features, the temporally shuffled or reversed clips may exhibit different spatiotemporal statistics from what is deemed as natural, thus corrupting the learned representations. SlowFast's explicit disentanglement of static appearance and dynamic motions allows temporally augmented clips of different motion patterns to thrive well. In contrast, our proposed temporal modeling method performs augmentation and discrimination in the multi-level feature space. The augmentation part, which is particularly analogous to the projection head of [12], is only utilized during training, hence the parameters do not affect backbone inference, avoiding unnatural sequences. To sum up, our simple temporal modeling operation is effective for both single-pathway and two-pathway backbones, while the conventional approach might be limited to two-pathway networks.

Number of Prototypes. We explore the influence of the hyper-parameter K, the number of prototypes, on action recognition. We show the results pretrained on UCF-101 and Kinetics-400 with a range of K values in Fig. 3. It demonstrates that it is not necessary to set K equal to the number of categories of a specific dataset, *e.g.*, 101 on UCF,
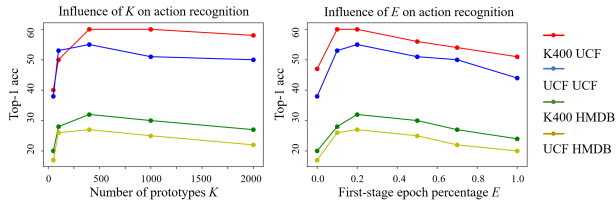
Figure 3. Ablation study on two hyper-parameters: K and E. The legend presents pretrained dataset and action recognition evaluation dataset, *e.g.*, the red line denotes pretraining on Kinetics-400 and evaluation on UCF-101.
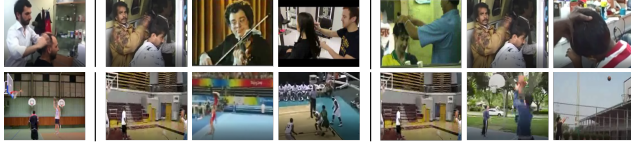

Figure 4. Retrieval of Top-3 similar samples in two distributions. Left: Query sample; Middle: Top-3 samples of instance-wise distribution, Right: Top-3 samples of semantic-wise distribution.

400 on K400. Instead, by setting K to a relatively larger number, the Top-1 accuracy is still comparatively high. It is worth noting that if K is too small (especially smaller than the number of categories), the performance drops significantly. Because when K is too small, the learned semantic-wise discrimination is too coarse-grained and fails to filter out the hard negatives when formulating the graph constraint. In summary, it is not difficult to set a reasonable value for K for pretraining. A comparatively large number is enough, and we set $K = 1000$ as default.

Two-stage Training Split. We formulate our multi-level feature optimization as a two-stage process. At the first stage, we only optimize high-level features to obtain good initialization of instance- and semantic-wise distribution. At the second stage, we jointly optimize all multi-level features. Here, we explore the influence of the stage split by training epochs. Fig. 3 compares different first stage training portion as percentage E of epochs, and we make several observations. First, when $E = 0$, *i.e.*, no first stage training, the supervision for lower-level features (in Eq. 9) is randomly initialized and this derails feature learning. Second, when E is large, optimization on lower-level features are not sufficient, hence weaker transferability of learned representations affects retrieval performance. Third, the performance is best when E is within the range of $[10\%, 20\%]$, a stable range for both UCF-101 and Kinetics-400.

4.5. Qualitative Analysis

Based on the inferred instance- and semantic-wise similarity distributions, we list the Top-3 most similar samples from each distribution based on the example query in Fig. 4. The results show that instance-wise similarity distribution provides samples with similar appearance or motion characteristics, while semantic-wise distribution
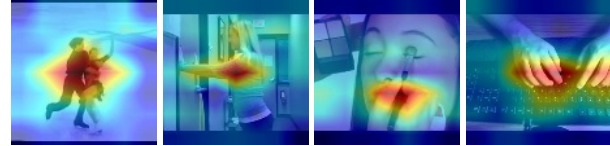
(a) Results of our temporal modeling approach.
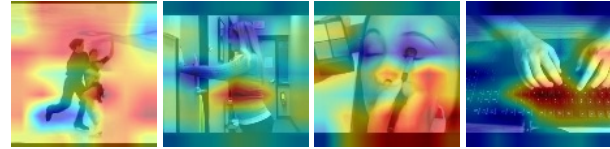
(b) Results of conventional temporal modeling approach.

Figure 5. CAM visualization of important motion areas. We use the heatmap to reveal how much temporal cues are contained in each spatial grid. Our approach learns these areas well.

provides samples of the same semantic category. The intersection of these two distributions leads to sample pairs that share both spatiotemporal characteristics and semantics. This demonstrates that the combined graph constraint could serve as a reliable self-supervisory signal that maintains unique instance-wise information and is able to discriminate different semantics.

To evaluate the temporal modeling performance, we use the pretrained backbone as a feature extractor, and train a linear classifier to discriminate temporally augmented features from the original. We provide the CAM [82] visualization in Fig. 5 to show how much temporal cues are contained in each region. It is clear that our temporal modeling strategy contributes to more accurate and discriminative motion areas, while conventional temporal augmentation is not able to perceive important motion cues well. For example, our method precisely focuses on the moving hands in the typing scene, but the conventional approach regards the keyboard as the motion key.

5. Conclusion

In this work, we propose a multi-level feature optimization framework for unsupervised video representation learning. We perform instance- and semantic-wise discrimination on high-level features, thereby employing reliable self-supervisory cues to optimize lower-level representations for improved generalization. Meanwhile, we also leverage multi-level features of various temporal spans for robust temporal modeling. Extensive experiments demonstrate that our learned representations achieve superior performance on a series of downstream tasks.

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] Yuki M Asano, Mandela Patrick, Christian Rupprecht, and Andrea Vedaldi. Labelling unlabelled videos from scratch with multi-modal self-supervision. *arXiv preprint arXiv:2006.13662*, 2020. 2, 3

[3] Yuki Markus Asano, Christian Rupprecht, and Andrea Vedaldi. Self-labelling via simultaneous clustering and representation learning. *arXiv preprint arXiv:1911.05371*, 2019. 2, 3

[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. 2, 3, 6

[5] Nadine Behrmann, Jurgen Gall, and Mehdi Noroozi. Unsupervised video representation learning by bidirectional feature prediction. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1670–1679, 2021. 3

[6] 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. 1, 3, 5, 6, 7

[7] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pages 961–970, 2015. 1

[8] Yue Cao, Zhenda Xie, Bin Liu, Yutong Lin, Zheng Zhang, and Han Hu. Parametric instance classification for unsupervised visual feature learning. *arXiv preprint arXiv:2006.14618*, 2020. 3

[9] 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

[10] 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. 2, 3, 4

[11] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017. 1, 5

[12] 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. 1, 2, 7

[13] Marco Cuturi. Sinkhorn distances: lightspeed computation of optimal transport. In *NIPS*, volume 2, page 4, 2013. 2, 3

[14] Junting Dong, Qing Shuai, Yuanqing Zhang, Xian Liu, Xiaowei Zhou, and Hujun Bao. Motion capture from internet videos. In *European Conference on Computer Vision*, pages 210–227. Springer, 2020. 1

[15] Zhiyuan Fang, Jianfeng Wang, Lijuan Wang, Lei Zhang, Yezhou Yang, and Zicheng Liu. Seed: Self-supervised distillation for visual representation. *arXiv preprint arXiv:2101.04731*, 2021. 2

[16] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6202–6211, 2019. 2

[17] 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. 1, 2, 3

[18] Chunhui Gu, Chen Sun, David A Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, et al. Ava: A video dataset of spatio-temporally localized atomic visual actions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6047–6056, 2018. 1

[19] Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 297–304. JMLR Workshop and Conference Proceedings, 2010. 2

[20] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1735–1742. IEEE, 2006. 2

[21] Tao Han, Junyu Gao, Yuan Yuan, and Qi Wang. Unsupervised semantic aggregation and deformable template matching for semi-supervised learning. *arXiv preprint arXiv:2010.05517*, 2020. 2

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

[23] Tengda Han, Weidi Xie, and Andrew Zisserman. Memory-augmented dense predictive coding for video representation learning. *arXiv preprint arXiv:2008.01065*, 2020. 2, 3, 5, 6, 7

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

[25] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Learning spatio-temporal features with 3d residual networks for action recognition. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 3154–3160, 2017. 5

[26] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Con-*

ference on Computer Vision and Pattern Recognition, pages 9729–9738, 2020. 1, 2, 4

[27] Kaiming He, Ross Girshick, and Piotr Dollár. Rethinking imagenet pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4918–4927, 2019. 6

[28] R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. *arXiv preprint arXiv:1808.06670*, 2018. 1, 2

[29] Md Amirul Islam, Matthew Kowal, Patrick Esser, Sen Jia, Bjorn Ommer, Konstantinos G Derpanis, and Neil Bruce. Shape or texture: Understanding discriminative features in cnns. *arXiv preprint arXiv:2101.11604*, 2021. 3

[30] Allan Jabri, Andrew Owens, and Alexei A Efros. Space-time correspondence as a contrastive random walk. *arXiv preprint arXiv:2006.14613*, 2020. 2, 3

[31] Simon Jenni, Givi Meishvili, and Paolo Favaro. Video representation learning by recognizing temporal transformations. *arXiv preprint arXiv:2007.10730*, 2020. 1, 3, 6

[32] Longlong Jing and Yingli Tian. Self-supervised spatiotemporal feature learning by video geometric transformations. *arXiv preprint arXiv:1811.11387*, 2(7):8, 2018. 6

[33] 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. 1, 3, 6

[34] Quan Kong, Wenpeng Wei, Ziwei Deng, Tomoaki Yoshinaga, and Tomokazu Murakami. Cycle-contrast for self-supervised video representation learning. *arXiv preprint arXiv:2010.14810*, 2020. 3, 5, 6

[35] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. Hmdb: a large video database for human motion recognition. In *2011 International conference on computer vision*, pages 2556–2563. IEEE, 2011. 5

[36] Junnan Li, Pan Zhou, Caiming Xiong, Richard Socher, and Steven CH Hoi. Prototypical contrastive learning of unsupervised representations. *arXiv preprint arXiv:2005.04966*, 2020. 1, 2, 3

[37] Xueting Li, Sifei Liu, Shalini De Mello, Xiaolong Wang, Jan Kautz, and Ming-Hsuan Yang. Joint-task self-supervised learning for temporal correspondence. *arXiv preprint arXiv:1909.11895*, 2019. 3

[38] Yan Li, Bin Ji, Xintian Shi, Jianguo Zhang, Bin Kang, and Limin Wang. Tea: Temporal excitation and aggregation for action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 909–918, 2020. 5

[39] Weiyao Lin, Huabin Liu, Shizhan Liu, Yuxi Li, Rui Qian, Tao Wang, Ning Xu, Hongkai Xiong, Guo-Jun Qi, and Nicu Sebe. Human in events: A large-scale benchmark for human-centric video analysis in complex events. *arXiv preprint arXiv:2005.04490*, 2020. 1

[40] Yang Liu, Samuel Albanie, Arsha Nagrani, and Andrew Zisserman. Use what you have: Video retrieval using representations from collaborative experts. *arXiv preprint arXiv:1907.13487*, 2019. 1

[41] Yang Liu, Keze Wang, Haoyuan Lan, and Liang Lin. Temporal contrastive graph for self-supervised video representation learning. *arXiv preprint arXiv:2101.00820*, 2021. 3

[42] 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. 6, 7

[43] Zelun Luo, Boya Peng, De-An Huang, Alexandre Alahi, and Li Fei-Fei. Unsupervised learning of long-term motion dynamics for videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2203–2212, 2017. 3

[44] Jiaqi Ma, Shuangrui Ding, and Qiaozhu Mei. Towards more practical adversarial attacks on graph neural networks. *arXiv preprint arXiv:2006.05057*, 2020. 3

[45] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2630–2640, 2019. 1

[46] 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. 1, 3

[47] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 1, 2

[48] Yingwei Pan, Ting Yao, Houqiang Li, and Tao Mei. Video captioning with transferred semantic attributes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6504–6512, 2017. 1

[49] AJ Piergiovanni and Michael S Ryoo. Representation flow for action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9945–9953, 2019. 5

[50] 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. 1, 2, 3, 7

[51] Jayanth Reddy Regatti, Aniket Anand Deshmukh, Eren Manavoglu, and Urun Dogan. Consensus clustering with unsupervised representation learning. *arXiv preprint arXiv:2010.01245*, 2020. 2

[52] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. *arXiv preprint arXiv:1412.6550*, 2014. 7

[53] 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. 5

[54] Chen Sun, Fabien Baradel, Kevin Murphy, and Cordelia Schmid. Learning video representations using contrastive bidirectional transformer. *arXiv preprint arXiv:1906.05743*, 2019. 6

[55] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. *arXiv preprint arXiv:1906.05849*, 2019. 1, 2

[56] Michael Tschannen, Josip Djolonga, Paul K Rubenstein, Sylvain Gelly, and Mario Lucic. On mutual information maximization for representation learning. *arXiv preprint arXiv:1907.13625*, 2019. 3

[57] Ruben Villegas, Jimei Yang, Seunghoon Hong, Xunyu Lin, and Honglak Lee. Decomposing motion and content for natural video sequence prediction. *arXiv preprint arXiv:1706.08033*, 2017. 3

[58] Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Anticipating visual representations from unlabeled video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 98–106, 2016. 3

[59] Carl Vondrick, Abhinav Shrivastava, Alireza Fathi, Sergio Guadarrama, and Kevin Murphy. Tracking emerges by colorizing videos. In *Proceedings of the European conference on computer vision (ECCV)*, pages 391–408, 2018. 3

[60] Bairui Wang, Lin Ma, Wei Zhang, and Wei Liu. Reconstruction network for video captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7622–7631, 2018. 1

[61] Jinpeng Wang, Yuting Gao, Ke Li, Xinyang Jiang, Xiaowei Guo, Rongrong Ji, and Xing Sun. Enhancing unsupervised video representation learning by decoupling the scene and the motion. *arXiv preprint arXiv:2009.05757*, 2020. 3, 6, 7

[62] Jinpeng Wang, Yuting Gao, Ke Li, Yiqi Lin, Andy J Ma, and Xing Sun. Removing the background by adding the background: Towards background robust self-supervised video representation learning. *arXiv preprint arXiv:2009.05769*, 2020. 3

[63] 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. 1, 3, 6, 7

[64] 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. 1, 3, 6, 7

[65] 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. 5

[66] Xiaolong Wang, Allan Jabri, and Alexei A Efros. Learning correspondence from the cycle-consistency of time. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2566–2576, 2019. 3

[67] 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. 1, 2, 4

[68] Chen Wei, Huiyu Wang, Wei Shen, and Alan Yuille. Co2: Consistent contrast for unsupervised visual representation learning. *arXiv preprint arXiv:2010.02217*, 2020. 2

[69] Longhui Wei, Lingxi Xie, Jianzhong He, Jianlong Chang, Xiaopeng Zhang, Wengang Zhou, Houqiang Li, and Qi Tian. Can semantic labels assist self-supervised visual representation learning? *arXiv preprint arXiv:2011.08621*, 2020. 2

[70] 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

[71] Jiahao Xie, Xiaohang Zhan, Ziwei Liu, Yew Soon Ong, and Chen Change Loy. Delving into inter-image invariance for unsupervised visual representations. *arXiv preprint arXiv:2008.11702*, 2020. 2

[72] 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. 1, 5

[73] Yuwen Xiong, Mengye Ren, and Raquel Urtasun. Loco: Local contrastive representation learning. *arXiv preprint arXiv:2008.01342*, 2020. 3, 4, 7

[74] 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/CVF Conference on Computer Vision and Pattern Recognition*, pages 10334–10343, 2019. 1, 3, 6

[75] Ting Yao, Yiheng Zhang, Zhaofan Qiu, Yingwei Pan, and Tao Mei. Seco: Exploring sequence supervision for unsupervised representation learning. *arXiv preprint arXiv:2008.00975*, 2020. 1, 3

[76] 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. 6, 7

[77] Mang Ye, Xu Zhang, Pong C Yuen, and Shih-Fu Chang. Unsupervised embedding learning via invariant and spreading instance feature. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6210–6219, 2019. 1, 2

[78] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? In *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014. 3

[79] Yufeng Zhang, Lianghui Ding, Yuxi Li, Weiyao Lin, Mingbi Zhao, Xiaoyuan Yu, and Yunlong Zhan. A regional distance regression network for monocular object distance estimation. *Journal of Visual Communication and Image Representation*, page 103224, 2021. 1

[80] Nanxuan Zhao, Zhirong Wu, Rynson WH Lau, and Stephen Lin. What makes instance discrimination good for transfer learning? *arXiv preprint arXiv:2006.06606*, 2020. 2, 3, 4

[81] Bolei Zhou, Alex Andonian, Aude Oliva, and Antonio Torralba. Temporal relational reasoning in videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 803–818, 2018. 5

[82] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on*

computer vision and pattern recognition, pages 2921–2929, 2016. 8

[83] Yixiong Zou, Shanghang Zhang, José MF Moura, Jian-Peng Yu, and Yonghong Tian. Revisiting mid-level patterns for distant-domain few-shot recognition. *arXiv preprint arXiv:2008.03128*, 2020. 3