Unsupervised Pre-training for Temporal Action Localization Tasks

Can Zhang[1]* Tianyu Yang[2] Junwu Weng[2] Meng Cao[1] Jue Wang[2] Yuexian Zou[1]✉
[1]School of Electronic and Computer Engineering, Peking University [2]Tencent AI Lab
zhangcan@pku.edu.cn tianyu-yang@outlook.com WE0001WU@e.ntu.edu.sg
mengcao@pku.edu.cn arphid@gmail.com zouyx@pku.edu.cn

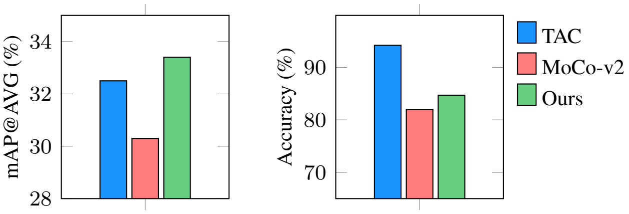
Abstract

Unsupervised video representation learning has made remarkable achievements in recent years. However, most existing methods are designed and optimized for video classification. These pre-trained models can be sub-optimal for temporal localization tasks due to the inherent discrepancy between video-level classification and clip-level localization. To bridge this gap, we make the first attempt to propose a self-supervised pretext task, coined as Pseudo Action Localization (PAL) to Unsupervisedly Pre-train feature encoders for Temporal Action Localization tasks (UP-TAL). Specifically, we first randomly select temporal regions, each of which contains multiple clips, from one video as pseudo actions and then paste them onto different temporal positions of the other two videos. The pretext task is to align the features of pasted pseudo action regions from two synthetic videos and maximize the agreement between them. Compared to the existing unsupervised video representation learning approaches, our PAL adapts better to downstream TAL tasks by introducing a temporal equivariant contrastive learning paradigm in a temporally dense and scale-aware manner. Extensive experiments show that PAL can utilize large-scale unlabeled video data to significantly boost the performance of existing TAL methods. Our codes and models will be made publicly available at https://github.com/zhang-can/UP-TAL.

1. Introduction

Model pre-training is an effective technique for training deep networks in many computer vision tasks. The core idea is to learn general representations on large-scale labeled or unlabeled data, and utilize the learned representations to improve the performance of downstream tasks with limited data. This is especially beneficial for tasks that require enormous human effort to annotate data, such as temporal action localization (TAL).

Despite the prevailing use of ready-made feature extractors [10, 48, 55] pre-trained on temporal action classification (TAC) in TAL, this pre-training strategy is sub-optimal as the *inherent discrepancy* between TAC and TAL exists. Without a doubt, this discrepancy impedes further performance improvement of TAL. Though some recent works [1, 61, 62] attempt to tackle this issue, they still rely on large-scale annotated video data. Recently, unsupervised pre-training has attracted great attention due to its potentials in exploiting large amounts of unlabeled data. Contrastive learning [15, 16, 25, 29, 43] is one of the most popular directions that focus on instance discrimination, which pulls instance-level positive pairs closer while repelling negative ones apart in the embedding space. To fill the gap between the upstream pre-training and the downstream tasks, recent contrastive learning methods focus on specifically designing pretext tasks for various downstream image tasks, *e.g.*, object detection [56,59,64], semantic segmentation [51,56], *etc*. In contrast, the progress of unsupervised pre-training in video domain is relatively lagging behind and most existing methods [2,28,32,44,45,53] are still designed and evaluated for classification tasks.



(a) Action Localization (TAL) (b) Action Classification (TAC)

Figure 1. **Comparison of Kinetics-400 pre-trained models by fine-tuning on downstream TAL (ActivityNet v1.3) and TAC (UCF101) datasets.** 'TAC' means supervised TAC pre-training, and we treat MoCo-v2 [16] with video input as our baseline. Instance-level discrimination is not well-aligned with TAL, thus unsupervised pre-training tailored for TAL is on demand.

In this paper, we make the first attempt on unsupervised pre-training for TAL tasks. One possible way [45] to achieve this is to directly extend the image contrastive learning idea to the video domain, where a video is treated

*Work done during an internship at Tencent AI Lab.

as an instance and the clips are regarded as views of instances. Those clip embeddings from the same video are pulled closer while those from different videos are pushed apart. Clearly, this way only focuses on instance (video-level) discrimination, *i.e.*, learning *time-invariant* features for specific video instances, which is required by TAC task in essence. In contrast, TAL expects the representations to be *equivariant to temporal translation and scale*. For example, if we change the start time and duration of an action instance in the input video, the output classification responses of TAC should be unchanged, while the output localization predictions of TAL need to be altered accordingly. The inherent discrepancy between these two tasks attracts our attention to question the suitability of the existing instance discrimination paradigm for TAL. Indeed, as shown in Fig. 1, such video-level discrimination is beneficial for TAC tasks, but not well-aligned with TAL tasks. So, it is desirable and challenging to design a new learning scheme that can be transferred well on TAL tasks.

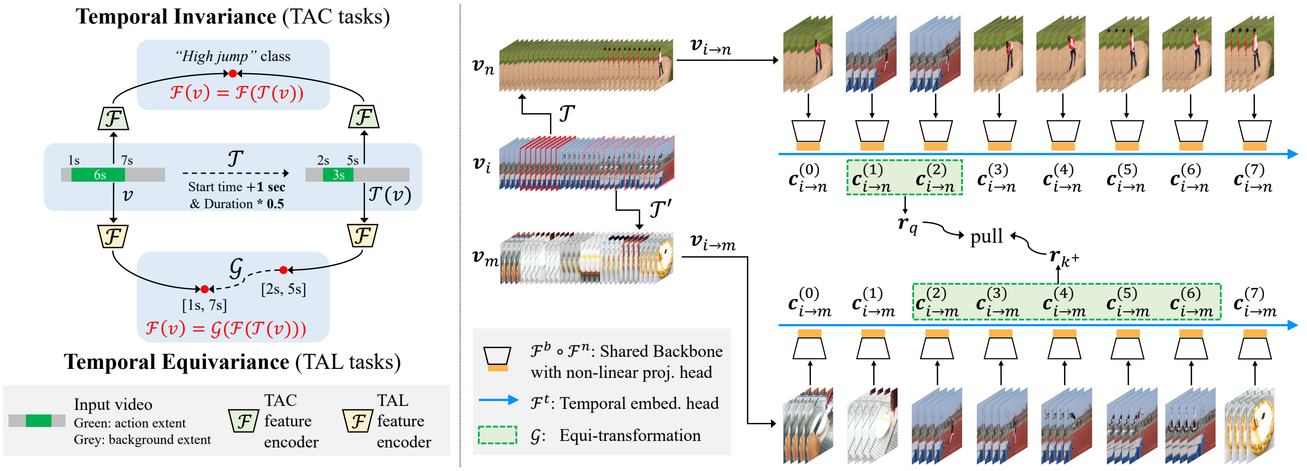
Motivated by the inherent discrepancy between TAC and TAL, we introduce *temporal equivariant* contrastive learning paradigm by designing a new unsupervised pretext task called *Pseudo Action Localization* (PAL). Specifically, to mimic the TAL-tailored data with temporal boundaries, we first construct our training set by transforming the existing large-scale TAC datasets in a cheap manner. We randomly crop two temporal regions with random temporal lengths and scales from one video as pseudo actions. Each of these regions includes multiple consecutive clips. Then we paste them onto different temporal positions of other randomly selected background videos. With the preset temporal transformation (paste location, clip length, sampling scale), the model is able to align the pseudo action features of two synthesized videos. Such transformation and alignment process are named as *input-level transformation* and *feature-level equi-transformation* in our paper. Moreover, to better align the upstream pre-training pipeline to the downstream TAL architecture, we follow the way of estimating temporal locations in TAL tasks [36, 38] by applying several layers of temporal convolutions to process the sequential clip-level features. Thereby, the information of surrounding background clips is highly involved in the final output features of pseudo action regions. With the *random* paste operation, the diversity of background-involvement is increased. Further, we propose to maximize the agreement between two aligned pseudo action region features such that the learned features are forced to focus on the most discriminative and background-irrelevant parts, thus enhancing their robustness and achieving the equivariance requirement in TAL.

We summarize our main contributions as follows: (1) To our best knowledge, this is the FIRST work focusing on unsupervised pre-training for temporal action localization tasks (UP-TAL). (2) We design an intuitive and effective self-supervised pretext task customized for TAL, called PAL. A time-equivariant contrastive learning paradigm is also introduced to perform transformed foreground discrimination, customized for TAL representation learning. (3) Extensive experiments on ActivityNet v1.3 [7], Charades-STA [22] and THUMOS'14 [31] datasets show that PAL transfers well on various downstream TAL-related tasks: Temporal Action Detection (TAD), Action Proposal Generation (APG) and Video Grounding (VG). Notably, our PAL even surpasses the supervised pre-training when using the same amount of video data.

2. Related work

Contrastive Video Representation Learning. Recently, contrastive learning [9, 15–17, 25, 29, 43] has gained increasing attention due to its outstanding performance. Essentially, these contrast-based methods focus on *instance discrimination* [58], *i.e.*, distinguishing each instance from the rest. Following this direction, recent researches [44, 45, 53, 65] extend the contrastive learning idea to the video domain, where clips from the same video are considered as positives and clips from the different videos as negatives. Besides, other directions, such as: dense future prediction [26, 27], cross-modal supervision [2, 28, 47], *etc*, have also been studied in the literature. Notably, most of these methods are designed for TAC tasks that learn time-invariant features. In contrast, we propose a novel pretext task tailored for TAL, which follows a temporal equivariant learning scheme. A concurrent work [32] also focuses on time-equivariant representation learning. Two clips from different videos but with the same relative transformation (overlap/order) are considered as positive pairs, which promotes detailed learning of motion patterns and thus is beneficial for TAC tasks. Our method differs essentially from it in the fact that positive pairs are constructed from two transformed regions (multiple clips) of the same foreground video but with different backgrounds. This facilitates the learning of TAL-friendly features such that they are robust to background interference but sensitive to temporal transformation (scale and location).

Temporal Action Localization (TAL) Tasks. Unlike TAC [10, 21, 35, 49, 50, 55, 69], the target of TAL is to temporally localize the action of interest in untrimmed videos. In general, TAL covers a range of tasks, such as: *Action Proposal Generation* (APG), *Temporal Action Detection* (TAD) and *Video Grounding* (VG), *etc*. APG aims at generating temporal proposals which are likely to contain human actions. Previous methods design temporal anchor instances for feature sequences [6, 30, 37] or directly predict boundary probabilities [36, 38]. TAD aims at predicting the temporal extent as well as the class labels of action instances. Most existing fully-supervised TAD methods [4, 11, 12, 36, 40, 60, 63] integrate the proposal gen-


Temporal Invariance (TAC tasks)
"High jump" class
$\mathcal{F}(v) = \mathcal{F}(\mathcal{T}(v))$
$\mathcal{F}$
1s 7s
6s
$\mathcal{T}$
2s 5s
3s
Start time +1 sec
& Duration * 0.5
v
$\mathcal{T}(v)$
$\mathcal{G}$
[2s, 5s]
[1s, 7s]
$\mathcal{F}(v) = \mathcal{G}(\mathcal{F}(\mathcal{T}(v)))$
Temporal Equivariance (TAL tasks)
Input video
Green: action extent
Grey: background extent
TAC feature encoder
TAL feature encoder
v_n
$v_{i\to n}$
v_i
$\mathcal{T}'$
v_m
$v_{i\to m}$
$c_{i\to n}^{(0)}$ $c_{i\to n}^{(1)}$ $c_{i\to n}^{(2)}$ $c_{i\to n}^{(3)}$ $c_{i\to n}^{(4)}$ $c_{i\to n}^{(5)}$ $c_{i\to n}^{(6)}$ $c_{i\to n}^{(7)}$
r_q
pull
r_{k^+}
$c_{i\to m}^{(0)}$ $c_{i\to m}^{(1)}$ $c_{i\to m}^{(2)}$ $c_{i\to m}^{(3)}$ $c_{i\to m}^{(4)}$ $c_{i\to m}^{(5)}$ $c_{i\to m}^{(6)}$ $c_{i\to m}^{(7)}$
$\mathcal{F}^b \circ \mathcal{F}^n$: Shared Backbone with non-linear proj. head
$\mathcal{F}^t$: Temporal embed. head
$\mathcal{G}$: Equi-transformation

(a) (b)

Figure 2. **(a) Schematic depiction of temporal invariance *vs.* temporal equivariance. (b) Overview of our PAL pretext task.** Given a video $\boldsymbol{v}_i$, we randomly sample two pseudo action regions from it and then paste them onto another two pseudo background videos at various temporal locations and scales. PAL learns temporal equivariant features by aligning pseudo action region features and maximizing the agreement between region features of the same video but with different backgrounds. Negatives are omitted for brevity.

eration and classification procedures in a unified network. Some recent works have also designed TAD algorithms under weaker supervision [42, 54, 67, 68, 72]. VG, *a.k.a*, text-to-video temporal grounding, aims to localize the time interval corresponding to a given text query. The current literature can be roughly divided into two categories, namely proposal-based [14,24,66,70] and proposal-free [3,8,39,41] architectures. For these TAL tasks, we choose three representative works (BMN [36], G-TAD [63] and LGI [41]) with officially released code to validate the efficacy of our PAL.

Supervised Pre-training for TAL. Due to the GPU memory constraint, the common practice in TAL is to first pre-train a feature encoder on large-scale trimmed TAC datasets, and then use it to extract frame-level or segment-level features in untrimmed TAL videos. Inevitably, this will result in a task discrepancy problem, since feature encoders are trained on TAC while used for TAL. This domain gap has not been fully studied though it is common in TAL. Recent advances try to bridge this gap through boundary type classification [61], foreground region classification [1] and end-to-end training [62]. Unfortunately, they all belong to the supervised pre-training paradigm and therefore rely on large-scale labeled videos. In contrast, we propose a novel method, for the first time (to our best knowledge), focusing on Unsupervised Pre-training for TAL (UP-TAL).

Cut-Paste for Data Synthesis. Cut-Paste, which cuts a part of one data sample and pastes it onto another sample, is found to be a useful data augmentation strategy when facing the data shortage issue. It has been widely adopted in supervised learning for object detection [18, 19], instance segmentation [20, 23], and self-supervised learning for image/video classification [52, 71], object detection [64] and anomaly detection [34], *etc*. The most recent work relevant to ours is BSP [61], which also synthesizes videos through temporal Cut-Paste. The essential difference lies in the fact that BSP *supervisedly* generates different types of temporal boundaries and learn to predict them to facilitate the learning of video features while our PAL synthesizes videos without using any label information and train the backbone by aligning pseudo action region features from two synthetic videos and maximizing their agreement with temporal equivariant contrastive learning.

3. Method

3.1. Intuition and Preliminaries

As mentioned in Sec. 1, the most essential difference between TAC and TAL is that the former requires *temporal invariance* while the latter desires *temporal equivariance* representations. This motivates us to question the suitability of the existing "TAC features for TAL" paradigm. Thus, in this section, we delve into the design of unsupervised pre-training customized for TAL, to reach the task alignment goal, *i.e.*, "TAL features for TAL".

For the TAC task, given a video $\boldsymbol{v}_i$ from a dataset $\boldsymbol{V} = \{\boldsymbol{v}_i\}_{i=1}^N$, the goal is to learn a feature encoding function $\mathcal{F}(\boldsymbol{v})$ with which the extracted representation is ensured to be insensitive to the temporal transformation $\mathcal{T}$, *i.e.* $\forall \boldsymbol{v} \in \boldsymbol{V} : \mathcal{F}(\mathcal{T}(\boldsymbol{v})) = \mathcal{F}(\boldsymbol{v})$, as illustrated in Fig. 2a top part. To achieve this objective, the learning strategy can be basically designed as pushing $\mathcal{F}(\mathcal{T}(\boldsymbol{v}))$ and $\mathcal{F}(\boldsymbol{v})$ close to each other in the feature space. To be more general, two random transformations $\mathcal{T}$ and $\mathcal{T}'$ are applied to $\boldsymbol{v}$ to implement the strategy, and contrastive learning [43] is involved to enforce the consistency:

$$\mathcal{F}(\mathcal{T}(\boldsymbol{v})) \overset{\text{pull}}{\rightarrow\leftarrow} \mathcal{F}(\mathcal{T}'(\boldsymbol{v})), \tag{1}$$

in which the identity mapping $\mathcal{T}_0(\boldsymbol{v}) = \boldsymbol{v}$ is also considered.

Under the scenario of TAL, we require $\mathcal{F}$ to be sensitive to the transformation $\mathcal{T}$, *i.e.* $\forall \boldsymbol{v} \in \boldsymbol{V} : \mathcal{F}(\mathcal{T}(\boldsymbol{v})) = \mathcal{T}(\mathcal{F}(\boldsymbol{v}))$, which can be re-written as $\mathcal{F}(\boldsymbol{v}) = \mathcal{G}(\mathcal{F}(\mathcal{T}(\boldsymbol{v})))$ and $\mathcal{G} \triangleq \mathcal{T}^{-1}$ (See Fig. 2a bottom part). Similar to Eqn. 1, we apply two random transformations to $\boldsymbol{v}$, and therefore have

$$\mathcal{F}(\boldsymbol{v}) = \mathcal{G}(\mathcal{F}(\mathcal{T}(\boldsymbol{v}))) = \mathcal{G}'(\mathcal{F}(\mathcal{T}'(\boldsymbol{v}))). \tag{2}$$

Intuitively, contrastive learning can be introduced here to model the temporal equivariance by forcing the features processed by two transformation pairs $(\mathcal{T}, \mathcal{G})$ and $(\mathcal{T}', \mathcal{G}')$ respectively to be analogous to one another:

$$\mathcal{G}(\mathcal{F}(\mathcal{T}(\boldsymbol{v}))) \overset{\text{pull}}{\rightarrow\leftarrow} \mathcal{G}'(\mathcal{F}(\mathcal{T}'(\boldsymbol{v}))). \tag{3}$$

In the following sections, a parameterized temporal transformation $\mathcal{T}$ tailored for TAL tasks is introduced. We delicately design a new self-supervised task called Pseudo Action Localization (PAL) with self-generated transformation signals $\mathcal{T}$, and apply the contrastive strategy to learn temporal translation and scale equivariance encoding $\mathcal{F}$.

3.2. Pseudo Action Localization

As illustrated in Fig. 2b, given a large-scale trimmed video dataset (*e.g.*, Kinetics [10]), we randomly select two temporal regions from one video (viewed as pseudo action regions), and then paste them onto another two videos (viewed as pseudo background) at various scales and locations. With the self-generated temporal locations and scales treated as prior during pre-training, the model is expected to localize the pseudo action regions from the synthesized new videos. Instead of directly predicting the paste locations and scales, we introduce the contrastive strategy to enforce the consistency between the features of two random regions defined by the priors for temporal equivariance representation learning, as illustrated in Eqn. 3.

In this pipeline, we first perform transformation $\mathcal{T}$ in the input space for TAL-tailored video generation (Sec. 3.2.1). Then we use backbone $\mathcal{F}$ and multiple heads to map the transformed videos into the feature space (Sec. 3.2.2). Next, equi-transformation $\mathcal{G}$ is applied to inverse the transformation $\mathcal{T}$ in the feature space (Sec. 3.2.3). We finally conduct region contrastive learning for TAL-customized pre-training (Sec. 3.2.4).

3.2.1 Input-Level Transformation

To learn the temporal equivariance encoding function $\mathcal{F}$, we define the transformation $\mathcal{T}$ as a video region sampling and *paste* operation. Specifically, given a video $\boldsymbol{v}_i$ as pseudo action video as well as a randomly selected video $\boldsymbol{v}_n$ as pseudo background, we first sample a random region from action video $\boldsymbol{v}_i$, and paste it onto the background video $\boldsymbol{v}_n$ to generate a synthesized video $\boldsymbol{v}_{i\rightarrow n}$. The input-level transformation $\mathcal{T}$ is then defined as follows:

$$\boldsymbol{v}_{i\rightarrow n}, s, e = \mathcal{T}(\boldsymbol{v}_i, \boldsymbol{v}_n), \tag{4}$$

where s and e represent the start and end *clip*[1] indices of the pseudo action region in the new video $\boldsymbol{v}_{i\rightarrow n}$.

To improve the robustness of the learned representation, we soften the paste operation in the implementation by changing it to a blending one with blending ratio β, which takes β of action region and mix it with $(1-\beta)$ of background region to generate the blended region. β is randomized from the range $[0.6, 1]$. Besides, spatial data augmentation is involved to increase the diversity of training data. Following the convention [28, 44, 45], we apply random cropping, horizontal flipping, Gaussian blurring and color jittering, and all are temporally consistent. In particular, instead of sampling action regions with fixed stride, we propose a *scale-aware sampling* strategy to add some randomness to the action timescale. Here, we refer to the timescale as how fast an action goes. It stems from the observation that an action video played at different speeds contains almost identical semantics. We simply model the timescale variation by sampling action region frames with different strides.

Overall, by this sample-and-paste way, our input-level transformation mimics the temporal *location* and *scale* variance in real untrimmed action videos, which also provides a strong supervision signal for TAL-tailored temporal equivariant contrastive learning.

3.2.2 Feature Encoding

Our feature encoder $\mathcal{F}$ contains a backbone $\mathcal{F}^b$ with nonlinear projection head $\mathcal{F}^n$ and a temporal embedding head $\mathcal{F}^t$, namely $\mathcal{F} = \mathcal{F}^b \circ \mathcal{F}^n \circ \mathcal{F}^t$. Formally, given the synthesized video $\boldsymbol{v}_{i\rightarrow n}$, the corresponding clip feature sequence $\{\boldsymbol{c}^{(j)}_{i\rightarrow n}\}^J_{j=1}$ is obtained by:

$$\{\boldsymbol{c}^{(j)}_{i\rightarrow n}\}^J_{j=1} = \mathcal{F}(\boldsymbol{v}_{i\rightarrow n}), \tag{5}$$

in which the backbone $\mathcal{F}^b$ is a clip-level encoder, and $\mathcal{F}^t$ is a video-level head for temporal modeling among clips. J is the number of sampled clips. It is noted that applying temporal convolutions ($\mathcal{F}^t$) on chronological clip-level features is crucial in our setting. This enables information aggregation among neighboring clips and therefore the features of pseudo action regions near the boundary can be

[1] Here we perform the temporal transformation in a clip-wise manner to align with the clip-level video encoder.

highly affected by the nearby background. In this way, our PAL can learn background-insensitive boundary features by maximizing the agreement (Sec. 3.2.4) between region features of the same video but influenced by different pseudo backgrounds.

3.2.3 Feature-Level Equi-Transformation

Recall that we aim at designing a TAL-tailored pre-training paradigm by extending the contrastive strategy to learn temporal equivariance representations. To this end, we propose to utilize additional free region-wise supervision in the form of inverse temporal transformations. In our case, the changes of pseudo action location in the input composited videos ($\boldsymbol{v}_{i\to n}$) will be reflected in the corresponding ones in their feature sequences ($\{\boldsymbol{c}_{i\to n}^{(j)}\}_{j=1}^{J}$) obtained by Eqn. 5.

To echo the input-level transformation $\mathcal{T}$ introduced in Sec. 3.2.1, we here define the feature-level equi-transformation $\mathcal{G}$ as an alignment operation. Formally, this feature alignment process is defined as:

$$\{\boldsymbol{c}_{i\to n}^{(j)}\}_{j=s}^{e} = \mathcal{G}(\{\boldsymbol{c}_{i\to n}^{(j)}\}_{j=1}^{J}, s, e), \tag{6}$$

then the region representation can be obtained by temporally averaging pooling the corresponding sequential clip-level features, *i.e.*, $\boldsymbol{r}_{i\to n}^{(s,e)} = \mathrm{TempAvgPool}(\{\boldsymbol{c}_{i\to n}^{(j)}\}_{j=s}^{e})$.

3.2.4 Contrastive Training Objective

Following the *transformation* $\mathcal{T}$ and *alignment* $\mathcal{G}$ operations introduced above, two pseudo action regions $[s, e]$ and $[s', e']$ from video $\boldsymbol{v}_i$ are extracted, and pasted onto two pseudo background videos $\boldsymbol{v}_n$ and $\boldsymbol{v}_m$ to obtain region representations $\boldsymbol{r}_{i\to n}^{(s,e)}$ and $\boldsymbol{r}_{i\to m}^{(s',e')}$. These two representations are set as the query and positive key pair $(\boldsymbol{r}_q, \boldsymbol{r}_{k+})$ in contrastive learning, namely $\boldsymbol{r}_q = \boldsymbol{r}_{i\to n}^{(s,e)}$ and $\boldsymbol{r}_{k+} = \boldsymbol{r}_{i\to m}^{(s',e')}$. The region features from other composited videos are viewed as negatives. Given the encoded query $\boldsymbol{r}_q$, positive key $\boldsymbol{r}_{k+}$ and negatives $\{\boldsymbol{r}_{k_i}\}_{i=1}^{K}$, the contrastive learning essentially encourages the query to be similar to the positive sample and dissimilar to the negative ones. Our PAL is a pretext task and independent of the detailed loss function, so we simply extend the InfoNCE [43] contrastive loss to ensure region consistency in this work:

$$\mathcal{L} = -\log \frac{\exp(\boldsymbol{r}_q \cdot \boldsymbol{r}_{k+}/\tau)}{\exp(\boldsymbol{r}_q \cdot \boldsymbol{r}_{k+}/\tau) + \sum_{i=1}^{K} \exp(\boldsymbol{r}_q \cdot \boldsymbol{r}_{k_i}/\tau)}, \tag{7}$$

where τ is a temperature hyper-parameter and K is the number of negative samples. Equipped with our proposed PAL, by minimizing the region contrastive loss, the encoding backbone $\mathcal{F}^b$ is encouraged to learn temporal equivariant features, which we believe is beneficial to the TAL tasks.

4. Experiments

4.1. Experimental Settings

To evaluate our proposed PAL, we follow the *pre-training* and *transferring* procedures: first pre-train the feature network on a large-scale trimmed dataset without category labels, then transfer the features pre-computed by the frozen backbone to the downstream TAL tasks.

4.1.1 Pre-training

Datasets. To have an apple-to-apple comparison with other self-supervised video representation learning methods, we use Kinetics [10] as the initial pre-training dataset, without using any labels. Kinetics is a large-scale trimmed action recognition benchmark. Each video has a single action class and lasts around 10 seconds. The typical version Kinetics-400 (K400) includes ∼300k videos with 400 human action classes, and the latest version Kinetics-700 (K700) contains ∼650k videos with 700 action classes.

Implementation Details. We choose I3D [10], the commonly used feature encoder in TAL, as the default backbone ($\mathcal{F}^b$) in our experiments. For temporal embedding head ($\mathcal{F}^t$), we employ two-layer temporal convolutions with a kernel size of 3 followed by ReLU activation function. We uniformly sample 8 clips (8 frames per clip) with a resolution of 112×112 for each video, and the max clip length of the pseudo action region is limited to 6. The range of the blending ratio β is set as $[0.6, 1.0]$. For our scale-aware sampling strategy, the sampling stride of frames within a clip is chosen from $[1, 4]$. Following [16], we also maintain a memory queue of 16,384 negative samples and use synchronized BN across all layers. We apply L2 norm to the output features from $\mathcal{F}^t$. The temperature τ is set to 0.07 for all experiments. For optimization, we train our PAL using the Adam algorithm with a weight decay of 10^{-5}. The initial learning rate is set as 10^{-4} and decreases by a factor of 10 when the validation loss saturates. The training takes 200 epochs in total with a batch size of 512 on 64 NVIDIA Tesla V100 GPUs.

4.1.2 Transferring to TAL tasks

Target TAL Tasks. We choose three popular temporal localization tasks to evaluate our PAL features: Temporal Action Detection (TAD), Action Proposal Generation (APG) and Video Grounding (VG).

Datasets. (1) *ActivityNet v1.3* [7] is a popular large-scale benchmark for TAD and APG tasks, including 10,024 training videos, 4,926 validation videos corresponded to 200 action classes. Each video contains 1.65 action instances on average; (2) *Charades-STA* [22] is commonly used for VG task, containing 12,408 and 3,720 text query pairs in training and test set, respectively. The average duration of videos

Table 1. **Comparison to state-of-the-art pre-training methods on the target tasks.** We use G-TAD [63] and BMN [36] as the evaluation methods for TAD and APG tasks, respectively. The results are conducted on ActivityNet v1.3 dataset. Rows highlighted in blue use fully-supervised pre-training. † represents results from [62]. * means our implementation. (TR: temporal resolution, SR: spatial resolution)

Method	**Modal**	**Dataset**	**Backbone**	**TR×SR^2**	**FLOPs**	**TAD Task (G-TAD [63])**				**APG Task (BMN [36])**			
				(per clip)	**(per clip)**	mAP@0.5	@0.75	@0.95	AVG	AR@1	@10	@100	AUC
CoCLR [28]	V+F	K400	S3D	32×128^2	47.2G	47.9	32.2	7.3	31.9	32.7	53.5	73.9	65.0
XDC [2]	V+A	IG65M	R(2+1)D-18	32×224^2	325.2G	48.4	32.6	7.6	32.3	33.2	54.1	74.0	65.4
MoCo-v2 [16] *	V	K400	I3D	8×112^2	3.6G	46.6	30.7	6.3	30.3	30.8	53.5	72.4	64.0
VideoMoCo [44]	V	K400	R(2+1)D-18	32×112^2	81.3G	47.8	32.1	7.0	31.7	31.8	53.9	72.8	65.1
RSPNet [13]	V	K400	R(2+1)D-18	16×112^2	40.6G	47.1	31.2	7.1	30.9	31.5	53.3	72.2	64.1
AoT [57] †	V	K400	TSM-Res50	8×224^2	33G	44.1	28.9	5.9	28.8	-	-	-	-
SpeedNet [5] †	V	K400	TSM-Res50	8×224^2	33G	44.5	29.5	6.1	29.4	-	-	-	-
PAL (Ours)	V	K400	I3D	8×112^2	3.6G	49.3	34.0	7.9	33.4	33.7	55.9	75.0	66.8
PAL (Ours)	V	K700	I3D	8×112^2	**3.6G**	**50.7**	**35.5**	**8.7**	**34.6**	**34.2**	**57.8**	**76.0**	**68.1**
TAC *	V	K400	I3D	8×112^2	3.6G	48.5	32.9	7.2	32.5	32.3	54.6	73.5	65.6
BSP [61]	V	K400	TSM-Res50	8×224^2	33G	50.9	35.6	8.0	34.8	33.7	57.4	75.5	67.6
LoFi-E2E [62]	V	K400+**ANet**	TSM-Res18	8×224^2	14.6G	50.4	35.4	8.9	34.4	-	-	-	-
TSP [1]	V	K400+**ANet**	R(2+1)D-34	16×112^2	76.4G	51.3	37.1	9.3	35.8	35.0	59.0	76.6	69.0

is 30 seconds and the maximum length of a text query is 10; (3) *THUMOS'14* [31] is a standard benchmark for TAD and APG tasks, containing 200 validation videos and 213 test videos of 20 action categories. The video length varies greatly, from less than a second to about 26 minutes. On average, each video contains ∼16 action instances.

Evaluation Metrics. We follow the standard evaluation protocol. For the TAD task, we report mean Average Precision (mAP) values under different temporal Intersection over Union (tIoU) thresholds. For the APG task, we report the Area Under the Curve (AUC) of the average recall *vs.* average number (AR-AN) of proposals per video. For the VG task, the top-1 recall at three tIoU thresholds and their mean value (mIoU) are reported.

Implementation Details. To validate the efficacy of our pre-training strategy, we retrain several state-of-the-art TAL methods by only replacing the original features with our PAL features. We choose those representative works with publicly-available codes. Specifically, we choose G-TAD [63] for TAD task, BMN [36] for APG task, and LGI [41] for VG task.

4.2. Main Results

In this section, we compare the performance of our PAL with other state-of-the-art pre-training approaches on three challenging TAL tasks. For those self-supervised methods designed for TAC tasks, we directly use their released pre-trained models to extract the video features for the downstream TAL task evaluations.

Temporal Action Detection (TAD) & Action Proposal Generation (APG). In Table 1, we report our TAD and APG results on ActivityNet v1.3 and compare them with state-of-the-art pre-training methods. When pre-trained on K400, our PAL consistently outperforms other self-supervised methods, which strongly demonstrates the effectiveness of our method. Although these self-supervised pre-training competitors have achieved promising results on TAC tasks, the task discrepancy issue still harms their transferability on TAL tasks, which verifies the necessity of our work. Compared to our baseline MoCo-v2 [16], which focuses on learning temporal invariant features, our proposed temporal equivariant learning scheme is more suitable for TAL, so it yields an improvement of +3.1% mAP@AVG and +2.8% AUC gains under the same settings. Notably, when using the same backbone (I3D) and pre-training dataset (K400), our unsupervised PAL even surpasses the supervised counterpart TAC by gains of +0.9% on mAP@AVG and +1.2% on AUC. It suggests that the proper use of data may benefit more than action label annotation information in TAL, which is consistent with our motivation. When pre-trained on a larger dataset K700, our PAL further improves the performance, showing its potential benefit of leveraging large-scale web videos. Compared to recent fully-supervised pre-training methods including BSP [61], LoFi-E2E [62] and TSP [1], our first attempt on unsupervised TAL pre-training achieves competitive results. Note that both LoFi-E2E [62] and TSP [1] use the downstream dataset ActivityNet (ANet) for feature pre-training which can lead to unfair comparison.

Video Grounding (VG). The VG results on Charades-STA are reported in Table 2. Note that the original LGI [41] exploits I3D features fine-tuned on the downstream Charades-STA dataset. For fair comparison, we retrain the LGI model using K400 pre-trained I3D features, without changing any hyper-parameters in the original codebase. Clearly, our PAL achieves the best VG performance under the unsupervised pre-training setting and even surpasses the supervised TAC trained features. Note that BSP [61] feature, which is pre-trained in a supervised

Table 2. **Comparison to state-of-the-art pre-training methods on the VG task.** We use LGI [41] as the evaluation method. The results are conducted on Charades-STA dataset. Rows highlighted in blue use fully-supervised pre-training. * Our implementation.

Method	**VG Task (LGI [41])**			
(K400 pre-trained)	R@0.3	R@0.5	R@0.7	mIoU
MoCo-v2 [16] *	54.2	40.9	21.1	38.7
VideoMoCo [44]	59.1	44.5	23.4	42.3
RSPNet [13]	55.8	41.5	21.4	39.6
PAL (Ours)	**63.7**	**50.0**	**27.2**	**46.8**
TAC *	61.6	46.8	24.6	44.3
BSP [61]	68.8	53.6	29.3	50.6

manner and has much more per-clip FLOPs (33G *vs.* 3.6G), performs better than ours as expected.

Overall, to our best knowledge, as the first unsupervised pre-training work customized for TAL, PAL consistently surpasses other unsupervised pre-training methods on three typical TAL tasks, demonstrating the efficacy of our idea.

4.3. Ablation Study

In this section, we conduct ablation experiments to fully understand the concept of our PAL. For ease of experimentation, all the ablation studies are conducted with 100 training epochs on K400 and evaluated on the TAD task.

Effectiveness of the key PAL components. In Table 3, we examine how each design in PAL affects the overall performance. We consider three key components in PAL: (1) dense sampling strategy to select multiple clips as region sample; (2) scale-aware sampling strategy to sample pseudo action regions with different strides; (3) paste operation to paste the selected regions onto the background videos. We start with the basic setting that does not involve any of the above designs, where only one clip is randomly sampled from each video and clip-level contrastive learning is performed. Then we introduce the dense sampling strategy to contrast region-level embeddings, which brings +0.5% improvement due to more temporal clues being included. Next, a scale-aware sampling strategy is applied along with dense sampling, leading to an overall +1.1% gain. This verifies that adding some randomness to the temporal scale facilitates representation learning. The biggest improvement is achieved after introducing paste operation. We infer that this is because attracting the region features of the same video but influenced by different backgrounds yields more background-insensitive features, which benefits localization tasks. Finally, pasting pseudo action regions onto different background videos further contributes a reasonable gain of +0.7% and the final improvement reaches +3.2% compared with baseline. In summary, by adding these key components step by step, the performance consistently boosts, verifying the effectiveness of our PAL.

Table 3. **Contribution of each design in PAL on TAL tasks.** Dense sampling strategy (dense), scale-aware sampling strategy (scale) and paste operation (paste) are involved step by step. All these designs contribute to the overall performance.

Exp.	**Setting**			**TAD Task**
	dense	scale	paste	mAP@AVG
#0 (*Baseline*)	✗	✗	✗	29.6
#1	✓	✗	✗	30.1 **(+0.5)**
#2	✓	✓	✗	30.7 **(+1.1)**
#3	✓	✓	same bkg.	32.1 **(+2.5)**
#4 (PAL)	✓	✓	diff bkg.	**32.8** **(+3.2)**

Table 4. **Ablation study on temporal embedding head.**

Num.	**Recep.**	**mAP@AVG**
0	1	30.8
1	3	32.1 **(+1.3)**
2	5	**32.8** **(+2.0)**
3	7	32.5 **(+1.7)**

Table 5. **Ablation study on paste ratios.**

β	**mAP@AVG**
1	32.2
$[0, 0.4]$	30.7 (-1.5)
$[0.4, 0.6]$	31.5 (-0.7)
$[0.6, 1.0]$	**32.8** **(+0.6)**
$[0, 1.0]$	31.3 (-0.9)

Number of the temporal embedding head layers. In Table 4, we experiment with the different numbers of the temporal embedding head layers. When the layer number is 0, the input clips are processed independently without temporal fusion, and therefore the surrounding background clips have no substantial effect on the action regions. It's obvious that attaching a single temporal convolution layer atop the backbone significantly boosts the performance (+1.3%). This verifies our hypothesis that introducing background semantics can help promote the localization power required by TAL. Since using two temporal embedding layers yields the best performance, we choose this setting as our default.

Hard paste *vs.* Soft paste. In PAL, β controls the paste ratio of pseudo action regions onto the background videos. We evaluate different β from 0 to 1. In particular, $\beta = 1$ means the "hard" paste and $\beta < 1$ is the "soft" paste. For the soft way, we test several representative intervals: $[0, 0.4]$, $[0.4, 0.6]$, $[0.6, 1.0]$ and $[0, 1.0]$, which indicates four cases, *i.e.*, background-dominated, half-and-half, action-dominated and purely random, respectively. As shown in Table 5, the $\beta \in [0.6, 1.0]$ setting outperforms hard paste and achieves the best result, partially because the action-dominated soft paste serves as an effective data augmentation strategy. So we use this setting by default.

Evaluation on THUMOS'14. THUMOS'14 is a relatively small-scale dataset compared with ActivityNet v1.3 (*cf.* Sec. 4.1.2). We list the experimental results in Table 6. Compared with the relative performance gain on ActivityNet v1.3, our improvement on THUMOS'14 is more prominent, which confirms the generalization ability of PAL under the small-scale data condition.

Table 6. **TAD results on small-scale THUMOS'14 dataset.**

Method	**TAD Task (G-TAD [41])**				
(K400, 100 epochs)	mAP@0.3	@0.4	@0.5	@0.6	@0.7
TAC *	44.6	37.3	29.5	18.8	9.5
MoCo-v2 [16] *	41.5	34.1	25.8	17.3	7.9
PAL (Ours)	**46.8**	**40.3**	**30.8**	**19.3**	**10.9**

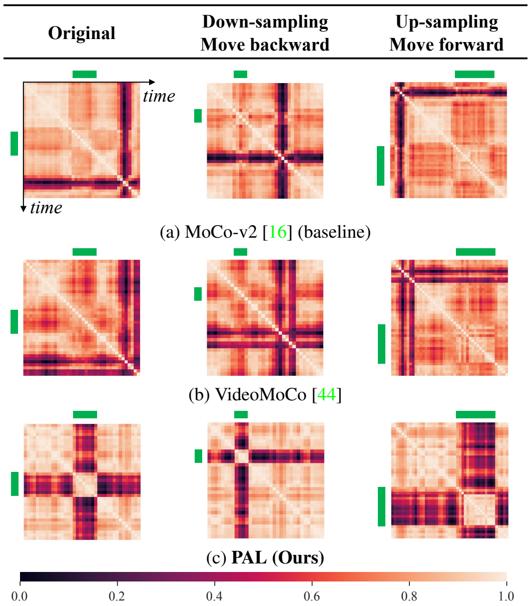
Table 7. **Comparison on TAC task.**

Method	**Backbone**	**TAC Task (Top-1 Acc.)**	
(K400, pre-trained)		**UCF101**	**HMDB51**
MoCo-v2 [16] *	3D-Res50	82.0	49.4
AoT [57]	T-CAM	79.4	-
SpeedNet [5]	S3D-G	81.1	48.8
VTHCL [65]	3D-Res50	82.1	49.2
VideoMoCo [44]	R(2+1)D-18	78.7	49.2
CoCLR [28]	S3D	**87.9**	**54.6**
PAL (Ours)	3D-Res50	84.7	52.5

Evaluation on TAC task. We investigate the transferring ability of our PAL on TAC downstream task. Following the common practice, all layers are fine-tuned end-to-end. The results are evaluated on UCF101 [46] and HMDB51 [33] datasets. Although our PAL feature is designed for TAL tasks, we observe in Table 7 that it still achieves competitive performance on TAC tasks. In detail, our PAL outperforms baseline MoCo-v2 by +2.7% & +3.1% at top-1 accuracy on UCF101 and HMDB51 respectively. Notably, it even exceeds the recently proposed VTHCL [65] with the same backbone.

4.4. Feature Visualization

Recall that PAL is proposed to guide the network in learning temporal translation and scale equivariance ability. To confirm this, we apply temporal transformations on the action instances in real-world videos and investigate whether these changes will be reflected in feature space accordingly. Specifically, given a video from ActivityNet v1.3, we first crop the action instance based on the temporal annotations, then re-sample the action instance with different temporal strides and insert them back into random temporal locations. Here, we consider two temporal transformations: (1) 2× *down-sampling* the action instance and moving *backward* along the time axis; and (2) 2× *up-sampling* the action instance and moving *forward* along the time axis. Next, MoCo-v2 [16] (baseline), VideoMoCo [44] and our PAL encoders are applied to extract features for the original video and the two transformed videos.



(a) MoCo-v2 [16] (baseline)

(b) VideoMoCo [44]

(c) **PAL (Ours)**

Figure 3. **Feature similarity visualization under different temporal transformations (2^{nd} & 3^{rd} columns) of ground-truth action instance.** The green bars represent the temporal extent of ground truth actions. Brighter color means higher similarity.

We visualize the cosine similarity between each clip features pair within the same video in Fig. 3. We also plot the ground-truth annotations (green bars) to indicate the action clips. As can be seen, MoCo-v2 and VideoMoCo learn time-invariant features that are insensitive to the temporal transformations. There is a high similarity between the pseudo action region and the background, while the salient area of PAL features changes accordingly. This confirms that our method successfully learns the time-equivariant characteristic, which is naturally more beneficial for TAL tasks. Besides, our introduced time-equivariant learning scheme can not only better separate the action and background clips, but also enable sharper contrast between action and its surrounding background clips. In this way, the clip features become more informative and boundary-aware, which facilitates localization. More visualization results can be found in our supplementary materials.

5. Discussion and Conclusion

This paper presents a new pretext task called Pseudo Action Localization (PAL), which is delicately designed to pre-train representations in an unsupervised manner for TAL tasks (UP-TAL). Motivated by the essential discrepancy between TAC and TAL, we also introduce a temporal scale and location equivariance learning scheme to facilitate better task alignment for the downstream transferring process. On a variety of downstream TAL tasks including temporal action detection, action proposal generation and video grounding, we demonstrate the effectiveness of our proposed method, which consistently surpasses its TAC counterpart and other unsupervised pre-training methods.

Acknowledgements. This paper was partially supported by the National Natural Science Foundation of China (NSFC) 62176008. Special acknowledgements are given to Aoto-PKUSZ Joint Lab for its support.

References

[1] Humam Alwassel, Silvio Giancola, and Bernard Ghanem. Tsp: Temporally-sensitive pretraining of video encoders for localization tasks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, 2021. 1, 3, 6

[2] Humam Alwassel, Dhruv Mahajan, Bruno Korbar, Lorenzo Torresani, Bernard Ghanem, and Du Tran. Self-supervised learning by cross-modal audio-video clustering. *Advances in Neural Information Processing Systems*, 33, 2020. 1, 2, 6

[3] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In *Proceedings of the IEEE international conference on computer vision*, pages 5803–5812, 2017. 3

[4] Yueran Bai, Yingying Wang, Yunhai Tong, Yang Yang, Qiyue Liu, and Junhui Liu. Boundary content graph neural network for temporal action proposal generation. In *European Conference on Computer Vision*, pages 121–137. Springer, 2020. 2

[5] Sagie Benaim, Ariel Ephrat, Oran Lang, Inbar Mosseri, William T Freeman, Michael Rubinstein, Michal Irani, and Tali Dekel. Speednet: Learning the speediness in videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9922–9931, 2020. 6, 8

[6] Shyamal Buch, Victor Escorcia, Chuanqi Shen, Bernard Ghanem, and Juan Carlos Niebles. Sst: Single-stream temporal action proposals. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 2911–2920, 2017. 2

[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. 2, 5

[8] Meng Cao, Long Chen, Mike Zheng Shou, Can Zhang, and Yuexian Zou. On pursuit of designing multi-modal transformer for video grounding. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9810–9823, 2021. 3

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

[10] 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, 2, 4, 5

[11] Yu-Wei Chao, Sudheendra Vijayanarasimhan, Bryan Seybold, David A Ross, Jia Deng, and Rahul Sukthankar. Rethinking the faster r-cnn architecture for temporal action localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1130–1139, 2018. 2

[12] Guang Chen, Can Zhang, and Yuexian Zou. Afnet: Temporal locality-aware network with dual structure for accurate and fast action detection. *IEEE Transactions on Multimedia*, 23:2672–2682, 2020. 2

[13] 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 Conference on Artificial Intelligence*, volume 1, 2021. 6, 7

[14] Shaoxiang Chen and Yu-Gang Jiang. Hierarchical visual-textual graph for temporal activity localization via language. In *European Conference on Computer Vision*, pages 601–618. Springer, 2020. 3

[15] 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

[16] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 1, 2, 5, 6, 7, 8

[17] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15750–15758, 2021. 2

[18] Nikita Dvornik, Julien Mairal, and Cordelia Schmid. Modeling visual context is key to augmenting object detection datasets. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 364–380, 2018. 3

[19] Debidatta Dwibedi, Ishan Misra, and Martial Hebert. Cut, paste and learn: Surprisingly easy synthesis for instance detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1301–1310, 2017. 3

[20] Hao-Shu Fang, Jianhua Sun, Runzhong Wang, Minghao Gou, Yong-Lu Li, and Cewu Lu. Instaboost: Boosting instance segmentation via probability map guided copy-pasting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 682–691, 2019. 3

[21] 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

[22] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. Tall: Temporal activity localization via language query. In *Proceedings of the IEEE international conference on computer vision*, pages 5267–5275, 2017. 2, 5

[23] Golnaz Ghiasi, Yin Cui, Aravind Srinivas, Rui Qian, Tsung-Yi Lin, Ekin D. Cubuk, Quoc V. Le, and Barret Zoph. Simple copy-paste is a strong data augmentation method for instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2918–2928, June 2021. 3

[24] Soham Ghosh, Anuva Agarwal, Zarana Parekh, and Alexander Hauptmann. ExCL: Extractive Clip Localization Using Natural Language Descriptions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1984–1990, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. 3

[25] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent: A new approach to self-supervised learning. *arXiv preprint arXiv:2006.07733*, 2020. 1, 2

[26] Tengda Han, Weidi Xie, and Andrew Zisserman. Video representation learning by dense predictive coding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019. 2

[27] Tengda Han, Weidi Xie, and Andrew Zisserman. Memory-augmented dense predictive coding for video representation learning. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*, pages 312–329. Springer, 2020. 2

[28] Tengda Han, Weidi Xie, and Andrew Zisserman. Self-supervised co-training for video representation learning. *Advances in Neural Information Processing Systems*, 33:5679–5690, 2020. 1, 2, 4, 6, 8

[29] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020. 1, 2

[30] Fabian Caba Heilbron, Juan Carlos Niebles, and Bernard Ghanem. Fast temporal activity proposals for efficient detection of human actions in untrimmed videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1914–1923, 2016. 2

[31] Haroon Idrees, Amir R Zamir, Yu-Gang Jiang, Alex Gorban, Ivan Laptev, Rahul Sukthankar, and Mubarak Shah. The thumos challenge on action recognition for videos "in the wild". *Computer Vision and Image Understanding*, 155:1–23, 2017. 2, 6

[32] Simon Jenni and Hailin Jin. Time-equivariant contrastive video representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9970–9980, October 2021. 1, 2

[33] 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. 8

[34] Chun-Liang Li, Kihyuk Sohn, Jinsung Yoon, and Tomas Pfister. Cutpaste: Self-supervised learning for anomaly detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9664–9674, 2021. 3

[35] Ji Lin, Chuang Gan, and Song Han. Tsm: Temporal shift module for efficient video understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7083–7093, 2019. 2

[36] Tianwei Lin, Xiao Liu, Xin Li, Errui Ding, and Shilei Wen. Bmn: Boundary-matching network for temporal action proposal generation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3889–3898, 2019. 2, 3, 6

[37] Tianwei Lin, Xu Zhao, and Zheng Shou. Temporal convolution based action proposal: Submission to activitynet 2017. *arXiv preprint arXiv:1707.06750*, 2017. 2

[38] Tianwei Lin, Xu Zhao, Haisheng Su, Chongjing Wang, and Ming Yang. Bsn: Boundary sensitive network for temporal action proposal generation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 3–19, 2018. 2

[39] Meng Liu, Xiang Wang, Liqiang Nie, Qi Tian, Baoquan Chen, and Tat-Seng Chua. Cross-modal moment localization in videos. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 843–851, 2018. 3

[40] Fuchen Long, Ting Yao, Zhaofan Qiu, Xinmei Tian, Jiebo Luo, and Tao Mei. Gaussian temporal awareness networks for action localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 344–353, 2019. 2

[41] Jonghwan Mun, Minsu Cho, and Bohyung Han. Local-global video-text interactions for temporal grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10810–10819, 2020. 3, 6, 7, 8

[42] Phuc Nguyen, Ting Liu, Gautam Prasad, and Bohyung Han. Weakly supervised action localization by sparse temporal pooling network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6752–6761, 2018. 3

[43] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 1, 2, 4, 5

[44] Tian Pan, Yibing Song, Tianyu Yang, Wenhao Jiang, and Wei Liu. Videomoco: Contrastive video representation learning with temporally adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11205–11214, 2021. 1, 2, 4, 6, 7, 8

[45] Rui Qian, Tianjian Meng, Boqing Gong, Ming-Hsuan Yang, Huisheng Wang, Serge Belongie, and Yin Cui. Spatiotemporal contrastive video representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6964–6974, 2021. 1, 2, 4

[46] 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. 8

[47] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. Videobert: A joint model for video and language representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7464–7473, 2019. 2

[48] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015. 1

[49] Zhigang Tu, Hongyan Li, Dejun Zhang, Justin Dauwels, Baoxin Li, and Junsong Yuan. Action-stage emphasized spatiotemporal vlad for video action recognition. *IEEE Transactions on Image Processing*, 28(6):2799–2812, 2019. 2

[50] Zhigang Tu, Wei Xie, Justin Dauwels, Baoxin Li, and Junsong Yuan. Semantic cues enhanced multimodality multistream cnn for action recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 29(5):1423–1437, 2018. 2

[51] Wouter Van Gansbeke, Simon Vandenhende, Stamatios Georgoulis, and Luc Van Gool. Unsupervised semantic segmentation by contrasting object mask proposals. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10052–10062, October 2021. 1

[52] Jinpeng Wang, Yuting Gao, Ke Li, Yiqi Lin, Andy J. Ma, Hao Cheng, Pai Peng, Feiyue Huang, Rongrong Ji, and Xing Sun. Removing the background by adding the background: Towards background robust self-supervised video representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11804–11813, June 2021. 3

[53] 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, 2

[54] Limin Wang, Yuanjun Xiong, Dahua Lin, and Luc Van Gool. Untrimmednets for weakly supervised action recognition and detection. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 4325–4334, 2017. 3

[55] 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. 1, 2

[56] Xinlong Wang, Rufeng Zhang, Chunhua Shen, Tao Kong, and Lei Li. Dense contrastive learning for self-supervised visual pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3024–3033, 2021. 1

[57] Donglai Wei, Joseph J Lim, Andrew Zisserman, and William T Freeman. Learning and using the arrow of time. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8052–8060, 2018. 6, 8

[58] 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

[59] Enze Xie, Jian Ding, Wenhai Wang, Xiaohang Zhan, Hang Xu, Peize Sun, Zhenguo Li, and Ping Luo. Detco: Unsupervised contrastive learning for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8392–8401, 2021. 1

[60] Huijuan Xu, Abir Das, and Kate Saenko. R-c3d: Region convolutional 3d network for temporal activity detection. In *Proceedings of the IEEE international conference on computer vision*, pages 5783–5792, 2017. 2

[61] Mengmeng Xu, Juan-Manuel Perez-Rua, Victor Escorcia, Brais Martinez, Xiatian Zhu, Li Zhang, Bernard Ghanem, and Tao Xiang. Boundary-sensitive pre-training for temporal localization in videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7220–7230, October 2021. 1, 3, 6, 7

[62] Mengmeng Xu, Juan-Manuel Perez-Rua, Xiatian Zhu, Bernard Ghanem, and Brais Martinez. Low-fidelity end-to-end video encoder pre-training for temporal action localization. *arXiv preprint arXiv:2103.15233*, 2021. 1, 3, 6

[63] Mengmeng Xu, Chen Zhao, David S Rojas, Ali Thabet, and Bernard Ghanem. G-tad: Sub-graph localization for temporal action detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10156–10165, 2020. 2, 3, 6

[64] Ceyuan Yang, Zhirong Wu, Bolei Zhou, and Stephen Lin. Instance localization for self-supervised detection pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3987–3996, 2021. 1, 3

[65] Ceyuan Yang, Yinghao Xu, Bo Dai, and Bolei Zhou. Video representation learning with visual tempo consistency. *arXiv preprint arXiv:2006.15489*, 2020. 2, 8

[66] Yitian Yuan, Tao Mei, and Wenwu Zhu. To find where you talk: Temporal sentence localization in video with attention based location regression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 9159–9166, 2019. 3

[67] Runhao Zeng, Chuang Gan, Peihao Chen, Wenbing Huang, Qingyao Wu, and Mingkui Tan. Breaking winner-takes-all: Iterative-winners-out networks for weakly supervised temporal action localization. *IEEE Transactions on Image Processing*, 28(12):5797–5808, 2019. 3

[68] Can Zhang, Meng Cao, Dongming Yang, Jie Chen, and Yuexian Zou. Cola: Weakly-supervised temporal action localization with snippet contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16010–16019, 2021. 3

[69] Can Zhang, Yuexian Zou, Guang Chen, and Lei Gan. Pan: Persistent appearance network with an efficient motion cue for fast action recognition. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 500–509, 2019. 2

[70] Hao Zhang, Aixin Sun, Wei Jing, and Joey Tianyi Zhou. Span-based localizing network for natural language video localization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6543–6554, Online, July 2020. Association for Computational Linguistics. 3

[71] Nanxuan Zhao, Zhirong Wu, Rynson WH Lau, and Stephen Lin. Distilling localization for self-supervised representation learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 10990–10998, 2021. 3

[72] Jia-Xing Zhong, Nannan Li, Weijie Kong, Tao Zhang, Thomas H Li, and Ge Li. Step-by-step erasion, one-by-one collection: a weakly supervised temporal action detector. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 35–44, 2018. 3