# Two-Stage Active Learning for Efficient Temporal Action Segmentation

Yuhao Su and Ehsan Elhamifar

Khoury College of Computer Sciences, Northeastern University, Boston, USA
{su.yuh,e.elhamifar}@northeastern.edu

**Abstract.** Training a temporal action segmentation (TAS) model on long and untrimmed videos requires gathering framewise video annotations, which is very costly. We propose a two-stage active learning framework to efficiently learn a TAS model using only a small amount of video annotations. Our framework consists of three components that work together in each active learning iteration. 1) Using current labeled frames, we learn a TAS model and action prototypes using a novel contrastive learning method. Leveraging prototypes not only enhances the model performance, but also increases the computational efficiency of both video and frame selection for labeling, which are the next components of our framework. 2) Using the currently learned TAS model and action prototypes, we select informative unlabeled videos for annotation. To do so, we find unlabeled videos that have low alignment scores to learned action prototype sequences in labeled videos. 3) To annotate a small subset of informative frames in each selected unlabeled video, we propose a video-aligned summary selection method and an efficient greedy search algorithm. By evaluation on four benchmark datasets (50Salads, GTEA, Breakfast, CrossTask), we show that our method significantly reduces the annotation costs, while consistently surpassing baselines over active learning iterations. Our method achieves comparable or better performance than other weakly supervised methods while using a small amount of labeled frames. We further extend our framework to a semi-supervised active learning setting. To the best of our knowledge, this is the first work studying active learning for TAS.

**Keywords:** Action segmentation · Video alignment · Active learning

## 1 Introduction

Automatic understanding of human activities in long and uncurated videos is crucial for many applications, such as healthcare, robotics, security and assistive technologies [22, 60, 100]. This has motivated many recent works on temporal action segmentation (TAS) [17,33,46,47,65,75,100,103], whose goal is to predict a label for each frame, hence partition a long video into non-overlapping segments.

Learning a TAS model, in principle, requires dense framewise annotations for many videos [17]. This is extremely costly given that untrimmed videos are often long, hence, annotators need to watch hours or days of video footage and

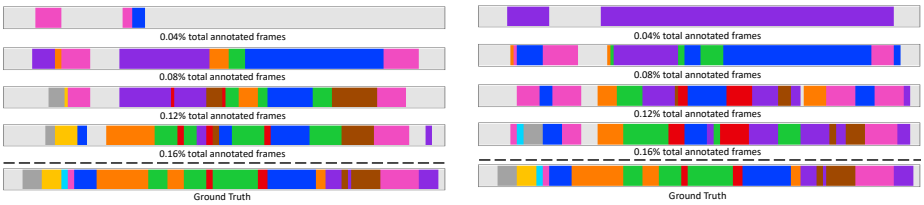


**Fig. 1:** Visualization of ours (left) and split-random baseline (right) model performance improvement by iteratively labeling more frames. This example corresponds to video 'rgb-04-1' in 50Salads. Note split-random baseline missed at least one large segment (blue one in middle right).

specify boundaries of different actions in every video. To reduce the annotation cost for learning, some of the existing works on TAS have focused on semi-supervised learning by using a small number of fully-annotated videos and many unannotated videos [18, 104, 117]. Yet, they still require fully labeling multiple long videos and it is not clear how to select the best videos for annotation and training. On the other hand, weakly-supervised approaches use weak annotations (e.g., ordered or unordered list of actions) in all training videos [13, 14, 21, 44, 55, 63, 73, 74, 89, 90, 99, 106, 123]. However, the performance of both categories of methods is significantly lower than the fully-supervised techniques. While time-stamp annotation [8, 50, 67, 69, 85] can achieve performance close to the fully-supervised setting. However, while a small number of annotated frames will be provided for training, they have been carefully selected in a very costly process of annotators watching every entire video and selecting frames that correspond to distinct actions in each video. In fact, a key limitation of such methods is the lack of effective video and frame selection for labeling to significantly reduce the annotation time and guide the labeling process efficiently.

Another approach to reduce the annotation cost is the conventional active learning (AL) setting, which iteratively selects unlabeled samples for assigning labels based on some utility functions [88]. However, almost all existing AL works have focused on sample classification, where each data sample has a single label. This is different from TAS, where each video sample has multiple action labels and frame labels depend on each other. Therefore, we need to address both video selection and frame selection for labeling. Recently, [86, 87] proposed a hybrid active learning approach for action detection, which is to specify spatiotemporal regions of an action in a video. Despite its success, their models only handle trimmed video clips, where each clip contains only one action and lasts a few seconds. On the other hand, in TAS, videos are long and have multiple actions, thus leveraging action temporal dependencies is crucial. Also videos often contain background (irrelevant and uninteresting actions not present in the pre-specified dictionary of actions), which makes learning challenging.

**Paper Contributions.** We develop a two-stage active learning framework to learn a temporal action segmentation (TAS) model in long and untrimmed videos using small annotations and with relatively small performance drop compared

to the fully-supervised TAS. To the best of our knowledge *this is the first work addressing active learning for TAS*. In our two-stage framework, we first select informative unlabeled videos (inter-video selection) followed by selecting and annotating informative frames within them (intra-video selection). For inter-video selection, we develop a method based on dynamic time warping [27] to select unlabeled videos with diverse and distinct action orderings with respect to the action orderings in the labeled videos. This leads to increased diversity among labeled videos and subsequently to more efficient TAS learning compared to a random video selection strategy. On the other hand, instead of labeling the entire selected videos, we propose a new alignment-based video summarization method to select and annotate a sequence of a small number of frames from each selected video that aligns well with the video, hence, can capture distinct actions. We repeat this iterative annotation process until the total budget or target performance met. This strategy not only improves the AL performance compared to other frame selection criteria (see Fig. 1), but also significantly reduces annotation effort. Our method also consists of action prototypes that will be learned in conjunction with TAS using a novel regularized contrastive action prototype loss. Action prototypes allow us to perform alignment more efficiently to perform video selection. By extensive evaluation on four benchmark datasets (GTEA [34], Breakfast [54], CrossTask [135], 50Salads [107]), we show that our framework significantly boosts performance over baselines. Moreover, we expand our framework to include a semi-supervised AL setting, and we achieve comparable or better performance than other weakly supervised methods with only a minimal number of labeled frames.

## 2 Related Works

**Temporal Action Segmentation.** Fully-supervised TAS methods have used dense framewise annotations for training, including recurrent networks [20, 46, 89, 102], temporal convolution nets [24, 33, 40, 43, 58, 59, 61, 65, 66, 84, 103, 105, 113, 114, 130], graph neural networks [46, 128], transformers [4, 6, 23, 25, 70, 94, 109, 110, 115, 125, 134], combination of transformers and CNNs [75] and diffusion models [68].

Multi-stage models [4, 33, 65, 75, 75, 84, 100, 102, 113, 119, 125] have been particularly effective because of their ability to capture and refine temporal context and reduce oversegmentation [17]. Notable examples include MS-TCN [33] and its variants, employing temporal convolution networks, and ASFormer [125], which use Transformers [111]. Semi-supervised approaches provide framewise labels for a subset of videos [18, 104, 117], while weakly-supervised methods rely weak annotation, examples are action set or transcript methods [13, 14, 21, 44, 55, 63, 73, 74, 89, 90, 99, 106, 123], single-frame labels [8, 50, 67, 69, 85], activity labels [18, 19], and from other sources (e.g., narrations, subtitles) [37, 97].

To further reduce reliance on annotation, unsupervised models [1, 5, 26, 29, 31, 41, 56, 93, 95, 95, 98, 101, 112, 118] often use videos from same activity at a time or use self-supervision from large-scale datasets (e.g. Howto100M [81]), treating

TAS as a downstream task. We argue that reducing annotation effort for TAS is important, however, instead of completely removing it, we propose an active learning framework to efficiently label only a subset of frames for a small number of videos to train TAS.

**Sequence Alignment.** Dynamic Time Warping (DTW) [82,92] and its variants, such as dropDTW [27], assume ordered matching between sequences and have achieved success in various applications [10, 13, 14, 16, 16, 27–29, 42, 101, 101, 124, 132]. To handle alignment of sequences that do not follow ordered matchings, other methods have been proposed, such as (restricted) edit distance [38, 51, 62, 99], order-preserving Wasserstein distance [108], temporal cycle consistency [30]. Our video-aligned summary selection method for intra-video selection is an extension of DTW for simultaneous summary selection and alignment.

**Active Learning.** The goal of active learning is to minimize the amount of data annotation by iteratively selecting samples, labeling them and training a model [88]. While query-based [3, 77, 79, 133] and stream-based [32, 53, 83] methods have been proposed, the current mainstream favors pool-based frameworks. These pool-based algorithms select a subset of samples from unlabeled data for annotation, mainly guided by two key criteria: uncertainty [7, 36, 39, 45, 49, 71, 71, 72, 76, 116, 126, 127, 131] and diversity [2, 78, 96]. Uncertainty-based methods select difficult samples about whose label the current model is most uncertain. Diversity-based methods ensure the chosen samples for labeling capture the distribution of the original data. Recent efforts in active learning have been expanded into video domain [48, 52, 64, 86, 87, 120, 136] and different settings, e.g., active fine-tuning [122] active self-supervised learning [9], active training [57] and active federated learning [11].
In the paper, we focus on untrimmed videos and explore two-stage active learning for action segmentation by capturing temporal diversity at various levels.

## 3 Active Temporal Action Segmentation Learning

**Problem Setting.** Assume we have a set of $N$ unlabeled videos with $A$ action classes, including background (i.e., irrelevant actions outside the pre-defined list of main action classes). For more efficient processing, we divide each video into clips, where each clip consists of a small number (e.g., 10 or 32) of consecutive frames. Each video $i$ has $T_i$ clips and consists of pre-extracted clip features, $\boldsymbol{X}_i = \left[\boldsymbol{x}_{i1}, \boldsymbol{x}_{i2}, \ldots, \boldsymbol{x}_{iT_i}\right]$, where $\boldsymbol{x}_{ij}$ denotes the feature of clip $j$ in video $i$. We denote the set of all videos by $\mathcal{X} = \{\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_N\}$.

Our two-stage active learning framework for TAS consists of $c$ iterations, where in each iteration, we select $m \ll N$ unlabeled videos and partially annotate $\rho \ll 1$ fraction of their clips. We denote the set of labeled video clips in iteration $t$ by $\mathcal{X}_l^{(t)}$ and the set of remaining unlabeled video clips by $\mathcal{X}_u^{(t)}$. It is important to note that our framework can work with any existing TAS architecture.

**Overview of Proposed Framework.** Our two-stage active TAS learning consists of the following components (see Fig. 2):

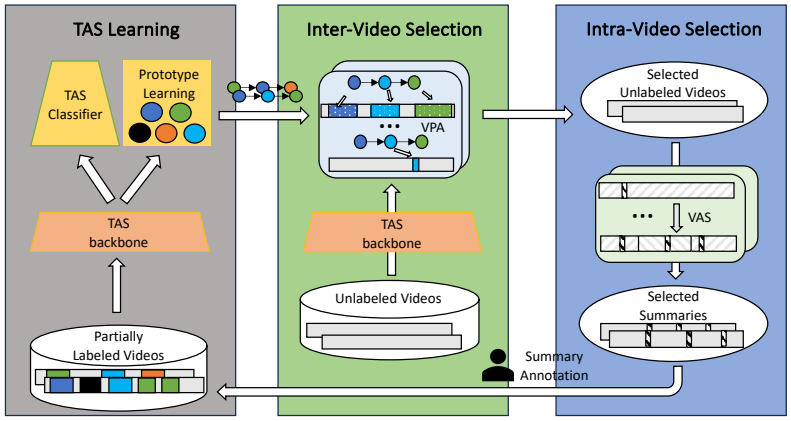


**Fig. 2:** Overview of our two-stage AL framework. Left: we first train a TAS model and construct action prototype sequences from partially labeled videos. Middle: Informative unlabeled videos are identified through our inter-video selection method (VPA). Right: We use intra-video selection method (VAS) to select summaries from selected unlabeled videos for annotation. We label selected clips by labeling middle frame of each clip. New partially labeled videos are added to the labeled pool for the next iteration.

— Using labeled clips $\mathcal{X}_l^{(t)}$ at iteration $t$, we train a TAS network and learn action prototypes. More specifically, for each action $a$ in labeled video clips, we learn a prototype $\boldsymbol{p}_a$. We use the learned TAS model and prototypes to construct a sequence of action prototypes for each partially labeled video. We use the prototype sequences to efficiently select informative unlabeled videos for annotation via our inter-video selection module. *The use of prototype sequences instead of entire videos significantly reduces the complexity of video selection.*

— Given the trained TAS model at iteration $t$, we select an informative subset of unlabeled videos from $\mathcal{X}_u^{(t)}$ for partial annotation. To do so, we use a method based on video-prototype sequence alignment (VPA) to select the most useful subset of unlabeled videos. *Our key idea is that an unlabeled video whose sequence of actions is sufficiently different from the sequences of actions in partially labeled videos, is expected to improve the TAS model the most*, once annotated and used for training. Therefore, we use the prototypes learned by our method to construct action prototype sequences for labeled videos and measure the alignment cost between each unlabeled video and each training prototype sequence. We will select the unlabeled videos with the highest alignment cost for annotation.

— For each selected unlabeled video, we find an informative subset of clips to annotate and to build the labeled clips for the next iteration $\mathcal{X}_l^{(t+1)}$. To do so, *we develop an alignment-based video summarization framework, which finds the summary of a video that best sequentially aligns with the video.* We refer to our method as video-aligned summarization (VAS). This is different from existing summarization methods [91, 129] that do not take into account the sequential nature of the summary with respect to the video. Given the NP-hardness of

the problem, we propose an efficient greedy algorithm to find the subset of informative clips. We use the label of the middle frame in each selected clip as the clip label and obtain $\mathcal{X}_l^{(t+1)}$ by adding the new annotated clips to $\mathcal{X}_l^{(t)}$. We next discuss each component in more details.

### 3.1 Learning Action Prototypes and TAS Model

Using annotated video clips at iteration $t$ of active learning, we jointly learn the parameters of a TAS model and action prototypes $\{\boldsymbol{p}_a\}$, representing all seen action classes. These prototypes not only facilitate TAS learning, but also allow constructing compact sequences, which we will use later for efficiently selecting unlabeled videos for annotation.

**Regularized Contrastive Learning of Action Prototypes.** Our action prototypes are learnable parameters that lie in the embedding space of TAS, each representing an action. Prototypes are learned with features $\boldsymbol{z}$ obtained from the model before the last layer. To learn them, we propose a loss, which we refer to as regularized contrastive action prototype loss, $\mathcal{L}_{\text{reg-cont}}$, which consists of three terms,

$$\mathcal{L}_{\text{reg-cont}} = \mathcal{L}_{\text{cont}} + \mathcal{L}_{\mu} + \mathcal{L}_{\sigma}. \tag{1}$$

Here, $\mathcal{L}_{\text{cont}}$ is a contrastive loss [15] which enforces that each prototype $\boldsymbol{p}_a$ must be close to labeled clips belonging to action $a$ (positive clips, $\mathcal{P}_a$) and far from labeled clips of other actions (negative clips, $\mathcal{N}_a$),

$$\mathcal{L}_{\text{cont}} = \sum_{a} \sum_{j \in \mathcal{P}_a} -\log\Big(\frac{f(\boldsymbol{p}_a, \boldsymbol{z}_j)}{f(\boldsymbol{p}_a, \boldsymbol{z}_j) + \sum_{j' \in \mathcal{N}_a} f(\boldsymbol{p}_a, \boldsymbol{z}_{j'})}\Big). \tag{2}$$

In the above, $f$ is the exponential cosine similarity function with a temperature parameter, $f(x, y) = \exp(\cos(x, y)/\gamma)$. The contrastive loss enforces that features from the same action to be mapped closely to the associated prototype hence to each other, while having sufficient separation to features of different actions. This leads to obtaining more discriminative action features. Beyond $\mathcal{L}_{\text{cont}}$, since active learning adds labeled clips incrementally, the TAS model may encounter different prototypes at various active learning iterations. This can lead to overfitting for some prototypes while underfitting for others. Therefore, we enforce that different prototypes maintain approximately equal levels of separation throughout the learning process. This leads to the following regularization losses

$$\mathcal{L}_{\mu} = \frac{1}{M} \sum_{a} \sum_{a' \neq a} \cos\left(\boldsymbol{p}_a, \boldsymbol{p}_{a'}\right), \;\; \mathcal{L}_{\sigma} = \frac{1}{M} \sum_{a} \sum_{a' \neq a} \big(\cos(\boldsymbol{p}_a, \boldsymbol{p}_{a'}) - \mathcal{L}_{\mu}\big)^2, \tag{3}$$

where $M = A(A-1)$ is the total number of prototype pairs. Experimentally, we observed that using these two regularization losses stabilizes the learning of prototypes and improves the performance. **TAS Learning.** To learn the

parameters of the TAS model, we use the standard cross-entropy and smoothing losses [33, 125] and our proposed regularized contrastive action prototype loss,

$$\mathcal{L} = \mathcal{L}_{\text{cross-entropy}} + \alpha \mathcal{L}_{\text{smooth}} + \beta \mathcal{L}_{\text{reg-cont}}, \tag{4}$$

where we set $\alpha = 0.15$ and $\beta = 0.1$ in our experiments.

### 3.2 Inter-Video Selection: Selecting from Unlabeled Videos

Given the learned action prototypes and TAS model using labeled clips $\mathcal{X}_l^{(t)}$ at iteration $t$, we then select a small number of most informative unlabeled videos from $\mathcal{X}_u^{(t)}$ for partial annotation. Our key idea is that the most difficult unlabeled videos for the current TAS model are the ones that have a different sequence of actions than the seen action sequences in the partially labeled videos. This comes from the fact that TAS learns to capture long-range action dependencies from training videos, therefore, unlabeled videos with different action orderings pose a challenge for the learned TAS model.

However, we do not have ground-truth sequence of actions in unlabeled videos. To address this challenge, we take each partially labeled video $i$, obtain its sequence of actions $(i_1, i_2, \ldots)$ from clip labels (see Fig. 2) and represent the video with the sequence of prototypes $\boldsymbol{P}_i = (\boldsymbol{p}_{i_1}, \boldsymbol{p}_{i_2}, \ldots, )$. We treat consecutive repetitions of the same action as one action and skip the background class/prototype. Our goal is to align each unlabeled video $\boldsymbol{X}_j$ with each $\boldsymbol{P}_i$ and select videos with the lowest alignment scores, since they represent different sequences of actions. To do so, we utilize drop-DTW [27], which aligns two sequences by allowing many to many matchings while allowing some elements from each sequence to stay unmatched (hence be dropped). Specifically, we adopt drop-DTW to the one-to-many matching, since several consecutive clips in $\boldsymbol{X}_j$ often correspond to the same action, hence must be matched with the same prototype (however, one clip cannot belong to more than one prototypes/actions) while background clips in $\boldsymbol{X}_j$ should be dropped.

For an unlabeled video $\boldsymbol{X}_j$, we measure its smallest drop-DTW cost to prototype sequences of partially labeled videos,

$$s_j = \min_i \text{drop-DTW}(\boldsymbol{P}_i, \boldsymbol{X}_j). \tag{5}$$

Once we compute $s_j$ for all unlabeled videos, we select $m$ videos with the highest alignment cost, denoted by $\Lambda^{(t)}$, since they represent the videos whose sequence of actions will be most dissimilar to all seen action sequences in labeled videos. We refer to our inter-video selection method as Video-Prototype sequence Alignment (VPA). Notice that using $\boldsymbol{P}_i$ instead of $\boldsymbol{X}_i$ significantly *reduces the alignment computational cost* since the number of action segments in a video is often much smaller than the video length, i.e., $|\boldsymbol{P}_i| \ll |\boldsymbol{X}_i|$. To compute the drop-DTW in (5), following [42], we measure the distance between each clip of $\boldsymbol{X}_j$ and each prototype in $\boldsymbol{P}_i$ via $-\log(\text{softmax}_1(\boldsymbol{P}_i^T \boldsymbol{X}_j / \gamma))$[1].

[1] Applying softmax operator over the first tensor dimension.

**Algorithm 1:** Video-Aligned summarization (VAS)

**Input:** Video clips $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n]$, subset size K

Initialize subset $\boldsymbol{S}^{(0)} = \varnothing$

**for** $j = 0, \ldots, K-1$ **do**

  $\boldsymbol{x}_{j+1} = \text{argmin}_{\boldsymbol{x} \in \boldsymbol{X}} \text{SVA}(\boldsymbol{X}, \boldsymbol{x} \cup \boldsymbol{S}^{(j)})$ (via Alg 2)

  $\boldsymbol{S}^{(j+1)} = \text{sort}(\boldsymbol{x}_{j+1} \cup \boldsymbol{S}^{(j)})$

$\boldsymbol{S} \leftarrow \boldsymbol{S}^{(K)}$

**Output:** Summary $\boldsymbol{S}$

**Algorithm 2:** Summary-Video Alignment (SVA)

**Input:** Video clips $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n]$, Summary $\boldsymbol{S} = [\boldsymbol{s}_1, \ldots, \boldsymbol{s}_K]$

Initialize cumulative cost matrix $\boldsymbol{M} \in \mathbb{R}^{(K+1)\times(n+1)}$ with $M[0,:] = \infty, M[:,0] = \infty, M[0,0] = 0$

Compute pairwise cosine dissimilarity cost matrix $\boldsymbol{D}$ between $\boldsymbol{X}$ and $\boldsymbol{S}$

**for** $i = 0, \ldots, K$ **do**

  **for** $j = 0, \ldots, n$ **do**

    $M[i+1, j+1] = D[i,j] + \min(M[i+1, j], M[i,j])$

Set alignment cost as $\text{SVA}(\boldsymbol{X}, \boldsymbol{S}) = M[K, n]$

**Result:** M, $\text{SVA}(\boldsymbol{X}, \boldsymbol{S})$

### 3.3 Intra-Video Selection: Selecting Clips for Annotation

The next step of our framework involves selecting a small subset of informative clips in each video in $\Lambda^{(t)}$ for annotation. To achieve this, we cast the problem as finding a summary $\boldsymbol{S}_i$ of each video $i \in \Lambda^{(t)}$, that best sequentially aligns with the video. In other words, if an entry in the summary $\boldsymbol{S}_i$ is matched with with clip $j$ in the video $\boldsymbol{X}_i$, the next entry in the summary must match with a clip after $j$. This ensures that the summary preserves the sequential structure of the video, hence captures its action ordering. Therefore, we propose to solve

$$\min_{\boldsymbol{S}_i \subseteq \boldsymbol{X}_i} \text{DTW}(\boldsymbol{S}_i, \boldsymbol{X}_i), \quad \text{s.t.} \quad |\boldsymbol{S}_i| \leq \rho |\boldsymbol{X}_i|, \tag{6}$$

which searches for a summary whose size is $\rho$ fraction of the video and has the minimum dynamic time warping (DTW) cost [82, 92] with the video. Here, we modify DTW to use a one-to-many matching (instead of many to many) so that each clip in $\boldsymbol{X}_i$ can be assigned to at most one representative in $\boldsymbol{S}_i$, while allowing each representative clip to encode multiple temporally close clips that have the same action label.

We expect that such a video-aligned summarization (VAS) method will select clips that correspond to different actions in the video. Notice that it is possible that the summary may not capture all actions in the video, hence annotations of the videos at a given active learning iteration may not contain clips from all actions. Thus, prototypes will only be learned for actions that have labeled video clips. However, given that inter-video selection will find videos with different

sequences of actions, the next iteration with a high likelihood will find and annotate videos with missing actions.

Notice that solving (6) is challenging and computationally complex due to the combinatorial search over all possible subsets of clips from a videos. Motivated by research on submodular function maximization [121], we propose a greedy algorithm which iteratively builds the subset $\boldsymbol{S}_i$. More specifically, in the first iteration of the greedy approach, we select a clip from $\boldsymbol{X}_i$ that has the lowest total cost for representing the entire video. In the next iteration, we choose another clip from $\boldsymbol{X}_i$ such that the two clips (when sorted by their indices) provide the best DTW alignment with the entire video and so on. Algorithm 1 shows the steps of our proposed greedy algorithm. Notice that to perform greedy selection, we need to align the video and the current summary set at each iteration of the greedy algorithm, which we perform by running one-to-many dynamic time warping using Algorithm 2. More details about complexity of our method are provided in the supplementary materials. Finally, we use the label of the middle frame of each selected clip as the clip label.

## 4 Experiments

### 4.1 Datasets

We use the following four datasets for evaluations. **1) 50Salads** [107] consists of 50 videos, with 17 actions that depict individuals making salads. On average, each video is roughly 6.4 minutes long. **2) GTEA** [35] contains 28 videos with 11 kitchen-related action classes performed by 4 subjects. Each video lasts around 1.5 minutes. **3) Breakfast** [54] comprises a total of 1,712 videos, depicting activities associated with breakfast preparation. There are a total of 48 distinct actions, with an average of 6 action instances in each video. **4) CrossTask** [135] consists of 2,750 YouTube videos, spanning 18 primary tasks and totaling 212 hours of video content. The duration of a video is around 6 minutes.

We have chosen *a wide range of datasets* to show the effectiveness of our method. The first three datasets are standard TAS benchmarks, where at most 20% of video frames are background. In contrast, CrossTask is collected from the Internet, and around 72% of video frames are background.

### 4.2 Implementation Details

For all datasets, we extracted video frame features by using the I3D [12] model, and obtained clip features by average pooling 32 consecutive frames [80, 101]; except in GTEA, we use 10 frames per clip since videos are much shorter compared to other datasets. For evaluation, similar to prior works [33, 101, 125], we upsample results and report: accuracy, segmental edit distance (edit), segmental f1 score at overlapping threshold 10%, 25%, 50%, denoted by f1@10, 25, 50. For modeling, We use ASFormer [125] as our backbone, with a lighter design. More implementation details are provided in the supplementary materials.

| | 50Salads | | | | | | GTEA | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| method | budget | acc | edit | f1@{10,25,50} | | | budget | acc | edit | f1@{10,25,50} | | |
| split-rand [87] | 0.16% | 49.0 | 39.8 | 48.0 | 42.0 | 24.7 | 0.5% | 45.2 | 54.1 | 56.1 | 47.9 | 25.5 |
| split-entropy [87] | 0.16% | 45.8 | 35.2 | 39.1 | 34.6 | 16.2 | 0.5% | 45.1 | 56.9 | 58.1 | 47.1 | 25.3 |
| equidistant [87] | 0.16% | 51.0 | 36.5 | 44.8 | 38.6 | 27.8 | 0.5% | 42.7 | 56.3 | 55.3 | 45.1 | 20.7 |
| coreset [122] | 0.16% | 38.4 | 26.1 | 29.1 | 24.9 | 13.8 | 0.5% | 43.1 | 50.3 | 50.7 | 41.7 | 23.4 |
| ours | 0.16% | **57.8** | **45.0** | **55.1** | **49.1** | **32.9** | 0.5% | **47.6** | **57.0** | **59.9** | **48.7** | **27.3** |
| *full* | *100%* | *82.4* | *73.2* | *82.8* | *80.3* | *67.4* | *100%* | *73.4* | *82.5* | *88.6* | *84.4* | *69.2* |
| | Breakfast | | | | | | CrossTask | | | | | |
| method | budget | acc | edit | f1@{10,25,50} | | | budget | acc | edit | f1@{10,25,50} | | |
| split-rand [87] | 0.16% | 61.8 | 56.9 | 61.1 | 55.1 | 39.4 | 0.16% | 73.4 | 33.2 | 29.7 | 22.7 | 11.2 |
| split-entropy [87] | 0.16% | 61.8 | 55.8 | 61.9 | 56.8 | 41.0 | 0.16% | 73.1 | 34.3 | 30.0 | 23.2 | 11.0 |
| equidistant [87] | 0.16% | 58.5 | 52.0 | 55.6 | 49.2 | 34.0 | 0.16% | 71.5 | 32.5 | 28.7 | 20.9 | 9.8 |
| coreset [122] | 0.16% | 61.0 | 56.0 | 60.6 | 55.9 | 40.5 | 0.16% | 73.0 | 32.3 | 27.8 | 22.2 | 10.9 |
| ours | 0.16% | **63.5** | **58.6** | **62.8** | **58.1** | **43.5** | 0.16% | **73.9** | **35.5** | **30.4** | **24.8** | **12.4** |
| *full* | *100%* | *76.4* | *73.2* | *75.6* | *72.0* | *57.9* | *100%* | *79.3* | *46.0* | *48.7* | *43.6* | *27.3* |

**Table 1:** Comparison of our proposed method with other baselines on four datasets.

**Active Learning settings.** During each active learning iteration, we select $m = 5\%$ of videos in the training set, and select $\rho = 25\%$ clips of each selected video. We label 1 frame per selected clip, resulting in a total frame percentage of AL iterations $\times\ m \times \rho$/clip-size. We randomly select videos in the first iteration when we initialize the model, and then use our proposed method for all future selections. We repeat iterations until the total budget is met or the target performance level is reached. Videos will not be repeatedly selected for labeling, and AL iterations will be stopped once all videos are selected.

### 4.3 Baseline Methods

Our work is the *first to explore active learning in TAS, so there is a lack of baselines. Thus, we design baselines from recent related research.* For intra-video selection, we use the following baselines: equidistant, split-random, split-entropy and coreset [96]. The first three baselines are adopted from recent research in active video action detection [87] while the last one was used in active fine-tuning [122]. Note we cannot use [87,122] directly as [87] is for trimmed videos and [122] is for fine tuning a pre-trained model. To be more specific, the equidistant method selects clips by skipping a fixed interval. In the case of split-random and split-entropy, we first split each video $i$ into equal $T_i/4$ intervals and then select a clip from each interval either randomly or based on the highest entropy [39]. Following [39], we enable drop-out and repeat inferences for 50 times to calculate the entropy of each clip. The Coreset method [96], treats clip selection as a K-center problem, which is solved using a naive greedy algorithm. All baselines use random selection for inter-video selection, and use the same TAS model as ours.

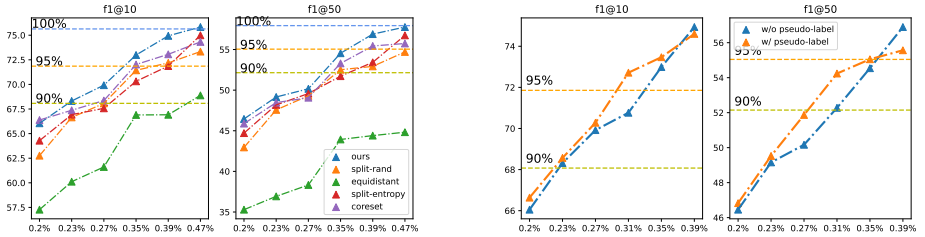


**Fig. 3:** (a) Continue AL iteration up to total 0.47% annotated frames, each horizontal line represents the percentage of fully-supervised performance. (b) Continue AL iteration with pseudo labels up to total 0.39% annotated frames, each horizontal line represents the percentage of fully-supervised performance.

### 4.4 Main Results

**Comparison with baselines.** Tab. 1 shows comparison of our method with four baselines on 50Salads, GTEA, Breakfast and CrossTask datasets. We fix the budget by running all methods for 4 active learning iterations, equating to total 0.16% frame budget (0.5% for GTEA). We also report the upper bound performance (*full*), where we use the annotations of all clips during training.

Notice on all datasets, the first three baselines have higher performance than coreset. We argue this is mainly because the first three baselines enforce a temporal separation for labeling, indicating that diverse temporal information is crucial for TAS learning. Secondly, we noticed coreset performs relatively better than equidistant on Breakfast and CrossTask, this is partly due to CrossTask and Breakfast being much larger than 50Salads and GTEA, and the lack of intra-video temporal information is compensated by labeling more videos.

Overall, our method performs the best across all these datasets, showing the effectiveness of our approach. More specifically, we improve f1@50 by 8.2%, 1.8%, 4.1% and 1.2% on 50Salads, GTEA, Breakfast and CrossTask, respectively. Moreover, by labeling only 0.16% frames (0.5% for GTEA), we achieve 57.8%, 69.1%, 80.1%, and 77.2% of fully-supervised performance on edit distance.

**Qualitative comparison.** Our qualitative analysis, as shown in Fig. 1, compares our method with the split-random baseline across four active learning iterations. A significant difference is evident when examining the segmentation quality of our model, especially when increasing the budget from 0.12% to 0.16%. Compared to the initially trained model and the split-random approach, our model's segmentation quality is substantially better, closely aligning with the ground truth and successfully segmenting all major components. In contrast, the split-random approach fails to identify at least one large segment (the blue one in the middle right), highlighting the superiority of our approach.

**Towards Full-supervision Performance.** In Fig. 3a, we continue active learning iterations aiming to achieve full-supervision performance on Breakfast. Our method surpasses all baselines over iterations. In particular, with 0.23% frames

| method | budget | acc | edit | f1@{10,25,50} | | |
|---|---|---|---|---|---|---|
| D-TSTAS (Timestamp) [69] | 0.29% | 65.7 | **75.8** | 71.3 | 69.3 | 50.7 |
| EM-gen (SkipTag) [85] | 0.29% | 64.1 | 59.9 | n/a | 57.3 | 45.2 |
| Ours | 0.23% | **73.4** | 71.1 | **74.4** | **70.1** | **54.3** |

**Table 2:** Comparison with timestamp and SkipTag supervision methods on Breakfast.

| Breakfast | MoF | IoU | IoD | CrossTask | MoF | IoU | IoD |
|---|---|---|---|---|---|---|---|
| TASL [73] | 47.8 | 35.2 | 46.1 | TASL [73] | 40.7 | 14.5 | 25.1 |
| POC [74] | 45.7 | 38.3 | - | POC [74] | 42.8 | 15.6 | - |
| MuCon [106] | 48.5 | **40.9** | 54 | MuCon [106] | - | - | - |
| Ours | **63.5** | 39.2 | 49.6 | Ours | **73.9** | **23.7** | **30.1** |

**Table 3:** Comparison with action set and transcript supervision methods.

labeled, our method achieves over 90% of the full-supervision performance in terms of f1@10, while other baselines require 0.27% labeled frames to achieve the same level of performance. Moreover, by labeling 0.35% and 0.47% frames, we attain 95% and 100% of the full-supervision performance, demonstrating the efficiency and effectiveness of our active learning method.

**AL in Semi-supervised Settings.** One might wonder if trained TAS model can generate pseudo-labels to assist learning in future AL iterations, and we show our method can extend to semi-supervised learning in Fig. 3b. Starting from the 5th iteration, in addition to newly annotated videos, we select 1.25% of videos with the lowest alignment costs and pseudo-label them. We add these pseudo labeled videos to labeled video set for future training. Comparison between learning "w/ pseudo-label" and standard "w/o pseudo-label" setting (orange v.s. blue) shows performance is improved with additional pseudo-labels. More specifically, under 0.31% labeled frames budget, learning with additional pseudo labels achieves close to 95% of fully-supervised performance on f1@50, while learning with labeled frames only achieves 90% of fully-supervised performance on f1@50. We further notice after 0.35% frames budget, our standard method outperforms semi-supervised one, likely due to noises introduced by pseudo-labels over iterations, resulting in slightly lower performance.

### 4.5 Comparison with Weakly Supervised Methods

**Comparison with timestamp supervision methods.** Timestamp supervision [8,50,67,69,85] methods reduce annotation cost by labeling one frame from each action in every video. Examples are D-TSTAS [69] and EM-Gen [85]. Additionally, EM-Gen [85] proposed a new form of supervision named SkipTag, allowing an annotator to randomly label K frames anywhere in the video, where K is the average number of action segments. Although SkipTag reduces the restrictions in timestamp supervision, it still requires all videos to be labeled. Unlike these methods, we do not require watching every entire video or ensuring each

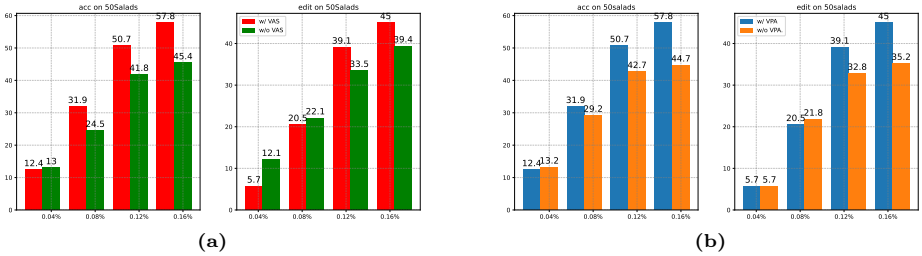


**Fig. 4:** (a) Ablation of our intra-video selection method (VAS) under 0.04%, 0.08%, 0.12% and 0.16% frame budget on 50Salads. (b) Ablation studies of our inter-video selection method (VPA) under 0.04%, 0.08%, 0.12% and 0.16% frame budget on 50Salads.

selected frame represents a different action segment, significantly simplifying the annotation process while striving for high efficiency and minimal budget.

In Tab. 2, we compare our approach against both D-TSTAS (Timestamp) and EM-Gen (SkipTag) on Breakfast. We adjust $\rho$ so we label K frames per selected video while keeping $m = 5\%$ for inter-video selection. Note that our approach annotates only a subset of videos during each iteration, while D-TSTAS and EM-Gen annotate all videos. Consequently, the annotation budget of our method is less than that of D-TSTAS and EM-Gen, and aligns with them only when all videos have been chosen for labeling. Breakfast contains roughly 3.6M frames [67] and K=6 [85], which leads to $(1712\times6)/3.6M$ frames $\approx 0.29\%$ budget for D-TSTAS and EM-Gen. For comparison with EM-Gen and D-TSTAS, when selecting 80% videos (0.23% frames), our method significantly outperforms both methods, e.g., surpassing EM-Gen by 9.1% and D-TSTAS by 3.6% on f1@50. In summary, our method performs better than both Timestamp and SkipTag supervision methods, while using less annotation budget.

**Comparison with action set and transcript supervision methods.** Action set and transcript supervision methods reduce annotation cost by learning from a list of unordered actions or ordered actions, respectively [17]. Quantitatively comparing supervision costs is challenging because obtaining transcripts or action sets typically requires an annotator to watch the entire video. Notably, [67] states that timestamp "does not require more time than annotating transcripts," whereas our supervision cost is lower than timestamp.

We train our method for four AL iterations on Breakfast and CrossTask datasets and compare with transcript methods (TASL [73], MuCon [106]) and an action set method (POC [74]). Tab. 3 demonstrates that our method achieves superior or comparable results while labeling only 0.16% of frames.

### 4.6 Ablation Studies

**Action prototype learning module effectiveness.** We evaluate the effectiveness of our action prototype learning module in Tab. 4a and Tab. 4b. For the first row, we use the average embedding features of each action to replace the

| $\mathcal{L}_{\text{cont}}$ | $\mathcal{L}_{\mu}$ | $\mathcal{L}_{\sigma}$ | acc | edit | f1@{10,25,50} | | |
|---|---|---|---|---|---|---|---|
| × | × | × | 50.0 | 44.4 | 52.2 | 46.4 | 26.8 |
| ✓ | × | × | 47.2 | 40.4 | 45.5 | 39.7 | 22.2 |
| ✓ | ✓ | × | 42.0 | 37.8 | 40.2 | 33.6 | 17.8 |
| ✓ | × | ✓ | 49.9 | 40.6 | 48.8 | 42.0 | 23.9 |
| ✓ | ✓ | ✓ | **57.8** | **45.0** | **55.1** | **49.1** | **32.9** |

**(a)** 50Salads

| $\mathcal{L}_{\text{cont}}$ | $\mathcal{L}_{\mu}$ | $\mathcal{L}_{\sigma}$ | acc | edit | f1@{10,25,50} | | |
|---|---|---|---|---|---|---|---|
| × | × | × | 46.3 | 56.6 | 58.9 | 48.7 | 27.2 |
| ✓ | × | × | 43.9 | **57.3** | 57.4 | 48.0 | 23.6 |
| ✓ | ✓ | × | 46.5 | 54.9 | 57.9 | 48.6 | 26.4 |
| ✓ | × | ✓ | 42.7 | 56.6 | 57.9 | 47.1 | 25.9 |
| ✓ | ✓ | ✓ | **47.6** | 57.0 | **59.9** | **48.7** | **27.3** |

**(b)** GTEA

**Table 4:** Loss ablation after 4 active learning iterations, all cases use our AL method.

prototypes for selection, and only keep cross entropy loss and smoothing loss for training. For other cases, we ablate each part in our proposed loss. We run our method for four iterations. Due to lack of labels, using the contrastive loss $\mathcal{L}_{\text{cont}}$ alone reduces the performance. In contrast, after adding our two regularizers ($\mathcal{L}_{\mu}$ and $\mathcal{L}_{\sigma}$), the performance is significantly improved.

**Intra-video selection effectiveness.** Next, we use split-random to replace our intra-video selection method (VAS), while keeping the inter-video selection (VPA) the same. Figure 4a shows accuracy and edit distance on 50Salads. Notice that VAS outperforms split-random in almost all cases, suggesting the effectiveness of our VAS, creating synergies with our inter-video selection.

**Inter-video selection effectiveness.** Here, we replace our inter-video selection (VPA) by random selection, while keeping the intra-video selection (VAS). As Figure 4b shows, our method starts from almost the same performance as random (since we also use random for initialization). With more budgets, our method performs better than random selection, suggesting that our proposed inter-video selection (VPA) can better capture useful videos for annotation by selecting action-wise diverse videos. Please refer to the supplementary material for additional experiments, discussions.

## 5 Limitation and Future Work

Our method is efficient, but the labeled dataset grows iteratively. Future research could explore continual active learning, maintaining a constant-size dataset by cycling videos, potentially reducing training time.

## 6 Conclusions

We proposed a two-stage active learning framework for TAS, consisting of three parts: i) Regularized action prototype module to learn discriminative prototypes for enhanced accuracy and computational efficiency. ii) Inter-video selection to select unlabeled videos with diverse sequence of actions. iii) Intra-video selection to identify informative clips in selected unlabeled videos for annotation. With extensive experiments on four datasets, we showed our approach consistently outperforms baselines over iterations, and achieve comparable or better performance than other weakly-supervised methods using minimum annotation.

## Acknowledgements

This work is sponsored by DARPA PTG (HR00112220001), NSF (IIS-2115110), ARO (W911NF2110276). Content does not necessarily reflect the position/policy of the Government.

## References

1. Aakur, S.N., Sarkar, S.: A perceptual prediction framework for self supervised event segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1197–1206 (2019)
2. Agarwal, S., Arora, H., Anand, S., Arora, C.: Contextual diversity for active learning. In: Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVI. p. 137–153 (2020)
3. Angluin, D.: Queries and concept learning. Machine learning **2**, 319–342 (1988)
4. Aziere, N., Todorovic, S.: Multistage temporal convolution transformer for action segmentation. Image and Vision Computing **128**, 104567 (2022)
5. Bansal, S., Arora, C., Jawahar, C.: My view is the best view: Procedure learning from egocentric videos. In: European Conference on Computer Vision (ECCV) (2022)
6. Behrmann, N., Golestaneh, S.A., Kolter, Z., Gall, J., Noroozi, M.: Unified fully and timestamp supervised temporal action segmentation via sequence to sequence translation. In: ECCV (2022)
7. Beluch, W.H., Genewein, T., Nürnberger, A., Köhler, J.M.: The power of ensembles for active learning in image classification. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 9368–9377 (2018)
8. Bueno-Benito, E., Vecino, B.T., Dimiccoli, M.: Leveraging triplet loss for unsupervised action segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops. pp. 4922–4930 (June 2023)
9. Cabannes, V., Bottou, L., Lecun, Y., Balestriero, R.: Active self-supervised learning: A few low-cost relationships are all you need. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 16274–16283 (October 2023)
10. Cao, K., Ji, J., Cao, Z., Chang, C.Y., Niebles, J.C.: Few-shot video classification via temporal alignment. IEEE Conference on Computer Vision and Pattern Recognition (2020)
11. Cao, Y.T., Shi, Y., Yu, B., Wang, J., Tao, D.: Knowledge-aware federated active learning with non-iid data. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 22279–22289 (October 2023)
12. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: IEEE Conference on Computer Vision and Pattern Recognition (2017)
13. Chang, C.Y., Huang, D.A., Sui, Y., Fei-Fei, L., Niebles, J.C.: D3tw: Discriminative differentiable dynamic time warping for weakly supervised action alignment and segmentation. IEEE Conference on Computer Vision and Pattern Recognition (2019)
14. Chang, X., Tung, F., Mori, G.: Learning discriminative prototypes with dynamic time warping. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8395–8404 (2021)

15. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.E.: A simple framework for contrastive learning of visual representations. International Conference on Machine learning (2020)
16. Cuturi, M., Blondel, M.: Soft-dtw: a differentiable loss function for time-series. International Conference on Machine learning (2017)
17. Ding, G., Sener, F., Yao, A.: Temporal action segmentation: An analysis of modern techniques. IEEE Transactions on Pattern Analysis and Machine Intelligence (2023)
18. Ding, G., Yao, A.: Leveraging action affinity and continuity for semi-supervised temporal action segmentation. In: European Conference on Computer Vision. pp. 17–32. Springer (2022)
19. Ding, G., Yao, A.: Temporal action segmentation with high-level complex activity labels. IEEE Transactions on Multimedia (2022)
20. Ding, L., Xu, C.: Tricornet: A hybrid temporal convolutional and recurrent network for video action segmentation. arXiv preprint arXiv:1705.07818 (2017)
21. Ding, L., Xu, C.: Weakly-supervised action segmentation with iterative soft boundary assignment. IEEE Conference on Computer Vision and Pattern Recognition (2018)
22. Donahue, G., Elhamifar, E.: Learning to predict activity progress by self-supervised video alignment. IEEE Conference on Computer Vision and Pattern Recognition (2024)
23. Du, D., Su, B., Li, Y., Qi, Z., Si, L., Shan, Y.: Do we really need temporal convolutions in action segmentation? In: 2023 IEEE International Conference on Multimedia and Expo (ICME). pp. 1014–1019. IEEE (2023)
24. Du, X., Mishra, B.D., Tandon, N., Bosselut, A., tau Yih, W., Clark, P., Cardie, C.: Be consistent! improving procedural text comprehension using label consistency. Annual Meeting of the North American Association for Computational Linguistics (2019)
25. Du, Z., Wang, Q.: Dilated transformer with feature aggregation module for action segmentation. Neural Processing Letters pp. 1–17 (2022)
26. Du, Z., Wang, X., Zhou, G., Wang, Q.: Fast and unsupervised action boundary detection for action segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3323–3332 (2022)
27. Dvornik, N., Hadji, I., Derpanis, K.G., Garg, A., Jepson, A.D.: Drop-dtw: Aligning common signal between sequences while dropping outliers. Neural Information Processing Systems (2021)
28. Dvornik, N., Hadji, I., Pham, H., Bhatt, D., Martinez, B., Fazly, A., Jepson, A.D.: Flow graph to video grounding for weakly-supervised multi-step localization. In: European Conference on Computer Vision. pp. 319–335. Springer (2022)
29. Dvornik, N., Hadji, I., Zhang, R., Derpanis, K.G., Wildes, R.P., Jepson, A.D.: Stepformer: Self-supervised step discovery and localization in instructional videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18952–18961 (2023)
30. Dwibedi, D., Aytar, Y., Tompson, J., Sermanet, P., Zisserman, A.: Temporal cycle-consistency learning. IEEE Conference on Computer Vision and Pattern Recognition (2019)
31. Elhamifar, E., Huynh, D.: Self-supervised multi-task procedure learning from instructional videos. European Conference on Computer Vision (2020)
32. Fang, M., Li, Y., Cohn, T.: Learning how to active learn: A deep reinforcement learning approach. arXiv preprint arXiv:1708.02383 (2017)

33. Farha, Y.A., Gall, J.: Ms-tcn: Multi-stage temporal convolutional network for action segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3575–3584 (2019)
34. Fathi, A., Ren, X., Rehg, J.M.: Learning to recognize objects in egocentric activities. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2011)
35. Fathi, A., Ren, X., Rehg, J.M.: Learning to recognize objects in egocentric activities. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2021)
36. Freytag, A., Rodner, E., Denzler, J.: Selecting influential examples: Active learning with expected model output changes. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part IV 13. pp. 562–577. Springer (2014)
37. Fried, D., Alayrac, J.B., Blunsom, P., Dyer, C., Clark, S., Nematzadeh, A.: Learning to segment actions from observation and narration. Annual Meeting of the Association for Computational Linguistics (2020)
38. Gabrys, R., Yaakobi, E., Milenkovic, O.: Codes in the damerau distance for deletion and adjacent transposition correction. IEEE Transactions on Information Theory (2017)
39. Gal, Y., Ghahramani, Z.: Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In: international conference on machine learning. pp. 1050–1059. PMLR (2016)
40. Gao, S.H., Han, Q., Li, Z.Y., Peng, P., Wang, L., Cheng, M.M.: Global2local: Efficient structure search for video action segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16805–16814 (2021)
41. Goel, K., Brunskill, E.: Learning procedural abstractions and evaluating discrete latent temporal structure. International Conference on Learning Representation (2019)
42. Hadji, I., Derpanis, K.G., Jepson, A.D.: Representation learning via global temporal alignment and cycle-consistency. IEEE Conference on Computer Vision and Pattern Recognition (2021)
43. Hampiholi, B., Jarvers, C., Mader, W., Neumann, H.: Depthwise separable temporal convolutional network for action segmentation. In: 2020 International Conference on 3D Vision (3DV). pp. 633–641. IEEE (2020)
44. Huang, D.A., Fei-Fei, L., Niebles, J.C.: Connectionist temporal modeling for weakly supervised action labeling. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14. pp. 137–153. Springer (2016)
45. Huang, S., Wang, T., Xiong, H., Huan, J., Dou, D.: Semi-supervised active learning with temporal output discrepancy. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3447–3456 (2021)
46. Huang, Y., Sugano, Y., Sato, Y.: Improving action segmentation via graph-based temporal reasoning. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
47. Ishikawa, Y., Kasai, S., Aoki, Y., Kataoka, H.: Alleviating over-segmentation errors by detecting action boundaries. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 2322–2331 (January 2021)

48. Ji, W., Liang, R., Zheng, Z., Zhang, W., Zhang, S., Li, J., Li, M., Chua, T.s.: Are binary annotations sufficient? video moment retrieval via hierarchical uncertainty-based active learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 23013–23022 (June 2023)
49. Joshi, A.J., Porikli, F., Papanikolopoulos, N.: Multi-class active learning for image classification. In: 2009 ieee conference on computer vision and pattern recognition. pp. 2372–2379. IEEE (2009)
50. Khan, H., Haresh, S., Ahmed, A., Siddiqui, S., Konin, A., Zia, M.Z., Tran, Q.H.: Timestamp-supervised action segmentation with graph convolutional networks. In: 2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 10619–10626. IEEE (2022)
51. Koide, S., Kawano, K., Kutsuna, T.: Neural edit operations for biological sequences. Advances in Neural Information Processing Systems (2018)
52. Kothawade, S., Ghosh, S., Shekhar, S., Xiang, Y., Iyer, R.: Talisman: targeted active learning for object detection with rare classes and slices using submodular mutual information. In: European Conference on Computer Vision. pp. 1–16. Springer (2022)
53. Krishnamurthy, V.: Algorithms for optimal scheduling and management of hidden markov model sensors. IEEE Transactions on Signal Processing **50**(6), 1382–1397 (2002)
54. Kuehne, H., Arslan, A., Serre, T.: The language of actions: Recovering the syntax and semantics of goal-directed human. IEEE Conference on Computer Vision and Pattern Recognition (2014)
55. Kuehne, H., Richard, A., Gall, J.: Weakly supervised learning of actions from transcripts. Computer Vision and Image Understanding Journal (2017)
56. Kukleva, A., Kuehne, H., Sener, F., Gall, J.: Unsupervised learning of action classes with continuous temporal embedding. IEEE Conference on Computer Vision and Pattern Recognition (2019)
57. Kye, S.M., Choi, K., Byun, H., Chang, B.: Tidal: Learning training dynamics for active learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 22335–22345 (October 2023)
58. Lea, C., Flynn, M.D., Vidal, R., Reiter, A., Hager, G.D.: Temporal convolutional networks for action segmentation and detection. IEEE Conference on Computer Vision and Pattern Recognition (2017)
59. Lea, C., Reiter, A., Vidal, R., Hager, G.D.: Segmental spatiotemporal cnns for fine-grained action segmentation. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14. pp. 36–52. Springer (2016)
60. Lee, S., Lu, Z., Zhang, Z., Hoai, M., Elhamifar, E.: Error detection in egocentric procedural task videos. IEEE Conference on Computer Vision and Pattern Recognition (2024)
61. Lei, P., Todorovic, S.: Temporal deformable residual networks for action segmentation in videos. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (2018). https://doi.org/10.1109/CVPR.2018.00705
62. Levenshtein, V.I.: Binary Codes Capable of Correcting Deletions, Insertions and Reversals. Soviet Physics Doklady **10** (1966)
63. Li, J., Lei, P., Todorovic, S.: Weakly supervised energy-based learning for action segmentation. International Conference on Computer Vision (2019)
64. Li, R., Zhang, B., Liu, J., Liu, W., Zhao, J., Teng, Z.: Heterogeneous diversity driven active learning for multi-object tracking. In: Proceedings of the IEEE/CVF

International Conference on Computer Vision (ICCV). pp. 9932–9941 (October 2023)
65. Li, S.J., AbuFarha, Y., Liu, Y., Cheng, M.M., Gall, J.: Ms-tcn++: Multi-stage temporal convolutional network for action segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence pp. 1–1 (2020). https://doi.org/10.1109/TPAMI.2020.3021756
66. Li, Y., Dong, Z., Liu, K., Feng, L., Hu, L., Zhu, J., Xu, L., Liu, S., et al.: Efficient two-step networks for temporal action segmentation. Neurocomputing **454**, 373–381 (2021)
67. Li, Z., Abu Farha, Y., Gall, J.: Temporal action segmentation from timestamp supervision. IEEE/CVF Conference on Computer Vision and Pattern Recognition (2021)
68. Liu, D., Li, Q., Dinh, A., Jiang, T., Shah, M., Xu, C.: Diffusion action segmentation. arXiv preprint arXiv:2303.17959 (2023)
69. Liu, K., Li, Y., Liu, S., Tan, C., Shao, Z.: Reducing the label bias for timestamp supervised temporal action segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6503–6513 (June 2023)
70. Liu, Z., Wang, L., Zhou, D., Wang, J., Zhang, S., Bai, Y., Ding, E., Fan, R.: Temporal segment transformer for action segmentation. arXiv preprint arXiv:2302.13074 (2023)
71. Liu, Z., Ding, H., Zhong, H., Li, W., Dai, J., He, C.: Influence selection for active learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9274–9283 (2021)
72. Liu, Z., Wang, J., Gong, S., Lu, H., Tao, D.: Deep reinforcement active learning for human-in-the-loop person re-identification. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 6122–6131 (2019)
73. Lu, Z., Elhamifar, E.: Weakly-supervised action segmentation and alignment via transcript-aware union-of-subspaces learning. International Conference on Computer Vision (2021)
74. Lu, Z., Elhamifar, E.: Set-supervised action learning in procedural task videos via pairwise order consistency. IEEE Conference on Computer Vision and Pattern Recognition (2022)
75. Lu, Z., Elhamifar, E.: Fact: Frame-action cross-attention temporal modeling for efficient action segmentation. IEEE Conference on Computer Vision and Pattern Recognition (2024)
76. Luo, W., Schwing, A., Urtasun, R.: Latent structured active learning. Advances in Neural Information Processing Systems **26** (2013)
77. Mahapatra, D., Bozorgtabar, B., Thiran, J.P., Reyes, M.: Efficient active learning for image classification and segmentation using a sample selection and conditional generative adversarial network. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 580–588. Springer (2018)
78. Mahmood, R., Fidler, S., Law, M.T.: Low budget active learning via wasserstein distance: An integer programming approach. arXiv preprint arXiv:2106.02968 (2021)
79. Mayer, C., Timofte, R.: Adversarial sampling for active learning. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 3071–3079 (2020)
80. Miech, A., Alayrac, J.B., Smaira, L., Laptev, I., Sivic, J., Zisserman, A.: End-to-end learning of visual representations from uncurated instructional videos. IEEE Conference on Computer Vision and Pattern Recognition (2020)

81. Miech, A., Zhukov, D., Alayrac, J.B., Tapaswi, M., Laptev, I., Sivic, J.: Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. International Conference on Computer Vision (2019)
82. Müller, M.: Information retrieval for music and motion, vol. 2. Springer (2007)
83. Narr, A., Triebel, R., Cremers, D.: Stream-based active learning for efficient and adaptive classification of 3d objects. In: 2016 IEEE International Conference on Robotics and Automation (ICRA). pp. 227–233. IEEE (2016)
84. Park, J., Kim, D., Huh, S., Jo, S.: Maximization and restoration: Action segmentation through dilation passing and temporal reconstruction. Pattern Recognition **129**, 108764 (2022)
85. Rahaman, R., Singhania, D., Thiery, A., Yao, A.: A generalized and robust framework for timestamp supervision in temporal action segmentation. In: Computer Vision–ECCV 2022: 17th European Conference (2022)
86. Rana, A., Rawat, Y.: Are all frames equal? active sparse labeling for video action detection. Advances in Neural Information Processing Systems **35**, 14358–14373 (2022)
87. Rana, A.J., Rawat, Y.S.: Hybrid active learning via deep clustering for video action detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18867–18877 (2023)
88. Ren, P., Xiao, Y., Chang, X., Huang, P.Y., Li, Z., Gupta, B.B., Chen, X., Wang, X.: A survey of deep active learning. ACM computing surveys (CSUR) **54**(9), 1–40 (2021)
89. Richard, A., Kuehne, H., Gall, J.: Weakly supervised action learning with rnn based fine-to-coarse modeling. IEEE Conference on Computer Vision and Pattern Recognition (2017)
90. Richard, A., Kuehne, H., Gall, J.: Action sets: Weakly supervised action segmentation without ordering constraints. IEEE Conference on Computer Vision and Pattern Recognition (2018)
91. Rochan, M., Wang, Y.: Video summarization by learning from unpaired data. IEEE Conference on Computer Vision and Pattern Recognition (2019)
92. Sakoe, H., Chiba, S.: Dynamic programming algorithm optimization for spoken word recognition. IEEE Transactions on Acoustics, Speech and Signal Processing **26** (1978)
93. Sarfraz, S., Murray, N., Sharma, V., Diba, A., Van Gool, L., Stiefelhagen, R.: Temporally-weighted hierarchical clustering for unsupervised action segmentation. IEEE/CVF Conference on Computer Vision and Pattern Recognition (2021)
94. Sener, F., Singhania, D., Yao, A.: Temporal aggregate representations for long-range video understanding. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVI 16. pp. 154–171. Springer (2020)
95. Sener, F., Yao, A.: Unsupervised learning and segmentation of complex activities from video. IEEE Conference on Computer Vision and Pattern Recognition (2018)
96. Sener, O., Savarese, S.: Active learning for convolutional neural networks: A core-set approach. In: International Conference on Learning Representations (2018), https://openreview.net/forum?id=H1aIuk-RW
97. Sener, O., Zamir, A.R., Savarese, S., Saxena, A.: Unsupervised semantic parsing of video collections. IEEE International Conference on Computer Vision (2015)
98. Shah, A., Lundell, B., Sawhney, H., Chellappa, R.: Steps: Self-supervised key step extraction and localization from unlabeled procedural videos. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 10375–10387 (October 2023)

99. Shen, Y., Elhamifar, E.: Semi-weakly-supervised learning of complex actions from instructional task videos. IEEE Conference on Computer Vision and Pattern Recognition (2022)
100. Shen, Y., Elhamifar, E.: Progress-aware online action segmentation for egocentric procedural task videos. IEEE Conference on Computer Vision and Pattern Recognition (2024)
101. Shen, Y., Wang, L., Elhamifar, E.: Learning to segment actions from visual and language instructions via differentiable weak sequence alignment. IEEE Conference on Computer Vision and Pattern Recognition (2021)
102. Singh, B., Marks, T.K., Jones, M., Tuzel, O., Shao, M.: A multi-stream bi-directional recurrent neural network for finegrained action detection. IEEE Conference on Computer Vision and Pattern Recognition (2016)
103. Singhania, D., Rahaman, R., Yao, A.: Coarse to fine multi-resolution temporal convolutional network. CoRR **abs/2105.10859** (2021), `https://arxiv.org/abs/2105.10859`
104. Singhania, D., Rahaman, R., Yao, A.: Iterative contrast-classify for semi-supervised temporal action segmentation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36, pp. 2262–2270 (2022)
105. Souri, Y., Farha, Y.A., Despinoy, F., Francesca, G., Gall, J.: Fifa: Fast inference approximation for action segmentation. In: DAGM German Conference on Pattern Recognition. pp. 282–296. Springer (2021)
106. Souri, Y., Fayyaz, M., Minciullo, L., Francesca, G., Gall, J.: Fast weakly supervised action segmentation using mutual consistency. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(10), 6196–6208 (2021)
107. Stein, S., McKenna, S.J.: Combining embedded accelerometers with computer vision for recognizing food preparation activities. In: Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing (2013)
108. Su, B., Hua, G.: Order-preserving wasserstein distance for sequence matching. IEEE Conference on Computer Vision and Pattern Recognition (2017)
109. Tang, Y., Zhang, X., Ma, L., Wang, J., Chen, S., Jiang, Y.G.: Non-local netvlad encoding for video classification. In: Proceedings of the European Conference on Computer Vision (ECCV) Workshops. pp. 0–0 (2018)
110. Tian, X., Jin, Y., Tang, X.: Local–global transformer neural network for temporal action segmentation. Multimedia Systems **29**(2), 615–626 (2023)
111. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. Neural Information Processing Systems (2017)
112. VidalMata, R.G., Scheirer, W.J., Kukleva, A., Cox, D., Kuehne, H.: Joint visual-temporal embedding for unsupervised learning of actions in untrimmed sequences. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 1238–1247 (2021)
113. Wang, D., Yuan, Y., Wang, Q.: Gated forward refinement network for action segmentation. Neurocomputing **407**, 63–71 (2020)
114. Wang, J., Du, Z., Li, A., Wang, Y.: Atrous temporal convolutional network for video action segmentation. In: 2019 IEEE International Conference on Image Processing (ICIP). pp. 1585–1589. IEEE (2019)
115. Wang, J., Wang, Z., Zhuang, S., Hao, Y., Wang, H.: Cross-enhancement transformer for action segmentation. Multimedia Tools and Applications pp. 1–14 (2023)

116. Wang, K., Zhang, D., Li, Y., Zhang, R., Lin, L.: Cost-effective active learning for deep image classification. IEEE Transactions on Circuits and Systems for Video Technology **27**(12), 2591–2600 (2016)
117. Wang, X., Zhang, S., Qing, Z., Shao, Y., Gao, C., Sang, N.: Self-supervised learning for semi-supervised temporal action proposal. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1905–1914 (2021)
118. Wang, Z., Chen, H., Li, X., Liu, C., Xiong, Y., Tighe, J., Fowlkes, C.: Sscap: Self-supervised co-occurrence action parsing for unsupervised temporal action segmentation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 1819–1828 (2022)
119. Wang, Z., Gao, Z., Wang, L., Li, Z., Wu, G.: Boundary-aware cascade networks for temporal action segmentation. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXV 16. pp. 34–51. Springer (2020)
120. Wanyan, Y., Yang, X., Chen, C., Xu, C.: Active exploration of multimodal complementarity for few-shot action recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6492–6502 (June 2023)
121. Wei, K., Iyer, R., Bilmes, J.: Submodularity in data subset selection and active learning. In: International conference on machine learning. pp. 1954–1963. PMLR (2015)
122. Xie, Y., Lu, H., Yan, J., Yang, X., Tomizuka, M., Zhan, W.: Active finetuning: Exploiting annotation budget in the pretraining-finetuning paradigm. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 23715–23724 (2023)
123. Xu, C., Elhamifar, E.: Deep supervised summarization: Algorithm and application to learning instructions. Neural Information Processing Systems (2019)
124. Yang, Y., Ma, J., Huang, S., Chen, L., Lin, X., Han, G., Chang, S.F.: Tempclr: Temporal alignment representation with contrastive learning. arXiv preprint arXiv:2212.13738 (2022)
125. Yi, F., Wen, H., Jiang, T.: Asformer: Transformer for action segmentation. In: The British Machine Vision Conference (BMVC) (2021)
126. Yoo, D., Kweon, I.S.: Learning loss for active learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 93–102 (2019)
127. Yuan, T., Wan, F., Fu, M., Liu, J., Xu, S., Ji, X., Ye, Q.: Multiple instance active learning for object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5330–5339 (2021)
128. Zhang, J., Tsai, P.H., Tsai, M.H.: Semantic2graph: Graph-based multi-modal feature fusion for action segmentation in videos (2022)
129. Zhang, K., Chao, W.L., Sha, F., Grauman, K.: Summary transfer: Exemplar-based subset selection for video summarization. IEEE Conference on Computer Vision and Pattern Recognition (2016)
130. Zhang, Y., Ren, K., Zhang, C., Yan, T.: Sg-tcn: Semantic guidance temporal convolutional network for action segmentation. In: 2022 International Joint Conference on Neural Networks (IJCNN). pp. 1–8. IEEE (2022)
131. Zhao, G., Dougherty, E., Yoon, B.J., Alexander, F., Qian, X.: Uncertainty-aware active learning for optimal bayesian classifier. In: International Conference on Learning Representations (ICLR 2021) (2021)

132. Zhou, F., Torre, F.: Canonical time warping for alignment of human behavior. Advances in neural information processing systems **22** (2009)
133. Zhu, J.J., Bento, J.: Generative adversarial active learning. arXiv preprint arXiv:1702.07956 (2017)
134. Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable detr: Deformable transformers for end-to-end object detection. arXiv preprint arXiv:2010.04159 (2020)
135. Zhukov, D., Alayrac, J.B., Cinbis, R.G., Fouhey, D., Laptev, I., Sivic, J.: Cross-task weakly supervised learning from instructional videos. IEEE Conference on Computer Vision and Pattern Recognition (2019)
136. Zolfaghari Bengar, J., Gonzalez-Garcia, A., Villalonga, G., Raducanu, B., Habibi Aghdam, H., Mozerov, M., Lopez, A.M., Van de Weijer, J.: Temporal coherence for active learning in videos. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops. pp. 0–0 (2019)