

Weakly-Supervised Online Action Segmentation in Multi-View Instructional Videos

Reza Ghoddoosian^{1,2*}, Isht Dwivedi¹, Nakul Agarwal¹, Chiho Choi¹, and Behzad Dariush¹

¹Honda Research Institute, USA

²VLM Lab, University of Texas at Arlington

reza.ghoddoosian@mavs.uta.edu, {idwivedi,nagarwal,cchoi,bdariush}@honda-ri.com

Abstract

This paper addresses a new problem of weakly-supervised online action segmentation in instructional videos. We present a framework to segment streaming videos online at test time using Dynamic Programming and show its advantages over greedy sliding window approach. We improve our framework by introducing the Online-Offline Discrepancy Loss (OODL) to encourage the segmentation results to have a higher temporal consistency. Furthermore, only during training, we exploit frame-wise correspondence between multiple views as supervision for training weakly-labeled instructional videos. In particular, we investigate three different multi-view inference techniques to generate more accurate frame-wise pseudo ground-truth with no additional annotation cost. We present results and ablation studies on two benchmark multi-view datasets, *Breakfast* and *IKEA ASM*. Experimental results show efficacy of the proposed methods both qualitatively and quantitatively in two domains of cooking and assembly.

1. Introduction

Action understanding in untrimmed instructional videos is important in many applications, where agents learn by observation of other agents performing complex tasks. Such videos are characterized by composition of a sequence of low-level atomic actions, e.g., *crack eggs* and *whisk eggs*, that form a high-level task, e.g., *making eggs*. This contextual dependency between actions as well as other attributes in instructional videos have inspired new research [5, 19, 33, 39, 48] that has advanced the field.

A fully-supervised training of these videos would require not only the labels for each action, but also their temporal assignment (start and end time) with ordering constraints. However, creating fully annotated clips with action

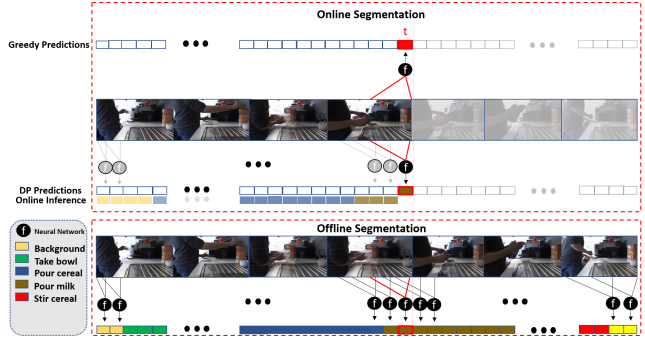


Figure 1. Top: online segmentation, where the frame of interest at time t is identified either greedily by the f function or through DP-based online inference based on current and past predictions. Bottom: Offline segmentation after observing the whole sequence.

assignments and labels on the temporal boundaries of individual actions is manually intensive and is therefore both time consuming and expensive. This limits the scale and practicality at which fully-supervised video datasets can be created. Furthermore, the subjective nature of labeling the start and end time of each action results in ambiguities and inconsistencies. In *weakly-supervised action segmentation* these limitations are addressed by using only the ordered sequence of action labels per video during training, and forgo subjective labeling of start and end time of each action.

Another important consideration in action understanding relates to requirements for processing the videos online versus offline, which is not addressed in existing weakly-supervised segmentation methods [6, 27, 47]. Online processing with low latency is an increasingly important part of interactive applications where real-time, or near real-time feedback is critical. For example, interactive applications such as human-robot interaction, error correction in manufacturing assembly, and virtual rehabilitation require immediate feedback from the intelligent system as the video streams arrive.

The work presented in this paper considers the two aforementioned aspects in action segmentation: weak-

*Work done during Reza's internship at Honda Research Institute, USA

supervision and online processing aimed at temporally partitioning videos into action segments. To our knowledge, our work is the first to address the problem of weakly supervised online action segmentation. Specifically, we present a framework to segment streaming instructional videos online at test time using Dynamic Programming (DP). We show the advantages of using DP as opposed to the greedy sliding window approach that are frequently used in previous online action understanding work [10, 16, 55] (Fig. 1).

We also introduce the Online-Offline Discrepancy Loss (OODL). Offline segmentation refers to inference after observing the video in its entirety. Offline segmentation is a non-causal procedure that is generally expected to be more accurate than its online counterpart that makes inference from partial observations. Indeed, there is a trade-off between accuracy of the recognized actions and low-latency (Sec. 6.2.1). The OODL loss uses the offline segmentation result as a reference and penalizes its difference with online segmentation results generated at each time step in the video. Effectively, this encourages the segmentation results inferred at different observation end points in the video to have higher temporal consistency with respect to each other.

Furthermore, due to lack of frame-level annotation in weakly-labeled videos, frame-wise correspondence between multiple synchronized views of the same recording can provide helpful cues about the temporal location of each action during training. Our work is the first to use the supervision of frame-level correspondence between different views for action segmentation. We compare three ways to exploit this multi-view correspondence to generate more accurate frame-level pseudo ground-truth for weakly-labeled videos. This is in contrast to previous segmentation methods [4, 6, 27], where different views are treated independently, discarding important multi-view information. Note that we only use the multi-view correspondence at training time and our method segments each video independently at test time with no access to other views. Also, our framework utilizes no additional annotation cost, as it is trained independent of the label and number of view points.

In summary, our main contributions are as follows:

- 1) We are the first to address the problem of weakly-supervised online action segmentation in instructional videos, and offer a DP-based framework.
- 2) We introduce the Online-Offline Discrepancy Loss (OODL). The OODL loss utilizes the offline segmentation result as a reference to train the online model by minimizing the difference between online and offline inference results.
- 3) We use frame-wise multi-view correspondence, during training only, to generate more accurate action pseudo-ground-truth in weakly-labeled videos with no additional annotation cost. Our work is the first to incorporate multi-view video understanding in action segmentation.
- 4) We present results and a detailed ablation study on two

benchmark multi-view datasets in domains of cooking and assembly: Breakfast [24] and IKEA [2]. We show quantitatively and qualitatively how our contributions consistently improve various suggested baselines on both datasets.

2. Related Work

Weakly-Supervised Action Segmentation. There has been extensive research in action segmentation of instructional videos under different forms of supervision, including fully-supervised [12, 17, 22, 41, 53], unsupervised [26, 40, 42], and time-stamp supervised [28] methods. Methods most similar to ours use only the sequence of action labels as the weak supervision in training [4, 6, 9, 18, 27, 37, 38, 46]. However, all previous methods consider offline segmentation of videos, where future frames are used to make predictions at the current frame. Specifically, [9] encodes the entire video first before decoding it to frame-level action scores. The work in [4, 6, 27, 38, 47] use Dynamic Programming (DP) to infer the most likely actions and their duration given the entire video. Our method also uses a DP-based framework, but to our knowledge, we are the first to introduce a weakly-supervised method to segment a streaming video in an online manner.

Online Action Understanding. Online action understanding has been studied in various problems such as online action detection [10, 55, 59], start of action detection [15, 44] or anticipation [1, 13, 14, 23, 30]. In the context of online action detection, [59] employs knowledge distillation to transfer information from offline to online models and [10, 36, 55, 56] introduce new neural networks to classify current actions in streaming videos using a sliding window approach. Others have focused solely on detecting the start of an ongoing action immediately [15, 44] or with a short delay [58]. However, past methods did not consider instructional videos and, more importantly, required frame-level labels to train.

Most similar to our work is WOAD [16] as the only weakly-supervised online action detection framework. WOAD [16] is different to our framework in two main ways: First, as a detection model, it is formulated to identify and localize occurrences of, typically, a single action in the input video, while we focus on instructional videos with a series of many unique actions. Second, during test time, we utilize Dynamic Programming and show in our experiments that it outperforms the greedy approach taken in [16].

Multi-View Action Understanding. Using video feeds from multiple view points has improved performance for different problems such as action recognition [29, 35, 49, 50, 52], person identification [11], anomaly detection [8], and video summarization [21, 31, 32]. Similar to our work, [20, 35, 45, 49] limit exploiting multiple views to training time only. In particular, [35, 49] focus on fully-supervised learning of trimmed videos. Meanwhile, [20] explores

unsupervised video-to-video alignment, but utilize partial frame-level labels for classification. In addition, [43, 45] study domain adaptation across 3rd and 1st person views. However, unlike us, they rely on view-specific labels for training. Others [29, 52, 54] use multiple data modalities as view points. Specifically, [54] introduces a semi-supervised and view-agnostic framework for trimmed video classification, where multiple modalities are fused to generate video pseudo labels. These pseudo labels are used along with a selected number of ground-truth labels to train a video classifier. In contrast, to our knowledge, we are the first to use multi-view for temporal segmentation in untrimmed videos without frame-level supervision.

3. Background

This section describes definitions and background concepts used henceforth. For more clarity, the supplementary material provides a table of all symbols used.

3.1. Problem Definition

During training, the input to our model is a video of length T represented by frame-level features \mathbf{x}_1^T and an ordered sequence of actions $\tau = (\tau_1, \tau_2, \dots, \tau_M)$ known as the transcript. M is the number of actions in a given video and can vary across videos. Information about the start and end time of each action is not known.

At test time, given the set of action labels in the dataset \mathbb{A} , the goal is to identify the action label $a_t \in \mathbb{A}$ at frame t for all $0 < t < T + 1$ based on only the past and current observations \mathbf{x}_1^t . The final result will be a sequence of N predicted segments identified online by their action a_n and duration l_n , where n refers to the n_{th} segment.

3.2. Offline Inference

Given the input features \mathbf{x}_1^T of the entire video, a common factorized formulation [27, 38] to model the posterior probability of the sequence of actions \mathbf{a}_1^N and their corresponding duration \mathbf{l}_1^N is given by:

$$p_{\text{off}}(\mathbf{a}_1^N, \mathbf{l}_1^N | \mathbf{x}_1^T) \approx p(\mathbf{x}_1^T | \mathbf{a}_1^N) p(\mathbf{l}_1^N | \mathbf{a}_1^N) p(\mathbf{a}_1^N). \quad (1)$$

To infer the most likely sequence of actions $\bar{\mathbf{a}}_1^N$ and their duration $\bar{\mathbf{l}}_1^N$ associated with the video transcript τ , we use

$$(\bar{\mathbf{a}}_1^N, \bar{\mathbf{l}}_1^N) = \underset{\mathbf{a}_1^N, \mathbf{l}_1^N}{\operatorname{argmax}} \{p_{\text{off}}(\mathbf{a}_1^N, \mathbf{l}_1^N | \mathbf{x}_1^T)\}, \quad (2)$$

$$= \underset{\mathbf{a}_1^N, \mathbf{l}_1^N}{\operatorname{argmax}} \left\{ \prod_{t=1}^T p(x_t | a_{n(t)}) \prod_{n=1}^N p(l_n | a_n) p(a_1^N) \right\}, \quad (3)$$

where $n(t)$ is the segment number at frame t . While training, $\bar{\mathbf{a}}_1^N = \tau$ and $N = M$, since the sequence of action labels is already given in the transcript. $p(x_t | a)$ is modeled by a GRU [7] and the Bayes rule as in [38]. The GRU

can be optionally replaced by any other neural network as a black box. $p(l | a)$ is a Poisson distribution modeling the duration of a given action and is parameterized by the mean length of action a . Finally, $p(\mathbf{a}_1^N) = 1$ if the sequence of action labels \mathbf{a}_1^N exist in the training set transcripts and 0 otherwise.

3.3. Offline Segmentation Energy Score

We revisit the definition of *energy score* \mathcal{E} introduced in offline segmentation [27]. Specifically, based on the inferred segments (Eq.3), we define $(\bar{\mathbf{a}}_1^N, \bar{\mathbf{l}}_1^N)$ as the unique valid path π^+ and $(\hat{\mathbf{a}}_1^N, \hat{\mathbf{l}}_1^N)$ as an invalid path $\pi^- \in \mathbb{P}^-$, where $\hat{\mathbf{a}}_n \in \mathbb{A} \setminus \{\bar{a}_n\}$ and \mathbb{P}^- is the set of all invalid paths given the inferred durations $\bar{\mathbf{l}}_1^N$. Accordingly, we define the segment-level energy score of the valid action \bar{a}_n with length \bar{l}_n at segment n as $e_n(\bar{a}_n, \bar{l}_n) = \prod_{t \in \eta(n)}^{\eta(n) + \bar{l}_n - 1} p(\bar{a}_n | x_t)$, and the segment-level energy score of an invalid action \hat{a}_n is given as $e_n(\hat{a}_n, \bar{l}_n) = \prod_{t \in \eta(n)}^{\eta(n) + \bar{l}_n - 1} p(\hat{a}_n | x_t)$. Here, $\eta(n)$ is a function that maps an input segment number to the starting frame number of that segment. Note that the start of each segment occurs immediately after the end of the previous segment and $p(a | x_t)$ is the output of the GRU. Further, in order to exclusively focus on hard invalid actions, the segment energy score of hard invalid actions denoted by $e_n^-(\hat{a}_n, \bar{l}_n)$ is defined as follows:

$$e_n^-(\hat{a}_n, \bar{l}_n) = \begin{cases} e_n(\hat{a}_n, \bar{l}_n), & \text{if } e_n(\hat{a}_n, \bar{l}_n) > e_n(\bar{a}_n, \bar{l}_n) \\ 1, & \text{otherwise} \end{cases}.$$

Finally, $\mathcal{E}_{\pi^+} = \prod_1^N e_n(\bar{a}_n, \bar{l}_n)$ is the energy score of the valid path, and $\mathcal{E}_{\pi^-} = \prod_1^N e_n^-(\hat{a}_n, \bar{l}_n)$ is the energy score of the invalid path π^- . Calculation of these energy scores is done in the log space using DP as explained in [27].

4. Weakly-Supervised Online Segmentation

In this section, we introduce our framework for causal action inference and present how the relation between online and offline inference is exploited to derive a loss function for weakly-supervised online action segmentation.

4.1. Online Inference

Since online action inference is a causal process, we cannot directly use Eq. 3 to infer the action label at the current frame t' . A straightforward causal solution is to employ the GRU in a sliding window fashion and apply $\operatorname{argmax} \{p(a_{t'} | x_t)\}$ as the output of the GRU with the highest probability [16]. However, as shown in Fig.5, this greedy approach does not consider the context and predictions of previous time steps and is therefore sub-optimal. In order to fully account for the past actions and their duration, we formulate the marginal causal (or online) probability $p_{\text{on}}(a_{t'} | \mathbf{x}_1^{t'})$ of the present action $a_{t'} = a_{n(t')}$ at segment

$n' = n(t')$ over all previous actions $\mathbf{a}_1^{n'-1}$ if $n' > 1$, and duration $\mathbf{l}_1^{n'}$. The inferred present action $\hat{a}_{t'}$ is derived as follows:

$$\hat{a}_{t'} = \operatorname{argmax}_{a_{t'} \in \mathbb{A}} \{p_{\text{on}}(a_{t'} | \mathbf{x}_1^{t'})\}, \quad (4)$$

$$= \operatorname{argmax}_{a_{n'} \in \mathbb{A}} \left\{ \sum_{\mathbf{a}_1^{n'-1}, \mathbf{l}_1^{n'}} p_{\text{on}}(\mathbf{a}_1^{n'}, \mathbf{l}_1^{n'} | \mathbf{x}_1^{t'}) \right\}. \quad (5)$$

To improve computational efficiency, we empirically approximated Eq. 5 by the maximum joint probability value:

$$\hat{a}_{t'} \approx \operatorname{argmax}_{a_{n'} \in \mathbb{A}} \left\{ \max_{\mathbf{a}_1^{n'-1}, \mathbf{l}_1^{n'}} p_{\text{on}}(\mathbf{a}_1^{n'}, \mathbf{l}_1^{n'} | \mathbf{x}_1^{t'}) \right\}. \quad (6)$$

Eq.6 involves two steps. The first is to find the most likely sequence of actions $\tilde{\mathbf{a}}_1^{n'}$ with duration $\tilde{\mathbf{l}}_1^{n'}$ until time t' . The second involves taking only the last segment label $\tilde{a}_{n'} = \text{pop}(\tilde{\mathbf{a}}_1^{n'})$ to infer the label of the current frame t' , where $\text{pop}()$ is a function that outputs the last element of a list. To execute the first step, online inference of the most likely sequence of past action segments $(\tilde{\mathbf{a}}_1^{n'}, \tilde{\mathbf{l}}_1^{n'})$ is formulated as $\operatorname{argmax}_{\mathbf{a}_1^{n'}, \mathbf{l}_1^{n'}} \{p_{\text{on}}(\mathbf{a}_1^{n'}, \mathbf{l}_1^{n'} | \mathbf{x}_1^{t'})\}$, where $p_{\text{on}}(\mathbf{a}_1^{n'}, \mathbf{l}_1^{n'} | \mathbf{x}_1^{t'})$ for $n' > 1^*$ is derived below:

$$p_{\text{on}}(\mathbf{a}_1^{n'}, \mathbf{l}_1^{n'} | \mathbf{x}_1^{t'}) = \Gamma(l_{n'} | a_{n'}) \prod_{t=1}^{t'} p(x_t | a_{n(t)}) \prod_{n=1}^{n'-1} p(l_n | a_n) \cdot p(\mathbf{a}_1^{n'}). \quad (7)$$

$p(\mathbf{a}_1^{n'}) = 1$ if $\mathbf{a}_1^{n'}$ is a sub-sequence of any of the transcripts in the training set, and 0 otherwise, and $\Gamma(l|a)$ is a half Poisson function to model the duration $l_{n'}$ of the current action $a_{n'}$ at the last observed segment, given by

$$\Gamma(l|a) = \begin{cases} 1 & \text{if } l < \lambda_a \\ \frac{\lambda_a^l \exp(-\lambda_a)}{l!} & \text{otherwise} \end{cases},$$

where λ_a is the estimated mean length of action a .

Inclusion of $\Gamma()$ in the online inference of the current action is essential as it accounts for the two following cases: First, using the full Poisson distribution of Eq.3 to model the duration of the current observed action leads to penalizing the current actions with a short duration, $l_{n'} < \lambda_{a_{n'}}$. However, since we do not have foresight about the duration of the current segment, any conclusion about the current segment length would be premature. Second, $\Gamma()$ still allows us to penalize the current action if its duration is longer than expected since this can be concluded solely based on the observed segment of the action.

At test time, the final online segmentation result in a streaming video when the current time t' changes from 1 to any given time T is the sequence of frame-level actions $(\hat{a}_1, \dots, \hat{a}_T)$, where each $\hat{a}_{t'} \leftarrow \tilde{a}_{n'} = \text{pop}(\tilde{\mathbf{a}}_1^{n'})$ is inferred by Eq.7 using the Viterbi algorithm.

*For $n' = 1$, the Poisson factor $p(l|a)$ is excluded.

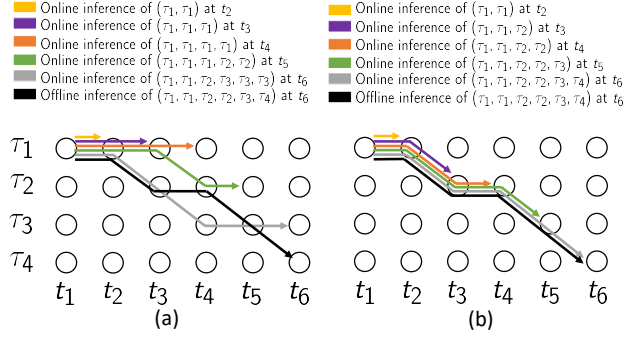


Figure 2. Given the video transcript $\tau=(\tau_1, \tau_2, \tau_3, \tau_4)$, the OODL loss encourages the online segmentation results in (a) to become a sub-sequence of the offline result for each time step as in (b).

4.2. Online-Offline Discrepancy Loss (OODL)

Offline action segmentation is expected to be more accurate than online segmentation because segments are inferred from information contained in the entire length of the video, including transcripts as well as prior knowledge of the video end. Thus, offline segmentation results provide a rich source of supervision for training online segmentation models. Ideally, the sequence of actions inferred online from the initial frame to any point in the video is expected to be a sub-sequence of the offline inference result as shown in Fig. 2. Consequently, this encourages all frame-level action sequences $\{\tilde{\mathbf{a}}_1^t\}_{t=1}^T$ to be temporally consistent, where each sequence $\tilde{\mathbf{a}}_1^t$ is inferred online at time t .

We present the Online-Offline Discrepancy Loss (OODL) $\mathcal{L}_{\text{OODL}}$ in Algorithm 1 to minimize the difference between online and offline segmentation scores. Specifically, we first use Eq.7 to infer the set of online paths $\{\tilde{\mathbf{a}}_1^t\}_{t=1}^T$ after $n(t)$ pairs of segment-level labels $(\tilde{\mathbf{a}}_1^{n(t)}, \tilde{\mathbf{l}}_1^{n(t)})$ are converted into t frame-level action labels $\tilde{\mathbf{a}}_1^t$ for each time step t . Then, we use the hinge loss function to penalize any online inference result that has a higher energy score \mathcal{E}_{on} than the energy score \mathcal{E}_{off} of the offline path $\bar{\mathbf{a}}_1^t \subseteq \bar{\mathbf{a}}_1^T$ inferred from Eq. 3. The OODL ultimately discourages all frame-level predictions that contribute to the discrepancy between the intermediate online inference results and the most likely sequence of actions inferred offline at the end of the video.

$\mathcal{L}_{\text{OODL}}$ is added to the baseline offline segmentation loss \mathcal{L}_b [27] to form our final loss function \mathcal{L}_f :

$$\mathcal{L}_f = \mathcal{L}_b + \mathcal{L}_{\text{OODL}}. \quad (8)$$

Minimizing the offline segmentation loss \mathcal{L}_b effectively corresponds to maximizing the decision margin between offline valid and hard invalid paths derived in Sec. 3.3.

$$\mathcal{L}_b = -\log(\mathcal{E}_{\pi^+}) + \log\left(\sum_{\pi^- \in \mathbb{P}^-} \mathcal{E}_{\pi^-}\right). \quad (9)$$

We iteratively utilize the offline and online segmentation pseudo labels inferred by Eq. 3, and 7, respectively, as well as the loss in Eq. 8 to train the GRU until convergence.

Algorithm 1 Online-Offline Discrepancy Loss (OODL)

Input: Video features \mathbf{x}_1^T of T frames and the offline inference result $(\bar{\mathbf{a}}_1^N, \bar{\mathbf{l}}_1^N)$ of N segments
Output: L_T as the OODL loss $\mathcal{L}_{\text{OODL}}$

- 1: **for** $t \leftarrow 1$ to T **do**:
- 2: $\tilde{\mathbf{a}}_1^{n(t)}, \tilde{\mathbf{l}}_1^{n(t)} = \text{argmax} \{p_{\text{on}}(\mathbf{a}_1^{n(t)}, \mathbf{l}_1^{n(t)} | \mathbf{x}_1^t)\} \triangleright \text{Eq. 7}$
- 3: $\tilde{\mathbf{a}}_1^t = (\tilde{a}_1, \dots, \tilde{a}_{n(t)}) = \text{convert}(\tilde{\mathbf{a}}_1^{n(t)}, \tilde{\mathbf{l}}_1^{n(t)})$
- 4: $\mathcal{E}_{\text{on}}(t) = \prod_{a_t \in \tilde{\mathbf{a}}_1^t} p(a_t | x_t)$
- 5: $\bar{\mathbf{a}}_1^T = (\bar{a}_1, \dots, \bar{a}_{n(T)}) = \text{convert}(\bar{\mathbf{a}}_1^N, \bar{\mathbf{l}}_1^N)$
- 6: $L_0 = 0$
- 7: **for** $t \leftarrow 1$ to T **do**:
- 8: $\mathcal{E}_{\text{off}}(t) = \prod_{a_t \in \bar{\mathbf{a}}_1^t} p(a_t | x_t)$
- 9: $d = \max(0, \log(\mathcal{E}_{\text{on}}(t)) - \log(\mathcal{E}_{\text{off}}(t)))$
- 10: $L_t = L_{t-1} + \frac{d}{t} \triangleright \text{Averaging } d \text{ over time } t$

return L_T

5. Multi-View Supervision

Due to lack of frame-level action labels at training time, it is imperative to maximize the functional capacity of the training data available. We do so by leveraging the correspondence between multiple unknown views to infer more accurate frame pseudo labels.

Concretely, consider a training set of K videos $\{v_i\}_{i=1}^K$ and their corresponding view adjacency matrix $V \in \mathbb{R}^{K \times K}$, where each element $v_{i,j}$ in V is 1 if v_i and v_j are two different views of the same recording, and 0 otherwise. During training, we pair each video v_i , as the anchor video, with an auxiliary video v_j , which is randomly sampled from the anchor view's adjacent set $\mathbb{V}_i = \{v_k | V_{i,k} = 1 \wedge k \neq i\}$. As shown in Fig. 3, each video pair is given as an input to a multiview-inference module to generate pseudo labels[†], which are used to train the GRU with respect to the anchor video i . In this section, we discuss three different multi-view inference techniques employed during training:

Sequence Voting (SV). Given synchronized video features ${}_i\mathbf{x}_1^T$ and ${}_j\mathbf{x}_1^T$ of any two given views, we define the result of voting as the sequence of actions $\bar{\mathbf{a}}_1^N$ with durations $\bar{\mathbf{l}}_1^N$ that have the highest product of sequence probability over both views:

$$(\bar{\mathbf{a}}_1^N, \bar{\mathbf{l}}_1^N) = \text{argmax}_{\mathbf{a}_1^N, \mathbf{l}_1^N} \{p(\mathbf{a}_1^N, \mathbf{l}_1^N | {}_i\mathbf{x}_1^T) p(\mathbf{a}_1^N, \mathbf{l}_1^N | {}_j\mathbf{x}_1^T)\}. \quad (10)$$

[†]Interchangeably named as offline valid path or offline inference result

In this case, the inferred sequence must have high probabilities (votes) in both views, as inconsistent probabilities (votes) diminish the overall score of any segmentation.

Probabilistic Inference (PI). Instead of combining multi-view results at the video level as in the SV technique, here we fuse frame-level scores to infer the sequence that maximizes the posterior probability $p(\mathbf{a}_1^N, \mathbf{l}_1^N | {}_i\mathbf{x}_1^T, {}_j\mathbf{x}_1^T)$ given the two views:

$$p(\mathbf{a}_1^N, \mathbf{l}_1^N | {}_i\mathbf{x}_1^T, {}_j\mathbf{x}_1^T) \approx p({}_i\mathbf{x}_1^T | \mathbf{a}_1^N) p({}_j\mathbf{x}_1^T | \mathbf{a}_1^N) p(\mathbf{l}_1^N | \mathbf{a}_1^N) p(\mathbf{a}_1^N). \quad (11)$$

The argmax of the above equation can be solved by integrating $p(x_t | a_{n(t)}) = p({}_i x_t | a_{n(t)}) p({}_j x_t | a_{n(t)})$ in Eq. 3.

Weighted Probabilistic Inference (WPI). The Probabilistic Inference model in Eq. 11 assumes equal contribution from each view. However, a more appropriate formulation is to compare the two views and provide a higher confidence weight on the more reliable view. Hence, we introduce the class agnostic confidence weight ${}_i c_t \in [0, 1]$ for the anchor view i at time t as follows:

$${}_i c_t, 1 - {}_i c_t = \text{Softmax} \left(\Phi_c([\Phi_f({}_i \mathbf{x}_{t-w}^t) \Phi_f({}_j \mathbf{x}_{t-w}^t)]) \right), \quad (12)$$

where $\Phi_f() : \mathbb{R}^{w \times F_1} \rightarrow \mathbb{R}^{F_2}$ is a function that embeds a temporal window of the past w frame features \mathbf{x}_{t-w}^t for each view independently, and $\Phi_c() : \mathbb{R}^{F_2} \rightarrow \mathbb{R}^2$ is the compare function that takes the concatenated view embeddings $[\Phi_f({}_i \mathbf{x}_{t-w}^t) \Phi_f({}_j \mathbf{x}_{t-w}^t)]$ and outputs the relative confidence weight of the anchor view i with respect to the auxiliary view j at time t . F_1 and F_2 are the dimensions of each frame feature and the window embedding respectively.

Having defined the view confidence weight ${}_i c_t$, we rewrite the likelihood $p(x_t | a_{n(t)})$ as follows and use Eq. 3 to infer the pseudo labels $(\bar{\mathbf{a}}_1^N, \bar{\mathbf{l}}_1^N)$:

$$p(x_t | a_{n(t)}) = p({}_i x_t | a_{n(t)})^{{}_i c_t} p({}_j x_t | a_{n(t)})^{(1 - {}_i c_t)}. \quad (13)$$

We incorporate a new loss \mathcal{L}_{vc} in our final loss function \mathcal{L}_f in order to learn the parameters of the view confidence weight ${}_i c_t(\theta_c)$, where θ_c is the set of all parameters in the compare and embedding functions of Eq. 12. In addition, θ_a denotes the set of parameters (i.e. GRU) required to predict frame-level action probability $p(a_t | x_t; \theta_a)$. Given the inferred pseudo labels $(\bar{\mathbf{a}}_1^N, \bar{\mathbf{l}}_1^N)$, we define the weighted energy score of the pseudo labels as :

$$\tilde{\mathcal{E}}_{\theta_c} = \prod_{t=1}^T p(\bar{a}_{n(t)} | {}_i x_t)^{{}_i c_t(\theta_c)} p(\bar{a}_{n(t)} | {}_j x_t)^{(1 - {}_i c_t(\theta_c))}, \quad (14)$$

where we freeze θ_a and allow $\mathcal{L}_{\text{vc}}(\theta_c) = -\log(\tilde{\mathcal{E}}_{\theta_c})$ to be optimized with respect to the view confidence weight ${}_i c_t(\theta_c)$, so that the weighted energy score $\tilde{\mathcal{E}}_{\theta_c}$ of the correct path $(\bar{\mathbf{a}}_1^N, \bar{\mathbf{l}}_1^N)$ is maximized:

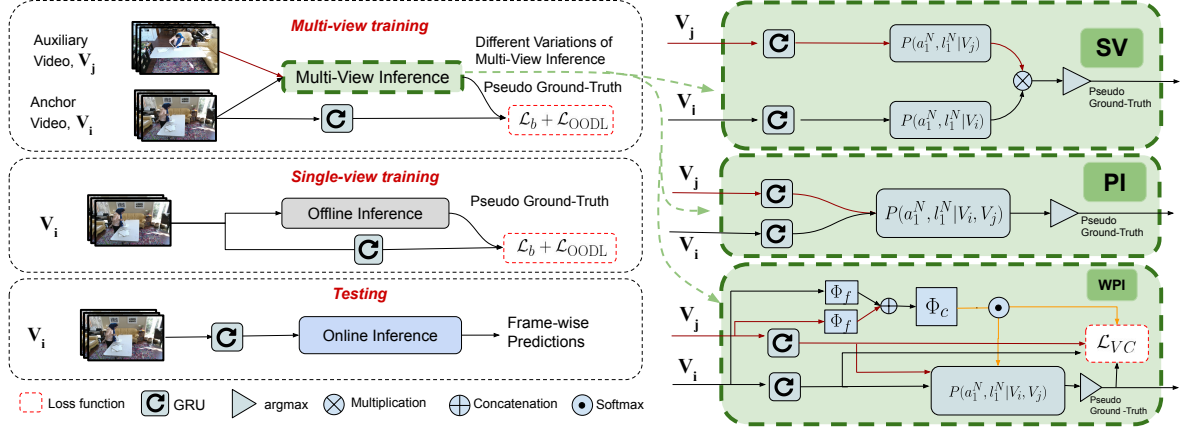


Figure 3. An overview of the single-view and multi-view training schemes can be seen on the left. More details of the three proposed multi-view inference techniques are depicted on the right. Notice how a single view is always used to segment the video at test time.

$$\mathcal{L}_f(\theta_c, \theta_a) = \mathcal{L}_b(\theta_a) + \mathcal{L}_{OODL}(\theta_a) + \mathcal{L}_{vc}(\theta_c). \quad (15)$$

Note that the embedding and compare functions, $\Phi_f()$ and $\Phi_c()$, are utilized only in training. Besides, \mathcal{L}_b and \mathcal{L}_{OODL} are computed using just the anchor video after taking the multi-view inference pseudo labels as their valid path and offline inference result respectively.

6. Experiments

Datasets. The *Breakfast Dataset (BD)* [24] contains approximately 1.7k cooking videos, recorded from multiple views, ranging from a few seconds to over ten minutes long. Both the angle and number of views differ across recordings. The dataset consists of 48 action labels demonstrating 10 breakfast dishes with a mean of 6.9 action segments per video. The evaluation metrics are calculated over four splits. The *IKEA ASM Dataset (IAD)* [2] has 371 recordings of assembly for four types of furniture. Each assembly is recorded from three consistent view points, providing 1113 total videos. Videos in this dataset contain a dense number of action segments (mean of ≈ 23) per shorter videos with a mean duration of 1.9 min. There are 32 action classes after combining the *NA* and *other* classes as *background*. We report results over 5 splits, where each split belongs to one of the five recording environments as suggested by [2].

Metrics. Similar to previous work [6, 9, 27], we use four metrics to evaluate performance: 1) *acc* is the frame-level accuracy averaged over all the videos. 2) *acc-bg* is the frame-level accuracy without the background frames. 3) *IoU* defined as the intersection over union, which is particularly useful for imbalanced datasets such as *IKEA*. 4) *IoD* denotes the intersection interval over the detected interval averaged across all videos. This metric tends to overrate over-segmentation results. As in [9], both *IoD* and *IoU* are calculated over non-background segments.

Implementation. We extracted I3D features [3] for the *IAD* dataset using TV-L1 optical flow [57] on a moving window of 16 frames. Final dimensions of the features were reduced to 400 by PCA. Meanwhile, for the *BD* dataset, we obtained the Fisher vectors [34] of iDT features [51] as in [25]. We implemented the embedding function $\Phi_f()$ as temporal convolution and max pooling, while two fully connected layers were used as the compare function $\Phi_c()$. Also, we set $F_2 = 64$ and $\omega = 21$. For a fair comparison we used the same random seed in all our experiments. The model was trained for around 70k and 6k iterations on the *BD* and *IAD* datasets, respectively, following the training setup of [27].

6.1. Comparison to the Baseline Methods

Baselines. We implemented the Greedy baseline following the strategy of [16], where a recurrent network is trained using the pseudo labels generated by an offline segmentation method. At test time, the network takes a greedy approach and identifies actions in a sliding window fashion. Also, DP_{on} represents the proposed online inference (Eq.7), and DP_{off} denotes the offline segmentation baseline of Eq.3.

Quantitative Results. Table 1 compares the Greedy and DP_{on} methods in online segmentation. The Greedy baseline shows poor performance specially in the *BD* dataset largely due to poor video quality that makes isolated predictions error-prone. However, Greedy shows decisively high *IoD* values. In general, high *IoD* with low *IoU* indicates over-segmentation, which leads to overrating the result. The presence of the Γ function in our online modeling is important. Its omission leads to about 1% and 3% drop in all metrics in the *BD* and *IAD* datasets, respectively. The best result is achieved by including the OODL loss and multi-view training. This leads to a 2.6% and 3.3% *IoU* improvement of the DP_{on} baseline in the *BD* and *IAD* datasets, respectively. Overall, improvements on the *IAD* dataset are

Table 1. Comparison of our multi-view supervised segmentation model with various baselines in online action segmentation. M refers to multi-view training. * We report the WPI and PI multi-view results for the *BD* and *IAD* datasets respectively.

Training	Test		Breakfast (%)				IKEA ASM (%)			
M	\mathcal{L}_{OO}	Inference	acc	acc-bg	IoU	IoD	acc	acc-bg	IoU	IoD
×	×	Greedy [16]	20.4	15.9	7.4	58.1	55.6	56.2	30.9	53.5
×	×	DP _{on} w/o Γ	34.3	31.4	21.4	45.1	52.8	54.6	30.0	39.3
×	×	DP _{on}	35.1	32.3	22.4	46.9	55.3	57.8	33.3	44.1
✓	✓	DP _{on} [*]	36.6	34.7	25.0	49.1	56.9	59.7	36.8	48.0
✓	✓	DP _{off} [*]	50.4	46.8	33.3	44.9	60.3	63.5	41.7	52.0

better represented by *IoU* since frame accuracy is dominated by the action “*spin leg*”, which occupies nearly 45% of the frames. We include offline results to show the performance gap between online and offline segmentation. The smaller gap between DP_{off} and the Greedy method in the *IAD* dataset highlights the challenge of weakly-supervised learning in videos with a dense number of action segments.

Qualitative Results. Incorporating multi-view supervision in training makes the GRU more robust to bad lighting, occlusion and scene variations as demonstrated in Fig. 4. Particularly, the top figure shows results of coffee table assembly by two people in the *IAD* dataset. This is a challenging case since nearly all tasks in both datasets are completed by one person. Hence, the baseline DP_{on} has missed the third instance of “*spinning the leg*”, which is correctly detected by our final segmentation model trained using \mathcal{L}_{OO} and multi-view inference. The bottom figure compares different segmentation methods under dark lighting and occlusion in a sample cooking video of the *BD* dataset. Notice the over-segmented results of the Greedy baseline in both cases. More examples included in the supplementary material.

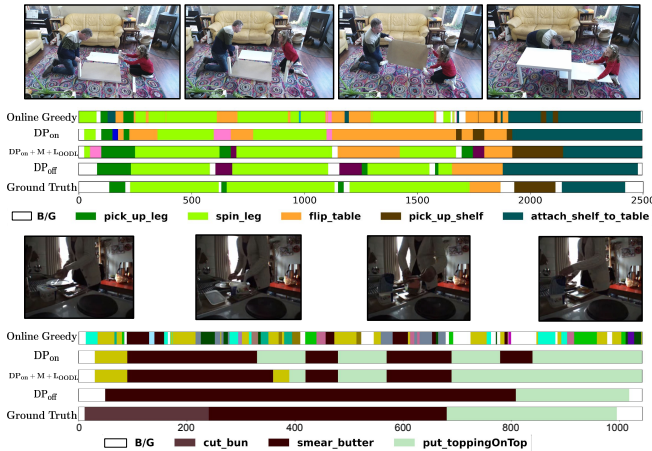


Figure 4. Segmentation results of various methods on the IKEA (top) and Breakfast (bottom) datasets. Legend is shown only for the ground-truth classes.

Table 2. Impact of the OODL loss on weakly-supervised online segmentation results for the *BD* and *IAD* datasets.

Training Approach	Breakfast (%)				IKEA ASM (%)			
	acc	acc-bg	IoU	IoD	acc	acc-bg	IoU	IoD
DP _{on}	35.1	32.3	22.4	46.9	55.3	57.8	33.3	44.1
DP _{on} + \mathcal{L}_{OODL}	35.5	32.5	23.4	48.0	55.3	57.9	34.3	45.5

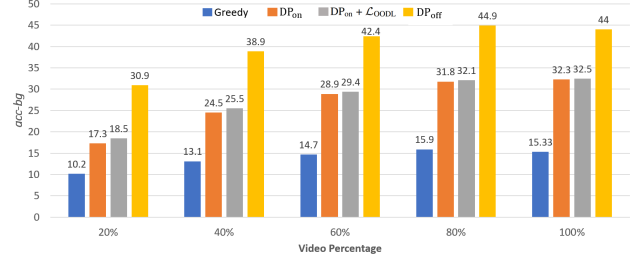


Figure 5. Average segmentation results (*acc-bg*) at five observation end points during the course of the video on the *BD* dataset.

rial.

6.2. Analysis and Ablation Study

All experiments in this section are reported as an average over all splits unless stated otherwise. Run-time and complexity, as limitations of the proposed algorithm, are discussed in the supplementary material.

6.2.1 Online-Offline Discrepancy Analysis

Impact of the OODL Loss. Addition of OODL loss leads to consistent improvement in both datasets as shown in Table 2. This improvement is manifested more vividly in *IoU* and *IoD* because *IoU*, in particular, is most appropriate in evaluating alignment quality between predicted and ground-truth segments.

Fig. 5 further demonstrates the role of the OODL loss in decreasing the online-offline segmentation discrepancy in the *BD* dataset. It shows the non-background frame accuracy of multiple segmentation approaches at five different observation end points in the video. Upon comparison of DP_{on} and DP_{on} + \mathcal{L}_{OODL} , it can be seen that the loss has improved mostly the early predictions in the video, where it is hardest to identify actions. This is mainly due to lack of past context in early stages of a task. With the passage of time, more information regarding the past actions becomes available. Consequently, this leads to more accurate online predictions of the current frame. On average, after observing the first 20% of the video, the performance of the DP_{on} is more similar to the Greedy baseline than the Offline model. However, the DP_{on} approach starts resembling the Offline model more just after the 60% point. In comparison, such a behavior is highlighted much less by the Greedy approach due to its limitation in capturing long-range past context.

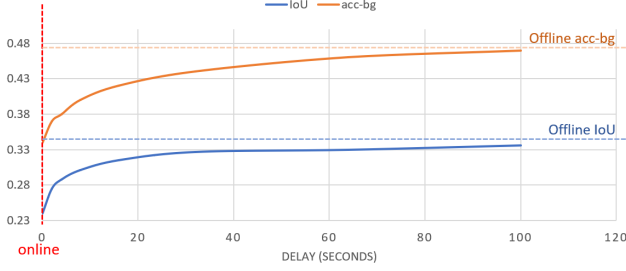


Figure 6. Accuracy vs. delay on split 2 of the *BD* dataset.

Evaluation of Semi-Online Segmentation. Online segmentation offers practical advantages over offline inference in interactive applications that require immediate feedback. However, this comes with a 10% and 13% compromise over *acc-bg* and *IoU*, respectively, as indicated in Fig. 6. In some applications, certain degree of latency can be tolerated. In order to evaluate the trade-off between latency and accuracy, we implemented a semi-online variation of our framework, where predictions are made after a fixed time delay. Fig. 6 shows that accuracy improves with larger latency and converges to the offline result. Importantly, we can achieve approximately 90% of the offline performance with 10 seconds of delay on the *BD* dataset.

6.2.2 Multi-View Supervision

Weakly-Supervised Online Segmentation. We evaluate the online segmentation performance of different multi-view inference techniques in Table 3. Regardless of the approach, using multi-view correspondence to generate pseudo ground-truth improves performance over the single-view method in all metrics and datasets. We also provide the *Fully-Supervised* baseline as the upper bound, where the pseudo ground-truth is 100% accurate.

Table 3. Comparison of online segmentation results under different pseudo ground-truth generation techniques (all with \mathcal{L}_{ODL}).

Training Approach	Breakfast (%)				IKEA ASM (%)			
	acc	acc-bg	IoU	IoD	acc	acc-bg	IoU	IoD
Single-View	35.5	32.5	23.4	48.0	55.3	57.9	34.3	45.5
SV	36.4	34.7	24.8	48.6	55.7	58.3	34.6	45.9
PI	36.2	34.2	24.4	48.1	56.9	59.7	36.8	48.0
WPI	36.6	34.7	25.0	49.1	56.4	59.0	35.9	47.2
<i>Fully-Supervised</i>	41.6	41.2	30.4	52.9	63.5	67.36	44.5	56.9

Table 4. Offline segmentation results with and without multi-view supervision (same multi-view approaches as in Table 1). [27] results obtained after running the authors’ code on our machine.

Model	Breakfast (%)				IKEA ASM (%)			
	acc	acc-bg	IoU	IoD	acc	acc-bg	IoU	IoD
CDFL [27]	49.2	44.2	31.0	43.7	59.9	62.0	39.5	50.4
Multi-View CDFL	50.4	46.8	33.3	44.9	60.3	63.5	41.7	52.0



Figure 7. Pseudo ground-truth generated by 3 multi-view inference techniques during training. The red bar above each single-view inference result (anchor/auxiliary) indicates their learned view confidence weight c_t at each frame. Each pair of frames corresponds to the time enclosed by its color-coded dashed box.

In the *BD* dataset, different views provide more complementary information as compared to the *IAD* dataset. This is attributed to many instances of challenging lighting or occlusion conditions in the *BD* dataset. Fig. 7 underscores this fact by showing the pseudo ground-truth generated during training. Specifically, the bread in the *anchor* view occludes the two actions of “cutting bread” and “smearing butter”. The resulting view confidence weight (red bar) of the *auxiliary* view becomes high for these frames. This allows the model to exploit the more visible view in the *auxiliary* video during these actions, while the *anchor* view is recognized as more reliable when the subject “takes an object”. Notice how the view confidence weights dictate the selection of single-view results to form the multi-view WPI outcome. Meanwhile, the three views of the *IAD* dataset remain fairly similar in terms of lighting and no considerable occlusion occurs in most videos. Hence, weighing views equally in the PI approach leads to the best results in the *IAD* dataset.

Weakly-Supervised Offline Segmentation. Table 4 shows how the advantage of multi-view training is further generalized to weakly-supervised “offline” segmentation. We selected the open source state-of-the-art offline segmentation method [27] as our baseline. For this experiment, we trained [27] twice, with and without multi-view supervision, under the same parameters and random seed.

7. Conclusion

We introduced a framework to address a new problem of weakly supervised online action segmentation in multi-view instructional videos. The proposed solutions are formulated with the insight that offline and multi-view results provide a rich source of supervision during training which in-turn improves performance of single view online segmentation models at test time. Extensive experiments on two benchmark datasets demonstrate efficacy of our algorithms.

8. Supplementary Material

8.1. Overview

In this supplementary material we first provide the definitions of all the terms used in the paper, explain complexity as a limitation, and then show and discuss more qualitative segmentation results of different methods.

8.2. Glossary of Symbols

We provide specific definitions of symbols in Table 5 for readers to refer to.

Table 5. Definitions of symbols used in the paper.

Symbol	Definition
\mathbb{A}	The set of all actions in the dataset
a_n	Action variable at segment n
a_t	Action variable at frame t
\hat{a}_t	Predicted action at time t in an online way
$(\bar{a}_1^N, \bar{l}_1^N)$	Offline inference result
$(\bar{a}_1^{n(t)}, \bar{l}_1^{n(t)})$	Online inference result until time t
$i c_t$	View confidence weight of video i (anchor) at time t
e_n	Energy score of segment n
\mathcal{E}_{π^+}	Energy score of the valid path π^+
\mathcal{E}_{π^-}	Energy score of the invalid path π^-
$\mathcal{E}_{\text{off}}(t)$	Energy score of the offline or valid path until time t
$\mathcal{E}_{\text{on}}(t)$	Energy score of the online path until time t
$\bar{\mathcal{E}}$	Weighted energy score of a path
F_1	Input feature dimension
F_2	Embedding dimension
K	Total number of videos in the data set
l_n	Duration variable of action a_n
\mathcal{L}_f	Final loss
\mathcal{L}_b	Baseline offline segmentation loss
\mathcal{L}_{vc}	View confidence loss to train WPI
M	Number of action labels in the transcript
N	Number of predicted segments in the video
$p_{\text{on}}()$	Causal probability
$p_{\text{off}}()$	Non-causal probability
\mathbb{P}^-	The set of all invalid paths
T	Total number of frames in the video
τ	Video transcript
v_i	Video i
\mathbb{V}_i	View adjacency set of video i
V	$K \times K$ view adjacency matrix
ω	Feature window size for Φ_f
\mathbf{x}_1^T	Sequence of T frame features
$i \mathbf{x}_1^T$	Features of video i
$\eta(n)$	Mapping function from segment to frame number
$\Gamma()$	Half Poisson function
λ_a	Estimated mean length of action a
Φ_c	Compare function
Φ_f	Feature embedding function
θ_c	Parameters of Φ_f and Φ_c
θ_a	Parameters of the action classifier, i.e. GRU
π^+	Valid path or offline segmentation action sequence
π^-	Invalid path

8.3. Limitation

Here we discuss the practical run-time and computational complexity of our method, both in test and training, and compare it with CDFL [27] as an offline baseline. The lower frame rate compared to a greedy approach and the lengthy training-time are notable limitations of the proposed online segmentation method. We hope further work can mitigate this limitation.

8.3.1 Runtime Frame Rate Analysis

Both our method and the greedy approach [16] compute optical flow (OF) and use the I3D network to extract features. We extracted features on 320×240 frames of the *BD* dataset recorded at 15 fps. On a single GeForce GTX 1080, the OF and I3D network process videos at 90 and 20 fps, respectively. Practically, if online inference is done every 15 frames, then our method segments videos at 10+ fps. While this is less than the 100 fps of the greedy alg., it leads to considerably more accurate results.

8.3.2 Computation Complexity of Online vs. Offline

The complexity of the proposed online inference to fully segment a video of length T and maximum transcript length of N is the same as the offline inference of CDFL ($O(T^2N)$). In other words, online inference at each time step takes $O(TN)$. This is due to DP as the inference at time t depends on the optimal results of previous time steps which have already been obtained as part of DP.

With this in mind, the training complexity of our method over K classes is the sum of complexities for the offline inference ($O(T^2N)$), baseline offline segmentation loss \mathcal{L}_b ($O(\Delta^2NK)$) and \mathcal{L}_{ODL} . $\Delta \ll T$ is a small window size of 10 [27]. A naive implementation of OODL has complexity of $O(T^2N)$. However, if the online and offline inferences are done together outside Alg.1 and $\mathcal{E}(t)$ is summed over segments rather than frames, the OODL complexity becomes $O(TN)$. Hence, regardless of the implementation choice, our overall training has the same complexity as CDFL ($O(T^2N + \Delta^2NK)$). Our time complexity during test time with M training transcripts is $O(T^2NM)$ which is also the same as that of CDFL. In order to quantitatively support our calculations, we tested both methods on the 4th split of the *BD* dataset. Our method and CDFL took 26 and 21 hrs to train, respectively. Meanwhile at test time, ours and CDFL took 2.7 and 2.4 hrs to run, respectively.

8.4. Qualitative Results

In Figure 8, we present two segmentation examples on the IKEA [2] (top) and Breakfast [24] (bottom) datasets. We demonstrate how training using multiple view points

has let to more robust segmentation results against full occlusion (top) and extremely bad lighting (bottom). Specifically, the top figure depicts a task where the subject assembles a “side table”. This assembly consists of four instances of “spinning leg”, where the last instance is fully occluded by the subject’s body. The baseline method DP_{on} , that is trained on a single viewpoint, misses most of the action, while training on multi-view correspondence and the OODL loss has enabled our final model ($DP_{on} + M + \mathcal{L}_{ODL}$) to capture nearly the full segment.

In the second example, the dark lighting makes it even hard for a human observer to recognize the ongoing action. Our final method is able to identify the action of “adding tea bag”, where both the offline method DP_{off} and online baselines DP_{on} fail. This is an interesting case, where our model is able to outperform even the offline method. One reason is the flexibility of the proposed online segmentation model in switching between different transcripts in a series of online inferences across different time steps. This allows the predicted sequence of actions to potentially come from a transcript not observed at training time. In contrast, in offline segmentation the sequence of inferred action labels is limited only to the training transcripts.

References

- [1] Yazan Abu Farha and Juergen Gall. Uncertainty-aware anticipation of activities. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019. 2
- [2] Yizhak Ben-Shabat, Xin Yu, Fatemeh Saleh, Dylan Campbell, Cristian Rodriguez-Opazo, Hongdong Li, and Stephen Gould. The ikea asm dataset: Understanding people assembling furniture through actions, objects and pose. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 847–859, 2021. 2, 6, 9
- [3] 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. 6
- [4] Chien-Yi Chang, De-An Huang, Yanan Sui, Li Fei-Fei, and Juan Carlos Niebles. D3tw: Discriminative differentiable dynamic time warping for weakly supervised action alignment and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3546–3555, 2019. 2
- [5] Chien-Yi Chang, De-An Huang, Danfei Xu, Ehsan Adeli, Li Fei-Fei, and Juan Carlos Niebles. Procedure planning in instructional videos. In *European Conference on Computer Vision*, pages 334–350. Springer, 2020. 1
- [6] Xiaobin Chang, Frederick Tung, and Greg Mori. Learning discriminative prototypes with dynamic time warping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8395–8404, 2021. 1, 2, 6
- [7] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014. 3
- [8] K Deepak, G Srivathsan, S Roshan, and S Chandrakala. Deep multi-view representation learning for video anomaly detection using spatiotemporal autoencoders. *Circuits, Systems, and Signal Processing*, 40(3):1333–1349, 2021. 2
- [9] Li Ding and Chenliang Xu. Weakly-supervised action segmentation with iterative soft boundary assignment. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6508–6516, 2018. 2, 6
- [10] Hyunjun Eun, Jinyoung Moon, Jongyoul Park, Chanho Jung, and Changick Kim. Learning to discriminate information for online action detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 809–818, 2020. 2
- [11] Chenyou Fan, Jangwon Lee, Mingze Xu, Krishna Kumar Singh, Yong Jae Lee, David J Crandall, and Michael S Ryoo. Identifying first-person camera wearers in third-person videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5125–5133, 2017. 2
- [12] Yazan Abu Farha and Jurgen Gall. Ms-tcn: Multi-stage temporal convolutional network for action segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3575–3584, 2019. 2

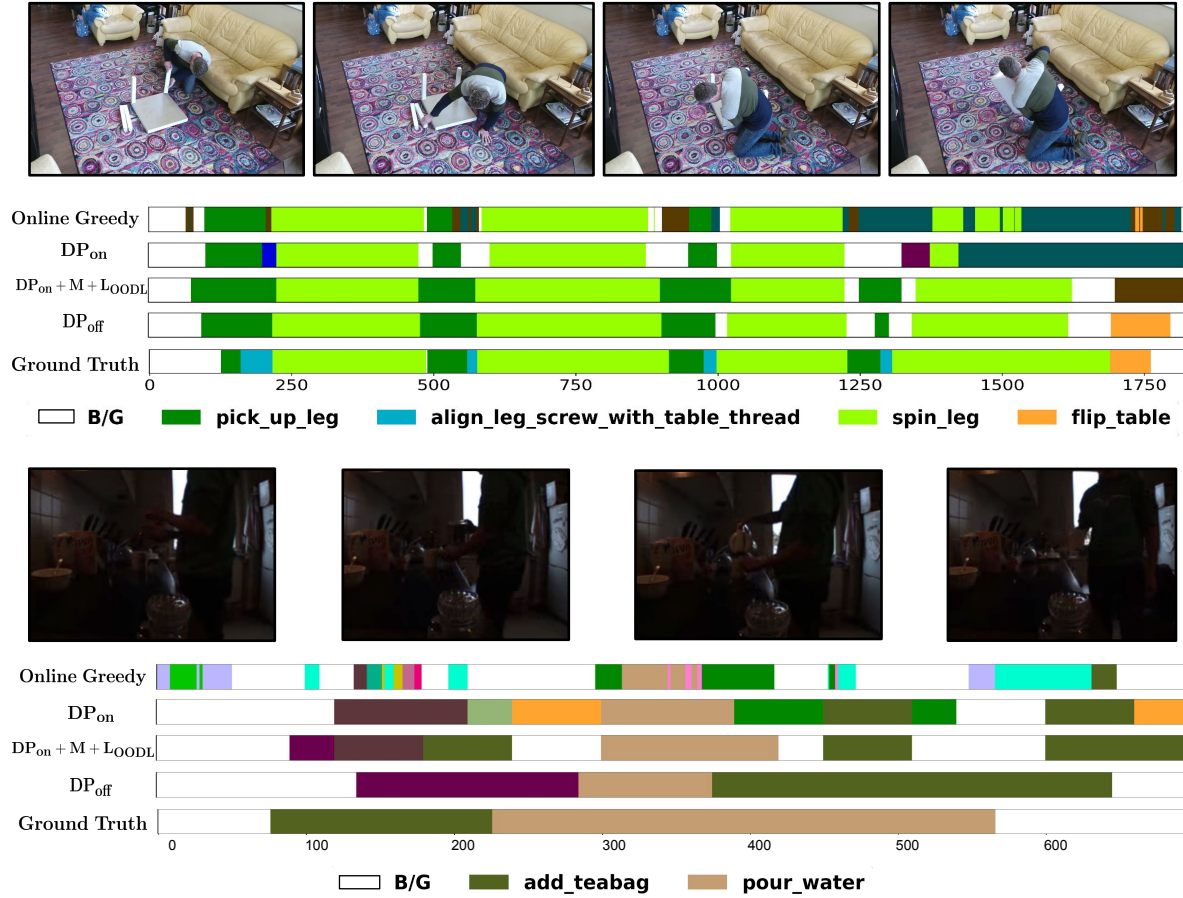


Figure 8. This figure shows segmentation results of various methods on the IKEA (top) and Breakfast (bottom) datasets. Subjects in the top and bottom figures assemble a side table and prepare tea respectively. Legend is shown only for the ground-truth classes.

- [13] Antonino Furnari and Giovanni Maria Farinella. What would you expect? anticipating egocentric actions with rolling-unrolling lstms and modality attention. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6252–6261, 2019. 2
- [14] Jiyang Gao, Zhenheng Yang, and Ram Nevatia. Red: Reinforced encoder-decoder networks for action anticipation. *arXiv preprint arXiv:1707.04818*, 2017. 2
- [15] Mingfei Gao, Mingze Xu, Larry S Davis, Richard Socher, and Caiming Xiong. Startnet: Online detection of action start in untrimmed videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5542–5551, 2019. 2
- [16] Mingfei Gao, Yingbo Zhou, Ran Xu, Richard Socher, and Caiming Xiong. Woad: Weakly supervised online action detection in untrimmed videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1915–1923, 2021. 2, 3, 6, 7, 9
- [17] Shang-Hua Gao, Qi Han, Zhong-Yu Li, Pai Peng, Liang Wang, and Ming-Ming Cheng. Global2local: Efficient structure search for video action segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16805–16814, 2021. 2
- [18] Reza Ghoddoosian, Saif Sayed, and Vassilis Athitsos. Action duration prediction for segment-level alignment of weakly-labeled videos. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2053–2062, 2021. 2
- [19] Reza Ghoddoosian, Saif Sayed, and Vassilis Athitsos. Hierarchical modeling for task recognition and action segmentation in weakly-labeled instructional videos. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1922–1932, January 2022. 1
- [20] Sanjay Haresh, Sateesh Kumar, Huseyin Coskun, Shahram N Syed, Andrey Konin, Zeeshan Zia, and Quoc-Huy Tran. Learning by aligning videos in time. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5548–5558, 2021. 2
- [21] Hsuan-I Ho, Wei-Chen Chiu, and Yu-Chiang Frank Wang. Summarizing first-person videos from third persons’ points of view. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 70–85, 2018. 2
- [22] Yuchi Ishikawa, Seito Kasai, Yoshimitsu Aoki, and Hirokatsu Kataoka. Alleviating over-segmentation errors by detecting action boundaries. In *Proceedings of the*

- IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2322–2331, 2021. 2
- [23] Qihong Ke, Mario Fritz, and Bernt Schiele. Time-conditioned action anticipation in one shot. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9925–9934, 2019. 2
- [24] Hilde Kuehne, Ali Arslan, and Thomas Serre. The language of actions: Recovering the syntax and semantics of goal-directed human activities. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 780–787, 2014. 2, 6, 9
- [25] Hilde Kuehne, Ali Arslan, and Thomas Serre. The language of actions: Recovering the syntax and semantics of goal-directed human activities. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 780–787, 2014. 6
- [26] Sateesh Kumar, Sanjay Haresh, Awais Ahmed, Andrey Konin, M Zeeshan Zia, and Quoc-Huy Tran. Unsupervised activity segmentation by joint representation learning and online clustering. *arXiv preprint arXiv:2105.13353*, 2021. 2
- [27] Jun Li, Peng Lei, and Sinisa Todorovic. Weakly supervised energy-based learning for action segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6243–6251, 2019. 1, 2, 3, 4, 6, 8, 9
- [28] Zhe Li, Yazan Abu Farha, and Jurgen Gall. Temporal action segmentation from timestamp supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8365–8374, 2021. 2
- [29] Yunyu Liu, Lichen Wang, Yue Bai, Can Qin, Zhengming Ding, and Yun Fu. Generative view-correlation adaptation for semi-supervised multi-view learning. In *European Conference on Computer Vision*, pages 318–334. Springer, 2020. 2, 3
- [30] Tahmida Mahmud, Mahmudul Hasan, and Amit K Roy-Chowdhury. Joint prediction of activity labels and starting times in untrimmed videos. In *Proceedings of the IEEE International conference on Computer Vision*, pages 5773–5782, 2017. 2
- [31] Jingjing Meng, Suchen Wang, Hongxing Wang, Junsong Yuan, and Yap-Peng Tan. Video summarization via multi-view representative selection. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 1189–1198, 2017. 2
- [32] Rameswar Panda and Amit K Roy-Chowdhury. Multi-view surveillance video summarization via joint embedding and sparse optimization. *IEEE Transactions on Multimedia*, 19(9):2010–2021, 2017. 2
- [33] Paritosh Parmar and Brendan Tran Morris. What and how well you performed? a multitask learning approach to action quality assessment. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 304–313, 2019. 1
- [34] Florent Perronnin and Christopher Dance. Fisher kernels on visual vocabularies for image categorization. In *2007 IEEE conference on computer vision and pattern recognition*, pages 1–8. IEEE, 2007. 6
- [35] AJ Piergiovanni and Michael S Ryoo. Recognizing actions in videos from unseen viewpoints. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4124–4132, 2021. 2
- [36] Sanqing Qu, Guang Chen, Dan Xu, Jinhu Dong, Fan Lu, and Alois Knoll. Lap-net: Adaptive features sampling via learning action progression for online action detection. *arXiv preprint arXiv:2011.07915*, 2020. 2
- [37] Alexander Richard, Hilde Kuehne, and Juergen Gall. Weakly supervised action learning with rnn based fine-to-coarse modeling. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 754–763, 2017. 2
- [38] Alexander Richard, Hilde Kuehne, Ahsan Iqbal, and Juergen Gall. Neuralnetwork-viterbi: A framework for weakly supervised video learning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 7386–7395, 2018. 2, 3
- [39] Marcus Rohrbach, Anna Rohrbach, Michaela Regneri, Sikandar Amin, Mykhaylo Andriluka, Manfred Pinkal, and Bernt Schiele. Recognizing fine-grained and composite activities using hand-centric features and script data. *International Journal of Computer Vision*, 119(3):346–373, 2016. 1
- [40] Saquib Sarfraz, Naila Murray, Vivek Sharma, Ali Diba, Luc Van Gool, and Rainer Stiefelhagen. Temporally-weighted hierarchical clustering for unsupervised action segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11225–11234, 2021. 2
- [41] Fadime Sener, Dipika Singhania, and Angela Yao. Temporal aggregate representations for long-range video understanding. In *European Conference on Computer Vision*, pages 154–171. Springer, 2020. 2
- [42] Fadime Sener and Angela Yao. Unsupervised learning and segmentation of complex activities from video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8368–8376, 2018. 2
- [43] Pierre Sermanet, Corey Lynch, Yevgen Chebotar, Jasmine Hsu, Eric Jang, Stefan Schaal, Sergey Levine, and Google Brain. Time-contrastive networks: Self-supervised learning from video. In *2018 IEEE international conference on robotics and automation (ICRA)*, pages 1134–1141. IEEE, 2018. 3
- [44] Zheng Shou, Junting Pan, Jonathan Chan, Kazuyuki Miyazawa, Hassan Mansour, Anthony Vetro, Xavier Giro-i Nieto, and Shih-Fu Chang. Online detection of action start in untrimmed, streaming videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 534–551, 2018. 2
- [45] Gunnar A Sigurdsson, Abhinav Gupta, Cordelia Schmid, Ali Farhadi, and Karteek Alahari. Actor and observer: Joint modeling of first and third-person videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7396–7404, 2018. 2, 3
- [46] Yaser Souri, Yazan Abu Farha, Fabien Despinoy, Gianpiero Francesca, and Juergen Gall. Fifa: Fast inference approxima-

- tion for action segmentation. In *DAGM German Conference on Pattern Recognition*, pages 282–296. Springer, 2021. 2
- [47] Yaser Souri, Mohsen Fayyaz, Luca Minciullo, Gianpiero Francesca, and Juergen Gall. Fast weakly supervised action segmentation using mutual consistency. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 1, 2
- [48] Yansong Tang, Dajun Ding, Yongming Rao, Yu Zheng, Danyang Zhang, Lili Zhao, Jiwen Lu, and Jie Zhou. Coin: A large-scale dataset for comprehensive instructional video analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1207–1216, 2019. 1
- [49] Shruti Vyas, Yogesh S Rawat, and Mubarak Shah. Multi-view action recognition using cross-view video prediction. In *ECCV*, pages 427–444. Springer, 2020. 2
- [50] Dongang Wang, Wanli Ouyang, Wen Li, and Dong Xu. Dividing and aggregating network for multi-view action recognition. In *ECCV*, pages 451–467, 2018. 2
- [51] Heng Wang and Cordelia Schmid. Action recognition with improved trajectories. In *Proceedings of the IEEE international conference on computer vision*, pages 3551–3558, 2013. 6
- [52] Lichen Wang, Zhengming Ding, Zhiqiang Tao, Yunyu Liu, and Yun Fu. Generative multi-view human action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6212–6221, 2019. 2, 3
- [53] Zhenzhi Wang, Ziteng Gao, Limin Wang, Zhifeng Li, and Gangshan Wu. Boundary-aware cascade networks for temporal action segmentation. In *European Conference on Computer Vision*, pages 34–51. Springer, 2020. 2
- [54] Bo Xiong, Haoqi Fan, Kristen Grauman, and Christoph Feichtenhofer. Multiview pseudo-labeling for semi-supervised learning from video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7209–7219, 2021. 3
- [55] Mingze Xu, Mingfei Gao, Yi-Ting Chen, Larry S Davis, and David J Crandall. Temporal recurrent networks for online action detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5532–5541, 2019. 2
- [56] Mingze Xu, Yuanjun Xiong, Hao Chen, Xinyu Li, Wei Xia, Zhuowen Tu, and Stefano Soatto. Long short-term transformer for online action detection. *Advances in Neural Information Processing Systems*, 34, 2021. 2
- [57] Christopher Zach, Thomas Pock, and Horst Bischof. A duality based approach for realtime tv-l1 optical flow. In *Joint pattern recognition symposium*, pages 214–223. Springer, 2007. 6
- [58] Bowen Zhang, Hao Chen, Meng Wang, and Yuanjun Xiong. Online action detection in streaming videos with time buffers. *arXiv preprint arXiv:2010.03016*, 2020. 2
- [59] Peisen Zhao, Lingxi Xie, Ya Zhang, Yanfeng Wang, and Qi Tian. Privileged knowledge distillation for online action detection. *arXiv preprint arXiv:2011.09158*, 2020. 2