# HOI-aware Adaptive Network for Weakly-supervised Action Segmentation

**Runzhong Zhang**[1] , **Suchen Wang**[1] , **Yueqi Duan**[2*] , **Yansong Tang**[2] ,
**Yue Zhang**[3] and **Yap-Peng Tan**[1]
[1]Nanyang Technological University
[2]Tsinghua University
[3]Beijing Jiaotong University
{runzhong001, suchen001}@e.ntu.edu.sg, duanyueqi@tsinghua.edu.cn,
tang.yansong@sz.tsinghua.edu.cn, 17112065@bjtu.edu.cn, eyptan@ntu.edu.sg

## Abstract

In this paper, we propose an HOI-aware adaptive network named AdaAct for weakly-supervised action segmentation. Most existing methods learn a fixed network to predict the action of each frame with the neighboring frames. However, this would result in ambiguity when estimating similar actions, such as pouring juice and pouring coffee. To address this, we aim to exploit temporally global but spatially local human-object interactions (HOI) as video-level prior knowledge for action segmentation. The long-term HOI sequence provides crucial contextual information to distinguish ambiguous actions, where our network dynamically adapts to the given HOI sequence at test time. More specifically, we first design a video HOI encoder that extracts, selects, and integrates the most representative HOI throughout the video. Then, we propose a two-branch HyperNetwork to learn an adaptive temporal encoder, which automatically adjusts the parameters based on the HOI information of various videos on the fly. Extensive experiments on two widely-used datasets including Breakfast and 50Salads demonstrate the effectiveness of our method under different evaluation metrics.

## 1 Introduction

Action segmentation aims to predict the action for every frame in the video. While previous methods have achieved remarkable performance in the fully-supervised setting [Kuehne *et al.*, 2016; Lea *et al.*, 2017; Rohrbach *et al.*, 2012; Singh *et al.*, 2016; Yi *et al.*, 2021; Park *et al.*, 2022], framewise annotation still requires huge labor costs and is hard to obtain. Therefore, action segmentation with weaker forms of supervision gradually gains its popularity in recent years. In particular, transcript supervision [Bojanowski *et al.*, 2014; Kuehne *et al.*, 2017; Huang *et al.*, 2016; Ding and Xu, 2018; Li *et al.*, 2019a; Lu and Elhamifar, 2021] provides an ordered list of actions occurring in the video without the starting and ending time, which significantly reduces the annotation costs and improves the applicability to a rapidly-growing number of videos on the Internet.

To learn from the transcript, previous approaches mainly follow the "generating-matching" pipeline [Richard *et al.*, 2018; Li *et al.*, 2019a; Lu and Elhamifar, 2021]. With the given training videos, they first apply a temporal encoder to generate framewise action probabilities, and then match the predicted probabilities sequence with the transcript based on Viterbi decoding or dynamic time warping. However, to estimate the action probability of frame $t$ during the generating step, most existing approaches only take a fixed number of neighbor frames around it [Richard *et al.*, 2018; Li *et al.*, 2019a], and feed such video clip features into an RNN-based [Chung *et al.*, 2014] architecture. In this case, the temporal encoder would fail to distinguish the attribute of similar actions such as pouring coffee and pouring juice, which may lead to counter-intuitive results of pouring coffee in a juice-making video. Although tremendous efforts have been made to remedy such ambiguity in the matching step, the results are still unsatisfying due to the inherent defect in the previous generating process.

In this paper, we address the ambiguity problem by designing an adaptive weakly-supervised action segmentation framework called AdaAct. Different from previous methods which take a series of fixed-length video clips as input successively (as shown in Figure 1 (a)), we exploit rich contextual information from temporally global but spatially local human-object interactions (HOI) throughout the whole video. Such HOI sequence further instructs the network as prior knowledge, where our temporal encoder can be dynamically adapted to it at the test time. As illustrated in Figure 1 (b), our method first extracts key interactions with objects at different video timestamps, such as the knife, orange, and squeezer. The obtained HOI sequence is further incorporated into the temporal encoder, thereby the network parameters of the encoder dynamically change with the HOI information on the fly. More specifically, we design a three-step video HOI encoder with the "extracting-selecting-integrating" process. We first apply a pre-trained HOI detector to extract positive interaction bounding boxes for the whole video and design a simple selecting algorithm to pick the most representative ones from them. Then, we explore the relations between these key HOI boxes and integrate them into a single feature vector via a transformer-based network. To dynamically adapt

*Corresponding author

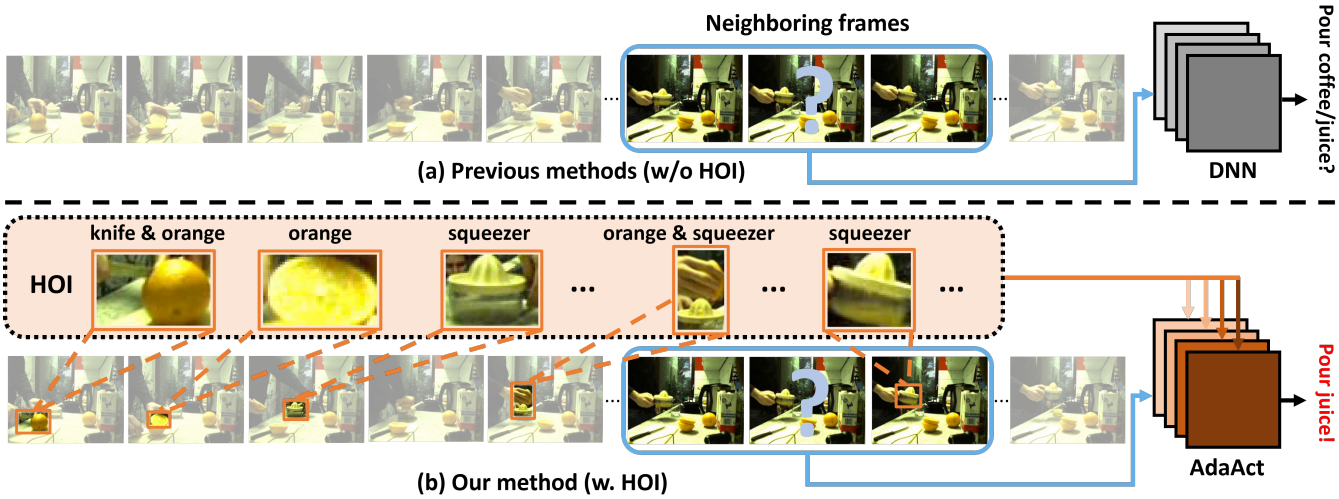


Figure 1: (a) Most existing methods estimate the action probability of frame $t$ using features of the adjacent frames (blue box in the figure). They are difficult to distinguish actions such as pouring juice, coffee, and water considering the high representation similarity. (b) Our method exploits temporally global but spatially local HOI information to learn an adaptive temporal encoder, which provides essential contextual information to distinguish similar actions. As shown in the figure, considering the interactions with a knife, orange, and squeezer in the video, the query action would be more likely to be pouring juice rather than coffee or water.

the network parameters, we propose a two-branch HyperNetwork that simultaneously learns HOI-dependent and HOI-independent knowledge. Our HOI-independent branch aims to unearth the general characteristics of instructional videos by iteratively updating a learnable embedding list throughout the training process. Such transferable information will later be encoded as a part of instruction used in the temporal encoder during the test time. On the other hand, our HOI-dependent branch takes the encoded feature vector from the video HOI detector as input and adapts the temporal encoder to the given HOI knowledge occurring in the video. Finally, late fusion is utilized to merge the knowledge from two branches, resulting in more precise action segmentation.

We summarize our key contributions as follows:

1) To our best knowledge, this is the first work to learn an adaptive temporal encoder for weakly-supervised action segmentation, where the parameters of the network are dynamically adapted according to the input video on the fly.
2) We propose to exploit temporally global but spatially local HOI information in weakly-supervised action segmentation, which provides essential contextual information to address the ambiguity problem of similar actions.
3) We validate our method on two challenging datasets, Breakfast and 50Salads, and achieve state-of-the-art results for both weakly-supervised action segmentation and alignment tasks.

## 2 Related Work

**Fully-supervised action segmentation.** Fully-supervised action segmentation methods learn action segments under the guidance of framewise annotations. Earlier attempts [Karaman *et al.*, 2014; Rohrbach *et al.*, 2012] applied action classification on the sliding window, followed by non-maximum suppression to filter out redundant predictions. However, these approaches failed to model the temporal dependency between action sequences. Kuehne *et al.* [2016] tackled this problem via the hidden Markov model, while Pirsiavash *et al.* [2014] applied context-free grammar to capture temporal structure. In recent years, various network architectures were proposed for learning long-range dependency. Lea *et al.* [2017] introduced an encoder-decoder architecture for action segmentation and detection. Lei *et al.* [2018] further applied deformable convolutions and residual stream. Farruha *et al.* [2019] and Li *et al.* [2020] introduced dilated temporal convolution and proposed a multi-stage temporal convolutional network, while various methods improved the multi-stage network using graph-based temporal reasoning [Huang *et al.*, 2020] or boundary-aware cascade network [Wang *et al.*, 2020b]. With the success of transformer-based models in computer vision, Yi *et al.* [2021] first introduced the transformer into the action segmentation task. Different from previous methods, Li *et al.* [2022] reformulated the action labels as text prompts and paired them with corresponding video clips, and co-trained the text encoder and the video encoder through a contrastive approach.

**Weakly-supervised action segmentation.** Many of the weakly-supervised methods utilize transcripts as supervision during training. Huang *et al.* [2016] first introduced the connectionist temporal classification framework to evaluate all possible matching between the videos and transcripts. Ding *et al.* [2018] started from the initial uniform mapping of the action transcript, and iteratively refined the transcript during the

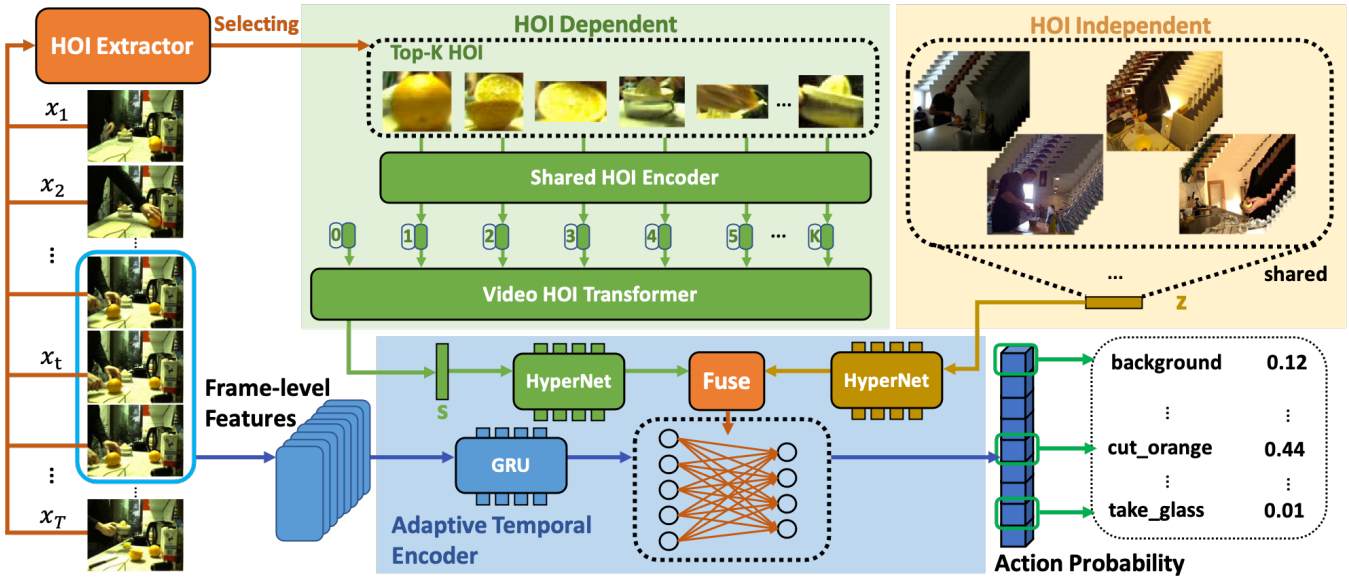


Figure 2: Overview of the network architecture. Our method simultaneously learns HOI-dependent knowledge $s$ from the video HOI encoder and HOI-independent knowledge $z$ from various videos across the dataset. The obtained knowledge is further incorporated through the two-branch HyperNetwork and late fusion, which generates the network parameters of the adaptive temporal encoder. In this way, our network dynamically adapts to the video contents when estimating the action probability of frame $t$, leading to better discrimination for similar actions.

training procedure. However, these methods fail to achieve end-to-end training. Richard *et al*. [2018] instead generated pseudo frame labels using the Viterbi algorithm and trained a classifier based on framewise cross-entropy loss. Li *et al*. [2019a] further extended the NN-Viterbi [Richard *et al*., 2018] by introducing a new constrained discriminative forward loss, which maximized the energy difference between valid and invalid segmentation of training videos. In D3TW, Chang *et al*. [2019] first applied a discriminative model for solving the degenerate sequence problem. As these methods have to search all the transcripts during testing and thus suffer from long inference time, Souri *et al*. [2021] proposed MuCon, a two-branch network that predicted both transcript and framewise label of action segmentation, and designed the mutual loss to ensure the consistency of representations. In recent years, different weakly-supervised settings besides the transcripts have been studied. Fayyaz *et al*. [2020] and Li *et al*. [2020] reduced the supervision level, assuming only the unordered list of actions is available for each training video. Inspired by the point supervision in semantic segmentation [Bearman *et al*., 2016], Li *et al*. [2021] trained a segmentation model using timestamps annotations, in which case only one arbitrary frame is annotated for each action. As these methods use different kinds of supervision for training, we do not directly compare them with our approach.

**Human object interaction.** The existing HOI detection can be mainly categorized into single-stage approaches [Liao *et al*., 2020; Wang *et al*., 2020a; Kim *et al*., 2020; Chen *et al*., 2021] and two-stage approaches [Li *et al*., 2019b; Zhang *et al*., 2021; Zhou and Chi, 2019; Zhou *et al*., 2020; Ulutan *et al*., 2020]. Single-stage approaches integrate bounding boxes detection and interaction recognition into a single model. Liao *et al*. [2020] and Wang *et al*. [2020a] first simultaneously generated bounding box candidates and interactions, and then outputted final predictions after the matching step. Chen *et al*. [2021] instead reformed the HOI detection as an adaptive set prediction problem. Compared with one-stage methods, two-stage approaches first detect humans and objects following the object detection pipeline and then apply an interaction model to analyze the relations of the bounding boxes. Qi *et al*. [2018] and Zhang *et al*. [2021] modeled the relations using graph neural network. Fang *et al*. [2018] emphasized the importance of human-part knowledge in HOI detection. Although different methods have been proposed in the image domain, research on video-level HOI detection is still under-exploited.

## 3 Methodology

Our goal is to address the weakly-supervised action segmentation problem under transcript supervision. Formally, we define each video with its supervision as a tuple $\{v, x_1^T, a_1^O, l_1^O\}$, where $v$ represents the video as a stack of raw frames, $x_1^T = [x_1, ..., x_T]$ denotes the unsupervised framewise features with length $T$, $a_1^O = [a_1, ..., a_O]$ indicates the transcript, an ordered list of $O$ actions occurred in the video, and $l_1^O = [l_1, ..., l_O]$ records the number of frames for each of the corresponding actions. Every action $a_o$ belongs to the set of $A$ action classes, namely $a_o \in \mathcal{A} = \{1, ..., A\}$. During the inference, the objective is to predict the optimal action list $\hat{a}$ and corresponding length $\hat{l}$ based on the framewise features $x$ of the video $v$.

In this paper, we propose an adaptive network named AdaAct that utilizes video-level HOI to distinguish similar actions. As shown in Figure 2, our method mainly consists

of a video HOI encoder and an adaptive temporal encoder. For the video HOI encoder, it first takes the input video and extracts all the valid interactions, then selects top-$K$ interactions by removing redundant and low-score detection. These interactions are finally integrated as HOI-dependent knowledge $s$. For the adaptive temporal encoder, it incorporates HOI-dependent knowledge $s$ with HOI-independent knowledge $z$ via a two-branch HOI-aware HyperNetwork [Ha *et al.*, 2016], which predicts the network parameters of the temporal encoder. In the following, we describe the video HOI encoder and the adaptive temporal encoder in detail, as well as the training strategy to learn these two models.

## 3.1 Video HOI Encoder

The goal of our video HOI encoder is to model the dependencies between key HOI through the whole video and encode them as HOI-dependent knowledge $s$. It mainly contains three levels from the bottom to the top: extracting, selecting, and integrating.

### Extracting

We take the video as input at the first level. Since the majority of HOI detection methods are developed only for image scenarios, we pre-process the video by down-sampling and extracting the raw frames under 15 FPS. After that, we employ the detector on every frame iteratively following the temporal order. To avoid introducing additional computation cost, here we follow the 100 Days of Hands [Shan *et al.*, 2020] with its weight frozen during training and testing. The model outputs the predictions as tuple $\{b_h, b_o, c, t\}$, where $b_h$ and $b_o$ represent the bounding boxes of hands and object, $c \in [0, 1]$ denotes the interaction confidence score, and $t$ indicates the timestamp of the frame.

### Selecting

Inspired by the non-maximum suppression (NMS) [Neubeck and Van Gool, 2006] used for filtering proposals in object detection, we propose a video-NMS algorithm to select top-$K$ object bounding boxes from the predictions pool. Different from the traditional NMS algorithm that filters the proposals only by the intersection over union (IoU), our method also adds the temporal constraint, so that the duplicates of highest-score proposals are removed based on IoU and time interval. After that, $K$ tuple predictions with the highest score are selected and ranked by timestamp order for the next step.

### Integrating

In the integrating step, we propose a ViT-based network to generate HOI-dependent knowledge $s$. To handle all the $K$ object bounding boxes $b_o$, we use a frozen ResNet50 [He *et al.*, 2016] and project them into a sequence of HOI embeddings $[e_1, ..., e_K]$. Following ViT's design, we append a learnable embedding $e_{token}$ before the sequence, the state of which serves as the HOI-dependent knowledge $s$ at the transformer output. We also add the 1D learnable position embeddings $P = [p_{token}, p_1, ..., p_K]$ to the HOI embeddings and feed the resulting sequence into the ViT network.

Given the input $E_0 = [e_{token}, e_1, ...e_K] + P$, the network conducts the following procedures for layer $n$ from 1 to $N$:

$$E'_n = \text{MSA}(\text{LN}(E_{n-1})) + E_{n-1}, \tag{1}$$

$$E_n = \text{MLP}(\text{LN}(E'_n)) + E'_n, \tag{2}$$

where MSA stands for the multi-head self-attention module, MLP represents multi-layer perceptron and LN denotes LayerNorm. The obtained HOI-dependent knowledge $s$ is then merged with HOI-independent knowledge $z$, which will be further explained in the following section.

## 3.2 Adaptive Temporal Encoder

For a fair comparison, we apply the GRU followed by a linear layer as the temporal encoder backbone in consistence with the previous methods [Richard *et al.*, 2018; Li *et al.*, 2019a; Lu and Elhamifar, 2021]. To instruct the temporal encoder with video-level knowledge, we employ the two-branch HOI-aware HyperNetwork [Ha *et al.*, 2016], a sub-network used to learn parameters for the temporal encoder in the action predicting process. Specifically, for the linear layer in the temporal encoder, its weights and bias are separately generated by feeding learnable embedding $z$ into the HOI-independent branch and $s$ into the HOI-dependent branch. The pipeline can be written as follows:

$$W = F(H_i(z), H_d(s)), \tag{3}$$

$$b = F(H'_i(z), H'_d(s)), \tag{4}$$

where $H_i$, $H_d$ represent the independent and dependent branches for weights generation, and $H'_i$, $H'_d$ are for bias. Finally, we apply the late fusion module $F$ to integrate information from the two branches. Instead of fixing the network during test time in typical deep learning networks, our method adaptively adjusts the network parameters by incorporating different video-level prior knowledge into framewise action prediction, thus eliminating the potential ambiguity occurring between similar actions.

### Multi-head HOI-independent Branch

Since the weight and bias can be considered as matrices with different dimensions, here we use the weight generation HyperNetwork as the example. We suppose the weight parameters generated from the HOI-independent branch are stored in matrix $W^z \in \mathbb{R}^{C_{out} \times A}$, where $C_{out}$ represents the frame representation dimension after processed by GRU. Therefore, the HOI-independent branch can be written as below:

$$W^z = H_i(z). \tag{5}$$

Instead of formulating the HOI-independent knowledge as a single vector, we initialize the embedding list $z = [z_1, ..., z_m]$, $z_i \in \mathbb{R}^D$. These vectors are fed into the two-layer linear network $H_i$, yielding $m$ different vectors with the same length $\frac{C_{out} \times A}{m}$. Finally, the outputs are reshaped and concatenated together as the $W^z$. To ensure the correctness of dimension, $C_{out}$ must be divisible by $m$. Formally, the network processes the following procedures:

$$W^z_i = \varphi(\text{MLP}(z_i)), i = 1, ..., m, \tag{6}$$

$$W = [W_1, ..., W_m], W_i \in \mathbb{R}^{\frac{C_{out}}{m} \times A}, \tag{7}$$

| **Breakfast** | MoF | MoF-BG | IoU | IoD |
|---|---|---|---|---|
| ECTC [Huang *et al.*, 2016] | 27.7 | - | - | - |
| HMM/RNN [Richard *et al.*, 2017] | 33.3 | - | - | - |
| TCFPN [Ding and Xu, 2018] | 38.4 | 38.4 | 24.2 | 40.6 |
| NN-Viterbi* [Richard *et al.*, 2018] | 41.9 | 38.9 | 33.3 | 42.8 |
| D3TW [Chang *et al.*, 2019] | 45.7 | - | - | - |
| CDFL* [Li *et al.*, 2019a] | 49.8 | 47.1 | 35.3 | 45.6 |
| MuCon [Souri *et al.*, 2021] | 49.0 | - | - | - |
| TASL* [Lu and Elhamifar, 2021] | 47.2 | 44.4 | 36.1 | 45.8 |
| AdaAct (Ours) | **51.2** | **48.3** | **36.3** | **46.4** |
| **50Salads** | MoF | MoF-BG | IoU | IoD |
| NN-Viterbi [Richard *et al.*, 2018] | 49.4 | - | - | - |
| CDFL [Li *et al.*, 2019a] | 54.7 | 49.8 | 31.5 | 40.4 |
| AdaAct (Ours) | **55.6** | **50.3** | **35.2** | **44.6** |

Table 1: Action segmentation results on the Breakfast and the 50Salads datasets. The dash line indicates that no prior result is available. We report the reproduced results for the methods with an asterisk.

where $\varphi(\cdot)$ represents the reshape operation. Compared with the original multi-head mechanism that uses different linear layers to project the same input, our method initializes a list of vectors and keeps the same network parameters.

**Multi-head HOI-dependent Branch**

Similar to the HOI-independent branch, we maintain the embedding list with the same size and separately sum them with the HOI-dependent knowledge $s$. The resulting vectors are projected by the two-layer linear network $H_d$, followed by the reshaping and concatenation to get the matrix $W^s$.

Finally, we generate the weight of the linear layer by element-wise multiplying $W^z$ and $W^s$:

$$W = W^z \odot W^s. \tag{8}$$

### 3.3 Transcript Decoding and Training

We formulate the action segmentation problem as finding the most likely labeling based on the video features. Specifically, the optimal $(\hat{a}_1^O, \hat{l}_1^O)$ can be obtained as follows:

$$\begin{aligned}
&\arg\max_{a_1^O, l_1^O} \left\{ p\left(a_1^O, l_1^O \mid x_1^T\right) \right\} \\
&= \arg\max_{a_1^O, l_1^O} \left\{ p\left(x_1^T \mid a_1^O, l_1^O\right) \cdot p\left(l_1^O \mid a_1^O\right) \cdot p\left(a_1^O\right) \right\} \\
&= \arg\max_{a_1^O, l_1^O} \left\{ \prod_{t=1}^{T} p\left(x_t \mid a_{o(t)}\right) \cdot \prod_{o=1}^{O} p\left(l_o \mid a_o\right) \cdot p\left(a_1^O\right) \right\}.
\end{aligned} \tag{9}$$

In the above formula, $p\left(x_t \mid a\right)$ can be further transformed:

$$p\left(x_t \mid a\right) \propto \frac{p\left(a \mid x_t\right)}{p(a)}, \tag{10}$$

where $p\left(a \mid x_t\right)$ is modeled by the output of our adaptive temporal encoder. For the modeling of $p\left(l_o \mid a_o\right)$ and $p\left(a_1^O\right)$, the same settings with previous work [Richard *et al.*, 2018] are utilized for the fair comparison.

| **Breakfast** | MoF | MoF-BG | IoU | IoD |
|---|---|---|---|---|
| ECTC [Huang *et al.*, 2016] | 35.0 | - | - | 45.0 |
| HMM/RNN [Richard *et al.*, 2017] | - | - | - | 47.3 |
| TCFPN [Ding and Xu, 2018] | 53.5 | 51.7 | 35.3 | 52.3 |
| D3TW [Chang *et al.*, 2019] | 57.0 | - | - | 56.3 |
| CDFL [Li *et al.*, 2019a] | 63.0 | 61.4 | 45.8 | 63.9 |
| MuCon [Souri *et al.*, 2021] | - | - | - | **66.2** |
| TASL [Lu and Elhamifar, 2021] | 64.1 | - | **49.9** | 64.7 |
| AdaAct (Ours) | **64.4** | **62.3** | **49.9** | 65.3 |
| **50Salads** | MoF | MoF-BG | IoU | IoD |
| CDFL [Li *et al.*, 2019a] | 68.0 | 65.3 | 45.5 | 58.7 |
| AdaAct (Ours) | **69.8** | **66.5** | **47.5** | **60.3** |

Table 2: Action alignment results on the Breakfast and the 50Salads datasets. The dash line indicates that no prior result is available.

Figure 3: HOI detection in the frying egg activity. Our model only encodes the representation of spatula, since there is no direct interaction between human hands and egg.

We apply the constrained discriminative forward loss proposed by [Li *et al.*, 2019a] for the network training, and provide detailed comparisons with the baseline method in the following section. It is worth noting that our method shows great flexibility and can be plugged into different existing methods.

## 4 Experiments

We validate our proposed method by comparing it with several state-of-the-art weakly-supervised action segmentation approaches, and discuss the effectiveness of each component in the following ablation studies.

### 4.1 Experimental Setup

**Datasets.** We conduct our experiments on two real-world instructional video datasets: Breakfast [Kuehne *et al.*, 2014] and 50Salads [Stein and McKenna, 2013]. The Breakfast dataset contains more than 1.7k videos of people performing 10 different cooking activities, such as preparing juice or preparing salad. The cooking activities are comprised of 48 fine-frained actions. Each video has 6.9 action segments on average, and the length of the video varies from several seconds to a few minutes. The 50Salads dataset has 50 long videos with 17 different action classes. On average, each video contains 20 action instances.

**Evaluation metrics.** We use the following four metrics for evaluation. (1) Mean over frame accuracy (**MoF**) is defined as the number of correctly predicted frames divided

| | cereals | coffee | fried-egg | juice | milk | pancake | salad | sandwich | scrambled-egg | tea | Total MoF |
|---|---|---|---|---|---|---|---|---|---|---|---|
| NN-Viterbi [Richard *et al.*, 2018] | 39.5 | 39.0 | 48.4 | 74.2 | 56.8 | 16.9 | 46.0 | 57.0 | 45.0 | **44.9** | 41.9 |
| CDFL [Li *et al.*, 2019a] | 37.9 | 37.0 | 54.1 | 75.8 | 58.2 | 31.1 | 27.4 | 38.8 | 43.4 | 34.6 | 49.8 |
| TASL [Lu and Elhamifar, 2021] | 51.8 | 43.6 | **59.2** | 74.2 | 56.5 | 24.9 | 46.0 | 58.8 | **50.4** | 42.1 | 47.2 |
| AdaAct (Ours) | **56.1** | **57.3** | 49.1 | **76.1** | **58.7** | **47.0** | **48.2** | **63.4** | 44.7 | 35.5 | **51.2** |

Table 3: Action segmentation performance on the Breakfast dataset. We report the mean over frame accuracy (MoF) of every cooking activity and across all the activities, where "cereals" indicates "making cereals", etc.

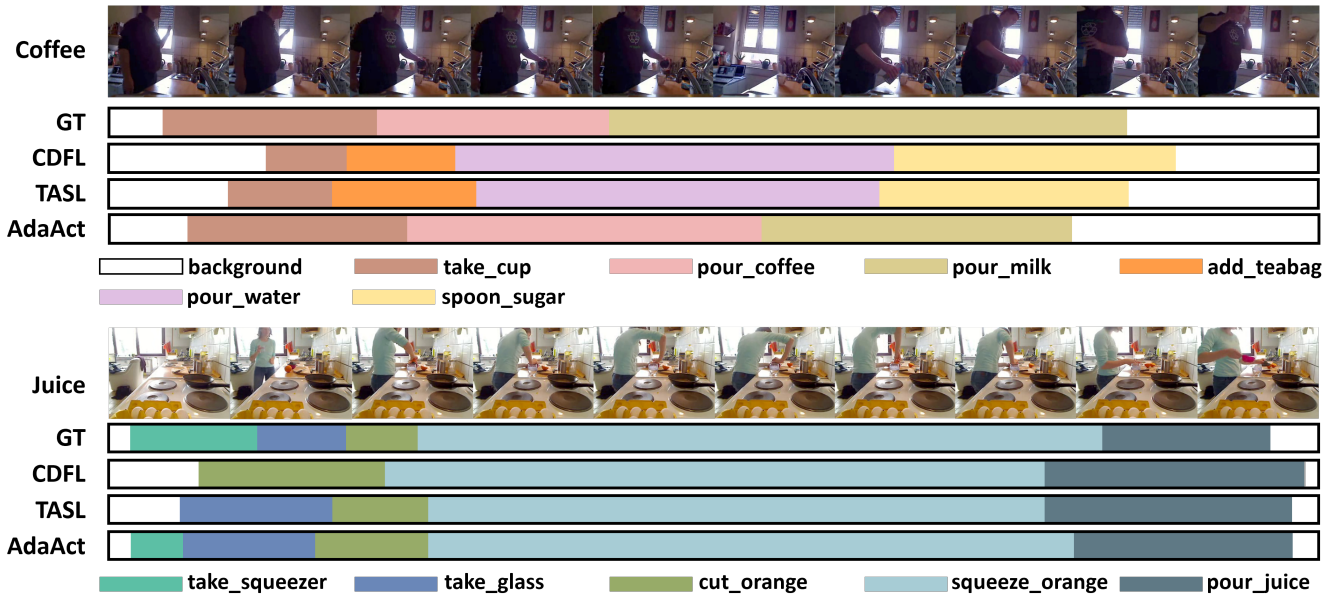


Figure 4: Action segmentation results of CDFL, TASL, and our approach on the coffee-making video (top) and juice-making video (bottom), where GT stands for the ground truth segmentation.

by the total number of frames. (2) Mean over frame accuracy without background (**MoF-BG**) removes the background frames when calculating MoF, thus eliminating the drawback when video contains long periods of irrelevant information. (3) Intersection over union (**IoU**) is calculated as $|GT \cap correct| \,/\, |GT \cup correct|$, where $GT$ stands for the ground truth frames and $correct$ denotes the correctly classified frames. (4) Intersection over detection (**IoD**) is defined as $|GT \cap correct| \,/\, |GT|$.

**Implementation details.** For the video HOI encoder, we follow the same HOI detector pre-trained on the 100K dataset as mentioned in 100 Days of Hands [Shan *et al.*, 2020]. We set 0.5 as the HOI detection threshold and pick $K = 10$ bounding boxes after the selection process. For the ViT network, we replace the image patching and linear projection steps with the ResNet50 backbone, leading to $10 \times 2048$ input size. We use $D = 128$ for the dimension of both HOI-dependent and HOI-independent knowledge, and set the multi-head number as 8. For the adaptive temporal encoder, we use the 64-hidden unit GRU. We maintain the learning rate of 0.01 with 12500 epochs through the training process.

## 4.2 Experimental Results

We report the experimental results for two tasks, namely action segmentation where only the video is available during the inference, and action alignment where both video and transcript are provided.

**Quantitative Results**

We quantitatively compare our method with prior works in this section. Table 1 reports the action segmentation results on two instructional video datasets under four evaluation metrics, where the best results are indicated in bold. We can observe that by introducing HOI-aware knowledge, our method exceeds state-of-the-art methods by 1.4% MoF and 1.2% MoF-BG on the Breakfast dataset, and 0.9% MoF and 0.5% MoF-BG on the 50Salads dataset. This validates that when only video is given during testing, our method learns rich video-level knowledge and instructs the decision-making of the temporal encoder, leading to significant performance improvement on both datasets.

Table 2 shows the action alignment results following the same metrics in Table 1. Notice that in this setting the transcript is available during inference, thus providing stronger video-level knowledge compared with learned HOI-aware knowledge in our method. Despite this, our method still outperforms existing approaches on both datasets and achieves +1.8% MoF and 1.2% MoF-BG improvement on the 50Salads, which proves that the HOI-aware knowledge also helps to refine the starts and ends of predicted actions in the video.

We also report the per-activity MoF results on the Break-


Figure 5: Visualization of HOI detection and corresponding action probability results. Only 4 frames with top-3 action predictions are shown due to the figure size limit. The correct predictions with the highest probability score are marked in red color.

fast dataset under the action segmentation setting. In Table 3, we can observe that our methods outperforms the baseline approaches for most of the cooking activities as expected. For those activities containing similar actions, such as pouring cereals in "cereals", pouring coffee in "coffee", and pouring milk in "pancake", our method achieves a large performance gain (+4.3% MoF in "cereals", +13.7% MoF in "coffee" and +15.9% MoF in "pancake"). This validates the effectiveness of our method in distinguishing ambiguous actions among different cooking activities.

However, we notice that our method still suffers from low performance in tea-making and egg-making ("fried-egg" and "scrambled-egg") videos. Since our method applies the constrained discriminative forward loss and follows the temporal encoder architecture in CDFL, our method is inevitably affected by its performance. From this perspective, we still outperforms CDFL for +0.9% in "tea" and +1.3% in "scrambled-egg". We also investigate the reason for performance dropping in the "fried-egg" activity. By visualizing the selected HOI bounding boxes in Figure 3, we observe that our HOI extractor tends to capture the spatula, while limited egg information is selected due to the long distance between itself and human hands. Despite the fact that people in the videos directly interact with the spatula for a long period of time, with the absence of egg detection, such HOI still could not provide enough information to instruct the temporal encoder. Therefore, the performance of our method instead degrades due to the noisy HOI-aware knowledge.

**Qualitative Results**

Figure 4 shows the action segmentation results of two videos on the Breakfast dataset. For the coffee-making video on the top, existing methods make the wrong predictions due to the high similarity of different pouring actions. In contrast, our method encodes strong semantic information in the HOI-aware knowledge, thus correctly classifying all the actions contained in the video. In the bottom juice-making case, when all the methods successfully capture the actions through the video, our method also shows higher accuracy in detecting the boundaries among different actions.

We also visualize how HOI detection helps to eliminate ambiguity in predicting action probability. For the cereals-making video in Figure 5, our HOI detector precisely captures the interactions with a cereal bag and a milk box at different timestamps. Without applying HOI-aware knowledge, the existing method suffers from low confidence when distinguishing "pour_cereals", "pour_water" and "pour_milk" in the first frame and makes the wrong prediction in the second. In contrast, our method both makes the correct predictions and widens the probability gaps among similar actions. In the last two frames, our method also greatly improves the "pour_milk" confidence, demonstrating that HOI-aware knowledge provides strong instruction for a better action probability estimation.

| HOI-dependent | HOI-independent | multi-head | MoF |
|---|---|---|---|
| | ✓ | ✓ | 47.5 |
| ✓ | | ✓ | 50.3 |
| ✓ | ✓ | | 49.4 |
| ✓ | ✓ | ✓ | **51.2** |

Table 4: Effect of each component in our method. We report the mean over frame accuracy (MoF) for action segmentation on the Breakfast dataset.

| knowledge dimension | MoF | MoF-BG | IoU | IoD |
|---|---|---|---|---|
| 32 | 45.5 | 42.8 | 33.1 | 42.9 |
| 64 | 48.7 | 46.0 | 34.8 | 45.2 |
| 128 | **51.2** | **48.3** | **36.3** | **46.4** |
| 256 | 46.9 | 44.0 | 33.7 | 44.0 |

Table 5: Effect of the HOI-dependent and HOI-independent knowledge dimension in our method. We report the mean over frame accuracy (MoF) for action segmentation on the Breakfast dataset.

### 4.3 Ablation Studies

We examine different components of our method and report the results in Table 4. The full model with the best performance is provided at the bottom for comparison. Introducing HOI-dependent knowledge leads to the most significant improvement of MoF by 3.7% and HOI-independent knowledge contributes to 0.9% MoF improvement, which demonstrates that both sources of knowledge are necessary for the HOI-aware understanding. In addition, applying the multi-head mechanism further achieves +1.8% MoF.

Table 5 shows how different dimensions of HOI-dependent/independent knowledge affect the action segmentation. As expected, either too small or large size would cause the performance to drop. The highest accuracy is achieved with 128 dimensions.

## 5 Conclusion

In this paper, we have proposed AdaAct, an HOI-aware adaptive network for video action segmentation under transcript supervision. Our method exploits essential contextual information from temporally global but spatially local human-object interactions, and dynamically adapts its network parameters according to the videos on the fly. AdaAct achieves state-of-the-art results on two instructional video datasets for both action segmentation and alignment tasks, and especially shows strong capability in distinguishing similar actions.

## Acknowledgements

This research is supported in part by the National Research Foundation of Singapore under the NRF Medium Sized Centre Scheme (CARTIN), and in part by the National Natural Science Foundation of China under Grant 62206147 and Grant 62206153. Any opinions, findings and conclusions expressed in this material are those of the author(s) and do not reflect the views of National Research Foundation, Singapore and National Natural Science Foundation, China.

## References

[Bearman *et al.*, 2016] Amy Bearman, Olga Russakovsky, Vittorio Ferrari, and Li Fei-Fei. What's the point: Semantic segmentation with point supervision. In *ECCV*, pages 549–565, 2016.

[Bojanowski *et al.*, 2014] Piotr Bojanowski, Rémi Lajugie, Francis Bach, Ivan Laptev, Jean Ponce, Cordelia Schmid, and Josef Sivic. Weakly supervised action labeling in videos under ordering constraints. In *ECCV*, pages 628–643, 2014.

[Chang *et al.*, 2019] 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 *CVPR*, pages 3546–3555, 2019.

[Chen *et al.*, 2021] Mingfei Chen, Yue Liao, Si Liu, Zhiyuan Chen, Fei Wang, and Chen Qian. Reformulating hoi detection as adaptive set prediction. In *CVPR*, pages 9004–9013, 2021.

[Chung *et al.*, 2014] 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.

[Ding and Xu, 2018] Li Ding and Chenliang Xu. Weakly-supervised action segmentation with iterative soft boundary assignment. In *CVPR*, pages 6508–6516, 2018.

[Fang *et al.*, 2018] Hao-Shu Fang, Jinkun Cao, Yu-Wing Tai, and Cewu Lu. Pairwise body-part attention for recognizing human-object interactions. In *ECCV*, pages 51–67, 2018.

[Farha and Gall, 2019] Yazan Abu Farha and Jurgen Gall. Ms-tcn: Multi-stage temporal convolutional network for action segmentation. In *CVPR*, pages 3575–3584, 2019.

[Fayyaz and Gall, 2020] Mohsen Fayyaz and Jurgen Gall. Sct: Set constrained temporal transformer for set supervised action segmentation. In *CVPR*, pages 501–510, 2020.

[Ha *et al.*, 2016] David Ha, Andrew Dai, and Quoc V Le. Hypernetworks. *arXiv preprint arXiv:1609.09106*, 2016.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[Huang *et al.*, 2016] De-An Huang, Li Fei-Fei, and Juan Carlos Niebles. Connectionist temporal modeling for weakly supervised action labeling. In *ECCV*, pages 137–153, 2016.

[Huang *et al.*, 2020] Yifei Huang, Yusuke Sugano, and Yoichi Sato. Improving action segmentation via graph-based temporal reasoning. In *CVPR*, pages 14024–14034, 2020.

[Karaman *et al.*, 2014] Svebor Karaman, Lorenzo Seidenari, and Alberto Del Bimbo. Fast saliency based pooling of fisher encoded dense trajectories. In *ECCV THUMOS Workshop*, volume 1, page 5, 2014.

[Kim *et al.*, 2020] Bumsoo Kim, Taeho Choi, Jaewoo Kang, and Hyunwoo J Kim. Uniondet: Union-level detector towards real-time human-object interaction detection. In *ECCV*, pages 498–514, 2020.

[Kuehne *et al.*, 2014] Hilde Kuehne, Ali Arslan, and Thomas Serre. The language of actions: Recovering the syntax and semantics of goal-directed human activities. In *CVPR*, pages 780–787, 2014.

[Kuehne *et al.*, 2016] Hilde Kuehne, Juergen Gall, and Thomas Serre. An end-to-end generative framework for video segmentation and recognition. In *WACV*, pages 1–8, 2016.

[Kuehne *et al.*, 2017] Hilde Kuehne, Alexander Richard, and Juergen Gall. Weakly supervised learning of actions from transcripts. *Computer Vision and Image Understanding*, 163:78–89, 2017.

[Lea *et al.*, 2017] Colin Lea, Michael D Flynn, Rene Vidal, Austin Reiter, and Gregory D Hager. Temporal convolutional networks for action segmentation and detection. In *CVPR*, pages 156–165, 2017.

[Lei and Todorovic, 2018] Peng Lei and Sinisa Todorovic. Temporal deformable residual networks for action segmentation in videos. In *CVPR*, pages 6742–6751, 2018.

[Li and Todorovic, 2020] Jun Li and Sinisa Todorovic. Set-constrained viterbi for set-supervised action segmentation. In *CVPR*, pages 10820–10829, 2020.

[Li *et al.*, 2019a] Jun Li, Peng Lei, and Sinisa Todorovic. Weakly supervised energy-based learning for action segmentation. In *ICCV*, pages 6243–6251, 2019.

[Li *et al.*, 2019b] Yong-Lu Li, Siyuan Zhou, Xijie Huang, Liang Xu, Ze Ma, Hao-Shu Fang, Yanfeng Wang, and Cewu Lu. Transferable interactiveness knowledge for human-object interaction detection. In *CVPR*, pages 3585–3594, 2019.

[Li *et al.*, 2020] Shi-Jie Li, Yazan AbuFarha, Yun Liu, Ming-Ming Cheng, and Juergen Gall. Ms-tcn++: Multi-stage temporal convolutional network for action segmentation. *TPAMI*, 2020.

[Li *et al.*, 2021] Zhe Li, Yazan Abu Farha, and Jurgen Gall. Temporal action segmentation from timestamp supervision. In *CVPR*, pages 8365–8374, 2021.

[Li *et al.*, 2022] Muheng Li, Lei Chen, Yueqi Duan, Zhilan Hu, Jianjiang Feng, Jie Zhou, and Jiwen Lu. Bridge-prompt: Towards ordinal action understanding in instructional videos. In *CVPR*, pages 19880–19889, 2022.

[Liao *et al.*, 2020] Yue Liao, Si Liu, Fei Wang, Yanjie Chen, Chen Qian, and Jiashi Feng. Ppdm: Parallel point detection and matching for real-time human-object interaction detection. In *CVPR*, pages 482–490, 2020.

[Lu and Elhamifar, 2021] Zijia Lu and Ehsan Elhamifar. Weakly-supervised action segmentation and alignment via transcript-aware union-of-subspaces learning. In *ICCV*, pages 8085–8095, 2021.

[Neubeck and Van Gool, 2006] Alexander Neubeck and Luc Van Gool. Efficient non-maximum suppression. In *ICPR*, volume 3, pages 850–855, 2006.

[Park *et al.*, 2022] Junyong Park, Daekyum Kim, Sejoon Huh, and Sungho Jo. Maximization and restoration: Action segmentation through dilation passing and temporal reconstruction. *Pattern Recognition*, 129:108764, 2022.

[Pirsiavash and Ramanan, 2014] Hamed Pirsiavash and Deva Ramanan. Parsing videos of actions with segmental grammars. In *CVPR*, pages 612–619, 2014.

[Qi *et al.*, 2018] Siyuan Qi, Wenguan Wang, Baoxiong Jia, Jianbing Shen, and Song-Chun Zhu. Learning human-object interactions by graph parsing neural networks. In *ECCV*, pages 401–417, 2018.

[Richard *et al.*, 2017] Alexander Richard, Hilde Kuehne, and Juergen Gall. Weakly supervised action learning with rnn based fine-to-coarse modeling. In *CVPR*, pages 754–763, 2017.

[Richard *et al.*, 2018] Alexander Richard, Hilde Kuehne, Ahsan Iqbal, and Juergen Gall. Neuralnetwork-viterbi: A framework for weakly supervised video learning. In *CVPR*, pages 7386–7395, 2018.

[Rohrbach *et al.*, 2012] Marcus Rohrbach, Sikandar Amin, Mykhaylo Andriluka, and Bernt Schiele. A database for fine grained activity detection of cooking activities. In *CVPR*, pages 1194–1201, 2012.

[Shan *et al.*, 2020] Dandan Shan, Jiaqi Geng, Michelle Shu, and David F Fouhey. Understanding human hands in contact at internet scale. In *CVPR*, pages 9869–9878, 2020.

[Singh *et al.*, 2016] Bharat Singh, Tim K Marks, Michael Jones, Oncel Tuzel, and Ming Shao. A multi-stream bi-directional recurrent neural network for fine-grained action detection. In *CVPR*, pages 1961–1970, 2016.

[Souri *et al.*, 2021] Yaser Souri, Mohsen Fayyaz, Luca Minciullo, Gianpiero Francesca, and Juergen Gall. Fast weakly supervised action segmentation using mutual consistency. *TPAMI*, 2021.

[Stein and McKenna, 2013] Sebastian Stein and Stephen J McKenna. Combining embedded accelerometers with computer vision for recognizing food preparation activities. In *UbiComp*, pages 729–738, 2013.

[Ulutan *et al.*, 2020] Oytun Ulutan, ASM Iftekhar, and Bangalore S Manjunath. Vsgnet: Spatial attention network for detecting human object interactions using graph convolutions. In *CVPR*, pages 13617–13626, 2020.

[Wang *et al.*, 2020a] Tiancai Wang, Tong Yang, Martin Danelljan, Fahad Shahbaz Khan, Xiangyu Zhang, and Jian Sun. Learning human-object interaction detection using interaction points. In *CVPR*, pages 4116–4125, 2020.

[Wang *et al.*, 2020b] Zhenzhi Wang, Ziteng Gao, Limin Wang, Zhifeng Li, and Gangshan Wu. Boundary-aware cascade networks for temporal action segmentation. In *ECCV*, pages 34–51, 2020.

[Yi *et al.*, 2021] Fangqiu Yi, Hongyu Wen, and Tingting Jiang. Asformer: Transformer for action segmentation. *arXiv preprint arXiv:2110.08568*, 2021.

[Zhang *et al.*, 2021] Frederic Z Zhang, Dylan Campbell, and Stephen Gould. Spatially conditioned graphs for detecting human-object interactions. In *ICCV*, pages 13319–13327, 2021.

[Zhou and Chi, 2019] Penghao Zhou and Mingmin Chi. Relation parsing neural network for human-object interaction detection. In *ICCV*, pages 843–851, 2019.

[Zhou *et al.*, 2020] Tianfei Zhou, Wenguan Wang, Siyuan Qi, Haibin Ling, and Jianbing Shen. Cascaded human-object interaction recognition. In *CVPR*, pages 4263–4272, 2020.