

# Timestamp-Supervised Action Segmentation with Graph Convolutional Networks

Hamza Khan      Sanjay Haresh      Awais Ahmed      Shakeeb Siddiqui  
 Andrey Konin      M. Zeeshan Zia      Quoc-Huy Tran

**Abstract**—We introduce a novel approach for temporal activity segmentation with timestamp supervision. Our main contribution is a graph convolutional network, which is learned in an end-to-end manner to exploit both frame features and connections between neighboring frames to generate dense frame-wise labels from sparse timestamp labels. The generated dense frame-wise labels can then be used to train the segmentation model. In addition, we propose a framework for alternating learning of both the segmentation model and the graph convolutional model, which first initializes and then iteratively refines the learned models. Detailed experiments on four public datasets, including 50 Salads, GTEA, Breakfast, and Desktop Assembly, show that our method is superior to the multi-layer perceptron baseline, while performing on par with or better than the state of the art in temporal activity segmentation with timestamp supervision.

## I. INTRODUCTION

Human activity understanding in videos has been an important research topic in the fields of robotics and computer vision, with various applications ranging from human-robot interaction, assisted living, healthcare, home automation to manufacturing [1]–[3]. With the significant research efforts in the last decade, one can expect great results for the task of action recognition [4], [5], where the input video is trimmed and captures a simple action. However, in many real-world applications [6], [7], we are often required to deal with untrimmed videos containing a complex activity. In this paper, we want to tackle one such problem, i.e., temporal activity segmentation, where given an input video capturing a complex activity, the task is to assign each frame of the video to one of the action/sub-activity classes.

The best performing methods for temporal activity segmentation are fully-supervised approaches [8]–[12], where frame-wise action class labels are required during training. However, these frame-wise labels are difficult and costly to obtain. In contrast, unsupervised approaches [13]–[19], which require little annotation, have been developed in the literature. Nevertheless, their performances are significantly worse than the above fully-supervised counterparts, which limits their practical applications. In order to balance between annotation efforts and segmentation accuracies, weakly-supervised approaches [20]–[27], which leverage different forms of weak supervision, have attracted notable research interest. Here, we are particularly interested in

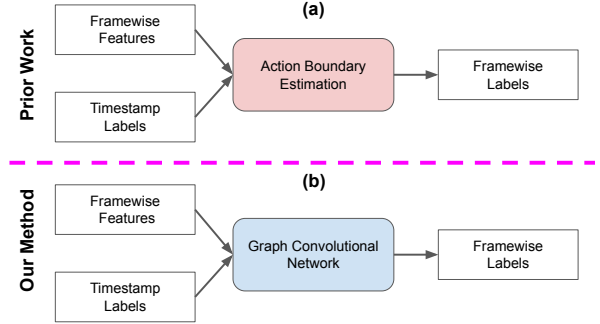


Fig. 1. (a) Li et al. [27] introduce a *heuristic* action boundary estimation module which uses frame-wise features and sparse timestamp labels to generate dense frame-wise labels for training a segmentation model. In particular, it detects action boundaries by minimizing distances between frame features and action centers. (b) In contrast, we propose a *learning* module based on graph convolutional network. More specifically, it leverages both frame features and connections between neighboring frames to convert sparse timestamp labels to dense frame-wise labels.

timestamp supervision [27] due to its superior accuracy as compared to other types of weak supervision [20]–[26].

In timestamp-supervised setup [27], [28], only one (random) frame is annotated for each action segment in a training video. Naively training the segmentation model with just these sparse timestamp labels leads to suboptimal results since the major (unlabeled) part of the video is not used effectively. In this work, we propose to learn a graph convolutional network to convert sparse timestamp labels to dense frame-wise labels for training the segmentation model (see Fig. 1(b)). Our work is motivated by [29], where a graph convolutional network is utilized for effectively propagating labels from only a few labeled nodes to the remaining unlabeled nodes. In our work, the graph convolutional network is learned in an end-to-end manner to exploit not only frame features but also connections between neighboring frames. This is in contrast to the heuristic action boundary estimation model in [27], which detects action boundaries by minimizing distances between frame features and action centers to generate dense frame-wise labels (see Fig. 1(a)). In addition, to effectively learn both the segmentation model and the graph convolutional model, we introduce an alternating learning framework, which first initializes the learned models from scratch and then iteratively improves them.

In summary, our contributions include:

- We present a novel method for activity segmentation with timestamp supervision. Our main idea is to learn a graph convolutional network, which exploits both

frame features and connections between neighboring frames to transform sparse timestamp labels to dense framewise labels. Moreover, we develop a framework for alternating learning of both the segmentation model and the graph convolutional network.

- Extensive evaluations demonstrate that our method outperforms the multi-layer perceptron baseline, and performs on par with or better than the state of the art in timestamp-supervised activity segmentation on 50 Salads, GTEA, Breakfast, and Desktop Assembly datasets.

## II. RELATED WORK

Below we summarize related works in temporal activity segmentation and graph convolutional networks with a focus on applications in video understanding problems.

**Fully-Supervised Activity Segmentation.** Temporal activity segmentation plays an important role in understanding human activities [30]–[32]. Fully-supervised methods reason about long-range dependencies in videos, which are typically tackled by using a two-stage approach of feature extraction and temporal reasoning with an HMM or RNN [8]–[10]. Recent approaches [11], [12] show that single-stage temporal convolutions are capable of modeling long-range dependencies effectively. Despite satisfactory results, the above approaches need dense framewise labels for fully-supervised training. Our method, on the other hand, requires weak supervision in the form of sparse timestamp labels.

**Weakly-Supervised Activity Segmentation.** Many works have focused on reducing the amount of annotations by using transcript supervision [20]–[23], i.e., the order of actions occurring in a given video. They mostly rely on aligning frames with transcripts by using different alignment techniques such as Dynamic Time Warping [23] and Viterbi [33]. Other approaches have also experimented with relaxing the ordering assumption by using only a set of actions for supervision, namely set supervision [24]–[26]. Although these algorithms are efficient in terms of significantly less annotation requirement, they perform notably worse than fully-supervised approaches. To alleviate that, Li et al. [27] propose to use timestamp supervision, i.e., for each action segment in a training video, only one (random) frame is annotated. Our method also uses timestamp supervision, however, instead of heuristic action boundary estimation in [27], we learn a graph convolutional network to generate dense framewise labels.

**Unsupervised Activity Segmentation.** Unsupervised activity segmentation has received considerable research interest due to little annotation requirement. Early methods [13]–[15] leverage narration cues from accompanying scripts to segment videos. Recent methods [16], [17] use pretext tasks for learning frame representations before clustering them by using  $K$ -means clustering to form action clusters. Furthermore, Li et al. [18] exploit action-level cues to boost the performance. More recently, Kumar et al. [19] perform frame representation learning and clustering in a joint framework via temporal optimal transport. Despite little annotation

requirement, the above methods perform significantly worse than fully-supervised or weakly-supervised approaches.

**Graph Convolutional Networks.** Graph convolutional networks, e.g., [29], have been shown to work more effectively on graph structured data in label-scarce settings than classical neural network architectures. They have also been applied recently to various video understanding problems such as action recognition [34], object-object reasoning [35], and action localization [36]. While the above works have focused on using graph convolutional networks for video understanding problems in fully-supervised settings, we leverage graph convolutional networks to generate dense framewise labels for activity segmentation in semi-supervised settings.

## III. OUR APPROACH

We present, in this section, the details of our approach. We first describe the task of temporal activity segmentation and timestamp supervision in Sec. III-A. Next, the proposed graph convolutional network module for generating dense framewise labels is introduced in Sec. III-B. Lastly, we propose a framework for alternating learning of the segmentation model and the graph convolutional network in Sec. III-C.

### A. Activity Segmentation with Timestamp Supervision

Temporal activity segmentation aims to associate each frame of an input video capturing a complex activity with one of the action/sub-activity classes. In particular, let us denote the input video as  $X = [x_1, x_2, \dots, x_T]$ , where  $T$  is the number of frames in  $X$ . The task is to obtain the action class for each frame of  $X$ , i.e.,  $A = [a_1, a_2, \dots, a_T]$  with  $a_i$  representing the action class for frame  $x_i$ . For fully-supervised methods [11], [12], [37], [38], a number of videos  $\mathcal{X} = \{X\}$  with corresponding framewise labels  $\mathcal{A} = \{A\}$  are given for training. These methods focus on designing an effective segmentation model to extract useful appearance and temporal cues in the videos and often produce satisfactory results. However, framewise annotations for all training videos are generally difficult and prohibitively expensive to obtain.

In this work, we are interested in the timestamp-supervised setting [27], [28]. Specifically, for each action segment in a training video, only one (random) frame is labeled. Assuming a training video  $X$  has  $N$  action segments (in general,  $N \ll T$ ), we can denote the timestamp labels for  $X$  as  $A_{TS} = [a_{t_1}, a_{t_1}, \dots, a_{t_N}]$ , where frame  $t_i$  belongs to the  $i$ -th action segment in  $X$ . According to [39], timestamp-supervised setting requires a significantly less (i.e., one sixth) labeling duration as compared to the above fully-supervised setting. Nevertheless, simply using the sparse timestamp labels for training the segmentation model (i.e., applying the classification loss with these sparse timestamp labels) leads to inferior results since the remaining (unlabeled) frames are not utilized effectively. To address that, we propose, in the next section, a graph convolutional network module for generating dense framewise labels from sparse timestamp labels so that all frames can be employed for training.

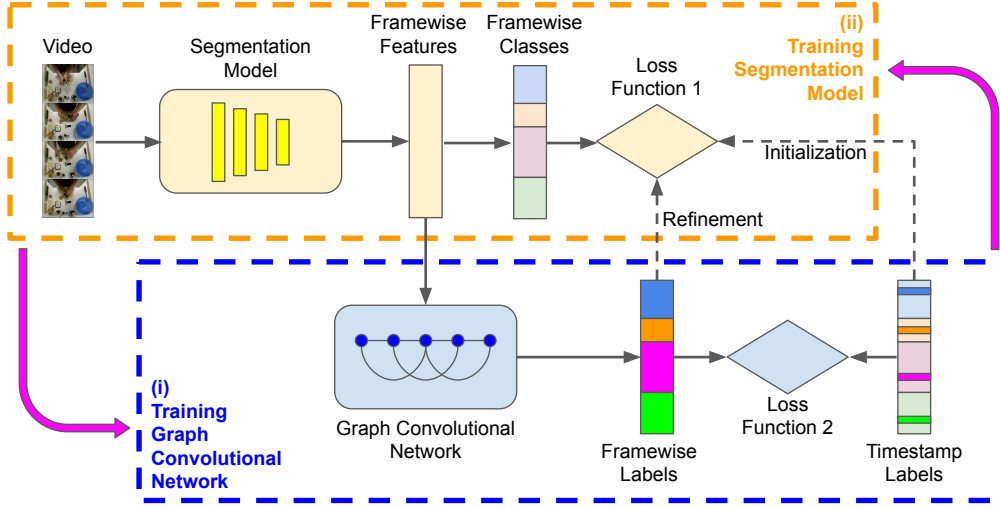


Fig. 2. Our alternating learning framework is shown above. During training, we first initialize the segmentation model by training it with sparse timestamp labels (namely, the *initialization* stage). After that, we alternate between (i) training the graph convolutional network, which takes frame-wise features from the segmentation model as input, uses sparse timestamp labels for supervision, and generates dense frame-wise labels as output, and (ii) training the segmentation model, which uses dense frame-wise labels from the graph convolutional network for supervision (namely, the *refinement* stage). During testing, the graph convolutional network is discarded while the segmentation model is employed to provide segmentation results.

### B. Graph Convolutional Network for Label Generation

Let us represent each video as a graph, where frames (along with their features) are nodes and there exist edges between neighboring frames (within a temporal window). Given only a few nodes with labels (i.e., sparse timestamp labels), we want to classify the remaining nodes in the graph. This problem, namely graph-based semi-supervised learning, is typically solved by minimizing the following loss function:

$$\mathcal{L} = \mathcal{L}_{class} + \lambda \mathcal{L}_{reg}, \quad (1)$$

where  $\mathcal{L}_{class}$  denotes the classification loss on the labeled nodes,  $\mathcal{L}_{reg}$  represents the *explicit* graph-based regularization term [40]–[43], and  $\lambda$  is the balancing parameter.

Inspired by [29], we encode the graph structure directly by using a graph neural network  $f(\mathbf{X}, \mathbf{A})$ . Here,  $\mathbf{X}$  denotes the node features,  $\mathbf{A}$  represents the adjacency matrix, and  $f$  is a graph neural network. That enables us to train  $f$  with only  $\mathcal{L}_{class}$  since  $\mathcal{L}_{reg}$  is *implicitly* embedded in  $f$ . Depending on both the node features and the adjacency matrix allows gradient information to be distributed from the labeled nodes to the unlabeled ones. Thus, the model is able to learn effective representations for both labeled and unlabeled nodes. In particular, we use a graph convolutional network with the below propagation rule:

$$\mathbf{H}_{l+1} = \sigma \left( \tilde{\mathbf{D}}^{-\frac{1}{2}} \tilde{\mathbf{A}} \tilde{\mathbf{D}}^{-\frac{1}{2}} \mathbf{H}_l \mathbf{W}_l \right). \quad (2)$$

Here,  $\tilde{\mathbf{A}} = \mathbf{A} + \mathbf{I}$  is the adjacency matrix with added self-connections (represented by the identity matrix  $\mathbf{I}$ ), while  $\tilde{\mathbf{D}}$  is the degree matrix of  $\tilde{\mathbf{A}}$ . Next,  $\mathbf{W}_l$ ,  $\sigma$ , and  $\mathbf{H}_l$  are the weight matrix for layer  $l$ , the activation function, and the activation matrix for layer  $l$  respectively. As demonstrated in [29], the above graph convolutional network can be derived from a first-order approximation of spectral graph convolutions [44].

As we empirically show in Sec. IV-A, using weighted graphs yields better results than using binary graphs. For weighted graphs, we define the edge weight between nodes  $i$  and  $j$  as the cosine similarity between corresponding features  $\mathbf{x}_i$  and  $\mathbf{x}_j$ , which is written as:

$$\mathbf{A}_{ij} = \frac{\mathbf{x}_i \cdot \mathbf{x}_j}{\|\mathbf{x}_i\| \|\mathbf{x}_j\|}, \quad (3)$$

where  $\cdot$  denotes the dot product.

### C. Alternating Learning Framework

In the following, we present our alternating learning framework for training the segmentation model and the graph convolutional network. In particular, we divide the training into two stages, i.e., the *initialization* stage and the *refinement* stage. During the *initialization* stage, we train only the segmentation model with sparse timestamp labels for  $\epsilon$  epochs ( $\gamma = 30$ ). Next, in the *refinement* stage, we perform  $\gamma$  iterations of alternating learning ( $\epsilon = 20$ ). For each iteration, we (i) first train the graph convolutional network for 300 epochs, where frame-wise features from the segmentation model are input, sparse timestamp labels are supervision signals, and dense frame-wise labels are output, and (ii) then train the segmentation model for 3 epochs, where dense frame-wise labels from the graph convolutional network are supervision signals. At testing, the graph convolutional network is discarded, and the segmentation model is used to produce the segmentation result. Fig. 2 summarizes our alternating learning framework. As we will show later in Sec. IV-C, our alternating learning framework for training the segmentation model and the graph convolutional network outperforms the joint learning counterpart.

We use the conventional combination of classification loss and smoothing loss [11], [12], [37], [38] for training the segmentation model and graph convolutional network. In

addition, we include the confidence loss [27] in the training of the segmentation model to boost its performance. We summarize those losses below.

**Classification Loss.** We apply the cross-entropy loss between the predicted probabilities and the (generated) action labels as:

$$\mathcal{L}_{class} = \frac{1}{T} \sum_t -\log \tilde{y}_{t,a}, \quad (4)$$

where  $T$  is the number of frames in the video and  $\tilde{y}_{t,a}$  is the predicted probability that frame  $x_t$  is assigned to action class  $a$ .

**Smoothing Loss.** We employ the smoothing loss to tackle the problem of over-segmentation as:

$$\mathcal{L}_{smooth} = \frac{1}{TC} \sum_{t,a} \tilde{\Delta}_{t,a}^2, \quad (5)$$

$$\tilde{\Delta}_{t,a} = \begin{cases} \Delta_{t,a}, & \Delta_{t,a} \leq \tau \\ \tau, & \Delta_{t,a} > \tau \end{cases}, \quad (6)$$

$$\Delta_{t,a} = |\log \tilde{y}_{t,a} - \log \tilde{y}_{t-1,a}|, \quad (7)$$

where  $C$  is the number of action classes in the activity and  $\tau = 4$  is the thresholding parameter.

**Confidence Loss.** We adopt the confidence loss from [27] to encourage the predicted probabilities to monotonically decrease as the distance to the timestamps increases as:

$$\mathcal{L}_{conf} = \frac{1}{T'} \sum_{a_{t_i} \in A_{TS}} \left( \sum_{t=t_{i-1}}^{t_{i+1}} \delta_{a_{t_i},t} \right), \quad (8)$$

$$\delta_{a_{t_i},t} = \begin{cases} \max(0, \log \tilde{y}_{t,a_{t_i}} - \log \tilde{y}_{t-1,a_{t_i}}), & t \leq t_i \\ \max(0, \log \tilde{y}_{t-1,a_{t_i}} - \log \tilde{y}_{t,a_{t_i}}), & t > t_i \end{cases}, \quad (9)$$

where  $t_i$  and  $a_{t_i}$  are the  $i$ -th timestamp and its corresponding action label,  $\tilde{y}_{t,a_{t_i}}$  is the predicted probability that frame  $x_t$  is assigned to action class  $a_{t_i}$ , and  $T' = 2(t_N - t_1)$  is the number of frames contributing to the loss.

**Final Losses.** The final losses  $\mathcal{L}_{seg}$  and  $\mathcal{L}_{graph}$  respectively for training the segmentation model and the graph convolutional network are written as follows:

$$\mathcal{L}_{seg} = \mathcal{L}_{class} + \alpha \mathcal{L}_{smooth} + \beta \mathcal{L}_{conf}, \quad (10)$$

$$\mathcal{L}_{graph} = \mathcal{L}_{class} + \alpha \mathcal{L}_{smooth}. \quad (11)$$

Here,  $\alpha = 0.15$  and  $\beta = 0.075$  are balancing parameters. We have tried adding the confidence loss  $\mathcal{L}_{conf}$  to the final loss  $\mathcal{L}_{graph}$  for training the graph convolutional network but did not get any performance gain.

#### IV. EXPERIMENTS

In this section, we benchmark our approach for timestamp-supervised activity segmentation on various datasets, including 50 Salads, GTEA, Breakfast, and Desktop Assembly.

**Implementation Details.** For a fair comparison, we follow [27] to adopt the multi-stage temporal convolutional network of [11] as our segmentation model. The I3D features [45] used as input in [27] are also used as input to our segmentation model. In addition, we implement a

two-layer graph convolutional network for label generation. The first layer maps the input features (64-dimensional vectors) to 32-dimensional vectors, which are subsequently passed through ReLU activation, the second layer, and lastly softmax classification. To construct the graph from an input video, we consider frames along with their features as nodes and connect each frame with its preceding 15 frames and succeeding 15 frames (yielding a temporal window size of 31) to form pairwise edges. The input features to the graph convolutional network are the output of the penultimate layer of the segmentation model and are 64-dimensional vectors. The segmentation model and the graph convolutional network are learned via backpropagation respectively through the losses in Sec. III-C. We implement our approach in PyTorch [46]. We use the ADAM optimizer [47] and a batch size of 8. The learning rate is set to 0.0005 and 0.01 respectively for the segmentation model and the graph convolutional network. We use a weight decay of 0.0005 only for the graph convolutional network. Our experiments are conducted with an Nvidia V100 GPU on Microsoft Azure.

**Competing Methods.** We compare our approach (namely “GCN”, short for Graph Convolutional Network) against the state-of-the-art method of [27]<sup>1</sup> (namely “ABE”, short for Action Boundary Estimation) for timestamp-supervised activity segmentation. Also, we add a fully-supervised baseline (namely “Baseline”), which has the same architecture as our segmentation model but is trained with ground truth dense framewise labels for full supervision. Lastly, we include the results of recent fully-supervised methods [11], [12], [37], [38], transcript-supervised methods [10], [20], [22], [23], [33], [48], [49], and set-supervised methods [24]–[26].

**Datasets.** We test the performance on four public datasets, namely 50 Salads [50], Breakfast [51], GTEA [52], and Desktop Assembly [19]. We summarize the datasets below:

- **50 Salads:** The dataset consists of 50 videos with actors preparing different kinds of salads. It contains 0.6M frames annotated with one of the 17 action classes, and the average video duration varies from 5 to 10 minutes.
- **Breakfast:** The dataset includes 1712 videos with actors preparing various types of breakfast. It consists of 3.6M frames annotated with one of the 48 action classes, and the average video duration ranges from a few seconds to several minutes.
- **GTEA:** The dataset consists of 28 videos with actors performing various kinds of daily activities. It contains 32K frames annotated with one of the 11 action classes, and the average video duration is 1 minute.
- **Desktop Assembly:** The dataset includes 76 videos with actors performing the desktop assembly activity. It consists of 59K frames annotated with one of the 22 action classes, and the average video duration is 1.5 minutes.

We use the same timestamp labels for our method and ABE.

**Metrics.** We follow [27] to report the results on five metrics, including framewise accuracy (Acc), edit distance (Edit), and

<sup>1</sup>We use the original code provided by the authors at <https://github.com/ZheLi2020/TimeStampActionSeg>



	Edge	Window	F1@10	F1@25	F1@50	Edit	Acc
<b>F</b>	Binary	3	73.7	70.5	59.1	65.7	73.6
		7	73.9	71.0	59.2	66.4	74.1
		17	74.2	71.1	58.9	66.5	74.7
		31	74.6	71.7	59.6	67.2	74.3
	Weighted	3	74.8	71.7	59.7	66.8	74.8
		7	74.9	72.1	60.3	67.3	74.8
		17	73.7	70.5	59.6	66.6	74.2
		31	<b>75.1</b>	<b>72.3</b>	<b>61.0</b>	<b>67.6</b>	<b>75.1</b>
<b>G</b>	Binary	3	78.4	73.9	58.7	72.0	64.7
		7	79.2	75.4	59.4	72.8	65.2
		17	79.8	76.3	59.2	73.3	65.4
		31	79.3	75.7	57.5	73.4	65.5
	Weighted	3	78.1	74.5	58.9	72.6	64.6
		7	79.6	75.9	60.1	73.0	65.5
		17	79.6	75.2	59.8	73.1	65.9
		31	<b>81.5</b>	<b>77.5</b>	<b>60.8</b>	<b>75.6</b>	<b>66.1</b>

TABLE I

ABLATION STUDY ON GRAPH CONSTRUCTION. BEST RESULTS ARE IN BOLD. **F** DENOTES 50 SALADS AND **G** DENOTES GTEA.

F1 scores with overlapping thresholds of 10%, 25%, and 50%. To reduce the impact of randomness, we use  $K$ -fold cross validation with  $K = 5$  for 50 Salads and  $K = 4$  for GTEA, Breakfast, and Desktop Assembly. Furthermore, for timestamp-supervised methods, including ours and ABE, we run each 3 times with 3 random seeds and report the *average* results over the 3 runs. We would like to note that this setup is different from the setup in [27], where the results are reported over a single run. For other types of methods, we simply obtain their results from [27].

#### A. Ablation Study on Graph Construction

We first conduct some ablation study experiments to understand the impacts of graph structures on our approach. In particular, we evaluate the performance of our method when using different types of graphs, i.e., binary graphs and weighted graphs, as well as various temporal window sizes, i.e., 3, 7, 17, and 31. For weighted graphs, the edge weight is defined as in Eq. 3. Tab. I presents the ablation study results on graph construction on the 50 Salads and GTEA datasets. From the results, temporal window size of 3 performs the worst, while larger temporal window sizes, which are able to capture longer-range dependencies despite having higher computational costs, often yield better results. Next, temporal window size of 31 has the best performance on both 50 Salads and GTEA datasets. Further, we have tried with larger temporal window sizes than 31, however, they yield little performance gains while having notably higher computational costs. In addition, by leveraging additional in-

	Method	F1@10	F1@25	F1@50	Edit	Acc
<b>F</b>	MLP	70.8	68.0	57.6	63.6	73.1
	GCN	<b>75.1</b>	<b>72.3</b>	<b>61.0</b>	<b>67.6</b>	<b>75.1</b>
<b>G</b>	MLP	77.8	74.4	59.1	72.2	65.6
	GCN	<b>81.5</b>	<b>77.5</b>	<b>60.8</b>	<b>75.6</b>	<b>66.1</b>

TABLE II

ABLATION STUDY ON MLP VS. GCN. BEST RESULTS ARE IN BOLD. **F** DENOTES 50 SALADS AND **G** DENOTES GTEA.

	$\gamma$	$\epsilon$	F1@10	F1@25	F1@50	Edit	Acc
<b>F</b>	0	50	43.5	36.6	24.6	42.2	38.2
	10	40	51.3	44.9	33.1	48.5	50.4
	30	20	<b>75.1</b>	<b>72.3</b>	<b>61.0</b>	<b>67.6</b>	<b>75.1</b>
	50	0	58.1	54.0	43.3	48.7	71.2
<b>G</b>	0	50	21.6	20.3	15.9	16.9	29.6
	10	40	22.8	21.7	16.9	17.3	29.9
	30	20	<b>81.5</b>	<b>77.5</b>	<b>60.8</b>	<b>75.6</b>	<b>66.1</b>
	50	0	63.0	58.4	45.1	56.2	57.1

TABLE III

ABLATION STUDY ON ALTERNATING LEARNING VS. JOINT LEARNING. BEST RESULTS ARE IN BOLD. **F** DENOTES 50 SALADS AND **G** DENOTES GTEA.

formation (i.e., similarity between frame features), weighted graphs usually outperform binary graphs on both 50 Salads and GTEA datasets. For all the remaining experiments in this section, we will use temporal window size of 31 and weighted graphs for our method.

#### B. Ablation Study on MLP vs. GCN

As mentioned in Sec. III-B, the graph convolutional network embeds the graph-based regularization term  $\mathcal{L}_{reg}$  implicitly, which helps it exploit cues available in the graph (i.e., frame features and connections between neighboring frames). In this section, we study the benefit of leveraging these cues via a graph convolutional network as compared to using a plain multi-layer perceptron network. Tab. II shows the results on the 50 Salads and GTEA datasets when using a graph convolutional network (GCN) and a multi-layer perceptron network (MLP) for label generation. Note that they have the *same* set of hyperparameters, e.g., number of layers, number of output channels, etc. As we can see from Tab. II, GCN outperforms MLP across all metrics on both datasets, which confirms the advantage of using GCN over MLP for label generation.

#### C. Ablation Study on Alternating Learning vs. Joint Learning

Here, we conduct experiments to support our choice of alternating learning over joint learning of the segmentation model and the graph convolutional network. In particular,

	Method	F1@10	F1@25	F1@50	Edit	Acc
<b>F</b>	Baseline	70.8	67.7	58.6	63.8	77.8
	MS-TCN [11]	76.3	74.0	64.5	67.9	80.7
	MS-TCN++ [12]	80.7	78.5	70.1	74.3	83.7
	BCN [37]	82.3	81.3	74.0	74.3	84.4
	ASRF [38]	<b>84.9</b>	<b>83.5</b>	<b>77.3</b>	<b>79.3</b>	<b>84.5</b>
<b>Ti</b>	ABE [27]	73.7	71.0	60.1	66.1	<b>76.0</b>
	GCN (Ours)	<b>75.1</b>	<b>72.3</b>	<b>61.0</b>	<b>67.6</b>	75.1
<b>Tr</b>	CDFL [48]	-	-	-	-	<b>54.7</b>
	NN-Viterbi [33]	-	-	-	-	49.4
	HMM-RNN [10]	-	-	-	-	45.5

TABLE IV

COMPARISON ON 50 SALADS. BEST RESULTS ARE IN BOLD. **F** DENOTES FULL SUPERVISION, **Ti** DENOTES TIMESTAMP SUPERVISION, AND **Tr** DENOTES TRANSCRIPT SUPERVISION.

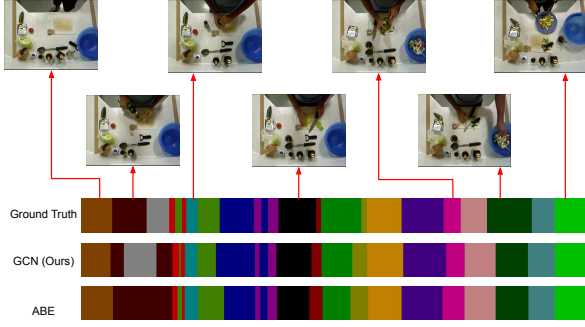


Fig. 3. Segmentation results on a 50 Salads video.

we compare the performance of joint learning (i.e., without the initialization stage or  $\{\gamma, \epsilon\} = \{0, 50\}$ ) with two versions of alternating learning (i.e.,  $\{\gamma, \epsilon\} = \{10, 40\}$  and  $\{\gamma, \epsilon\} = \{30, 20\}$ ). Moreover, we compare with the naive learning of the segmentation model using only sparse timestamp labels (i.e., without the refinement stage or  $\{\gamma, \epsilon\} = \{50, 0\}$ ). The results on the 50 Salads and GTEA datasets are shown in Tab. III. It is evident from the results that alternating learning outperforms both joint learning and naive learning by large margins. Next, the bad performance of joint learning is likely because as we train both models from scratch, during the first few iterations, the features extracted by the segmentation model and hence the framewise labels predicted by the graph convolutional network are not meaningful. As a result, the joint learning framework is unstable and yields worse results than the alternating learning one. Lastly, the worse performance of naive learning is likely because unlabeled frames in the video is not utilized effectively. In the following experiments, we will use alternating learning with  $\{\gamma, \epsilon\} = \{30, 20\}$  for our method.

	Method	F1@10	F1@25	F1@50	Edit	Acc
<b>F</b>	Baseline	85.1	82.7	69.6	79.6	76.1
	MS-TCN [11]	85.8	83.4	69.8	79.0	76.3
	MS-TCN++ [12]	88.8	85.7	76.0	83.5	<b>80.1</b>
	BCN [37]	88.5	87.1	77.3	<b>84.4</b>	79.8
	ASRF [38]	<b>89.4</b>	<b>87.8</b>	<b>79.8</b>	83.7	77.3
<b>Ti</b>	ABE [27]	77.7	73.8	58.1	72.1	<b>67.7</b>
	GCN (Ours)	<b>81.5</b>	<b>77.5</b>	<b>60.8</b>	<b>75.6</b>	66.1

TABLE V

COMPARISON ON GTEA. BEST RESULTS ARE IN BOLD. **F** DENOTES FULL SUPERVISION AND **Ti** DENOTES TIMESTAMP SUPERVISION.

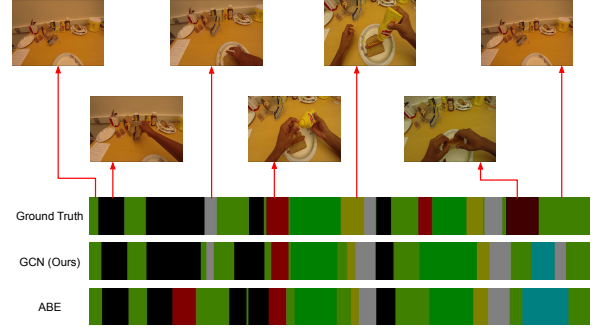


Fig. 4. Segmentation results on a GTEA video.

#### D. Comparison on 50 Salads

We now compare the performance of our approach against that of the state-of-the-art timestamp-supervised method of [27] (ABE), on the 50 Salads dataset. In addition, we include the results of fully-supervised methods [11], [12], [37], [38] and transcript-supervised methods [10], [33], [48]. Tab. IV presents the results. As we see from Tab. IV, our approach outperforms ABE on F1@10, F1@25, F1@50, and Edit, while performing on par with ABE on Acc. Lower F1 and Edit indicate that ABE is more likely to suffer from over-segmentation, since over-segmented results may have high Acc but yield low F1 and Edit. In addition, timestamp-supervised methods produce better results than transcript-supervised ones. Fig. 3 visualizes some qualitative results of our approach and ABE, where our result is closer to the ground truth. Please see also our supplementary video<sup>2</sup>.

#### E. Comparison on GTEA

Tab. V presents the results of timestamp-supervised methods, i.e., ours and ABE [27], and fully-supervised methods [11], [12], [37], [38] on the GTEA dataset. Similar to the results on 50 Salads, our approach achieves higher F1@10, F1@25, F1@50, and Edit, and lower Acc than ABE on GTEA. This also shows our method is less suffering from over-segmentation as compared to ABE. In addition, some

<sup>2</sup>Our supplementary video is available at <https://youtu.be/tvV3soPMTIo>

	Method	F1@10	F1@25	F1@50	Edit	Acc
<b>F</b>	Baseline	69.9	64.2	51.5	69.4	68.0
	MS-TCN [11]	52.6	48.1	37.9	61.7	66.3
	MS-TCN++ [12]	64.1	58.6	45.9	65.6	67.6
	BCN [37]	68.7	65.5	55.0	66.2	<b>70.4</b>
	ASRF [38]	<b>74.3</b>	<b>68.9</b>	<b>56.1</b>	<b>72.4</b>	67.6
<b>Ti</b>	ABE [27]	67.4	60.8	44.9	<b>68.5</b>	<b>63.1</b>
	GCN (Ours)	<b>67.9</b>	<b>61.0</b>	<b>45.3</b>	67.0	61.4
<b>Tr</b>	CDFL [48]	-	-	-	-	<b>50.2</b>
	MuCon [49]	-	-	-	-	47.1
	$D^3$ TW [23]	-	-	-	-	45.7
	NN-Viterbi [33]	-	-	-	-	43.0
	TCFPN [22]	-	-	-	-	38.4
	HMM-RNN [10]	-	-	-	-	33.3
	ECTC [20]	-	-	-	-	27.7
<b>S</b>	SCT [25]	-	-	-	-	<b>30.4</b>
	SCV [26]	-	-	-	-	30.2
	Action Sets [24]	-	-	-	-	23.3

TABLE VI

COMPARISON ON BREAKFAST. BEST RESULTS ARE IN BOLD. **F** DENOTES FULL SUPERVISION, **Ti** DENOTES TIMESTAMP SUPERVISION, **Tr** DENOTES TRANSCRIPT SUPERVISION, AND **S** DENOTES SET SUPERVISION.



Fig. 5. Segmentation results on a Breakfast video.

qualitative results of our method and ABE are plotted in Fig. 4, where our result is closer to the ground truth.

#### F. Comparison on Breakfast

The results of methods with different forms of supervision, including full supervision [11], [12], [37], [38], timestamp supervision (ours and ABE [27]), transcript supervision [10], [20], [22], [23], [33], [48], [49], and set supervision [24]–[26], on the Breakfast dataset are shown in Tab. VI. From Tab. VI, our approach obtains similar results as ABE (i.e., our method has higher F1@10, F1@25, and F1@50, and lower Edit and Acc). In addition, timestamp-supervised methods yield better results than other weakly-supervised ones. Fig. 5

	Method	F1@10	F1@25	F1@50	Edit	Acc
<b>F</b>	Baseline	90.2	87.2	76.9	<b>89.9</b>	79.4
	ASRF [38]	<b>91.4</b>	<b>90.3</b>	<b>83.2</b>	86.6	<b>81.9</b>
<b>Ti</b>	ABE [27]	89.8	86.4	67.5	<b>88.3</b>	71.7
	GCN (Ours)	<b>90.4</b>	<b>88.0</b>	<b>75.1</b>	87.3	<b>77.1</b>

TABLE VII

COMPARISON ON DESKTOP ASSEMBLY. BEST RESULTS ARE IN BOLD. **F** DENOTES FULL SUPERVISION AND **Ti** DENOTES TIMESTAMP SUPERVISION.

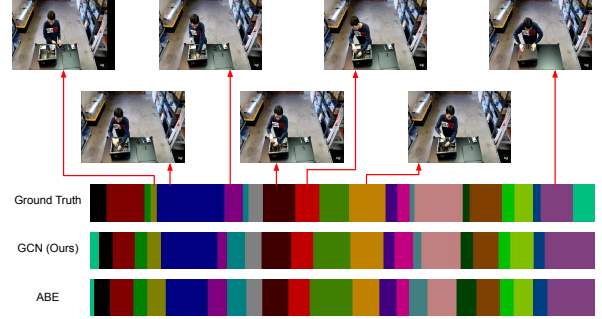


Fig. 6. Segmentation results on a Desktop Assembly video.

plots some qualitative results of our approach and ABE, where the two methods perform similarly.

#### G. Comparison on Desktop Assembly

Tab. VII presents the results of timestamp-supervised methods, namely ours and ABE [27], and fully-supervised methods, including the state-of-the-art method of [38], on the Desktop Assembly dataset. From the results, our approach outperforms ABE on F1@10, F1@25, F1@50, and Acc by substantial margins, while performing on par with ABE on Edit. In addition, Fig. 6 visualizes some qualitative results of our approach and ABE, where our result is closer to the ground truth.

## V. CONCLUSION

We propose, in this paper, a novel method for timestamp-supervised activity segmentation, which utilizes a graph convolutional network for generating dense frame-wise labels from sparse timestamp labels. The graph convolutional network is learned in an end-to-end fashion to leverage not only frame features but also connections between neighboring frames. Moreover, we present a framework for alternating learning of both the segmentation model and the graph convolutional model. We show that our method is superior to the multi-layer perceptron baseline, while performing on par with or better than the state of the art in timestamp-supervised activity segmentation on 50 Salads, GTEA, Breakfast, and Desktop Assembly. Our future work will explore the use of deep supervision [53]–[57] or self-supervised losses [58], [59] for improving the performance.

## REFERENCES

- [1] L. D. Riek, “Healthcare robotics,” *Communications of the ACM*, vol. 60, no. 11, pp. 68–78, 2017.
- [2] T. Iqbal and L. D. Riek, “Human-robot teaming: Approaches from joint action and dynamical systems,” *Humanoid robotics: A reference*, pp. 2293–2312, 2019.
- [3] A. Konin, S. N. Syed, S. Siddiqui, S. Kumar, Q.-H. Tran, and M. Z. Zia, “Retroactivity: Rapidly deployable live task guidance experiences,” in *IEEE International Symposium on Mixed and Augmented Reality Demonstration*, 2020.
- [4] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri, “A closer look at spatiotemporal convolutions for action recognition,” in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2018, pp. 6450–6459.
- [5] X. Wang, R. Girshick, A. Gupta, and K. He, “Non-local neural networks,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7794–7803.
- [6] Y.-W. Chao, S. Vijayanarasimhan, B. Seybold, D. A. Ross, J. Deng, and R. Sukthankar, “Rethinking the faster r-cnn architecture for temporal action localization,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1130–1139.
- [7] D. Gong, L. Liu, V. Le, B. Saha, M. R. Mansour, S. Venkatesh, and A. v. d. Hengel, “Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 1705–1714.
- [8] H. Kuehne, J. Gall, and T. Serre, “An end-to-end generative framework for video segmentation and recognition,” in *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2016, pp. 1–8.
- [9] B. Singh, T. K. Marks, M. Jones, O. Tuzel, and M. Shao, “A multi-stream bi-directional recurrent neural network for fine-grained action detection,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1961–1970.
- [10] A. Richard, H. Kuehne, and J. Gall, “Weakly supervised action learning with rnn based fine-to-coarse modeling,” in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2017, pp. 754–763.
- [11] Y. A. Farha and J. Gall, “Ms-tcn: Multi-stage temporal convolutional network for action segmentation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3575–3584.
- [12] S.-J. Li, Y. AbuFarha, Y. Liu, M.-M. Cheng, and J. Gall, “Ms-tcn++: Multi-stage temporal convolutional network for action segmentation,” *IEEE transactions on pattern analysis and machine intelligence*, 2020.
- [13] J. Malmaud, J. Huang, V. Rathod, N. Johnston, A. Rabinovich, and K. Murphy, “What’s cookin’? interpreting cooking videos using text, speech and vision,” *arXiv preprint arXiv:1503.01558*, 2015.
- [14] O. Sener, A. R. Zamir, S. Savarese, and A. Saxena, “Unsupervised semantic parsing of video collections,” in *Proceedings of the IEEE International conference on Computer Vision*, 2015, pp. 4480–4488.
- [15] J.-B. Alayrac, P. Bojanowski, N. Agrawal, J. Sivic, I. Laptev, and S. Lacoste-Julien, “Unsupervised learning from narrated instruction videos,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4575–4583.
- [16] A. Kukleva, H. Kuehne, F. Sener, and J. Gall, “Unsupervised learning of action classes with continuous temporal embedding,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 12066–12074.
- [17] R. G. VidalMata, W. J. Scheirer, A. Kukleva, D. Cox, and H. Kuehne, “Joint visual-temporal embedding for unsupervised learning of actions in untrimmed sequences,” in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021, pp. 1238–1247.
- [18] J. Li and S. Todorovic, “Action shuffle alternating learning for unsupervised action segmentation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 12628–12636.
- [19] S. Kumar, S. Hareesh, A. Ahmed, A. Konin, M. Z. Zia, and Q.-H. Tran, “Unsupervised activity segmentation by joint representation learning and online clustering,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.
- [20] D.-A. Huang, L. Fei-Fei, and J. C. Niebles, “Connectionist temporal modeling for weakly supervised action labeling,” in *European Conference on Computer Vision*. Springer, 2016, pp. 137–153.
- [21] H. Kuehne, A. Richard, and J. Gall, “Weakly supervised learning of actions from transcripts,” *Computer Vision and Image Understanding*, vol. 163, pp. 78–89, 2017.
- [22] L. Ding and C. Xu, “Weakly-supervised action segmentation with iterative soft boundary assignment,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6508–6516.
- [23] C.-Y. Chang, D.-A. Huang, Y. Sui, L. Fei-Fei, and J. C. 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*, 2019, pp. 3546–3555.
- [24] A. Richard, H. Kuehne, and J. Gall, “Action sets: Weakly supervised action segmentation without ordering constraints,” in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2018, pp. 5987–5996.
- [25] M. Fayyaz and J. Gall, “Sct: Set constrained temporal transformer for set supervised action segmentation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 501–510.
- [26] J. Li and S. Todorovic, “Set-constrained viterbi for set-supervised action segmentation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10820–10829.
- [27] Z. Li, Y. Abu Farha, and J. Gall, “Temporal action segmentation from timestamp supervision,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 8365–8374.
- [28] D. Moltisanti, S. Fidler, and D. Damen, “Action recognition from single timestamp supervision in untrimmed videos,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 9915–9924.
- [29] T. N. Kipf and M. Welling, “Semi-supervised classification with graph convolutional networks,” *arXiv preprint arXiv:1609.02907*, 2016.
- [30] H. Pirsiavash and D. Ramanan, “Parsing videos of actions with segmental grammars,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 612–619.
- [31] F. Caba Heilbron, V. Escorcia, B. Ghanem, and J. Carlos Niebles, “Activitynet: A large-scale video benchmark for human activity understanding,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 961–970.
- [32] D. Damen, H. Doughty, G. M. Farinella, S. Fidler, A. Furnari, E. Kazakos, D. Moltisanti, J. Munro, T. Perrett, W. Price, *et al.*, “Scaling egocentric vision: The epic-kitchens dataset,” in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 720–736.
- [33] A. Richard, H. Kuehne, A. Iqbal, and J. Gall, “Neuralnetwork-viterbi: A framework for weakly supervised video learning,” in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2018, pp. 7386–7395.
- [34] R. Herzig, E. Levi, H. Xu, H. Gao, E. Brosh, X. Wang, A. Globerson, and T. Darrell, “Spatio-temporal action graph networks,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 2019, pp. 0–0.
- [35] S. Hareesh, S. Kumar, M. Z. Zia, and Q.-H. Tran, “Towards anomaly detection in dashcam videos,” in *2020 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, pp. 1407–1414.
- [36] R. Zeng, W. Huang, M. Tan, Y. Rong, P. Zhao, J. Huang, and C. Gan, “Graph convolutional networks for temporal action localization,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 7094–7103.
- [37] Z. Wang, Z. Gao, L. Wang, Z. Li, and G. Wu, “Boundary-aware cascade networks for temporal action segmentation,” in *European Conference on Computer Vision*. Springer, 2020, pp. 34–51.
- [38] Y. Ishikawa, S. Kasai, Y. Aoki, and H. Kataoka, “Alleviating over-segmentation errors by detecting action boundaries,” in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021, pp. 2322–2331.
- [39] F. Ma, L. Zhu, Y. Yang, S. Zha, G. Kundu, M. Feiszli, and Z. Shou, “Sf-net: Single-frame supervision for temporal action localization,” in *European conference on computer vision*. Springer, 2020, pp. 420–437.
- [40] X. Zhu, Z. Ghahramani, and J. D. Lafferty, “Semi-supervised learning using gaussian fields and harmonic functions,” in *Proceedings of the 20th International conference on Machine learning (ICML-03)*, 2003, pp. 912–919.



- [41] D. Zhou, O. Bousquet, T. Lal, J. Weston, and B. Schölkopf, "Learning with local and global consistency," *Advances in neural information processing systems*, vol. 16, 2003.
- [42] M. Belkin, P. Niyogi, and V. Sindhwani, "Manifold regularization: A geometric framework for learning from labeled and unlabeled examples," *Journal of machine learning research*, vol. 7, no. 11, 2006.
- [43] J. Weston, F. Ratle, H. Mobahi, and R. Collobert, "Deep learning via semi-supervised embedding," in *Neural networks: Tricks of the trade*. Springer, 2012, pp. 639–655.
- [44] D. K. Hammond, P. Vandergheynst, and R. Gribonval, "Wavelets on graphs via spectral graph theory," *Applied and Computational Harmonic Analysis*, vol. 30, no. 2, pp. 129–150, 2011.
- [45] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6299–6308.
- [46] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," 2017.
- [47] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.
- [48] J. Li, P. Lei, and S. Todorovic, "Weakly supervised energy-based learning for action segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 6243–6251.
- [49] Y. Sourì, M. Fayyaz, L. Minciullo, G. Francesca, and J. Gall, "Fast weakly supervised action segmentation using mutual consistency," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.
- [50] S. Stein and S. J. McKenna, "Combining embedded accelerometers with computer vision for recognizing food preparation activities," in *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing*, 2013, pp. 729–738.
- [51] H. Kuehne, A. Arslan, and T. 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 (CVPR)*, June 2014.
- [52] A. Fathi, X. Ren, and J. M. Rehg, "Learning to recognize objects in egocentric activities," in *CVPR 2011*. IEEE, 2011, pp. 3281–3288.
- [53] C.-Y. Lee, S. Xie, P. Gallagher, Z. Zhang, and Z. Tu, "Deeply-supervised nets," in *Artificial intelligence and statistics*. PMLR, 2015, pp. 562–570.
- [54] C. Li, M. Zeeshan Zia, Q.-H. Tran, X. Yu, G. D. Hager, and M. Chandraker, "Deep supervision with shape concepts for occlusion-aware 3d object parsing," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 5465–5474.
- [55] C. Li, M. Z. Zia, Q.-H. Tran, X. Yu, G. D. Hager, and M. Chandraker, "Deep supervision with intermediate concepts," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 8, pp. 1828–1843, 2018.
- [56] M. E. Fathy, Q.-H. Tran, M. Z. Zia, P. Vernaza, and M. Chandraker, "Hierarchical metric learning and matching for 2d and 3d geometric correspondences," in *Proceedings of the european conference on computer vision (ECCV)*, 2018, pp. 803–819.
- [57] B. Zhuang, Q.-H. Tran, G. H. Lee, L. F. Cheong, and M. Chandraker, "Degeneracy in self-calibration revisited and a deep learning solution for uncalibrated slam," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2019, pp. 3766–3773.
- [58] H. Mobahi, R. Collobert, and J. Weston, "Deep learning from temporal coherence in video," in *Proceedings of the 26th Annual International Conference on Machine Learning*, 2009, pp. 737–744.
- [59] S. Hareesh, S. Kumar, H. Coskun, S. N. Syed, A. Konin, Z. Zia, and Q.-H. Tran, "Learning by aligning videos in time," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 5548–5558.