End-to-End Semi-Supervised Learning for Video Action Detection

Akash Kumar Yogesh Singh Rawat
Center for Research in Computer Vision
University of Central Florida
akash_k@knights.ucf.edu yogesh@crcv.ucf.edu

Abstract

*In this work, we focus on semi-supervised learning for video action detection which utilizes both labeled as well as unlabeled data. We propose a simple **end-to-end consistency based** approach which effectively utilizes the unlabeled data. Video action detection requires both, action class prediction as well as a spatio-temporal localization of actions. Therefore, we investigate two types of constraints, **classification consistency**, and **spatio-temporal consistency**. The presence of predominant background and static regions in a video makes it challenging to utilize spatio-temporal consistency for action detection. To address this, we propose two novel regularization constraints for spatio-temporal consistency; 1) **temporal coherency**, and 2) **gradient smoothness**. Both these aspects exploit the **temporal continuity** of action in videos and are found to be effective for utilizing unlabeled videos for action detection. We demonstrate the effectiveness of the proposed approach on two different action detection benchmark datasets, UCF101-24 and JHMDB-21. In addition, we also show the effectiveness of the proposed approach for video object segmentation on the Youtube-VOS dataset which demonstrates its **generalization capability** to other tasks. The proposed approach achieves competitive performance by using merely **20%** of annotations on UCF101-24 when compared with recent fully supervised methods. On UCF101-24, it improves the score by **+8.9%** and **+11%** at 0.5 f-mAP and v-mAP respectively, compared to supervised approach.*

1. Introduction

We have seen a great progress in video action classification [5, 9, 39, 43, 45], where the availability of large-scale datasets is one of the enabling factor [17, 19, 38]. Video action detection on the other hand is much more challenging where spatio-temporal localization is performed on the video. In addition, obtaining large-scale datasets for this problem is even more challenging as annotating each frame is a huge time and cost intensive task.

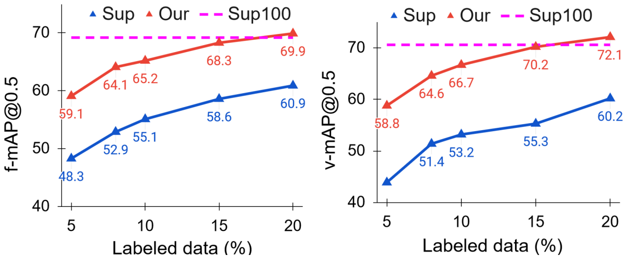

Figure 1. A comparison of proposed semi-supervised method with supervised baseline showing absolute gain in f-mAP and v-mAP for varying number of labelled samples on UCF-101-24 dataset. The proposed method outperforms supervised baseline and using merely 20% of labeled samples, matches the performance of fully supervised method trained on 100% labels. Sup is supervised and Sup100 is supervised with with 100% labels.

In this work, we focus on semi-supervised learning for video action detection which makes use of a small set of annotated samples along with several unlabeled samples. For annotated set, we have video-level class labels as well as frame-level localizations. To the best of our knowledge, this is the *first work* which focuses on semi-supervised learning for video action detection.

Semi-supervised learning has been successfully studied for image classification [3, 36, 44] with some recent works in object detection [10, 15, 16, 41, 46]. Pseudo-labeling [14] and consistency regularization [36, 44, 46] are two main approaches used for semi-supervised learning. Pseudo-labeling utilize unlabeled samples with high confidence score to improve the performance and rely on several iterations of training which is computationally expensive. On the other hand, consistency regularization relies on single-step training where small input perturbations are used for robust learning. Training a video action detection model is already *computationally expensive* due to high-dimensional input, therefore we propose a *consistency-based* approach for an efficient solution.

Video action detection requires a sample level class prediction as well as a spatio-temporal localization on each frame. Therefore, we investigate two different consistency constraints to utilize unlabeled samples; *classification consistency* and *spatio-temporal localization consistency*. Consistency regularization for classification has been found very effective [3,36], however, it often relies on a rich set of augmentations. Extending these augmentations to the video domain for spatio-temporal consistency is not always feasible.

We propose a simple formulation for spatio-temporal consistency where it is computed for each pixel in the video. Extending traditional consistency objective to spatio-temporal domain could capture pixel level variations, but it fails to capture any *temporal constraints* as the consistency is computed independently for each pixel. To address this issue, we explore *temporal continuity* of actions in videos. We argue that motion has some temporal continuity and we attempt to utilize this to regularize the spatio-temporal consistency. We investigate two different ways to capture motion continuity, *temporal coherence* and *gradient smoothness*. Temporal coherence aims at refining the uncertain boundary regions that distinguish foreground and background, and, gradient smoothness enforces temporally consistent localization.

The proposed method is trained end-to-end utilizing both labeled and unlabeled samples without the need for any iterations which makes it efficient. We demonstrate its effectiveness with an extensive set of experiments on two different datasets, UCF101-24 and JHMDB-21. We show that with *limited labels* it can achieve competitive performance when compared with *fully-supervised* methods outperforming all the *weakly-supervised* approaches. In addition, we also demonstrate the *generalization capability* of the proposed method on Youtube-VOS for video object segmentation. We make the following contributions in this work,

- We propose a *simple end-to-end approach* for semi-supervised video action detection. To the best of our knowledge, this is the *first* work focusing on this problem.
- We investigate two different consistency regularization approaches for video action detection; *classification consistency* and *spatio-temporal consistency*.
- We propose two novel regularization constraints for spatio-temporal consistency, *temporal coherency* and *gradient smoothness*, which focus on the *temporal continuity* of actions in videos.

2. Related Work

2.1. Video Action Detection

Video action detection has made significant progress in recent years [22, 27, 32, 35, 50, 52, 53], which is mainly attributed to the development of convolutional neural networks. Earlier attempts started from 2D proposals and then moved towards 3D proposals, the authors in [13] extended the 2D box proposals to 3D cuboids for locating actions in videos. Similarly, [18] utilizes a sequence of frames, and, outputs an anchor cuboid for action localization. In [50], the authors propose to update rough proposals progressively during the training that proves out to be effective. To take advantage of a longer temporal sequence, the authors in [37] utilize a recurrent approach with the help of a Conv LSTM. Some approaches also rely on optical flow, [12, 53], however, it incurs additional computational cost.

Most of these existing methods utilize a proposal-based approach [12, 13, 50, 53] which requires a two-step process and makes these methods complex. In this work, we utilize a simple architecture as an action detection network, which is an end-to-end approach based on capsule routing [7]. Although the authors in [7] propose a simple architecture, the requirement of 3D routing makes it computationally expensive. Therefore, we use a modified model as our baseline action detection network and utilize a 2D routing [30] instead to make it computationally efficient.

Weakly-supervised action detection Video action detection requires annotations on every frame of a video for localization. To alleviate this high annotation cost, recently some weakly-supervised approaches have been proposed [1, 6, 8]. In [6], the authors explore the impact of different levels of supervision for action detection. The authors in [8] utilize siamese similarity over frames and localize actions with the help of actor proposals generated by object detectors. Similarly, the authors in [1] detect actions using an off-the-shelf human detector trained on image datasets [23] with the help of multiple instance learning. Although weakly-supervised approaches reduce the annotation cost on every frame, the performance of these methods is still far from fully supervised approaches. Moreover, it requires class labels for all the samples and also relies on additional bounding boxes localization from state-of-the-art detectors such as Detectron [11] and Faster-RCNN [29].

2.2. Semi-Supervised Learning

Semi-supervised learning utilizes a finite number of labeled samples along with a large number of unlabeled samples. Generally, there are two most prominent approaches for semi-supervised learning, pseudo-labeling [4] and consistency regularization [20, 28, 33, 42]. Pseudo-labeling [14, 21] is an iterative process which makes it iterative approach is computationally expensive and not well suited for video action detection.

Consistency regularization makes use of perturbations on input data and attempts to minimize the difference between predictions from augmented versions same sample

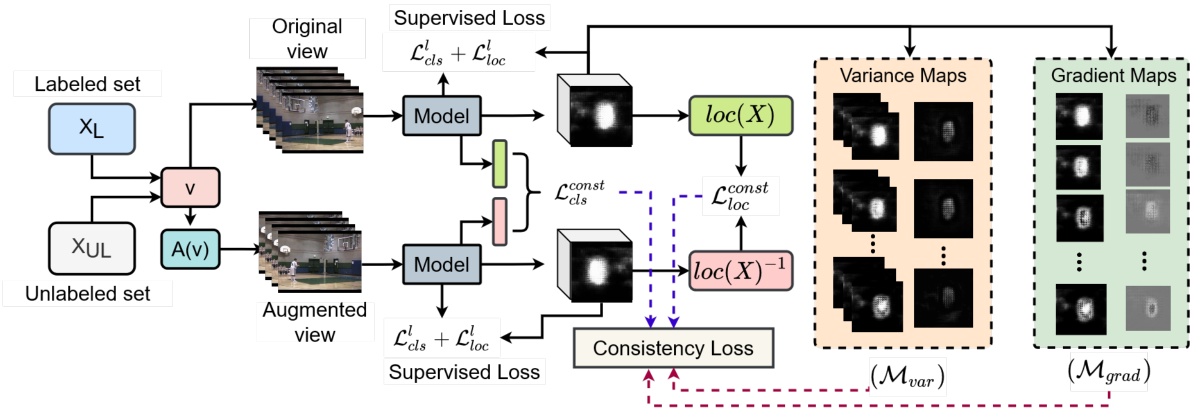

Figure 2. Overview of our proposed approach. Original and augmented view of the input video is passed through the network. The activations at the penultimate layer of classifier head are considered for classification consistency and the spatio-temporal localization is considered for localization consistency. The *attention* masks $\mathcal{M}_{var}$ and $\mathcal{M}_{grad}$ are computed for temporal coherence and gradient smoothness using the spatio-temporal localization. In addition, traditional supervised classification and localization loss is computed for the labelled samples.

[2, 3, 15, 36, 49]. This makes learning simpler in comparison with pseudo-labeling. Recently, it has been explored for the task of semi-supervised image object detection [10, 15, 16, 41, 46]. Most of the work follows setup for the teacher-student network. Inspired by the simplicity of π-model consistency based approaches and its success in classification and object detection, we propose a consistency based approach for video action detection. Also, there is *no existing* work on semi-supervised video action detection to the best of our knowledge.

3. Approach

Given a video $v = (v_1, v_2..., v_n)$ with n frames, we want to perform spatio-temporal localization which provides a class label p for the whole video and localization map l on each frame v_i. Localization map l can be pixel-wise prediction [17] or a bounding-box [38]. In semi-supervised learning, the dataset is consists of a labeled (D_L) and an unlabeled (D_{UL}) set. Let's denote the whole training set with X, labeled subset as $X_L : \{v_l^0, v_l^1, ..., v_l^N\}$ and unlabeled subset as $X_U : \{v_u^0, v_u^1, ..., v_u^N\}$. We want to utilize both these sets to train an action detection model M.

Each training sample v is augmented to get a second view $v^{'}(A(v))$. The action detection model M is used to predict a class label and spatio-temporal localization cls, $loc = M(v)$ for each sample v. A traditional supervised loss is computed for classification ($\mathcal{L}^l_{cls}$) and localization ($\mathcal{L}^l_{loc}$) for a labeled sample. We utilize consistency regularization for both labeled and unlabeled samples. We calculate the difference between a sample (v_u) and its augmented view ($v^{'}_u$) for consistency. We investigate two different consistency loss for action detection, classification ($\mathcal{L}^{const}_{cls}$) and spatio-temporal ($\mathcal{L}^{const}_{loc}$). An overview of the proposed approach is shown in figure Fig. 2. Next, we go through in detail about the action detection model M and these two consistency regularization loss terms.

3.1. Action detection model

We propose a simple action detection model (M) based on VideoCapsuleNet [7]. VideoCapsuleNet is a 3D convolution based encoder-decoder architecture which utilizes 3D capsule routing for detecting and localizing actions in a video. Although it is a simple architecture, the use of 3D capsule routing increases the computation overhead significantly. We propose to use 2D routing [31] instead of 3D routing after pooling the temporal dimension of features and found it to be more efficient without much performance drop. We utilize this adapted model in our experiments. This model M provides a classification prediction p, and, a spatio-temporal localization l for an input video.

3.2. Classification consistency

We want the classification prediction for a sample and its augmented view to be similar. For that, we looked into the output of the latent features of the original view $feat(X)$ and the augmented view $feat(X^{'})$ from the network. The intuition is that the variation in the distribution should be minimal. To enforce this, we employed Jenson-Shannon divergence (JSD) to compute the difference between them. Using JSD, the classification consistency loss ($\mathcal{L}^{const}_{cls}$) is defined as:

$$\mathcal{L}^{const}_{cls} = \mathcal{L}_{JSD} = JSD(feat(X), feat(X^{'})). \qquad (1)$$

3.3. Spatio-temporal consistency

In this consistency constraint, the network learns to detect spatio-temporal localization for multiple views of a video. Using a sample (v), the action detection network (M) outputs a localization map $(l(v))$, which is a pixel-wise prediction, where each pixel has a probability of either action or not action. If we augment the original sample (v), the model should be able to consistently predict the action region $(l(v^{'}))$. Using spatio-temporal consistency, we propose to bring these predictions close to each other. For this, we need to evaluate a pixelwise difference between the two predicted localization maps of augmented view $(loc(X^{'}))$ and the original view $(loc(X))$.

To compare the predictions, we need to inverse the data augmentation for the augmented view $(loc(X^{'}))$ so that mapping between the pixel locations are same while calculating the difference. To minimize this difference in predictions, we use L2 loss. The spatio-temporal consistency loss $(\mathcal{L}_{loc}^{const})$ is defined as,

$$\mathcal{L}_{loc}^{const} = \mathcal{L}_{L2} = L2(loc(X), (loc(X^{'})^{-1})), \tag{2}$$

where $loc(X^{'})^{-1}$ indicates reversal of augmentations.

The spatio-temporal consistency defined above (Eq. (2)) only captures the spatial variance for different predicted localization maps, and, doesn't enforce any temporal constraints. Thus, it effectively works similar to any consistency-based object-detection for images. However, we have a third dimension in videos, the temporal dimension, and moving along this dimension, we can enforce *continuity* and *smoothness* constraints. It means that the predictions should not only be continuous, but the transition across each frame should also be smooth as well.

Therefore, we explore *temporal continuity* of actions in a video to effectively utilize spatio-temporal consistency. We focus on two different aspects of temporal continuity, *temporal coherency* and *gradient smoothness*. Temporal coherency captures the relative change in the boundary region of actions across time and helps in refining the detection boundaries. On the other hand, gradient smoothness helps in the detection of abrupt changes in predictions across time.

Temporal coherence Temporal coherence is described as the relative displacement of the foreground pixels (action region) in the temporal dimension over a finite amount of frames (f_n). We compute the variance of the pixels in the current frame by measuring the relative shift in its position in future and past frames. This pixel-wise variance is computed for all the pixels in a video and is termed as variance map $\mathcal{M}_{var}$. The variance map $\mathcal{M}_{var}$ of a video attend to *short-term fine-grained changes* concentrating on the continuity of predictions. Analyzing variance of a particular frame, it will have two distinct regions (Fig. 2), *unambiguous*, and *ambiguous*. If a model is confident that a pixel is an action or non-action, we call it *unambiguous* otherwise we describe it as *ambiguous*. Since the model is already confident on unambiguous regions, we look into the latter. Some of these ambiguous regions will depict the boundaries connecting the foreground and background. Using the variance map we aim to give more *attention* to these regions. This will help the model exploit the ambiguity in spatio-temporal dimensions.

We utilize the variance map as attention to regularize the spatio-temporal consistency loss. This regularized loss $\mathcal{L}_{var}^{const}$ is defined as

$$\mathcal{L}_{var}^{const} = w.(\mathcal{M}_{var} \odot \mathcal{L}_{L2}) + (1 - w).(\mathcal{L}_{L2}), \tag{3}$$

where, mask $\mathcal{M}_{var}$ is calculated as:

$$\mathcal{M}_{var} = \frac{\sum_{i=1}^{n}(loc_i - \mu_n)^2}{n}. \tag{4}$$

Here, loc_i represents the localization on frame i for which variance is computed, and n represents the total number of frames. μ_n represents the average of n frames. w indicates the weight factor for temporal coherency and non-attentive L2 loss. However, at the beginning of training, the model will only have primitive knowledge of spatial localization of actions. Therefore, in the initial phase of training, we start with $w = 0$ where every pixel in the video has equal importance. As the training progresses, the model can recognize the coarse localization of actions, but, is still unsure of boundary regions. Therefore, we exponentially ramp-up the weight (w) of temporal coherence attention mask (M_{var}) used for L2 loss throughout the training, subsequently, reducing the effect of non-attentive L2 loss. Finally, to exploit longer temporal information, we make use of augmented view. We reversed the spatial augmentation and flip it temporally, and attach it to the original view except for the last and first frame and calculate the variance for this longer clip. Since this new clip can be used to make a repetitive cycle, we have coined the term *cyclic variance* in our paper.

Gradient Smoothness Taking a deeper look into the temporal aspects of localization, the transition of actor localization should be smooth. To maintain this smoothness constraint, we analyze the change in output localization probability score maps using second-order gradients. Gradient reflects the change in direction. The first-order gradient of a spatio-temporal region along the temporal dimension provides a temporal gradient flow map. Since the offset is small in the temporal dimension, the first-order gradient map should be smooth. Taking the second-order gradient signifies the change in the first-order gradient. As the offset is small, the second-order gradient should be zero. The spikes in the second-order gradient map determine the change in the continuity of the temporal gradient flow map. We utilize this map $\mathcal{M}_{grad}$ as an *attention* to enforce the *long-term smoothness* of spatio-temporal localization. We calculate the gradient smoothness consistent loss as

$$\mathcal{L}_{grad}^{const} = (\mathcal{M}_{grad} \odot \mathcal{L}_{L2}), \tag{5}$$

		UCF101-24				JHMDB-21			
Consistency		f-mAP		v-mAP		f-mAP		v-mAP	
CC	LC	0.2	0.5	0.2	0.5	0.2	0.5	0.2	0.5
		85.1 ± 0.95	61.0 ± 0.75	91.3 ± 0.25	61.1 ± 1.25	87.9 ± 0.65	61.1 ± 1.40	92.5 ±0.85	59.5 ±0.20
✓		87.8 ± 0.85 (↑ 2.7)	65.0 ± 0.90 (↑ 4.0)	93.7 ± 0.60 (↑ 2.4)	66.2 ± 1.10 (↑ 5.1)	89.1 ± 1.32 (↑ 1.2)	62.9 ± 2.14 (↑ 1.8)	94.1 ± 0.60 (↑ 1.6)	61.2 ±2.43 (↑ 1.7)
	✓	89.6 ± 0.30 (↑ 4.5)	69.8 ± 0.05 (↑ 8.8)	95.2 ± 0.15 (↑ 3.9)	71.8 ± 0.05 (↑ 10.7)	89.0 ± 1.70 (↑ 1.1)	63.4 ± 1.90 (↑ 2.3)	94.8 ± 0.60 (↑ 2.3)	61.6 ± 1.70 (↑ 2.1)
✓	✓	89.1 ± 0.85 (↑ 4.0)	69.5 ± 0.65 (↑ 8.5)	95.1 ± 0.30 (↑ 3.8)	71.8 ± 0.50 (↑ 10.7)	89.2 ± 2.35 (↑ 1.3)	63.6 ± 2.45 (↑ 2.5)	94.4 ± 0.67 (↑ 1.9)	62.8 ±1.95 (↑ 3.3)

Table 1. Performance on UCF101-24 and JHMDB-21 datasets with inclusion of individual and combined consistency losses. The first row indicates supervised training results. Here CC and LC denotes classification and localization consistency.

where mask $\mathcal{M}_{grad}$ is calculated as

$$\mathcal{M}_{grad} = \frac{\partial^2(loc)}{\partial t^2} \text{where} \frac{\partial(loc)}{\partial t} = \frac{loc_{t+1} - loc_{t-1}}{2}. \tag{6}$$

Here, the first order partial derivative $\frac{\partial(loc)}{\partial z}$ is approximated using a central difference derivative mask.

3.4. Overall training objective

To formalize the final training objective, we have supervised losses and consistency losses. We calculate the supervised loss for classification ($\mathcal{L}^l_{cls}$) and localization ($\mathcal{L}^l_{loc}$). For consistency, we have classification ($\mathcal{L}^{const}_{cls}$), spatio-temporal ($\mathcal{L}^{const}_{loc}$), temporal coherency ($\mathcal{L}^{const}_{var}$) and gradient smoothness loss ($\mathcal{L}^{const}_{grad}$). The overall supervised loss is computed as

$$\mathcal{L}_{labeled} = \mathcal{L}^l_{cls} + \mathcal{L}^l_{loc}, \tag{7}$$

and the combined consistency loss is computed as

$$\mathcal{L}_{const} = \lambda_1 \mathcal{L}^{const}_{cls} + \lambda_2 (\mathcal{L}^{const}_{var}, \mathcal{L}^{const}_{grad}), \tag{8}$$

where λ_1 and λ_2 are weight parameters for classification and spatio-temporal consistency respectively. Finally, the overall training objective is a combination of these two,

$$\mathcal{L}_{total} = \mathcal{L}_{labeled} + \lambda \mathcal{L}_{const}. \tag{9}$$

Here (λ) is a weight parameter used for consistency loss.

4. Experiments

Datasets For our action detection experiments, we use UCF101-24 [38] and JHMDB-21 [17] datasets. **UCF101-24** contains 3207 untrimmed videos. The number of training and testing videos is 2284 and 923 respectively. It contains 24 action classes. These classes mainly belong to sports and are sub-sampled from the original UCF101 dataset which contains 101 action classes. The original resolution of clips is 320x240. The action duration covers almost 78% of the total duration of the video. **JHMDB-21** contains 928 videos categorized into 21 action classes. These classes are similar to sports scenes as UCF. It's a trimmed dataset where the action is happening during the whole video duration. The frame resolution is the same as UCF101-24. To show that our approach can be generalized across other domains, we perform experiments on **YouTube-VOS** dataset as well. This dataset has 3471 training videos and 589 videos for evaluation.

Implementation Details In our experiments, the height, and width of the input frame is 224x224. Our batch size is eight. In each batch, the ratio of labeled to unlabeled samples is 1:1. So, out of eight clips in a batch, four are samples from the labeled subset and the remaining four from the unlabeled subset. Then, they are randomly shuffled. The number of frames per clip is eight. We choose the frames with a skip rate of 2. The distribution of labeled and unlabeled samples for our experiments is 20/80 for UCF101-24 and 30/70 for JHMDB-21 datasets. We use I3D [5] as a backbone with pretrained weights from Kinetics [19] and Charades [34].

Training details We use Adam optimizer with an initial learning rate of 1e-4, and a scheduler decay rate of 0.1, if training loss doesn't improve in the last five epochs. We train the model for 100 epochs on UCF101-24 and 50 epochs for JHMDB-21. The lambda value for consistency loss is set to 0.1. The parameters λ_1 and λ_2 in Eq. (8) are set to 0.3 and 0.7. We train the network for 100 epochs on UCF101-24 and for 50 epochs on JHMDB-21. To calculate the temporal coherency of each frame in a clip, 2 future and 2 past frames are picked where we use the localization of the augmented view. This serves as the attention mask to calculate L2 loss. For gradient smoothness, we calculate the mask values for L2 loss using second order gradient of spatio-temporal prediction on a single clip in the temporal dimension. For labeled samples, we used the margin loss [7] as classification loss ($\mathcal{L}^l_{cls}$) and binary cross-entropy plus dice loss to measure localization loss ($\mathcal{L}^l_{loc}$).

Evaluation metrics We compute frame-metric average precision (f-mAP) and video-metric average precision (v-mAP) scores to evaluate the action detection performance. f-mAP calculates the score based on how many frames overlap with the ground truth frames given the IoU, and, v-mAP

Experiment				UCF101-24		JHMDB-21	
V	G	VC	L2	f-mAP@0.5	v-mAP@0.5	f-mAP@0.5	v-mAP@0.5
✓				68.3	70.3	61.9	61.4
		✓		68.4 (↑ 0.1)	70.8 (↑ 0.5)	63.0 (↑ 1.1)	61.5 (↑ 0.1)
✓			✓	68.8 (↑ 0.5)	71.6 (↑ 1.3)	63.3 (↑ 1.4)	62.4 (↑ 1.0)
		✓	✓	**69.9** (↑ 1.6)	**72.1** (↑ 1.8)	**64.4** (↑ 2.5)	**63.5** (↑ 2.1)
	✓			69.6	72.4	63.2	63.1
	✓		✓	69.4 (↓ 0.2)	72.0 (↓ 0.4)	63.1 (↓ 0.1)	62.2 (↓ 0.9)

Table 2. An analysis of the effect of temporal constraints on consistency regularization using UCF101 - 20% and JHMDB-21 - 30% labeled subset. V, G, VC and L2 stands for Variance, Gradient, Cyclic variance

scores based on video overlapping. We have shown the results for frame-mAP and video-mAP at 0.2 and 0.5.

Baselines To compare our work with existing approaches we extend a few semi-supervised image classification approaches to videos. Especially, we looked into pseudo-label [14], MixMatch [3] and Consistency-based Object Detection (Co-SSD (CC)) [15]. Pseudo-label requires multiple iterations of training, whereas, MixMatch is dependent on stochastic data augmentations. The ratio of labeled to the unlabeled subset is 20 to 80, consistent for all our experiments. The augmentation strategy is the same as mentioned in the paper [3]. We generate two views, a weak and a strong one, and follow the same procedure as the original paper. Training details for MixMatch and pseudo-label are mentioned in the supplementary.

4.1. Results

First, we analyse the classification and spatio-temporal consistency losses. The results are shown in Table 1.

UCF101-24 From Table 1, we can see, when we apply classification consistency on action features, we get a major improvement in both f-mAP and v-mAP, a boost of 4-5.1% at 0.5 over the supervised approach. Next, we investigated consistency based on spatio-temporal localization. In general, spatio-temporal consistency outperforms the classification consistency. Especially at v-mAP@0.5, there's a 10.7% jump in performance over the supervised baseline. This proves that spatio-temporal consistency enforces the network to learn better features. Finally, the combination of both, outperforms the classification by a margin of 4.5-5.6% at 0.5 metrics, however, relative to spatio-temporal the performance is almost similar. This shows that later has a greater influence than classification consistency.

JHMDB-21 JHMDB-21 is a relatively smaller dataset which leads to a small number of videos per class and the problem of over-fitting. Thus, we examine the performance for different subsets, and, finally, for our work, we used 30% of the dataset as a labeled subset, which amounts to 189 labeled samples and 471 unlabeled samples. Relative to supervised training on 30%, we get a boost of approximately 1-2% for both classification and localization consistency (Table 1). Training with the combination of both consistencies did provide us a gain of roughly 1% over single consistency on v-mAP@0.5.

We observe in Table 1 that combining classification and spatio-temporal consistency doesn't have a significant impact. Since the classification consistency performance is lower as compared to the spatio-temporal, we only rely on spatio-temporal consistency for further experiments.

Next, we analyze the impact of temporal constraints on consistency regularization. From Table 1 and 2, we evaluate the performance gain for non-attentive L2 versus *temporal coherency* plus non-attentive L2. For UCF101-24, f-mAP and v-mAP at 0.5, the later outperforms non-attentive L2 by a margin of 0.1 and 0.3% respectively. For JHMDB-21, we see a better margin of improvement, with a boost of 1% for f-mAP and 2% for v-mAP. Coming to *gradient smoothness*, v-mAP at 0.5 for UCF101-24 and JHMDB-21 beat the non-attentive L2 by 0.6% and 1.5% respectively. This corroborates our claim that *temporal coherency* and *gradient smoothness* indeed proves out to be effective to enforce temporal continuity constraints.

4.2. Comparison

We first compare the proposed approach with the semi-supervised baselines followed by existing works on weakly-supervised learning and supervised learning. This is the first work on semi-supervised video action detection to the best of our knowledge, therefore for a fair comparison, we introduce several standardized semi-supervised baselines. This includes two major sub-areas: consistency (MixMatch, Co-SSD(CC)), and, pseudo-label.

Semi-supervised For comparison with semi-supervised approaches, we extended MixMatch, pseudo-label, and Co-SSD to video action detection. The performance of MixMatch is lowest amongst all (Table 3). Compared to pseudo-label at v-mAP@0.5, our approach outperformed by a margin of 5-6% on UCF101-24 and 7-8% on JHMDB-21. Co-SSD outperformed the pseudo-label approach, however, we beat that approach with a margin of 4-5% for both the datasets. We show this comparison in Fig. 3 for different percentages of labeled samples.

Weakly-supervised: These methods [1, 6, 8] use 100% of the class labels while training, as well as a state-of-the-art actor detector, to get bounding boxes across the whole video. On the other hand, we did not use any bounding box or class label information for 80% of the data. We beat the best-reported scores on UCF101-24 by a margin of more

Method	Backbone		UCF101-24			JHMDB-21		
			f-mAP	v-mAP		f-mAP	v-mAP	
	2-D	3-D	0.5	0.2	0.5	0.5	0.2	0.5
Fully-Supervised								
Singh *et al.* [35] †	✓		-	73.5	46.3	-	73.8	72.0
Kalogeitan *et al.* [18]	✓		69.5	76.5	49.2	65.7	74.2	73.7
Yang *et al.* [50]†	✓		75.0	76.6	-	-	-	-
Song *et al.* [37]†	✓		72.1	77.5	52.9	65.5	74.1	73.4
Zhao and Snoek [53]†	✓		-	78.5	50.3	-	-	74.7
Li *et al.* [22]	✓		78	82.8	53.8	70.8	77.3	70.2
Hou *et al.* [13]		✓	41.4	47.1	-	61.3	78.4	76.9
Gu *et al.* [12]†		✓	76.3	-	59.9	73.3	-	78.6
Sun *et al.* [40]		✓	-	-	-	77.9	-	80.1
Pan *et al.* [26]		✓	84.3	-	-	-	-	-
Duarte *et al.* [7]		✓	78.6	97.1	80.3	64.6	95.1	-
Ours		✓	69.2	95.3	71.9	68.1	96.8	68.4
Weakly-Supervised								
Mettes *et al.* [25]	✓		-	37.4	-	-	-	-
Mettes and Snoek [24]	✓		-	41.8	-	-	-	-
Cheron *et al.* [6]		✓	-	43.9	17.7	-	-	-
Escorcia *et al.* [8]		✓	45.8	19.3	-	-	-	-
Arnab *et al.* [1]		✓	-	61.7	35.0	-	-	-
Zhang *et al.* [51]		✓	30.4	45.5	17.3	65.9	77.3	50.8
Semi-Supervised								
MixMatch [3]		✓	20.2	60.2	13.8	7.5	46.2	5.8
Psuedo-label [14]		✓	64.9	93.0	65.6	57.4	90.1	57.4
Co-SSD(CC) [15]		✓	65.3	93.7	67.5	60.7	94.3	58.5
Ours		✓	**69.9**	**95.7**	**72.1**	**64.4**	**95.4**	**63.5**

Table 3. Comparison with existing supervised and weakly supervised works along with the semi-supervised baselines on UCF101- 24 and JHMDB-21 . † denotes approach uses Optical flow.

than 35% approximately. A comparison is shown in Table 3.

Supervised: Table 3 shows a comparison with several existing supervised action detection approaches. We observe that with only 20% labeled data for UCF101-24, our scores at v-mAP@0.2 and 0.5 outperform all of the approaches. f-mAP@0.5 is better than most of the approaches apart from [50], [37], and, [12]. However, all of them have used optical flow as a second modality. Optical flow works as an extra supervisory signal. For JHMDB-21, we were able to beat some of the approaches at f-mAP@0.5 and v-mAP@0.2.

5. Ablation Study

To get a deeper insights on how *attention* constraint helped in improving the accuracy, we did a study on individual components of *temporal coherence* and *gradient smoothness* modules. Since JHMDB-21 is a small dataset, we analyze the results for three different seed variations for each of the components on JHMDB-21. The ablation scores are shown in Table 2. For JHMDB, the scores are mean of three runs. The standard deviation is in supplementary.

Temporal coherence: From Table 2, when we apply only attention mask, we outperform the supervised baseline by a good margin. Using cyclic variance improved the score by 0.1-0.5% for UCF101-24 and 0.1-1.1% for JHMDB-21 dataset. Incorporating non-attentive L2 loss, we see a good increment of roughly 1% for both datasets. Finally, moving onto the cyclic variance alongwith non-attentive L2 we got an additional boost of 1%. This demonstrates that not only temporal variance is helping, but longer temporal information (cyclic variance) also compliments the base score as well. We also notice that the margin of improvement is higher when we are utilizing a pixel-wise prediction mask.

Gradient smoothness: Following the paths of temporal coherence, we first employed gradient smoothness with non-attentive L2. However, this did not improve the score further as compared to standalone gradient smoothness loss. Using only that, we see it outperforms non-attentive L2 on

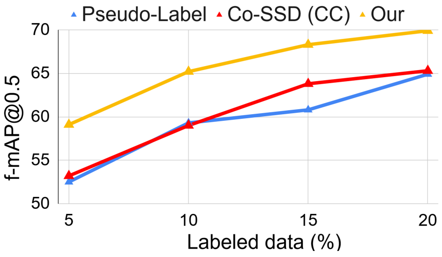

Figure 3. Comparison of pseudo-label [14], Co-SSD (CC) [15], and ours approach for 5, 10, and, 15 and 20% of labeled data.

Dataset	f-mAP (%)		v-mAP (%)	
	0.2	0.5	0.2	0.5
Trimmed	90.2	69.9	96.2	72.1
Untrimmed	90.1	69.5	96.0	71.3

Table 4. Performance comparison when we use untrimmed videos instead of trimmed videos where the action is occurring in all the video frames. We observe small to negligible performance drop which indicates that our method can also utilize untrimmed videos.

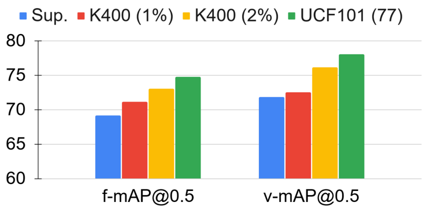

Figure 4. We observe performance gain when additional unlabeled videos were used from external sources, such as Kinetic-400 (K400) and UCF-101 (other 77 action classes from UCF-101).

v-mAP@0.5 for both datasets. This was expected as *gradient smoothness* focuses on the whole clip compared to *temporal coherency*.

6. Discussions

In this section, we discuss some of the queries pertaining to semi-supervised activity detection in general.

Does the amount of unlabeled samples matters? For this study, we keep the amount of labeled samples constant to 20%. Then, we increase the amount of unlabeled samples from 20% (1x) to 80% (4x). (Fig. 5) We see a constant gain in performance for all the metrics. This shows that more the number of unlabeled samples, better will be the performance.

What is the impact of using untrimmed dataset instead of trimmed? In trimmed videos, all the frames in a video sample contains action, however, in untrimmed, there may be some frames without any actions. In our experiments we assume the availability of trimmed videos for UCF101-24. We experimented with a scenario where the unlabeled videos can be untrimmed. The dataset has some videos with no activity at all. The evaluation is shown in Table 4. We only observe a marginal to negligible drop in performance which shows the robustness of the proposed spatio-temporal consistency for untrimmed videos. This is also evident from our next set of experiments where we utilize additional untrimmed unlabeled videos from external datasets.

Method	Semi-Sup.	Avg	J_S	J_U	F_S	F_U
LSTM [47]		10.1	11.6	10.1	9.6	9.2
	✓	36.8	43.1	31.4	40.8	31.8
Sup. (100%)		47.9	55.7	39.6	55.2	41.3

Table 5. Evaluation on Youtube-VOS dataset. We have used 10% data for supervised approach. The bottom row shows the results for supervised training on 100% data.

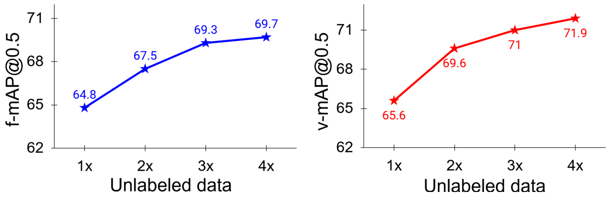

Figure 5. Performance with varying the amount of unlabeled data.

Can additional data support as a supervisory signal? Lastly, we explored how can we utilize rest of the actions (not labelled) from UCF101 and outperform the supervised accuracy. In Fig. 4 we observe that additional video samples even from Kinetics dataset helps in improving the performance. However, the gain is more significant when the videos are from a similar distribution, UCF-101 in this scenario.

6.1. Generalization to video object segmentation

We have also shown that our approach is generalizable across different task. For VOS dataset [48], we see an overall improvement of 30% over the supervised baseline (Table 5). We have shown the score using *temporal coherency* consistency loss. Training details are in supplementary.

7. Conclusion

In this work, we propose a novel end-to-end approach for *semi-supervised video action detection*. To best of our knowledge, this is the *first attempt* in semi-supervised learning for action detection. We propose the use of *consistency regularization* for an efficient and effective detection performance. We demonstrate the positive impact of *temporal coherency* and *gradient smoothness* constraint for spatio-temporal localization. The proposed approach achieves significant performance boost over *supervised baselines* with limited labels and *outperforms* weakly supervised methods.

References

[1] A. Arnab, Chen Sun, Arsha Nagrani, and C. Schmid. Uncertainty-aware weakly supervised action detection from untrimmed videos. *ArXiv*, abs/2007.10703, 2020. 2, 6, 7

[2] David Berthelot, Nicholas Carlini, E. D. Cubuk, Alexey Kurakin, Kihyuk Sohn, Han Zhang, and Colin Raffel. Remixmatch: Semi-supervised learning with distribution alignment and augmentation anchoring. *ArXiv*, abs/1911.09785, 2019. 3

[3] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. 1, 2, 3, 6, 7

[4] Avrim Blum and Tom Mitchell. Combining labeled and unlabeled data with co-training. In *Proceedings of the Eleventh Annual Conference on Computational Learning Theory*, COLT' 98, page 92–100, New York, NY, USA, 1998. Association for Computing Machinery. 2

[5] 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 (CVPR)*, July 2017. 1, 5

[6] Guilhem Chéron, Jean-Baptiste Alayrac, Ivan Laptev, and Cordelia Schmid. A flexible model for training action localization with varying levels of supervision. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. 2, 6, 7

[7] Kevin Duarte, Yogesh S Rawat, and Mubarak Shah. Videocapsulenet: A simplified network for action detection. *Advances in Neural Information Processing Systems*, 2018. 2, 3, 5, 7

[8] Victor Escorcia, C. D. Dao, Mihir Jain, Bernard Ghanem, and Cees G. M. Snoek. Guess where? actor-supervision for spatiotemporal action localization. *Comput. Vis. Image Underst.*, 192:102886, 2020. 2, 6, 7

[9] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6201–6210, 2019. 1

[10] Qiang feng Zhou, Chaohui Yu, Zhibin Wang, Qi Qian, and Hao Li. Instant-teaching: An end-to-end semi-supervised object detection framework. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4079–4088, 2021. 1, 3

[11] Ross Girshick, Ilija Radosavovic, Georgia Gkioxari, Piotr Dollár, and Kaiming He. Detectron. `https://github.com/facebookresearch/detectron`, 2018. 2

[12] Chunhui Gu, Chen Sun, David A. Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, Cordelia Schmid, and Jitendra Malik. Ava: A video dataset of spatio-temporally localized atomic visual actions. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6047–6056, 2018. 2, 7

[13] Rui Hou, Chen Chen, and Mubarak Shah. Tube convolutional neural network (t-cnn) for action detection in videos. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 5823–5832, 2017. 2, 7

[14] Dong hyun Lee. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. 1, 2, 6, 7, 8

[15] Jisoo Jeong, Seungeui Lee, Jeesoo Kim, and Nojun Kwak. Consistency-based semi-supervised learning for object detection. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. 1, 3, 6, 7, 8

[16] Jisoo Jeong, Vikas Verma, Minsung Hyun, Juho Kannala, and Nojun Kwak. Interpolation-based semi-supervised learning for object detection. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11597–11606, 2021. 1, 3

[17] H. Jhuang, J. Gall, S. Zuffi, C. Schmid, and M. J. Black. Towards understanding action recognition. In *International Conf. on Computer Vision (ICCV)*, pages 3192–3199, Dec. 2013. 1, 3, 5

[18] Vicky Kalogeiton, Philippe Weinzaepfel, Vittorio Ferrari, and Cordelia Schmid. Action tubelet detector for spatio-temporal action localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 4415–4423, 2017. 2, 7

[19] Will Kay, João Carreira, K. Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, T. Back, A. Natsev, Mustafa Suleyman, and Andrew Zisserman. The kinetics human action video dataset. *ArXiv*, abs/1705.06950, 2017. 1, 5

[20] Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. *ArXiv*, abs/1610.02242, 2017. 2

[21] Hengduo Li, Zuxuan Wu, Abhinav Shrivastava, and Larry Davis. Rethinking pseudo labels for semi-supervised object detection. *ArXiv*, abs/2106.00168, 2021. 2

[22] Yixuan Li, Zixu Wang, Limin Wang, and Gangshan Wu. Actions as moving points. In *arXiv preprint arXiv:2001.04608*, 2020. 2, 7

[23] Tsung-Yi Lin, M. Maire, Serge J. Belongie, James Hays, P. Perona, D. Ramanan, Piotr Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 2

[24] Pascal Mettes and Cees G. M. Snoek. Pointly-supervised action localization. *International Journal of Computer Vision*, 127:263–281, 2018. 7

[25] Pascal Mettes, Cees G. M. Snoek, and Shih-Fu Chang. Localizing actions from video labels and pseudo-annotations. *ArXiv*, abs/1707.09143, 2017. 7

[26] Junting Pan, Siyu Chen, Mike Zheng Shou, Yu Liu, Jing Shao, and Hongsheng Li. Actor-context-actor relation network for spatio-temporal action localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 464–474, 2021. 7

[27] Xiaojiang Peng and Cordelia Schmid. Multi-region two-stream r-cnn for action detection. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, pages 744–759, Cham, 2016. Springer International Publishing. 2

[28] Antti Rasmus, Harri Valpola, Mikko Honkala, Mathias Berglund, and Tapani Raiko. Semi-supervised learning with ladder network. *ArXiv*, abs/1507.02672, 2015. 2

[29] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28:91–99, 2015. 2

[30] Sara Sabour, Nicholas Frosst, and Geoffrey E Hinton. Dynamic routing between capsules. *arXiv preprint arXiv:1710.09829*, 2017. 2

[31] Sara Sabour, Nicholas Frosst, and Geoffrey E. Hinton. Dynamic routing between capsules. *ArXiv*, abs/1710.09829, 2017. 3

[32] Suman Saha, Gurkirt Singh, Michael Sapienza, Philip H.S. Torr, and Fabio Cuzzolin. Deep learning for detecting multiple space-time action tubes in videos. *British Machine Vision Conference (BMVC) 2016*, 2016. 2

[33] Mehdi S. M. Sajjadi, Mehran Javanmardi, and Tolga Tasdizen. Regularization with stochastic transformations and perturbations for deep semi-supervised learning. In *NIPS*, 2016. 2

[34] Gunnar A. Sigurdsson, Gül Varol, X. Wang, Ali Farhadi, I. Laptev, and A. Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. *ArXiv*, abs/1604.01753, 2016. 5

[35] Gurkirt Singh, Suman Saha, Michael Sapienza, Philip H. S. Torr, and Fabio Cuzzolin. Online real-time multiple spatiotemporal action localisation and prediction. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. 2, 7

[36] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 596–608. Curran Associates, Inc., 2020. 1, 2, 3

[37] Lin Song, Shiwei Zhang, Gang Yu, and Hongbin Sun. Tacnet: Transition-aware context network for spatio-temporal action detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 2, 7

[38] K. Soomro, A. Zamir, and M. Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *ArXiv*, abs/1212.0402, 2012. 1, 3, 5

[39] Jonathan C. Stroud, David A. Ross, Chen Sun, Jia Deng, and Rahul Sukthankar. D3d: Distilled 3d networks for video action recognition. In *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 614–623, 2020. 1

[40] C. Sun, Abhinav Shrivastava, Carl Vondrick, K. Murphy, R. Sukthankar, and C. Schmid. Actor-centric relation network. *ArXiv*, abs/1807.10982, 2018. 7

[41] Yihe Tang, Weifeng Chen, Yijun Luo, and Yuting Zhang. Humble teachers teach better students for semi-supervised object detection. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3131–3140, 2021. 1, 3

[42] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *NIPS*, 2017. 2

[43] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6450–6459, 2018. 1

[44] Qizhe Xie, Zihang Dai, E. Hovy, Minh-Thang Luong, and Quoc V. Le. Unsupervised data augmentation for consistency training. *arXiv: Learning*, 2020. 1

[45] Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and K. Murphy. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In *ECCV*, 2018. 1

[46] Mengde Xu, Zheng Zhang, Han Hu, Jianfeng Wang, Lijuan Wang, Fangyun Wei, Xiang Bai, and Zicheng Liu. End-to-end semi-supervised object detection with soft teacher. *ArXiv*, abs/2106.09018, 2021. 1, 3

[47] N. Xu, L. Yang, Yuchen Fan, Jianchao Yang, Dingcheng Yue, Yuchen Liang, Brian L. Price, Scott D. Cohen, and Thomas S. Huang. Youtube-vos: Sequence-to-sequence video object segmentation. *ArXiv*, abs/1809.00461, 2018. 8

[48] N. Xu, L. Yang, Yuchen Fan, Dingcheng Yue, Yuchen Liang, Jianchao Yang, and Thomas S. Huang. Youtube-vos: A large-scale video object segmentation benchmark. *ArXiv*, abs/1809.03327, 2018. 8

[49] Xiangli Yang, Zixing Song, Irwin King, and Zenglin Xu. A survey on deep semi-supervised learning. *ArXiv*, abs/2103.00550, 2021. 3

[50] Xitong Yang, Xiaodong Yang, Ming-Yu Liu, Fanyi Xiao, Larry S. Davis, and Jan Kautz. Step: Spatio-temporal progressive learning for video action detection. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 264–272, 2019. 2, 7

[51] Shiwei Zhang, Lin Song, Changxin Gao, and Nong Sang. Glnet: Global local network for weakly supervised action localization. *IEEE Transactions on Multimedia*, 22(10):2610–2622, 2020. 7

[52] Jiaojiao Zhao, Xinyu Li, Chunhui Liu, Bing Shuai, Hao Chen, Cees Snoek, and Joseph Tighe. Tuber: Tube-transformer for action detection. *ArXiv*, abs/2104.00969, 2021. 2

[53] Jiaojiao Zhao and Cees G. M. Snoek. Dance with flow: Two-in-one stream action detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 2, 7