LEARNING FROM STUDENT'S MISTAKES: IMPROVING MEAN TEACHER FOR END-TO-END SEMI-SUPERVISED VIDEO ACTION DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

In this work, we focus on semi-supervised learning for video action detection. We present ***Enhanced Mean Teacher***, a simple end-to-end student-teacher based framework which relies on pseudo-labels to learn from unlabeled samples. A limited amount of data make the teacher prone to unreliable boundaries while detecting the spatio-temporal actions. We propose a novel ***auxiliary module***, which learns from students' mistakes on labeled samples and improves the spatio-temporal pseudo-labels generated by the teacher on unlabeled set. The proposed framework utilize ***spatial*** and ***temporal augmentations*** to generate pseudo-labels where both ***classification*** as well as ***spatio-temporal consistencies*** are used to train the model. We evaluate our approach on two action detection benchmark datasets, UCF101-24, and JHMDB-21. On UCF101-24, our approach outperforms the supervised baseline by an approximate margin of **19%** on *f-mAP@0.5* and **25%** on *v-mAP@0.5*. Using merely 10-15% of the annotations in UCF-101-24, the proposed approach provides a competitive performance compared to the supervised baseline trained on 100% annotations. We also evaluate the effectiveness of Enhanced Mean Teacher for video object segmentation demonstrating its generalization capability to other tasks in the video domain.

1 INTRODUCTION

Video action detection is a challenging problem with several real-world applications in security, assistive-living, robotics, and autonomous-driving. What makes the task of video action detection challenging is the requirement of spatio-temporal localization in addition to video-level activity classification. This also requires annotations on each video frame, which can be cost and labor intensive. In this work, we focus on semi-supervised learning (SSL) to develop label efficient method for video action detection.

SSL is an active area of research and existing approaches can be broadly categorized into iterative proxy-label Rizve et al. (2020) and consistency based Tarvainen & Valpola (2017) approaches. Proxy-label approaches mostly require multiple iterations over the dataset, which is unsuitable for the video domain due to long training cycles. On the other hand, most of the consistency-based approaches are end-to-end and require a single instance of run through the dataset for training. This results in an efficient train time which motivated us to develop a consistency based approach.

Most of the existing work in this direction is focused on image classification (Rasmus et al., 2015; Tarvainen & Valpola, 2017; Sajjadi et al., 2016; Laine & Aila, 2017b) with some recent works focusing on object detection (Xu et al., 2021; Jeong et al., 2021; Tang et al., 2021; feng Zhou et al., 2021; Chen et al., 2022; Liu et al., 2022; 2021) as well. In the video domain most of the efforts are focused on video classification task (Jing et al., 2021; Singh et al., 2021; Xiao et al., 2022; Xu et al., 2022). Amongst consistency-based approaches, we observe that Mean teacher (Tarvainen & Valpola, 2017) based approaches dominate and have successfully outperformed other consistency based methods. Thus, motivated by the success of student teacher learning in image classification and object detection, we adapt this approach for spatio-temporal detection in the video domain.

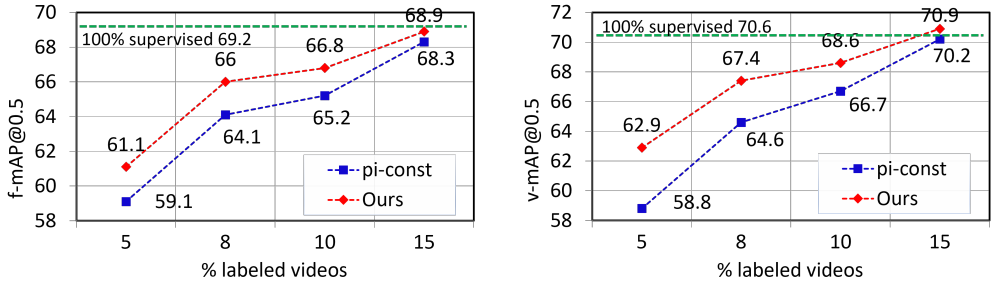

Figure 1: Comparison between Enhanced Mean Teacher and the state-of-the-art method Kumar & Rawat (2022) with varying proportions of labeled data in UCF101-24. Enhanced Mean Teacher provides comparable performance with 100% fully supervised approach with only 15% labels.

Video action detection, in contrast to image classification and object detection, poses additional challenges for semi-supervised learning. It is a more complex task that combines both classification and frame-wise action localization. It requires spatio-temporal localization of actions which suffers more degradation in performance under limited availability of labels. With the introduction of motion along the temporal dimension, the predictions need to be smooth and coherent. Therefore, it is hard to generate high-quality spatio-temporal pseudo-labels for semi-supervised learning.

To overcome these challenges, we propose ***Enhanced Mean Teacher***, a simple end-to-end framework where the student is trained using pseudo-labels generated by the teacher. The teacher and student are given different augmented input videos and the student gradually updates the teacher via Exponential Moving Average Tarvainen & Valpola (2017). In the proposed framework, 1) we study both ***classification*** and ***spatio-temporal consistencies*** to effectively utilize pseudo-labels generated by the teacher, 2) we propose a novel ***auxiliary module*** which learns from the student's mistakes on labeled samples and enhance the teacher by improving the spatio-temporal pseudo-labels generated on unlabeled set, and 3) we utilize several augmentation strategies spanning ***spatial*** as well as ***temporal augmentations*** to focus on both appearance and motion aspects of video in action detection which helps in generating effective spatio-temporal pseudo-labels from unlabeled samples.

In summary, we make the following contributions in this work:

- We propose a simple end-to-end *Enhanced Mean Teacher* for semi-supervised video action detection. To the best of our knowledge, this is the *first work* demonstrating the use of student-teacher network for video action detection.
- We propose a *novel auxiliary module*, which learns from the student's mistakes to improve the teacher and provide a better supervisory signal under limited labeled samples.
- We perform an extensive empirical analysis of consistency losses for this task which covers both *classification* as well as *spatio-temporal* aspects of video action detection.
- We present insights into the behavior of our framework with several augmentation techniques, including both *spatial* and *temporal augmentations*.

We perform a comprehensive empirical evaluation of the proposed approach on two different action detection benchmarks. Our study demonstrates significant improvement over supervised baselines by a good margin, consistently outperforming the sate-of-the-art approach. We also demonstrate the generalization capability of the proposed approach to video object segmentation.

2 Related Work

Video action detection Video action detection comprises two tasks: action classification and spatio-temporal localization. Some of the initial attempts to solve this problem focused on extending image-based object detectors such as RCNN (Ren et al., 2015), where detection is performed at frame-level which is later used for video-level activity classification Gkioxari et al. (2018); Yang et al. (2019); Hou et al. (2017); Yang et al. (2017); Li et al. (2018); Peng & Schmid (2016). The proposals on each frame are aggregated together to form an activity tube for spatio-temporal localization. The use of 3D convolutions to learn these proposals improved the performance as temporal context is utilized Hou et al. (2017). Most of these methods make use of a two-stage process, where

localization is performed with the help of a proposal network which is classified into activities in the second stage Gkioxari et al. (2018); Yang et al. (2019); Hou et al. (2017); Yang et al. (2017). Recently, some encoder-decoder based approaches have been developed which simplify the two-stage video action detection process Duarte et al. (2018); Zhao et al. (2022). In a more recent work Kumar & Rawat (2022), the authors further simplify VideoCapsuleNet Duarte et al. (2018) to reduce computation cost with minor performance trade-off. In this work, we make use of this optimized approach as our base model for video action detection.

Weakly-supervised learning Some recent works have proposed weakly-supervised approaches to overcome the high labeling cost for video action detection (Escorcia et al., 2020; Arnab et al., 2020; Mettes & Snoek, 2018; Zhang et al., 2020; Chéron et al., 2018; Mettes et al., 2017). These approaches require either video-level annotations or annotations only on selected frames. However, they rely on external actor detectors (Ren et al., 2015; Girshick et al., 2018; Liu et al., 2016) which introduces additional learning constraints. Even with the use of per-frame annotations along with video-level labels, the performance is far from satisfactory when compared with supervised baselines. In our work, we only use a subset of labeled videos that are fully annotated and demonstrate competitive performance when compared with supervised methods.

Semi-supervised learning In recent years, semi-supervised approaches have shown great promise in label efficient learning. Most of the efforts are focused on classification tasks where sample level annotation is required, such as object recognition Liu et al. (2022); feng Zhou et al. (2021) and video classification Singh et al. (2021); Xu et al. (2022). These efforts can be broadly categorized into iterative pseudo-labeling (Lee et al., 2013) and consistency-based (Berthelot et al., 2019; Sohn et al., 2020) learning. Consistency-based approaches are more efficient in terms of model training as the learning is performed in a single step in contrast to several iterations in iterative pseudo-labeling Rizve et al. (2020). The initial efforts in the consistency-based approach developed the PI-model where training is performed using consistency between augmented versions of a sample Laine & Aila (2017a). Mean teacher Tarvainen & Valpola (2017) improved this further where the pseudo-labels generated by the teacher are used to train a student. This approach has been successfully explored for both image classification Ke et al. (2019) as well as object detection Xu et al. (2021); Tang et al. (2021); feng Zhou et al. (2021); Chen et al. (2022); Liu et al. (2022; 2021). Different from all these, we focus on videos where the temporal dimension adds more complexity to the problem. There are some recent works focusing on videos, but they are limited to video classification task Jing et al. (2021); Singh et al. (2021); Xiao et al. (2022); Xu et al. (2022) where spatio-temporal localization is not required. We focus on video action detection, which requires spatio-temporal localization on every frame of the video in addition to video level class predictions. More recently, a PI-based consistency approach (Kumar & Rawat, 2022) has been explored for semi-supervised video action detection. Different from this, we propose a Mean Teacher based approach which achieves better performance.

3 METHODOLOGY

An overview of the proposed approach is illustrated in Figure 2. As shown in this Figure, Enhanced Mean Teacher consists of a student-teacher structure where the teacher model generates pseudo-labels using weak augmentations for the student who learns from these pseudo-labels on strongly augmented samples. In addition, the teacher also learns from the student's mistakes on labeled samples to improve its pseudo-labels with the help of an auxiliary module which is trained jointly.

Problem formulation Given a set of labeled samples $X_L : \{x_i, y_i\}_{i=0}^{i=N_l}$ and an unlabeled subset $X_U : \{x_i\}_{i=0}^{i=N_u}$, where x denotes a video and y is corresponding annotation with N_t labeled and N_u unlabeled samples, our goal is to train an action detection model (M) using both labeled and unlabeled data. The labeled videos are annotated with a ground-truth class and frame-level spatio-temporal localization denoted as y_t and f_t respectively. Each video sample (x_i), either from labeled or unlabeled set, is passed through a temporal augmenter (A_t) followed by a spatial augmenter (A_s) to obtain the strong (x_s) and weak (x_t) augmented views. We use the same action detection model M as a teacher (M_t) and as a student (M_s). Each of these models have two outputs; action classification logits, t_{cls} and s_{cls}, and raw spatio-temporal localization map, t_{loc} and s_{loc}, respectively for teacher and student. Similar to the action detection model, we have teacher (Aux_t) and student

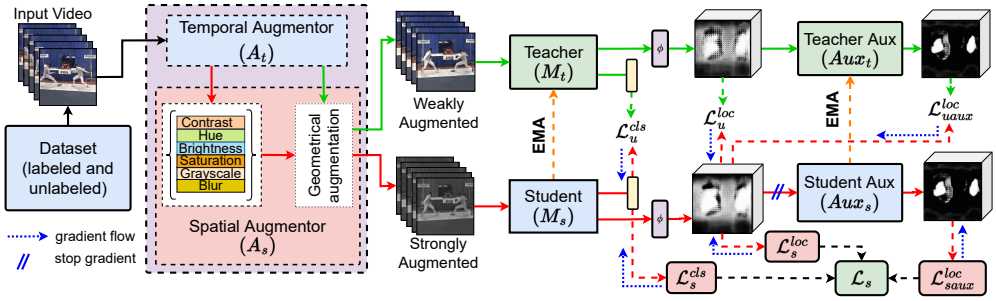

Figure 2: An overview of Enhanced Mean Teacher. The teacher model produce pseudo-labels using weak augmentation which are used to train the student model on strong augmentations. The teacher's pseudo-labels are further improved using Auxiliary module which learns from student's mistakes.

(Aux_s) auxiliary networks which learn from the student's mistakes. We pass $\phi(t_{loc})$ and $\phi(s_{loc})$ to auxiliary networks, Aux_t and Aux_s, which provide the transformed localization maps, ta_{loc} and sa_{loc} respectively, where a Sigmoid function (ϕ) is used to normalize the raw localization outputs.

In the next subsections, we first go over the differences between Enhanced Mean Teacher and Mean Teacher (Tarvainen & Valpola, 2017) setup, then, discuss the proposed auxiliary module, how augmented views are generated for teacher and student, and, finally, the losses used to train the model.

3.1 Enhanced Mean Teacher

Enhanced Mean Teacher follows a student-teacher training scheme. In each training iteration, both labeled and unlabeled videos are randomly sampled to train the student model. Similar to Mean Teacher, we use the teacher's prediction as a pseudo-label for the student model which attends to a stronger perturbed version of the video. To accomplish this, we generate two augmented views: a weak augmentation for the teacher, and a strong augmentation for the student. Strong augmentations make the student more robust and generalize to different variations, and, weak augmentation helps in providing a confident pseudo label by the teacher. Different from Mean Teacher, we also generate spatio-temporal pseudo-labels which are required for video action detection. In addition to spatial augmentations, we also rely on temporal augmentations which help in generating effective pseudo-labels for spatio-temporal localization. In our work, the capacity of teacher and student models are the same and we train the teacher and student model jointly from end to end. Similar to Mean Teacher, teacher's model parameters $(\theta_{teacher})$ are updated via Exponential Moving Average (EMA) of the student's model parameters $(\theta_{student})$ with a decay rate of β. This update can be defined as,

$$\theta_{teacher} = \beta\theta_{teacher} + (1-\beta)\theta_{student} \tag{1}$$

Apart from spatio-temporal pseudo-labels and temporal augmentations, Enhanced Mean Teacher also utilize an Auxiliary module Aux, which learns from the mistakes of the student network and transfer that learning to the teacher for generating better pseudo-labels.

3.2 Auxiliary Module

The performance of a Mean Teacher model relies on the quality of the pseudo-labels generated by the teacher. In the case of spatio-temporal pseudo-labels, it can be challenging for the model to generate high-quality pseudo-labels with a limited amount of labeled samples and the model is prone to make mistakes. We propose to learn from these mistakes on labeled samples and transfer that learning to the teacher model in order to improve the generated pseudo-labels.

Enhanced Mean Teacher utilizes an auxiliary module that helps the main action detection model to improve the localization prediction. The auxiliary module is based on a UNet-3D architecture Ronneberger et al. (2015). However, instead of a traditional RGB input which is generally used as an input for the UNet architecture (Ronneberger et al., 2015), we take the normalized single channel teacher's spatio-temporal localization prediction t_{loc} as the input. With M_t already approximating the action location, the task of Aux_t is to further improve the action boundary and temporal coherency. The auxiliary module also helps in amplifying the confidence of the action region which in turn serves as a better supervisory signal for the student (s_{loc}). With similar motivation, the authors in Pham et al. (2021) proposed to utilize feedback from student for teacher based on meta-learning

with additional computation cost. This was mainly focused on image classification and requires two-step training. The proposed auxiliary module is designed for spatio-temporal pseudo-labels and can be trained jointly with the base model in end-to-end training.

The model parameters (θ_{taux}) for Aux_t are updated via EMA of Aux_s parameters (θ_{saux}) with the same decay rate β as describe in Equation 1. The update is defined as,

$$\theta_{taux} = \beta\theta_{taux} + (1 - \beta)\theta_{saux} \tag{2}$$

The auxiliary network Aux_s does not use any pre-trained weights and is trained on labeled samples. The student's prediction will be more distorted than the teacher's because of the strong data augmentation. This in turn will help Aux_t to provide a more confident pseudo label on a weakly augmented localization map ($\phi(t_{loc})$). The base action detection model and auxiliary module are trained jointly but the gradients from the auxiliary module are not used to update the base model. This ensures that the improvement of pseudo-labels is not dependent on the input video and is sample agnostic. This helps in learning a transformation that generalizes well to unlabeled samples.

3.3 AUGMENTATIONS

We study both spatial and temporal augmentations to generate weak and strong views for a sample. A video sample is passed through temporal (A_t) and spatial (A_s) augmenters sequentially to get the two views. The temporal augmenter A_t has three modes; 1) normal: the skip rate is consistent and we follow the sequential flow of frames, 2) temporal reversal: instead of the original order, we flip the order of frames temporally, and, 3) random selection: the frames are sampled sequentially but at random skip rates. For each sample, one of these modes is randomly selected. After temporal augmentation, the video passes through the spatial augmenter A_s. The strong view includes the following augmentations: Gaussian blur, horizontal flip, grayscale, hue, saturation, brightness, and contrast, whereas, the weak view goes through one geometrical transformation, which performs horizontal flipping. Some recent supervised approaches in the video action detection domain (Li et al., 2020; Pan et al., 2021; Zhao et al., 2022) incorporate similar spatial augmentations, however, the proposed temporal augmentations have not been studied for this task. The use of temporal augmentations also differ from the semi-supervised video action classification works (Jing et al., 2021; Singh et al., 2021; Xiao et al., 2022; Xu et al., 2022) where the complex spatial augmentation approaches for images Cubuk et al. (2020); Sohn et al. (2020) are extended to videos.

3.4 LOSSES

The objective function of Enhanced Mean Teacher has two parts: supervised loss $\mathcal{L}_s$ and unsupervised loss $\mathcal{L}_u$. For both supervised and unsupervised losses, we have classification ($\mathcal{L}_s^{cls}, \mathcal{L}_u^{cls}$) and localization loss ($\mathcal{L}_s^{loc}, \mathcal{L}_u^{loc}$) respectively. The final objective function $\mathcal{L}$ is defined as $\mathcal{L} = \mathcal{L}_s + \lambda\mathcal{L}_u$ where λ is a weight parameter for unsupervised loss. For the supervised losses, we use Spread Loss for classification and BinaryCrossEntroply plus Logits and Dice Loss for localization as in baseline action detection model (Kumar & Rawat, 2022). We calculate the supervised loss on the labeled subset of student's predictions ($\mathcal{L}_s^{cls}, \mathcal{L}_s^{loc}$) and student's auxiliary network predictions ($\mathcal{L}_{saux}^{loc}$), and unsupervised loss on labeled plus unlabeled subset. Next, we discuss unsupervised losses in detail, which mainly constitutes of classification and localization-level consistency.

Classification consistency To minimize the difference between teachers' prediction t_{cls} and student's prediction s_{cls}, we use classification consistency. We compute the consistency of the predicted probability distribution for the video class where we utilize logits instead of class labels. This was based on our preliminary experiments where we observed that the consistency between logits is more effective when compared with the consistency between soft-labels. This is different from Mean Teacher where the predicted labels were found more effective for image classification. We utilize Jenson-Shennon Divergence (JSD) for this and the loss is computed as $\mathcal{L}_u^{cls} = JSD(t_{cls}, s_{cls})$.

Localization-level consistency Spatio-temporal localization consistency focus on the localization of actions on every frame of the input video. It is computed at pixel-level on each frame between the teacher's and the student's predicted localization map. This is different from semi-supervised object detection task Xu et al. (2021); feng Zhou et al. (2021); Liu et al. (2021) where regression is performed on the coordinates of a bounding-box for frame-level detection. Our approach is agnostic

Table 1: Comparison of performance on UCF101-24 on *baseline setup*. First row is supervised score on labeled subset. CLS and LOC is classification and spatio-temporal consistency. Blue shows the absolute gain from supervised, and, green shows best performing configuration.

		f-mAP		v-mAP	
CLS	LOC	0.2	0.5	0.2	0.5
		78.4 ± 1.20	50.2 ± 2.35	86.1 ± 1.40	46.2 ± 2.75
✓		87.4 ± 0.45 (↑ 9.0)	63.8 ± 0.25 (↑ 13.6)	93.7 ± 0.65 (↑ 7.6)	64.0 ± 1.75 (↑ 17.8)
	✓	87.4 ± 0.60 (↑ 9.0)	65.4 ± 1.50 (↑ 15.2)	94.2 ± 0.45 (↑ 8.2)	65.7 ± 1.70 (↑ 19.5)
✓	✓	87.0 ± 0.10 (↑ 8.6)	65.4 ± 1.00 (↑ 15.2)	93.1 ± 0.15 (↑ 7.0)	66.5 ± 1.75 (↑ 20.3)

to the type of detections; it can be used for both bounding-box and pixel-wise detection for spatio-temporal localization. We utilize L2 distance to compute the spatio-temporal consistency. The pixel-wise difference between localization map could be either between raw logits (t_{loc}, s_{loc}) or between the normalized outputs $(\phi(t_{loc}), (\phi(s_{loc}))$.

We have two localization-level consistency losses; 1) between teacher's (M_t) and student's (M_s) raw logits predictions $\mathcal{L}_u^{loc} = L2(s_{loc}, t_{loc})$, and, 2) between the sigmoid of auxiliary teacher network (Aux_t) and student's (M_s) predicted localization maps, $\mathcal{L}_{uaux}^{loc} = L2(\phi(s_{loc}), \phi(ta_{loc}))$.

Finally, the overall training objective for Enhanced Mean Teacher is defined as,

$$\mathcal{L} = \mathcal{L}_s + \lambda \mathcal{L}_u = (\mathcal{L}_s^{cls} + \mathcal{L}_s^{loc} + \mathcal{L}_{saux}^{loc}) + \lambda(\mathcal{L}_u^{cls} + \mathcal{L}_u^{loc} + \mathcal{L}_{uaux}^{loc}) \tag{3}$$

4 Experiments

Action detection and auxiliary model architecture Following recent work on semi-supervised action detection (Kumar & Rawat, 2022), we use VideoCapsuleNet (Duarte et al., 2018) as our base action detection model. It is a simple encoder-decoder based architecture that utilizes capsule routing. Different from the original model, we use 2D routing instead of 3D routing which makes it computationally efficient. This also maintains consistency with the previous work (Kumar & Rawat, 2022) and enables a fair comparison. For the auxiliary module, we use a 3D UNet (Ronneberger et al., 2015) architecture with a depth of 3 layers with 8, 16, and 32 channels respectively. More details about the architecture are provided in supplementary.

Datasets We use three publicly available datasets to perform our experiments, two for video action detection (UCF101-24 and JHMDB-21), and one to show generalization on YouTube-VOS (Xu et al., 2018b). UCF101-24 contains 3k videos which is split between training and test as 2.1k and 0.9k respectively. JHMDB-21 has 900 videos with 600 for training and 300 for testing. The resolution of the video is 320x240 for both of the datasets. The number of classes in UCF101-24 is 24 and in JHMDB-21 it's 21. UCF101-24 is an untrimmed dataset whereas JHMDB-21 is a trimmed dataset. The labeled and unlabeled subset for UCF101-24 is divided in the ratio of 10:90 and for JHMDB-21 it's 30:70. The distribution of the number of training and evaluation videos on YouTube-VOS is 3471 and 589 respectively. We use the 10:90 ratio for labeled to unlabeled subsets.

Implementation details We train Enhanced Teacher Model for 50 epochs. The batch size is set to 8 where the number of samples from both labeled and unlabeled subsets are the same. We use Adam optimizer with an initial learning rate set to 0.0001. The value of β for the main and auxiliary teacher's EMA parameters update is set to 0.97 which follows previous works. The value of λ for the unsupervised loss weight is set to 0.1 which was determined empirically. Next, we discuss the augmenter setup to generate the two augmented views. Firstly, a consistent same area of 224x224 is cropped from the video for both views. During training time, cropping is random, whereas during testing we take a center crop. Both the videos are passed through A_t to maintain the frame ids. After that, to get the strong augmented view, it is passed through a series of spatial augmentations. Finally, both the weak and strong augmented view is passed through a geometrical transformation (Random Horizontal Flip) which keeps the localization coherent. We use a normalized random probability map to apply the series of spatial augmentations. More details are provided in the supplementary.

Evaluation metrics We evaluate the proposed approach using frame metric average precision (f-mAP) and video metric average precision (v-mAP). f-mAP is computed by summing over all the

Table 2: Comparison with previous state-of-the art approaches on fully supervised, weakly-supervised and semi-supervised learning on UCF101-24 and JHMDB-21. † shows approach using Optical Flow as second modality, and, $*$ shows the modified version of original work (Duarte et al., 2018). The last row shows the score for supervised labeled subset, that is 10% for UCF101-24 and 30% for JHMDB-21. Best score on each metric is underlined.

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								
Kalogeiton et al. (2017)	✓		69.5	76.5	49.2	65.7	74.2	73.7
Song et al. (2019)†	✓		72.1	77.5	52.9	65.5	74.1	73.4
Zhao & Snoek (2019)†	✓		-	78.5	50.3	-	-	74.7
Li et al. (2020)	✓		78.0	82.8	53.8	70.8	77.3	70.2
Hou et al. (2017)		✓	41.4	47.1	-	61.3	78.4	76.9
Sun et al. (2018)		✓	-	-	-	77.9	-	80.1
Pan et al. (2021)		✓	84.3	-	-	-	-	-
Zhao et al. (2022)		✓	83.2	83.3	58.4	-	87.4	82.3
Duarte et al. (2018)		✓	78.6	97.1	80.3	64.6	95.1	-
Duarte et al. (2018)*		✓	69.2	95.3	71.9	68.1	96.8	68.4
Weakly-Supervised								
Mettes & Snoek (2018)	✓		-	41.8	-	-	-	-
Chéron et al. (2018)		✓	-	43.9	17.7	-	-	-
Escorcia et al. (2020)		✓	45.8	19.3	-	-	-	-
Arnab et al. (2020)		✓	-	61.7	35.0	-	-	-
Zhang et al. (2020)		✓	30.4	45.5	17.3	65.9	77.3	50.8
Semi-Supervised								
MixMatch Berthelot et al. (2019)		✓	10.3	54.7	4.9	7.5	46.2	5.8
Psuedo-label Lee et al. (2013)		✓	59.3	89.9	58.3	57.4	90.1	57.4
Co-SSD(CC)Jeong et al. (2019)		✓	60.2	91.3	64.0	60.7	94.3	58.5
Kumar & Rawat (2022)		✓	65.2	93.1	66.7	64.4	95.4	63.5
Tarvainen & Valpola (2017) *		✓	56.6	90.6	56.9	60.9	92.3	58.6
Ours		✓	66.8	94.6	68.6	67.4	95.0	67.2
Supervised baseline		✓	47.8	84.9	43.3	64.3	93.2	61.9

frames with an IoU greater than a certain threshold. Similarly, for v-mAP 3D IoU is utilized instead of frame-level IoU. We compute these scores at multiple thresholds where we have shown results at 0.2 and 0.5 in the main paper with all other thresholds in the supplementary.

4.1 Results and comparisons

Preliminary Analysis We first study a baseline setup of Mean Teacher where we investigate the proposed classification and spatio-temporal consistency independently as well as in conjunction. Since the focus of this study is not to examine different hyperparameters settings, we set the weight to 1 when both consistencies are used. These preliminary evaluations are shown in Table 1, where we can observe that there's an improvement with both classification and localization consistency over the baseline. On UCF101-24, the approximate gain at f-mAP@0.5 and v-mAP@0.5 is 14-15% and 17-20% respectively and the gain at 0.2 threshold is between 7-9%. Similarly, on JHMDB-21, the gain is between 2-5% at both thresholds. JHMDB-21 is a very small dataset, therefore the improvement in performance with these consistencies is not as good as UCF101-24 due to limited amount of unlabeled samples. To compare the size of these two datasets, UCF101-24 10% is roughly 200+ videos, and, on the other hand, JHMDB-21 30% is 180+ videos. Thus, UCF101 has a clear advantage when it comes to having a large amount of unlabeled data compared to JHMDB-21. Apart from this, JHMDB-21 has pixel-wise annotations as compared to UCF101-24 which has bounding-box annotations, which also makes this dataset more challenging. Considering the improvement on both datasets, we show that our approach is suitable for both these variations.

Looking into each consistency loss from Table 1, we observe that standalone spatio-temporal localization consistency outperform classification consistency. when both classification and localization consistencies are applied, there's no gain on the performance. Thus, we specifically focus on the spatio-temporal localization consistency in our proposed model.

Table 3: Effect of using Auxiliary module. L2: Base model teacher's logits, Aux-L is auxiliary teacher's logits and Aux-S is sigmoid of auxiliary teacher's output. First three rows use only single loss between teacher and student. Last two rows shows the score for combination of consistency between main model and student, and auxiliary network and student.

			UCF101-24				JHMDB-21			
L2	Aux-L	Aux-S	f-mAP		v-mAP		f-mAP		v-mAP	
			0.2	0.5	0.2	0.5	0.2	0.5	0.2	0.5
✓			87.6	66.6	94.5	67.4	90.7	65.8	94.8	65.1
	✓		86.4	66.0	91.9	68.2	91.5	67.4	93.6	67.2
		✓	87.2	64.9	92.7	65.7	88.4	65.7	92.8	65.2
✓	✓		85.6	65.1	91.6	66.2	91.7	67.0	98.4	65.3
✓		✓	89.0	66.8	94.6	68.6	90.5	64.7	96.9	64.4

Comparison with state-of-the-art approaches In Table 2, we compare our work with recent supervised, weakly-supervised, and semi-supervised approaches. Going through each of them in detail, *firstly* in **supervised** scenario, with only 10% labeled data we outperform all the 2D-based approaches on v-mAP. Amongst 3D-based methods, we outperform a few of them and show competitive performance with others. Most of the 2D approaches use optical flow as the second modality, whereas, our work is based on the architecture which uses only a single modality as input. *Secondly*, amongst **weakly-supervised** approaches, our approach surpasses the state-of-the-art on both datasets by a good margin. On UCF101-24, our approach outperforms by an approximate margin of 20-30% at the 0.5 thresholds. On JHMDB-21, we have an absolute improvement of 1.5% on f-mAP@0.5 and 16.4% on v-mAP@0.5 compared to Zhang et al. (2020). Lastly, looking into **semi-supervised** approaches, the first two rows show the performance of two image-based approaches, the third row shows the score based on one object detection approach, and lastly, Kumar & Rawat (2022) is a pi-consistency based approach for video action detection. The setup for image and object detection-based approaches is similar to Kumar & Rawat (2022). Berthelot et al. (2019) is a heavy augmentation-based approach that is not able to generalize well with this less amount of videos. Apart from that, we beat the pseudo-label based approach on all thresholds. Compared with the semi-supervised object detection approach, we outperform it by 4-6% on UCF101-24 and 7-9% on JHMDB-21. Lastly, compared with a parallel approach to semi-supervised video action detection, we have a gain of 1.6% on f-mAP@0.5 +0.9% on v-mAP@0.5 on the UCF101-24 dataset. On JHMDB-21, the gain is 3.0% and 3.7% at f-mAP@0.5 and v-mAP@0.5 respectively.

4.2 Ablations

Base vs Enhanced Mean Teacher: Here, we compare the base Mean Teacher, Table 1 second row, with the proposed Enhanced Mean Teacher approach. There's a performance boost of 10-12% on UCF101-24 and 6-8% on JHMDB-21 at the 0.5 thresholds. This shows the effectiveness of our proposed framework on top of the base Mean Teacher.

Effectiveness of auxiliary module: In Table 3, we observe that Aux-L, the logit loss between the teacher's auxiliary output and the student's model, has the best score among the three variations on both UCF101-24 and JHMDB-21. This demonstrates that the proposed pseudo label refinement boost the student's model performance. Further, when the base and auxiliary module losses are combined, although the gain is not much on JHMDB-21, it still beats the baseline model consistency loss only (L2). On UCF101-24, L2+Aux-S outperforms all the other variations.

Effectiveness of augmentations: We analyze the performance gain by applying different types of augmentations. In Fig. 3, we can observe that incorporating both spatial and temporal augmentations outperforms a standalone set of augmentations. This further supports our hypothesis that spatio-temporal augmentation makes the model more robust. Looking into UCF101-24 vs JHMDB-21, the temporal augmentation alone seems not as effective as spatial augmentation for JHMDB-21. This can be attributed to the nature of pixel-wise annotations in JHMDB-21 where interpolation between frames can be challenging for the proposed temporal augmentations.

4.3 Discussion and analysis

We further answer some of the important set of questions pertaining to Enhanced Mean Teacher approach for semi-supervised activity detection in this section.

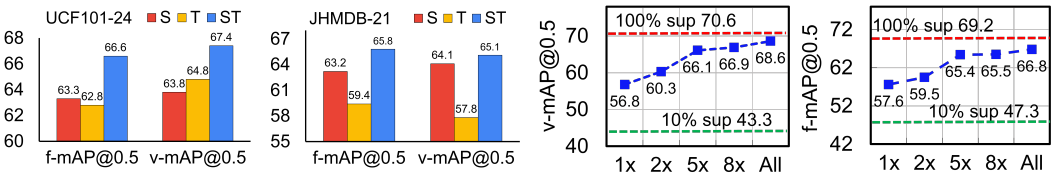

Figure 3: Left two histograms show the effect of different type of augmentations on performance for UCF101-24 and JHMDB-21. S: Spatial, T: Temporal, ST: Spatio-Temporal. Plots on the right shows improvement in v-mAP and f-mAP with increase in unlabeled samples for UCF101-24.

Table 4: Performance comparison on Youtube-VOS (Xu et al., 2018b). J_s and J_u are Jaccard on seen and unseen categories. Similarly, F_s and F_u are boundary metric on seen and unseen categories.

Approach	Avg	J_S	J_U	F_S	F_U
Xu et al. (2018a)†	10.1	11.6	10.1	9.6	9.2
Kumar & Rawat (2022)	36.8	43.1	31.4	40.8	31.8
Ours	42.1	49.1	35.9	47.5	36.0
Xu et al. (2018a)†(100%)	47.9	55.7	39.6	55.2	41.3

Does an increase in unlabeled data helps? In this experiment setting, we keep the labeled percentage of data fixed to 10% and increase the amount of unlabeled data from 1x i.e. (1/10th of 90%) all the way to 10x (90%). From Fig. 3, we can observe that there is a difference of approximately 10% in 1x and 10x which demonstrates that the amount of unlabeled data plays a significant role.

Qualitative analysis of auxiliary module: We analyze the effect of the auxiliary module by comparing the output of the localization map of the main model and the auxiliary module. We observe that the teacher's auxiliary network predictions provide more emphasis to the activity region on the frame (Figure 4 supplementary). At the same time, the network is generating a distinctive boundary between foreground and background compared to teacher and student's predictions.

Burn-in vs end-to-end? Some recent approaches have shown the benefit of pre-training on labeled set for model initialization (Liu et al., 2021). In this experiment, we analyzed the effect of burn-in on proposed Enhanced Mean Teacher and pre-trained the model on labeled set and use it for combined training on labeled and unlabeled sets. We observe that f-mAP/v-mAP@0.5 decreased by 0.2 to 66.6 and 68.4 respectively, and there was no gain to Enhanced Mean Teacher with burn-in weights.

4.4 Generalization to Video Object Segmentation

We further demonstrate the generalization capability of Enhanced Mean Teacher. We outperform the supervised baseline by an absolute margin of 33% on average. Compared to the previous semi-supervised approach (Kumar & Rawat, 2022), our approach shows a gain of 5-7% on all metrics.

5 Conclusion

In this work, we propose Enhanced Mean Teacher, a novel student-teacher based method for semi-supervised video action detection. Different from prior works in student-teacher, it is focused on videos where spatio-temporal localization is required in addition to video level class scores. We investigate both spatial and temporal aspects of a video for designing consistency and augmentation strategies for Enhanced Mean Teacher. We also propose a novel auxiliary module which learn from the student's mistakes and transfer that knowledge to the teacher to generate better pseudo labels for the student. We demonstrate the effectiveness of Enhanced Mean Teacher on two different action detection datasets with extensive set of experiments. Furthermore, we also show its performance on video object segmentation validating its generalization capability to other tasks in video domain.

6 Limitations

The distribution of data in labeled and unlabeled set is mostly balanced in the studied action detection datasets. An imbalance in class distribution could affect the proposed model and lead to a biased performance. The progress of minority classes will not go hand in hand compared to the developments in the field of computer vision. However, we believe this problem is common for most of the semi-supervised approaches and it is an important direction which should be further studied.

REFERENCES

A. Arnab, Chen Sun, Arsha Nagrani, and C. Schmid. Uncertainty-aware weakly supervised action detection from untrimmed videos. *ArXiv*, abs/2007.10703, 2020. 3, 7

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 (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/1cd138d0499a68f4bb72bee04bbec2d7-Paper.pdf. 3, 7, 8

Binghui Chen, Pengyu Li, Xiang Chen, Biao Wang, Lei Zhang, and Xian-Sheng Hua. Dense learning based semi-supervised object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4815–4824, June 2022. 1, 3

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 (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper/2018/file/53fde96fcc4b4ce72d7739202324cd49-Paper.pdf. 3, 7

Ekin Dogus Cubuk, Barret Zoph, Jon Shlens, and Quoc Le. Randaugment: Practical automated data augmentation with a reduced search space. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 18613–18624. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/d85b63ef0ccb114d0a3bb7b7d808028f-Paper.pdf. 5

Kevin Duarte, Yogesh S Rawat, and Mubarak Shah. Videocapsulenet: A simplified network for action detection. *Advances in Neural Information Processing Systems*, 2018. 3, 6, 7

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. 3, 7

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)*, pp. 4079–4088, 2021. 1, 3, 5

Ross Girshick, Ilija Radosavovic, Georgia Gkioxari, Piotr Dollár, and Kaiming He. Detectron. https://github.com/facebookresearch/detectron, 2018. 3

Georgia Gkioxari, Ross Girshick, Piotr Dollár, and Kaiming He. Detecting and recognizing human-object interactions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8359–8367, 2018. 2, 3

Rui Hou, Chen Chen, and Mubarak Shah. Tube convolutional neural network (t-cnn) for action detection in videos. In *IEEE International Conference on Computer Vision*, 2017. 2, 3, 7

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 (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/d0f4dae80c3d0277922f8371d5827292-Paper.pdf. 7

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)*, pp. 11597–11606, 2021. 1

Longlong Jing, Toufiq Parag, Zhe Wu, Yingli Tian, and Hongcheng Wang. Videossl: Semi-supervised learning for video classification. *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1109–1118, 2021. 1, 3, 5

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)*, pp. 4415–4423, 2017. doi: 10.1109/ICCV.2017.472. 7

Zhanghan Ke, Daoye Wang, Qiong Yan, Jimmy S. J. Ren, and Rynson W. H. Lau. Dual student: Breaking the limits of the teacher in semi-supervised learning. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6727–6735, 2019. 3

Akash Kumar and Yogesh Singh Rawat. End-to-end semi-supervised learning for video action detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022. 2, 3, 5, 6, 7, 8, 9

Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. *ArXiv*, abs/1610.02242, 2017a. 3

Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. *ArXiv*, abs/1610.02242, 2017b. 1

Dong-Hyun Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, volume 3, pp. 896, 2013. 3, 7

Dong Li, Zhaofan Qiu, Qi Dai, Ting Yao, and Tao Mei. Recurrent tubelet proposal and recognition networks for action detection. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 303–318, 2018. 2

Yixuan Li, Zixu Wang, Limin Wang, and Gangshan Wu. Actions as moving points. In *arXiv preprint arXiv:2001.04608*, 2020. 5, 7

W. Liu, Dragomir Anguelov, D. Erhan, Christian Szegedy, Scott E. Reed, Cheng-Yang Fu, and A. Berg. Ssd: Single shot multibox detector. In *ECCV*, 2016. 3

Yen-Cheng Liu, Chih-Yao Ma, Zijian He, Chia-Wen Kuo, Kan Chen, Peizhao Zhang, Bichen Wu, Zsolt Kira, and Peter Vajda. Unbiased teacher for semi-supervised object detection. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021. 1, 3, 5, 9

Yen-Cheng Liu, Chih-Yao Ma, and Zsolt Kira. Unbiased teacher v2: Semi-supervised object detection for anchor-free and anchor-based detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9819–9828, June 2022. 1, 3

Pascal Mettes and Cees G. M. Snoek. Pointly-supervised action localization. *International Journal of Computer Vision*, 127:263–281, 2018. 3, 7

Pascal Mettes, Cees G. M. Snoek, and Shih-Fu Chang. Localizing actions from video labels and pseudo-annotations. *ArXiv*, abs/1707.09143, 2017. 3

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*, pp. 464–474, 2021. 5, 7

Xiaojiang Peng and Cordelia Schmid. Multi-region two-stream r-cnn for action detection. In *European conference on computer vision*, pp. 744–759. Springer, 2016. 2

Hieu Pham, Zihang Dai, Qizhe Xie, and Quoc V Le. Meta pseudo labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11557–11568, 2021. 4

Antti Rasmus, Harri Valpola, Mikko Honkala, Mathias Berglund, and Tapani Raiko. Semi-supervised learning with ladder network. *ArXiv*, abs/1507.02672, 2015. 1

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pp. 91–99, 2015. 2, 3

Mamshad Nayeem Rizve, Kevin Duarte, Yogesh S Rawat, and Mubarak Shah. In defense of pseudo-labeling: An uncertainty-aware pseudo-label selection framework for semi-supervised learning. In *International Conference on Learning Representations*, 2020. 1, 3

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *ArXiv*, abs/1505.04597, 2015. 4, 6

Mehdi S. M. Sajjadi, Mehran Javanmardi, and Tolga Tasdizen. Regularization with stochastic transformations and perturbations for deep semi-supervised learning. In *NIPS*, 2016. 1

Ankit Singh, Omprakash Chakraborty, Ashutosh Varshney, Rameswar Panda, Rogério Schmidt Feris, Kate Saenko, and Abir Das. Semi-supervised action recognition with temporal contrastive learning. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10384–10394, 2021. 1, 3, 5

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 (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 596–608. Curran Associates, Inc., 2020. URL `https://proceedings.neurips.cc/paper/2020/file/06964dce9addb1c5cb5d6e3d9838f733-Paper.pdf`. 3, 5

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. 7

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

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)*, pp. 3131–3140, 2021. 1, 3

Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *NIPS*, 2017. 1, 2, 3, 4, 7

Junfei Xiao, Longlong Jing, Lin Zhang, Ju He, Qi She, Zongwei Zhou, Alan Yuille, and Yingwei Li. Learning from temporal gradient for semi-supervised action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3252–3262, June 2022. 1, 3, 5

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, 5

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, 2018a. 9

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, 2018b. 6, 9

Yinghao Xu, Fangyun Wei, Xiao Sun, Ceyuan Yang, Yujun Shen, Bo Dai, Bolei Zhou, and Stephen Lin. Cross-model pseudo-labeling for semi-supervised action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2959–2968, June 2022. 1, 3, 5

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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 264–272, 2019. 2, 3

Zhenheng Yang, Jiyang Gao, and Ram Nevatia. Spatio-temporal action detection with cascade proposal and location anticipation. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2017. 2, 3

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. doi: 10.1109/TMM.2019.2959425. 3, 7, 8

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. 7

Jiaojiao Zhao, Yanyi Zhang, Xinyu Li, Hao Chen, Bing Shuai, Mingze Xu, Chunhui Liu, Kaustav Kundu, Yuanjun Xiong, Davide Modolo, Ivan Marsic, Cees G. M. Snoek, and Joseph Tighe. Tuber: Tubelet transformer for video action detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13598–13607, June 2022. 3, 5, 7

A APPENDIX

A.1 QUALITATIVE ANALYSIS OF AUXILIARY MODULE

To visualize the effectiveness of the auxiliary, we show the predicted localization maps in Fig. 4 and 5. The figure shows the example from JHMDB dataset where the ground truth is a pixelwise detection, and. from UCF101-24 where it's bounding box. Firstly, looking into JHMDB-21, comparing across last three columns, we can see that the localization map of the teacher's auxiliary model is more confident about the actor location and the background. Main model teacher and student prediction is almost same since the weights of student is copied to the teacher. There's minor difference due to the strongly augmented view for the student network.

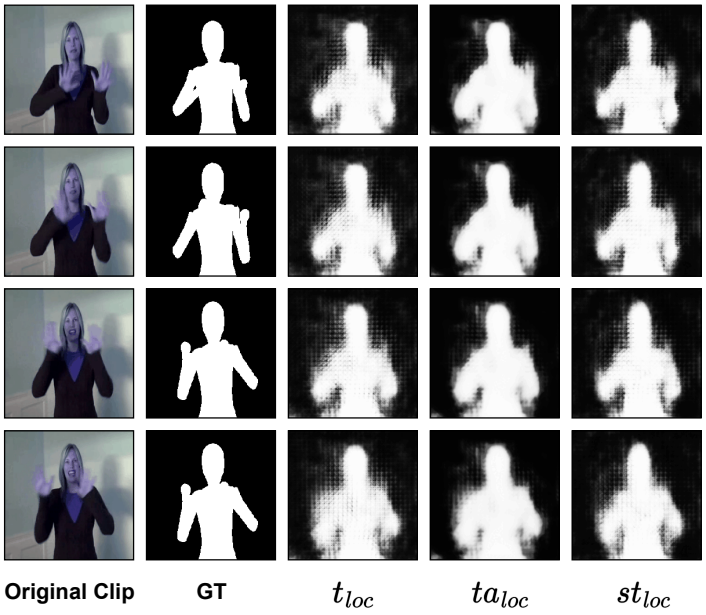

Figure 4: This figure shows the original clip, it's ground truth localization map and main and auxiliary network's predictions. These are consecutive frames processed by the model. Sequentially, this is what columns are depicting: original video, GT: Ground truth, t_{loc}: M_t's prediction, ta_{loc}: Aux_t's prediction, st_{loc}: M_s's prediction. The example is taken from JHMDB-21 dataset.

In UCF101-24 example 5, although the bounding box position looks same across frames, it's one of the challenging conditions. Here, not only the motion of actor is fast, there's camera motion as well. Contrasting main model's teacher and student predictions against Auxiliary teacher's output, the figure shows that Auxiliary teacher output is more confident where the action is happening and what is background. On the other hand, teacher's and student's prediction are more distorted and have less confidence.

A.2 AUXILIARY ARCHITECTURE

In our work, we use a modified version of UNet 3D architecture. UNet3D is a simple extension of it's 2D version. 2D Convolution block is replaced by a 3D convolution block and the upsample mode is *trilinear* instead of *bilinear*. Original UNet 3D model has a lot of trainable parameters. To reduce the extra overhead of trainable paramters, we reduce the depth of our 3D UNet architecture. Original UNet has 5 channels depth, and the variation in depth goes like this, $32 \rightarrow 64 \rightarrow 128 \rightarrow 256 \rightarrow 512 \rightarrow 256 \rightarrow 128 \rightarrow 64 \rightarrow 32$. In our case, we reduce the number of channels. Our architecture have this variation, $8 \rightarrow 16 \rightarrow 32 \rightarrow 16 \rightarrow 8$. This brings down the number of trainable parameters to approximately 21k. We also changed the auxiliary model with various depth and compared the performance. We tried the variation with adding and removing one more depth. The performance

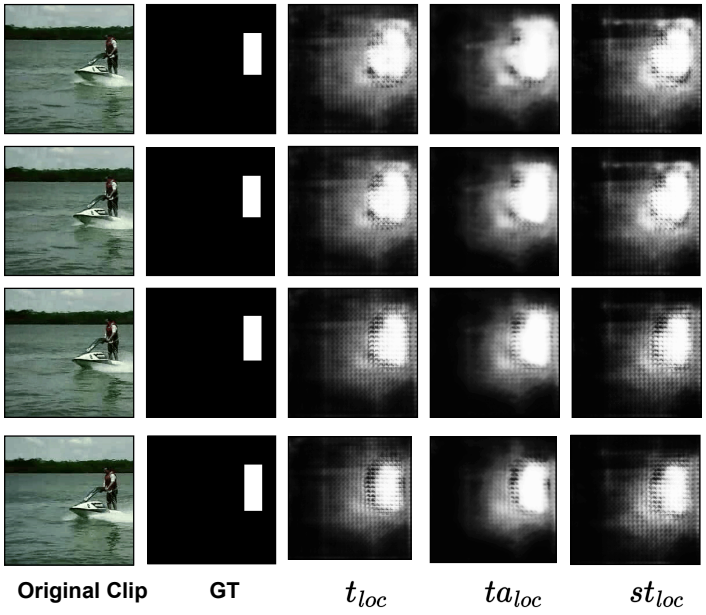

Figure 5: This figure shows the original clip, it's ground truth localization map and main and auxiliary network's predictions. These are consecutive frames processed by the model. Sequentially, this is what columns are depicting: original video, GT: Ground truth, t_{loc}: M_t's prediction, ta_{loc}: Aux_t's prediction, st_{loc}: M_s's prediction. The example is taken from UCF101-24 dataset.

Table 5: Modification in depth of 3D Unet architecture. The number of channels are 8, 16, 32, 64. We pick first two for the auxiliary network's depth of 2, first three for auxiliary network's depth 3 and so on.

	2	3	4
fmap@0.5	66.8	**66.8**	66.4
vmap@0.5	68.4	**68.6**	68.0

is shown in Table 5. From the Table, we can see that the network with depth three have the best performance on both f-mAP and v-mAP at 0.5 thresholds.

A.3 Additional Results

In our work, we run on three different seeds and then calculated mean and variance over that. We will discuss those results here.

Baseline Setup: JHMDB-21 In Table 6, we show the scores on *baseline setup* on JHMDB-21 dataset with three different seed runs. There's performance gain from the supervised score for each standalone consistency and when both consistency are applied simultaneously. This is crucial because JHMDB-21 is a challenging dataset for two prominent readons: 1)Less number of videos available for training, and, 2)Ground truth is a pixelwise detection, instead of bounding box detection. We see that there's approximately 1% gain f-mAP@0.2 and 2-5% gain on f-mAP@0.5. Similarly, on v-mAP, at 0.2 threshold, the gain is 2-3% and at 0.5, performance boost up by 4-5%.

Effect of Auxiliary Module To demonstrate the effective of auxiliary network, we perform different seed runs on JHMDB-21 dataset. From Table 7, we show that calculating the spatio-temporal localization consistency loss between teacher's auxiliary network and student's prediction helps in all cases except one. In Aux-S case, the sigmoid layer constraints the value between 0 and 1, and, thus the loss value is very low, and the penalty is less, therefore, the loss value is low. On the other hand, if we use only logits, Aux-L case, we see a performance boost of 2-3% on f-mAP and v-

Table 6: Comparison of performance on JHMDB-21 on *baseline setup*. First row is supervised score on labeled subset. CLS and LOC is classification and spatio-temporal localization consistency. Blue shows the absolute gain from supervised, and, green shows the which consistency has the most performance boost.

		f-mAP		v-mAP	
		0.2	0.5	0.2	0.5
CLS	LOC	88.4 ± 1.70	61.3 ± 2.30	92.9 ± 1.00	59.5 ± 2.75
✓		89.7 ± 1.45 (↑ 1.3)	66.5 ± 2.10 (↑ 5.2)	95.1 ± 0.65 (↑ 2.2)	64.7 ± 2.75 (↑ 5.2)
	✓	89.2 ± 0.95 (↑ 0.8)	64.3 ± 1.95 (↑ 3.0)	95.8 ± 1.05 (↑ 2.9)	63.4 ± 3.20 (↑ 3.9)
✓	✓	88.8 ± 0.70 (↑ 0.4)	63.5 ± 1.30 (↑ 2.2)	95.6 ± 1.15 (↑ 2.7)	63.9 ±1.20 (↑ 4.4)

Table 7: Effect of using Auxiliary module on JHMDB-21 dataset. L2: Base model teacher's logits, Aux-L is auxiliary teacher's logits and Aux-S is sigmoid of auxiliary teacher's output. First three rows use only single loss between teacher and student. Last two rows shows the score for combination of consistency between main model and student, and auxiliary network and student.

			f-mAP		v-mAP	
L2	Aux-L	Aux-S	0.2	0.5	0.2	0.5
✓			89.2 ± 0.95	64.3 ± 1.95	95.8 ± 1.05	63.4 ± 3.20
	✓		90.9 ± 0.45	66.1 ± 1.30	94.9 ± 1.20	66.2 ± 1.20
		✓	88.5 ± 1.30	65.0 ± 0.90	94.4 ± 1.30	65.1 ± 0.35
✓	✓		90.3 ± 1.40	65.8 ± 1.10	97.4 ± 0.95	65.2 ± 0.25
✓		✓	89.5 ± 0.85	64.1 ± 0.45	96.9 ± 0.40	64.4 ± 1.20

Table 8: Effect of using Auxiliary module. Comparison between *baseline setup* and Enhanced Mean Teacher on UCF101-24 dataset on more thresholds. L2: Base model teacher's logits, and Aux-S is sigmoid of auxiliary teacher's output. First row use only single loss between teacher and student. Second row shows the score for combination of consistency between main model and student, and auxiliary network and student.

L2	Aux-S	f-mAP				v-mAP			
		0.1	0.2	0.3	0.5	0.1	0.2	0.3	0.5
✓		90.4	87.6	83.3	66.6	97.6	94.4	87.0	67.4
✓	✓	91.6	89.0	84.8	66.8	97.5	94.6	89.1	68.6

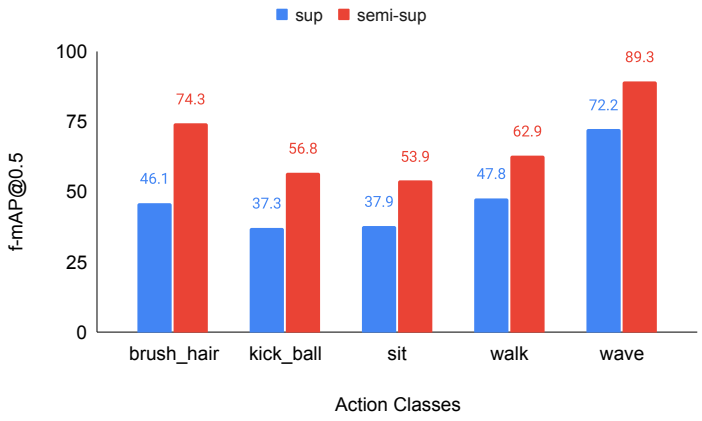

Figure 6: This figure shows the top 5 classes which has the most improvement on f-mAP@0.5 on our proposed semi-supervised approach compared to the supervised counterpart on JHMDB-21 dataset.

mAP at 0.5 thresholds. When two consistency losses are combined, that is consistency loss between base teacher's logits and student's logits plus, and, between, Auxiliary logits/sigmoid and student's logits/sigmoid output, there's still performance gain on all threshold.

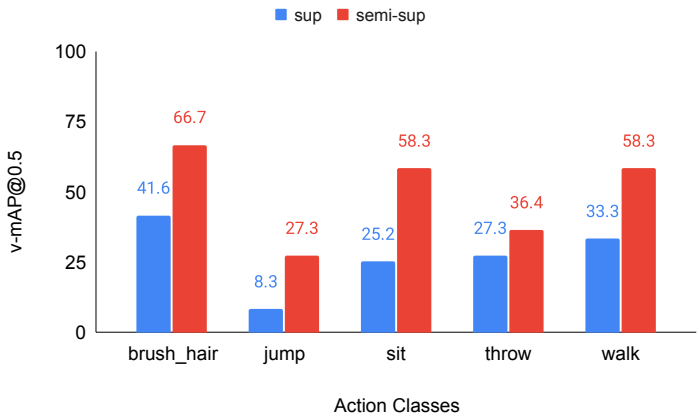

Figure 7: This figure shows the top 5 classes which has the most improvement on v-mAP@0.5 on our proposed semi-supervised approach compared to the supervised counterpart on JHMDB-21 dataset.

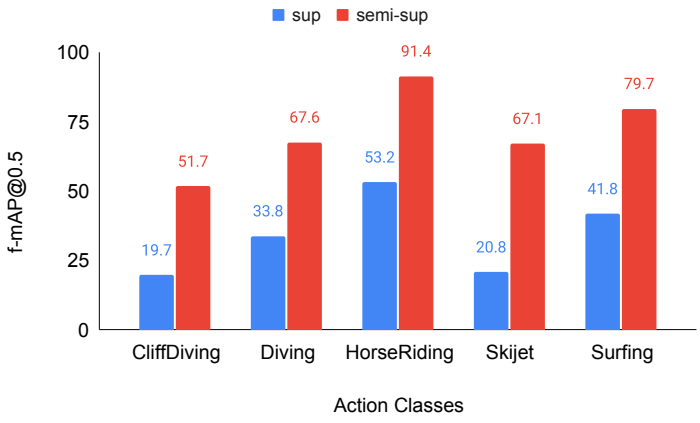

Figure 8: This figure shows the top 5 classes which has the most improvement on f-mAP@0.5 on our proposed semi-supervised approach compared to the supervised counterpart on UCF101-24 dataset.

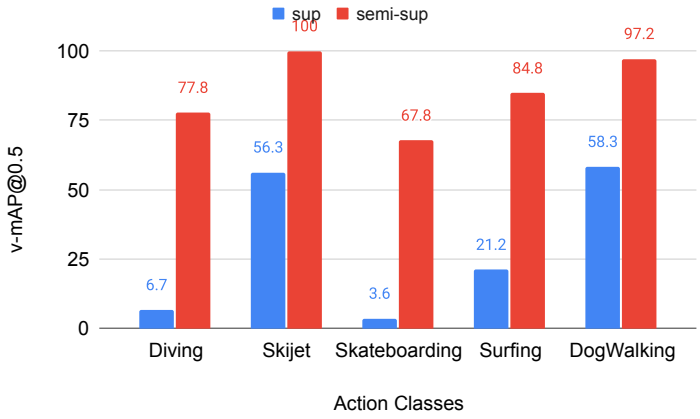

Figure 9: This figure shows the top 5 classes which has the most improvement on v-mAP@0.5 on our proposed semi-supervised approach compared to the supervised counterpart on UCF101-24 dataset.

Comparison on more thresholds Here, we extend ablation table on auxiliary module and compare the *baseline setup* and our proposed approach on UCF101-24. We show results on two more thresholds at 0.1 and 0.3. Table 8 shows that we beat the *baseline setup* at more thresholds.

A.4 Classwise Performance Analysis

In this study, we deep diver into the f-mAP and v-mAP of different classes. Here, we discuss performance at specific threshold of 0.5. From the figures 6 and 7, we show that the classes with the most improvement with our Enhanced Mean teacher approach. The classes with improvement on f-mAP@0.5 and v-mAP@0.5 are `brush_hair, kick_ball, sit, walk, wave` and `brush_hair, jump, sit, throw, walk` respectively. Some of these classes have very fast motion. Most improvement on those classes shows that our approach is more robust to motion changes and the predictions are more temporally coherent.

We extend this analysis to UCF101-24 dataset as well. From Fig. 8 and 9, classes with the most gain are `CliffDiving, Diving, HorseRiding, Skijet, Surfing`, and, `Diving, Skijet, Skateboarding, Surfing, DogWalking`. A major boost in `Diving` and `Surfing` also corroborates our claim that Enhanced Mean Teacher is less susceptible large motion changes and also small objects.

A.5 Data Augmentation Details

In our work, we have different set of augmentations for weak and strong augmented view. For weakly augmented, we only do Random Horizontal Flip, whereas, for strong, we use Color Jitter (Hue, Brightness, Saturation, Contrast), Grayscale and Gaussian Blur. In Table 9, we show how we select random parameters for each of them. We have not search for best hyperparameter settings for data augmentation in our work.

Table 9: Details about selection of random parameters for spatial augmentations.

Strong Augmentations			
Type	Probability	Random Value	Explanation
Contrast	0.7	0.8	Random uniform selection between [0.6, 1.4)
Hue	0.7	0.05	Random uniform selection between [-0.1, 0.1)
Brightness	0.7	0.9	Random uniform selection between [0.6, 1.4)
Saturation	0.7	0.7	Random uniform selection between [0.6, 1.4)
Grayscale	0.6	-	-
Gaussian Blur	0.5	σ_x=0.1, σ_y=2.0	Kernel size=(3, 3)
Weak + Strong Augmentation			
Horizontal Flip	0.5	-	-