Stable Mean Teacher for Semi-supervised Video Action Detection

Akash Kumar, Sirshapan Mitra, Yogesh Singh Rawat

Center for Research in Computer Vision, University of Central Florida
{akash.kumar, sirshapan.mitra, yogesh}@ucf.edu

Abstract

In this work, we focus on semi-supervised learning for video action detection. Video action detection requires spatio-temporal localization in addition to classification and a limited amount of labels makes the model prone to unreliable predictions. We present ***Stable Mean Teacher***, a simple end-to-end student-teacher based framework that benefits from ***improved*** and ***temporally consistent*** pseudo labels. It relies on a novel ***ErrOr Recovery (EoR)*** module which learns from students' mistakes on labeled samples and transfers this to the teacher to improve pseudo-labels for unlabeled samples. Moreover, existing spatio-temporal losses does not take temporal coherency into account and are prone to temporal inconsistencies. To overcome this, we present ***Difference of Pixels (DoP)***, a simple and novel constraint focused on temporal consistency which leads to coherent temporal detections. We evaluate our approach on four different spatio-temporal detection benchmarks, UCF101-24, JHMDB21, AVA, and Youtube-VOS. Our approach outperforms the supervised baselines for action detection by an average margin of **23.5%** on UCF101-24, **16%** on JHMDB21, and, **3.3%** on AVA. Using merely 10% and 20% of data, it provides a competitive performance compared to the supervised baseline trained on 100% annotations on UCF101-24 and JHMDB21 respectively. We further evaluate its effectiveness on AVA for scaling to large-scale datasets and Youtube-VOS for video object segmentation demonstrating its ***generalization capability*** to other tasks in the video domain.

Code —
https://github.com/AKASH2907/stable-mean-teacher
Project Page — https://akash2907.github.io/smt_webpage

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 requires annotations on each video frame, which can be cost and time intensive. In this work, we focus on semi-supervised learning (SSL) to develop label efficient method for video action detection.

Semi-supervised learning (SSL) is an active research area with two prominent approaches: iterative proxy-label (Rizve et al. 2020) and consistency based methods (Tarvainen and Valpola 2017). Iterative proxy-label methods, although effective, are not suitable for the video domain due to their lengthy training cycles. On the other hand, consistency-based approaches offer end-to-end solutions, requiring only a single pass through the dataset for training. While most of the existing research in this area has focused on image classification (Rasmus et al. 2015; Tarvainen and Valpola 2017; Sajjadi, Javanmardi, and Tasdizen 2016; Laine and Aila 2017) and object detection (Xu et al. 2021; Chen et al. 2022; Liu, Ma, and Kira 2022), limited efforts have been made in the video domain with works only focusing on classification (Jing et al. 2021; Xu et al. 2022; Kumar et al. 2023). We also observe that Mean Teacher (Tarvainen and Valpola 2017) based approaches have demonstrated superior performance among consistency-based methods. Building upon the success of student-teacher learning in the image domain, we extend it to the video domain for spatio-temporal detection tasks.

Video action detection, in contrast to classification and object detection, poses additional challenges for semi-supervised learning. It is a complex task that combines both classification and spatio-temporal localization which suffers performance degradation under limited availability of labels. Moreover, the detections have to be temporally coherent in addition to spatial correctness. Therefore, it is challenging to generate high-quality spatio-temporal pseudo-labels for videos. To overcome these challenges, we propose ***Stable Mean Teacher***, a simple end-to-end framework. It is an adaptation of Mean Teacher where we study both *classification* and *spatio-temporal consistencies* to effectively utilize the pseudo-labels generated for unlabeled videos.

Stable Mean Teacher consists of a novel ***ErrOr Recovery (EoR)*** module which learns from the student's mistakes on labeled samples and transfers this learning to the teacher for improving the spatio-temporal pseudo-labels generated on unlabeled set. EoR improves pseudo labels but ignores temporal coherency which is important for action detection. To overcome this, we introduce ***Difference of Pixels (DoP)***, a simple and novel constraint that focuses on temporal coherence and helps in generating consistent spatio-temporal pseudo-labels from unlabeled samples.

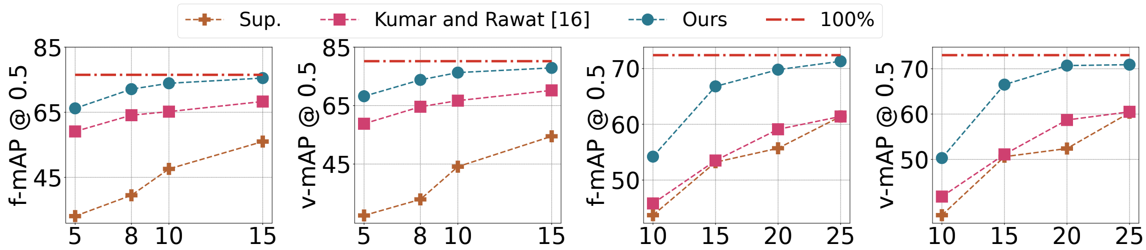

Figure 1: **Performance overview:** Stable Mean Teacher provides comparable performance with 10% (**UCF101-24**; left two plots) and 20% (**JHMDB-21**; right two plots) labels when compared with fully supervised approach which is trained on 100% annotations. It consistently outperforms existing state-of-the-art (2022) and supervised baseline on both f-mAP and v-mAP with a good margin on both UCF101-24 and JHMDB-21 at all different percentages of labeled set. *x-axis* shows the annotation percentage in each plot.

In summary, we make the following contributions:

- We propose ***Stable Mean Teacher***, a simple end-to-end approach for semi-supervised video action detection.
- We propose a novel ***ErrOr Recovery (EoR)*** module, which learns from the student's mistakes and helps the teacher in providing a better supervisory signal under limited labeled samples.
- We propose ***Difference of Pixels (DoP)***, a simple and novel constraint that focuses on temporal consistencies and leads to coherent spatio-temporal predictions.

We perform a comprehensive evaluation on three different action detection benchmarks. Our study demonstrates significant improvement over supervised baselines, consistently outperforming the state-of-the-art approach for action detection (Figure 1). We also demonstrate the generalization capability of our approach to video object segmentation.

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 are based on image-based object detectors such as RCNN (Ren et al. 2015) and DETR (Carion et al. 2020), where detection at frame-level is used for video-level activity classification (Yang et al. 2019; Hou, Chen, and Shah 2017; Yang, Gao, and Nevatia 2017; Dave et al. 2022; Zhao et al. 2022; Chen et al. 2023; Ntinou, Sanchez, and Tzimiropoulos 2024). Most approaches involve two stages, where localization is performed using a region proposal network which is classified into activities in the second stage (Gkioxari et al. 2018; Yang et al. 2019; Hou, Chen, and Shah 2017; Yang, Gao, and Nevatia 2017). Recently, some encoder-decoder based approaches have been developed on CNN (Duarte, Rawat, and Shah 2018) and transformer-based (Zhao et al. 2022; Wu et al. 2023; Chen et al. 2023; Ntinou, Sanchez, and Tzimiropoulos 2024) backbones which simplify the two-stage video action detection process. However, transformer-based backbones are complex and heavy, involving multiple modules. In a recent work (Kumar and Rawat 2022; Singh et al. 2024), the authors further simplify VideoCapsuleNet (Duarte, Rawat, and Shah 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 in weakly-supervised learning attempts to overcome the high labeling cost for action detection (2020; 2020; 2018; 2020; 2018; 2017). These approaches require either video-level annotations or annotations only on a few frames. However, they rely on external detectors (Ren et al. 2015; Liu et al. 2016; Carion et al. 2020) 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 have shown great promise in label efficient learning. Most of the efforts are focused on classification tasks (Tarvainen and Valpola 2017; Ke et al. 2019) where sample level annotation is required, such as object recognition (Liu, Ma, and Kira 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 efficient as the learning is performed in a single step in contrast to several iterations in iterative pseudo-labeling (Rizve et al. 2020). Mean teacher (Tarvainen and Valpola 2017) is a strong consistency-based approach where the pseudo-labels generated by the teacher are used to train a student in both image classification (Tarvainen and Valpola 2017) as well as object detection (Xu et al. 2021; Tang et al. 2021; feng Zhou et al. 2021; Chen et al. 2022; Liu, Ma, and Kira 2022; Liu et al. 2021, 2022). In (Pham et al. 2021), the authors proposed to utilize feedback from students for teachers based on meta-learning which requires two-step training with additional computation cost.

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 (Jing et al. 2021; Singh et al. 2021; Xiao et al. 2022; Xu et al. 2022) and temporal action

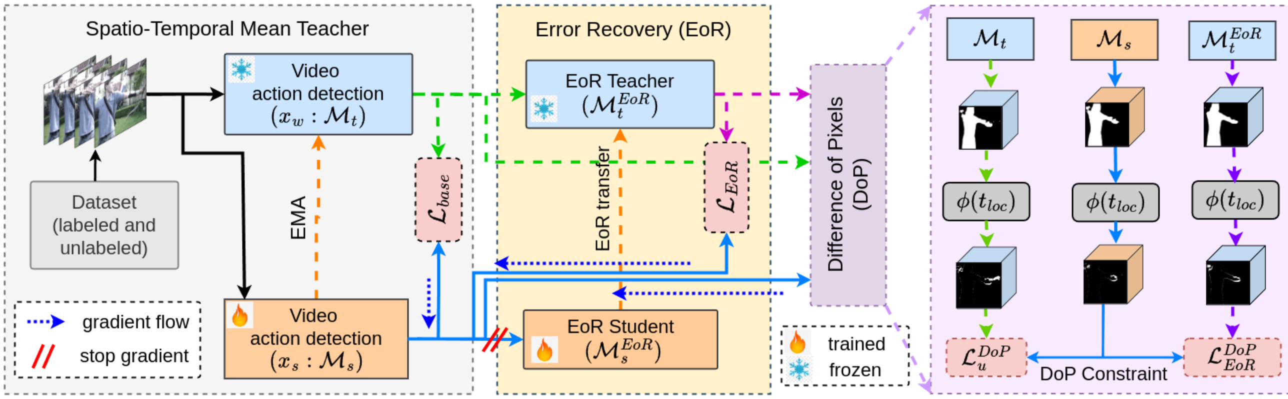

Figure 2: **Overview of Stable Mean Teacher.** The two key components to improve the quality of spatio-temporal pseudo label: 1) *Error Recovery*: refines the spatial action boundary, 2) *DoP constraint*: induces *temporal coherency* on predicted spatio-temporal pseudo labels.

localization (Ji, Cao, and Niebles 2019; Wang et al. 2021; Nag et al. 2022) where per frame dense spatio-temporal annotations 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 and Rawat 2022) has been explored for semi-supervised video action detection. Different from this, we propose a Mean Teacher based approach adapted for video activity detection task which achieves better performance.

Methodology

Problem formulation Given a set of labeled samples $X_L : \{x_i, y_i, f_i\}_{i=0}^{i=N_l}$ and an unlabeled subset $X_U : \{x_i\}_{i=0}^{i=N_u}$, where x is a video and y and f corresponds to class label and frame level annotation with N_l labeled and N_u unlabeled samples. The labeled videos are annotated with a ground-truth class and frame-level spatio-temporal localization denoted as y_t and f_t respectively. Our goal is to train an action detection model (M) using both labeled and unlabeled data.
Overview An overview of the proposed approach is illustrated in Figure 2. As shown in this Figure, Stable Mean Teacher follows a student-teacher approach adapted for video action detection task where the teacher model generates pseudo-labels using weak augmentations for the student who learns from these pseudo-labels on strongly augmented samples. Each video sample (x_i) is augmented to generate two views: strong (x_s) and a weak (x_w). We use the same action detection model M as a teacher $(\mathcal{M}_t)$ and as a student $(\mathcal{M}_s)$. Each of these models has 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. To generate a better and more confident spatio-temporal pseudo-label, the teacher learns from the student's mistakes on labeled samples to improve its pseudo-labels with the help of an *Error Recovery (EoR)* module which is trained jointly. We pass t_{loc} and s_{loc} to Error Recovery module, $(\mathcal{M}_t^{EoR})$ and $(\mathcal{M}_s^{EoR})$, which generates refined localization maps, t_{loc}^{EoR} and s_{loc}^{EoR} respectively. To further induce temporal coherency in the predicted spatio-temporal pseudo label, we apply *Difference of Pixels (DoP)* constraint for temporal refinement on the pseudo label for the student.
Background We use Mean Teacher (2017), a student-teacher training scheme as our baseline approach. In Table 1, we show that baseline Mean teacher works, however, it only exploits classification consistency, whereas video action detection task requires optimizing both classification and spatio-temporal localization task simultaneously.

To address this issue, we *adapt* (2017) for action detection to formulate our base model with capability to generate spatio-temporal pseudo-labels required for this task. Similar to Mean Teacher (2017), we use the teacher's prediction as a pseudo-label for the student model which attends to a strong perturbed version of the video. The 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, $\theta_{teacher} = \beta\theta_{teacher} + (1-\beta)\theta_{student}$. This base setup is trained using both classification and spatio-temporal loss and is defined as $\mathcal{L}_{base} = \mathcal{L}_{base}^{cls} + \mathcal{L}_{base}^{loc}$. where, $\mathcal{L}_{base}^{cls}$ represents classification loss and $\mathcal{L}_{base}^{loc}$ represents spatio-temporal localization loss. Moving forward, we refer STMT as *the base model* in our work.

Stable Mean Teacher

The performance of the base model relies on the quality of the pseudo-labels generated by $(\mathcal{M}_t)$. However, with limited labels, since the model focuses on two tasks: classification and localization simultaneously, it relies on the samples available per class. This limits the generalization capability of the model $(\mathcal{M}_t)$ to generate high-quality pseudo-labels. To address this issue, we propose an Error Recovery module to improve the localization in class-agnostic learning.

Error Recovery (EoR) The Error Recovery module $(\mathcal{M}^{EoR})$ focuses on correcting mistakes of the student model in spatio-temporal localization. These mistakes are

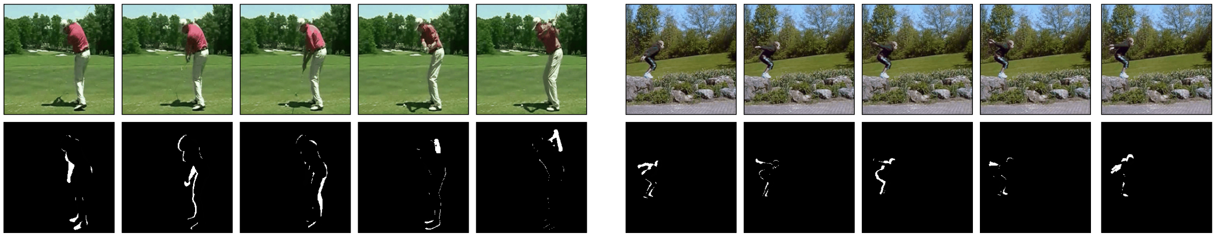

Figure 3: **Visualization of Difference of Pixels (DoP).** First row shows the RGB frames, second row shows the pixel difference map of ground truth along the temporal dimension. We show two scenarios: *Left:* Static: constant background; actor in motion, and *Right:* Dynamic: changing background; actor in motion. Temporal difference emphasizes the variation of boundary pixels between consecutive frames.

approximated by an EoR module which attempts to recover these in class-agnostic learning. This is advantageous since model solely focuses on localization task disregarding specific action class. This in turns enriches the model's ability to localize actors accurately which potentially generates a better pseudo label for the student's model ($\mathcal{M}_s$). The base student model first tries to localize the actor and this is passed as input to the EoR module. The EoR module only focus on refining the localization without worrying about the type of activity. Therefore it can be trained in a class-agnostic manner. EoR module is first trained to recover the student's spatio-temporal mistakes on labeled samples with strong augmentations. Once trained, it is used to recover the mistakes on unlabeled samples with weak augmentations to improve the pseudo labels generated by the teacher. This in turn improves pseudo labels for the student to learn from unlabeled samples. The model parameters (θ_t^{EoR}) for $\mathcal{M}_t^{EoR}$ are updated via EMA of $\mathcal{M}_s^{EoR}$ parameters (θ_s^{EoR}) with the same decay rate β. The update is defined as, $\theta_t^{EoR} = \beta\theta_t^{EoR} + (1-\beta)\theta_s^{EoR}$. The Error Recovery module $\mathcal{M}_s^{EoR}$ does not use any pre-trained weights and is jointly trained with the base model on labeled samples in end-to-end learning. The student's prediction will be more distorted than the teacher's due to strong augmentations. This will help $\mathcal{M}_t^{EoR}$ to provide a more confident pseudo label on a weakly augmented sample. The loss ($\mathcal{L}_{EoR}$) is calculated between student's base model ($\mathcal{M}_s$) output and the refined pseudo label by $\mathcal{M}_t^{EoR}$, $\mathcal{L}_{EoR} = MSE(\mathcal{M}_t^{EoR}(t_{loc}), s_{loc})$, where MSE is Mean-Squared-Error.

Difference of Pixels (DoP) The Error Recovery module enhances the spatial localization of pseudo labels. However, in the context of videos, predictions need to maintain *consistency* over time. To ensure this temporal coherency across frames, we introduce a novel training constraint named Difference of Pixels (DoP). This approach is motivated by the limitations of conventional loss functions that primarily emphasize frame or pixel accuracy, often neglecting temporal coherency. DoP bridges this gap by focusing on pixel movement within videos (Figure 3), and optimizes the accuracy of pixel difference across frames with ($\mathcal{L}_{DoP}$),

$$\begin{aligned}\mathcal{L}_{DoP} = \mathcal{L}_u^{DoP} + \mathcal{L}_{EoR}^{DoP} &= MSE(\phi(t_{loc}), \phi(s_{loc})) \\ &+ MSE(\phi(t_{loc}^{EoR}), \phi(s_{loc})).\end{aligned} \tag{1}$$

where, ϕ denotes temporal difference,

$$\phi(x^{f'}) = x_{loc}^{f+1} - x_{loc}^{f} \tag{2}$$

and, x_{loc}^{f} means localization map at frame f. This strategy enforces stronger *temporal coherency* within spatio-temporal predictions and enhances the quality of pseudo-labels produced by the networks.

Role of EoR and DoP: The base model provides a rough estimate of activity area. EoR recognizes fine-grained errors in spatial boundaries and helps as enhanced class-agnostic supervision to improve student's ($\mathcal{M}_s$) spatio-temporal localization (Figure 4 left). However, EoR loss focuses on spatio-temporal localization without any temporal coherency which is important for action detection. This is where DoP plays a role and helps in the temporal coherency of localization across frames. DoP constraint helps to generate a smooth flow of localization along temporal dimension as it enforces consistency on *displacement of pixels*.

Gradient flow: The base model and Error Recovery module are trained jointly but the gradients from the Error Recovery module are not used to update the base model. If the gradients are allowed to update the base model, then it will also impact the prediction of the base model (discussed in the ablation study). This will be the same as adding more parameters to the model, which is not our goal. Our objective on the other hand is to learn from the mistakes of the base model and not to improve it. This also ensures that the improvement of pseudo-labels is not dependent on the input video and is class agnostic. The Error Recovery module will only have access to the prediction of the base model without any knowledge of the input video. This helps in learning a transformation that generalizes well to unlabeled samples.

Augmentations: We study both spatial and temporal augmentations to generate weak and strong views. First, temporal and then spatial augmentations are applied to the input video. Augmenting in this sequence makes the process *computationally efficient* as for spatial augmentation we only perform augmentation of required frames instead of augmenting all the video frames. The weak augmentation includes only horizontal flipping whereas strong augmentation includes color jitter, gaussian blur, and grayscale.

Learning Objectives The objective function of Stable Mean Teacher has two parts: supervised ($\mathcal{L}_s$) and unsuper-

Method	Backbone	UCF101-24 Annot.	f-mAP	v-mAP		JHMDB21 Annot.	f-mAP	v-mAP	
Fully-Supervised		%	0.5	0.2	0.5	%	0.5	0.2	0.5
TACNet (Song et al. 2019)†	RN-50		72.1	77.5	52.9		65.5	74.1	73.4
MOC (Li et al. 2020)	DLA-34		78.0	82.8	53.8		70.8	77.3	70.2
ACAR-Net (Pan et al. 2021)	SF-R50		84.3	-	-		77.9	-	80.1
VideoCapsuleNet (Duarte, Rawat, and Shah 2018)	I3D		78.6	97.1	80.3		64.6	95.1	-
YOWO (Köpüklü, Wei, and Rigoll 2019)	ResNext-101		80.4	75.8	48.8		74.4	85.7	58.1
TubeR (Zhao et al. 2022)	I3D		83.2	83.3	58.4		-	87.4	82.3
STMixer (Wu et al. 2023)	SF-R101NL		83.7	-	-		86.7	-	-
EVAD (Chen et al. 2023)	ViT-B		85.1	-	-		90.2	-	-
BMVIT (Ntinou, Sanchez, and Tzimiropoulos 2024)	ViT-B		90.7	-	-		88.4	-	-
Weakly-Supervised									
PSAL (Mettes and Snoek 2018)	RN-50		-	41.8	-		-	-	-
Cheron *et al.* (2018)	RN-50		-	43.9	17.7		-	-	-
GuessWA (Escorcia et al. 2020)	IRv2		45.8	19.3	-		-	-	-
UAWS (Arnab et al. 2020)	RN-50		-	61.7	35.0		-	-	-
GLNet (Zhang et al. 2020)	I3D		30.4	45.5	17.3		65.9	77.3	50.8
Semi-Supervised									
MixMatch (Berthelot et al. 2019)††	I3D	10%	10.3	54.7	4.9	30%	7.5	46.2	5.8
Pseudo-label (Lee et al. 2013)	I3D	10%	59.3	89.9	58.3	20%	55.3	87.6	52.0
ISD (Jeong et al. 2021)	I3D	10%	60.2	91.3	64.0	20%	57.8	90.2	57.0
E2E-SSL (Kumar and Rawat 2022)	I3D	10%	65.2	91.8	66.7	20%	59.1	93.2	58.7
Baseline Mean Teacher (Tarvainen and Valpola 2017)	I3D	10%	67.3	92.7	70.5	20%	56.3	88.8	52.8
Stable Mean Teacher (Ours)	I3D	10%	**73.9**	**95.8**	**76.3**	20%	**69.8**	**98.8**	**70.7**
Supervised baseline	I3D	10%	53.5	77.2	49.7	20%	55.7	93.9	52.4

Table 1: **Comparison with previous state-of-the art approaches** on fully, weakly and semi-supervised learning on UCF101-24 and JHMDB21. † shows approach using Optical flow as second modality. The last row shows the score for supervised labeled subset, that is 10% for UCF101-24 and 20% for JHMDB21. Best score on each metric is underlined. RN-50, SF-R50/101 and IRv2 is ResNet-50, SlowFast-R50/101, and, InceptionResNetV2 respectively. (2019)†† suffers from cold-start problem below 30% on JHMDB21.

vised ($\mathcal{L}_u$). Supervised loss has classification ($\mathcal{L}_s^{cls}$) and localization ($\mathcal{L}_s^{loc}$) and follows losses from (2018). Unsupervised loss comprises of three parts: 1) Base model (STMT) loss ($\mathcal{L}_{base}$) which incorporates both classification ($\mathcal{L}_{base}^{cls}$) and localization ($\mathcal{L}_{base}^{loc}$), 2) Error Recovery loss ($\mathcal{L}_{EoR}$), and 3) DoP loss ($\mathcal{L}_{DoP}$). We calculate the supervised loss on the labeled subset of student's predictions ($\mathcal{L}_s^{cls}, \mathcal{L}_s^{loc}$) and student's Error Recovery module predictions, and unsupervised loss on labeled plus unlabeled subset. We have two unsupervised losses: **a)** ***classification consistency***: it minimizes the difference between teachers' prediction t_{cls} and student's prediction s_{cls} using Jenson-Shennon Divergence (JSD), and, **b)** ***localization consistency***: computes pixel-level difference on each frame between teacher (t_{loc}, t_{loc}^{EoR}) and student s_{loc} localization maps using MSE. Finally, the overall loss for Stable Mean Teacher is defined as,

$$\mathcal{L} = \mathcal{L}_s + \lambda \mathcal{L}_u = \mathcal{L}_s + \lambda(\mathcal{L}_{base} + \mathcal{L}_{EoR} + \mathcal{L}_{dop}) \quad (3)$$

where λ is a weight parameter for unsupervised losses.

Experiments

Datasets: We use four benchmark datasets to perform our experiments; UCF101-24 (2012), JHMDB21 (2013), and AVA v2.2 (AVA)(2018) for action detection, and YouTube-VOS (2018c) to show generalization on video segmentation (VOS). UCF101-24 consists of 3207 videos, split into 2284 for training and 923 for testing. JHMDB21 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 JHMDB21 it's 21. AVA consists of 299 videos, each lasting 15 minutes. The dataset is divided into 211K clips for training and 57K clips for validation. Annotations are provided at 1 FPS with bounding boxes and labels. We report our performance on 60 action classes following standard evaluation protocols (Pan et al. 2021; Zhao et al. 2021). The distribution of the number of training, validation, and evaluation videos on YouTube-VOS-2019 (2018a) is 3471, 507, and 541 respectively.

Labeled and unlabeled setup: The labeled and unlabeled subset for UCF101-24 and Youtube-VOS is divided in the ratio of 10:90 and for JHMDB21 it's 20:80. For the AVA dataset, we use 50% of the dataset for semi-sup setup. We utilize a 10:40 split between labeled to unlabeled ratio. We perform our experiments with 10%/20% labeled set for UCF101-24/JHMDB21 instead of 20%/30% (as in (Kumar and Rawat 2022)), as the performance is already close to fully supervised training when using 20%/30% in these datasets. These scores are shown in the supplementary.

Implementation details We train the model for 50 epochs with a batch size of 8 where the number of samples from both labeled and unlabeled subsets are the same. The value

Method	Backbone	Pretraining	K	FPS	$\mathcal{A}$	mAP	GFLOPs
Non real-time spatio-temporal action detector							
WOO (2021)	SF-R101	K600	8	-	100%	28.3	252
SE-STAD (2023)	SF-R101	K400	8	-	100%	29.3	165
TubeR (2021)	CSN-152	IG-65M	32	3	100%	29.7	120
STMixer (2023)	CSN-152	IG-65M	32	3	100%	31.7	120
EVAD (2023)	ViT-B	K400	16	-	100%	32.3	243
BMViT (2024)	ViT-B	K400, MAE	16	-	100%	31.4	350
Real-time spatio-temporal action detector							
YOWO (2019)	ResNext-101	K400	16	35	100%	17.9	44
YOWOv2-N (2023)	Shufflev2-1.0x	K400	16	40	100%	12.6	1.3
Ours(YOWOv2-N)	Shufflev2-1.0x	K400	16	40	10%	8.5	1.3
Sup. baseline	Shufflev2-1.0x	K400	16	40	10%	5.2	1.3

Table 2: **Evaluation on AVA dataset.** K is the length of input video clip. $\mathcal{A}$ denotes annotation percent. mAP denotes f-mAP@0.5. YOWO2-N denotes nano version.

			UCF101-24		JHMDB-21	
$\mathcal{L}_{base}$	$\mathcal{L}_{EoR}$	$\mathcal{L}_{DoP}$	v@0.5	f@0.5	v@0.5	f@0.5
✓			74.5	72.2	62.0	61.8
✓	✓		75.9	73.1	68.1	68.3
✓		✓	75.4	72.6	64.5	62.9
✓	✓	✓	76.3	73.9	70.7	69.8

Table 3: **Ablations:** Effectiveness of *Error Recovery module* and *Difference of Pixels*. $\mathcal{L}_{base}$: training without $\mathcal{L}_{EoR}$ and $\mathcal{L}_{DoP}$. v@0.5: v-mAP@0.5, f@0.5:f-mAP@0.5.

of β for EMA parameters update is set to 0.99 which follows prior works (2022; 2021). The value of λ for the unsupervised loss is set to 0.1 determined empirically.

Base model and Error Recovery model architecture: We use VideoCapsuleNet (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 (2022) and enables a fair comparison. For the Error Recovery module, we use a 3D UNet (2015) architecture with a depth of 4 layers with 16, 32, 64, and 128 channels respectively.

Evaluation metrics: For spatio-temporal video localization, we evaluate the proposed approach similar to previous works (2016; 2015) on frame metric average precision (f-mAP) and video metric average precision (v-mAP). f-mAP is computed by summing over all the frames with an IoU greater than a threshold per class. Similarly, for v-mAP 3D IoU is utilized instead of frame-level IoU. We show results at 0.2 and 0.5 thresholds in the main paper with other thresholds results provided in the supplementary. For VOS, we show results on Jaccard (J) and Boundary (F) metrics.

Results

Comparison with semi-supervised: In Table 1, *semi-supervised* results highlight the limitations of image-based and object detection approaches, with (2019) struggling to generalize with limited video data. Our method outperforms the pseudo-label approach across all thresholds and surpasses the semi-supervised object detection method by 12-14% on UCF101-24 and 9-12% on JHMDB21 using 10%

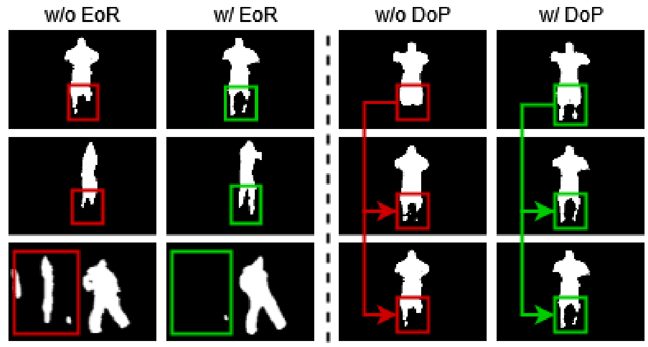

Figure 4: **Qualitative analysis for EoR and DoP:** Left side illustrates the effectiveness of *Error Recovery* module on multiple samples, with improvement in action boundary precision and it also helps in suppressing background noise. On the right hand, we demonstrate how *DoP constraint* induces temporal coherency in predictions for sequence of video frames.

less data. Compared to a parallel video action detection method (Kumar and Rawat 2022), we achieve gains of 8.7% (f-mAP@0.5) and 9.6% (v-mAP@0.5) on UCF101-24, and 5.4% and 7.2%, respectively, on JHMDB21, also with 10% less data. Against our base model, our approach shows consistent improvements of up to 1.8% on UCF101-24 and 8.7% on JHMDB21 across various metrics.

Comparison with supervised and weakly-supervised: In the *supervised* scenario, our method, with only 10% labeled data, surpasses all 2D-based approaches in v-mAP (Table 1) and demonstrates competitive performance against 3D-based methods, outperforming several. Notably, unlike most 2D methods that incorporate optical flow, our architecture relies solely on a single modality. In the *weakly-supervised* setting, our approach achieves state-of-the-art results on both datasets. On UCF101-24 (Table 1), we outperform the best method (Arnab et al. 2020; Zhang et al. 2020) by approximately 35% at the 0.5 threshold. On JHMDB21, we achieve an absolute improvement of 3.9% on f-mAP@0.5 and 19.9% on v-mAP@0.5.

Scaling to large-scale dataset: To evaluate the scalability of our approach, we perform experiments on the AVA dataset a large-scale dataset. AVA is not spatio-temporal as against UCF101-24 with only sparse frame-level annotations available. In Table 2, we run on a real-time spatio-temporal approach and show our approach improves on supervised by 3.3% on YOWOv2-N with only 10% labeled data.

Ablation Studies

Impact of Error Recovery module: The significance of the EoR module is evident in Table 3, with a performance boost of 6% on JHMDB-21 and 1% on UCF101-24 compared to the baseline. This improvement stems from refined pseudo labels (t^{EoR}_{loc}), enabling better localization of activity regions for the student (s_{loc}). The larger gain on JHMDB-21 reflects its higher complexity due to pixel-level ground truth compared to bounding box-level annotations in UCF101-24. Ex-

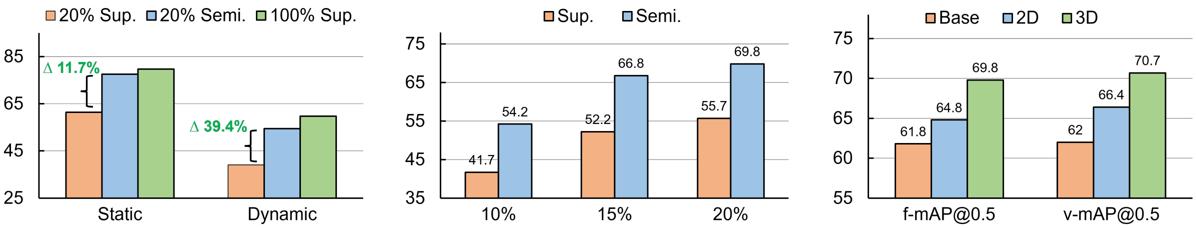

Figure 5: ***Analyzing Stable Mean Teacher:*** *(Left)* ***Static vs dynamic scenes:*** Dynamic scenes are more challenging than static scenes, however, the relative boost in performance for dynamic is 27.7% more than in the case of static scene scenarios. Δ denotes relative change at v-mAP@0.5. *(Middle)* ***Annotation percent:*** Moving towards right to left on the x-axis, the gain in performance (f-mAP@0.5) increases. It indicates the approach is more effective in low label regime. *(Right)* ***Error Recovery architectures:*** The performance of 3D Error Recovery architecture outperforms the 2D based architecture.

tending the analysis to v-mAP@0.5:0.95 reveals improvements of 2.5% (f-mAP) and 3.9% (v-mAP) in mean IoU, highlighting the module's ability to enhance finer boundary localization.

Effect of DoP constraint: The Difference of Pixels (DoP) constraint enhances temporally coherent pseudo labels, as shown in Table 3. It improves performance by 0.5-0.8% with the base model (STMT) and 1-3% when combined with the Error Recovery module on UCF101-24 and JHMDB-21. The DoP constraint has a greater impact on v-mAP than f-mAP, highlighting its role in improving *temporal coherency*. Additionally, it achieves a 1% increase in mean IoU for f-mAP@0.5:0.95 and 2% for v-mAP@0.5:0.95 when used independently, demonstrating its effectiveness.

Qualitative analysis: In Figure 4, we analyze the effectiveness of each component qualitatively. DoP makes the predictions coherent across time and the Error Recovery module helps to generate better fine-grained predictions. We show more qualitative analysis in the supplementary.

Discussion and Analysis

Static vs dynamic scenes: Activities can be classified into static (constant background) and dynamic (changing background) scenes. Dynamic actions are more challenging due to simultaneous changes in actor and background. Our approach demonstrates a significant improvement of **11.7%** for static and **39.4%** for dynamic actions at v-mAP@0.5 (Figure 5 (left)), highlighting its robustness in dynamic.

Effectiveness of approach in low-label regime: As shown in Figure 5 (middle), our approach achieves a 40% higher gain at 15% labeled data compared to 20%, demonstrating its efficacy in leveraging unlabeled data in low-label scenes.

Importance of gradient stopping: Error Recovery module utilizes grayscale maps to localize the actor whereas main model uses RGB frames to classify and localize the action. Since the task of Error Recovery module is to be class-agnostic, if the gradient of Error Recovery module flows back into the main network, it degrades the quality of pseudo labels generated by the main model. This further degrades the refinement procedure by Error Recovery module. We observe performance degrades by approximately 3% without stopping the gradient flow.

Additional parameters doesn't help: Adding EoR module parameters to the base model improves performance slightly (e.g., f-mAP@0.5 of 62.3 on JHMDB-21 with 20% data), but lags behind our proposed approach by $\tilde{7}$%. This shows that merely increasing parameters offers limited benefit without leveraging effective design principles (Table 1).

Method	Annot.	Avg	J_S	J_U	F_S	F_U
Xu (2018b) †	10%	10.1	11.6	10.1	9.6	9.2
Kumar *et al.* (2022)	10%	36.8	43.1	31.4	40.8	31.8
Ours	5%	38.2	45.3	32.0	43.2	32.2
Ours	10%	41.3	48.2	35.0	46.7	35.4
Xu (2018b)	100%	47.9	55.7	39.6	55.2	41.3

Table 4: **Generalization capability:** Performance comparison on Youtube-VOS. J_s and J_u are Jaccard on seen and unseen categories; F_s and F_u are boundary metric on seen and unseen categories. † shows 10% supervised results.

Generalization to video object segmentation (VOS) We demonstrate the generalization capability of Stable Mean Teacher on video object segmentation (VOS) using the YouTube-VOS dataset (Table 4). Our method outperforms the supervised baseline by an average margin of 31% and achieves a 4-6% improvement over the semi-supervised approach (2022) across all metrics. Notably, even with only 5% labeled data, our approach surpasses (2022), highlighting its effectiveness in low-label settings.

Conclusion

We propose *Stable Mean Teacher*, a novel student-teacher approach for semi-supervised action detection. Stable Mean Teacher relies on a novel *Error Recovery* module which learns from student's mistakes and transfer that knowledge to the teacher for generating better pseudo labels for the student. It also benefits from *Difference of Pixels*, a simple constraint which enforces temporal coherency in the spatio-temporal predictions. We demonstrate the effectiveness of Stable Mean Teacher on three action detection datasets with extensive set of experiments. Furthermore, we also show its performance on VOS task validating its *generalization capability* to other dense prediction tasks in videos.

References

Arnab, A.; Sun, C.; Nagrani, A.; and Schmid, C. 2020. Uncertainty-Aware Weakly Supervised Action Detection from Untrimmed Videos. *ArXiv*, abs/2007.10703.

Berthelot, D.; Carlini, N.; Goodfellow, I.; Papernot, N.; Oliver, A.; and Raffel, C. A. 2019. MixMatch: A Holistic Approach to Semi-Supervised Learning. In Wallach, H.; Larochelle, H.; Beygelzimer, A.; d'Alché-Buc, F.; Fox, E.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; and Zagoruyko, S. 2020. End-to-End Object Detection with Transformers. *ArXiv*, abs/2005.12872.

Chen, B.; Li, P.; Chen, X.; Wang, B.; Zhang, L.; and Hua, X.-S. 2022. Dense Learning Based Semi-Supervised Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 4815–4824.

Chen, L.; Tong, Z.; Song, Y.; Wu, G.; and Wang, L. 2023. Efficient video action detection with token dropout and context refinement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10388–10399.

Chen, S.; Sun, P.; Xie, E.; Ge, C.; Wu, J.; Ma, L.; Shen, J.; and Luo, P. 2021. Watch only once: An end-to-end video action detection framework. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 8178–8187.

Chéron, G.; Alayrac, J.-B.; Laptev, I.; and Schmid, C. 2018. A flexible model for training action localization with varying levels of supervision. In Bengio, S.; Wallach, H.; Larochelle, H.; Grauman, K.; Cesa-Bianchi, N.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Dave, I.; Scheffer, Z.; Kumar, A.; Shiraz, S.; Rawat, Y. S.; and Shah, M. 2022. GabriellaV2: Towards Better Generalization in Surveillance Videos for Action Detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops*, 122–132.

Duarte, K.; Rawat, Y. S.; and Shah, M. 2018. Videocapsulenet: A simplified network for action detection. *Advances in Neural Information Processing Systems*.

Escorcia, V.; Dao, C. D.; Jain, M.; Ghanem, B.; and Snoek, C. G. M. 2020. Guess Where? Actor-Supervision for Spatiotemporal Action Localization. *Comput. Vis. Image Underst.*, 192: 102886.

feng Zhou, Q.; Yu, C.; Wang, Z.; Qian, Q.; and Li, H. 2021. Instant-Teaching: An End-to-End Semi-Supervised Object Detection Framework. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 4079–4088.

Finn, C.; Goodfellow, I.; and Levine, S. 2016. Unsupervised learning for physical interaction through video prediction. *arXiv preprint arXiv:1605.07157*.

Gkioxari, G.; Girshick, R.; Dollár, P.; and He, K. 2018. Detecting and recognizing human-object interactions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 8359–8367.

Gu, C.; Sun, C.; Ross, D. A.; Vondrick, C.; Pantofaru, C.; Li, Y.; Vijayanarasimhan, S.; Toderici, G.; Ricco, S.; Sukthankar, R.; Schmid, C.; and Malik, J. 2018. AVA: A Video Dataset of Spatio-Temporally Localized Atomic Visual Actions. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6047–6056.

Hou, R.; Chen, C.; and Shah, M. 2017. Tube convolutional neural network (T-CNN) for action detection in videos. In *IEEE International Conference on Computer Vision*.

Jeong, J.; Verma, V.; Hyun, M.; Kannala, J.; and Kwak, N. 2021. Interpolation-based Semi-supervised Learning for Object Detection. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 11597–11606.

Jhuang, H.; Gall, J.; Zuffi, S.; Schmid, C.; and Black, M. J. 2013. Towards understanding action recognition. In *International Conf. on Computer Vision (ICCV)*, 3192–3199.

Ji, J.; Cao, K.; and Niebles, J. C. 2019. Learning Temporal Action Proposals With Fewer Labels. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 7072–7081.

Jing, L.; Parag, T.; Wu, Z.; Tian, Y.; and Wang, H. 2021. VideoSSL: Semi-Supervised Learning for Video Classification. *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 1109–1118.

Ke, Z.; Wang, D.; Yan, Q.; Ren, J. S. J.; and Lau, R. W. H. 2019. Dual Student: Breaking the Limits of the Teacher in Semi-Supervised Learning. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 6727–6735.

Köpüklü, O.; Wei, X.; and Rigoll, G. 2019. You only watch once: A unified cnn architecture for real-time spatiotemporal action localization. *arXiv preprint arXiv:1911.06644*.

Kumar, A.; Kumar, A.; Vineet, V.; and Rawat, Y. S. 2023. Benchmarking self-supervised video representation learning. *arXiv preprint arXiv:2306.06010*.

Kumar, A.; and Rawat, Y. S. 2022. End-to-End Semi-Supervised Learning for Video Action Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Laine, S.; and Aila, T. 2017. Temporal Ensembling for Semi-Supervised Learning. *ArXiv*, abs/1610.02242.

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

Li, Y.; Wang, Z.; Wang, L.; and Wu, G. 2020. Actions as Moving Points. In *arXiv preprint arXiv:2001.04608*.

Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S. E.; Fu, C.-Y.; and Berg, A. 2016. SSD: Single Shot MultiBox Detector. In *ECCV*.

Liu, Y.-C.; Ma, C.-Y.; Dai, X.; Tian, J.; Vajda, P.; He, Z.; and Kira, Z. 2022. Open-Set Semi-Supervised Object Detection. In *European Conference on Computer Vision*.

Liu, Y.-C.; Ma, C.-Y.; He, Z.; Kuo, C.-W.; Chen, K.; Zhang, P.; Wu, B.; Kira, Z.; and Vajda, P. 2021. Unbiased Teacher for Semi-Supervised Object Detection. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Liu, Y.-C.; Ma, C.-Y.; and Kira, Z. 2022. 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)*, 9819–9828.

Mettes, P.; and Snoek, C. G. M. 2018. Pointly-Supervised Action Localization. *International Journal of Computer Vision*, 127: 263–281.

Mettes, P.; Snoek, C. G. M.; and Chang, S.-F. 2017. Localizing Actions from Video Labels and Pseudo-Annotations. *ArXiv*, abs/1707.09143.

Nag, S.; Zhu, X.; Song, Y.-Z.; and Xiang, T. 2022. Semi-Supervised Temporal Action Detection with Proposal-Free Masking. In *European Conference on Computer Vision*.

Ntinou, I.; Sanchez, E.; and Tzimiropoulos, G. 2024. Multiscale vision transformers meet bipartite matching for efficient single-stage action localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18827–18836.

Pan, J.; Chen, S.; Shou, M. Z.; Liu, Y.; Shao, J.; and Li, H. 2021. Actor-context-actor relation network for spatio-temporal action localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 464–474.

Pham, H.; Dai, Z.; Xie, Q.; and Le, Q. V. 2021. Meta pseudo labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11557–11568.

Rasmus, A.; Valpola, H.; Honkala, M.; Berglund, M.; and Raiko, T. 2015. Semi-Supervised Learning with Ladder Network. *ArXiv*, abs/1507.02672.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, 91–99.

Rizve, M. N.; Duarte, K.; Rawat, Y. S.; and Shah, M. 2020. In Defense of Pseudo-Labeling: An Uncertainty-Aware Pseudo-label Selection Framework for Semi-Supervised Learning. In *International Conference on Learning Representations*.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. *ArXiv*, abs/1505.04597.

Sajjadi, M. S. M.; Javanmardi, M.; and Tasdizen, T. 2016. Regularization With Stochastic Transformations and Perturbations for Deep Semi-Supervised Learning. In *NIPS*.

Singh, A.; Chakraborty, O.; Varshney, A.; Panda, R.; Feris, R. S.; Saenko, K.; and Das, A. 2021. Semi-Supervised Action Recognition with Temporal Contrastive Learning. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10384–10394.

Singh, A.; Rana, A. J.; Kumar, A.; Vyas, S.; and Rawat, Y. S. 2024. Semi-supervised Active Learning for Video Action Detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(5): 4891–4899.

Sohn, K.; Berthelot, D.; Carlini, N.; Zhang, Z.; Zhang, H.; Raffel, C. A.; Cubuk, E. D.; Kurakin, A.; and Li, C.-L. 2020. FixMatch: Simplifying Semi-Supervised Learning with Consistency and Confidence. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M. F.; and Lin, H., eds., *Advances in Neural Information Processing Systems*, volume 33, 596–608. Curran Associates, Inc.

Song, L.; Zhang, S.; Yu, G.; and Sun, H. 2019. TACNet: Transition-Aware Context Network for Spatio-Temporal Action Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Soomro, K.; Zamir, A.; and Shah, M. 2012. UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild. *ArXiv*, abs/1212.0402.

Sui, L.; Zhang, C.-L.; Gu, L.; and Han, F. 2023. A simple and efficient pipeline to build an end-to-end spatial-temporal action detector. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 5999–6008.

Tang, Y.; Chen, W.; Luo, Y.; and Zhang, Y. 2021. Humble Teachers Teach Better Students for Semi-Supervised Object Detection. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 3131–3140.

Tarvainen, A.; and Valpola, H. 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *NIPS*.

Wang, X.; Zhang, S.; Qing, Z.; Shao, Y.; Gao, C.; and Sang, N. 2021. Self-Supervised Learning for Semi-Supervised Temporal Action Proposal. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 1905–1914.

Weinzaepfel, P.; Harchaoui, Z.; and Schmid, C. 2015. Learning to track for spatio-temporal action localization. In *Proceedings of the IEEE international conference on computer vision*, 3164–3172.

Wu, T.; Cao, M.; Gao, Z.; Wu, G.; and Wang, L. 2023. STMixer: A One-Stage Sparse Action Detector. *ArXiv*, abs/2303.15879.

Xiao, J.; Jing, L.; Zhang, L.; He, J.; She, Q.; Zhou, Z.; Yuille, A.; and Li, Y. 2022. Learning From Temporal Gradient for Semi-Supervised Action Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 3252–3262.

Xu, M.; Zhang, Z.; Hu, H.; Wang, J.; Wang, L.; Wei, F.; Bai, X.; and Liu, Z. 2021. End-to-End Semi-Supervised Object Detection with Soft Teacher. *ArXiv*, abs/2106.09018.

Xu, N.; Yang, L.; Fan, Y.; Yang, J.; Yue, D.; Liang, Y.; Price, B.; Cohen, S.; and Huang, T. 2018a. Youtube-vos: Sequence-to-sequence video object segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 585–601.

Xu, N.; Yang, L.; Fan, Y.; Yang, J.; Yue, D.; Liang, Y.; Price, B. L.; Cohen, S. D.; and Huang, T. S. 2018b. YouTube-VOS: Sequence-to-Sequence Video Object Segmentation. *ArXiv*, abs/1809.00461.

Xu, N.; Yang, L.; Fan, Y.; Yue, D.; Liang, Y.; Yang, J.; and Huang, T. S. 2018c. YouTube-VOS: A Large-Scale Video Object Segmentation Benchmark. *ArXiv*, abs/1809.03327.

Xu, Y.; Wei, F.; Sun, X.; Yang, C.; Shen, Y.; Dai, B.; Zhou, B.; and Lin, S. 2022. Cross-Model Pseudo-Labeling for Semi-Supervised Action Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2959–2968.

Yang, J.; and Dai, K. 2023. Yowov2: A stronger yet efficient multi-level detection framework for real-time spatio-temporal action detection. *arXiv preprint arXiv:2302.06848*.

Yang, X.; Yang, X.; Liu, M.-Y.; Xiao, F.; Davis, L. S.; and Kautz, J. 2019. Step: Spatio-temporal progressive learning for video action detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 264–272.

Yang, Z.; Gao, J.; and Nevatia, R. 2017. Spatio-temporal action detection with cascade proposal and location anticipation. In *Proceedings of the British Machine Vision Conference (BMVC)*.

Zhang, S.; Song, L.; Gao, C.; and Sang, N. 2020. GLNet: Global Local Network for Weakly Supervised Action Localization. *IEEE Transactions on Multimedia*, 22(10): 2610–2622.

Zhao, J.; Li, X.; Liu, C.; Bing, S.; Chen, H.; Snoek, C. G.; and Tighe, J. 2021. Tuber: Tube-transformer for action detection. *arXiv preprint arXiv:2104.00969*.

Zhao, J.; Zhang, Y.; Li, X.; Chen, H.; Shuai, B.; Xu, M.; Liu, C.; Kundu, K.; Xiong, Y.; Modolo, D.; Marsic, I.; Snoek, C. G. M.; and Tighe, J. 2022. TubeR: Tubelet Transformer for Video Action Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 13598–13607.