Semi-supervised Active Learning for Video Action Detection

Ayush Singh[1], Aayush J Rana[2], Akash Kumar[2], Shruti Vyas[2], Yogesh Singh Rawat[2]

[1] IIT (ISM) Dhanbad
[2]University of Central Florida
ayush.s.18je0204@cse.iitism.ac.in, {aayushjungbahadur.rana, akash.kumar, shruti, yogesh}@ucf.edu

Abstract

In this work, we focus on label efficient learning for video action detection. We develop a novel semi-supervised active learning approach which utilizes both labeled as well as unlabeled data along with informative sample selection for action detection. Video action detection requires spatio-temporal localization along with classification, which poses several challenges for both active learning (*informative sample selection*) as well as semi-supervised learning (*pseudo label generation*). First, we propose ***NoiseAug***, a simple augmentation strategy which effectively selects informative samples for video action detection. Next, we propose ***fft-attention***, a novel technique based on high-pass filtering which enables effective utilization of pseudo label for SSL in video action detection by emphasizing on relevant activity region within a video. We evaluate the proposed approach on three different benchmark datasets, UCF-101-24, JHMDB-21, and Youtube-VOS. First, we demonstrate its effectiveness on video action detection where the proposed approach outperforms prior works in semi-supervised and weakly-supervised learning along with several baseline approaches in both UCF101-24 and JHMDB-21. Next, we also show its effectiveness on Youtube-VOS for video object segmentation demonstrating its *generalization capability* for other dense prediction tasks in videos.

Introduction

Video understanding is an essential task for security, automation, and robotics (Rizve et al. 2021a) as video data enables information extraction for detection (Hou, Sukthankar, and Shah 2017; Yang et al. 2019), recognition (Hara, Kataoka, and Satoh 2018; Kumar et al. 2023), tracking (Vondrick et al. 2018), and scene understanding (Lei et al. 2018). Video understanding in general requires a large amount of labeled data to train an effective model. Collecting such data for dense prediction tasks such as action detection is even more challenging as it requires spatio-temporal annotations on every frame of the video. In this work, we focus on label efficient learning for video action detection.

Existing works on label efficient learning for video action detection have primarily focused on weakly supervised learning, semi-supervised learning, and active learning. Weakly supervised methods often underperform compared to supervised methods, often requiring externally trained object detector to address the detection aspect of action detection (Chéron et al. 2018; Weinzaepfel, Martin, and Schmid 2016). Recently, semi-supervised learning (SSL) (Kumar and Rawat 2022) and active learning methods (Rana and Rawat 2022) have shown promising performance. However, they also have their own limitations. SSL relies on randomly selected sub-samples, which can result in non-informative sample selection and suboptimal models. On the other hand, active learning aims to address this issue by selecting only informative samples for training. Nevertheless, it suffers from a cold-start problem, which makes it challenging to train a good model with limited labels which are initially available.

We propose a unified approach for video action detection by bridging the gap between semi-supervised learning (SSL) and active learning. We address the challenges of the cold-start problem in active learning by using an SSL technique to train a reliable initial model. Similarly, we resolve the need for informative training sample for SSL using optimized selection via active learning. Our student-teacher-based SSL framework benefits from active learning's informative sample selection, offering the advantages of both SSL and active learning for improved video action detection.

Video action detection needs to perform both spatio-temporal localization and classification. Solving this task using limited labels pose two distinct challenges; 1) determining the informativeness of samples, and 2) generating high-quality pseudo-labels. To address the first challenge, we propose ***NoiseAug***, a simple and novel augmentation strategy designed to estimate sample informativeness in video action detection. Model-driven AL used in existing works for sample selection often perturbs the model via regularization (Gal and Ghahramani 2016; Heilbron et al. 2018; Aghdam et al. 2019), which limits the extent of perturbation since too much perturbation will affect the network negatively. Therefore, we propose data-driven AL and use varying degree of data augmentations while maintaining video integrity. By isolating the type of augmentation seen by the model during training and selection step, we can focus on the relevant regions and reduce bias from training samples which is a common problem in active learning with limited labeled set (Pardo et al. 2021; Gal and Ghahramani 2016; Aghdam et al. 2019).

Active learning enables cost-effective labeling by selecting informative samples and subsequently helps improve model

performance, benefiting scenarios where data annotation is expensive or time-consuming. However, active learning also suffers from cold start when the initial model is trained using very limited labeled data (Houlsby, Hernández-Lobato, and Ghahramani 2014; Prabhu, Dognin, and Singh 2019). This leads to the second challenge, generating high-quality pseudo labels for semi-supervised learning (SSL). To tackle this, we introduce ***fft-attention***, a novel technique based on high-pass filtering that emphasizes on activity regions and their edges within a video. Fft-attention improves the prediction of activity regions and enhances the quality of pseudo labels for SSL in video action detection.

In summary, we make the following contributions,

- We propose a novel semi-supervised active learning framework for video action detection which provides label efficient solution. To the best of our knowledge, this is the first work focusing on this problem.
- We propose *NoiseAug*, a novel noise based augmentation for video data perturbation which helps in informative sample selection.
- We propose *fft-attention*, a novel high pass filter which helps in estimating action and non-action regions for effective pseudo label generation in semi-supervised learning.

We evaluate the proposed approach on two different video action detection benchmarks and compare with several baselines, including existing semi-supervised and weakly-supervised approaches outperforming all prior works. We also demonstrate its effectiveness on Youtube-VOS for video object segmentation showing the generalization capability to other dense prediction tasks in videos.

Related Work

Video action detection Video action detection is a complex and challenging task (Yang et al. 2019; Pan et al. 2021; Sun et al. 2018; Li et al. 2020), where the goal is to perform spatio-temporal action detection in a given video. Most prior works use fully-supervised approach where all the samples are annotated spatio-temporally. The recent works have rapidly improved performance due to improved networks (Hara, Kataoka, and Satoh 2018; Szegedy et al. 2017; He et al. 2016) and increased data availability (Soomro, Zamir, and Shah 2012; Jhuang et al. 2013). However, getting large dataset with spatio-temporal annotation is costly. Weakly-supervised learning is an alternative which uses partially annotated data over the entire dataset to train action detection models (Mettes, Snoek, and Chang 2017; Mettes and Snoek 2019; Chéron et al. 2018; Escorcia et al. 2020; Arnab et al. 2020; Zhang et al. 2019). These methods rely on external pre-trained object detector (Ren et al. 2015) and often fall behind significantly on performance compared to the fully-supervised methods.

Semi-supervised learning Semi-supervised learning utilizes both labeled and unlabelled samples for training (Sohn et al. 2020; Berthelot et al. 2019b,a; Tarvainen and Valpola 2017; Oliver et al. 2018; Miyato et al. 2018; Yang et al. 2021; Schiappa, Rawat, and Shah 2022), generally using regularization (Rasmus et al. 2015; Tarvainen and Valpola 2017; Sajjadi, Javanmardi, and Tasdizen 2016; Laine and Aila 2017) or pseudo-labeling (Li et al. 2021; Lee 2013; Rizve et al. 2021b) methods for classification (Berthelot et al. 2019b,a; Rizve et al. 2021b) and detection (Kumar and Rawat 2022; Rosenberg, Hebert, and Schneiderman 2005). For video action detection, using pseudo-labeling approach for semi-supervised learning becomes costly and difficult with limited labels (Zhang, Zhao, and Wang 2022; Schiappa, Rawat, and Shah 2022). The pseudo-labeling approach also assumes that a pre-trained object detector or region proposal is available (Ren et al. 2020; Zhang, Zhao, and Wang 2022). A better option is to use consistency regularization which relies on the model itself to moderate the learning (Berthelot et al. 2019b; Kumar and Rawat 2022; Sajjadi, Javanmardi, and Tasdizen 2016; Tarvainen and Valpola 2017; Jeong et al. 2019), generally using perturbations in input or model. We use a combination of consistency regularization via strong and weak augmentation of labeled and unlabeled samples using mean-teacher setup (Tarvainen and Valpola 2017). One of the challenges in consistency based SSL for video action detection is having too much noise from background regions of a video, as seen in (Kumar and Rawat 2022). To this end, we focus on consistency of relevant action regions while suppressing large backgrounds present in videos with our fft-attention based filter approach.

Active learning Labeling a large video dataset for action detection task is expensive as a lot of frames must be annotated spatio-temporally for each video. Active learning (Pardo et al. 2021) enables selecting samples for annotation by estimating the usefulness of each sample to the underlying task. It is used to iteratively select a subset of data for annotation on various tasks as image classification (Wang et al. 2016), image object detection (Aghdam et al. 2019; Pardo et al. 2021) and video temporal localization (Heilbron et al. 2018) with only few studies done for video action detection (Rana and Rawat 2022). The sample selection in AL is done using uncertainty (Liu et al. 2019), entropy (Aghdam et al. 2019), core-set selection (Sener and Savarese 2017) or mutual-information (Kirsch, Van Amersfoort, and Gal 2019). While there have been some prior works that combine AL and SSL for object detection and segmentation task (Elezi et al. 2022; Rangnekar, Kanan, and Hoffman 2023), we are the first to propose a unified SSL active learning framework for spatio-temporal video action detection to best of our knowledge. We use data perturbation via noise based augmentation to get the model's uncertainty, using that as an estimate of usefulness for each sample in our AL strategy.

Proposed Method

We introduce a semi-supervised active learning approach for video action detection, where active learning is employed to select samples during each training iteration, and SSL is utilized for model training. Our proposed NoiseAug strategy enhances sample selection in active learning. We use a student teacher based approach for SSL where we extend Mean Teacher (Tarvainen and Valpola 2017) for video action detection by incorporating fft-attention based filtering for effective pseudo-label training on unlabeled data. Mean

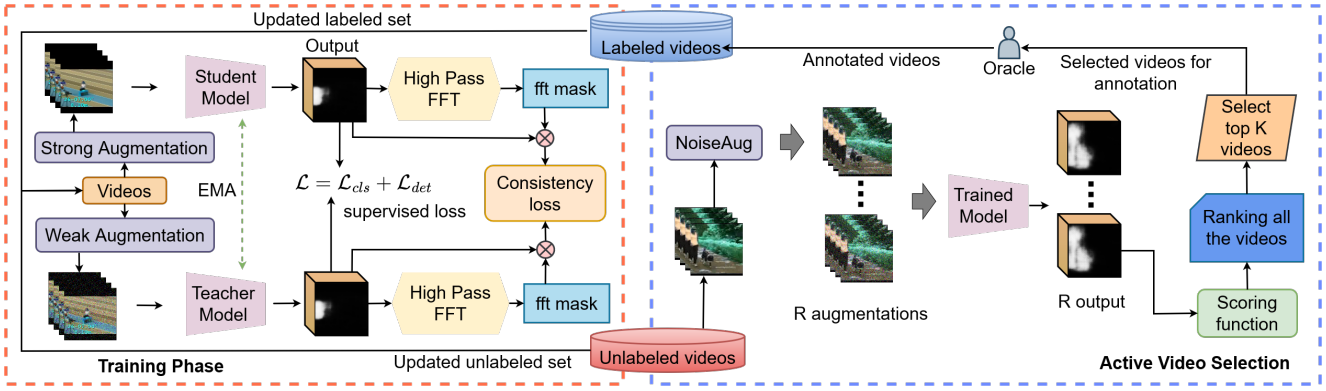

Figure 1: *Overview of our proposed approach:* During the training phase, we take the labeled and unlabeled data at equal ratio to train the model together. We apply strong and weak augmentations to all input samples. All detection output is passed through our FFT filter to get a weight mask, which is used to compute the final consistency loss between teacher and student model output. During the active video selection phase, we take trained student model and pass R variants with NoiseAug and score the sample. We select top K videos for further annotation and cycle back to the next training phase.

Teacher approach relies on strong and weak augmentations to generate pseudo-labels that can be used to train augmented data. However, video data has a lot of unrelated background region that adds extra noise to this process. We propose fft-attention that helps focus on relevant activity region in videos and improves pseudo-label based training. An overview of the proposed approach is shown in Figure 1.

Problem formulation: Given a dataset $\mathcal{D}$ with N unlabeled videos $\mathcal{X}^U = \{x_1^U, ..., x_N^U\}$ and Q labeled videos $\mathcal{X}^L = \{x_1^L, ..., x_Q^L\}$, we train an action detection model M for θ weights using $\mathcal{D} = \{\mathcal{X}^L, \mathcal{X}^U\}$ using a semi-supervised approach. We assume that the initial labeled set is much smaller than the unlabeled set such that $Q << N$. Once we have a trained model M, we use it for selecting more samples to annotate in the AL step. For each sample in $\mathcal{X}^U$, we prepare $\mathcal{V}$ variations by using proposed noise augmentation and use the model M to estimate each sample's utility value. Once we do that for all $\mathcal{X}^U$, we rank them and select samples within budget for further annotation which makes a new labeled set $\mathcal{X}_2^L$ for training new model. First, we describe the AL approach for sample selection from $\mathcal{X}^U$ for further annotation in next subsection.

Active Learning for Sample Selection

We use a trained model M to get high utility samples from $\mathcal{X}^U$ for further annotation. This is an important step to increase labeled data under limited budget such that model training improves significantly over random sample selection approach. The model M estimates prediction uncertainty of detection among all variations of a sample which in turn gives a sample level score on its usefulness. One of the key challenge to get this uncertainty is to avoid training bias of the network. Since the labeled training set $\mathcal{X}^L$ is often small initially ($Q << N$), network can be easily over-fitted to have bias as well as become robust to the training augmentations. Our aim is to provide simple augmentation that is unique and not seen in the training time, which encourages the network to better estimate uncertainty of unlabeled samples.

NoiseAug: The goal is to find samples that maximize the model's performance for action detection. We follow prior AL works (Gal and Ghahramani 2016; Jain and Kapoor 2009) and use uncertainty as a measure for scoring and selecting samples. Prior works use Monte Carlo method (Gal and Ghahramani 2016) or multi-layer output (Aghdam et al. 2019) to compute per pixel informativeness (uncertainty, entropy). In contrast, we leverage on noise invariance to evaluate a sample's utility. We measure the variance in uncertainty of predictions from the model for different noise infused variants of the same sample. We generate multiple variations $\mathcal{V}_\phi$ of the same sample v using noise augmentation given as,

$$\mathcal{V}_\phi^i = v^{[T\times H\times W\times C]} \odot \mathcal{N}^i(0,1)^{[T\times H\times W\times C]} \quad (1)$$

where, v is a sample with dimensions $[T \times H \times W \times C]$ and $\mathcal{N}^i(0,1)$ is a Gaussian distribution of the same dimension used as noise. We use their Hadamard product to get the final augmented variation $\mathcal{V}_\phi^i$. We repeat this process R times to get a set of noise augmented variations $\mathcal{V}_\phi = \{\mathcal{V}_\phi^1, \mathcal{V}_\phi^2, ..., \mathcal{V}_\phi^R\}$.

Sample selection: We use NoiseAug to measure uncertainty using model's confidence in each of the noise variants. We utilize the temporal aspect of a video for comparing pixel level uncertainty. We average the pixel values for neighboring frames as another form of regularization for uncertainty measure. We define the temporal average function as,

$$Avg(f_i[p]) = \frac{1}{T}\sum_{t=i-T/2}^{i+T/2} M(f_t[p];\theta) \quad (2)$$

where, for a pixel p in a given frame f_i at the i^{th} temporal location, we average the prediction of model M with θ weights for same pixel location p over neighboring T frames. Then we compute the uncertainty score for sample v as,

$$\mathcal{U}(v) = \sum_{i=1}^{F}\sum_{p=(1,1)}^{P_{[x,y]}} -log(Avg(f_i[p])) \quad (3)$$

where, F is total frames with $[x, y]$ size, M is the model with θ weights that gives prediction for the pixel p in i^{th} frame.

The sample's informativeness as a whole is a reflection of how consistent the model is to all of its noise variants. Ideally, the model M is trained to be noise invariant. Any sample with high variance in uncertainty indicates that the network is not doing well for that sample when noise is introduced. Thus, we use Equation 3 on each of the R variations to get the variance in uncertainty given as,

$$S = Var(\mathcal{U}[\mathcal{V}_\phi^1), \mathcal{U}(\mathcal{V}_\phi^2), ..., \mathcal{U}(\mathcal{V}_\phi^R)] \tag{4}$$

where, S is the informativeness score, $Var()$ gives the variance for uncertainty of all augmented variants $\mathcal{V}$ of sample v. For each AL round, we pick top **K** samples for annotation such that our labeled videos becomes $\mathcal{X}_2^L = \{x_1^L, ..., x_{Q+\mathbf{K}}^L\}$ and our unlabeled set becomes $\mathcal{X}_2^U = \{x_1^U, ..., x_{N-\mathbf{K}}^U\}$. We use this new set of data $\mathcal{D}_2 = \{\mathcal{X}_2^L, \mathcal{X}_2^U\}$ to train a new action detection model in the next round and continue this until we exhaust our total annotation budget.

Semi-Supervised Learning

To leverage the entire training dataset $\mathcal{D} = \{\mathcal{X}^L, \mathcal{X}^U\}$ with both labeled set $\mathcal{X}^L$ and unlabeled set $\mathcal{X}^U$, we use SSL approach that uses mean-teacher based regularization to train using the unlabeled data. Mean-teacher trains unlabeled data using pseudo-labels predicted from a teacher model on an augmented variation of the data. We use supervised loss on weak and strong augmented variations for the labeled data. For the unlabeled data, we use the prediction from teacher network M_t to generate pseudo-labels that can be used to train the student network M_s using supervised loss. Along with that, we also follow prior work (Kumar and Rawat 2022) to use mean squared error based consistency loss for training M_s. For a given video v with F frames, we apply different degrees of augmentation following mean-teacher setup to obtain v'. The consistency loss is then computed as,

$$\mathcal{L}_{cons}^{MSE} = \frac{1}{F}\sum_{i=1}^{F}\frac{1}{x \cdot y}\sum_{p=(1,1)}^{P_{[x,y]}} \|M_s(f_p^i; \theta_s) - M_t(f_p^{i'}; \theta_t)\|^2{}_2 \tag{5}$$

where, M_t and M_s are teacher and student models with θ_t and θ_s weights respectively that gives spatio-temporal detection for i^{th} frame. We compute MSE value for each pixel p for frame f^i of $[x, y]$ size. This general form of MSE based consistency loss gives equal weight for all P pixels in the frame, which is non-ideal for spatio-temporal detection as we only want to focus on certain action regions in each frame f.

It is preferable to focus on relevant regions without using manually designed heuristics (pre-computed regions (Ren et al. 2015)). We want to reduce model uncertainty for specific areas with lower prediction quality, specifically the edges of an actor. To this end, we propose using a high pass filter that will reduce low frequency areas and give more focus on the high frequency area such as edges. Next, we define the high pass filter we use for selective focus.

FFT based high pass filter In order to separate the low and high frequency areas, we apply a FFT based high pass filter. We are trying to focus more on the edges of the predicted detection regions while suppressing other areas. We assume that the high frequency areas (edges) are harder for the network to learn due to quick changes at the edges in a video. Thus, we identify such regions and give them higher weight than the easier regions during training to increase model's consistency. A FFT high pass filter finds the edges and attenuates lower frequency, keeping the non-edge regions with lower weight. We define the FFT high pass filter function as,

$$HPF(f) = FFT(M(f; \theta)) \tag{6}$$

where, M is the model with weight θ, $FFT()$ is the FFT function that gives the filtered output for a frame f. For a given video v and its augmented variant v' with F frames, the per frame consistency using the FFT filter is,

$$FC(f, f', W) = \frac{1}{x.y}\sum_{p=(1,1)}^{P_{[x,y]}} \|M_s(f_p; \theta_s) - M_t(f_p'; \theta_t)\|_2^2 \cdot W_p \tag{7}$$

where, $FC(f, f', W)$ is the frame-wise consistency function that takes frame f, f' and weight W, all of $[x, y]$ size. We modify the MSE computation from Equation 5 to use pixel-wise weight W on the computed MSE value of pixel p. Then we redefine the overall consistency loss as,

$$\mathcal{L}_{cons}^{HPF} = \frac{1}{F}\sum_{i=1}^{F} FC(f^i, f^{i'}, HPF(f^i)) \tag{8}$$

$$\mathcal{L}_{cons}^{HPF'} = \frac{1}{F}\sum_{i=1}^{F} FC(f^i, f^{i'}, HPF(f^{i'})) \tag{9}$$

where, we compute the per-frame consistency with HPF as weight from both f and f' frames of v and v' videos.

Temporal consistency We use the temporal information of subsequent frames to improve the consistency loss for training. While the HPF based consistency computes spatial consistency for each frame, it does not use temporal consistency information. To enforce temporal consistency, we use temporal average function from Equation 2, which changes the model's output for Equation 6 and 7 from $M(f; \theta) \rightarrow Avg(f)$.

Overall Training Objective

We train the model M_s with $\mathcal{D} = \{\mathcal{X}^L, \mathcal{X}^U\}$ which consists of both labeled and unlabeled data. We use supervised loss on labeled data for classification $\mathcal{L}_{cls}$ and detection $\mathcal{L}_{det}$. For the unlabeled data, we use M_t to get pseudo-label for training M_s in a supervised fashion. For unsupervised M_s training, we use the proposed FFT high pass filter based consistency loss. Our training objective is given as,

$$\mathcal{L}_{cons}^{overall} = \lambda_1 \mathcal{L}_{cons}^{HPF} + \lambda_2 \mathcal{L}_{cons'}^{HPF} \tag{10}$$

$$\mathcal{L} = \mathcal{L}_{cls} + \mathcal{L}_{det} + \mathcal{L}_{cons}^{overall} \tag{11}$$

where, λ_1 and λ_1 are loss weights given to the consistency loss which varies for unlabeled samples following prior SSL works (Sohn et al. 2020; Kumar and Rawat 2022).

Method	Backbone		UCF101-24				JHMDB-21			
			Label	f-mAP	v-mAP		Label	f-mAP	v-mAP	
	2D	3D	%	0.5	0.2	0.5	%	0.5	0.2	0.5
Fully-Supervised										
Kalogeitan *et al.*(Kalogeiton et al. 2017)	✓			69.5	76.5	49.2		65.7	74.2	73.7
Song *et al.*(Song et al. 2019)†	✓			72.1	77.5	52.9		65.5	74.1	73.4
Li *et al.*(Li et al. 2020)	✓			78.0	82.8	53.8		70.8	77.3	70.2
Gu *et al.*(Gu et al. 2018)†		✓		76.3	-	59.9		73.3	-	78.6
Duarte *et al.*(Duarte, Rawat, and Shah 2018)		✓		78.6	97.1	80.3		64.6	95.1	-
Pan *et al.*(Pan et al. 2021)		✓		84.3	-	-		-	-	-
Zhao *et al.* (Zhao et al. 2022)		✓		83.2	83.3	58.4		-	87.4	82.3
Wu *et al.* (Wu et al. 2023)		✓		83.7	-	-		86.7	-	-
Weakly-Supervised										
Mettes *et al.*(Mettes, Snoek, and Chang 2017)	✓			-	37.4	-		-	-	-
Mettes and Snoek (Mettes and Snoek 2019)	✓			-	41.8	-		-	-	-
Cheron *et al.*(Chéron et al. 2018)		✓		-	43.9	17.7		-	-	-
Escorcia *et al.*(Escorcia et al. 2020)		✓		45.8	19.3	-		-	-	-
Arnab *et al.*(Arnab et al. 2020)		✓		-	61.7	35.0		-	-	-
Zhang *et al.*(Zhang et al. 2019)		✓		30.4	45.5	17.3		65.9	77.3	50.8
Semi-Supervised										
MixMatch (Berthelot et al. 2019b)		✓	20%	20.2	60.2	13.8	30%	7.5	46.2	5.8
Psuedo-label (Lee 2013)		✓	20%	64.9	93.0	65.6	30%	57.4	90.1	57.4
Co-SSD (CC)(Jeong et al. 2019)		✓	20%	65.3	93.7	67.5	30%	60.7	94.3	58.5
PI-consistency (Kumar and Rawat 2022)		✓	20%	69.9	95.7	72.1	30%	64.4	95.4	63.5
Ours (M-T SSL)		✓	20%	69.8	94.9	72.2	30%	68.5	98.4	68.0
Ours (M-T SSL + AL)		✓	20%	**72.0**	**96.3**	**74.5**	30%	**70.7**	**98.8**	**71.7**
Supervised baseline		✓	20%	59.8	91.6	59.2	30%	59.4	96.5	60.4

Table 1: *Comparison with existing works*: We compare with existing supervised and weakly supervised works along with the semi-supervised baselines on UCF101- 24 and JHMDB-21. † denotes method using Optical flow.

Experiments

Datasets We conduct our experiments on three video datasets, UCF101-24 (Soomro, Zamir, and Shah 2012) and JHMDB-21 (Jhuang et al. 2013). UCF101-24 consists of 24 classes with a total of 3207 untrimmmed videos with bounding box annotations. JHMDB-21 dataset has 21 classes from a total of 928 videos with pixel-level annotations. Both UCF101-24 and JHMDB-21 are focused on action detection task. We further generalize our approach on YouTube-VOS dataset, a video object segmentation task, which has temporally sparse pixel-wise mask annotation for specific objects. It has 3471 videos for training with 65 object categories.

Evaluation metrics Following prior action detection works (Peng and Schmid 2016) we evaluate the f-mAP and v-mAP scores at different IoU thresholds for UCF101-24 and JHMDB-21. The f-mAP is computed from spatial IoU for each frame per class and averaged for all frames to get precision score at given IoU. Similarly, v-mAP is computed using spatio-temporal IoU for each video per class and averaged. For VOS task, we compute the average IoU and boundary similarity score following (Xu et al. 2018b).

Implementation details: We use the PyTorch to build our models and train them on single 16GB GPU. For action detection, we use VideoCapsuleNet (Duarte, Rawat, and Shah 2018; Kumar and Rawat 2022; Rana and Rawat 2022), with margin-loss for classification and BCE loss for detection. The network input is 8 RGB frames of size 224×224. We use a batch size of 8 for training with equal ratio of labeled and unlabeled sample per batch. We use the Adam optimizer (Kingma and Ba 2014) with a learning rate of $1e-4$. We use EMA update at rate of 0.996. We train UCF101-24 for 80 epochs and JHMDB-51 for 50 epochs. **Hyperparameters:** We use a temporal block of $T = 3$ frames for the temporal average function in Equation 2. The loss weights for the consistency loss are $\lambda_1 = 0.5$ and $\lambda_2 = 0.5$ in Equation 10 and $\lambda_3 = [0.01 \rightarrow 0.1]$ increased over warmup range. Please refer to supplementary for more details. **Active Learning:** We take $R = 8$ different noise added variations for each video v to get the sample informativeness score S in Equation 4. We select 5% and 2% new samples for UCF101-24 and 10% for JHMDB-21 in each AL round. We take temporal average over $T = 3$ frames in Equation 2.

Active Learning Baselines

We compare our proposed AL approach with baselines on both UCF101-24 and JHMDB-21. We use random selection, MC uncertainty (Gal and Ghahramani 2016), MC entropy (Aghdam et al. 2019) as selection baselines to compare with the proposed AL selection. All baselines use same backbone

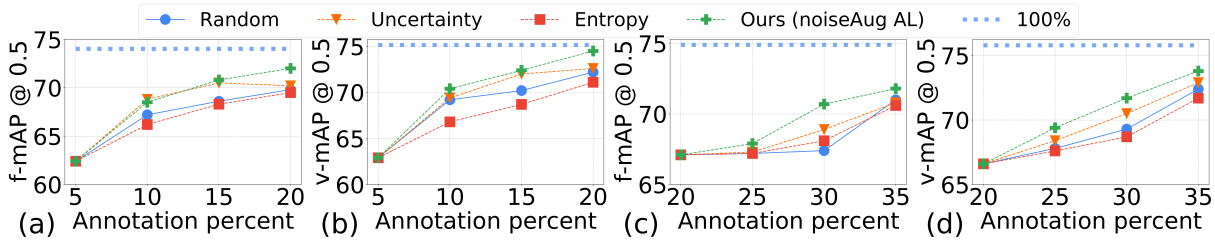

Figure 2: *Analysis on selection criteria*: We compare our proposed AL selection with other selection baselines using the same SSL training setup on UCF101-24 (a-b) and JHMDB-21 (c-d).

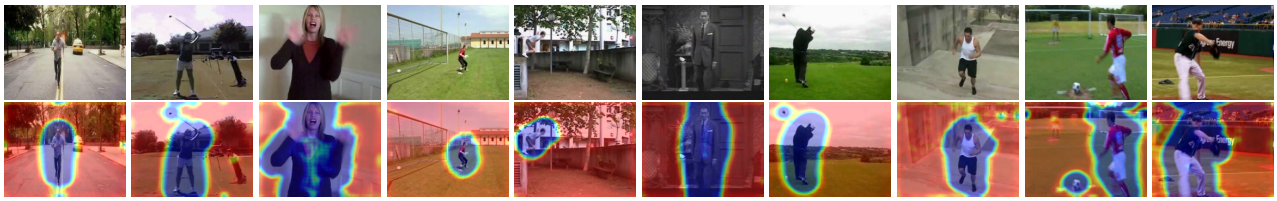

Figure 3: *Qualitative analysis of FFT based high pass filter:* We show the input frames (first row) and corresponding weights (bottom row) using proposed FFT filter. The FFT method gives higher weight towards the edges of detected action regions while suppressing background. Red: low weight, blue: medium weight, green: high weight

.

			UCF101-24						JHMDB-21					
			10%		15%		20%		20%		25%		30%	
C	M-T	FFT	f-mAP	v-mAP	f-mAP	v-mAP	f-mAP	v-mAP	f-mAP	v-mAP	f-mAP	v-mAP	f-mAP	v-mAP
✓			62.4	62.5	64.6	65.3	66.5	68.7	62.6	59.3	63.1	62.9	63.4	64.2
	✓		67.2	68.6	68.4	69.5	69.2	71.9	61.5	63.3	65.6	65.3	66.0	67.4
	✓	✓	68.5	70.4	70.8	72.4	72.0	74.5	67.1	66.6	65.2	67.8	67.4	69.3

Table 2: *Ablations*: We show effectiveness of different components used in SSL training. We evaluate the effect of consistency based SSL (C), mean-teacher (M-T) setup and proposed FFT filter during the training. We report f-mAP and v-mAP @ 0.5.

as ours for fair comparison, with results in Figure 2.

UCF101-24 We begin with 5% labeled data and increment by 5% in every AL cycle as shown in Figure 2(a-b). With more data, we notice that compared to baseline selection methods, our AL method is consistently performing better. We also notice a cold start problem for MC entropy as the model is not performing well for most samples in initial round of 10%, using only model entropy leads to non-optimum sample selection in future rounds. Our AL approach uses noise augmentation to estimate the model uncertainty along with temporal averaging to utilize the temporal consistency expected in videos, which leads to better sample informativeness scores that is more reflective of the model's need.

JHMDB-21 Due to the dataset being smaller with only 660 training videos and the detection task being harder with pixel-wise semantic segmentation, we initialize the training with 20% labeled data (198 videos) for all methods. Each round increases labeled data by 5% videos until we reach 35% data. The quantitative results are shown in figure 2(c-d). Similar with UCF101-24, we see that our AL method consistently outperforms baseline selection methods.

SSL Baselines

We compare the effect of different SSL techniques to show why the proposed mean-teacher setup is optimum for video understanding task. We use mean-teacher SSL and compare results with consistency based SSL in Table 2. We observe that mean-teacher based SSL outperforms consistency based SSL for all dataset, showing that the controlled weight update of teacher using EMA and pseudo-label from teacher using augmented data better regulates unlabeled training.

Comparison With the State-of-the-Art

We compare to prior works using fully, weakly and semi-supervised approach on UCF101-24 and JHMDB-21 in Table 1. Compared to the weakly supervised methods, we perform significantly better for both dataset. The semi-supervised methods are closer in performance with our approach. We show that our method with only mean-teacher (M-T) SSL (no AL selection) performs better than existing SSL methods. The temporal consistency component and focusing on edges using FFT high pass filter enables our model to weight the relevant regions appropriately compared to background regions, which gives our method a competitive edge. Furthermore, when we use the proposed noise augmentation

Method	Data	Avg	J_S	J_U	F_S	F_U
Random	10%	10.1	11.6	10.1	9.6	9.2
PI-consistency	10%	36.8	43.1	31.4	40.8	31.8
Ours	10%	39.3	46.1	33.7	43.9	33.5
Random	20%	34.7	42.8	29.0	38.7	28.3
Ours	20%	41.6	49.6	34.2	47.7	35.6

Table 3: *Generalization capability*: Evaluation on Youtube-VOS dataset. We use same backbone following (Xu et al. 2018a) for our and random method. 10% results for PI-consistency are reported from (Kumar and Rawat 2022).

S/W	noiseAug	5%		10%		15%		20%	
		f	v	f	v	f	v	f	v
✓		62.4	62.9	61.9	62.3	63.6	63.4	64.0	64.2
	✓	62.4	62.9	68.5	70.4	70.8	72.4	72.0	74.5

Table 4: *Effectiveness of NoiseAug*: Comparison of different augmentations used for AL selection for strong/weak augmentation from mean teacher SSL training and proposed noiseAug. [f: f-mAP, v: v-mAP @ 0.5]

based AL to do sample selection and train using SSL, we perform better than prior weakly and semi-supervised methods with $+2.4\%$ v-mAP@0.5 for UCF101-24 and $+8.2\%$ v-mAP@0.5 for JHMDB-21 over prior best score. We also show the generalization on YouTube-VOS in Table 3.

Ablations

Effect of FFT We evaluate the usefulness of FFT filter as weights for putting more emphasis on regions around an action. We train the teacher and student model for action detection without FFT and compare with our baseline in Table 2. We observe a drop in performance when the FFT filter is not used as a weight to compute the consistency loss, showing that the regions selected using FFT have more relevance to action detection. We also see how the FFT filter emphasises relevant regions during training in Figure 3.

Augmentations for AL consistency We use noise augmentation for our AL selection strategy, where we use $R = 8$ noise augmented variants of the video v to compute uncertainty variance. To validate the effectiveness of this augmentation, we use the same strong/weak augmentation setup used for mean-teacher SSL training process (details in suppementary). As we observe in Table 4, using the proposed noise based augmentation provides useful information for sample selection compared to using strong/weak augmentation from SSL step. The network is already trained with strong/weak augmentation, making the predictions more robust for such augmentations in the AL selection step. Varying levels of noise-based augmentation enable the network to encounter different sample variations in AL selection from training, leading to improved uncertainty estimation for new samples.

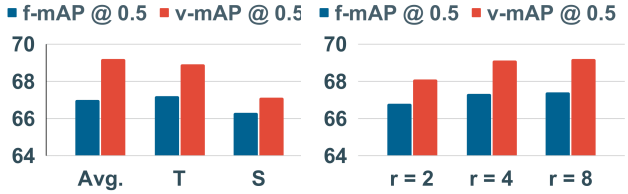

Figure 4: *FFT Analysis:* Left: FFT filter on teacher(**T**)-student(**S**) for SSL training. Right: Effect of radius (**r**) on FFT filter. Both are on UCF101-24 for 10% labeled data.

Discussion and Analysis

FFT as a high pass filter For a given frame f of a video, the model performance $M(f;\theta)$ can be categorized in four main types: High-Confidence and High-Consistency (HCF-HC), High-Confidence and Low-Consistency (HCF-LC), Low-Confidence and High-Consistency (LCF-HC) and Low-Confidence and Low-Consistency (LCF-LC). The simple consistency loss from Equation 5 focuses on regions with lower consistency (either HCF-LC or LCF-LC). The model is not able to predict confidently for LCF-LC samples due to lack of similar supervised training samples.

Ideally we do not want to give high weight for LCF-LC samples in unsupervised setting as model is not able to predict anything for such samples. In contrary, for unsupervised setting we would prefer to have more weights on HCF-LC samples as model is confident but inconsistent. Using a high pass filter enables this, as it gives higher weight for high confidence (HCF) regions and filters out low confidence (LCF) regions. This is demonstrated in Figure 3, where the general consistency loss gives equal weight to large background region and small action region, making it hard for network to learn with noise augmentation. Our FFT based approach gives higher weights on action regions (specifically the edges) which the model can improve on more than background.

Radius for FFT filter Fft-attention relies on the radius of the filter which affects the value for each pixel from FFT. While small radius looks at local window, it has higher sensitivity based on local changes. Conversely, larger radius looks at larger neighborhood but dilates the effect in return. We analyze different radius for FFT filter (Figure 4) and found it to be robust within a range.

Conclusion

We present a unified semi-supervised active learning approach for spatio-temporal video action detection, particularly in scenarios where obtaining labels is costly. We show that using noise as augmentation to compute the informativeness of each sample improves the sample selection for active learning. We also introduce the use of FFT based high pass filter to focus more on relevant activity regions for SSL consistency. Our proposed approach is characterized by its simplicity and can be easily generalized to other dense prediction tasks in videos.

Acknowledgments

This research is based upon work supported in part by the Office of the Director of National Intelligence (Intelligence Advanced Research Projects Activity) via 2022-21102100001 and in part by University of Central Florida seed funding. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, or the US Government. The US Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

References

Aghdam, H. H.; Gonzalez-Garcia, A.; Weijer, J. v. d.; and López, A. M. 2019. Active learning for deep detection neural networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3672–3680.

Arnab, A.; Sun, C.; Nagrani, A.; and Schmid, C. 2020. Uncertainty-aware weakly supervised action detection from untrimmed videos. In *European Conference on Computer Vision*, 751–768. Springer.

Berthelot, D.; Carlini, N.; Cubuk, E. D.; Kurakin, A.; Sohn, K.; Zhang, H.; and Raffel, C. 2019a. ReMixMatch: Semi-Supervised Learning with Distribution Alignment and Augmentation Anchoring. *ArXiv*, abs/1911.09785.

Berthelot, D.; Carlini, N.; Goodfellow, I.; Papernot, N.; Oliver, A.; and Raffel, C. A. 2019b. 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.

Chéron, G.; Alayrac, J.-B.; Laptev, I.; and Schmid, C. 2018. A flexible model for training action localization with varying levels of supervision. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, 950–961.

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

Elezi, I.; Yu, Z.; Anandkumar, A.; Leal-Taixe, L.; and Alvarez, J. M. 2022. Not all labels are equal: Rationalizing the labeling costs for training object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14492–14501.

Escorcia, V.; Dao, C. D.; Jain, M.; Ghanem, B.; and Snoek, C. 2020. Guess where? Actor-supervision for spatiotemporal action localization. *Computer Vision and Image Understanding*, 192: 102886.

Gal, Y.; and Ghahramani, Z. 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning*, 1050–1059. PMLR.

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.

Hara, K.; Kataoka, H.; and Satoh, Y. 2018. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 6546–6555.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Heilbron, F. C.; Lee, J.-Y.; Jin, H.; and Ghanem, B. 2018. What do i annotate next? an empirical study of active learning for action localization. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 199–216.

Hou, R.; Sukthankar, R.; and Shah, M. 2017. Real-Time Temporal Action Localization in Untrimmed Videos by Sub-Action Discovery. In *BMVC*, volume 2, 7.

Houlsby, N.; Hernández-Lobato, J. M.; and Ghahramani, Z. 2014. Cold-start active learning with robust ordinal matrix factorization. In *International conference on machine learning*, 766–774. PMLR.

Jain, P.; and Kapoor, A. 2009. Active learning for large multi-class problems. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 762–769. IEEE.

Jeong, J.; Lee, S.; Kim, J.; and Kwak, N. 2019. Consistency-based Semi-supervised Learning for Object detection. 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.

Jhuang, H.; Gall, J.; Zuffi, S.; Schmid, C.; and Black, M. J. 2013. Towards understanding action recognition. In *Proceedings of the IEEE international conference on computer vision*, 3192–3199.

Kalogeiton, V.; Weinzaepfel, P.; Ferrari, V.; and Schmid, C. 2017. Action Tubelet Detector for Spatio-Temporal Action Localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, 4415–4423.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kirsch, A.; Van Amersfoort, J.; and Gal, Y. 2019. Batchbald: Efficient and diverse batch acquisition for deep bayesian active learning. *Advances in neural information processing systems*, 32.

Kumar, A.; Kumar, A.; Vineet, V.; and Rawat, Y. S. 2023. A Large-Scale Analysis on Self-Supervised Video Representation Learning. 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*, 14700–14710.

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

Lee, D.-H. 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.

Lei, J.; Yu, L.; Bansal, M.; and Berg, T. L. 2018. Tvqa: Localized, compositional video question answering. *arXiv preprint arXiv:1809.01696*.

Li, H.; Wu, Z.; Shrivastava, A.; and Davis, L. 2021. Rethinking Pseudo Labels for Semi-Supervised Object Detection. *ArXiv*, abs/2106.00168.

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

Liu, Z.; Wang, J.; Gong, S.; Lu, H.; and Tao, D. 2019. Deep reinforcement active learning for human-in-the-loop person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 6122–6131.

Mettes, P.; and Snoek, C. G. 2019. Pointly-supervised action localization. *International Journal of Computer Vision*, 127(3): 263–281.

Mettes, P.; Snoek, C. G.; and Chang, S.-F. 2017. Localizing actions from video labels and pseudo-annotations. *arXiv preprint arXiv:1707.09143*.

Miyato, T.; Maeda, S.-i.; Koyama, M.; and Ishii, S. 2018. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 41(8): 1979–1993.

Oliver, A.; Odena, A.; Raffel, C. A.; Cubuk, E. D.; and Goodfellow, I. 2018. Realistic evaluation of deep semi-supervised learning algorithms. *Advances in neural information processing systems*, 31.

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.

Pardo, A.; Xu, M.; Thabet, A.; Arbeláez, P.; and Ghanem, B. 2021. BAOD: budget-aware object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1247–1256.

Peng, X.; and Schmid, C. 2016. Multi-region Two-Stream R-CNN for Action Detection. In Leibe, B.; Matas, J.; Sebe, N.; and Welling, M., eds., *Computer Vision – ECCV 2016*, 744–759. Cham: Springer International Publishing. ISBN 978-3-319-46493-0.

Prabhu, A.; Dognin, C.; and Singh, M. 2019. Sampling bias in deep active classification: An empirical study. *arXiv preprint arXiv:1909.09389*.

Rana, A. J.; and Rawat, Y. S. 2022. Are all Frames Equal? Active Sparse Labeling for Video Action Detection. In *Advances in Neural Information Processing Systems*.

Rangnekar, A.; Kanan, C.; and Hoffman, M. 2023. Semantic Segmentation with Active Semi-Supervised Learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 5966–5977.

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 Cortes, C.; Lawrence, N.; Lee, D.; Sugiyama, M.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

Ren, Z.; Yu, Z.; Yang, X.; Liu, M.-Y.; Schwing, A. G.; and Kautz, J. 2020. UFO^2: A Unified Framework Towards Omni-supervised Object Detection. In *European Conference on Computer Vision*, 288–313. Springer.

Rizve, M. N.; Demir, U.; Tirupattur, P.; Rana, A. J.; Duarte, K.; Dave, I. R.; Rawat, Y. S.; and Shah, M. 2021a. Gabriella: An online system for real-time activity detection in untrimmed security videos. In *2020 25th International Conference on Pattern Recognition (ICPR)*, 4237–4244. IEEE.

Rizve, M. N.; Duarte, K.; Rawat, Y. S.; and Shah, M. 2021b. In defense of pseudo-labeling: An uncertainty-aware pseudo-label selection framework for semi-supervised learning. In *International Conference on Learning Representations*.

Rosenberg, C.; Hebert, M.; and Schneiderman, H. 2005. Semi-supervised self-training of object detection models.

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

Schiappa, M. C.; Rawat, Y. S.; and Shah, M. 2022. Self-supervised learning for videos: A survey. *arXiv preprint arXiv:2207.00419*.

Sener, O.; and Savarese, S. 2017. Active learning for convolutional neural networks: A core-set approach. *arXiv preprint arXiv:1708.00489*.

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. R.; and Shah, M. 2012. UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*.

Sun, C.; Shrivastava, A.; Vondrick, C.; Murphy, K.; Sukthankar, R.; and Schmid, C. 2018. Actor-Centric Relation Network. *ArXiv*, abs/1807.10982.

Szegedy, C.; Ioffe, S.; Vanhoucke, V.; and Alemi, A. A. 2017. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Tarvainen, A.; and Valpola, H. 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30.

Vondrick, C.; Shrivastava, A.; Fathi, A.; Guadarrama, S.; and Murphy, K. 2018. Tracking emerges by colorizing videos. In *Proceedings of the European conference on computer vision (ECCV)*, 391–408.

Wang, K.; Zhang, D.; Li, Y.; Zhang, R.; and Lin, L. 2016. Cost-effective active learning for deep image classification. *IEEE Transactions on Circuits and Systems for Video Technology*, 27(12): 2591–2600.

Weinzaepfel, P.; Martin, X.; and Schmid, C. 2016. Human action localization with sparse spatial supervision. *arXiv preprint arXiv:1605.05197*.

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

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

Yang, X.; Song, Z.; King, I.; and Xu, Z. 2021. A Survey on Deep Semi-supervised Learning. *ArXiv*, abs/2103.00550.

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.

Zhang, H.; Zhao, X.; and Wang, D. 2022. Semi-supervised Learning for Multi-label Video Action Detection. In *Proceedings of the 30th ACM International Conference on Multimedia*, 2124–2134.

Zhang, S.; Song, L.; Gao, C.; and Sang, N. 2019. Glnet: Global local network for weakly supervised action localization. *IEEE Transactions on Multimedia*, 22(10): 2610–2622.

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.