ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-6654-0540-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICASSP43922.2022.9746817

SPATIO-TEMPORAL MOTION AGGREGATION NETWORK FOR VIDEO ACTION DETECTION

Hongcheng Zhang, Xu Zhao

Department of Automation, Shanghai Jiao Tong University, Shanghai, China

ABSTRACT

Recognizing action patterns and detecting action instances are vital for spatial temporal action detection task, which aims to recognize the actions of interest in untrimmed videos and localize them in both space and time. The mainstream action tubelet detectors, however, ignore the conflicts in features between localization and classification, and use localization features for temporal modeling, which leads to ineffective action classification. In this paper, we propose the ***Spatio-Temporal Motion Aggregation*** mechanism for integrating the local motion feature from a short term snippet and the longer spatio-temporal information to predict the action category. We design the ***Class-Agnostic Center Localization*** module to perform action instance center localization in the Class-Agnostic manner. Besides, ***Movement and Size Regression*** is proposed for movement estimation and spatial extent detection by using Gaussian kernels to encode training samples. These three modules work together to generate the tubelet detection results, which could be further linked to yield video-level tubes with a matching strategy. Our detector achieves the state-of-the-art performance in both frame-mAP and video-mAP metrics, on the UCF-24 and JHMDB datasets.

Index Terms— video understanding, video action detection, spatio-temporal action detection, anchor-free detector

1. INTRODUCTION

As a crucial problem in video understanding, spatial temporal action detection aims to localize the action instances in both space and time, at the same time predicts their action labels. Spatial temporal action detection can be widely applied in numerous scenarios, such as autonomous driving, video surveillance and advanced video search engines, etc.

The mainstream methods of spatial temporal action detection can be divided into **frame-level** methods and **tubelet-level** methods, which perform action detection on either per-frame or tubelet, and then generate action tubes by linking the detection results.

Frame-level methods have been explored in [1, 2, 3, 4, 5, 6], which utilize 2D detectors to detect 2D boxes from each frame, and then classify the corresponding spatio-temporal features RoI-pooled over actor proposals for action instance recognition. These frame-level methods fail to well capture the correlation between adjacent frames in temporal dimension and thus are less effective for detecting action tubes in video level. Tubelet-level methods perform action detection at the clip (i.e., a short video snippet) level to leverage temporal correlation information which output the regressed tubelets (i.e., a short sequence of bounding boxes) with a sequence of frames as input [7, 8, 9, 10, 11, 12]. These methods perform action detection on consecutive multiple frames, which leverage the temporal correlation to acheive promising performance in video-level detection.

Though tubelet-level methods have achieved promising results on standard benchmarks, there are still some challenges to be solved. Firstly, tubelet-level methods are not effective in action modeling, since the action recognition relies on the spatial feature extracted by 2D CNNs from a short receptive field in temporal dimension. And performing localization and classification tasks in one network will lead to the conflicts in features, as demonstrated in the field of image object detection [13, 14]. This problem becomes more severe when performing video action detection with temporal dimension, as simply taking a longer clip as input will benefit the action classifier to obtain longer temporal information though, but make action localization more challenging. Secondly, tubelet-level methods are mostly anchor-based, which leads to a huge number of pre-defined anchor boxes in spatio-temporal dimension and increase both memory cost and training time, as demonstrated in [8, 12].

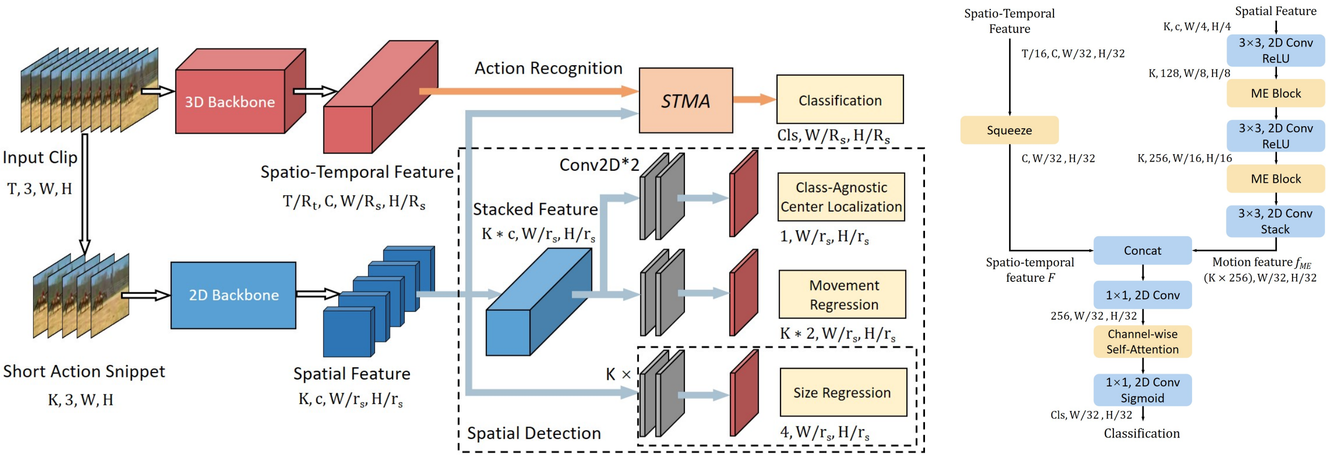
To tackle these challenges, we propose a tubelet-level method in an anchor-free manner, termed as Spatial Temporal Motion Aggregation Network and our contributions are as follows.

(1) To avoid the conflicts in features between localization and classification in tubelet-level methods, we design Class-Agnostic Center Localization and Spatio-Temporal Motion Aggregation to perform center localization and action classification tasks respectively, instead of performing them in one module. Class-Agnostic Center Localization generates the objectness probability heatmap and locate the object center without predicting the category. Spatio-Temporal Motion Aggregation integrates the local motion feature from the action snippet and the spatial temporal information from the longer temporal receptive field to perform the action classification.

(2) We adopt an anchor-free detection strategy that using a movement branch to regress center offset and a size branch to get the spatial extent of bounding box in the clip based on [12]. Actually, treating the detected center point of each frame as one training sample is not optimal for detection, since a mini perturbation of the center localization will lead to an inaccurate movement regression and size regression in the inference. Inspired by recent anchor-free detector [15], we treat all pixels in a Gaussian-area as training samples both for movement regression and size regression, thus can make the network not sensitive to the position of key frame's center.

(3) We perform experiments on two challenging action tube detection benchmarks of UCF-24 [16] and JHMDB [17]. Our method outperforms the existing state-of-the-art tubelet-level approaches in both frame-mAP and video-mAP on these two datasets.

This work has been funded in part by the NSFC grants 62176156, 61673269, and Shanghai Engineering Research Center of Intelligent Control and Management.



(a) (b)

Fig. 1. The pipeline of Spatio-Temporal Motion Aggregation network which consists of STMA mechanism and spatial detection branch, as shown in (a). The implementation details of STMA mechanism are shown in (b), best view in colors.

2. APPROACH

2.1. Framework Overview

Action tubelet detection aims at localizing a short sequence of bounding boxes from an input clip and recognizing its action category as well. As shown in Fig. 1(a), Given a T frames clip $I \in \mathbb{R}^{T\times 3\times W\times H}$ with resolution of $W \times H$, we sample K consecutive frames from the input clip to generate a **short action snippet** $I_{snippet} \in \mathbb{R}^{K\times 3\times W\times H}$. First, we input the whole clip I into a 3D backbone to extract spatio-temporal feature $F \in \mathbb{R}^{\frac{T}{R_t}\times C\times\frac{W}{R_s}\times\frac{H}{R_s}}$, and input the short action snippet into a weight shared 2D backbone to extract spatial feature $f \in \mathbb{R}^{K\times c\times\frac{W}{r_s}\times\frac{H}{r_s}}$, simultaneously. R_s and r_s, R_t, C and c are the spatial down-sample rate, temporal down-sample rate, and channel dimension of feature F and f, respectively. Then, we design two branches to perform tubelet detection in an anchor-free manner. The first branch is **Action Recognition**, which is defined over all frames in the clip. **Spatio-Temporal Motion Aggregation** mechanism is used in this branch to aggregate the local motion information from the short action snippet and the longer spatio-temporal information from the entire clip to predict action category. The second branch is **Spatial Detection**, which consists of three parts: **Class-Agnostic Center Localization**, **Movement Regression** and **Size Regression** and performs the spatial detection on each frame in the short action snippet. These two branches collaborate together to yield tubelet detection results from the input clip, which will be further linked to form action tube in a long untrimmed video by following a common linking strategy.

2.2. Spatio-Temporal Motion Aggregation

In action recognition branch, we propose **Spatio-Temporal Motion Aggregation (STMA)** mechanism to integrate the local motion feature in the short term snippet and the longer temporal information extracted from 3D backbone, as shown in Fig. 1(b). To obtain richer semantic information for classification and make the features match in spatial scale, we use convolutional layer with stride 2 to conduct down-sampling in feature f. Then, we adopt **Motion Excitation (ME)** to capture the motion information from adjacent frames based on the feature level. Motion feature f_m is modeled following the similar operation presented in [18, 19] as shown in Eq.(1). The motion feature is concatenated to each other according to the temporal dimension with zero-padding in the last element, as follows:

$$f_m = \mathrm{Conv}(f(t+1)) - f(t), f_m \in \mathbb{R}^{c_1\times\frac{W}{r_s}\times\frac{H}{r_s}}. \quad (1)$$

$$f_M = [f_m(1), ..., f_m(K-1), 0], f_M \in \mathbb{R}^{K\times c_1\times\frac{W}{r_s}\times\frac{H}{r_s}}. \quad (2)$$

The f_M is processed by spatial average pooling, convolutional layer and sigmoid to generate a mask $M \in \mathbb{R}^{K\times c_1\times 1\times 1}$. The mask M is multiplied element-wise with the input, which is added to the input as a residual. Then, we concatenate the motion feature f_{ME} and F to perform a channel-wise self-attention for feature aggregation. Finally, we feed the fused feature to a convolutional layer to predict the action category and use cross entropy as classification loss.

2.3. Class-Agnostic Center Localization

Class-Agnostic Center Localization aims to locate the center positions of action instances for key (center) frame in short snippet. In order to decouple the center localization and classification, it is worth noting that we perform center localization in a class-agnostic manner instead of locating center and predicting action category together in one branch [8, 12, 20]. Based on the short-term snippet feature representation extracted by 2D backbone $f \in \mathbb{R}^{K\times c\times\frac{W}{r_s}\times\frac{H}{r_s}}$, we locate the key frame center by estimating a objectness probabilities heatmap $\hat{P} \in [0,1]^{\frac{W}{r_s}\times\frac{H}{r_s}}$, and $\hat{P}_{xy}$ represents the probability of an action instance centered at position (x, y). We use a Gaussian kernel to generate the heatmap $P_{xy} = \exp(-\frac{(x-x_i)^2+(y-y_i)^2}{2\sigma^2})$ for the i^{th} action instance's key frame center in the current action snippet. We leverage a focal loss for training the probability heatmap, as represented in Eq.(3):

$$L_{ct} = -\frac{1}{N}\sum_{x,y}\begin{cases}(1-\hat{P}_{xy})^{\alpha}\log(\hat{P}_{xy}) & \text{if } P_{xy}=1\\ (1-P_{xy})^{\beta}(\hat{P}_{xy})^{\alpha}\log(1-\hat{P}_{xy}) & \text{otherwise}\end{cases} \quad (3)$$

2.4. Gaussian kernel based Regression

Inspired by 2D object detection method [15], we utilize the pixels inside a Gaussian kernel area as the training samples, instead of single pixel at the object center. We leverage a different Gaussian kernel $G_i(x,y) = \exp(-\frac{(x-x_i)^2+(y-y_i)^2}{2{\sigma_r}^2})$ from center localization, where (x_i, y_i) represents the center of key frame location for the i^{th} action instance in current snippet. And the non-zero part in G_i named as Gaussian area A_i and each pixel in A_i will be treated as a training sample for both movement and size regression.

Movement regression. Movement Regression has been explored by [12] previously, which captures the correlation between adjacent frames in temporal dimension by regressing the encoded center offsets. To improve the regression performance, we adopt the training samples from Gaussian area A_n based on the action instances' center instead of single center point. We concatenate multi-frame features along channel dimension as the input of Movement Regression, which outputs a movement prediction map $\hat{M} \in \mathbb{R}^{(2\times K)\times \frac{W}{r_s}\times \frac{H}{r_s}}$, as shown in Fig. 1(a). For i^{th} action instance from each frame in current snippet of length K, the regression target is defined as the offset between the center of the current frame and the key frame, as follows:

$$m_j^i = (x_j^i - x_{key}^i, y_j^i - y_{key}^i), j = 1, 2, \dots, K. \tag{4}$$

where (x_j^i, y_j^i) is the bounding box center of i^{th} action instance at j^{th} frame. We optimize the movement map $\hat{M}$ at the Gaussian area for training and use the L1 loss as follows:

$$L_{mov} = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{N_i}\sum_{(x,y)\in A_i} w_{x,y}\left|\hat{M}_{x,y} - m^i\right|. \tag{5}$$

where n and N_i are the number of action instances and the number of samples for i^{th} instance, respectively. And w is the sample weight according to their area to balance the instances with different size.

Size regression. To determine the action instance's box size for each frame in the action snippet and output a tubelet of length K. Size regression takes each frame's feature $f \in \mathbb{R}^{c\times \frac{W}{r_s}\times \frac{H}{r_s}}$ as input and regresses a size prediction map in a frame-level manner. For each frame in the input snippet, given pixel (x, y) in Gaussian area A_i, the regression target is defined as the center offset from (xr_s, yr_s) to four sides of the bounding box on image scale, represented as $(w_l, h_t, w_r, h_b)_{x,y}^i$. The predicted box $\hat{B} = (\hat{x}_1, \hat{y}_1, \hat{x}_2, \hat{y}_2)$ at (x, y) can be represented as:

$$\begin{aligned} \hat{x}_1 = xr_s - w_l,\ \hat{y}_1 = yr_s - h_t, \\ \hat{x}_2 = xr_s + w_r,\ \hat{y}_2 = yr_s + h_b. \end{aligned} \tag{6}$$

Then we use a GIoU loss related to response of the Gaussian distribution area A_i. So the formula of the Size Regression loss can be described as follows:

$$L_{size} = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{N_i}\sum_{(x,y)\in A_i} w_{x,y}\mathrm{GIoU}(\hat{B}_{x,y}^i, B^i). \tag{7}$$

where $\hat{B}_{x,y}^i$ stands for the predicted box at position (x, y) and B^i is the corresponding i^{th} ground-truth box on image scale.

Total loss. The total loss is composed of classification loss L_{cls}, center localization loss L_{ct} and regression loss L_{reg}, weighted by scalars. We set $w_{cls} = 0.01$, $w_{ct} = 1$, $w_{reg} = 0.5$, $\gamma = 0.1$ in our experiment, and the formula is as follows:

$$L_{reg} = L_{mov} + \gamma L_{size}. \tag{8}$$

$$L_{total} = w_{cls}L_{cls} + w_{ct}L_{ct} + w_{reg}L_{reg}. \tag{9}$$

2.5. Inference

After the training, we can get the results of action instances' centers from class-agnostic center localization. Specifically, we treat the top $N = 10$ maximums in the estimated heatmap P as centers. With N action instance centers, we can recognize its action label according to classification results from STMA. Then we use the positions of these centers to gather the results of movement regression, which are used for calculate the other frames' center in the short snippet. Finally, we can get bounding box at each action instance's center of each frame as a tubelet. After getting the clip-level detection results, we link these tubelets into final action tubes across time by using common linking algorithm as [7, 12] for fair comparison.

3. EXPERIMENT

3.1. Experimental Setting

Datasets and metric. We perform experiments on the UCF-24 [16] and JHMDB [17] datasets and adopt the common settings in data processing as the previous tubelet-level methods [7, 12]. We utilize frame mAP and video mAP to evaluate detection performance.

Training details. We choose the DLA34 [22] as our 2D backbone with COCO [23] pretrain and 3D ResNext-101 [24] as our 3D backbone with K400 [25] pretrain. We set the input clip length $T = 16$ and length of short action snippet $K = 5$. We set the spatial downsample rate $r_s = 4$ for 2D backbone, temporal downsample rate $R_t = 16$ and spatial downsample rate $R_s = 32$ for 3D backbone. To reduce training time and memory cost, we only input rgb frames resized to 288 × 288 for training instead of two-stream manner. We adopt the cosine annealing learning rate strategy with an initial learning rate 0.0001 and set the batch size to 24. And we train the entire network end-to-end with the Adam optimizer for 16 epochs on UCF-24 [16] dataset and 12 epochs on JHMDB[17] dataset, which is performed on 8 RTX 2080Ti GPUs.

3.2. Comparison with the State-of-the-Art methods

In this section, we compare our method with the existing state-of-the-art spatio-temporal action detection methods on the UCF-24 and JHMDB datasets as shown in Table 1. For a fair comparison, we also report two-stream results of these methods. On JHMDB dataset, our method outperforms other tubelet methods [8, 7, 12, 20] for video-mAP and get comparable performance to these two-stream based methods on UCF-24 dataset for both frame-mAP and video-mAP. This result confirms that our method is effective for localizing precise tubelets from clips, and can effectively perform action recognition by aggregating motion and spatio-temporal information.

3.3. Ablation study

Effectiveness of Spatio-Temporal Motion Aggregation (STMA). In order to verify the effectiveness of STMA, and independently explore the impact and contribution of each module, we conducted

Method	UCF-24					JHMDB				
	Frame-mAP@0.5	Video-mAP				Frame-mAP@0.5	Video-mAP			
		@0.2	@0.5	@0.75	@0.5:0.95		@0.2	@0.5	@0.75	@0.5:0.95
Single-stream(RGB)										
ACT [7]	-	-	41.8	-	18.5	-	-	60.0	-	34.0
T-CNN [9]	41.4	47.1	-	-	-	-	78.4	76.9	-	-
TACNet [20]	-	-	45.0	-	19.4	-	-	64.5	-	35.1
Dance with Flow [11]	-	-	45.6	-	20.2	-	-	63.6	-	38.0
MOC [12]	73.1	78.8	51.0	27.1	26.5	-	-	-	-	-
Ours	**78.8**	**83.3**	54.1	**29.7**	28.1	77.6	**81.5**	**80.7**	**73.2**	**60.4**
Two-stream(RGB+FLOW)										
ACT [7]	-	77.2	51.4	22.7	25.0	65.7	74.2	73.7	52.1	44.8
STEP [8]	75.0	76.6	-	-	-	-	-	-	-	-
TACNet [20]	72.1	77.5	52.9	21.8	24.1	65.5	74.1	73.4	52.5	44.8
Dance with Flow [11]	-	78.5	50.3	22.2	24.5	-	-	74.7	53.3	45.0
AVA baseline [21]	76.3	-	**59.9**	-	-	73.3	-	78.6	-	-
ACRN [6]	-	-	-	-	-	**77.9**	-	80.1	-	-
MOC [12]	78.0	82.8	53.8	29.6	**28.3**	70.8	77.3	77.2	71.7	59.1

Table 1. Spatial Temporal action detection performance comparison with state-of-the-art methods on UCF-24 and JHMDB, measured by both frame-mAP and video-mAP on different IoU thresholds and average mAP.

sufficient ablation experiments on UCF-24 dataset, as shown in Table 2. We explored the impact of Class-Agnostic Center Localization (CAL), Motion Excitation (ME) and Spatio-Temporal Aggregation (STA) on the performance of both video mAP and frame mAP. We adopt single stream model from MOC [12] as our baseline model, the other settings are the same as mentioned in Section 3.1. From the comparison of "CAL" , "CAL+ME", "CAL+ME+STA", we can find that ME benifits action recognition and improves the performance of action detection with the local motion feature. And the STA is conducive to our model to better integrate long-term spatio-temporal feature with the local motion information, while CAL can ensure that long-term spatio-temporal information will not interfere with the performance of spatial detection.

method	STMA strategy			F-mAP@0.5	V-mAP@0.2	V-mAP@0.5
	CAL	ME Block	STA			
base				72.8	77.8	50.2
CAL	✓			73.1	77.5	50.6
CAL+ME	✓	✓		75.4	79.0	51.2
CAL+ME+STA	✓	✓	✓	77.2	80.6	52.6
CAL+ME+GR	✓	✓		76.1	80.1	51.9
CAL+ME+STA+GR	✓	✓	✓	**78.8**	**83.3**	**54.1**

Table 2. Ablation study on Spatio-Temporal Motion Aggregation on UCF-24 dataset to elaborate effects of different settings.

method	Gaussian Kernel		F-mAP@0.5	V-mAP@0.2	V-mAP@0.5
	Size	Movement			
base			72.8	77.8	50.2
Size Regression	✓		74.4	77.4	50.8
Movement Regression		✓	73.9	78.3	50.7
Size + Movement	✓	✓	74.7	79.2	51.1

Table 3. Ablation study on Gaussian kernel based Regression. It shows that using Gaussian kernel based area for training benefits both size and movement regression in action detection task.

Study on Gaussian kernel based regression. To prove the effectiveness of Gaussian kernel based Regression (GR), we performed ablation study on the UCF-24 dataset, as shown in Table 3. Based on the baseline mentioned in Sec 3.3, we compared the action detection performance of GR on Size Regression, Movement Regression and both of them. The experiment result shows that taking the samples from the Gaussian kernel region as training samples can improve the detection performance on movement and size regression.

tubelet length	F-mAP@0.5	V-mAP@0.2	V-mAP@0.5
$K = 1$	77.9	75.8	48.4
$K = 3$	78.6	77.1	50.6
$K = 5$	78.8	83.3	54.1
$K = 7$	78.3	81.3	52.9

Table 4. Ablation study on the tubelet length K.

Study on tubelet length. To explore the influence of tubelet length K on detection, we set up different tubelet length K to perform ablation experiments as shown in Table 4. The result shows that the performance is optimal when $K = 5$. And when $K = 1$, our method is equivalent to a frame-level detector, thus less effective to detect action tubes in video-level. The detection performance declines when K is greater than 5, since the spatial displacement in temporal dimension will make movement and size regression hard.

4. CONCLUSION

In this paper, we proposed Spatio-Temporal Motion Aggregation Network, which performs tubelet-level video action detection in an anchor-free manner. STMA mechanism integrates the local motion feature from a short term snippet and the longer spatio-temporal information to perform action recognition more accurately. Besides, Gaussian kernel based training strategy makes the movement and size regression not sensitive to the deviation in center localization, which improves the detection performance. Experiments conducted on two benchmarks including UCF-24 and JHMDB have validated the state-of-the-art performance of our method.

5. REFERENCES

[1] Georgia Gkioxari and Jitendra Malik, “Finding action tubes,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 759–768.

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

[3] Suman Saha, Gurkirt Singh, Michael Sapienza, Philip HS Torr, and Fabio Cuzzolin, “Deep learning for detecting multiple space-time action tubes in videos,” *arXiv preprint arXiv:1608.01529*, 2016.

[4] Gurkirt Singh, Suman Saha, Michael Sapienza, Philip HS Torr, and Fabio Cuzzolin, “Online real-time multiple spatiotemporal action localisation and prediction,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 3637–3646.

[5] Jianchao Wu, Zhanghui Kuang, Limin Wang, Wayne Zhang, and Gangshan Wu, “Context-aware rcnn: A baseline for action detection in videos,” in *European Conference on Computer Vision*. Springer, 2020, pp. 440–456.

[6] Chen Sun, Abhinav Shrivastava, Carl Vondrick, Kevin Murphy, Rahul Sukthankar, and Cordelia Schmid, “Actor-centric relation network,” in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 318–334.

[7] Vicky Kalogeiton, Philippe Weinzaepfel, Vittorio Ferrari, and Cordelia Schmid, “Action tubelet detector for spatio-temporal action localization,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 4405–4413.

[8] Xitong Yang, Xiaodong Yang, Ming-Yu Liu, Fanyi Xiao, Larry S Davis, and Jan Kautz, “Step: Spatio-temporal progressive learning for video action detection,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 264–272.

[9] Rui Hou, Chen Chen, and Mubarak Shah, “Tube convolutional neural network (t-cnn) for action detection in videos,” in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 5822–5831.

[10] Suman Saha, Gurkirt Singh, and Fabio Cuzzolin, “Amtnet: Action-micro-tube regression by end-to-end trainable deep architecture,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 4414–4423.

[11] Jiaojiao Zhao and Cees GM Snoek, “Dance with flow: Two-in-one stream action detection,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 9935–9944.

[12] Yixuan Li, Zixu Wang, Limin Wang, and Gangshan Wu, “Actions as moving points,” in *European Conference on Computer Vision*. Springer, 2020, pp. 68–84.

[13] Yue Wu, Yinpeng Chen, Lu Yuan, Zicheng Liu, Lijuan Wang, Hongzhi Li, and Yun Fu, “Rethinking classification and localization for object detection,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 10186–10195.

[14] Guanglu Song, Yu Liu, and Xiaogang Wang, “Revisiting the sibling head in object detector,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 11563–11572.

[15] Zili Liu, Tu Zheng, Guodong Xu, Zheng Yang, Haifeng Liu, and Deng Cai, “Training-time-friendly network for real-time object detection,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020, vol. 34, pp. 11685–11692.

[16] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah, “A dataset of 101 human action classes from videos in the wild,” *Center for Research in Computer Vision*, vol. 2, no. 11, 2012.

[17] Hueihan Jhuang, Juergen Gall, Silvia Zuffi, Cordelia Schmid, and Michael J Black, “Towards understanding action recognition,” in *Proceedings of the IEEE international conference on computer vision*, 2013, pp. 3192–3199.

[18] Boyuan Jiang, MengMeng Wang, Weihao Gan, Wei Wu, and Junjie Yan, “Stm: Spatiotemporal and motion encoding for action recognition,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 2000–2009.

[19] Yan Li, Bin Ji, Xintian Shi, Jianguo Zhang, Bin Kang, and Limin Wang, “Tea: Temporal excitation and aggregation for action recognition,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 909–918.

[20] Lin Song, Shiwei Zhang, Gang Yu, and Hongbin Sun, “Tacnet: Transition-aware context network for spatio-temporal action detection,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 11987–11995.

[21] Chunhui Gu, Chen Sun, David A Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, et al., “Ava: A video dataset of spatio-temporally localized atomic visual actions,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6047–6056.

[22] Fisher Yu, Dequan Wang, Evan Shelhamer, and Trevor Darrell, “Deep layer aggregation,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 2403–2412.

[23] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick, “Microsoft coco: Common objects in context,” in *European conference on computer vision*. Springer, 2014, pp. 740–755.

[24] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh, “Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet?,” in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2018, pp. 6546–6555.

[25] Joao Carreira and Andrew Zisserman, “Quo vadis, action recognition? a new model and the kinetics dataset,” in *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6299–6308.