

# TACNet: Transition-Aware Context Network for Spatio-Temporal Action Detection

Lin Song<sup>1\*</sup> Shiwei Zhang<sup>2\*</sup> Gang Yu<sup>3</sup> Hongbin Sun<sup>1†</sup>

<sup>1</sup> Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University.

<sup>2</sup> Artificial Intelligence and Automation, Huazhong University of Science and Technology.

<sup>3</sup> Megvii Inc. (Face++).

{stevengrove@stu, hsun@mail}.xjtu.edu.cn, swzhang@hust.edu.cn, yugang@megvii.com

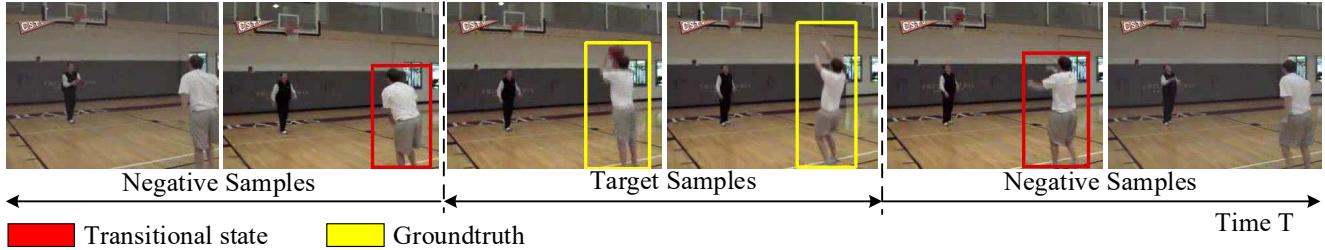


Figure 1: Diagram of transitional state. There are some ambiguous states around but not belong to the target actions, and it is hard to distinguish them. We define the states as “transitional state” (red boxes). If we can effectively distinguish these states, we can improve the ability of temporal extent detection.

## Abstract

Current state-of-the-art approaches for spatio-temporal action detection have achieved impressive results but remain unsatisfactory for temporal extent detection. The main reason comes from that, there are some ambiguous states similar to the real actions which may be treated as target actions even by a well-trained network. In this paper, we define these ambiguous samples as “transitional states”, and propose a Transition-Aware Context Network (TACNet) to distinguish transitional states. The proposed TACNet includes two main components, i.e., temporal context detector and transition-aware classifier. The temporal context detector can extract long-term context information with constant time complexity by constructing a recurrent network. The transition-aware classifier can further distinguish transitional states by classifying action and transitional states simultaneously. Therefore, the proposed TACNet can substantially improve the performance of spatio-temporal action detection. We extensively evaluate the proposed TACNet on UCF101-24 and J-HMDB datasets. The experimental results demonstrate that TACNet obtains competitive performance on JHMDB and significantly outperforms

the state-of-the-art methods on the untrimmed UCF101-24 in terms of both frame-mAP and video-mAP.

## 1. Introduction

Action detection focuses both on classifying the actions present in a video and on localizing them in space and time. It has been receiving more and more attention from researchers because of its various applications. Action detection has already served as a critical technology in anomaly detection, human-machine interaction, video monitoring, etc. Currently, most of action detection approaches [8, 15, 19, 23] separate the spatio-temporal detection into two stages, i.e., spatial detection and temporal detection. These approaches adopt the detectors based on deep neural networks [4, 12] to spatially detect action in the frame level. Then, they construct temporal detection by linking frame-level detections and applying some objective functions, such as maximum subarray method [15], to create spatio-temporal action tubes. Because these methods treat video frames as a set of independent images, they can not exploit the temporal continuity of videos. Thus, their detection results are actually unsatisfactory.

To address this issue, ACT [10] employs a stacked strategy to exploit the short-term temporal continuity for clip-

\*indicates equal contribution.

†indicates corresponding author.

level detection and significantly improves the performance of spatio-temporal action detection. However, ACT still can not extract long-term temporal context information, which is critical to the detection of many action instances, such as “long jump”. Moreover, due to the two separate stages in action detection, ACT can not thoroughly address the time error induced by ambiguous samples, which are illustrated as red boxes in Figure 1. In this paper, the ambiguous sample is defined as “transitional state”, which is close to the action duration but does not belong to the action. According to the error analysis of ACT detector, 35%-40% of the total errors [10] are time errors, which are mainly caused by transitional states. Therefore, to further improve the performance of spatio-temporal action detection, it is critical to extract long-term context information and distinguish the transitional states in a video sequence.

The above observations motivate this work. In particular, we propose a Transition-Aware Context Network (TACNet) to improve the performance of spatio-temporal action detection. The proposed TACNet includes two main components, i.e., temporal context detector and transition-aware classifier. The temporal context detector is designed based on standard SSD framework, but can encode the long-term context information by embedding several multi-scale Bi-directional Conv-LSTM [11] units. To the best of our knowledge, this is the first work to combine Conv-LSTM with SSD to construct a recurrent detector for action detection. The transition-aware classifier can distinguish transitional states by classifying action and action states simultaneously. More importantly, we further propose a common and differential mode scheme to accelerate the convergence of TACNet. Therefore, the proposed TACNet can not only extract long-term temporal context information but also distinguish the transitional states. We test the proposed TACNet on UCF101-24 [21] and J-HMDB [9] datasets and achieve a remarkable improvement in terms of both frame- and video-level metrics on both datasets.

In summary, we make the following three contributions:

- we propose a temporal context detector to extract long-term temporal context information efficiently with constant time complexity;
- we design a transition-aware classifier, which can distinguish transitional states and alleviate the temporal error of spatio-temporal action detection;
- we extensively evaluate our TACNet in untrimmed videos from the UCF-24 dataset and achieve state-of-the-art performance.

## 2. Related Work

Spatio-temporal action detection methods can generally be classified into two categories: weakly and fully supervised

methods. Although we concentrate on fully supervised methods in this paper, weakly supervised methods have also achieved significant improvement in recent years. The purpose of these methods is to detect actions only with video-level labels but without frame-level bounding box annotations. These methods can significantly reduce annotation costs and are more suitable to process large unannotated video data. Multi-Instance Learning (MIL) is one of the frequently-used approaches for weakly supervised spatio-temporal action detection. In [20], Siva *et al.* transforms the weakly supervised action detection as a MIL problem. They globally optimize both inter- and intra-class distances to locate interested actions. Multi-fold MIL scheme is then proposed in [6] to prevent training from prematurely locking onto erroneous object detection. Recently, deep model and attention mechanisms are also employed in deep model based weakly supervised methods. The methods in [18, 11] apply attention mechanism to focus on key volumes for action detection. Besides, Mettes *et al.* [14, 13] proposes to apply point annotations to perform action detection.

Compared with weakly supervised methods, fully supervised methods can leverage bounding box level annotations to achieve remarkable performance for spatio-temporal action detection. Many approaches are proposed to construct action tubes. Gkioxari *et al.* [5] firstly proposes to apply the linking algorithm on frame-level detection to generate action tubes. Peng *et al.* [15] improves frame-level action detection by stacking optical flow over several frames and also proposes a maximum subarray method for temporal detection. Weinzaepfel *et al.* [22] improves the linking algorithm by using a tracking-by-detection method. Singh *et al.* [19] designs online algorithm to incrementally generate action tubes for real-time action detection. However, these methods do not explore the temporal information of actions, and the performance is still unsatisfactory. To encode the temporal information, Saha *et al.* [16] and Hou *et al.* [8] extend classical region proposal network (RPN) to 3D RPN, which generates 3D region proposals spanned by several successive video frames. Becattini *et al.* [1] adopts LSTM to predict action progress. Zhu *et al.* [25] proposes a two-stream regression network to generate temporal proposals. In contrast, Kalogeiton *et al.* [10] stacks the feature maps of multiple successive frames by SSD detector to predict score and regression on anchor cuboid and achieves the state-of-the-art performance. Therefore, this paper uses ACT [10] as the baseline to compare and evaluate the action detection performance of TACNet.

## 3. Transition-Aware Context Network

### 3.1. TACNet framework

Figure 2 illustrates the overall framework of TACNet, which mainly consists of two parts, i.e., two-stream tem-

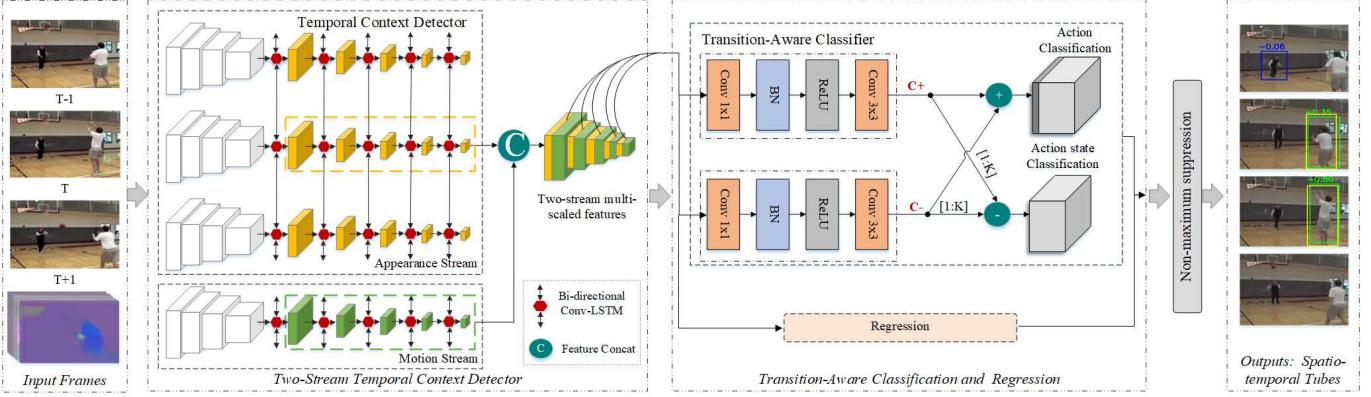


Figure 2. Overall framework of the proposed TACNet. TACNet mainly contains two modules: temporal context detector and transition-aware classifier. In the temporal context detector, we embed several multi-scale Conv-LSTM [11] units in the standard SSD detector [12] to extract temporal context. Based on the recurrent action detector, the transition-aware classifier is designed to simultaneously detect the action categories and states. Then, we can correctly localize the temporal boundaries for the target actions.

poral context detection and transition-aware classification & regression. Although the framework is similar to most of the previous methods, especially ACT detector, temporal context detector and transition-aware classifier are proposed to significantly improve the capability of extracting long-term temporal context information and distinguishing transitional states respectively. For the temporal context detector, we adopt two-stream SSD to construct action detection as the ACT detector does. In addition, to extract long-term temporal context information, we embed several Bi-directional Conv-LSTM (Bi-ConvLSTM) [11] unites into different feature maps with different scales. The Bi-directional ConvLSTM architecture can keep the spatial layout of the feature maps, and is of benefit to performing spatial localization. In the transition-aware classifier, to distinguish transitional states, we propose two branches to classify the actions and action states simultaneously. Moreover, we further design a common and differential mode scheme, inspired by the basic concepts in the signal processing domain [2], to accelerate the convergence of the overall TACNet. Associating with the regression module, transition-aware classifier can spatially detect the actions and temporally predict the temporal boundaries in the meantime. In addition, the proposed method can be embedded into various detection frameworks, and this work is based on the SSD due to its effectiveness and efficiency.

### 3.2. Temporal Context Detector

Long-term temporal context information is critical to spatio-temporal action detection. The standard SSD performs action detection from multiple feature maps with different scales in the spatial level, but it does not consider temporal context information. To extract temporal context, we embed Bi-ConvLSTM unit into SSD framework to design a recurrent detector. As a kind of LSTM, ConvLSTM

not only can encode long-term information, but also is more suitable to handle spatio-temporal data like video. Because the ConvLSTM unit can preserve the spatial structure of frames over time by replacing the fully connected multiplicative operations in an LSTM unit with convolutional operations. Therefore, it is reasonable to employ ConvLSTM units into our framework to extract long-term temporal information.

In particular, we embed a Bi-ConvLSTM unit between every two layers of adjacent scales in SSD to construct the proposed temporal context detector, as shown in Figure 2. The proposed module considers the input sequences in both forward and backward directions which adopts a pair of temporal-symmetric ConvLSTM for these two directions. The Bi-ConvLSTM can extract a pair of features for each scale in a frame. These features are concatenated and transformed by a  $1 \times 1$  convolutional layer to eliminate the redundancy of channels. By this means, the proposed temporal context detector can leverage the advantage of SSD and extract long-term temporal context information as well. Moreover, we also make two modifications to the Bi-ConvLSTM unit: (i) we replace activation function  $tanh$  by  $ReLU$ , which can improve the performance slightly according to experimental results; (ii) we apply 2D dropout between inputs and hidden states to avoid over-fitting.

Compared with the ACT, our method is also efficient in terms of computational cost. ACT applies a sliding window with stride 1, and takes  $n$  stacked frames as input for the processing of each frame. Therefore, the computational complexity of ACT is  $O(n)$ . On the contrary, we constantly process each frame twice. Therefore, supposing  $n$  is the number of stacked frames, the computational cost of ACT and the proposed temporal context detector is  $O(n)$  and  $O(1)$  respectively. We can find that the computational cost gap increases when  $n$  grows, especially  $n$  can be vast when

considering long-term temporal information.

### 3.3. Transition-Aware Classifier

Proposals in the transitional state have a similar appearance to target actions and may easily confuse the detectors. Most of the existing methods do not provide explicit definition on these proposals but rely on post-processing algorithms to prune them or simply treat them as background. However, since these proposals are much different from background (e.g., scenes and other objects), treating them as background enlarges intra-class variance and limits the detection performance. In this paper, we propose a transition-aware classifier to perform the action category classification and the prediction of transitional state simultaneously.

To simultaneously predict action categories and action states, we firstly define a pair of scores, i.e.  $\mathbf{c}^+ = [c_0^+, c_1^+, \dots, c_K^+]$  and  $\mathbf{c}^- = [c_1^-, c_2^-, \dots, c_K^-]$ , where  $K$  is the number of categories and  $c_0^+$  is the score of background. The scores  $\mathbf{c}^+$  and  $\mathbf{c}^-$  denote the action classification scores and transitional state classification scores respectively. We should note that the transitional state scores have no background category. In the transition-aware classifier, we apply two classifiers to predict these two scores, as shown in Figure 3. Based on these definitions, we formulate the training targets of target actions and transitional states as following:

For an active sample of category  $i$ , the training target should meet Eq. 1:

$$c_i^+ > c_i^- \text{ and } c_i^+ > c_j^+, \forall j \neq i \quad (1)$$

where  $i, j \in [1, 2, 3, \dots, K]$ .

While for a transitional sample of category  $i$ , the training target should meet Eq. 2:

$$c_i^+ < c_i^- \quad (2)$$

In general, we can directly train TACNet based on these targets. However, we find that it is hard to converge when learning these two targets. The reason may come from that the  $\mathbf{c}^+$  and  $\mathbf{c}^-$  are inter-coupled, which leads to the interaction between them. For example, a transitional sample of the category  $i$  tries to simultaneously minimize  $c_i^+$  and maximize  $c_i^-$  to meet Eq.2, which will hurt the distribution of category prediction  $\mathbf{c}^+$ .

To solve this issue, we take advantage of the knowledge in signal processing domain [2] that inter-coupled signal can be decoupled by a pair of individual branches, i.e., common mode and differential mode branches. Inspired by these concepts, we design a common and differential scheme to train our network. In particular, the proposed transition-aware classifier still outputs  $\mathbf{c}^+$  and  $\mathbf{c}^-$ , but the difference is that we use  $\mathbf{c}^+ + \mathbf{c}^-$  to predict action category (upper branch in Figure 3) and  $\mathbf{c}^+ - \mathbf{c}^-$  to predict action

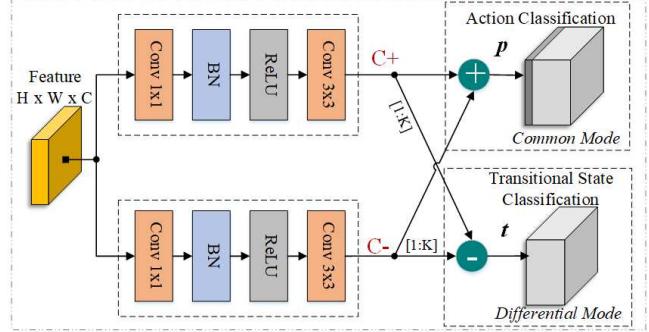


Figure 3. Diagram of the transition-aware classifier. We propose a common and differential mode scheme to decouple the inter-coupled features into two branches, which predict action category and action states respectively.

state (lower branch in Figure 3). We formulate the decoupled targets as following:

$$p_i = \frac{e^{c_i^+ + c_i^-}}{\sum_{j \in [0, K]} e^{c_j^+ + c_j^-}}, \forall i \in [0, K] \quad (3)$$

$$t_i = \frac{e^{c_i^+ - c_i^-}}{e^{c_i^+ - c_i^-} + 1}, \forall i \in [1, K]. \quad (4)$$

where  $p_i$  denotes the probability of being classified as category  $i$  (Eq.3), which is independent to action states. However, we should note that  $t_i$  denotes the probability of active state but not transitional state (Eq.4). The probability of a transitional state is calculated as  $1 - t_i$ .

When a sample in transitional state tries to minimize  $\mathbf{c}^+ - \mathbf{c}^-$ , we suppose that  $\mathbf{c}^+$  changes to  $\mathbf{c}^+ - \lambda$ , and  $\mathbf{c}^-$  will change to  $\mathbf{c}^- + \lambda$  (since these two branches share the same gradient magnitude). The category prediction  $\mathbf{c}^+ + \mathbf{c}^-$  will have no change. By this means, the predictions of the action category and action state are decoupled. In our experiments, we find that the network can converge easily and the prediction has a very slight effect on each other.

### 3.4. Training and Inference Procedures

In the training phase of TACNet, we denote  $P$  as the positive samples which have more than 0.5 IoU with at least one ground-truth bounding box. We apply  $T$  and  $\mathbf{v}$  to denote the positive sample set of transitional states and the corresponding predicted action categories respectively. However, there are no annotations for transitional states in the existing untrimmed datasets so far. According to the definition of transitional states, we propose a “simple-mining” strategy to label the positive samples of the transition states. In detail, the detected boxes are treated as transitional state samples when their scores meet  $c_i^+ > c_0^+$ , but their corresponding frames have no ground-truth annotations. These transitional state samples are further applied to train TACNet.

We employ the same loss function as SSD for regression  $\mathcal{L}_{\text{reg}}$ . Besides, we introduce the classification loss  $\mathcal{L}_{\text{cls}}$  and transitional loss  $\mathcal{L}_{\text{trans}}$  based on the classification scores  $\mathbf{p}$  and action state scores  $\mathbf{t}$  illustrated in Eq.5:

$$\begin{aligned}\mathcal{L}_{\text{cls}} &= -\sum_{j \in P} \log p_y^j - \sum_{j \in G \setminus P} \log p_0^j, \\ \mathcal{L}_{\text{trans}} &= -\sum_{j \in P} \log t_y^j - \sum_{j \in T} \log (1 - t_{\mathbf{v}_j}^j), \\ \mathbf{v}_j &= \arg \max_{i \in [1, K]} (p_i^j), T = \{j \mid \mathbf{v}_j > 0, j \in U \setminus G\},\end{aligned}\quad (5)$$

where  $U$  and  $G$  refers to the set of all available anchors and anchors in the images which have groundtruth annotations respectively,  $p_i^j$  is the predicted probability of  $j$ -th anchor with category  $i$ , and  $t_i^j$  is the probability of  $j$ -th transitional anchor with predicted category  $i$ .

We optimize the proposed TACNet with the combined loss:

$$\mathcal{L} = \frac{1}{N_p} (\mathcal{L}_{\text{cls}} + \mathcal{L}_{\text{reg}}) + \frac{1}{N_t} \mathcal{L}_{\text{trans}}, \quad (6)$$

where  $N_p$  and  $N_t$  denote the number of positive samples  $P$  and transitional samples  $T$  respectively. We train each stream of our network in an end-to-end manner. The experiments demonstrate that the branches of classification and transition can be optimized simultaneously.

In the inference phase, TACNet takes video clips as input and output three items: spatial detection boxes, classification scores, and action state scores. To construct spatio-temporal action tubes, we first apply the greedy algorithm [19] on the frame-level detections using category scores to construct candidate tubes. Secondly, we apply action state predictions to perform temporal detection. In the experiments, we find that the action state scores are discontinuous, and hence apply the watershed segmentation algorithm [24] to trim candidate tubes for temporal detection. Besides, inspired by ACT [10], we introduce a micro-tube refinement (MR) procedure which constructs a micro-tube for proposals in the same spatial position of adjacent frames to average the predicted category scores. The score of each box is set to the product of category score and action state score.

## 4. Experiments

### 4.1. Experimental setup

We evaluate TACNet on two datasets: UCF101-24 [21] and J-HMDB [9]. The UCF101-24 dataset contains 3207 videos for 24 classes. Approximate 25% of all the videos are untrimmed. The J-HMDB dataset contains 928 videos with 33183 frames for 21 classes. Videos in this dataset are

Table 1. Performance Comparison on J-HMDB.

Method	Frame-mAP	Video-mAP			
		0.2	0.5	0.75	0.5:0.95
SSD	49.5	60.9	60.1	41.5	33.8
TCD <sup>1</sup>	54.1	65.0	64.5	45.3	35.1
TS <sup>2</sup> +SSD	56.4	70.9	70.3	48.3	42.2
TS+TCD	61.5	74.1	73.4	52.5	44.8
<b>TS+TCD+MR<sup>3</sup></b>	<b>65.5</b>	<b>74.1</b>	<b>73.4</b>	<b>52.5</b>	<b>44.8</b>

<sup>1</sup> TCD: Temporal Context Detector;

<sup>2</sup> TS: Two-Stream;

<sup>3</sup> MR: Micro-tube refinement.

trimmed to actions. Hence we only use it to evaluate the spatial detection of the proposed TACNet.

We apply the metrics of *frame-mAP* and *video-mAP* [5] to evaluate TACNet at frame-level and video-level respectively. The frame-mAP and video-mAP measure the area under the precision-recall curve of detections for each frame and action tubes respectively. Therefore, frame-mAP measures the ability of classification and spatial detection in a single frame, and video-mAP can also evaluate the performance of temporal detection. The prediction is correct when its overlap with ground-truth box/tube above a certain threshold and predicted category is correct. In this paper, we apply constant IoU threshold (0.5) to evaluate frame-mAP and variable IoU thresholds (i.e 0.2, 0.5, 0.75 and 0.5:0.95) to evaluate video-mAP.

We provide some implementation details in TACNet as follows. Input frames are resized to 300x300. The clip size  $L$  is set as 16 for training and inference. The number of flow images  $S$  is set as 5 for each frame. The probability of 2D dropout is set as 0.3. In the training phase, we stack 32 clips as a mini-batch and apply data augmentation of color jitter, cropping, rescaling, and horizontal flipping. We use the hard-negative-mining strategy that only the hardest negatives up to the same quantity of positives are kept to calculate the loss. Following the previous works, we separate training the appearance stream from the motion stream. To train the appearance branch, we set the learning rate as 0.01 for initial learning and decrease it by 10 times in every 20 epochs. We employ a warmup scheme [7] with the learning rate of 0.001 in the training phase. To train the motion branch, we take the parameters from the appearance branch as the initial parameters and set the initial learning rate as 0.001. Furthermore, we fine-tune the fusion network with a consistent learning rate of 0.0001. Besides, we optimize TACNet by stochastic gradient descent (SGD) with momentum of 0.9. In the inference phase, the number of micro-tube frames is set as 8 for micro-tube refinement procedure.

### 4.2. Analysis of Temporal Context Detector

We perform several experiments to evaluate the effectiveness of the proposed temporal context detector un-

Table 2. Performance Comparison on UCF101-24.

Method	F-mAP	Video-mAP			
		0.2	0.5	0.75	0.5:0.95
SSD	65.3	69.1	39.3	16.3	18.4
TCD	67.5	72.2	45.0	17.4	19.4
TS+SSD	66.5	74.3	47.5	19.2	21.0
TS+TCD	68.7	77.3	50.6	20.9	23.0
TS+TCD+BG <sup>1</sup>	68.3	77.0	48.9	20.8	22.4
TS+SSD+TAC <sup>2</sup>	67.1	74.5	49.0	20.1	21.8
TS+TCD+TAC	69.7	77.5	52.9	21.8	24.1
<b>TS+TCD+TAC+MR</b>	<b>72.1</b>	<b>77.5</b>	<b>52.9</b>	<b>21.8</b>	<b>24.1</b>

<sup>1</sup> BG: This method simply treats transitional states as background, and can be treat as hard-mining method;

<sup>2</sup> TAC: Transition-Aware Classifier;

der different configurations on J-HMDB and UCF101-24 dataset. We report the frame-level and video-level performance in Table 1 - 2.

From the results, we can find that temporal context detector significantly outperforms standard SSD on both datasets. On the dataset of J-HMDB, the proposed temporal context detector with a two-stream configuration obtains the improvements of 5.1% and 3.1% the in terms of frame-mAP and video-mAP (IoU threshold is 0.5) respectively. The improvements on UCF101-24 are 2.2% and 3.1% respectively. These results clearly demonstrate that temporal context information can effectively improve performance.

#### 4.3. Analysis of Transition-Aware Classifier

We perform several experiments to evaluate the transition-aware classifier on the untrimmed UCF101-24 dataset. We report the experimental results with different configurations in Table 2, and the per-class performance of untrimmed categories in Table 4. We present some visualized analysis of transition-aware classifier in Figure 4.

In Table 2, the results of improvement prove its effectiveness of the proposed transition-aware classifier with different settings. It should be noted that the hard-mining method (“BG” in the table) leads to a slight performance drop, which means that simply treating transition states as the background is unreasonable. In contrast, the transition-aware classifier achieves significant improvement in terms of both metrics. In particular, it improvements performance by 2.3% in terms of video-mAP when IoU is 0.5. Therefore, the results can clearly demonstrate that it is crucial to define transitional states and the proposed transition-aware classifier can well distinguish these states.

In Table 4, we can find that transition-aware classifier can obtain obvious improvement in the untrimmed videos. Especially, we outperform the baseline without a transitional-aware classifier and ACT by 1.7% and 7.8% in terms of video-mAP respectively. The transition-aware classifier achieves 8% improvement in terms of temporal

detection when only considering temporal IoU (“Temporal” in the table). Therefore, these results demonstrate the capability of a transition-aware classifier for temporal extent detection.

In Figure 4, we take a “Volleyball” action instance as an example to intuitively show the reason for the performance improvement of a transition-aware classifier. We can find that it is difficult to distinguish transitional states by only considering classification scores. However, action state scores can help to trim the actions well temporally. More results can be found in Figure 5. In Figure 5, we can find that the proposed classifier can also distinguish the action states for multiple instances.

Table 4. Per-class performance of untrimmed categories on UCF101-24.

Category	Frame-mAP		Video-mAP				
	Spatio-temporal <sup>1</sup>		Temporal <sup>2</sup>		Tw	To	
	Tw <sup>3</sup>	To <sup>4</sup>	Tw	To			
Basketball	44.0	34.8	5.5	0.3	0.0	25.8	9.3
Dunk	57.1	53.0	18.9	5.3	1.2	88.2	76.6
CliffDiving	74.9	74.4	42.9	45	39.9	84.4	84.4
CricketBowl	39.7	42.4	3.6	0.9	1.1	25.5	14.0
Diving	85.9	82.3	52.1	43.2	26.1	84.7	84.7
GolfSwing	58.9	55.8	65.9	49.5	51.0	70.4	70.4
LongJump	58.6	59.0	50.9	46.2	71.1	68.0	66.2
PoleVault	64.0	63.8	57.2	60.7	44.6	77.4	77.4
TennisSwing	46.4	40.2	2.2	0.1	0.5	9.6	8.0
Volleyball	50.6	44.1	13.4	0.9	0.0	48.8	12.0
Mean-AP	<b>58.0</b>	55.0	<b>31.3</b>	25.2	23.5	<b>58.3</b>	50.3

<sup>1</sup> Spatio-temporal: evaluating the performance in terms of standard video-mAP;

<sup>2</sup> Temporal: evaluating the performance again ground-truth by only considering temporal IoU while without spatial IoU;

<sup>3</sup> Tw: results with transition-aware classifier;

<sup>4</sup> To: results without a transition-aware classifier.

Table 5. Performance of different detector and base model on UCF101-24 dataset without micro-tube refinement.

Method	Model	Frame-mAP	Video-mAP			
			0.2	0.5	0.75	0.5:0.95
SSD	VGG16	69.7	77.5	52.9	21.8	24.1
DSSD	VGG16	70.1	77.5	53.0	22.1	24.5
SSD	Resnet50	72.0	78.9	54.4	23.0	24.8
DSSD	Resnet50	<b>74.6</b>	<b>79.2</b>	<b>54.6</b>	<b>23.3</b>	<b>25.4</b>

#### 4.4. Exploration on advanced backbone

In this section, we explore different detectors and models using our method. We respectively replace the detector and base model with Deconvlution-SSD [3] and Resnet-50 respectively, and show the results in Table 5. All the models are pretrained on the ImageNet. Two streams, temporal context detector, and transition-aware classifier are also ap-

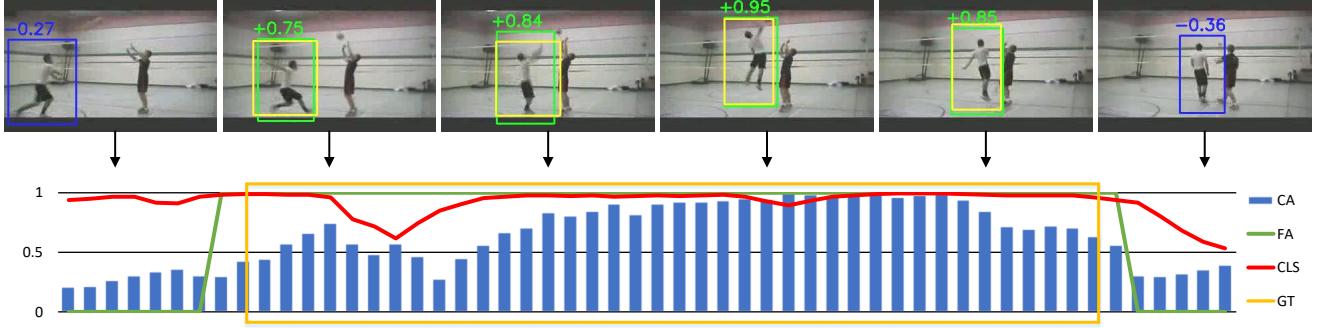


Figure 4. Visualized analysis of transition-aware classifier when taking “Basketball pitch” action as an example. Top row: detection boxes with their corresponding action state scores; bottom row: the procedure with different predictions including Coarse action state score (CA), Refined action state score (FA), Classification score (CLS), and compared to Ground-truth (GT). Based on these predictions, we temporally trim the detection whose FA score is larger than 0.5, while the others are treated as transitional samples.

Table 3. Comparison with the state-of-the-art on J-HMDB (trimmed) and UCF101 (untrimmed)

Method	J-HMDB				UCF101-24 (Full) <sup>1</sup>				UCF101-24 (Untrimmed) <sup>1</sup>	
	F-mAP	Video-mAP			F-mAP	Video-mAP			F-mAP	Video-mAP
		0.2	0.5	0.75		0.2	0.5	0.75		0.5
Saha [17]	-	72.6	71.5	43.3	40.04	-	66.7	35.9	7.9	14.4
Peng [15]	58.5	74.3	73.1	-	-	65.7	73.5	32.1	2.7	7.3
Singh [19]	-	73.8	72.0	44.5	41.6	-	73.5	46.3	15.0	20.4
Hou [8]	61.3	<b>78.4</b>	<b>76.9</b>	-	-	41.4	47.1	-	-	-
Becattini [1]	-	-	-	-	-	-	67.0	35.7	-	-
Kalogeiton [10]	<b>65.7</b>	74.2	73.7	52.1	44.8	69.5	76.5	49.2	19.7	23.4
Ours	65.5	74.1	73.4	<b>52.5</b>	<b>44.8</b>	<b>72.1</b>	<b>77.5</b>	<b>52.9</b>	<b>21.8</b>	<b>24.1</b>
									<b>52.1</b>	<b>23.5</b>
										<b>58.0</b>
										<b>31.3</b>

<sup>1</sup> UCF101-24 is a mixture dataset which is made up of untrimmed categories and trimmed categories, thus we evaluate our approaches in two criterions to fully illustrate the performance gain on untrimmed videos.

plied. We can find that the employment of advanced models can further improve performance.

#### 4.5. Comparison with the state-of-the-art

We compare TACNet with state-of-the-art methods in terms of both frame-mAP and video-mAP, and the results are shown in Table 3. From the results, we see that TACNet outperforms all these previous methods on temporally untrimmed UCF101-24 dataset in terms of both metrics. In particular, we surpass ACT, which is the current state-of-the-art, by 3.7% in terms of video-level when the IoU threshold is 0.5. On the trimmed J-HMDB dataset, ACT and T-CNN outperform TACNet in terms of frame-mAP and video-mAP respectively. We think there are two reasons: (i) these two methods directly generate action tubes, which are suitable for the trimmed dataset; (ii) J-HMDB dataset is relatively simple, especially for localization since only one instance in each frame. However, for action detection, video-mAP is more suitable to evaluate the performance than frame-mAP in both spatial and temporal domain, and in terms of video-mAP TACNet obtains a competitive performance with ACT. On the more challenging untrimmed

UCF101-24 dataset, we can find that TACNet significantly outperforms T-CNN and ACT. It even improves the performance against T-CNN by 28.5% and 30.4% in terms of frame-mAP and video-mAP when IoU is 0.5. Therefore, the superior performance of TACNet demonstrates the importance of long-term temporal context information and transitional state detection.

#### 5. Conclusions

This paper aims to improve the performance of action detection. In particular, we find that it is critical to extract long-term temporal context information and distinguish transitional states. Based on these observations, we propose a TACNet which consists of a temporal context detector and a transitional-aware classifier. We extensively explore TACNet on two public datasets. From the experimental results, we find TACNet can significantly improve the performance and surpass the state-of-the-art on the challenging untrimmed dataset. The performance improvements of TACNet come from both temporal detection and transition-aware method. In future work, we will continue our exploration on how to further improve temporal

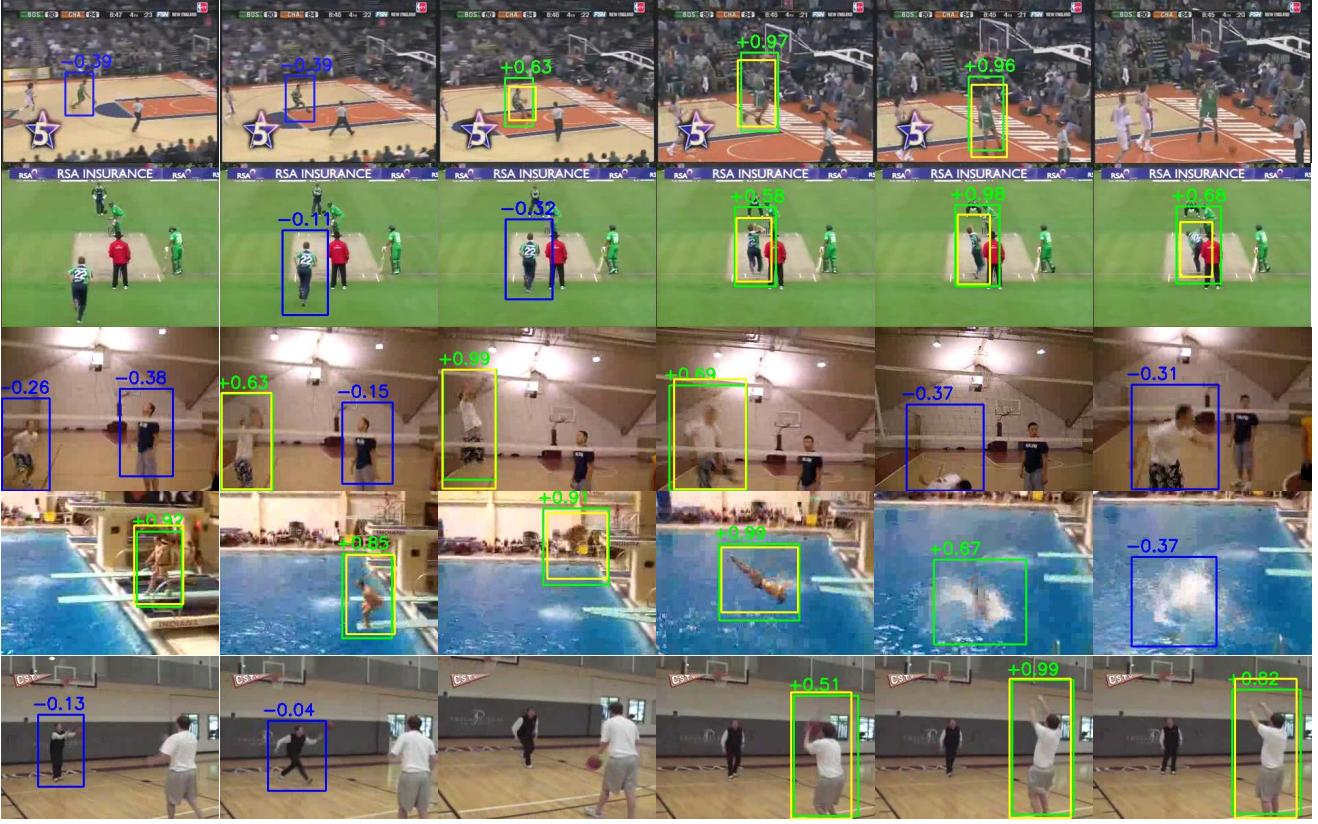


Figure 5. Predicted results of the proposed TACNet in four states: (a) Background (w/o box), (b) Transitional state (blue box), (c) Active state (green box) and (d) Ground-truth (yellow box).

detection by considering the relation between actors and their surrounding peoples or objects.

## 6. Acknowledgment

This research was supported by the National Key R&D Program of China (No. 2017YFA0700800).

## References

- [1] F. Becattini, T. Uricchio, L. Ballan, L. Seidenari, and A. Del Bimbo. Am i done? predicting action progress in videos. *arXiv preprint arXiv:1705.01781*, 2017.
- [2] D. E. Bockelman and W. R. Eisenstadt. Combined differential and common-mode scattering parameters: Theory and simulation. *IEEE transactions on microwave theory and techniques*, 43(7):1530–1539, 1995.
- [3] C.-Y. Fu, W. Liu, A. Ranga, A. Tyagi, and A. C. Berg. DSSD: Deconvolutional single shot detector. *arXiv preprint arXiv:1701.06659*, 2017.
- [4] R. Girshick. Fast r-cnn. In *ICCV*, pages 1440–1448, 2015.
- [5] G. Gkioxari and J. Malik. Finding action tubes. In *CVPR*, pages 759–768, 2015.
- [6] R. Gokberk Cinbis, J. Verbeek, and C. Schmid. Multi-fold mil training for weakly supervised object localization. In *CVPR*, pages 2409–2416, 2014.
- [7] P. Goyal, P. Dollár, R. Girshick, P. Noordhuis, L. Wesolowski, A. Kyrola, A. Tulloch, Y. Jia, and K. He. Accurate, large minibatch sgd: training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*, 2017.
- [8] R. Hou, C. Chen, and M. Shah. Tube convolutional neural network (t-cnn) for action detection in videos. In *ICCV*, 2017.
- [9] H. Jhuang, J. Gall, S. Zuffi, C. Schmid, and M. J. Black. Towards understanding action recognition. In *ICCV*, pages 3192–3199, 2013.
- [10] V. Kalogeiton, P. Weinzaepfel, V. Ferrari, and C. Schmid. Action tubelet detector for spatio-temporal action localization. In *ICCV*, 2017.
- [11] Z. Li, K. Gavrillyuk, E. Gavves, M. Jain, and C. G. Snoek. Videolstm convolves, attends and flows for action recognition. *Computer Vision and Image Understanding*, 166:41–50, 2018.
- [12] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg. Ssd: Single shot multibox detector. In *ECCV*, pages 21–37. Springer, 2016.
- [13] P. Mettes, C. G. Snoek, and S.-F. Chang. Localizing actions from video labels and pseudo-annotations. *arXiv preprint arXiv:1707.09143*, 2017.
- [14] P. Mettes, J. C. van Gemert, and C. G. Snoek. Spot on: Action localization from pointly-supervised proposals. In

*ECCV*, pages 437–453. Springer, 2016.

- [15] X. Peng and C. Schmid. Multi-region two-stream r-cnn for action detection. In *ECCV*, pages 744–759, 2016.
- [16] S. Saha, G. Singh, and F. Cuzzolin. Amtnet: Action-micro-tube regression by end-to-end trainable deep architecture. *ICCV, Oct*, 2, 2017.
- [17] S. Saha, G. Singh, M. Sapienza, P. H. Torr, and F. Cuzzolin. Deep learning for detecting multiple space-time action tubes in videos. *arXiv preprint arXiv:1608.01529*, 2016.
- [18] S. Sharma, R. Kiros, and R. Salakhutdinov. Action recognition using visual attention. *arXiv preprint arXiv:1511.04119*, 2015.
- [19] G. Singh, S. Saha, M. Sapienza, P. Torr, and F. Cuzzolin. Online real-time multiple spatiotemporal action localisation and prediction. In *CVPR*, pages 3637–3646, 2017.
- [20] P. Siva and T. Xiang. Weakly supervised action detection. In *BMVC*, volume 2, page 6, 2011.
- [21] K. Soomro, A. R. Zamir, and M. Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.
- [22] P. Weinzaepfel, Z. Harchaoui, and C. Schmid. Learning to track for spatio-temporal action localization. In *ICCV*, pages 3164–3172, 2015.
- [23] G. Yu and J. Yuan. Fast action proposals for human action detection and search. In *CVPR*, pages 1302–1311, 2015.
- [24] Y. Zhao, Y. Xiong, L. Wang, Z. Wu, X. Tang, and D. Lin. Temporal action detection with structured segment networks. In *ICCV*, volume 8, 2017.
- [25] H. Zhu, R. Vial, and S. Lu. Tornado: A spatio-temporal convolutional regression network for video action proposal. In *CVPR*, pages 5813–5821, 2017.