

An End-to-end 3D Convolutional Neural Network for Action Detection and Segmentation in Videos

Rui Hou, *Student Member, IEEE*, Chen Chen, *Member, IEEE*, and Mubarak Shah, *Fellow, IEEE*

Abstract—Deep learning has been demonstrated to achieve excellent results for image classification and object detection. However, the impact of deep learning on video analysis (e.g. action detection and recognition) has not been that significant due to complexity of video data and lack of annotations. In addition, training deep neural networks on large scale video datasets is extremely computationally expensive. Previous convolutional neural networks (CNNs) based video action detection approaches usually consist of two major steps: frame-level action proposal generation and association of proposals across frames. Also, most of these methods employ two-stream CNN framework to handle spatial and temporal features separately. In this paper, we propose an end-to-end 3D CNN for action detection and segmentation in videos. The proposed architecture is a unified deep network that is able to recognize and localize action based on 3D convolution features. A video is first divided into equal length clips and next for each clip a set of tube proposals are generated based on 3D CNN features. Finally, the tube proposals of different clips are linked together and spatio-temporal action detection is performed using these linked video proposals. This top-down action detection approach explicitly relies on a set of good tube proposals to perform well and training the bounding box regression usually requires a large number of annotated samples. To remedy this, we further extend the 3D CNN to an encoder-decoder structure and formulate the localization problem as action segmentation. The foreground regions (*i.e.* action regions) for each frame are segmented first then the segmented foreground maps are used to generate the bounding boxes. This bottom-up approach effectively avoids tube proposal generation by leveraging the pixel-wise annotations of segmentation. The segmentation framework also can be readily applied to a general problem of video object segmentation. Extensive experiments on several video datasets demonstrate the superior performance of our approach for action detection and video object segmentation compared to the state-of-the-arts.

Index Terms—Action detection, action segmentation, CNN, 3D convolutional neural networks, deep learning, tube proposal

1 INTRODUCTION

THE goal of action detection is to detect every occurrence of a given action within a long video, and to localize each detection both in space and time. Deep learning learning based approaches have significantly improved video action recognition performance. Compared to action *recognition*, action *detection* is a more challenging task due to flexible volume shape and large spatio-temporal search space.

Previous deep learning based action detection approaches first detect frame-level action proposals by popular proposal algorithms [1], [2] or by training proposal networks [3]. Then the frame-level action proposals are associated across frames to determine final action detection through tracking based approaches. Moreover, in order to capture both spatial and temporal information of an action, two-stream networks (a spatial CNN and an optical flow CNN) are used. In this manner, the spatial and motion information are processed separately.

Region Convolution Neural Network (R-CNN) for object detection in images was proposed by Girshick *et al.* [4]. It was followed by a fast R-CNN proposed in [5], which includes the classifier as well. Later, Faster R-CNN [6] was developed by introducing a region proposal network. It has been extensively used to produce excellent results for object detection in images. A natural generalization of the R-CNN from 2D images to 3D spatio-temporal volumes is to study their effectiveness for the problem of action detection in videos. A straightforward spatio-

temporal generalization of the R-CNN approach would be to treat action detection in videos as a set of 2D image detection using Faster R-CNN. However, unfortunately, this approach does not take the temporal information into account and is not sufficiently expressive to distinguish between actions.

Inspired by Faster R-CNN [6], we propose Tube Convolutional Neural Network (T-CNN) for action detection by leveraging the descriptive power of 3D CNN. To better capture the spatio-temporal information of video, we exploit 3D CNN since it is able to capture motion characteristics in videos and shows promising results on video action recognition. In our approach, an input video is divided into equal length clips first. Then, the clips are fed into Tube Proposal Network (TPN) and a set of tube proposals are obtained. Next, tube proposals from each video clip are linked according to their actionness scores and overlap between adjacent proposals to form a complete tube proposal for spatio-temporal action localization in the video. Finally, the Tube-of-Interest (ToI) pooling is applied to the linked action tube proposal to generate a fixed size feature vector for action label prediction.

Our T-CNN is generalization of Faster R-CNN. However, Faster R-CNN relies on a set of anchor boxes with different sizes and aspect ratios, and applies them in a sliding window fashion for object detection. Bounding box regression is used to refine the anchor boxes in terms of position and size for setting a tight boundary around the object. Such object detection methods are considered as top-down detectors, which are similar in spirit to the top-down, class-specific segmentation algorithms [7], [8], [9] which use the shape of the deformed model of a known object to estimate the desired segmentation. Although top-down detection methods have achieved promising results, they still face several

• The authors are with the Center for Research in Computer Vision, University of Central Florida, Orlando, FL, 32816. E-mail: houray@gmail.com; chenchen870713@gmail.com; shah@crcv.ucf.edu

issues. First, bounding boxes only provide coarse localization information as compared to pixel-wise segmentation map. Second, to train bounding box regression, a large amount of training data is necessary. Also, since human actions are complex with dynamic visual patterns, it would require variable templates for each frame of a video for top-down algorithms, leading to the challenge of preparing a set of good anchor boxes priors.

Low-level features like intensity, orientation, and motion can render measurement of actionness for every pixel in image. Therefore, we also propose to explore an alternative strategy for action detection using bottom-up action segmentation. Assuming pixel-wise annotations (action foreground and background) are available, an encoder-decoder network structure, which is commonly used for image semantic segmentation [10], using 3D CNN is developed to predict pixel-wise labels for each video frame. Then the pixel-wise predictions (*i.e.* action segmentation map) for each frame are used to infer the corresponding bounding boxes by finding the boundary of the foreground map. After bounding box generation, ToI pooling is applied to feature tubes for action recognition. We dub the segmentation based approach as Segmentation T-CNN (ST-CNN). The general frameworks of T-CNN and ST-CNN are compared in Figure 1. The main difference between ST-CNN and T-CNN is that ST-CNN avoids generating tube proposals by treating action detection as a binary (*i.e.* foreground action and background) video segmentation task.

To the best of our knowledge, this is the first work to exploit 3D CNN for video action detection (*i.e.* localization and recognition) in an end-to-end fashion. As an extension to our ICCV paper [11], the main contributions of this paper are summarized as follows. First, we improve the T-CNN framework by proposing a bottom-up approach, ST-CNN, which leverages pixel-wise action segmentation for action detection. Second, we also use ST-CNN for video object segmentation and show competitive results comparing to the state-of-the-arts. Third, we conduct a thorough comparison between T-CNN and ST-CNN on action detection, and provide insights in terms of performance and generalization ability. Moreover, through extensive experiments, we demonstrate that our end-to-end 3D CNN approach has the capacity to achieve superior performance over the popular two-stream network pipeline for video analysis tasks, *e.g.* action detection and video object segmentation.

The remainder of this paper is organized as follows. Section 2 reviews related work on action recognition and detection. Section 3 discusses the challenges of generalizing R-CNN from 2D to 3D. Section 4 provides detailed procedures of the proposed T-CNN. In Section 5, we introduce the bottom-up approach, ST-CNN, for action detection and segmentation. The experimental results on action detection and video object segmentation are presented in Section 6. Finally, Section 7 concludes the paper.

2 RELATED WORK

Object detection. For object detection in images, Girshick *et al.* propose Region CNN (R-CNN) [4]. In their approach region proposals are extracted using selective search. Then the candidate regions are warped to a fixed size and fed into CNN to extract CNN features. Finally, SVM model is trained for object classification. A fast version of R-CNN, Fast R-CNN, is presented in [5]. Compared to the multi-stage pipeline of R-CNN, fast R-CNN incorporates object classifier in the network and trains object classifier and bounding box regressor simultaneously. Region of

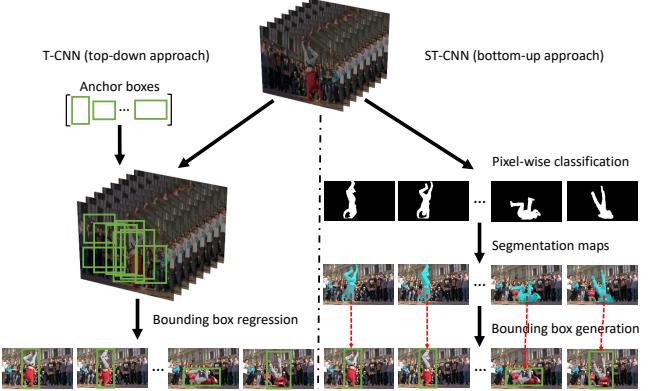


Fig. 1. Top-Down (T-CNN) and Bottom-Up (ST-CNN) approaches. The top-down approach takes the whole frame as the search space with some predefined bounding box templates, filters candidate bounding box proposals based on their confidence scores, and performs bounding box regression on the selected proposals. While, the Bottom-Up approach directly operates on pixels. It takes only segments within frames as candidates and agglomerates them as bounding box detection.

interest (RoI) pooling layer is introduced to extract fixed-length feature vectors for bounding boxes with different sizes. After that, Faster R-CNN [6] introduces a Region Proposal Network (RPN) to replace selective search for proposal generation. RPN shares full image convolutional features with the detection network, thus the proposal generation is almost cost-free. A recent extension of Faster R-CNN is the Mask R-CNN [12], which adds another branch for predicting object mask within each bounding box detection. SSD (Single Shot MultiBox Detector) [13] and YOLO (you only look once) [14], [15] detectors eliminate proposal generation and detect objects in a single network, thus greatly improving the detection speed. These state-of-the-art object detectors are served as the basis for recent action detection methods [3], [16], [17].

3D CNN. Convolutional neural networks have been demonstrated to achieve excellent results for action recognition [18]. Karpathy *et al.* [19] explore various frame-level fusion methods over time. Ng *et al.* [20] use recurrent neural network employing the CNN feature. Since these approaches only use frame based CNN features, the temporal information is neglected. Simonyan *et al.* [21] propose the two-stream CNN approach for action recognition. Besides a classic CNN which takes images as an input, it has a separate network for optical flow. Moreover, Wang *et al.* fuse the trajectories and CNN features. Although these methods, which take hand-crafted temporal feature as a separate stream, show promising performance on action recognition, they do not employ end to end deep network and require separate computation of optical flow and optimization of the parameters. 3D CNN is a logical solution to this issue. Ji *et al.* [22] propose a 3D CNN based human detector and head tracker to segment human subjects in videos. Sun *et al.* [23] propose a factorization of 3D CNN and exploit multiple ways to decompose convolutional kernels. Tran *et al.* [24] use 3D CNN for large scale action recognition. Leveraging the 3D CNN model [24] trained on Sport1M video dataset [19], the 3D CNN framework has been exploited for temporal action localization in untrimmed videos [25], [26], [27], [28], [29]. However, to the best of our knowledge, we are the first to exploit 3D CNN for action *detection*.

Action detection. Compared to action recognition, action detection is a more challenging problem [30], which has been an

active area of research. Before the deep learning era, the top-down based approaches dominated the topic first. Ke *et al.* [31] present an approach for event detection in crowded videos by matching event models to over-segmented spatio-temporal volumes. Lan *et al.* [32] treat the spatial location of human in the video as a latent variable, and perform recognition and detection simultaneously. Tian *et al.* [33] develop spatio-temporal deformable parts model [34] to detect actions in videos. Oneata *et al.* [35] and Desai *et al.* [36] propose sliding-windows based approaches.

Another line of research explores the bottom-up strategy. Bottom-up based approaches replace the exhaustive search, which is a common strategy in top-down approaches, with selective search. The selective search has a much smaller search space than the exhaustive one. In other words, selective search based approaches take less time to perform “search” in frames. Therefore, with the same processing time, these approaches are able to take advantage of more advanced feature representations such as Fisher vector or VLAD instead of basic HOG and HOF. Ma *et al.* [37] use a linear SVM on a bag of hierarchical space-time segments representation. They build a spatio-temporal feature pyramid followed by latent SVM. Jain *et al.* [38] expand the 2D super-pixels to 3D super-voxels. Super-voxels are merged according to appearance and motion costs to produce multiple segmentation layers. Authors in [39] and [40] over segment a video into super-voxels, which are merged into action proposals. The proposals are used to determine action labels for detection. Lu *et al.* [41] propose a hierarchical Markov Random Field (MRF) model to connect different levels of the super-voxel hierarchies for action detection.

The success of deep learning algorithms, particularly CNN, in images paved the way to use them for action detection in video. Authors in [1] extract frame-level action proposals using selective search and link them using Viterbi algorithm. While in [2] frame-level action proposals are obtained by EdgeBox [42] and linked by a tracking algorithm. Motivated by Faster R-CNN [6], two-streams R-CNNs for action detection is proposed in [3], [17], where a spatial RPN and a motion RPN are used to generate frame-level action proposals. Kalogeiton *et al.* [43] extend the spatial only proposal to spatio-temporal proposal by inducing action tubelet. Singh *et al.* [16] adopt SSD to perform online spatio-temporal action localization in real-time. However, these deep learning based approaches treat the spatial and temporal features of a video separately by training two-stream 2D CNN. Therefore, the temporal consistency in videos is not well explored in the network. In contrast, in the proposed T-CNN, we determine action tube proposals directly from input videos and extract compact and more effective spatio-temporal features using 3D CNN.

It is also worth noting that the existing deep learning based approaches operate in a top-down fashion, which requires exhaustive search in the whole frame followed by narrowing down to several appropriate bounding boxes. Clearly, a bottom-up based approach is more efficient. We propose a segmentation based action detection approach (ST-CNN), which accumulates foreground pixels together to form segmentation regions in all frames. And we are the first to generalize the encoder-decoder style network from 2D to 3D for action detection and segmentation.

Video object segmentation. Video object segmentation aims to delineate the foreground object(s) from the background in all frames. Several CNN based approaches have been proposed for this task. Existing methods [44], [45], [46] leverage two-stream pipeline to train a RNN [47] model or learn motion patterns for

TABLE 1

Network architecture of T-CNN. We refer kernel with shape $d \times h \times w$ where d is the kernel depth, h and w are height and width. Output matrix with shape $C \times D \times H \times W$ where C is the number of channels, D is the number of frames (depth), H and W are the height and width of feature maps.

name	kernel dims ($d \times h \times w$)	output dims ($C \times D \times H \times W$)
conv1	$3 \times 3 \times 3$	$64 \times 8 \times 300 \times 400$
max-pool1	$1 \times 2 \times 2$	$64 \times 8 \times 150 \times 200$
conv2	$3 \times 3 \times 3$	$128 \times 8 \times 150 \times 200$
max-pool2	$2 \times 2 \times 2$	$128 \times 4 \times 75 \times 100$
conv3a	$3 \times 3 \times 3$	$256 \times 4 \times 75 \times 100$
conv3b	$3 \times 3 \times 3$	$256 \times 4 \times 75 \times 100$
max-pool3	$2 \times 2 \times 2$	$256 \times 2 \times 38 \times 50$
conv4a	$3 \times 3 \times 3$	$512 \times 2 \times 38 \times 50$
conv4b	$3 \times 3 \times 3$	$512 \times 2 \times 38 \times 50$
max-pool4	$2 \times 2 \times 2$	$512 \times 1 \times 19 \times 25$
conv5a	$3 \times 3 \times 3$	$512 \times 1 \times 19 \times 25$
conv5b	$3 \times 3 \times 3$	$512 \times 1 \times 19 \times 25$
toi-pool2	—	$128 \times 8 \times 8 \times 8$
toi-pool5	—	$512 \times 1 \times 4 \times 4$
1x1 conv	—	8192
fc6	—	4096
fc7	—	4096

moving object segmentation [46]. While taking inspiration from these works, we are the first to present a 3D CNN based deep framework for video object segmentation in a fully automatic manner without human selected feature.

3 GENERALIZING R-CNN FROM 2D TO 3D

Generalizing R-CNN from 2D image regions to 3D video tubes is challenging due to the asymmetry between space and time. Different from images which can be cropped and reshaped into a fixed size, videos vary widely in temporal dimension. Therefore, we divide input videos into fixed length (8 frames) clips, so that video clips can be processed with a fixed-size CNN architecture. Also, clip based processing mitigates the cost of GPU memory.

To better capture the spatio-temporal information in video, we exploit 3D CNN for action proposal generation and action recognition. One advantage of 3D CNN over 2D CNN is that it captures motion information by applying convolution in both time and space. Since 3D convolution and 3D max pooling are utilized in our approach, not only in the spatial dimension but also in the temporal dimension, the size of video clip is reduced while relevant information is captured. As demonstrated in [24], the temporal pooling is important for recognition task since it better models the spatio-temporal information of video and reduces background noise. However, the temporal order is lost in the process. That means if we arbitrarily change the order of the frames in a video clip, the resulting 3D max-pooled feature cube will be the same. This is problematic in action *detection*, since it relies on the feature cube to get bounding boxes for the original frames. To this end, we incorporate temporal skip pooling to retain temporal order residing in the original frames. More details are provided in the next section.

Since a video is processed clip by clip, action tube proposals with various spatial and temporal sizes are generated for different clips. These clip proposals need to be linked into a sequence of tube proposal, which is used for action label prediction and localization. To produce a fixed length feature vector, we propose a new pooling layer – **Tube-of-Interest** (ToI) pooling layer. The

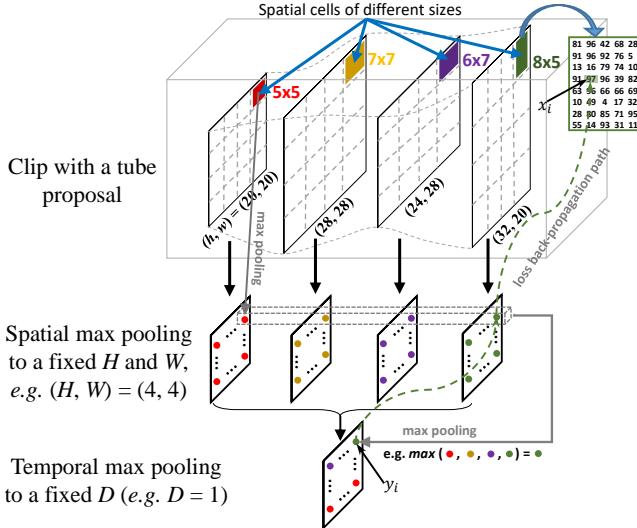


Fig. 2. Tube of interest pooling. In this example four feature maps of different sizes are first pooled spatially to 4×4 then pooled temporally to a fixed size of 1. The cell size for spatial pooling in each feature map varies depending on the size of the feature map. “ \rightarrow ” indicates the loss back-propagation path of a ToI output, y_i (●), to its corresponding input, x_i .

ToI pooling layer is a 3D generalization of Region-of-Interest (RoI) pooling layer of R-CNN. The classic max pooling layer defines the kernel size, stride and padding which determine the shape of the output. In contrast, for RoI pooling layer, the output shape is fixed first, then the kernel size and stride are determined accordingly. Compared to RoI pooling which takes 2D feature map and 2D regions as input, ToI pooling deals with feature cube and 3D tubes. Denote the size of a feature cube as $d \times h \times w$, where d , h and w respectively represent the depth, height and width of the feature cube. A ToI in the feature cube is defined by a d -by-4 matrix, which is composed of d boxes distributed in d feature maps. The boxes are defined by a four-tuple $(x_1^i, y_1^i, x_2^i, y_2^i)$ that specifies the top-left and bottom-right corners in the i -th feature map. Since the d bounding boxes may have different sizes, aspect ratios and positions, in order to apply spatio-temporal pooling, pooling in spatial and temporal domains are performed separately. First, the $h \times w$ feature maps are divided into $H \times W$ bins, where each bin corresponds to a cell with size of approximately $h/H \times w/W$. In each cell, max pooling is applied to select the maximum value. Second, the spatially pooled d feature maps are temporally divided into D bins. Similar to the first step, d/D adjacent feature maps are grouped together to perform the standard temporal max pooling. As a result the fixed output size of ToI pooling layer is $D \times H \times W$. A graphical illustration of ToI pooling is presented in Figure 2.

Back-propagation of ToI pooling layer routes the derivatives from output back to the input as shown in Figure 2. Assume x_i is the i -th activation to the ToI pooling layer, and y_j is the j -th output. Then the partial derivative of the loss function (L) with respect to each input variable x_i can be expressed as:

$$\frac{\partial L}{\partial x_i} = \frac{\partial L}{\partial y_j} \frac{\partial y_j}{\partial x_i} = \sum_j [i = f(j)] \frac{\partial L}{\partial y_j}. \quad (1)$$

Each pooling output y_j has a corresponding input position i . We use a function $f(\cdot)$ to represent the argmax selection from the

ToI. Thus, the gradient from the next layer $\partial L / \partial y_j$ is passed back to only that neuron which achieves the max $\partial L / \partial x_i$.

4 T-CNN FOR ACTION DETECTION

As shown in Figure 3, our T-CNN is an end-to-end deep learning framework that takes video clips as input. The core component is the Tube Proposal Network (TPN) (see Figure 4) to produce tube proposals for each clip. Linked tube proposal sequences are used for spatio-temporal action detection and action recognition in the video.

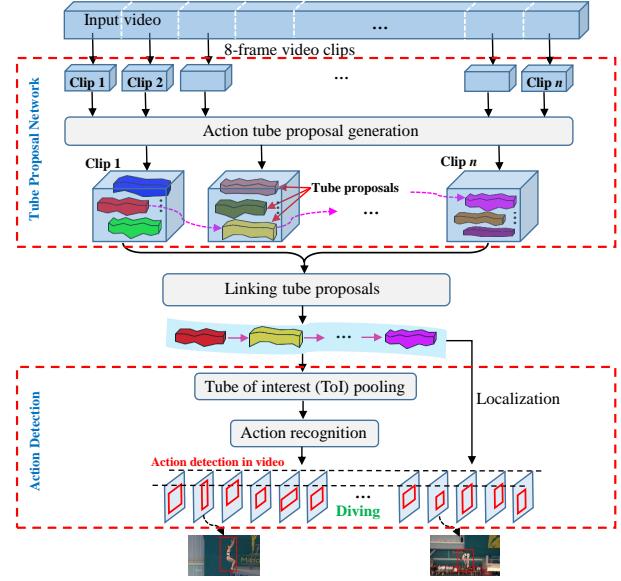


Fig. 3. An overview of the proposed Tube Convolutional Neural Network (T-CNN) for action detection. First, an input video is divided into equal length clips of 8 frame each and fed to Tube Proposal Network to generate tube proposals. Next, these proposals are then linked into larger tubes covering full actions and fed to Action Detection network. Finally, Action Detection network employs ToI pooling to recognize and localize the action.

4.1 Tube Proposal Network

For a 8-frame video clip, 3D convolution and 3D pooling are used to extract spatio-temporal feature cube. In 3D CNN, convolution and pooling are performed spatio-temporally. Therefore, the temporal information of the input video is preserved. Our 3D CNN consists of seven 3D convolution layers and four 3D max-pooling layers. We denote the kernel shape of 3D convolution/pooling by $d \times h \times w$, where d , h , w are depth, height and width, respectively. In all convolution layers, the kernel sizes are $3 \times 3 \times 3$, padding and stride remain as 1. The numbers of filters are 64, 128 and 256 respectively in the first 3 convolution layers and 512 in the remaining convolution layers. The kernel size is set to $1 \times 2 \times 2$ for the first 3D max-pooling layer, and $2 \times 2 \times 2$ for the remaining 3D max-pooling layers. The details of network architecture are presented in Table 1. We use the C3D model [24] as the pre-trained model and finetune it on each dataset in our experiments.

After 3D conv5, the temporal size is reduced to 1 frame (*i.e.* feature cube with depth $D = 1$). In the feature cube, each frame/slice consists of a number of channels specified in Table 1. Here, we drop the number of channels for ease of explanation.

Following Faster R-CNN, we generate bounding box proposals based on the conv5 feature cube¹.

Anchor bounding boxes selection. In Faster R-CNN, the bounding box dimensions are hand picked, *i.e.* 9 anchor boxes with 3 scales and 3 aspect ratios. We can directly adopt the same anchor boxes in our T-CNN framework. However, it has been shown in [15] that if we choose better bounding box priors for the network, it helps the network learn better to predict good detections. Therefore, instead of choosing hand-picked anchor boxes, we apply k-means clustering on bounding boxes in the training set to learn 12 anchor boxes (*i.e.* cluster centers). This data driven anchor box selection approach is adaptive to different datasets.

Each bounding box (bbx) is associated with an “actionness” score, which measures the probability that the bbx corresponds to a valid action. We assign a binary class label (of being an action or not) to each bounding box. Bounding boxes with actionness scores smaller than a threshold are discarded. In the training phase, the bbx which has an Intersection-over-Union (IoU) overlap higher than 0.7 with any ground-truth bbx, or has the highest IoU overlap with a ground-truth box (the later condition is considered in case the former condition may find no positive sample) is taken as a positive bounding box proposal.

Temporal skip pooling. Bounding box proposals generated from conv5 feature tube can be used for frame-level action detection by bounding box regression. However, due to temporal max pooling (8 frames to 1 frame), the temporal order of the original 8 frames is lost. Therefore, we use temporal skip pooling to inject the temporal order for frame-level detection. Specifically, we map each positive bounding box generated from conv5 feature cube into conv2 feature cube which has 8 feature frames/slices. Since these 8 feature slices correspond to the original 8 frames in a video clip, the temporal order is preserved. As a result, if there are 5 bounding boxes in conv5 feature cube for example, 5 scaled bounding boxes are mapped to each conv2 feature slice at the corresponding locations. This creates 5 tube proposals as illustrated in Figure 4, which are paired with the corresponding 5 bounding box proposals for frame-level action detection. To form a fixed feature maps, ToI pooling is applied to the variable size tube proposals as well as the bounding box proposals. Since a tube proposal covers 8 frames in Conv2, the ToI pooled bounding box from Conv5 is duplicated 8 times to form a tube. We then apply L2 normalization to the paired two tubes, and vectorize and concatenate them. Since we use the C3D model [24] as the pre-trained model, we apply a 1x1 convolution to match the input dimension of fc6. Three fully-connected (FC) layers process each descriptor and produce the output: displacement of height, width and 2D center of each bounding box (“bbox”) in each frame. The regression loss measures the differences between ground-truth and predicted bounding boxes, represented by a 4D vector (Δ_{center_x} , Δ_{center_y} , Δ_{width} , Δ_{height}). The sum of them for all bounding boxes is the regression loss of the whole tube. Finally, a set of refined tube proposals by adding the displacements of height, width and center are generated as an output from the TPN representing potential spatio-temporal action localization of the input video clip.

1. Since the depth of conv5 feature cube is reduced to 1, the ToI pooling after that essentially reduces to RoI pooling, a special case of ToI. For consistency, we use ToI throughout the entire framework.

4.2 Linking Tube Proposals

A set of tube proposals are obtained for each video clip after the TPN. We then link these tube proposals to form a sequence of proposals for spatio-temporal action localization of the entire video. Each tube proposal from different clips can be linked in a tube proposal sequence (*i.e.* video tube proposal) for action detection. However, not all combinations of tube proposals can correctly capture the complete action. For example, a tube proposal in one clip may contain the action and a tube proposal in the following clip may only capture the background. Intuitively, the content within the selected tube proposals should capture an action and connected tube proposals in any two consecutive clips should have a large temporal overlap. Therefore, two criteria are considered when linking tube proposals: actionness and overlap scores. Each video proposal is then assigned a score defined as follows:

$$S = \frac{1}{m} \sum_{i=1}^m Actionness_i + \frac{1}{m-1} \sum_{j=1}^{m-1} Overlap_{j,j+1}, \quad (2)$$

where $Actionness_i$ denotes the actionness score of the tube proposal from the i -th clip, $Overlap_{j,j+1}$ measures the overlap between the linked two proposals respectively from the j -th and $(j+1)$ -th clips, and m is the total number of video clips. As shown in Figure 4, each bounding box proposal from conv5 feature cube is associated with an actionness score. These actionness scores are inherited by the corresponding tube proposals from conv2 feature cube. The overlap between two tube proposals is calculated based on the IoU (Intersection Over Union) of the last frame of the j -th tube proposal and the first frame of the $(j+1)$ -th tube proposal. The first term of S computes the average actionness score of all tube proposals in a video proposal and the second term computes the average overlap between the tube proposals in every two consecutive video clips. Therefore, we ensure the linked tube proposals can encapsulate the action and at the same time have temporal consistency. An example of linking tube proposals and computing scores is illustrated in Figure 5. We choose a number of linked proposal sequences with highest scores in a video (see more details in Sec. 6.1).

4.3 Action Detection

After linking tube proposals, we get a set of linked sequences of tube proposal, which represent potential action instances. The next step is to classify these linked tube proposal sequences. The tube proposals in the linked sequences may have different sizes. In order to extract a fixed length feature vector from each of the linked proposal sequence, our proposed ToI pooling is utilized. Then the ToI pooling layer is followed by two FC layers and a drop-out layer. The dimension of the last FC layer is $N + 1$ (N action classes and 1 background class).

5 BOTTOM-UP ACTION DETECTION IN VIDEOS

The proposed T-CNN approach is able to detect actions in videos and put bounding boxes for localization in each frame. As discussed before, T-CNN as a top-down approach relies on exhaustive search in the whole frame and appropriate bounding boxes selection. It has been shown that bottom-up approaches which directly operate on group of pixels *e.g.* through super-voxel or super pixel segmentation are more efficient for action detection. Also, it is obvious that pixel-wise action segmentation

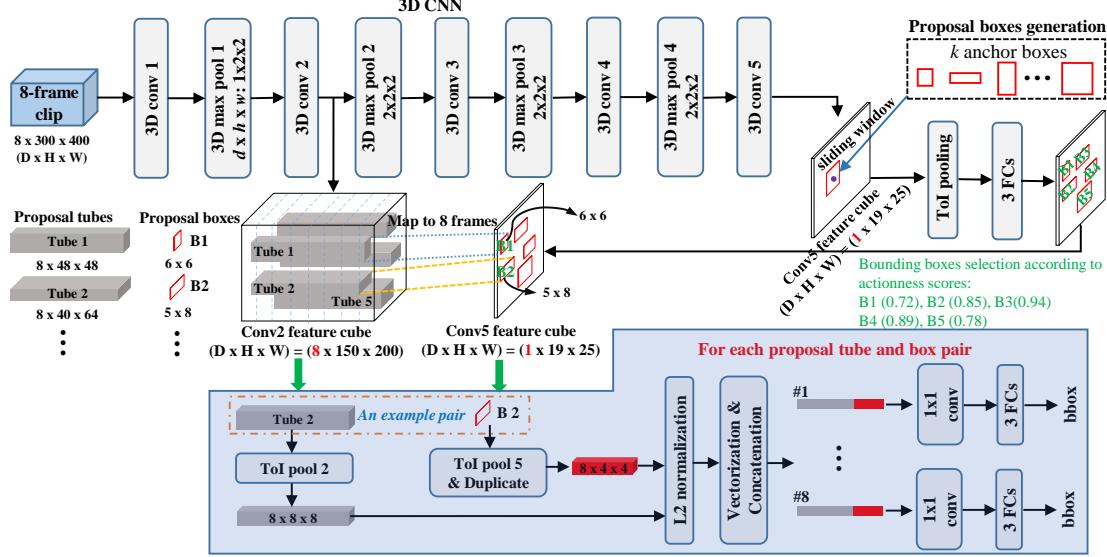


Fig. 4. The tube proposal network (TPN) takes a 8-frame clip as input and applies 3D convolution and max-pooling to extract spatio-temporal features. conv5 feature cube are used to generate bounding box proposals. Those with high actionness scores are mapped to conv2 feature cube (contains 8 frames information) at the corresponding positions to get tube proposals. Each proposal tube and box pair are aggregated after separate ToI pooling. Then bounding box regression is performed for each frame. Note that the channel dimension is omitted in this figure (see Table 1).

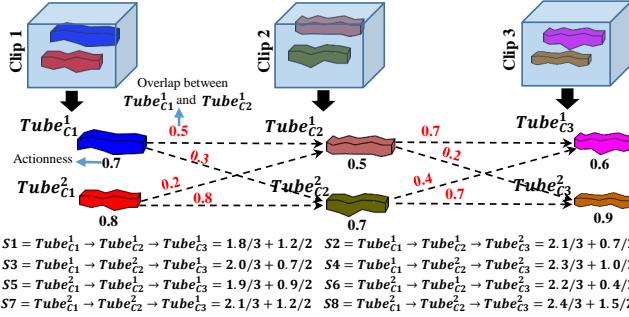


Fig. 5. An example of linking tube proposals in video clips using network flow. In this example, there are three video clips and each has two tube proposals, resulting in 8 video proposals. Each video proposal has a score, e.g., S_1, S_2, \dots, S_8 , which is computed according to Eq. (2).

maps provide finer human silhouettes than bounding boxes, since bounding may also include background pixels.

To achieve this goal, we develop a ST-CNN (Segmentation Tube CNN) approach to automatically localize and segment the silhouette of an actor for action detection. Figure 6 shows the network structure of the proposed ST-CNN. It is an end-to-end 3D CNN, which builds upon an encoder-decoder structure like the SegNet [10] for image semantic segmentation. A video is divided into 8-frame clips as input to the network. On the encoder side, 3D convolution and max pooling are performed. Due to 3D max pooling, the spatial and temporal sizes are reduced. In order to generate the pixel-wise segmentation map for each frame in the original size, 3D upsampling is used in the decoder to increase the resolution of feature maps. To capture spatial and temporal information at different scales, a concatenation with the corresponding feature maps from the encoder is employed after each 3D upsampling layer. Finally, a segmentation branch is used for pixel-wise prediction (*i.e.* background or action foreground) for each frame in a clip. Recognition branch takes the segmentation maps (output of the segmentation branch), where

TABLE 2
Network architecture of ST-CNN. Note that conv6 and conv7 (in the segmentation branch) are applied to each frame of the concat1 layer feature cube ($112 \times 8 \times 240 \times 320$) to produce 8 segmentation maps.

name	kernel dims ($d \times h \times w$)	output dims ($C \times D \times H \times W$)
conv1	$3 \times 3 \times 3$	$64 \times 8 \times 240 \times 320$
max-pool1	$1 \times 2 \times 2$	$64 \times 8 \times 120 \times 160$
conv2	$3 \times 3 \times 3$	$128 \times 8 \times 120 \times 160$
max-pool2	$2 \times 2 \times 2$	$128 \times 4 \times 60 \times 80$
conv3a	$3 \times 3 \times 3$	$256 \times 4 \times 60 \times 80$
conv3b	$3 \times 3 \times 3$	$256 \times 4 \times 60 \times 80$
max-pool3	$2 \times 2 \times 2$	$256 \times 2 \times 30 \times 40$
conv4a	$3 \times 3 \times 3$	$512 \times 2 \times 30 \times 40$
conv4b	$3 \times 3 \times 3$	$512 \times 2 \times 30 \times 40$
max-pool4	$2 \times 2 \times 2$	$512 \times 1 \times 15 \times 20$
conv5a	$3 \times 3 \times 3$	$512 \times 1 \times 15 \times 20$
conv5b	$3 \times 3 \times 3$	$512 \times 1 \times 15 \times 20$
upsample4	$3 \times 3 \times 3$	$64 \times 2 \times 30 \times 40$
conv4c	$3 \times 3 \times 3$	$448 \times 2 \times 30 \times 40$
upsample3	$3 \times 3 \times 3$	$64 \times 4 \times 60 \times 80$
conv3c	$3 \times 3 \times 3$	$448 \times 4 \times 60 \times 80$
upsample2	$3 \times 3 \times 3$	$64 \times 8 \times 120 \times 160$
conv2c	$3 \times 3 \times 3$	$128 \times 8 \times 120 \times 160$
upsample1	$3 \times 3 \times 3$	$48 \times 8 \times 240 \times 320$
conv1c	$3 \times 3 \times 3$	$64 \times 8 \times 240 \times 320$
conv6	1×1	$4096 \times 8 \times 240 \times 320$
conv7	1×1	$2 \times 8 \times 240 \times 320$
toi-pool	—	$112 \times 8 \times 8 \times 8$
fc6	—	4096
fc7	—	4096

the foreground segmentation maps (action regions) are converted into bounding boxes, and the feature cube of the last concatenation layer “concat1”, to extract the feature tube of the action volume. ToI pooling is applied to the feature tube and followed by three FC layers for action recognition. The detailed network configuration is presented in Table 2.

Compared to T-CNN, ST-CNN is a bottom-up approach and has the following advantages for action detection.

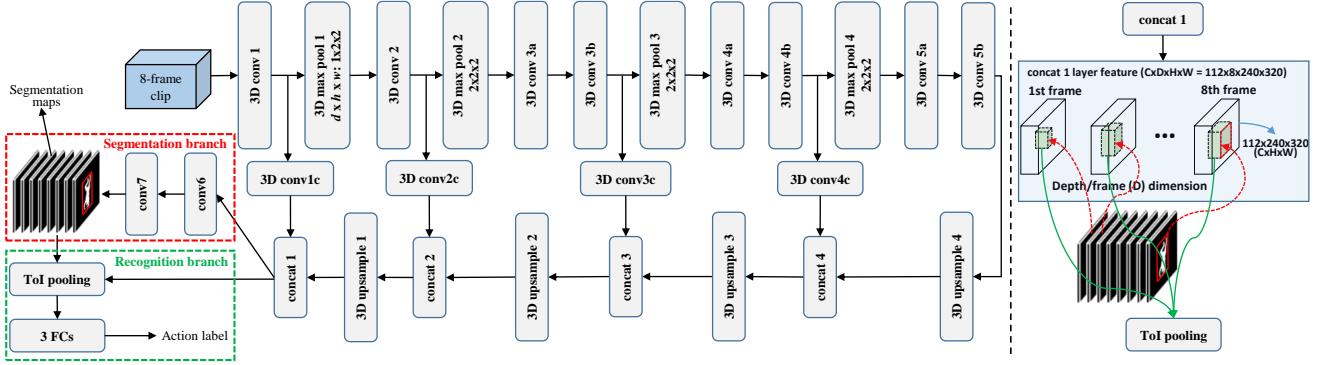


Fig. 6. The framework of ST-CNN for action detection. An encoder-decoder 3D CNN architecture is employed. The segmentation branch produces a binary segmentation map (action foreground vs. background) for each frame of the input video clip. The foreground pixels are used to infer frame-wise bounding boxes (action localization), based on which the feature tube is extracted from “concat1” feature cube as an input to the recognition branch (see the right part of this figure for details).

- ST-CNN provides pixel-level localization for actions (*i.e.* action segmentation maps), which are more accurate than bounding boxes for action localization. Moreover, bounding boxes can be easily derived from the segmentation maps.

- ST-CNN avoids bounding box regression used in T-CNN, which is hard to converge. Specifically, to learn a sufficient regression model, large amount of data is required. In contrast, less data is required for binary (foreground vs. background) classification and convergence is easier to reach. In our case, lack of annotated training video data causes difficulty in training the bounding boxes regression model. Since ST-CNN uses a classification model, which predicts foreground/background label for each pixel, the number of training samples is drastically increased, since one detection bounding box or action segmentation map includes hundreds of pixels.

Connections between T-CNN and ST-CNN. Although we do not employ TPN on ST-CNN as in T-CNN, both frameworks explicitly use 3D convolutions as building blocks to learn spatio-temporal features for action detection. The encoder part of ST-CNN shares the same network architecture of T-CNN for feature extraction. Moreover, ToI pooling is employed in both frameworks to handle variable bounding box sizes in tubes.

5.1 3D Upsampling Layer

A common method of recovering higher image resolution from max-pooling is to use un-pooling [48] followed by a convolution layer. The locations of the maximum values within max-pooling regions are recorded. The un-pooling layer places the values from the layer above to appropriate positions, according to the recorded max-pooling locations, as an approximation of reverse max-pooling. Deconvolution [49] is another approach to increase the resolution feature map, and has been explored in various tasks such as semantic segmentation [49], optical flow estimation [50], etc. A graphical illustration of the 3D un-pooling followed by 3D convolution is presented in Figure 7.

Compared to un-pooling and deconvolution, sub-pixel convolution proposed in [51] shows better performance for image super-resolution and is more computationally efficient. Therefore, we adopt the sub-pixel convolution approach [51] and generalize it to 3D. An example of the 3D sub-pixel convolution operation to upscale an input low resolution feature cube by a factor of 2 in both spatial and temporal dimensions is demonstrated in Figure 8.

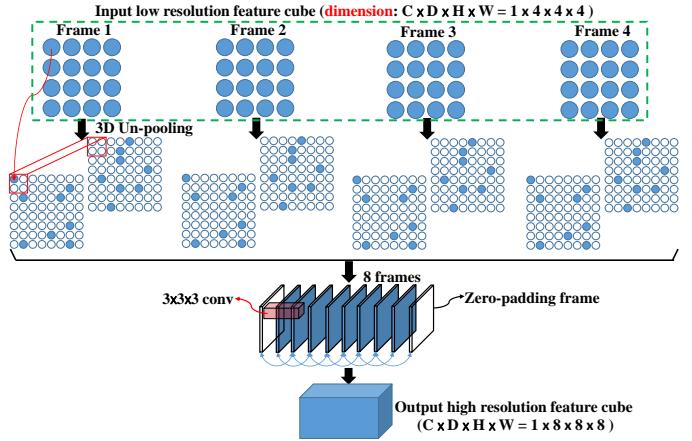


Fig. 7. The 3D “un-pooling + convolution” operation with a upscale factor of 2 in both spatial and temporal dimensions. In this example, the input is a low resolution (LR) input feature cube ($C \times D \times H \times W = 1 \times 4 \times 4 \times 4$). The un-pooling operation is used to generate two upscaled frames ($H \times W = 8 \times 8$) for each input frame. (Note that the un-pooling locations are randomly selected for demonstration purpose.) Then 3D convolution is applied to the resulting 8 frames and 2 zero-padding frames to produce the high resolution (HR) feature cube ($C \times D \times H \times W = 1 \times 8 \times 8 \times 8$).

Formally, suppose we want to upsample the LR feature maps (or cube) by p_d , p_h and p_w times in depth (frame), height and width, respectively. A convolution layer is first applied on the LR feature maps, P^{LR} with dimension $(C \times D \times H \times W = C^L \times D^L \times H^L \times W^L)$, to expand the channels from C^L to $C^H = p_d \times p_h \times p_w \times C^L$. In the example of Figure 8, $C^L = 1$ and the upscale factor is 2, *i.e.* $p_d = 2$, $p_h = 2$ and $p_w = 2$. We denote the channel expanded feature maps as \hat{P}^{LR} . The spatial ($H \times W$) and temporal (D) resolutions of \hat{P}^{LR} are the same as those of P^{LR} . Second, subsequent channel-to-space&depth transpose layer is placed on top of it to upsample the spatial and temporal dimensions by (p_h, p_w) and p_d times, respectively. Let $P_{c,i,j,k}^{HR}$ ($C \times D \times H \times W = C^H \times D^H \times H^H \times W^H$) be the pixel value in the HR feature maps, located at (c, i, j, k) denoting the position (channel, depth/frame, height, width). The pixels in the HR feature maps P^{HR} are mapped from \hat{P}^{LR} according to:

$$P_{c,i,j,k}^{HR} = \hat{P}_{c',i',j',k'}^{LR}, \quad (3)$$

where $c \in \{0, \dots, C^H - 1\}$, $i \in \{0, \dots, D^H - 1\}$, $j \in$

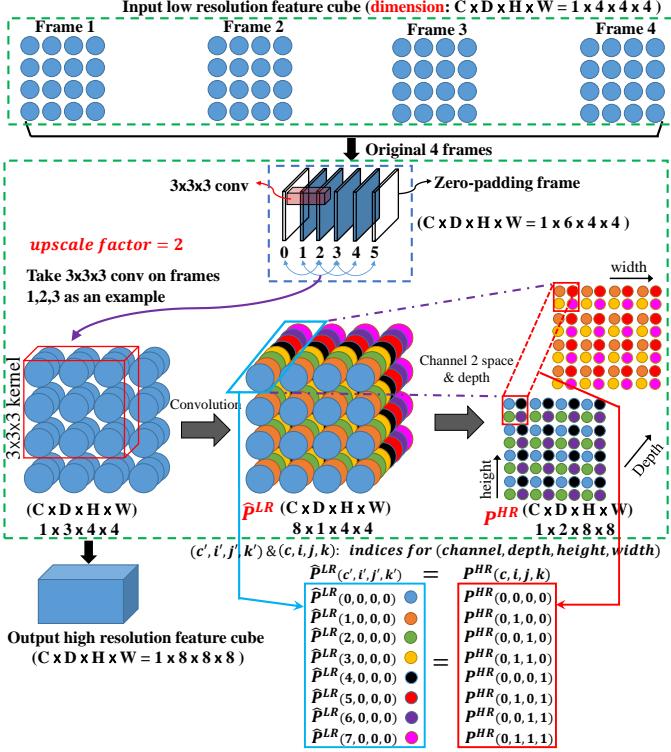


Fig. 8. The 3D sub-pixel convolution approach for feature cube upscaling. In this example, given the LR input feature cube with dimension ($C \times D \times H \times W = 1 \times 4 \times 4 \times 4$). It is first added with two zero-padding frames. Then the $3 \times 3 \times 3$ convolution operation is applied to those frames (i.e. 0-2, 1-3, 2-4, and 3-5). Take the 3D convolution on frames 1-3 as an example. We use 8 3D convolution kernels to generate \hat{P}^{LR} with 8 channels. Then channel to space & depth reshape is performed, following Eqs. 3 and 4, to produce P^{HR} , where both the spatial dimension ($H \times W$) and the temporal dimension (D) are increased by a factor of 2. Therefore, the operations (3D convolution and channel-to-space&depth reshape) on the input frames (0-2, 1-3, 2-4, and 3-5) result in an upscaled feature cube ($C \times D \times H \times W = 1 \times 8 \times 8 \times 8$).

$\{0, \dots, H^H - 1\}$, and $k \in \{0, \dots, W^H - 1\}$. Indices c' , i' , j' and k' for \hat{P}^{LR} are defined as follows:

$$\begin{cases} c' = c \cdot p_d \cdot p_h \cdot p_w + \text{mod}(i, p_d) + p_w \cdot \text{mod}(j, p_h) \\ \quad + p_w \cdot p_h \cdot \text{mod}(k, p_w) \\ i' = \lfloor i/p_d \rfloor \\ j' = \lfloor j/p_h \rfloor \\ k' = \lfloor k/p_w \rfloor \end{cases} \quad (4)$$

Note that 3D convolution is performed in the LR feature space in the sub-pixel convolution approach, whereas it is applied on the un-pooled sparse feature cube (in HR space) in the “un-pooling + convolution” strategy. Therefore, the 3D sub-pixel convolution method is able to integrate more information in 3D convolution for feature cube upscaling.

6 EXPERIMENTS

To verify the effectiveness of the proposed 3D end-to-end deep learning framework for action detection and segmentation, we evaluate our approach on three trimmed video action detection datasets including UCF-Sports [52], J-HMDB [53], UCF-101 [54]; one un-trimmed video action detection dataset – THUMOS’14 [55]; and a video object segmentation dataset – DAVIS’16 [56]. In the following, we provide the implementation details of our

networks, datasets and experiment settings, as well as show experimental comparisons of our method to the state-of-the-arts.

6.1 Implementation Details

6.1.1 T-CNN for action detection

The TPN and recognition network share weights in their common layers. Due to memory limitation, in the training phase, each video is divided into overlapping 8-frame clips with resolution 300×400 and temporal stride 1. When training the TPN, each anchor box is assigned a binary label. An anchor box which has the highest IoU overlap with a ground-truth box, or an anchor box that has an IoU overlap higher than 0.7 with any ground-truth box is assigned a positive label, and others are assigned negative labels. In each iteration, 4 clips are fed into the network. Since the number of background boxes (i.e. negative boxes) is much larger than that of action boxes, we randomly select some of the negative boxes to balance the number of positive and negative samples in a batch. For recognition network training, we choose 40 linked proposal sequences with highest scores in a video as Tubes of Interest.

T-CNN model is trained in an alternative manner. First, we **initialize the TPN** based on the pre-trained model in [24], and use the generated proposals to **initialize the recognition network**. Next, the weights tuned by the recognition network are used to **update the TPN**. Finally, the tuned weights and proposals from the TPN are used for **finalizing the recognition network**. For UCF-Sports and J-HMDB, the learning rate of the networks is initialized as 10^{-3} and decreased to 10^{-4} after 30k batches. Training terminates after 50k batches. For UCF-101 and THUMOS’14, the learning rate is initialized as 10^{-3} and decreased to 10^{-4} after 60k batches. Training terminates after 100k batches.

During testing, each video is divided into non-overlapping 8-frame clips. If the number of frames in the video cannot be divided by 8, we pad zeros after the last frame to make it divisible. 40 tube proposals with highest actionness confidence through TPN are chosen for the linking process. Non-maximum suppression (NMS) is applied to the linked proposals to get the final action detection results.

6.1.2 ST-CNN for action detection

We use the same resolution 300×400 of the input videos without cropping and resizing. The detection task uses both losses (segmentation loss and recognition loss) as shown in Figure 6. In the training phase, similar to T-CNN, videos are divided into overlapping 8-frame clips with temporal stride 1. For the J-HMDB dataset, the ground-truth annotations provide pixel-wise binary labels, where foreground is considered as positive and background as negative. These pixel-wise annotations are used for minimizing the segmentation loss. Once the segmentation is generated for each frame, a bounding box is obtained by including positive pixels (i.e. foreground) in each frame. Bounding boxes in each video clip form tubes of interest. Through ToI max pooling, the feature cube size is fixed for action recognition. The learning rate is initialized at 10^{-4} and decreased to 10^{-5} after 498,000 batches where each batch contains 4 clips. Training terminates after 500,000 batches. During testing, the incoming video is divided into non-overlapping 8-frame clips first. Then, the segmentation branch predict the foreground/background for all the pixels (**action segmentation – localization**). Finally, we take the boundary of predicted foreground pixels in each frame to form the Tube-of-Interest, which is fed into the ToI pooling layer to predict the action label (**action recognition**).

6.1.3 ST-CNN for video object segmentation

Since ST-CNN already includes a segmentation loss, it can be easily applied to the video object segmentation task, *i.e.* segmenting the primary foreground object in each video frame. The benchmark dataset, DAVIS’16 [56], is employed for evaluation. In this experiment, only segmentation loss is considered as shown in Figure 6 since action recognition is not involved. Similar to the detection task, each batch contains 4 clips, but the learning rate is initialized at 10^{-3} and decreased to 10^{-4} after 100,000 batches. The training process terminates after 120,000 batches.

6.2 Action Detection Experimental Results

6.2.1 UCF-Sports

UCF-Sports dataset contains 150 short videos of 10 different sport classes. Videos are trimmed and the action and bounding boxes annotations are provided for all frames. We follow the standard training and test split defined in [32] to carry out the evaluation.

We use the usual IoU criterion and generate ROC curve in Figure 9(a) when overlap equals to $\alpha = 0.2$. Figure 9(b) illustrates AUC (Area-Under-Curve) measured with different overlap criterion. In direct comparison, our T-CNN clearly outperforms all the competing methods shown in the plot. We are unable to directly compare the detection accuracy against Peng *et al.* [3] in the plot, since they do not provide the ROC and AUC curves. As shown in Table 3, the frame level mAP of our approach outperforms theirs in 8 actions out of 10. Moreover, by using the same metric, the video mAP of our approach reaches 95.2 ($\alpha = 0.2$ and 0.5), while they report 94.8 ($\alpha = 0.2$) and 94.7 ($\alpha = 0.5$).

6.2.2 J-HMDB

J-HMDB consists of 928 videos with 21 different actions. All the video clips are well trimmed. There are three train-test splits and the evaluation is done on the average results over the three splits. The experimental results comparison is shown in Table 4. We report our results using 3 metrics: frame-mAP, the average precision of detection at frame level as in [1]; video-mAP, the average precision at video level as in [1] with IoU threshold $\alpha = 0.2$ and $\alpha = 0.5$. It is evident that our T-CNN consistently outperforms the state-of-the-art approaches in terms of all three evaluation metrics.

6.2.3 THUMOS’13 (UCF 101)

UCF-101 dataset with 101 actions is commonly used for action recognition. For action detection task, a subset (THUMOS’13) of 24 action classes and 3,207 videos have spatio-temporal annotations [16]. Similar to other methods, we perform the experiments on the first train/test split only. We report our results in Table 5 with 3 metrics: frame-mAP, video-mAP ($\alpha = 0.2$) and video-mAP ($\alpha = 0.5$). Our approach again yields the best performance. Moreover, we also report the action recognition results of T-CNN on the above three datasets in Table 6.

6.2.4 THUMOS’14

To further validate the effectiveness of our T-CNN approach for action detection, we evaluate it using the untrimmed videos from the THUMOS’14 dataset [55]. The THUMOS’14 spatio-temporal localization task consists of 4 classes of actions: BaseballPitch, golfSwing, TennisSwing and ThrowDiscus. There are about 20 videos per class and each video contains 500 to 3,000 frames. The videos are divided into validation set and test set, but only

videos in the test set have spatial annotations provided by [58]. Therefore, we use samples corresponding to those 4 actions in UCF-101 with spatial annotations to train our model.

In untrimmed videos, there often exist other unrelated actions besides the action of interests. For example, “walking” and “picking up a golf ball” are considered as unrelated actions when detecting “GolfSwing” action. We denote clips which have positive ground truth annotation as positive clips, and the other clips as negative clips (*i.e.* clips containing unrelated actions). If we randomly select negative samples for training, the number of boxes on unrelated actions is much smaller than that for background boxes (*i.e.* boxes capturing only background). Thus the trained model will have no capability to distinguish action of interest and unrelated actions.

To this end, we introduce a so called negative sample mining process. Specifically, when initializing the TPN, we only use positive clips. Then we apply the model on the whole training video (consisting of both positive and negative clips). Most false positives in negative clips should include unrelated actions to help our model learn the difference between action of interest and unrelated actions. Therefore we select boxes in negative clips with highest scores as **hard negatives**, since low scores probably related to the image background and those background clips do not contribute to the model to distinguish between action of interest and unrelated actions. In the TPN updating procedure, we choose 32 boxes which have IoU with any ground truth greater than 0.7 as positive samples and randomly pick another 16 samples as negative. We also select 16 samples from the hard negative pool as negatives. Therefore, we efficiently train a model, which is able to distinguish not only action of interest from background, but also action of interest from unrelated actions.

The mean ROC curves of different methods for THUMOS’14 action detection are plotted in Figure 9(c). Our method without negative mining performs better than the baseline method Sultani *et al.* [58]. Additionally, with negative mining, the performance is further boosted. For qualitative results, we shows examples of detected action tubes in videos from UCF-Sports, JHMDB, UCF-101 (24 actions) and THUMOS’14 datasets (see Figure 10). Each block corresponds to a different video that is selected from the test set. We show the highest scoring action tube for each video.

6.2.5 T-CNN vs. ST-CNN for Action Detection

We evaluate the proposed ST-CNN method on the J-HMDB dataset using two performance measures, *i.e.* bounding box action detection and pixel-wise segmentation maps, since both bounding box and segmentation maps annotations are provided for this dataset. For action detection evaluation, we generate the bounding box for each frame based on the enclosing box of the predicted segmentation map. We report the video mean average precision (mAP) results of our method and several state-of-the-art detection approaches in Table 4. ST-CNN achieves the best results and outperforms T-CNN by 0.8%. We also compare ST-CNN with a baseline method [41] in terms of action segmentation, and our method performs significantly better than the baseline, leading to almost 10% improvement of IoU. Some qualitative action segmentation and action detection results of our ST-CNN approach on the J-HMDB dataset are illustrated in Figure 11. We also show the detection results of T-CNN for comparison. The predicted segmentation maps of ST-CNN are well aligned with the body contours, which results in more accurate bounding boxes than T-CNN (*e.g.* “brush hair”, “catch” and “pick” as shown in

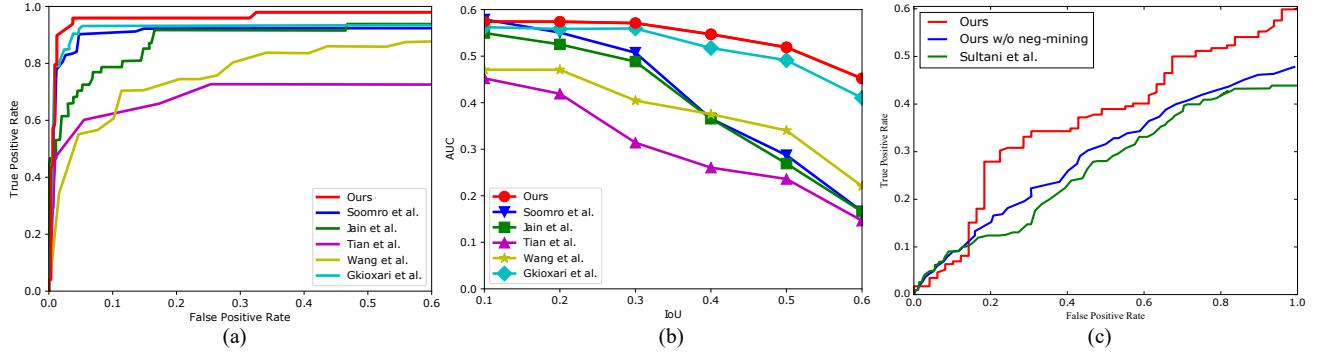


Fig. 9. The ROC and AUC curves for UCF-Sports dataset [52] are shown in (a) and (b), respectively. The results are shown for Jain *et al.* [38] (green), Tian *et al.* [33] (purple), Soomro *et al.* [39] (blue), Wang *et al.* [57] (yellow), Gkioxari *et al.* [1] (cyan) and Proposed Method (red). (c) shows the mean ROC curves for four actions of THUMOS'14. The results are shown for Sultani *et al.* [58] (green), the proposed T-CNN (red) and T-CNN without negative mining (blue).

TABLE 3
mAP for each class of UCF-Sports. The IoU threshold α for frame m-AP is fixed to 0.5.

	Diving	Golf	Kicking	Lifting	Riding	Run	SkateB.	Swing	SwingB.	Walk	mAP
Gkioxari <i>et al.</i> [1]	75.8	69.3	54.6	99.1	89.6	54.9	29.8	88.7	74.5	44.7	68.1
Weinzaepfel <i>et al.</i> [2]	60.71	77.55	65.26	100.00	99.53	52.60	47.14	88.88	62.86	64.44	71.9
Peng <i>et al.</i> [3]	96.12	80.47	73.78	99.17	97.56	82.37	57.43	83.64	98.54	75.99	84.51
Kalogeiton <i>et al.</i> [43]	—	—	—	—	—	—	—	—	—	—	87.7
Ours (T-CNN)	84.38	90.79	86.48	99.77	100.00	83.65	68.72	65.75	99.62	87.79	86.7
Ours (ST-CNN*)	70.9	89.1	90.7	89.7	99.6	71.1	80.4	89.3	86.7	77.5	84.5

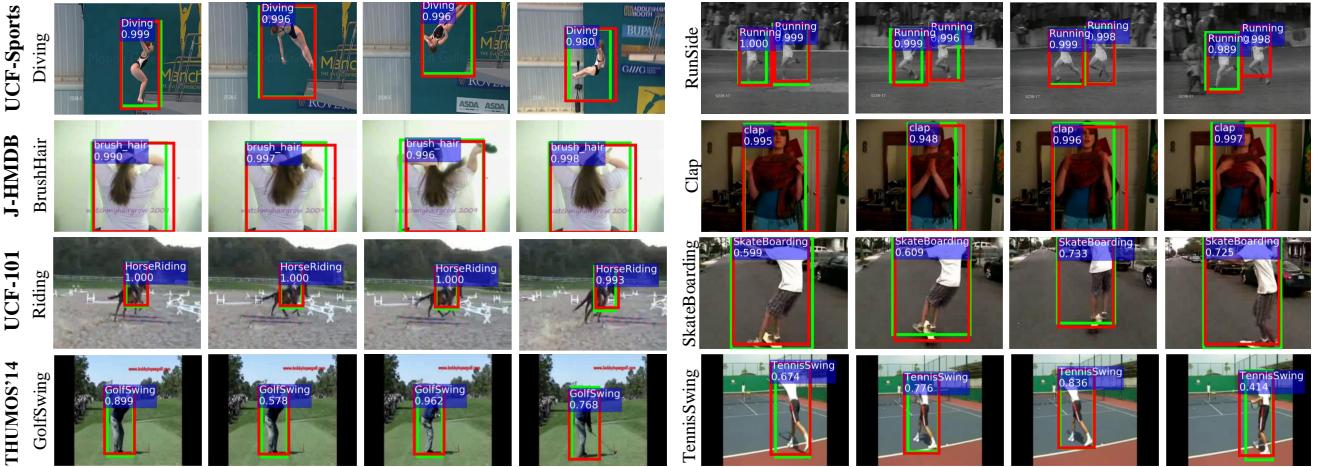


Fig. 10. Action detection results obtained by T-CNN on UCF-Sports, JHMDB, UCF-101 and THUMOS'14. Red boxes show the detections in the corresponding frames, and green boxes show ground truth. The predicted labels are overlaid.

Figure 11). Although the segmentation noise may cause imperfect bounding box predictions (see the last two example frames of “climb”, where incorrect foreground appear on the hand rail), some post-processing techniques such as conditional random fields (CRFs) can be employed to smooth out the segmentation noise to achieve better segmentation maps.

We compare our 3D sub-pixel convolution with the “unpooling + convolution”, denoted by Ours (un-pool) in Table 4. We replace our upsample convolution layer with transpose convolution layer and keep other settings unchanged. By using upsample convolution, we achieve about 5% accuracy gain compared to transpose convolution based approach.

Since the J-HMDB dataset provides the annotations of pixel-

wise segmentation maps, ST-CNN can leverage such pixel-wise labeling information to train the model. However, the other action detection datasets may only have bounding box annotations for actions, such as UCF-Sports. How well the trained ST-CNN model on J-HMDB is able to generalize to other datasets for action detection with no additional bounding box training? To validate its performance, we directly apply the trained ST-CNN on J-HMDB to the UCF-Sports dataset for generating action segmentation maps. The tightest bounding box that covers the foreground segmentation in each frame is considered as the action bounding box prediction. Beside bounding box prediction, since the UCF-Sports dataset consists of different action categories, we need to train the action recognition classifier on UCF-Sports to

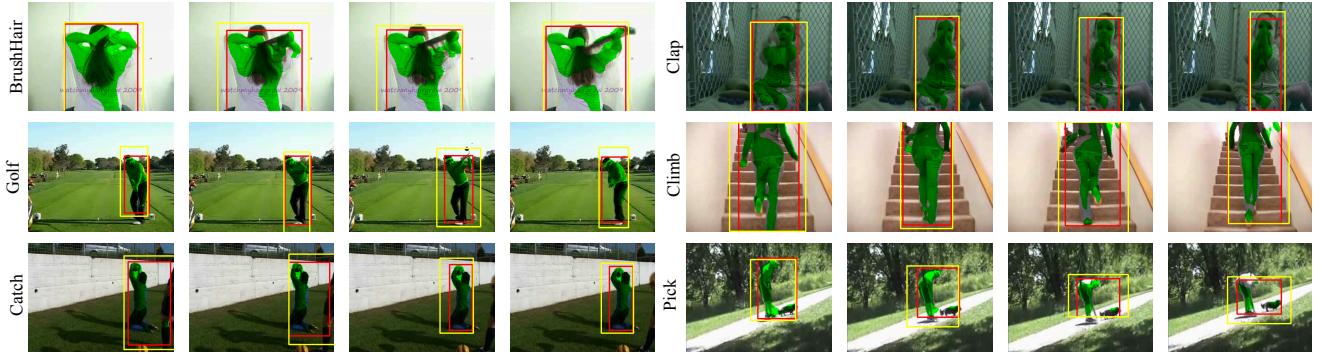


Fig. 11. Action segmentation and detection results obtained by ST-CNN on the J-HMDB dataset. Green pixel-wise segmentation maps show the predictions of ST-CNN, and the red boxes show the bounding boxes generated from the segmentation maps. Yellow boxes represent the detection results obtained by T-CNN for comparison.

TABLE 4

Comparison of the state-of-the-art approaches on J-HMDB. The IoU threshold α for frame m-AP is fixed to 0.5.

	f.-mAP ($\alpha = 0.5$)	v.-mAP ($\alpha = 0.2$)	v.-mAP ($\alpha = 0.5$)
Gkioxari <i>et al.</i> [1]	36.2	—	53.3
Weinzaepfle <i>et al.</i> [2]	45.8	63.1	60.7
Peng <i>et al.</i> [3]	58.5	74.3	73.1
Kalogeiton <i>et al.</i> [43]	65.7	74.2	73.7
Singh <i>et al.</i> [16]	—	73.8	72.0
Ours (T-CNN)	61.3	78.4	76.9
Ours (ST-CNN)	64.9	78.6	78.3
Ours (un-pool)	57.1	71.6	73.9

TABLE 5

Comparison of the state-of-the-art on UCF-101 (24 actions). The IoU threshold α for frame m-AP is fixed to 0.5.

IoU th.	f.-mAP	video-mAP			
		0.05	0.1	0.2	0.3
Weinzaepfle <i>et al.</i> [2]	35.84	54.3	51.7	46.8	37.8
Peng <i>et al.</i> [3]	65.73	78.76	77.31	72.86	65.7
Kalogeiton <i>et al.</i> [43]	67.1	—	—	77.2	—
Singh <i>et al.</i> [16]	—	—	—	73.5	—
Ours	67.3	78.2	77.9	73.1	69.4

TABLE 6

Action recognition results of our T-CNN approach on the four datasets.

	Accuracy (%)
UCF-Sports	95.7
J-HMDB	67.2
UCF-101 (24 actions)	94.4

predict which action the tubes belong to. We follow the same experimental settings as in Section 6.2.1, and report the results in Table 3. We see that although ST-CNN is trained on a different dataset (J-HMDB), it still achieves a frame level mAP of 84.5%, which is comparable to T-CNN with a mAP of 86.7%. Note that T-CNN is trained on UCF-Sports for bounding box prediction. The results show that the ST-CNN model trained on J-HMDB is able to well distinguish between action regions and background, demonstrating its high capacity of generalizing to cross-dataset action detection. Moreover, ST-CNN outperforms other competing action detection approaches.

6.3 Video Segmentation Experiments

Densely Annotated VIdeo Segmentation (DAVIS) 2016 dataset is specifically designed for the task of video object segmentation. It consists of 50 videos with 3455 annotated frames. Consistent with most prior work, we conduct experiments on the 480p videos with a resolution of 854×480 pixels. 30 videos are taken for training and 20 for validation.

For this experiment, we only need the segmentation branch of our ST-CNN network. Before the training process, we gather the videos from three datasets: JHMDB [53], DAVIS'16 [56] (only training set) and SegTrackv2 [59]. We resize all the frames and annotations to fixed dimensions of 320×240 . If there are multiple foreground objects in the annotation, we merge them as a single foreground object and consider other pixels as background. We train the model for $100k$ iterations and fix the learning rate at 10^{-3} . We call this trained model as our pre-trained model for video segmentation.

We then use this pre-trained model and finetune it on DAVIS'16 training set for $8k$ iterations with learning rate 10^{-4} and $2k$ more iterations with learning rate 10^{-5} . Due to GPU memory constraint, we are not able to process the high resolution videos in 3D CNN. In the training phase, we resize input videos to 320×240 and adjust ground truth annotations to the corresponding resolution. During testing, test videos are resized to 320×240 and fed into the network. To make our results comparable to ground truth, the 320×240 segmentation maps from the network are upscaled to 854×480 by bi-linear interpolation. Finally, our segmentation boundaries are smoothed by fully connected conditional random field [60].

Due to limited number of training videos in DAVIS'16 dataset, we also perform data augmentation to increase the number of video clips for training. Specifically, the following data augmentation techniques are considered:

- **Illumination Modification.** We translate the video pixels from RGB space to HSV space, since the Saturation (S) channel corresponds to “lightness”, while value (V) channel determines the “brightness”. The V channel value v is randomly altered to v' via $v' = av$ where $a \in 1 \pm 0.1$.
- **Background replacement with Foreground.** Half of the foreground (either top/down or left/right) is replaced by background. We use the nearest pixels to interpolate the region.
- **Clip Flipping and Shifting.** We randomly select clips to flip horizontally, and shift the image by 1 pixel.

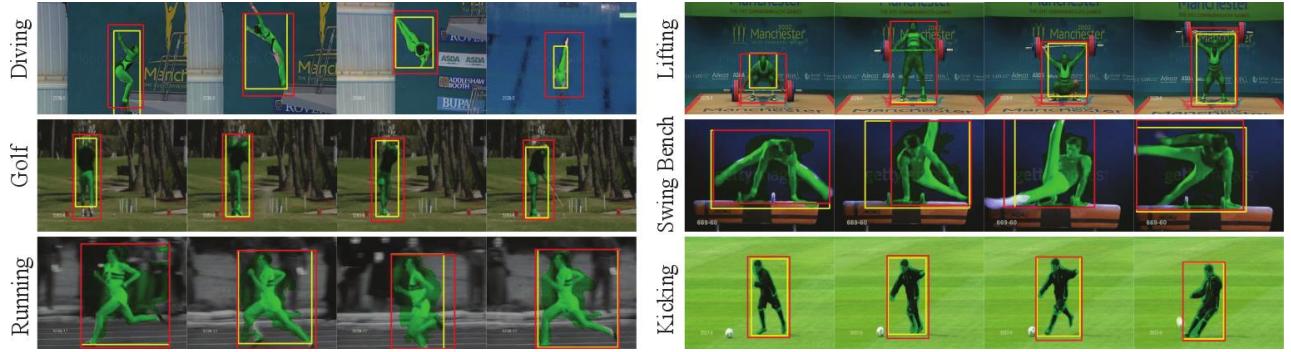


Fig. 12. Transfer learning from J-HMDB to UCF-Sports. We compare ST-CNN action segmentation (green) and bounding box detection (yellow) with ground truth (red) on selected frames of testing videos.

We adopt the same evaluation setting as reported in [56]. There are three parts. **Region Similarity** \mathcal{J} , which is obtained by IoU between the prediction and the ground-truth segmentation map. **Contour Accuracy** \mathcal{F} measures the contours accuracy. **Temporal Stability** \mathcal{T} tracks the temporal consistency in a video. For the first two evaluation, we report the mean, recall and decay. For the third one, we report the average. We compare our results with several unsupervised implementations, since our approach does not require any manual annotation or prior information about the object to be segmented, which is defined as unsupervised segmentation. It is different from the semi-supervised approaches which assume the ground truth segmentation map of the first frame of a test video is given. Apparently, unsupervised segmentation is a much harder task, but is more practical since it does not require any human labelling effort during testing once the segmentation model has been trained. We compare our method with the state-of-the-art unsupervised approaches in Table 7. According to the results, our method achieves the best performance in all performance metrics. Compared to ARP [61], the previous state-of-the-art unsupervised approach, our method achieves 5% gain in contour accuracy (\mathcal{F}) and 15% gain in temporal stability (\mathcal{T}), demonstrating that 3D CNN can effectively take advantage of the temporal information in video frames to achieve temporal segmentation consistency.

Figure 13 shows the quantitative results per video sequence of our approach and the next three top performing methods on DAVIS dataset: ARP [61], LVO [47] and FSEG [45]. Our approach performs the best on low contrast videos including Blackswan, Car-Roundabout and Scooter-Black and achieves competitive results on other videos. Figure 14 presents the qualitative results on four video sequences. In the first row, our results are the most accurate. Our method is the only one which can detect the rims. In the second row, ARP performs the best in suppressing background. However, only our approach detects both legs. The third row shows that only our method is able to accurately segment the tail of the camel. The last row is a very challenging video because of the smoke and small initial size of the car. ARP misses part of the car, while LVO and FSEG mis-classify part of background as moving object. However, our method segments out the car completely and accurately from the background smoke in the scene.

6.4 Computational Cost

We carry out our experiments on a workstation with one GPU (Nvidia GTX Titan X). Given a 40-frames video, T-CNN pipeline takes 1.1 seconds to generate tube proposals, 0.03 seconds to link

tube proposals in a video and 0.9 seconds to predict action label. ST-CNN takes only 0.7 seconds to detect actions (including about 0.6 seconds for segmentation), which is 3 times faster than T-CNN. In ST-CNN, the video clips are only fed into the network once and the detection results are obtained, while in T-CNN, the clips are fed into TPN first to get the tube proposals. Then both clips and tubes are used in the recognition network to predict the label. Since the input to the first stage are clips with fixed duration, while the input to the second stage is whole videos with various duration, it is hard to share weights in their common layers. In contrast, in ST-CNN, there is only one stage since we avoid the proposal generation part. Therefore, ST-CNN is more efficient than T-CNN.

7 CONCLUSION

In this paper we propose an end-to-end 3D CNN based pipeline for action detection in videos. It exploits 3D CNN to extract effective spatio-temporal features and perform action segmentation and recognition. We explored two approaches to locate the action. In the first, coarse proposal boxes are densely sampled based on the 3D convolutional feature cube and linked together (T-CNN). In the second, pixels in each frame are segmented into foreground/background and foreground pixels are aggregated into action segments (ST-CNN). Extensive experiments on several benchmark datasets demonstrate the strength of our approach for spatio-temporal action detection as well as segmentation compared to the state-of-the-arts.

ACKNOWLEDGMENTS

The project was supported by Award No. 2015-R2-CX-K025, awarded by the National Institute of Justice, Office of Justice Programs, U.S. Department of Justice. The opinions, findings, and conclusions or recommendations expressed in this publication are those of the author(s) and do not necessarily reflect those of the Department of Justice.

REFERENCES

- [1] G. Gkioxari and J. Malik, "Finding action tubes," in *CVPR*, 2015.
- [2] P. Weinzaepfel, Z. Harchaoui, and C. Schmid, "Learning to track for spatio-temporal action localization," in *ICCV*, 2015.
- [3] X. Peng and C. Schmid, "Multi-region two-stream r-cnn for action detection," in *ECCV*, 2016.
- [4] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *CVPR*, 2014.

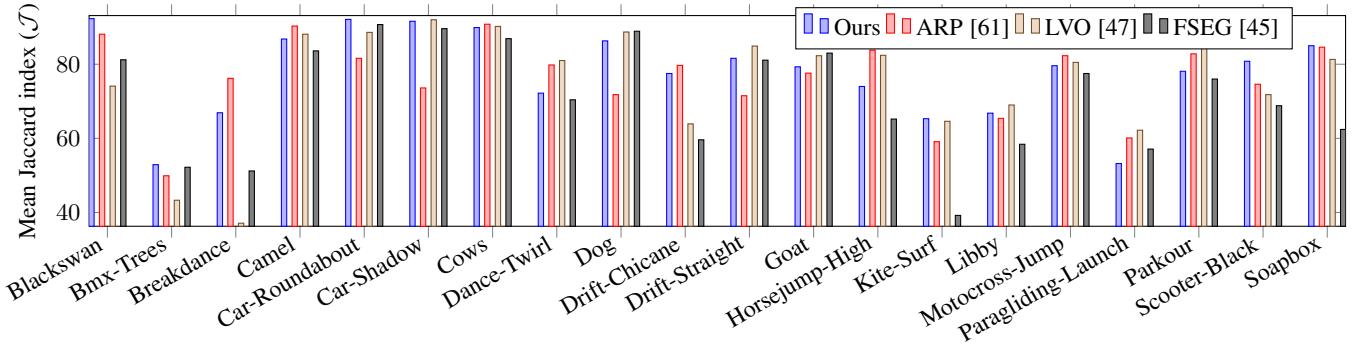
Fig. 13. Comparison of Mean Jaccard index (\mathcal{J}) of different approaches on each of the sequences independently.

TABLE 7
Overall results of region similarity (\mathcal{J}), contour accuracy (\mathcal{F}) and temporal stability (\mathcal{T}) for different approaches. \uparrow means the more the better and \downarrow means the less the better.

Measure		ARP [61]	FSEG [45]	LMP [46]	FST [62]	CUT [63]	NLC [64]	MSG [65]	KEY [66]	CVOS [67]	TRC [68]	SAL [69]	Ours
\mathcal{J}	Mean \uparrow	76.2	70.7	70.0	55.8	55.2	55.1	53.3	49.8	48.2	47.3	39.3	77.6
	Recall \uparrow	91.1	83.5	85.0	64.9	57.5	55.8	61.6	59.1	54.0	49.3	30.0	95.2
	Decay \downarrow	7.0	1.5	1.3	0.0	2.2	12.6	2.4	14.1	10.5	8.3	6.9	2.3
\mathcal{F}	Mean \uparrow	70.6	65.3	65.9	51.1	55.2	52.3	50.8	42.7	44.7	44.1	34.4	75.5
	Recall \uparrow	83.5	73.8	79.2	51.6	61.0	51.9	60.0	37.5	52.6	43.6	15.4	94.7
	Decay \downarrow	7.9	1.8	2.5	2.9	3.4	11.4	5.1	10.6	11.7	12.9	4.3	4.9
\mathcal{T}	Mean \downarrow	39.3	32.8	57.2	36.6	27.7	42.5	30.1	26.9	25.0	39.1	66.1	22.0

- [5] R. Girshick, “Fast r-cnn,” in *ICCV*, 2015.
- [6] S. Ren, K. He, R. Girshick, and J. Sun, “Faster r-cnn: Towards real-time object detection with region proposal networks,” in *NIPS*, 2015.
- [7] E. Borenstein and S. Ullman, “Class-specific, top-down segmentation,” in *ECCV*. Springer, 2002.
- [8] J. Winn and N. Jojic, “Locus: Learning object classes with unsupervised segmentation,” in *ICCV*, vol. 1. IEEE, 2005.
- [9] Z. Tu, X. Chen, A. L. Yuille, and S.-C. Zhu, “Image parsing: Unifying segmentation, detection, and recognition,” *IJCV*, vol. 63, no. 2, 2005.
- [10] V. Badrinarayanan, A. Kendall, and R. Cipolla, “Segnet: A deep convolutional encoder-decoder architecture for image segmentation,” *arXiv preprint arXiv:1511.00561*, 2015.
- [11] H. Rui, C. Chen, and M. Shah, “Tube convolutional neural network (t-cnn) for action detection in videos,” in *ICCV*, 2017.
- [12] K. He, G. Gkioxari, P. Dollár, and R. Girshick, “Mask r-cnn,” *arXiv preprint arXiv:1703.06870*, 2017.
- [13] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, “Ssd: Single shot multibox detector,” in *ECCV*, 2016.
- [14] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You only look once: Unified, real-time object detection,” in *CVPR*, 2016.
- [15] R. Joseph and F. Ali, “Yolo9000: Better, faster, stronger,” in *CVPR*, 2017.
- [16] G. Singh, S. Saha, M. Sapienza, P. Torr, and F. Cuzzolin, “Online real time multiple spatiotemporal action localisation and prediction,” in *ICCV*, 2017.
- [17] V. Kalogeiton, P. Weinzaepfel, V. Ferrari, and C. Schmid, “Joint learning of object and action detectors,” in *ICCV*, 2017.
- [18] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” *Nature*, vol. 521, no. 7553, 2015.
- [19] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, “Large-scale video classification with convolutional neural networks,” in *CVPR*, 2014.
- [20] J. Yue-Hei Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici, “Beyond short snippets: Deep networks for video classification,” in *CVPR*, 2015.
- [21] K. Simonyan and A. Zisserman, “Two-stream convolutional networks for action recognition in videos,” in *NIPS*, 2014.
- [22] S. Ji, W. Xu, M. Yang, and K. Yu, “3d convolutional neural networks for human action recognition,” *TPAMI*, vol. 35, no. 1, 2013.
- [23] L. Sun, K. Jia, D.-Y. Yeung, and B. E. Shi, “Human action recognition using factorized spatio-temporal convolutional networks,” in *ICCV*, 2015.
- [24] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, “Learning spatiotemporal features with 3d convolutional networks,” in *ICCV*, 2015.
- [25] Z. Shou, D. Wang, and S.-F. Chang, “Temporal action localization in untrimmed videos via multi-stage cnns,” in *CVPR*, 2016.
- [26] Z. Shou, J. Chan, A. Zareian, K. Miyazawa, and S.-F. Chang, “Cdc: Convolutional-de-convolutional networks for precise temporal action localization in untrimmed videos,” in *CVPR*, 2017.
- [27] H. Xu, A. Das, and K. Saenko, “R-c3d: Region convolutional 3d network for temporal activity detection,” in *ICCV*, 2017.
- [28] S. Buch, V. Escorcia, B. Ghanem, L. Fei-Fei, and J. C. Niebles, “End-to-end, single-stream temporal action detection in untrimmed videos,” in *BMVC*, 2017.
- [29] S. Buch, V. Escorcia, C. Shen, B. Ghanem, and J. C. Niebles, “Sst: Single-stream temporal action proposals,” in *CVPR*, 2017.
- [30] M. Jain, J. C. van Gemert, and C. G. Snoek, “What do 15,000 object categories tell us about classifying and localizing actions?” in *CVPR*, 2015.
- [31] Y. Ke, R. Sukthankar, and M. Hebert, “Event detection in crowded videos,” in *ICCV*, 2007.
- [32] T. Lan, Y. Wang, and G. Mori, “Discriminative figure-centric models for joint action localization and recognition,” in *ICCV*, 2011.
- [33] Y. Tian, R. Sukthankar, and M. Shah, “Spatiotemporal deformable part models for action detection,” in *CVPR*, 2013.
- [34] P. Felzenszwalb, D. McAllester, and D. Ramanan, “A discriminatively trained, multiscale, deformable part model,” in *CVPR*, 2008.
- [35] D. Oneata, J. Verbeek, and C. Schmid, “Efficient action localization with approximately normalized fisher vectors,” in *CVPR*, 2014.
- [36] C. Desai and D. Ramanan, “Detecting actions, poses, and objects with relational phraselets,” in *ECCV*, 2012.
- [37] S. Ma, J. Zhang, N. Ikizler-Cinbis, and S. Sclaroff, “Action recognition and localization by hierarchical space-time segments,” in *ICCV*, 2013.
- [38] M. Jain, J. Van Gemert, H. Jégou, P. Bouhoumy, and C. G. Snoek, “Action localization with tubelets from motion,” in *CVPR*, 2014.
- [39] K. Soomro, H. Idrees, and M. Shah, “Action localization in videos through context walk,” in *CVPR*, 2015.
- [40] D. Oneata, J. Revaud, J. Verbeek, and C. Schmid, “Spatio-temporal object detection proposals,” in *ECCV*, 2014.
- [41] J. Lu, J. J. Corso *et al.*, “Human action segmentation with hierarchical supervoxel consistency,” in *CVPR*, 2015.
- [42] G. Zhu, F. Porikli, and H. Li, “Tracking randomly moving objects on edge box proposals,” *arXiv preprint arXiv:1507.08085*, 2015.



Fig. 14. Qualitative results of the proposed approach (red), ARP (yellow), LVO (cyan) and FSEG (magenta) on selected frames from DAVIS dataset.

- [43] V. Kalogeiton, P. Weinzaepfel, V. Ferrari, and C. Schmid, “Action tubelet detector for spatio-temporal action localization,” in *ICCV*, 2017.
- [44] J. Cheng, Y.-H. Tsai, S. Wang, and M.-H. Yang, “Segflow: Joint learning for video object segmentation and optical flow,” *arXiv preprint arXiv:1709.06750*, 2017.
- [45] S. D. Jain, B. Xiong, and K. Grauman, “Fusionseg: Learning to combine motion and appearance for fully automatic segmentation of generic objects in videos,” *arXiv preprint arXiv:1701.05384*, 2017.
- [46] P. Tokmakov, K. Alahari, and C. Schmid, “Learning motion patterns in videos,” *arXiv preprint arXiv:1612.07217*, 2016.
- [47] ———, “Learning video object segmentation with visual memory,” *arXiv preprint arXiv:1704.05737*, 2017.
- [48] M. D. Zeiler and R. Fergus, “Visualizing and understanding convolutional networks,” in *ECCV*. Springer, 2014.
- [49] H. Noh, S. Hong, and B. Han, “Learning deconvolution network for semantic segmentation,” in *ICCV*, 2015.
- [50] P. Fischer, A. Dosovitskiy, E. Ilg, P. Häusser, C. Hazırbaş, V. Golkov, P. van der Smagt, D. Cremers, and T. Brox, “Flownet: Learning optical flow with convolutional networks,” *arXiv preprint arXiv:1504.06852*, 2015.
- [51] W. Shi, J. Caballero, F. Huszár, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang, “Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network,” in *CVPR*, 2016.
- [52] M. Rodriguez, A. Javed, and M. Shah, “Action mach: a spatio-temporal maximum average correlation height filter for action recognition,” in *CVPR*, 2008.
- [53] H. Jhuang, J. Gall, S. Zuffi, C. Schmid, and M. J. Black, “Towards understanding action recognition,” in *ICCV*, 2013.
- [54] Y.-G. Jiang, J. Liu, A. Roshan Zamir, I. Laptev, M. Piccardi, M. Shah, and R. Sukthankar, “THUMOS challenge: Action recognition with a large number of classes,” */ICCV13-Action-Workshop/*, 2013.
- [55] Y.-G. Jiang, J. Liu, A. Roshan Zamir, G. Toderici, I. Laptev, M. Shah, and R. Sukthankar, “THUMOS challenge: Action recognition with a large number of classes,” <http://crcv.ucf.edu/THUMOS14/>, 2014.
- [56] F. Perazzi, J. Pont-Tuset, B. McWilliams, L. Van Gool, M. Gross, and A. Sorkine-Hornung, “A benchmark dataset and evaluation methodology for video object segmentation,” in *CVPR*, 2016.
- [57] L. Wang, Y. Qiao, and X. Tang, “Video action detection with relational dynamic-poselets,” in *ECCV*, 2014.
- [58] W. Sultani and M. Shah, “What if we do not have multiple videos of the same action? – video action localization using web images,” in *CVPR*, June 2016.
- [59] R. Almomani and M. Dong, “Segtrack: A novel tracking system with improved object segmentation,” in *ICIP*. IEEE, 2013.
- [60] P. Krähenbühl and V. Koltun, “Efficient inference in fully connected crfs with gaussian edge potentials,” in *NIPS*, 2011.
- [61] Y. J. Koh and C.-S. Kim, “Primary object segmentation in videos based on region augmentation and reduction,” in *CVPR*, 2017.
- [62] A. Papazoglou and V. Ferrari, “Fast object segmentation in unconstrained video,” in *ICCV*, 2013.
- [63] M. Keuper, B. Andres, and T. Brox, “Motion trajectory segmentation via minimum cost multicut,” in *ICCV*, 2015.
- [64] A. Faktor and M. Irani, “Video segmentation by non-local consensus voting,” in *BMVC*, vol. 2, no. 7, 2014.
- [65] T. Brox and J. Malik, “Object segmentation by long term analysis of point trajectories,” in *ECCV*, 2010.
- [66] Y. J. Lee, J. Kim, and K. Grauman, “Key-segments for video object segmentation,” in *ICCV*, 2011.
- [67] B. Taylor, V. Karasev, and S. Soatto, “Causal video object segmentation from persistence of occlusions,” in *CVPR*, 2015.
- [68] K. Fragkiadaki, G. Zhang, and J. Shi, “Video segmentation by tracing discontinuities in a trajectory embedding,” in *CVPR*, 2012.
- [69] W. Wang, J. Shen, and F. Porikli, “Saliency-aware geodesic video object segmentation,” in *CVPR*, 2015.



Rui Hou received his BSc from Beijing University of Posts and Telecommunications in 2010 and MS degree in Electrical Engineering from Arizona State University. He joined Center for Research in Computer Vision at University of Central Florida in 2012, where he is currently working toward the PhD degree in Computer Science. His research interests include action recognition/detection in video and semantic segmentation in video.



Chen Chen received the PhD degree in the Department of Electrical Engineering at the University of Texas at Dallas in 2016. He is currently a Post-Doc in the Center for Research in Computer Vision at University of Central Florida. His research interests include compressed sensing, signal and image processing, and computer vision. He is an associated editor for KSII Transactions on Internet and Information Systems, and Signal, Image and Video Processing.



Mubarak Shah, the Trustee chair professor of computer science, is the founding director of the Center for Research in Computer Vision at the University of Central Florida. He is a fellow of IEEE, AAAS, IAPR and SPIE. He is an editor of an international book series on video computing, editor-in-chief of Machine Vision and Applications journal, and an associate editor of ACM Computing Surveys journal. He was the program cochair of the CVPR in 2008, an associate editor of the IEEE TPAMI.