Spatiotemporal Deformable Models for Long-Term Complex Activity Detection

Salman Khan  
Oxford Brookes University  
United Kingdom  
19052999@brookes.ac.uk

Fabio Cuzzolin  
Oxford Brookes University  
United Kingdom  
fabio.cuzzolin@brookes.ac.uk

Abstract

Long-term complex activity recognition and localisation can be crucial for the decision-making process of several autonomous systems, such as smart cars and surgical robots. Nonetheless, most current methods are designed to merely localise short-term action/activities or combinations of atomic actions that only last for a few frames or seconds. In this paper, we address the problem of long-term complex activity detection via a novel deformable, spatiotemporal parts-based model. Our framework consists of three main building blocks: (i) action tube detection, (ii) the modelling of the deformable geometry of parts, and (iii) a sparsity mechanism. Firstly, action tubes are detected in a series of snippets using an action tube detector. Next, a new 3D deformable RoI pooling layer is designed for learning the flexible, deformable geometry of the constellation of parts. Finally, a sparsity strategy differentiates between activated and deactivate features. We also provide temporal complex activity annotation for the recently released ROAD autonomous driving dataset and the SARAS-ESAD surgical action dataset, to validate our method and show the adaptability of our framework to different domains. As they both contain long videos portraying long-term activities they can be used as benchmarks for future work in this area.

1 Introduction

Complex activity recognition is attracting much attention in the computer vision research community due to its significance for a variety of real-world applications, such as autonomous driving [7, 8], surveillance [26], medical robotics [55] or team sports analysis [19]. In autonomous driving, for instance, it is extremely important that the vehicle understands dynamic road scenes, in order e.g. to accurately predict the intention of pedestrians and forecast their trajectories to inform appropriate decisions. In surveillance, group activities rather than actions performed by individuals need to be monitored. Robotic assistant surgeons need to understand what the main surgeon is doing throughout a complex surgical procedure composed by many short-term actions and events [42], in order to suitably assist them.

Methods recently published on action or activity recognition and localisation can be broadly divided into two categories; single atomic action [27, 34, 17, 49] and multiple atomic action recognition/localisation [20, 47, 52, 44, 28, 23]. The first class of methods only focuses on identifying the start and end of an action performed in a short video that only includes a single instance. Among the datasets used to validate this type of methods are, e.g., UCF-101 [43] or Charades [37]. The second class of methods consider videos which contain a number of atomic actions or repetitions of the same action multiple times with different duration. Methods in this category do address complex activity recognition, as their aim is to understand an overall, dynamic scene by detecting and identifying its constituent components. The datasets used for complex activity detection are Epic-kitchens [10], THUMOS14 [21] and ActivityNet v1.3 [6]. Both classes of methods are geared towards merely recognising and localising short term action or activities that lasts for only a few frames or seconds.

Unlike all existing methods, in this work we present a complex activity detection framework capable of recognising long-term activities, validated in autonomous driving and surgical robotics (as illustrated in Figure 1) but
Figure 1: Our complex activity recognition framework workflow for two videos from the ROAD (autonomous driving) and SARAS (surgical robotics) datasets. The overall concept of detecting complex activities via a part-based deformation mechanism is the same in both cases. However, in ROAD background frames exist which do not contain any complex road activity. In contrast, the SARAS complex activities are actually the phases of a surgical procedure, which are contiguous without the need for a background label.

The main contributions of this paper are therefore:

- A novel framework for long-term complex activity recognition and localisation.
- An original deformable 3D RoI pooling approach for pooling part features to create activity representations.
- A sparsity mechanism for coping with a variable number of parts of relative importance across different instances of the same semantic activity.
- Crucially, an effort to augment two newly-released datasets in the autonomous driving and robotic surgery domains to make them into suitable benchmarks for future work on complex activity detection.
2 Related Work

Complex activity recognition. Most recent work on complex activity recognition concerns scalar sensors \[5, 45, 53\] or combination of both scalar and vision sensors \[1, 23\]. Recently, though, several vision-based complex activity recognition methods have been proposed \[20, 47, 52, 44, 28, 23\] with the goal of understanding an overall scene by recognising and segmenting atomic actions. These methods can be further divided into (i) sliding windows approaches \[46, 36\], in which an activity classifier is applied to each snippet, and (ii) boundaries analyses \[48, 15\], in which a model is trained to identify the start and end time of each action. Overall, current activity recognition methods are geared to recognise short-term activities via a combination of small atomic actions.

Unlike existing approaches, our objective is to understand long-term activities in dynamic scenes. A good example is the phases a surgical procedure can be broken into, which are of long duration and whose detection is crucial to inform the decision making of automated robotic assistants.

Action/activity temporal localisation. Significant progress has been made in recent years toward temporal action localisation in untrimmed videos. The main goal of temporal action localisation is to remove the irrelevant background from the input videos and identify the start and end instants of the action of interest. Temporal localisation methods can be divided into supervised \[2, 20\] versus weakly/unsupervised ones \[35, 30, 51\]. While the former require temporal annotations to learn a model able to identify the start and end of an action, in weakly-supervised settings the model is learned from partial annotation.

Our goal is the same as temporal activity localisation, with the difference that we wish to target complex, long duration events articulated into a possibly large number of parts, whose internal structure cannot be leveraged by naive temporal segmentation methods.

Deformable parts-based models. Deformable part-based models have been used by the research community for more than a decade \[12, 13, 14, 18\] for objection detection and segmentation. Following the rapid development of Convolution Neural Networks (CNNs), Gir-
shick et al. first recognised that deformable part-based models can be implemented for object detection in a CNN formulation, in which each convolution pyramid is fed to a distance transform pooling and a geometric filter layer. The main limitation of this method is that it is not end-to-end trainable and requires a heuristic selection of part sizes and components. A subsequent end-to-end deformable CNN formulation was proposed in [9], which uses two new CNN layers (deformable convolution and deformable RoI pooling) that reproduce the functionalities of traditional part deformation. The latest version of deformable CNN is Deformable ConvNets v2 [54], which introduces a modulation mechanism in both deformable convolution and RoI pooling.

To the best of our knowledge, all deformable models proposed to date focus on either object or short-term action detection, whereas here we propose a deformable part-based mechanism for detecting complex long-term activities.

3 Proposed Method

3.1 Motivation

Whereas modern complex activity recognition and localisation methods focus on single action or collections of short-term atomic actions, the issue of detecting complex activities articulated into a number of atomic actions and spanning longer intervals of time is still rather unexplored. Crucial to the identification of such activities is being able to model the relationship/coordination among individual actions (performed either by the same person or by multiple individuals). Deformable models that represented activities as graphs of atomic parts provide such a framework. The existing part-based methods are investigated for action detection and prediction tasks, but their potential for the recognition of long-term activities has not been explored so far.

The workflow of our approach is illustrated in Figure 2. The input video is divided into fixed-sized snippets which are fed to an action tube detector, followed by a novel 3D deformable RoI pooling stage which computes parts features, which are later sparsified to allow flexibility for activity detection.

3.2 Action Tube Detection

To provide a fixed-size representation for the instances of atomic actions composing a complex activity, we adopt AMTNet [32]. AMTNet is a two-stream online action tube detector that uses both RGB and optical flow information. Here, however, we only use the RGB stream. The main rationale for using AMTNet is that it generates tubes in an incremental manner while preserving a fixed-size representation, as required by our feature extraction stage.

Architectural Details. AMTNet uses VGG-16 [38] as baseline CNN feature extractor. The last two fully-connected layers of VGG-16 are replaced by two convolutional layers, and add four extra convolutional layers at the end. AMTNet takes sequence of RGB frames as an input with a fixed temporal interval $\Delta$ between consecutive frames, i.e., $\{f_t, f_{t+\Delta}\}$. The input to AMTNet is in the format $[BS \times Sq \times D \times H \times W]$, where $BS$ is the batch size, $Sq$ is the sequence length (in this case a pair), $D$ is the dimensionality (equal to 3 as we are dealing with RGB frames), while $H$ and $W$ are the height and weight of each frame ($300 \times 300$ in our case). As typical in action detection, AMTNet uses both a classification and a regression layer for recognition and detection, respectively, with the goal of predicting action ‘micro-tubes’ defined by pairs of consecutive detections. The method predicts bounding boxes for a pair of frames separated by fixed gap $\Delta$, while the bounding boxes for intermediate frames are generated by interpolation. In this work, atomic action instances are represented as 3 micro-tubes with $\Delta = 3$ for an overall tube length of 12 frames, aligned with our snippet length.

Complete action tubes are incrementally generated by AMTNet by temporally linking the micro-tubes predicted by the network (see Figure 3).

3.3 3D Part-Based Deformable Layer

The main building block and feature extractor of our framework is a novel 3D deformable RoI pooling layer which encodes the spatiotemporal geometry of the action tubes which correspond to the activity parts. This is an extension of the existing standard deformable RoI pooling layer [9] that has the ability to extract and learn features from an action tube rather than a 2D bounding box. The principle of our 3D deformable RoI pooling operation is shown in Figure 4. Like the classical deformable RoI pooling layer, our module also includes standard RoI
Figure 3: Illustration of AMTNet [32] micro- and complete action tube generation process. (a) AMTNet generates micro-action tubes by predicting bounding boxes for the start and the end frame of a snippet. Predictions for a fixed number of intermediate frames are constructed by bilinear interpolation. (b) Complete action tubes are created by temporally linking micro action tubes by dynamic programming.

Figure 4: Structure of our 3D deformable RoI pooling layer, which takes feature maps and action tube locations as input and arranges them into a fixed-size grid of parts (here illustrated for size $3 \times 3$). For each part, an offset is generated and multiplied by the original part (tube) feature to output the final component features.

3.4 Component-wise Sparsity

Of the features extracted by the action detector and collected in the RoI feature map, some are active and discriminative of the dynamics of the action and activity, whereas others are quite static and do not contribute to the recognition process. Thus, in our approach we identify those components which actually contribute to discriminating the activity (‘activated’), and neglect those which do not (‘deactivated’).

To identify activated and deactivated components we used an incremental dimensional reduction method similar to the feature selection variant of Covariance-free Incremental Partial Least Squares (CIPLS) [24]. In the original method Variable Importance in Projection (VIP) is used for estimating the importance of each feature component. In our framework we adapt the VIP approach to finding the importance of each of the component of the feature maps generated by our 3D deformable RoI pooling layer. We then select a fraction of components with higher importance for further learning.

3.5 Complete Framework

The complete framework is the concatenation of the aforementioned three modules. Firstly, we divide the video

pooling (used in all region proposal-based object detection methods), a fully connected layer, and offsets.

Firstly, standard RoI pooling is applied to the provided feature maps and bounding boxes locations forming an action tube, by subdividing the tube into a pooled feature map grid of fixed-size in both the spatial and the temporal dimension. Next, normalised offsets are generated for these feature maps using a fully-connected layer, which are then transformed using the element-wise product with original RoI’s width and height. Offsets are also multiplied by a scalar value to modulate the magnitude of offsets (empirically set to 0.1) which makes the it invariant to different sizes of RoI. In our framework, this layer takes the VGG features extracted by AMTnet and the each detected action tube separately as an input and returns an overall feature map which encodes both the appearance and shape of each atomic action.
Algorithm 1: Complex Activity Detection

1. **Input**: Input Video $V$ divided into $N$ number of Snippets $S_i$ consisting of fixed $M$ number of frames $F^{1,2,3,...,M}$.
2. **Initialization**: Load pretrained action tube detector $ATD$ on desire dataset, import our 3D deformable RoI pooling $DRoI$ and sparsity $Sp$ modules.
   
3. for $i \in 1,2,3,...,N$ do
   
4. Extract features and detect action tubes $B_A^{M \times K \times 5}, X_B^{M \times 64 \times 300 \times 300} \leftarrow ATD(S_i)$
   
5. $X_{DRoI}^{K \times 64 \times M \times 7 \times 7} \leftarrow DRoI(B_A^{M \times K \times 5}, X_B^{M \times 64 \times 300 \times 300})$;
   
6. $Sp(X_{DRoI}^{K \times 64 \times M \times 7 \times 7})$;
   
7. $X_{Fc1}^{K \times 5 \times 512} \leftarrow Fc1(\text{Sp}X_{Fc1}^{K \times 32 \times 7 \times 7 \times 7})$;
   
8. $X_{Fc2}^{2048} \leftarrow Fc2(\text{Sp}X_{Fc2}^{K \times 5 \times 512})$;
   
9. $CA_i \leftarrow \text{Softmax}(X_{Fc2}^{2048})$;
   
10. if $CA_{i-1} == CA_i$ then
   
11.   Same activity;
   
12. else
   
13.   if $CA_{i-1} == CA_{i+1}$ then
   
14.     Same activity;
   
15. else
   
16.     End of activity;
   
17. start new activity;
   
end

16. **Output**: Start and end frames with activity labels

$V$ into $N$ Snippets $S_{1,2,3,...,N}$, with each Snippet $S_i$ consisting of a fixed ($M$) number of frames: $S_i = F^{1,2,3,...,M}$. Each snippet is passed to the action tube detection module $AT$ which returns $K$ action tubes each composed by 12 bounding boxes with labels $B_A$ and intermediate VGG features $X_A$ represented as $B_A^{M \times K \times 5}, X_A^{M \times 64 \times 300 \times 300} \in AT$. Action tube locations and features are then passed to our 3D deformable RoI pooling layer $DRoI$ which returns a fixed-sized (i.e., $7 \times 7$) grid of components whose dimensionality is equal to the number (64) of convolutional layers: $X_{DRoI}^{K \times 64 \times M \times 7 \times 7}$. The number of components of each tube is then reduced by selecting only the activated component via our sparsity module $Sp$, using a fixed proportion (i.e., 50%), yielding $X_{Sp}^{K \times 32 \times M \times 7 \times 7} \in Sp$. Finally, the latter features are fed to two fully-connected layers with batch normalization and ReLU followed by a Softmax classifier to classify the snippet into their respective activity category. For localisation we use a sliding window approach with a dual verification mechanism. Namely, since we target long term activities, if there is a random false positive or a random false negative between two same class snippets we simply ignore it and consider it as a same activity. The complete process is described in Algorithm [1].

**Implementation**: Before training our overall architecture, we separately train AMTNet for action tube detection over both datasets. Note that we had to design from scratch suitable data loaders for the two datasets, as the format of the annotation there is completely different from that of the original datasets AMTNet was validated upon. As mentioned, our 3D RoI pooling layer includes a temporal dimension to learn and extract tube parts of a tube rather than of a 2D object. In our experiments we also convert a more recent version of deformable RoI pooling called ‘modulated’ deformable RoI pooling to the 3D case. For sparsity we adapt the incremental dimensionality reduction method in [24] to reduce the number of tube components. All three modules are eventually concatenated into a single architecture with two fully-connected layers and a Softmax classifier at the end. Our architecture is implemented using the PyTorch [29] deep learning library with OpenCV and Scikit-learn. For training we used a machine equipped with 4 Nvidia GTX 1080 GPUs with 12GB VRAM each.

4 Experimental Results

4.1 Datasets and Evaluation Metrics

In this paper we used two datasets for evaluating our approach, both already annotated at video level for action tubes detection which is a prerequisite for our system to be able to perform complex activity recognition.

**ROAD** [39]: ROAD (The Road event Awareness Dataset for Autonomous Driving) is annotated for road action and event detection. Each event is described in terms of three different labels: (road) agent (e.g. cyclist, bus), action performed by the agent (e.g. turning left, right), and event location (w.r.t. the autonomous vehicle). The ROAD dataset consists of total 22 videos carefully selected from the Oxford RobotCar Dataset because of their
diverse weather and lighting conditions. ROAD comprises 560K bounding boxes in 122K annotated frames with 560K agent labels, 640K action labels and 499K location labels.

For this work we augmented the annotation of the ROAD dataset for complex road activity detection. We used a total of 19 videos with an average duration of 8 minutes each, in which 12 were selected for training and the remaining 7 for testing. We temporally annotated the ROAD videos by specifying the start and end frame for six different classes of complex road activities we inferred from video inspection. For example, a ‘Negotiating intersection’ can be defined which is made up of the following ‘atomic’ events: Autonomous Vehicle (AV)-move + Vehicle traffic light / Green + AV-stop + Vehicle(s) / Stopped / At junction+ AV-move. Each activity class with statistics is listed and briefly described in Table 1.

SARAS [3]: ESAD (Endoscopic Surgeon Action Detection Dataset) is a benchmark devised for surgeon action detection in real-world endoscopic surgery videos captured as part of the SARAS (Smart Autonomous Robotic Assistant Surgeon) project. In ESAD, surgeon actions are divided into 21 different categories and annotated with the help of professional surgeons. For this work we took a step forward and annotated ESAD in terms of complex activities corresponding to the different phases of the surgical procedure portrayed by the videos (namely, radical prostatectomy). For example, Phase # 3 corresponds to ‘Bladder neck transection’, in which a scissor cuts the neck of the bladder until it is transected. Complete details about each phase and their statistics are given in Table 2. For more details on ESAD please refer to [3].

Evaluation Metrics: For the evaluation of action tube detection performance we use standard frame/video mean Average Precision (mAP) with different IoU thresholds $\delta$, namely 0.2, 0.3, 0.5, 0.75, on both datasets. Complex activity recognition is evaluated using classification accuracy, precision, recall and F-score. For complex activity localisation we use the standard protocol mAP over the temporal dimension used by almost all relevant methods.

4.2 Results on Action Tube Detection

In this section, we show the evaluation of action tubes detector AMTNet as it is an important bit of our framework and the overall accuracy depends upon it. A detailed comparative analysis of AMTNet over different action detection datasets can be found in the original paper [32].
Table 3: Action tube detection performance on both the ROAD and SARAS-ESAD datasets. Both Frame-mAP and Video-mAP at different IoU thresholds are reported for evaluation.

| Methods / IoU threshold | ROAD     | SARAS-ESAD |
|-------------------------|----------|------------|
|                        | 0.2      | 0.3 | 0.5 | 0.75 | 0.2 | 0.3 | 0.5 | 0.75 |
| Singh et al. [39] (frame-mAP) | - | - | 25.9 | - | - | - | - | - |
| Singh et al. [39] (video-mAP) | 17.5 | - | 4.6 | - | - | - | - | - |
| Bawa et al. [3] (frame-mAP) | - | - | - | - | - | 24.3 | 12.2 | - |
| AMTNet (frame-mAP) | 22.3 | 18.1 | 15.4 | 11.0 | 30.4 | 24.6 | 18.7 | 7.9 |
| AMTNet (video-mAP) | 11.6 | 7.9 | 3.8 | - | 13.7 | 10.1 | 8.8 | 5.4 |

while in this section our focus is on the performance of AMTNet on our two datasets of choice. We show these results because none of them was ever used to test AMTNet. The detailed results are reported in Table 3 where we also compare AMTNet with the baselines proposed by the papers which released the two datasets. The ROAD baseline, termed 3D RetinaNet [39], reports detailed results for action tube detection there at both frame-mAP and video-mAP level. The ESAD baseline [3], a vanilla implementation of RetinaNet, only provides frame-level results. Thus, to provide a comparison between our method and both baselines we calculated both frame-mAP and video-mAP results on both datasets. As it can be seen in Table 3, AMTNet performed better than [3] on SARAS while it is inferior to [39] on ROAD. Again, the main rationale for using AMTNet is that it can provide a fixed-size representation for the tubes as required by our framework, motivating us to compromise on accuracy.

4.3 Complex Activity Recognition

In this section, we provide the detail analysis of complex activity classification for both ROAD and SARAS datasets. The performance of each class in both datasets with four evaluation metrics are illustrated in Figure 5. From the results, it is clear that in ROAD dataset, there is a variation between the performance activities. In other words, the recognition accuracy of activities that occur too often i.e., waiting in a queue is very higher than the other with very low occurrence i.e., sudden appearance. This fluctuation of activity occurrence makes the ROAD dataset more challenging during the training. Unlike the ROAD, in SARAS each of the activity contains enough samples for training but diversity in phases also play a challenging role in recognition.

4.4 Complex Activity Localisation

To show the effectiveness of our proposed method we performed extensive experiments for localization of activities in long-term scenarios. This section also shows the ef-
Table 4: Comparative analysis of ROAD activity localization with different number of maximum action tubes selection at five different IoU thresholds $\alpha$.

| # of Tubes | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 |
|------------|-----|-----|-----|-----|-----|
| 3          | 0.71| 0.69| 0.63| 0.57| 0.55|
| 5          | 0.82| 0.73| 0.66| 0.61| 0.58|
| 6          | 0.77| 0.71| 0.65| 0.60| 0.56|
| 7          | 0.69| 0.64| 0.59| 0.51| 0.48|

Table 5: Comparative analysis of SARAS activity localization with different number of maximum action tubes selection at five different IoU thresholds $\alpha$.

| # of Tubes | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 |
|------------|-----|-----|-----|-----|-----|
| 3          | 0.67| 0.63| 0.61| 0.58| 0.56|
| 5          | 0.65| 0.62| 0.58| 0.57| 0.55|
| 6          | 0.58| 0.54| 0.52| 0.51| 0.49|
| 7          | 0.53| 0.51| 0.50| 0.48| 0.45|

4.5 Limitations and Future Work

The main limitation of this work is that it relies on action tube detection and, from our results, the existing tube detectors are not reliable enough to perform well over challenging real-world datasets such as those we chose as benchmarks. Clearly, if the tube detector misses an important atomic action this will affect the overall activity detection performance. Another limitation of our current implementation is the tube components selection are purely on the basis of the current sample without any extra supervision or learning.

In the future our primary target will be the design of a more accurate action tube detector with the ability to perform better in challenging scenarios such as those portrayed in ROAD or SARAS. We will also aim at further building on our 3D deformable RoI pooling layer to devise a model able to jointly perform both part-wise feature extraction and the identification of activated and deactivate components for each tube.

Further down the line, an interesting generalisation of deformable models is constituted by heterogenous graphs in which nodes (rather than correspond all to action tubes) may be associated with any relevant element of a dynamic scene, such as object, agent, action, location, and their attributes (e.g. red, fast, drivable, etc) with the aim of achieving a truly complete description of complex dynamic events.

5 Conclusions

In this paper we presented a long-term spatiotemporal complex activity detection framework which uses a part-based deformable model. Our approach is based on three
Table 6: ROAD activity localization performance of each activity with different number of action tubes at standard fixed threshold of 0.5 IoU for mAP.

| # of Tubes | Negotiating in-intersection | Negotiating pedestrian crossing | Waiting in queue | Merging into vehicle lane | Sudden appearance | Walking middle of road |
|------------|-----------------------------|-------------------------------|-----------------|---------------------------|------------------|-----------------------|
| 3          | 0.51                        | 0.66                          | 0.83            | 0.78                      | 0.21             | 0.34                  |
| 5          | 0.54                        | 0.68                          | 0.85            | 0.79                      | 0.25             | 0.37                  |
| 6          | 0.52                        | 0.69                          | 0.86            | 0.80                      | 0.19             | 0.35                  |
| 7          | 0.40                        | 0.61                          | 0.72            | 0.69                      | 0.19             | 0.28                  |

Table 7: SARAS activity localization performance of each activity with different number of action tubes at standard fixed threshold of 0.5 IoU for mAP.

| # of Tubes | Phase#1 | Phase#2 | Phase#3 | Phase#4 | Phase#5 | Phase#6 | Phase#7 | Phase#8 |
|------------|---------|---------|---------|---------|---------|---------|---------|---------|
| 3          | 0.64    | 0.48    | 0.77    | 0.69    | 0.36    | 0.81    | 0.16    | 0.64    |
| 5          | 0.61    | 0.47    | 0.74    | 0.68    | 0.37    | 0.80    | 0.17    | 0.62    |
| 6          | 0.53    | 0.41    | 0.68    | 0.63    | 0.32    | 0.70    | 0.13    | 0.56    |
| 7          | 0.50    | 0.36    | 0.59    | 0.58    | 0.29    | 0.66    | 0.11    | 0.49    |

building blocks: action tube detection, part-based deformable RoI pooling for feature extraction, and a sparsity module which reduced the number of features by activating the truly discriminative components of the feature vectors. Besides proposing this novel framework, we also temporally annotated two recently released benchmark datasets (ROAD and ESAD) in terms of long-term complex activities. Both datasets come with video-level action tube annotation that make them suitable candidates to be used as benchmarks for future work in this area. We thoroughly evaluated our method, showing the effectiveness of our 3D part-based deformable model approach for the detection of complex activities.

References

[1] Mohammad M Arzani, Mahmood Fathy, Ahmad A Azirani, and Ehsan Adeli. Switching structured prediction for simple and complex human activity recognition. *IEEE transactions on cybernetics*, 2020.

[2] Yueran Bai, Yingying Wang, Yunhai Tong, Yang Yang, Qiuye Liu, and Junhui Liu. Boundary content graph neural network for temporal action proposal generation. In *European Conference on Computer Vision*, pages 121–137. Springer, 2020.

[3] Vivek Singh Bawa, Gurkirt Singh, Francis KapingA, Alice Leporini, Carmela Landolfo, Armando Stabile, Francesco Setti, Riccardo Muradore, Elettra Olarei, Fabio Cuzzolin, et al. Esad: Endoscopic surgeon action detection dataset. *arXiv preprint arXiv:2006.07164*, 2020.

[4] Harkirat Singh Behl, Michael Sapienza, Gurkirt Singh, Suman Saha, Fabio Cuzzolin, and Philip HS Torr. Incremental tube construction for human action detection. *arXiv preprint arXiv:1704.01358*, 2017.

[5] Pratool Bharti, Debraj De, Striram Chellapappan, and Sajal K Das. Human: Complex activity recognition with multi-modal multi-positional body sensing. *IEEE Transactions on Mobile Computing*, 18(4):857–870, 2018.

[6] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pages 961–970, 2015.

[7] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11621–11631, 2020.

[8] Fanta Camara, Nicola Bellotto, Serhan Cosar, Florian Weber, Dimitris Nathanael, Matthias Althoff, Jingyuan Wu, Johannes Ruenz, André Dietrich, Gustav Markkula, et al. Pedestrian models for autonomous driving part ii: high-level models of human behavior. *IEEE Transactions on Intelligent Transportation Systems*, 2020.

[9] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 764–773, 2017.

[10] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Da-
vide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Scaling egocentric vision: The epic-kitchens dataset. In Proceedings of the European Conference on Computer Vision (ECCV), pages 720–736, 2018.

[11] Rocco De Rosa, Nicolò Cesa-Bianchi, Ilaria Gori, and Fabio Cuzzolin. Online action recognition via nonparametric incremental learning. In BMVC, 2014.

[12] P. Felzenszwalb, D. McAllester, and D. Ramanan. A discriminatively trained, multiscale, deformable part model. In 2008 IEEE Conference on Computer Vision and Pattern Recognition, pages 1–8, 2008.

[13] Pedro Felzenszwalb, David McAllester, and Deva Ramanan. A discriminatively trained, multiscale, deformable part model. In 2008 IEEE conference on computer vision and pattern recognition, pages 1–8. IEEE, 2008.

[14] Pedro F Felzenszwalb, Ross B Girshick, and David McAllester. Cascade object detection with deformable part models. In 2018 IEEE Conference on Computer Vision and Pattern Recognition, pages 2241–2248. IEEE, 2010.

[15] Jiyang Gao, Zhenheng Yang, Kan Chen, Chen Sun, and Ram Nevatia. Turn tap: Temporal unit regression network for temporal action proposals. In Proceedings of the IEEE international conference on computer vision, pages 3628–3636, 2017.

[16] Ross Girshick, Forrest Iandola, Trevor Darrell, and Jitendra Malik. Deformable part models are convolutional neural networks. In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, pages 437–446, 2015.

[17] Guoqiang Gong, Xinghan Wang, Yadong Mu, and Qi Tian. Learning temporal co-attention models for unsupervised video action localization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 9819–9828, 2020.

[18] Junwei Liang, Lu Jiang, Juan Carlos Niebles, Alexander G Hauptmann, and Li Fei-Fei. Peeking into the future: Predicting future person activities and locations in videos. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 5725–5734, 2019.

[19] Fuchen Long, Ting Yao, Zhaofan Qiu, Xinmei Tian, Jiebo Luo, and Tao Mei. Learning to localize actions from moments. arXiv preprint arXiv:2008.13705, 2020.

[20] Chenxu Luo and Alan L Yuille. Grouped spatial-temporal aggregation for efficient action recognition. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 5512–5521, 2019.

[21] Haroon Iddrees, Amir R Zamir, Yu-Gang Jiang, Alex Gorbun, Ivan Laptev, Rahul Sukthankar, and Mubarak Shah. The thumos challenge on action recognition for videos “in the wild”. Computer Vision and Image Understanding, 155:1–23, 2017.

[22] Sujoy Paul, Sourya Roy, and Amit K Roy-Chowdhury. Wttalc: Weakly-supervised temporal activity localization and classification. In Proceedings of the European Conference on Computer Vision (ECCV), pages 563–579, 2018.

[23] Suman Saha, Gurkirt Singh, and Fabio Cuzzolin. Amt-net: Action-micro-tube regression by end-to-end trainable
deep architecture. In Proceedings of the IEEE International Conference on Computer Vision, pages 4414–4423, 2017.

[32] Suman Saha, Gurkirt Singh, and Fabio Cuzzolin. Two-stream amnet for action detection. arXiv preprint arXiv:2004.01494, 2020.

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

[34] Baifeng Shi, Qi Dai, Yadong Mu, and Jingdong Wang. Weakly-supervised action localization by generative attention modeling. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 1009–1019, 2020.

[35] Zheng Shou, Hang Gao, Lei Zhang, Kazuyuki Miyazawa, and Shih-Fu Chang. Autoclos: Weakly-supervised temporal action localization in untrimmed videos. In Proceedings of the European Conference on Computer Vision (ECCV), pages 154–171, 2018.

[36] Zheng Shou, Dongang Wang, and Shih-Fu Chang. Temporal action localization in untrimmed videos via multi-stage cnns. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 1049–1058, 2016.

[37] Gunnar A Sigurdsson, Gil Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In European Conference on Computer Vision, pages 510–526. Springer, 2016.

[38] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556, 2014.

[39] Gurkirt Singh, Stephen Akrigg, Manuele Di Maio, Valentina Fontana, Reza Javanmard Alitappeh, Suman Saha, Kossar Jeddisaravi, Farzad Yousefi, Jacob Culley, Tom Nicholson, et al. Road: The road event aware-ness dataset for autonomous driving. arXiv preprint arXiv:2102.11585, 2021.

[40] Gurkirt Singh, Suman Saha, and Fabio Cuzzolin. Predicting action tubes. In Proceedings of the European Conference on Computer Vision (ECCV) Workshops, pages 0–0, 2018.

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

[42] Vivek Singh Bawa, Gurkirt Singh, Francis KapingA, Inna Skarga-Bandurova, Elettra Oleari, Alice Leporini, Carmela Landolfo, Pengfei Zhao, Xi Xiang, Gongning Luo, et al. The saras endoscopic surgeon action detection (esad) dataset: Challenges and methods. arXiv e-prints, pages arXiv–2104, 2021.

[43] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. arXiv preprint arXiv:1212.0402, 2012.

[44] SwathiKiran Sudhakaran, Sergio Escalera, and Oswald Lanz. Lst: Long short-term attention for egocentric action recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 9954–9963, 2019.

[45] Nirmalya Thakur and Chia Y Han. An improved approach for complex activity recognition in smart homes. In International Conference on Software and Systems Reuse, pages 220–231. Springer, 2019.

[46] Limin Wang, Yu Qiao, and Xiaoou Tang. Action recognition and detection by combining motion and appearance features. THUMOS14 Action Recognition Challenge, 1(2):2, 2014.

[47] Chao-Yuan Wu, Christoph Feichtenhofer, Haoqi Fan, Kaiming He, Philipp Krahenbuhl, and Ross Girshick. Long-term feature banks for detailed video understanding. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 284–293, 2019.

[48] Huijuan Xu, Abir Das, and Kate Saenko. R-c3d: Region convolutional 3d network for temporal action detection. In Proceedings of the IEEE international conference on computer vision, pages 5783–5792, 2017.

[49] Mengmeng Xu, Chen Zhao, David S Rojas, Ali Thabet, and Bernard Ghanem. G-tad: Sub-graph localization for temporal action detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 10156–10165, 2020.

[50] Chuxu Zhang, Dongjin Song, Chao Huang, Ananthram Swami, and Nitesh V Chawla. Heterogeneous graph neural network. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pages 793–803, 2019.

[51] Da Zhang, Xiyang Dai, and Yuan-Fang Wang. Metal: Min-imum effort temporal activity localization in untrimmed videos. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 3882–3892, 2020.

[52] Hang Zhao, Antonio Torralba, Lorenzo Torresani, and Zhicheng Yan. Hacs: Human action clips and segments dataset for recognition and temporal localization. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 8668–8678, 2019.

[53] Xiaoankang Zhou, Wei Liang, I Kevin, Kai Wang, Hao Wang, Laurence T Yang, and Qun Jin. Deep-learning-enhanced human activity recognition for internet of health-
[54] Xizhou Zhu, Han Hu, Stephen Lin, and Jifeng Dai. Deformable convnets v2: More deformable, better results. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 9308–9316, 2019.

[55] Aneeq Zia, Andrew Hung, Irfan Essa, and Anthony Jarc. Surgical activity recognition in robot-assisted radical prostatectomy using deep learning. In International Conference on Medical Image Computing and Computer-Assisted Intervention, pages 273–280. Springer, 2018.