

Modeling Spatio-Temporal Human Track Structure for Action Localization

Guilhem Chéron · Anton Osokin · Ivan Laptev · Cordelia Schmid

Abstract This paper addresses spatio-temporal localization of human actions in video. In order to localize actions in time, we propose a recurrent localization network (RecLNet) designed to model the temporal structure of actions on the level of person tracks. Our model is trained to simultaneously recognize and localize action classes in time and is based on two layer gated recurrent units (GRU) applied separately to two streams, i.e. appearance and optical flow streams. When used together with state-of-the-art person detection and tracking, our model is shown to improve substantially spatio-temporal action localization in videos. The gain is shown to be mainly due to improved temporal localization. We evaluate our method on two recent datasets for spatio-temporal action localization, UCF101-24 and DALY, demonstrating a significant improvement of the state of the art.

Keywords Action localization · Recurrent network · Video.

1 Introduction

Successful action recognition will help us drive our cars, prevent crime, search our video collections and will eventually enable robots to serve us at home. Such applications require action localization, i.e. identifying when the action happens and who is performing the action. Most of the current methods and benchmarks for action recognition, how-

Guilhem Chéron * †
E-mail: guilhem.cheron@inria.fr

Anton Osokin ‡ *
E-mail: aosokin@hse.ru

Ivan Laptev *
E-mail: ivan.laptev@inria.fr

Cordelia Schmid †
E-mail: cordelia.schmid@inria.fr

* Inria Paris, ENS, CNRS, UMR 8548 (Willow team)

† Inria Grenoble, LJK, CNRS, Univ. Grenoble Alpes (Thoth team)

‡ National Research University Higher School of Economics, Moscow

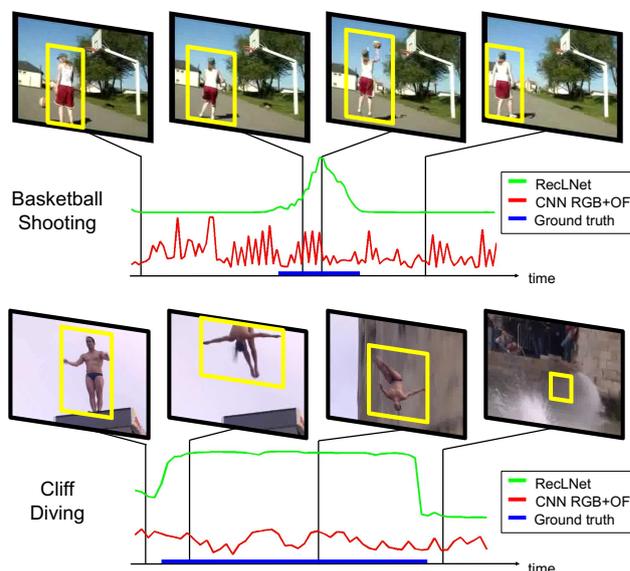


Fig. 1 Spatio-temporal action localization using a CNN baseline (red) and our RecLNet (green) both applied on the level of person tracks. Our approach provides accurate temporal boundaries when the action happens.

ever, only address action classification [11, 52], i.e. assuming temporally segmented action intervals as input.

Identifying the beginning and the end of an action naturally suggests the need of temporal models for video sequences. Sequence models have previously been explored for sound, speech and text understanding. In particular, recurrent neural network models (RNNs) have recently shown success for speech recognition [10] and text generation [58] as well as for image and video captioning [12, 27, 64]. RNNs have also been explored for action classification in video [12, 41], but have shown limited improvements for this task so far.

Action classification may not require sophisticated temporal models if classes can be distinguished solely by the

presence of action-specific features. On the other hand, if the structure of the video is required as the output, explicit spatio-temporal models of the video can be beneficial. Recent work [36, 38, 53, 69, 70] has indeed shown improvements in temporal action localization achieved with recurrent models of video sequences. Here we develop and investigate recurrent models for spatio-temporal action localization.

Our goal is to localize the acting person in the video frame and to identify temporal boundaries of corresponding actions. To this end, we propose a recurrent localization network (RecLNet) with gated recurrent units (GRU) [9] for modeling actions on the level of person tracks. Our method starts by person detection and tracking similar to [16, 43, 49, 66]. Differently to previous work, we train our RecLNet to score actions and detect temporal boundaries within each person track (see Figure 1). This scoring is achieved by two-stream recurrent units exploiting appearance and motion FastRCNN [15] features pooled from person boxes, while final detections are obtained using our simple and effective temporal localization method composed of filtering and thresholding. We provide a thorough experimental evaluation and analyze the impact of recurrence on spatio-temporal action localization by making the following contributions:

- we show spatio-temporal action localization improvement supported by our RecLNet trained on a track-level and compare different standard and recurrent architectures;
- we empirically diagnose *temporal* localization as being a weakness of existing methods which our method is able to correct;
- our method is complementary to most recent works [26] mainly focusing on increasing spatial boxes precision and we identify the spatial aspect as being our principal room for improvement;
- results are reported on the two largest datasets for our task, namely UCF101-24 detection [56] and DALY [67], for both of these datasets our method results in significant improvements over the state of the art.

The rest of the paper is organized as follows. Section 2 reviews the related work on action classification, temporal and spatio-temporal localization. Section 3 introduces our RecLNet model, its architecture and our threshold temporal localization technique. Section 5 presents our experimental and qualitative results. Section 6 finally draws conclusions.

2 Related work

Our work is mostly related to methods for human action classification, temporal action localization and spatio-temporal action detection in video.

Action classification. The majority of action recognition methods targets clip-level video classification, i.e. the as-

ignment of video clips to a closed set of action classes. Recent datasets for this task include UCF-101 [56], HMDB [33], ActivityNet [20] and Sports-1M [28]. Local space-time features such as HOG, HOF and IDT [35, 65] have shown initial progress for this task. More recently CNN and RNN-based approaches have been investigated to learn video representations for action recognition. A combination of motion and appearance information learned by two separate CNN networks has been proposed in [52]. Alternative methods based on spatio-temporal convolutions have been studied in [25, 59] and more recently using the C3D [61] and I3D [7] architectures. Such method [7] takes advantage of transfer learning by training 3D architectures on a large video datasets [29]. Several works used RNNs for aggregating video information along time [2, 12, 41, 45, 53, 57]. Currently best performing action classification methods combine IDT features with CNN-based representations of motion and appearance [11, 13, 14, 63] or pure 3D CNN architecture [7]. RNNs have shown promise for the task of gesture recognition in [45] but did not show significant improvements for the more general task of action classification. In this work, we compute appearance and motion CNN features extracted on the level of person tracks and use them as input to our RecLNet for action localization.

Temporal localization. Temporal action localization aims both to classify and identify temporal extents of actions in longer video clips. Recent datasets for this task include THUMOS [24], Activity-Net [20] and MPII Cooking [48]. Methods for joint action classification and temporal segmentation have explored dynamic programming [21] and temporal grammars [46]. More recently several RNN-based methods have shown gains for action localization in [3, 38, 53, 69, 70]. For example, [38, 53, 70] explore variants of LSTM to improve per-frame action classification whereas the method in [69] learns to directly predict action boundaries. An alternative approach based on 3D CNNs and temporal action proposals has shown competitive results in [51] while [71] extends proposals in time and segments them in stages to evaluate their “completeness” based on structured pyramid pooling. Here we use GRU as recurrent units for temporal modeling of actions in our RecLNet. Unlike previous work on temporal action localization, however, we address a much more challenging task of localizing actions in space and time.

Spatio-temporal detection. Spatio-temporal action detection aims to find locations of actions in space and time. The list of datasets for this task is limited: UCF101-24 – a subset of UCF-101 [56] used in the THUMOS challenge [17], the DALY dataset [67]. Other datasets that we are aware of have either a very limited number of examples or consists of temporally trimmed videos. Some of the earlier works explore volumetric features and 3D sliding window detectors [31, 34]. More recent methods have extended ideas of object proposals in still images to action proposals in video [16, 42,

62]. The common strategy in [16, 43, 49, 54, 66, 67] is to localize actions in each frame with per-frame action or person detectors and to link resulting bounding boxes into continuous tracks. The temporal localization is then achieved by the temporal sliding windows [66, 67] or dynamic programming [43, 49, 50, 54]. Instead of relying on per-frame detections, Saha et al. [50] regress pairs of successive frames and Hou et al. [23] generate clip proposals from 3D feature maps. An approach of Zolfaghari et al. [72] uses a CNN to sequentially fuse human pose features in addition to the standard appearance and optical flow modalities. Most recent method [26] relies on SSD detector [37] adapted to spatio-temporal anchors and generates human tracks with tubelets linking. [55, 68] investigate unsupervised spatio-temporal action localization but this is outside of the scope of our work. In this paper, we propose a recurrent localization network (RecLNet) that both classifies and localizes actions within tracks supported by a thresholding and filtering method. Our analysis shows that RecLNet provides significant gains due to accurate temporal localization and its complementarity to the recent method [26] with more accurate spatial localization but approximate temporal localization. Our resulting approach outperforms the state of the art in spatio-temporal action detection [26, 44, 49, 54, 67] on two challenging benchmarks for this task.

3 Action localization

This section presents our method for spatio-temporal action localization. The overview of the method is illustrated in Figure 2. A spatially localized person track (Figure 2, row 1) is passed to appearance and optical flow feature extractors (Figure 2, row 2). These descriptors feed our localization network composed of 3 layers for each stream and 1 fusion layer. In each stream, the first layer (Figure 2, row 3) normalizes either appearance or flow features before sending them to a stack of two GRU layer units. The GRU outputs from both stack levels and both streams are concatenated (Figure 2, row 5) and converted by a fully-connected layer (Figure 2, row 6) to action probabilities (Figure 2, last row).

In the following, we first briefly present the inputs of our method, namely the human tracks and their associated features. Then, we introduce our spatio-temporal action localization method based on the recurrent localization network (RecLNet). Finally, we discuss how to post-process the action detection scores in order to output final spatio-temporal human action tubes.

3.1 Person tracks

We obtain human tracks as in [49, 67] by first running an action or person detector in each video frame, and then link-

ing detections into human tracks spanning the entire video clip. Our method aims at segmenting the track in time to obtain temporal action boundaries. For this purpose, we associate to each time frame its corresponding bounding-box in the human track. This box is used as a pooling region to extract per-frame descriptors. Such features are obtained by Fast-RCNN [15] appearance and flow ROI-pooling. The details on state-of-the-art human tracks and their associated features used in this work will be discussed in Section 4.

3.2 Temporal action localization

This section presents our model for action localization and its training procedure on the level of person tracks. In our work, we choose to adopt a model with memory links, like in a recurrent neural network (RNN), which we call recurrent localization network (RecLNet), to temporally localize actions. RNNs have been demonstrated to successfully model sequential data especially for language tasks such as speech recognition [10], machine translation [4] or image captioning [27]. Given this success, we believe that recurrent networks are well-suited for modeling temporal sequences of appearance and motion in person tracks. Our final RecLNet model is composed of gated recurrent units (GRU) described below.

The LSTM and GRU architectures. Features from a human track of length T can be seen as an input sequence $\mathbf{x} = (x_1, \dots, x_T)$ where for each x_i we aim to provide action activation h_i , forming the output $\mathbf{h} = (h_1, \dots, h_T)$. To generate such output, we investigate two types of recurrent networks for our RecLNet, namely LSTM and GRU as defined below.

In the long short-term memory (LSTM) architecture [22], one memory cell and three ‘gates’ give LSTM the ability of discovering long-range temporal relationships by reducing the vanishing gradient problem compared to vanilla RNN. This is a useful property in our task since we need to handle particularly long human tracks. The LSTM cell takes as input features x_t at time step t together with the output h_{t-1} at the previous time step $t - 1$ and operates as follows:

$$\begin{aligned} f_t &= \sigma(W_f x_t + U_f h_{t-1} + b_f) \\ i_t &= \sigma(W_i x_t + U_i h_{t-1} + b_i) \\ o_t &= \sigma(W_o x_t + U_o h_{t-1} + b_o) \\ c_t &= f_t \odot c_{t-1} + i_t \odot \tanh(W_c x_t + U_c h_{t-1} + b_c) \\ h_t &= o_t \odot \tanh(c_t) \end{aligned} \quad (1)$$

where σ is the sigmoid function and f_t, i_t, o_t, c_t are the ‘forget gate’, ‘input gate’, ‘output gate’ and ‘memory cell’, respectively. Matrices W, U and vectors b denote the parameters of the cell.

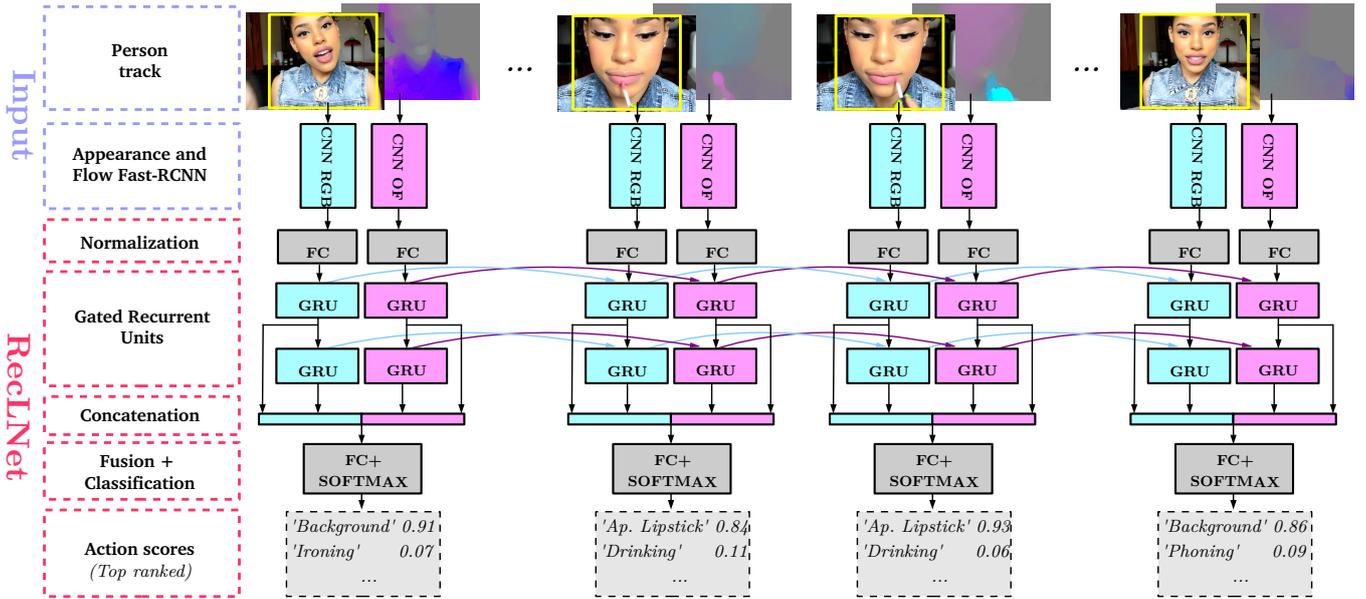


Fig. 2 Our ReCLNet approach for spatio-temporal action localization. The input is a person track where for each spatially localized actor bounding-box, we extract appearance (RGB) and optical flow (OF) CNN features. Each stream is normalized by a fully-connected layer and fed into a two-layer GRU. The outputs from both GRU levels and from both streams are concatenated then classified with a fully-connected layer combined with softmax scoring. Outputs are class probabilities for each frame.

The GRU [9] cell differs from LSTM by the absence of the output gate and operates as follows:

$$\begin{aligned}
 z_t &= \sigma(W_z x_t + U_z h_{t-1} + b_z) \\
 r_t &= \sigma(W_r x_t + U_r h_{t-1} + b_r) \\
 h_t &= z_t \odot h_{t-1} \\
 &\quad + (1 - z_t) \odot \tanh(W_h x_t + U_h (r_t \odot h_{t-1}) + b_h)
 \end{aligned}
 \tag{2}$$

where z_t and r_t are the ‘update gate’ and ‘reset gate’, respectively. The GRU cell is simpler, has less parameters and has shown some improvement on video tasks as in [60]. An empirical comparison of LSTM and GRU cells is given in Section 5.

We define the **Recurrent Localization Network (ReCLNet)** as a multi-class recurrent network trained to classify actions against background. As shown in Figure 2, at each time step, appearance (RGB) and optical flow (OF) Fast-RCNN networks take bounding boxes of human tracks as object proposals and extract features. Each stream (RGB and OF) is processed independently by feeding Fast-RCNN outputs (FC7 layers) to our ReCLNet which produces action scores at each time step. Each stream of our ReCLNet consists of a fully-connected layer that normalizes the input features (appearance or flow) and a stack of two GRU layers. Note the second GRU layer takes as input the output of the first one while their output is concatenated to a $2 \times M$ -dimensional stream output (where M is the memory size). Finally the appearance and flow branch outputs are concatenated and a last fully-connected layer with softmax converts the recurrent output to an action probability for all actions (and back-

ground). The network is trained using the standard negative log-likelihood loss w.r.t. all C action classes and background boxes:

$$L(W) = - \sum_{i=1}^N \sum_{c=1}^{C+1} \mathbb{1}\{y_i = c\} \log P(y_i = c | x_i, W), \tag{3}$$

with probabilities defined by softmax:

$$P(y_i = c | x_i, W) = \frac{e^{f_c(x_i, W)}}{\sum_{c'=1}^{C+1} e^{f_{c'}(x_i, W)}}. \tag{4}$$

Here $f_c(x_i, W)$ is the output of the last fully-connected layer corresponding to class c , x_i and y_i denote features and labels at time step i , symbol W denotes network parameters, N is the number of boxes in the training set, C is the number of action classes, $\mathbb{1}\{\cdot\}$ is the indicator function.

To train ReCLNet, we set the ground-truth targets y_i in the following way. We assign label c to a frame bounding-box detection from a person track if it overlaps more than 0.3 spatial IoU with one ground truth annotation from action label c , otherwise this input is considered as background. Having low IoU threshold allows us to get more positives.

Appearance and optical flow fusion. Single-stream networks are first independently trained then we consider three fusion methods to combine their appearance (RGB) and optical flow (OF) outputs. The *average* simply averages both softmax RGB and OF network outputs. Both next fusion methods are trained on top of the two stream networks (their weights remaining fixed). The *gating layer* learns per-class

weights to multiply both network outputs before their original softmax layer and applies softmax after summing weighted class outputs. The third method, *fusion layer*, trains a fully-connected classification layer (followed by softmax) on top of the concatenated memory units outputs from each stream network (concatenation layer in Figure 2).

3.3 Post-processing of human action tracks

RecLNet output represents action scores at each time step of a human track. While being spatially localized, these scored tracks have to be segmented in time in order to produce the final spatio-temporal detections. For the track ranking, it is also necessary to score each of the final detections.

In this section, we then describe our method to get the final spatio-temporal detections and their associated score. Temporal localization is performed within each track.

Threshold. Let a be the action to segment and τ a human track spanning from frame 1 to T of scores $\mathbf{s}_\tau = (s_1^a, \dots, s_T^a)$ with s_t^a the score associated to the person box at time t . The goal of temporal segmentation is to extract from τ one or several sub-tracks ν (of time interval $[t_1^\nu, t_2^\nu] \subseteq [1, T]$) as final spatio-temporal detections.

For this purpose, we first get smoother scores \hat{s}_i^a by applying a median window filtering on the s_i^a . Then, we temporally segment the track by selecting consecutive boxes with scores above a certain threshold θ , while others are rejected. More formally, considering that consecutive boxes $(b_{t_0}, \dots, b_{t-1})$ from time t_0 to $t-1$ have already been added to the sub-track ν , the next box b_t is added to ν if $\hat{s}_t^a \geq \theta$. Otherwise, ν ends and is returned as final detection. This method allows to break the initial track into several sub-track candidates of arbitrary lengths and is then able to capture several repetitive action instances happening on the same human track (like *drinking*, *applying make up*, *playing harmonica*, see Figure 4 and Figure 5). Here, model outputs must be smooth in order to get accurate action temporal boundaries. The recurrent units (like GRU, LSTM) are then well-suited for this localization method while appearance and flow CNNs output is generally noisier. Our temporal localization technique is referred to as *threshold*.

Temporally segmented track scoring. In order to rank detections and perform the average-precision (AP) evaluation, we need to set a score r_ν for each final spatio-temporal detection ν . Following, e.g., [49], we define r_ν as the average of the top 40 action scores s_i^a contained in ν . The same scores are also used for non-maximum-suppression (NMS) of spatio-temporal detection candidates based on their scores and overlap. For NMS, we use spatio-temporal Intersection-over-Union¹ as overlap criterion.

¹ The spatio-temporal IoU between two tracks is defined as a product of temporal IoU between the time segments of the tracks and average spatial IoU on the frames where both tracks are present.

3.4 Implementation details

RecLNet parameters. The FC7 output of each Fast-RCNN is 4096-dimensional and the first fully-connected layers convert each stream to a 1024-dimensional vector. The memory size M is equal to 256 and the last fully-connected layer input is 1024-dimensional ($2 \times 2 \times M$, i.e. the memory stack from both GRU layers of both streams).

Training. Both appearance and flow branches of RecLNet are separately trained using the Adam optimizer [32] with a weight decay set to $5 \cdot 10^{-4}$ to avoid overfitting. Note that to train the single-stream networks, we halve the input dimension of the last fully-connect layer of RecLNet (concatenation layer in Figure 2). Training batches contain 100 different tracks of temporal length 20. Backpropagation through time (BPTT) is then performed every 20 time steps.

Detection. In all experiments, the NMS overlap threshold is set 0.2, the action localization *threshold* θ is set to 0.1 and the median window size is 25.

Optical flow. To obtain the optical flow data, we compute horizontal and vertical flow for each consecutive pair of frames using the approach of [6]. Following [16, 66], flow maps are saved as 3-channels images corresponding to optical flow in x and y direction and its magnitude with all the values restricted to the interval $[0, 255]$.

4 Experimental setup

This section describes UCF101-24 [56] and DALY [67] datasets used for evaluation of our method. For both datasets, we provide experimental details on the use of ground truth annotation and data pre-processing.

4.1 UCF101-24

The original version of the UCF-101 dataset [56] is designed for action classification and contains 13321 videos for 101 action classes. The task of spatio-temporal action localization is defined on a subset of 24 action classes (selected by [17]) in 3207 videos. We refer to this subset as ‘‘UCF101-24’’. Each instance of an action is manually annotated by a person track with the temporal interval corresponding to the interval of an action. In our training and testing, we use recently corrected ground truth tracks [49].² Each UCF-101 video contains actions of a single class.

Short vs. long classes. While some of the UCF101-24 action classes are short (‘basketball dunk’ or ‘tennis swing’)

² <https://github.com/gurkirt/corrected-UCF101-Annots>

other continuous actions (‘biking’ or ‘rope climbing’) typically last for the full duration of the video. To better evaluate the temporal localization on UCF101-24, we define a subset with short action classes that on average last less than a half of the video length. This subset contains six actions (*Basketball*, *Basketball Dunk*, *Cricket Bowling*, *Salsa Spin*, *Tennis Swing* and *Volleyball Spiking*) and we call them “Short classes”. In Section 5, we evaluate localization for all 24 action classes and for Short classes separately.

Performance evaluation. To evaluate the detection performance, we use the standard spatio-temporal IoU criterion defined by the UCF101-24 benchmark. The detected action tube is considered to be correct if and only if its intersection with the ground-truth tube is above the criterion threshold and if both tubes belong to the same action class. Duplicate detections are considered as false positives and are penalized by the standard precision-recall measure. The overall performance on the UCF101-24 dataset is compared in terms of mean average precision (mAP) value.

Action-specific human tracks. To enable direct comparison of our method with [49], for UCF101-24 experiments we use the same person tracks as in [49]. These tracks are obtained by linking per-frame action detections with dynamic programming (DP). Action detections are obtained with the Fast-RCNN method [15] trained with appearance and flow input for the task of spatial action localization. As our model performs its own temporal localization, we do not run the temporal segmentation of [49] (second DP pass) and keep 5 action proposals per action covering the whole video. Per-frame input features for our ReLNet are obtained from the same Fast-RCNN detector used for track detection.

4.2 DALY

DALY [67] is a recent large-scale dataset for action localization containing 510 videos (31 hours) of 10 different daily activities such as ‘brushing teeth’, ‘drinking’ or ‘cleaning windows’. There is only one split containing 31 train and 20 test videos per class. The average length of the videos is 3min 45s. In contrast to UCF101-24, all actions are short w.r.t. the full video length, making the task of temporal action localization more challenging. DALY may contain multiple action classes in the same video. The DALY dataset provides ground-truth temporal boundaries for all the action instances and the spatial annotation (bounding boxes) for few keyframes of each instance.

Human tracks and associated features. To enable direct comparison with [67] we use tracks provided by [67]. An accurate Faster-RCNN [47] person detector is trained on the large MPII human pose dataset [1]. The tracks are obtained by linking human detections with a tracking-by-detection approach. Similarly to UCF101-24, the appearance and flow

features are obtained with a Fast-RCNN detector trained to detect actions on the annotated DALY frames. The action scores from this detector will be used as a baseline in Section 5 and referred to as *CNN RGB+OF*. Following [39], we compensate the annotation bias by adapting human box sizes to action annotation. To achieve this, we train a linear regression of bounding boxes. For each DALY annotated keyframe from the training set we associate its most overlapping bounding-box from the human tracks (with at least 0.5 IoU) to train the linear model.

DALY tracks labeling. To compensate for the sparse action annotation in DALY, we extend ground truth to all frames of the actions with automatic tracking. We use the Siamese Fully-Convolutional Network for online non-class-specific tracking method³ [5] initialized on all ground-truth keyframes. For a given action instance, we aggregate all tracks at each frame by a median bounding box. To assess the quality of generated ground-truth tracks, we computed the track proposals recall (92.2% at 0.3) within the action time interval. Also, the IoU overlap between ground-truth tracks and annotated keyframes is greater than 0.8 (resp. 0.5) for 91.4% (resp. 99.0%) of keyframes.

5 Experiments

In this section, we first evaluate the impact of the *threshold* temporal localization, recurrent architecture and the fusion method (Section 5.1). We then evaluate the potential gain due to *action classification* (Section 5.2) followed by an extensive analysis on *temporal localization* (Section 5.3). Next, we show an improvement if using I3D and temporal person tracks (Section 5.4). We compare our ReLNet to the state of the art on UCF101-24 and DALY datasets (Section 5.5). Section 5.6 concludes this experimental part by presenting qualitative results.

5.1 Impact of localization method, recurrent architecture and fusion

Localization method.

In [49], temporal localization is performed using the Viterbi algorithm on top of action tracks spanning the whole video. To temporally trim these tracks, one binary label (action vs. background) is associated to each bounding-box by maximizing an energy function where unary potentials represent box action scores while pairwise potentials control the final action length (or track smoothness). This smoothness score is weighted per class and then might over control the track length towards the action durations seen on the training set while, in our method, median filtering encodes smoothness

³ <https://github.com/bertinetto/siamese-fc>

Method	UCF101-24			DALY		
	OF	RGB	RGB+OF	OF	RGB	RGB+OF
CNN [49]	59.3	55.2	60.4	-	-	-
CNN	-	-	-	9.9	11.9	13.4
FC	60.7	57.7	64.0	11.5	13.6	16.1
LSTM	65.0	58.7	66.5	13.1	11.9	16.2
GRU	67.0	59.5	67.1	14.4	14.2	17.4

Table 1 Performance on UCF101-24 and DALY for different recurrent architectures (*LSTM* and *GRU*) and baselines with no temporal connections (*CNN* and *FC*). Models are evaluated with flow (OF) and/or appearance (RGB) features as input (RGB+OF averages both stream outputs) for spatio-temporal localization at IoU = 0.3 (mAP).

more explicitly and suffers less from the dataset biases described in 4.1.

In Table 2, we refer to the track action scores from [49] as *CNN RGB+OF*, since they come from two Fast-RCNN where appearance and optical flow outputs have been fused. We recall that to enable a direct comparison with [49] we are using the same human tracks as [49] but score them differently (see Section 4 for details). Here, we compare the Viterbi algorithm originally used in [49] for temporal localization and our *threshold* method on original scores (*CNN RGB+OF*). Interestingly we observe that our *threshold* technique improves over their original results by +4.9% and conclude that thresholding combined with median filtering works better than this typically used localization method while being simpler. The next paragraph shows the impact of differently scoring, among others, tracks from [49]. In the following, only our *threshold* temporal localization is used.

Method	Localization	UCF101-24
CNN RGB+OF	Viterbi [49]	55.5
CNN RGB+OF	threshold	60.4

Table 2 Localization method analysis on UCF101-24. We compare the temporal localization of [49] which uses *Viterbi* algorithm to our *threshold* method. We apply the two methods on original detection scores from [49] (*CNN RGB+OF*). Evaluation is spatio-temporal action localization at IoU 0.3 (mAP).

Recurrent architecture. Table 1 compares models based on LSTM and GRU units (see Section 3 for details). As a baseline we use a model with no temporal connections but similar architecture: the recurrent units are replaced by a stack of 2 fully-connected layers with non-linearity and has the same number of parameters as the GRU unit. This additional fully-connected classifier is referred to as *FC*. For UCF101-24 evaluation, we again report original track scores (*CNN* [49] as in Table 2) while for DALY, *CNN* represents the Fast-RCNN scores similarly retrained as authors [67] on their human tracks (see Section 4.2 for details). Input modalities are either, optical flow features (*OF*), appearance features (*RGB*) or both (*RGB+OF*) where action scores from

the two stream outputs are averaged in this case. Spatio-temporal action localization is evaluated on UCF101-24 and DALY datasets at spatio-temporal IoU of 0.3.

We first observe that training an additional classifier (*FC*) is better than directly taking Fast-RCNN outputs (*CNN*) and improves performance by +3.6% and +2.7% on UCF101-24 and DALY respectively when both appearance and optical flow streams are used (*RGB+OF*). Also, non-recurrent baselines (*CNN* and *FC*) perform worse than the recurrent variants (*LSTM* and *GRU*). When using both *RGB+OF* as input, the gain due to recurrence is +6.7% and +4.0% on UCF101-24 and DALY respectively when comparing *GRU* to *CNN* and +3.1% and +1.3% when comparing *GRU* to *FC*. Interestingly, recurrent unit improvement is larger on optical flow +6.3% on UCF101-24 and +2.9% on DALY when comparing *GRU* to *FC*. This shows that temporal memory links are beneficial for better spatio-temporal action localization performance. We also note, as this is often the case when working with videos (e.g. [60]), that *GRU* outperforms *LSTM* while being a simpler model.

This experiment has shown *GRU* memory unit achieves the best action localization accuracy. In the following, *GRU* will then be used in our recurrent model for determining the action temporal extent in all experiments.

Fusion method. Table 3 explores the different fusion strategies described in Section 3 to combine appearance (RGB) and flow (OF) features in our localization model (with *GRU* layer which was validated in the previous experiment). We first note that all fusion methods improve action localization results on both UCF101-24 and DALY. However, the simple averaging method is not able to capture feature complementarity between appearance and flow features especially on UCF101-24 where it gets only +0.1% improvement compared to the best performing OF features. The *gating layer* is able to take advantage of features combination by substantially improving on both datasets. Finally, the *fusion layer* better captures features complementarity and improves action localization mAP on UCF101-24 and DALY respectively by +2% and +5.3% compared to the single flow features and by +9.5% and +5.5% compared to appearance features.

In the following, our final recurrent model uses the *fusion layer* and is referred to as *RecLNet* in all experiments.

Finally, when using our *threshold* method, re-scoring the tracks with our RecLNet instead of taking Fast-RCNN outputs (*CNN*) as in [49, 67] improves spatio-temporal action localization results by +8.6% and +4.0% on UCF101-24 and DALY respectively (from Table 1). The improvement is even larger (+13.5%) when comparing RecLNet accuracy (69.0%) to the original result using scores and temporal localization from [49] (55.5% on UCF101-24 from Table 2).

Given that we are using the same tracks, this first gives an insight that our method improvement compared to [49, 67] might be due to *temporal* localization. This question is further studied in the next sections.

Features	Fusion	UCF101-24	DALY
OF	-	67.0	14.4
RGB	-	59.5	14.2
RGB+OF	<i>average</i>	67.1	17.4
RGB+OF	<i>gat. layer</i>	69.0	18.1
RGB+OF	<i>fusion layer</i>	69.0	19.7

Table 3 Performance on UCF101-24 and DALY for different features and fusion strategies *average*, *gating layer* and *fusion layer* (see Section 3 for details). Models with *GRU* memory units take optical flow (OF) and/or appearance (RGB) features as input and are evaluated for spatio-temporal localization at IoU = 0.3 (mAP).

5.2 Analyzing the performance gain: *action classification*

Spatio-temporal action localization is composed of *spatial localization*, *action classification* and *temporal localization*. In this section we focus on action classification and analyze its performance given person tracks and pre-defined temporal action boundaries.

Evaluating action classification. Here, we do not require temporal localization by restricting the track to the ground-truth time interval in order to evaluate the performance gain due to action classification. As described in Section 3, the track scoring is obtained with the average over the top 40 action scores s_i^a in the trimmed interval. Table 4 shows our method (*RecLNet*) increases action classification by approximately +2% compared to scores from [49] (*CNN RGB+OF*) when evaluated on UCF101-24 classes. This improvement is moderate compared to the +13.5% spatio-temporal localization boost of Section 5.1.

IoU	0.2	0.3	0.4
CNN RGB+OF [49]	84.3	81.6	75.6
RecLNet	86.9	83.7	77.6

Table 4 Performance on UCF101-24 for clips trimmed to the ground-truth temporal interval (mAP at IoU 0.2, 0.3 and 0.4).

Table 5 shows the same experiments on DALY. We observe that our *RecLNet* method performs slightly worse (from -0.1% to -0.9%) compared to the *CNN RGB+OF* baseline (see Section 4.2 for details). This might be due to DALY track labeling which, contrary to UCF101-24, is obtained by automatic tracking of ground-truth sparse annotations (see Section 4.2) which introduces some noise and can slightly affect action classification. We, therefore, conclude that the spatio-temporal action localization gain of our method over the baseline, demonstrated in Section 5.1, cannot come from action classification.

IoU	0.2	0.3	0.4
CNN RGB+OF	65.5	64.8	63.6
RecLNet	65.4	64.5	62.7

Table 5 Performance on DALY for clips trimmed to the ground-truth temporal interval (mAP at IoU 0.2, 0.3 and 0.4).

Overall, this experiment confirms results in the literature [41] that recurrent units (RNN, LSTM, GRU) do not really improve the performance for action classification. The spatial localization (the person boxes positions) being fixed, the gain of our method for spatio-temporal action localization has to come from better temporal localization. This observation is analyzed in detail in Section 5.3.

5.3 Analyzing the performance gain: *temporal localization*

In this section, we show the main gain of RecLNet comes from better temporal localization and demonstrate its complementarity with the state of the art [26]. We first motivate this study by explaining the UCF101-24 bias, then a per-class analysis compares our RecLNet to [26] and to the method we build on [49]. Finally, we investigate the potential room for improvement.

UCF101-24 bias. As described in Section 4.1, UCF101-24 contains only six short action classes lasting less than half of the video duration, while 17 actions (74% of the dataset classes) span at least 70% of the video length in which 11 of them (48% of the dataset classes) span even more than 90% of the video length. UCF101-24 results are then biased toward long actions which do not require temporal localization. Indeed, tracks spanning the whole video already achieves good temporal localization for these classes (note that such detections would totally fail on the DALY dataset). To avoid this bias and to focus on challenging cases of temporal localization, we next compare our method to [26, 49] on the subset of short classes along with whole dataset.

	Basketball	BasketballDunk	Cricketbowling	SalsaSpin	Volleyballspiking	Tenniswing	mAP-short	Biking	Cliffdiving	Diving	Fencing	FigureSkating	Golfswing	Horsebackriding	Judothrowing	Lombumping	PoleVault	RollerClimbing	Shuttlecock	Skiing	ShiJi	Soccerjuggling	Surfing	Tabletennisbackhand	Tabletennisforehand	mAP
[49]	9.8	11.9	7.3	18.8	1.0	27.2	12.7	56.1	67.1	82.9	80.5	99.7	59.8	88.1	64.1	40.1	49.6	78.4	85.1	70.7	92.7	84.9	55.4	36.9	67.1	55.5
RecLNet	69.4	38.8	50.8	25.4	26.5	53.0	44.0	65.0	75.6	93.4	80.6	99.6	70.3	93.0	55.4	65.6	56.9	99.3	91.6	84.8	89.3	92.3	51.6	47.5	80.5	69.0
[26]	18.0	44.1	23.5	20.7	1.0	39.1	24.4	72.0	82.0	84.8	81.3	99.2	76.7	94.6	60.8	82.5	90.4	92.4	92.0	84.0	80.4	91.4	65.9	62.3	76.8	67.3

Table 6 Per-class AP for IoU 0.3 on UCF101-24 comparing our RecLNet to state of the art [49, 26]. We report mean AP for short classes, *mAP-short*, and all classes, *mAP*. The 6 short classes are reported on the left (before *mAP-short* column).

Per-class performance on UCF101-24. Table 6 compares the per-class results of our approach with [49] since we are using their human tracks and with the best performing state-of-the-art method [26] (see Section 5.5 for comparison with the state of the art). We report mAP for short classes only (*mAP-short*) and also for all classes (*mAP*). We outperform [49] by +31.3% on average on short classes and by +59.6% on “Basketball”. As we have already seen previously, the overall improvement is +13.5%. We can observe that for some “long” actions the performance drops slightly. Similarly, we outperform the state of the art [26] by only +1.7% overall while we reach a large improvement of around +20% on short classes.

IoU	All classes			Short classes		
	0.3	0.5	0.75	0.3	0.5	0.75
[26]	67.3	51.4	22.7	24.4	2.6	0.0
RecLNet	69.0	46.5	10.3	44.0	6.4	0.0

Table 7 Performance on UCF101-24 when differentiating short classes from others. Spatio-temporal action localization (mAP) is evaluated at IoU 0.3, 0.5 and 0.75).

Table 7 now compares our model to the best performing method [26] at higher IoU. We observe that even if short classes get extremely difficult to detect at highest IoU (0% mAP at IoU 0.75), at IoU 0.5, our model still outperforms the state of the art by +3.8% on short classes while loosing −4.9% overall (Section 5.5 studies this latter result when comparing to the state of the art).

These experiences on per-class performance show that our model is able to improve [49] by producing more accurate temporal action localization. Also, while [26] takes advantage from strong features and spatially accurate person boxes to get excellent accuracy on long classes (e.g. compared to [49]), its performance on short classes suffers from approximate temporal localization. This demonstrates the potential complementarity between our RecLNet and the current best performing method [26]. The room for improvement of RecLNet will be studied in the next section.

Correct temporal loc.	Correct spatial loc.	Correct class	class mAP @0.75	
			short	all
-	-	-	0.0	10.3
✓	-	-	7.7	15.8
-	✓	-	11.1	54.0
-	-	✓	0.0	17.0
✓	-	✓	11.1	23.8
-	✓	✓	17.2	60.8
✓	✓	-	41.7	70.5
✓	✓	✓	43.2	73.9

Table 8 RecLNet performance on UCF101-24 short and all classes when considering different components to be correct. Spatio-temporal action localization (mAP) is evaluated at IoU 0.75). **Spatial** localization is the largest room for improvement of our method.

Toward action localization improvement. This paragraph analyzes how to improve the spatio-temporal action localization performance of our model especially when evaluating at very high IoU (0.75). To distinguish the potential improvements, we can consider *action classification*, *spatial localization* and/or *temporal localization* as being correct. We proceed as follow. When we suppose perfect *action classification*, we set final spatio-temporal detection score r_v (see Section 3.3) to 0 for all false positives (note this is equivalent to the recall). Let o_s (resp. o_t) the spatial IoU overlap of a final detection with its ground-truth. Thereby, when considering perfect *spatial* (resp. *temporal*) localization, if o_s (resp. o_t) is greater than 0.3, it is then set to 1 in the spatio-temporal IoU computation. We choose 0.3 as it seems fair enough 1) not to get a track completely shifted in time ($o_t \simeq 0$) for which the spatial IoU o_s would have been computed on only one frame or a few and 2) to ensure that the track at least approximately spatially “follows” ($o_s \neq 0$) the person. Table 8 presents the combinations of these assumptions. First, assuming correct *temporal localization* or *action classification* only improves our model performance by respectively +5.5% and +6.7%. However, the *spatial* assumption is by far the best room for improvement of our model (+43.7%) and achieves 54.0% accuracy. It also shows our method gets 0% accuracy on short classes mostly because of inaccurate human track bounding-boxes as the spatial assumption improves it to 11.1%. Of course, combining several assumptions further improve the performance

(note that last line does not reach 100% since the track recall is not 100% and that the above “0.3 criterion” eliminates some track candidates). However, we observe that combining correct *temporal localization* and *action classification* (23.8%) is still far from the single *spatial localization* assumption (54.0%). This experiment confirms that improving spatial human detection is the direction to take further enhancements and demonstrates that temporal localization is already a strong component of RecLNet.

This section has shown that the UCF101-24 bias is not in favour of our model while the short classes are by far the hardest to localize and then are source of potentially large improvement for current action localization methods. Also, we observed that our model already benefits from accurate temporal localization while its considerable room for improvement is the spatial localization component (the temporal localization being the one with the less potential for RecLNet). Consequently, the state of the art [26], which mostly relies on spatially more accurate human tracks and better features than the ones we build on [49] (as previously compared in Table 8), and our model, which greatly improves temporal localization, are definitely complementary.

5.4 Improved tracks and descriptors

By using the same tracks and features as input, the previous sections have shown our RecLNet is able to correct the temporal localization weakness of state-of-the-art methods. Here, we substitute these tracks and features (introduced in Section 3.1 and 4) by improved models. In the following, we first explain the track modification then the features substitution and finally analyze their impact.

Person action tracks with temporal integration. In order to improve person tracking, we integrate temporal information in the detector. Supported by our analysis in Section 5.3, showing that better spatial localization would greatly improve our results, and by [26] that shows large detection improvement by stacking features coming from several neighboring frames, we design a new approach for tracks. In the same spirit as [26] who stacks frame features in the SSD detector [37], we adapt Faster-RCNN [47] to perform accurate detection by temporally integrating stacks of K images. We introduce the following modifications to the Faster-RCNN pipeline. First, the inference pass, producing the feature map on which the ROI-pooling applies, is performed independently on K consecutive frames. The K output feature maps are stacked along the channel dimension and will serve at computing scores and regressions. Second, the RPN computes at each anchor location a global objectness score for the stack and K regressions. Third, anchors labels are computed based on the mean overlap between the K ground-truth boxes and the K regressed proposals. Fourth, the 3D proposals are further regressed to action proposals (tubelets)

by computing one action score and K regressions per class. Tubelets are linked into action tracks using the code of [26]. This is an online method that iteratively aggregates tubelets sorted by action score to a set of current links based on spatio-temporal overlap. The detector is used on the ResNet-101 architecture [19] and we choose $K = 5$ since it is a good trade-off between tubelet quality and efficiency. Since only sparse keyframes are spatially annotated on DALY, automatic tracking is performed to propagate the ground truth (see Section 4.2). Here, stacks used at train time on DALY are then composed of one annotated keyframe and its 2 frames tracked forward and backward. These action tracks are referred to as *stack* in the following.

I3D features. Recent results have shown large action recognition improvement using the I3D [8] architecture. We replace the per-frame features by descriptors extracted with the I3D RGB and flow networks both trained on the Kinetics dataset [30]. We extract features after the 7-th inception block, before the max-pooling, as it is a good balance between deepness, for strong classification results, and accurate resolution, for precise track pooling. After this block, the temporal receptive field is roughly one hundred frames. As input, we use a spatial resolution of 320×240 pixels resulting in feature maps of size 20×15 with 832 channels. These feature maps are extracted at intervals of 4 frames. The box temporally aligned with the middle of the interval is used as pooling region to obtain a 832-dimensional I3D descriptor representing the track at this time step.

method	feat.	tracks	UCF101-24			DALY		
			0.3	0.4	0.5	0.1	0.2	0.3
RecLNet	f. based	f. based	69.0	57.5	46.5	30.2	25.4	19.7
RecLNet	I3D	f. based	72.3	63.2	50.8	37.9	33.9	26.8
RecLNet	I3D	<i>stack</i>	77.4	68.5	57.5	41.1	37.6	31.0
FC	I3D	f. based	71.8	62.3	49.8	35.5	32.1	25.0
FC	I3D	<i>stack</i>	75.4	67.4	56.3	38.7	34.9	28.6

Table 9 Spatio-temporal action localization (mAP) performance on UCF101-24 (at IoU 0.3, 0.4 and 0.5) and DALY (at IoU 0.1, 0.2 and 0.3) when substituting the frame-based features by I3D and the frame-based tracks by *stack*. Results are reported for RecLNet and the fully connected classifier (FC).

Experimental evaluation. Here, frame-based features and frame-based tracks we refer to are the ones described in Section 4. Table 9 first shows that when substituting the baseline features by the I3D descriptors our RecLNet model obtains a performance boost of around 4% on UCF101-24 (+4.3% at IoU 0.5) and 7-8% on DALY (+7.1% at IoU 0.3). This result is inline with [18]. When we further replace the frame-based tracks by ours obtained with stack-Faster-RCNN (*stack*), RecLNet results get a second improvement of 5-7% on UCF101-24 and 3-4% on DALY. This gain of

IoU	0.05	0.1	0.2	0.3	0.4	0.5	0.6	0.75
[66]	54.3	51.7	46.8	37.8	-	-	-	-
[23]	54.7	51.3	47.1	39.2	-	-	-	-
[72]	65.2	59.5	47.6	38.0	-	-	-	-
[43]	54.5	50.4	42.3	32.7	-	-	-	-
[67]	71.1	-	58.9	-	-	-	-	-
[44]	78.8	77.3	72.9	65.7	-	-	-	-
[49]	79.1	76.6	66.8	55.5	46.4	35.9	26.8	-
[40]	-	-	34.8	-	-	-	-	-
[54]	-	-	73.5	-	-	46.3	-	15.0
[50]	-	71.3	63.1	51.6	-	33.1	-	-
[26]	-	-	76.5	65.2	-	49.2	-	19.7
[18]	-	-	-	-	-	59.9	-	-
ReLNet	83.0	81.7	77.0	69.0	57.5	46.5	36.7	10.3
ReLNet++	86.6	86.1	83.4	77.4	68.5	57.5	46.0	23.9

Table 10 Comparison to the state of the art on UCF101-24 (mAP for IoU values ranging from 0.05 to 0.6).

performance validates what was shown by [26], i.e., integrating temporal information in the detector increases the track spatial precision leading to action localization improvement (as we observed in Section 5.3). For comparison we also report the fully connected classifier (FC) results as in Section 5.1. Given the large temporal receptive field of I3D features mentioned earlier, we notice ReLNet still improves the performance, e.g., it gets +2.0% and +2.4% mAP at IoU 0.3 when using the *stack* tracks on UCF101-24 and DALY respectively. In the following, the ReLNet model with the I3D features and the improved tracks (*stack*) is referred to as ReLNet++.

5.5 Comparison to the state of the art

UCF101-24. Table 10 compares ReLNet and ReLNet++ models to [18, 23, 26, 40, 43, 44, 49, 50, 54, 66, 67, 72] on UCF101-24 spatio-temporal action localization. ReLNet is our recurrent network with GRU memory units and *fusion layer* to combine optical flow and appearance features performing localization with our *threshold* approach and ReLNet++ uses I3D features and our improved *stack* tracks. ReLNet already significantly outperforms all other methods for all IoU thresholds below 0.5. Due to our significantly improved temporal localization, our approach outperforms [49] on which we build by +9.9% at high IoU (0.6). We also outperform [44], despite the fact that they use more sophisticated features combining a number of different human parts as well as multi-scale training and testing. Both these components are complementary to our approach. Our significant boost in performance can be explained by the fact that most current action localization methods rely on spatial features, for example per-frame CNN descriptors for optical flow and appearance, to detect action both spatially and temporally. However, such features are not designed

scores	localization	0.1	0.2	0.3
[67]	<i>sliding window</i>	-	14.5	-
CNN RGB+OF	<i>threshold</i>	22.9	17.8	13.4
ReLNet	<i>threshold</i>	30.2	25.4	19.7
ReLNet++	<i>threshold</i>	41.1	37.6	31.0

Table 11 State of the art on DALY (mAP for IoU 0.1, 0.2 and 0.3). Localization indicates the method used for temporal localization.

Action classes	CNN RGB+OF	ReLNet
<i>App. MakeUp Lips</i>	3.4	13.6
<i>Brushing Teeth</i>	7.9	19.3
<i>Cleaning Floor</i>	14.3	24.7
<i>Cleaning Windows</i>	6.5	10.9
<i>Drinking</i>	20.8	13.1
<i>Folding Textile</i>	7.2	16.6
<i>Ironing</i>	25.4	32.7
<i>Phoning</i>	7.6	19.9
<i>Playing Harmonica</i>	33.4	34.4
<i>Taking Photos/Videos</i>	7.3	12.2
mAP	13.4	19.7

Table 12 Per-class performance on DALY (mAP at IoU 0.3).

for temporal localization. By relying on an accurate temporal model such as our recurrent ReLNet, we can clearly improve temporal detection and methods we build on that do not focus on this aspect. Recent method [26], relying on more accurate spatial localization obtained by spatio-temporal cuboid regression, outperforms our model at high IoU. Indeed, they benefit from more precise human track bounding-boxes than the one we build on. Such an approach is used by ReLNet++, the enhanced version of ReLNet. This is shown to improve the state of the art by a large margin outperforming [26] by +12.2%, +8.3% and +4.2% at IoU 0.3, 0.5 and 0.75 respectively. We also mention very recent work [18] which only reports results at IoU 0.5 and achieves similar performance.

DALY. The recent DALY dataset [67] is very challenging for spatio-temporal localization because it contains only short actions compared to the video length making temporal segmentation crucial and difficult. Table 11 compares our results to the state of the art. We can see that our ReLNet approach significantly outperforms [67] by more than 10%. We also report results with our CNN scores *CNN RGB+OF* and temporal thresholding of the action scores without using ReLNet. Interestingly, the baseline approach outperforms the state of the art, but performs significantly worse than our (*ReLNet*) method. Also, Table 12 shows ReLNet improves over the baseline for all classes except *Drinking*. Again, since we build on tracks from [67] and extract similar features (see Section 3), the improvement should be attributed to the more accurate temporal localization achieved

by our RecLNet model. RecLNet++ further improves the results and outperforms [67] by +23.1%.

5.6 Qualitative results

Good temporal localization needs precise evolution of action score with clear temporal boundaries. Having smooth scores also help our *threshold* localization method presented in Section 3.3. Here, we provide qualitative results to illustrate that our RecLNet possesses these properties and is well-suited for such localization.

Figure 3 shows scores from [49] (red curves) versus our RecLNet response (green curves) for the 5 action track proposals used on UCF101-24. Thus, there are in total 10 curves per graph, one curve per track for the two compared methods. A bold curve (resp. dashed curve) shows that its corresponding track overlaps with more (resp. less) than 0.5 spatial IoU with a ground-truth action annotation. The ground-truth action interval is represented by the horizontal blue bar below the x-axis. The x-axis represents time. For each graph, we show the frame corresponding to the maximum RecLNet response (localized by the vertical yellow line) with its associated human detection (yellow box).

We show scores for 3 short actions *basketball*, *tennis swing* and *volleyball spiking*, and 2 long actions, *cliff diving* and *diving*. We can observe that the scores of the RecLNet correspond to the temporal boundaries more precisely than the CNN scores of [49]. Scores from [49] are often flat and have high values throughout the entire video length (Figure 3, row 3). In general, they are not precise temporally (e.g., Figure 3, row 4). Also, a lot of dashed lines are highly scored (Figure 3, rows 3 and 5) which adds false positives. Finally, these graphs show that it is easy to set a threshold on RecLNet responses in order to obtain good temporal localization, while scores from [49] (and scores from CNNs in general) are often imprecise temporally.

Similarly, Figures 4, 5 show qualitative results for temporal action localization on the DALY dataset, comparing again our RecLNet (green curves) and the *CNN RGB+OF* baseline (red curves). Each curve corresponds to one person track in the video. Each row corresponds to one action class. The last (third) column in Figures 4, 5 corresponds to failure cases or examples where our RecLNet model does not improve temporal localization. Similarly to results for the UCF101-24 dataset, the two first columns shows that scores of RecLNet are better aligned with the ground truth action boundaries compared to scores of the *CNN RGB+OF* baseline. At the same time, dashed lines indicate many high-scoring false positive detections for the *CNN RGB+OF* method (e.g. Figure 4, row 1, plot 2 and Figure 5, row 2, plot 2). The last column in Figures 4, 5 indicates that DALY is a challenging dataset with many difficult examples where, most

of the time, both methods fail to detect the correct temporal action boundaries.

6 Conclusion

This paper shows that training a recurrent model (GRU) on the level of person tracks for modeling the temporal structure of actions improves localization in time significantly. Building on current state-of-the-art methods that obtain very good results for spatial localization, RecLNet improves significantly the detection of action time boundaries and hence the overall performance of spatio-temporal localization. As demonstrated in our analysis, improving spatial precision of human tracks, inspired by most recent methods, and using better features further improve our results (RecLNet++). Our method outperforms the state of the art on the two challenging datasets, namely UCF101-24 and DALY.

Acknowledgements

This work was supported in part by ERC grants ACTIVIA and ALLEGRO, the MSR-Inria joint lab, the Louis Vuitton ENS Chair on Artificial Intelligence, an Amazon academic research award, and an Intel gift.

References

1. M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *CVPR*, 2014. 6
2. M. Baccouche, F. Mamalet, C. Wolf, C. Garcia, and A. Baskurt. Sequential deep learning for human action recognition. In *International Workshop on Human Behavior Understanding*, 2011. 2
3. T. Bagautdinov, A. Alahi, F. Fleuret, P. Fua, and S. Savarese. Social scene understanding: End-to-end multi-person action localization and collective activity recognition. In *CVPR*, 2017. 2
4. D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. In *CoRR*, 2014. 3
5. L. Bertinetto, J. Valmadre, J. F. Henriques, A. Vedaldi, and P. H. Torr. Fully-convolutional siamese networks for object tracking. In *BMVC*, 2016. 6
6. T. Brox, A. Bruhn, N. Papenbergh, and J. Weickert. High accuracy optical flow estimation based on a theory for warping. In *ECCV*, 2004. 5
7. J. Carreira and A. Zisserman. Quo vadis, action recognition? A new model and the kinetics dataset. In *CVPR*, 2017. 2

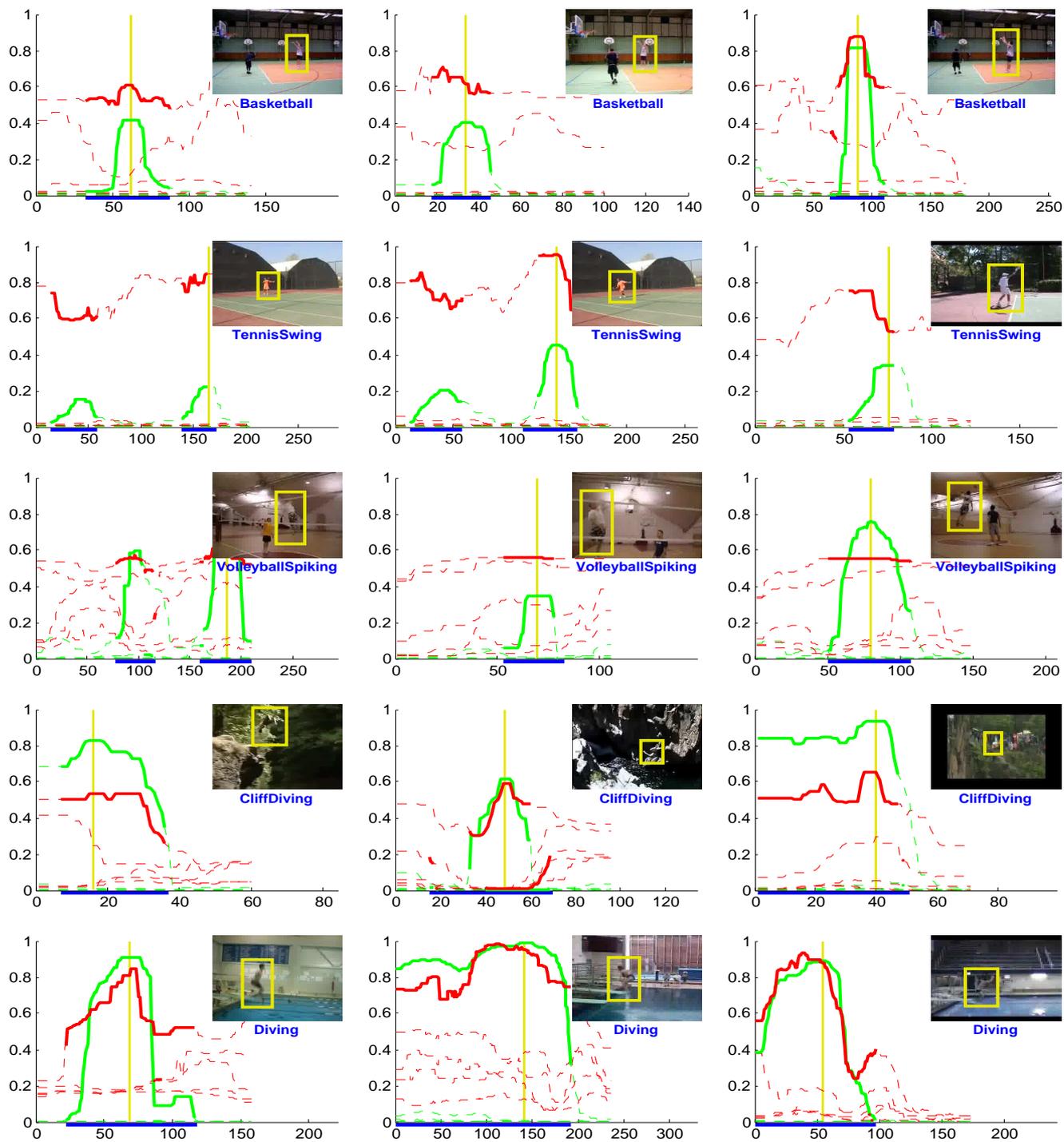


Fig. 3 Qualitative results for temporal localization on UCF101-24. Each curve represents a human track, green curves correspond to our RecLNet scores and red ones to scores from [49]. A bold curve (resp. dashed curve) shows that the track overlaps with more (resp. less) than 0.5 spatial IoU with a ground-truth action. The horizontal blue bar represents the ground-truth time segment. The video frame corresponds to the maximum RecLNet response (localized at the vertical yellow line) with its associated human detection (yellow box). The x-axis represents frame numbers.



Fig. 4 Qualitative results for temporal localization on DALY. Each curve represents a human track, green curves correspond to our RecLNet scores and red ones to scores from CNN features CNN_{RGB+OF} . A bold curve (resp. dashed curve) shows that the track overlaps with more (resp. less) than 0.5 spatial IoU with a ground-truth action. The horizontal blue bar represents the ground-truth time segment. The video frame corresponds to the maximum RecLNet response (localized at the vertical yellow line) with its associated human detection (yellow box). The last column corresponds to failure cases or examples where our RecLNet model does not improve temporal localization. Each row corresponds to one action class, namely *Applying Make Up On Lips*, *Brushing Teeth*, *Cleaning Floor*, *Cleaning Windows* and *Drinking*. The x-axis represents frame numbers.



Fig. 5 Qualitative results for temporal localization on DALY. Each curve represents a human track, green curves correspond to our ReCLNet scores and red ones to scores from CNN features CNN_{RGB+OF} . A bold curve (resp. dashed curve) shows that the track overlaps with more (resp. less) than 0.5 spatial IoU with a ground-truth action. The horizontal blue bar represents the ground-truth time segment. The video frame corresponds to the maximum ReCLNet response (localized at the vertical yellow line) with its associated human detection (yellow box). The last column (red boxes) corresponds to failure cases or examples where our ReCLNet model does not improve temporal localization. Each row corresponds to one action class, namely *Folding Textile*, *Ironing*, *Phoning*, *Playing Harmonica* and *Taking Photos Or Videos*. The x-axis represents frame numbers.

8. J. Carreira and A. Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, 2017. 10
9. K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *CoRR*, 2014. 2, 4
10. G. E. Dahl, D. Yu, L. Deng, and A. Acero. Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. In *IEEE Transactions on Audio, Speech, and Language Processing*, 2012. 1, 3
11. C. R. de Souza, A. Gaidon, E. Vig, and A. M. López. Sympathy for the details: dense trajectories and hybrid classification architectures for action recognition. In *ECCV*, 2016. 1, 2
12. J. Donahue, L. A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *CVPR*, 2015. 1, 2
13. C. Feichtenhofer, A. Pinz, and R. P. Wildes. Spatiotemporal residual networks for video action recognition. In *NIPS*, 2016. 2
14. C. Feichtenhofer, A. Pinz, and A. Zisserman. Convolutional two-stream network fusion for video action recognition. In *CVPR*, 2016. 2
15. R. Girshick. Fast r-cnn. In *ICCV*, 2015. 2, 3, 6
16. G. Gkioxari and J. Malik. Finding action tubes. In *CVPR*, 2015. 2, 3, 5
17. A. Gorban, H. Idrees, Y.-G. Jiang, A. Roshan Zamir, I. Laptev, M. Shah, and R. Sukthankar. THUMOS challenge: Action recognition with a large number of classes. <http://www.thumos.info/>, 2015. 2, 5
18. C. Gu, C. Sun, S. Vijayanarasimhan, C. Pantofaru, D. A. Ross, G. Toderici, Y. Li, S. Ricco, R. Sukthankar, C. Schmid, et al. Ava: A video dataset of spatio-temporally localized atomic visual actions. In *CoRR*, 2017. 10, 11
19. K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 10
20. F. C. Heilbron, V. Escorcia, B. Ghanem, and J. C. Niebles. ActivityNet: A large-scale video benchmark for human activity understanding. In *CVPR*, 2015. 2
21. M. Hoai, Z.-Z. Lan, and F. De la Torre. Joint segmentation and classification of human actions in video. In *CVPR*, 2011. 2
22. S. Hochreiter and J. Schmidhuber. Long short-term memory. In *Neural computation*, 1997. 3
23. R. Hou, C. Chen, and M. Shah. Tube convolutional neural network (t-cnn) for action detection in videos. *ICCV*, 2017. 3, 11
24. H. Idrees, A. R. Zamir, Y.-G. Jiang, A. Gorban, I. Laptev, R. Sukthankar, and M. Shah. The THUMOS challenge on action recognition for videos “in the wild”. In *Computer Vision and Image Understanding*, 2016. 2
25. S. Ji, W. Xu, M. Yang, and K. Yu. 3D convolutional neural networks for human action recognition. In *ICML*, 2010. 2
26. V. Kalogeiton, P. Weinzaepfel, V. Ferrari, and C. Schmid. Action tubelet detector for spatio-temporal action localization. In *ICCV*, 2017. 2, 3, 8, 9, 10, 11
27. A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, 2015. 1, 3
28. 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. 2
29. W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, M. Suleyman, and A. Zisserman. The kinetics human action video dataset. *CoRR*, 2017. 2
30. W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, et al. The kinetics human action video dataset. In *CoRR*, 2017. 10
31. Y. Ke, R. Sukthankar, and M. Hebert. Efficient visual event detection using volumetric features. In *ICCV*, 2005. 2
32. D. Kingma and J. Ba. Adam: A method for stochastic optimization. *CoRR*, 2014. 5
33. H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. HMDB: a large video database for human motion recognition. In *ICCV*, 2011. 2
34. I. Laptev and P. Pérez. Retrieving actions in movies. In *ICCV*, 2007. 2
35. I. Laptev, M. Marszałek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *CVPR*, 2008. 2
36. J. Liu, A. Shahroudy, D. Xu, and G. Wang. Spatiotemporal LSTM with trust gates for 3D human action recognition. In *ECCV*, 2016. 2
37. 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. 3, 10
38. S. Ma, L. Sigal, and S. Sclaroff. Learning activity progression in LSTMs for activity detection and early detection. In *CVPR*, 2016. 2
39. M. Mathias, R. Benenson, M. Pedersoli, and L. Van Gool. Face detection without bells and whistles. In *ECCV*, 2014. 6
40. P. Mettes, J. C. van Gemert, and C. G. M. Snoek. Spot on: Action localization from pointly-supervised proposals. In *ECCV*, 2016. 11
41. J. Y.-H. Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici. Beyond short snippets: Deep networks for video classification. In *CVPR*, 2015. 1, 2, 8

42. D. Oneata, J. Revaud, J. Verbeek, and C. Schmid. Spatio-temporal object detection proposals. In *ECCV*, 2014. [2](#)
43. X. Peng and C. Schmid. Multi-region two-stream R-CNN for action detection. In *ECCV*, 2016. [2](#), [3](#), [11](#)
44. X. Peng and C. Schmid. Multi-region two-stream R-CNN for action detection. In *HAL*, 2017. [3](#), [11](#)
45. L. Pigou, A. van den Oord, S. Dieleman, M. Van Herreweghe, and J. Dambre. Beyond temporal pooling: Recurrence and temporal convolutions for gesture recognition in video. In *IJCV*, 2017. [2](#)
46. H. Pirsiavash and D. Ramanan. Parsing videos of actions with segmental grammars. In *CVPR*, 2014. [2](#)
47. S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *NIPS*, 2015. [6](#), [10](#)
48. M. Rohrbach, A. Rohrbach, M. Regneri, S. Amin, M. Andriluka, M. Pinkal, and B. Schiele. Recognizing fine-grained and composite activities using hand-centric features and script data. In *IJCV*, 2016. [2](#)
49. S. Saha, G. Singh, M. Sapienza, P. H. Torr, and F. Cuzzolin. Deep learning for detecting multiple space-time action tubes in videos. In *BMVC*, 2016. [2](#), [3](#), [5](#), [6](#), [7](#), [8](#), [9](#), [10](#), [11](#), [12](#), [13](#)
50. S. Saha, G. Singh, and F. Cuzzolin. Amtnet: Action-micro-tube regression by end-to-end trainable deep architecture. *ICCV*, 2017. [3](#), [11](#)
51. Z. Shou, D. Wang, and S.-F. Chang. Temporal action localization in untrimmed videos via multi-stage CNNs. In *CVPR*, 2016. [2](#)
52. K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. In *NIPS*, 2014. [1](#), [2](#)
53. B. Singh, T. K. Marks, M. Jones, O. Tuzel, and M. Shao. A multi-stream bi-directional recurrent neural network for fine-grained action detection. In *CVPR*, 2016. [2](#)
54. G. Singh, S. Saha, M. Sapienza, P. Torr, and F. Cuzzolin. Online real time multiple spatiotemporal action localisation and prediction. In *CoRR*, 2017. [3](#), [11](#)
55. K. Soomro and M. Shah. Unsupervised action discovery and localization in videos. In *ICCV*, 2017. [3](#)
56. K. Soomro, A. R. Zamir, and M. Shah. UCF101: A dataset of 101 human actions classes from videos in the wild. In *CoRR*, 2012. [2](#), [5](#)
57. N. Srivastava, E. Mansimov, and R. Salakhutdinov. Unsupervised learning of video representations using LSTMs. In *CoRR*, 2016. [2](#)
58. I. Sutskever, J. Martens, and G. E. Hinton. Generating text with recurrent neural networks. In *ICML*, 2011. [1](#)
59. G. W. Taylor, R. Fergus, Y. LeCun, and C. Bregler. Convolutional learning of spatio-temporal features. In *ECCV*, 2010. [2](#)
60. P. Tokmakov, K. Alahari, and C. Schmid. Learning video object segmentation with visual memory. In *ICCV*, 2017. [4](#), [7](#)
61. D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning spatiotemporal features with 3D convolutional networks. In *ICCV*, 2015. [2](#)
62. J. C. van Gemert, M. Jain, E. Gati, and C. G. M. Snoek. APT: Action localization proposals from dense trajectories. In *BMVC*, 2015. [3](#)
63. G. Varol, I. Laptev, and C. Schmid. Long-term temporal convolutions for action recognition. *PAMI*, 2017. [2](#)
64. O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. In *CVPR*, 2015. [1](#)
65. H. Wang and C. Schmid. Action recognition with improved trajectories. In *ICCV*, 2013. [2](#)
66. P. Weinzaepfel, Z. Harchaoui, and C. Schmid. Learning to track for spatio-temporal action localization. In *ICCV*, 2015. [2](#), [3](#), [5](#), [11](#)
67. P. Weinzaepfel, X. Martin, and C. Schmid. Human action localization with sparse spatial supervision. In *CoRR*, 2016. [2](#), [3](#), [5](#), [6](#), [7](#), [8](#), [11](#), [12](#)
68. J. Yang and J. Yuan. Common action discovery and localization in unconstrained videos. In *ICCV*, 2017. [3](#)
69. S. Yeung, O. Russakovsky, G. Mori, and L. Fei-Fei. End-to-end learning of action detection from frame glimpses in videos. In *CVPR*, 2016. [2](#)
70. J. Yuan, B. Ni, X. Yang, and A. A.Kassim. Temporal action localization with pyramid of score distribution features. In *CVPR*, 2016. [2](#)
71. Y. Zhao, Y. Xiong, L. Wang, Z. Wu, X. Tang, and D. Lin. Temporal action detection with structured segment networks. *ICCV*, 2017. [2](#)
72. M. Zolfaghari, G. L. Oliveira, N. Sedaghat, and T. Brox. Chained multi-stream networks exploiting pose, motion, and appearance for action classification and detection. *ICCV*, 2017. [3](#), [11](#)