Deep Reinforcement Learning for Weak Human Activity Localization

Wanru Xu, Zhenjiang Miao, *Member, IEEE*, Jian Yu, and Qiang Ji, *Fellow, IEEE*

***Abstract*— Human activity localization aims at recognizing contents and detecting locations of activities in video sequences. With an increasing number of untrimmed video data, traditional activity localization methods always suffer from two major limitations. First, detailed annotations are needed in most existing methods, *i*.e., bounding-box annotations in every frame, which are both expensive and time consuming. Second, the search space is too large for 3D activity localization, which requires generating a large number of proposals. In this paper, we propose a unified deep Q-network with weak reward and weak loss (DWRLQN) to address the two problems. Certain weak knowledge and weak constraints involving the temporal dynamics of human activity are incorporated into a deep reinforcement learning framework under sparse spatial supervision, where we assume that only a portion of frames are annotated in each video sequence. Experiments on UCF-Sports, UCF-101 and sub-JHMDB demonstrate that our proposed model achieves promising performance by only utilizing a very small number of proposals. More importantly, our DWRLQN trained with partial annotations and weak information even outperforms fully supervised methods.**

***Index Terms*— Activity localization, deep reinforcement learning, weak constraint, weak supervision.**

I. Introduction

LOCATING human activity in untrimmed videos is an important and challenging task, whose goal is to recognize what activity the video contains and detect where the activity happens. In the recent years, a number of human activity localization methods have been proposed [36], [45], [52]. Most of these existing methods are fully supervised and have two limitations. 1) A vast number of detailed framewise annotations are necessary in the training process to retain model's generalization capability. However, it is expensive and time consuming to obtain these annotations manually. 2) A large number of proposals are extracted by exhaustively searching through entire videos to improve the model's discriminating ability. It is inefficient to complete such search in videos and to extract such a mass of proposals, and it significantly increases computation costs as well. Therefore, current localization methods cannot be used for large datasets and real-time applications.

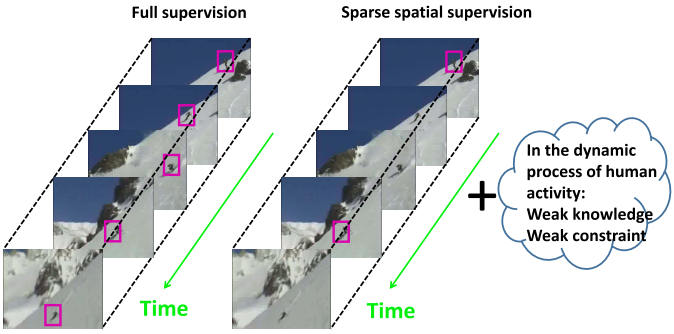


Fig. 1. Comparison of full supervision and sparse spatial supervision for activity localization. (Left) Full supervision: detailed bounding-box annotations in every frame are available for training. (Right) Sparse spatial supervision: only a portion of bounding boxes or even only one bounding box per video sequence is provided. To complement these incomplete annotations, some weak knowledge and weak constraints are mined and integrated.

In this paper, we propose a unified deep Q-network with weak reward and weak loss (DWRLQN) to locate human activities effectively and efficiently, which appropriately addresses the above limitations. First, to reduce the requirement for annotations, we integrate some intrinsic dynamic information of activity videos and train the DWRLQN under weak supervision rather than full supervision. Second, to reduce both the search space and the number of proposals, we learn an optimal search strategy by deep reinforcement learning instead of exhaustive search. Different from the weakly supervised object localization, with only the image-level labels and without the bounding-box annotations, the weakly supervised activity localization task means that besides the activity category, only a portion of bounding boxes or even only one bounding box per video sequence is available for training as shown in Fig.1 (Right). Superior to fully supervised methods, which require detailed bounding-box annotations in every frame for training as shown in Fig.1 (Left), we utilize a kind of weakly supervised method for activity localization, called sparse spatial supervision. In addition to these incomplete annotations, we make full use of some prior knowledge and underlying constraints in the human activity sequence to improve training performance.

Manuscript received May 26, 2018; revised January 31, 2019, May 1, 2019 and September 11, 2019; accepted September 14, 2019. Date of publication September 26, 2019; date of current version November 7, 2019. This work was supported in part by the NSFC under Grant 61672089, Grant 61273274, and Grant 61572064, in part by the China Postdoctoral Science Foundation under Grant 2019M650469, and in part by the National Key Technology R&D Program of China under Grant 2012BAH01F03. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Riccardo Leonardi. *(Corresponding author: Wanru Xu.)*

W. Xu and Z. Miao are with the Institute of Information Science, Beijing Jiaotong University, Beijing 100044, China, and also with the Beijing Key Laboratory of Advanced Information Science and Network Technology, Beijing 100044, China (e-mail: xuwanru@bjtu.edu.cn; zjmiao@bjtu.edu.cn).

J. Yu is with the School of Computer and Information Technology, Beijing Jiaotong University, Beijing 100044, China (e-mail: jianyu@bjtu.edu.cn).

Q. Ji is with the Department of Electrical and Computer Engineering, Rensselaer Polytechnic Institute, Troy, NY 12180 USA (e-mail: qji@ecse.rpi.edu).

Digital Object Identifier 10.1109/TIP.2019.2942814

1057-7149 © 2019 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

Authorized licensed use limited to: The Libraries at Rensselaer Polytechnic Institute. Downloaded on January 13,2023 at 19:29:15 UTC from IEEE Xplore. Restrictions apply.

Inspired by the success of DeepMind [25], [27] in playing Atari games and Chinese Go [33], [43], we find that deep reinforcement learning is a good framework for activity localization under sparse spatial supervision. First, reinforcement learning is one of the typical weakly supervised learning methods, where some prior knowledge can be used to define its reward function instead of employing accurate annotations. For example, the reward for Chinese Go play [43] is the final win, not the judgement of each move; thus, it is not necessary to evaluate each move by experts. Second, a localization agent optimized by a deep Q-network can successively detect activities by exploring a small number of potential space regions. Following a human perception procedure, the agent focuses attention on those regions with rich information and gradually searches the target activity. Therefore, we propose a unified DWRLQN to train the agent and incorporate some prior knowledge to complement the limited annotations. In this paper, the activity localization problem is formulated as a Markov decision process and solved by the reinforcement learning methodology. In the proposed DWRLQN, the state is extracted from the current visible region; the reward is defined as the improvement of localization accuracy for annotated frames and calculated by weak knowledge for unannotated frames; the loss function is composed of the Bellman equation and some weak constraints. Specifically, we train a localization agent that starts from the whole region of each frame and is capable of deciding where to focus next to search for a better location via predefined actions. To the best of our knowledge, this is the first deep reinforcement learning model for human activity localization with minimal supervision effort.

Overall, the contributions of this paper can be concluded as follows: 1) Complete human activity localization task by a unified deep Q-network, which proposes an effective and efficient way to extract proposals following an optimized search strategy instead of an exhaustive search. 2) Sparse spatial supervision is employed to train DWRLQN, which integrates some weak knowledge and weak constraints involved in the temporal dynamics of the human activity sequence to enhance the model's description capability and discriminating ability. 3) Experiments on three public datasets (*i.e.* UCF-Sports, UCF-101 and sub-JHMDB) show that our proposed DWRLQN trained with partial annotations and weak information even outperforms fully supervised approaches.

II. Related Work

In this section, we review related works on human activity localization under full supervision and weak supervision. We also introduce some deep reinforcement learning methods used in the computer vision area.

A. Fully Supervised Activity Localization

In fully supervised learning, human activity localization is commonly approached by spatio-temporal proposal matching [13], [32], namely, the classical proposal + classification framework [36], [45], [52]. The main idea is first to generate a large number of spatio-temporal candidates and then score them by a pre-trained classifier to determine the final detection result. Traditionally, proposals are generated by a multi-scale sliding window [8], [20], which is effective but not efficient. In [44], a temporal sliding window is used to generate several candidates and then train an SVM to rank them by combining motion and appearance features. Meanwhile, several efforts aim at reducing the computational cost, such as the spatio-temporal branch-and-bound algorithm [54]. The activity proposal extraction model would be another alternative to reduce the search space significantly. The 2D deformable part model is extended to a 3D spatio-temporal DPM for activity localization in [41], where the most discriminative 3D sub-volumes are selected as individual parts and their spatio-temporal relations are also learned. The naive Bayes-based mutual information maximization (NBMIM) approach [53] is introduced to find the optimal sub-volume in video space for efficiently detecting activity. In addition to these sub-volumes, which cannot handle moving activities, spatio-temporal activity tubes [42] are extracted as detection candidates by a structured output regression with a max-path search. Analogously, a series of spatio-temporal video tubes [52] are considered as activity candidates that are generated via a greedy method by computing an actionness score. Activity proposals are defined as 2D+t sequences of bounding boxes called tubelets in [15], which are generated by hierarchically merging super-voxels. A multi-region two-stream R-CNN model [28] is introduced to locate activity in realistic videos, which starts from frame-level detection based on an R-CNN [10] and then links frame-level detections to obtain the final video-level detection. Although these proposal extraction models achieve promising performance, they are still time consuming and conflict with the human perception procedure.

B. Weakly Supervised Activity Localization

Although annotating each frame in a video sequence with bounding boxes is unrealistic for a large-scale dataset, the concept of reducing spatial supervision for activity localization has so far received little attention. A novel method for learning a discriminative sub-window classifier from weakly annotated data is proposed in [14]. In [35], the weakly supervised learning task is formulated as a multiple instance learning problem, which optimizes both inter-class and intra-class distance globally. In [48], they first extract human tubes by a state-of-the-art human detector trained with a large number of annotated human images, and then select positive and negative tubes under very sparse spatial supervision. A matrix completion method is proposed in [26] for human activity recognition and localization using weakly supervised multi-label learning, where non-rectangular spatio-temporal discriminative regions are extracted as activity proposals by clustering with texture and motion features. An effective method [39] is introduced for temporal activity localization using cross-domain transfer between web frames and video images, whose inputs are noisy image labels and weak video labels. In summary, although above methods [50] are proposed for weak activity localization, they also require a large quantity of annotated data from other sources.

Authorized licensed use limited to: The Libraries at Rensselaer Polytechnic Institute. Downloaded on January 13,2023 at 19:29:15 UTC from IEEE Xplore. Restrictions apply.

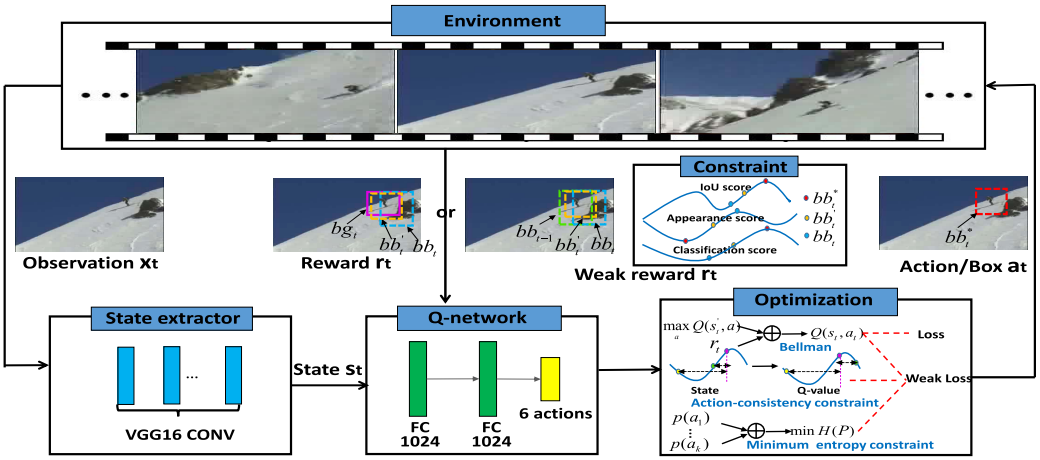


Fig. 2. Illustration of our proposed deep Q-network with weak reward and weak loss (DWRLQN) for human activity localization under sparse spatial supervision. Two cases exist: 1) frame with annotations: the reward function is measured by the ground truth, and the loss function is calculated by the Bellman equation; 2) frame without annotations: the weak reward and weak loss are adopted to train the DWRLQN model, which are only defined by some weak knowledge and weak constraints.

C. Deep Reinforcement Learning in Computer Vision

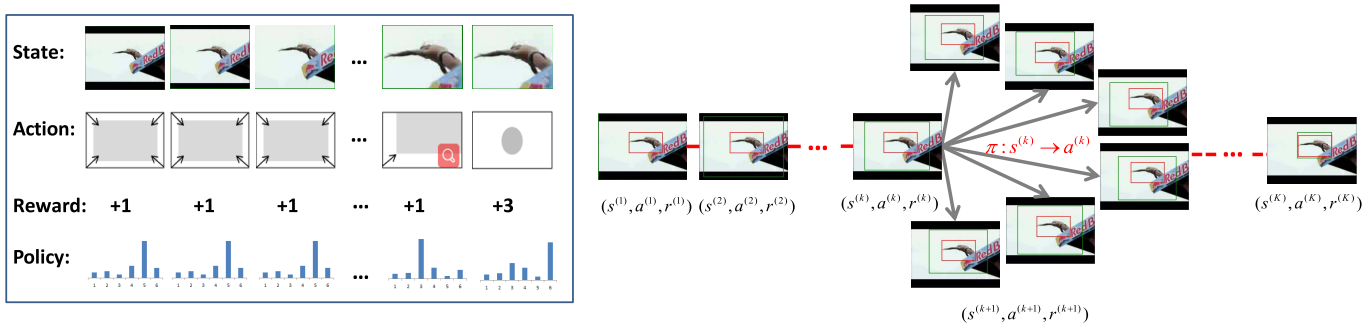
To address the above problems, we draw inspiration from recent deep reinforcement learning, which attempts to implement attentional processing in a deep learning framework. Currently, increasingly more problems of computer vision have been formulated in the deep reinforcement learning framework, such as object detection [1], [3], [17], recognition [4] and tracking [6], [55], [56]. A fully end-to-end approach for temporal activity localization is presented in [51] and a searching agent is trained by reinforcement learning, which directly learns to predict the temporal bounds of activities. An active model for object localization is proposed in [3], where an agent is allowed to focus attention on candidate regions to find the location of target object accurately and rapidly. An effective tree-structured reinforcement learning (Tree-RL) method [17] is introduced to sequentially locate objects by fully exploiting both historical search paths and current observations. Similarly, a hierarchical object detection method guided by a deep reinforcement learning agent is proposed in [1], whose procedure is iterated, providing a hierarchical image analysis. A collaborative deep reinforcement learning method [19] is proposed for multiple objects detection jointly by treating each detector as an agent. Meanwhile, a novel multi-agent deep Q-learning algorithm is utilized to learn inter-agent communication, which effectively exploits such beneficial contextual information. A novel neural network tracking model is proposed in [56], which comprises three sub-models: a CNN extracting features from each frame, an RNN constructing temporal states, and a reinforcement learning agent making decisions for locating target object. A template selection strategy constructed by deep reinforcement learning for visual tracking is introduced in [6], which is adopted to automatically select the best template for tracking a certain frame in video sequences. An action-decision network (ADNet) [55] is applied to control a novel tracker by sequentially pursuing actions, which is trained by supervised learning as well as deep reinforcement learning. Instead of exhaustively training individual detectors for all possible relationships, a deep variation-structured reinforcement learning (VRL) framework [21] is proposed to sequentially discover object relationships and attributes in 2D images. An attention-aware deep reinforcement learning (ADRL) method [4] is introduced for video-based face recognition, which aims to discard some misleading frames and select the most informative frames to represent face videos. Such a deep reinforcement learning method follows the human perception procedure and actually is a coarse-to-fine search process, where observation and refinement are performed iteratively.

Compared to these fully supervised models, we take full advantage of the weakly supervised property in reinforcement learning and propose a unified DWRLQN for human activity localization under sparse spatial supervision. In this paper, our aim is to refine hypotheses about where an activity occurs by incorporating incomplete annotations with some weak knowledge and weak constraints.

III. Deep Q-Network With Weak Reward and Weak Loss for Activity Localization

The goal of weak activity localization is to quickly detect the activity location with minimal supervision effort, which is defined as a series of bounding boxes $\{x_1^t, y_1^t, x_2^t, y_2^t\}_1^T$, and each of them is represented by the coordinates of its two corners. As depicted in Fig.2, a video sequence is considered as the environment, where an optimized localization agent finds a sequence of actions to progressively transform the detected bounding box using a set of pre-defined actions in the spatial domain until reaching the target activity. At each search step, according to the predicted action, the state is updated, and then the agent decides how to choose the next action in terms of the new state. The search procedure iteratively continues until the target activity is localized. In this paper, we develop a DWRLQN to train the localization agent, whose input is the current observation and output is the probability of each action. In particular, the DWRLQN leverages some

Authorized licensed use limited to: The Libraries at Rensselaer Polytechnic Institute. Downloaded on January 13,2023 at 19:29:15 UTC from IEEE Xplore. Restrictions apply.



(a) The concept of the tuple (S, A, R) and localization policy π. (b) The way to obtain the optimal search trajectory by MDP-based model.

Fig. 3. The illustration of the Markov decision process in activity localization task, including the tuple (S, A, R), policy π and the optimal search trajectory.

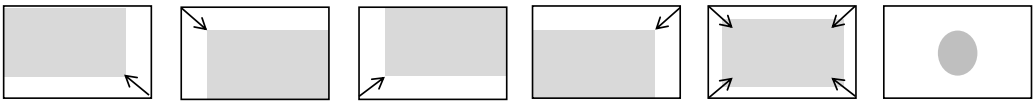


Fig. 4. Illustration of the actions adopted in DWRLQN, which are utilized to transform bounding boxes. The first five are scaling actions and the last one is a terminate action to stop the search process.

weak information for human activity localization under sparse spatial supervision, where the ground truth is used to train the model for annotated frames and the weak information are used to define the reward and loss function for unannotated frames.

A. Activity Localization as a Markov Decision Process

In this paper, we cast the problem of weak activity localization as a Markov decision process (MDP). Typically, an MDP [40] is defined as a tuple (S, A, R) that is decided by a policy π, and the interpretation of each item in activity localization task is shown in Fig.3(a).

- S represents a finite set of states, and $s^{(k)} \in S$ is the state representation at search step k, which encodes the RGB information of the currently visible region and is composed of the features extracted inside the bounding box. We introduce a deep model to capture the appearance and extract the visual state for each bounding box bb, which is denoted as $f_s(I(bb))$, where $I(bb)$ denotes the image patch inside the bounding box. As shown in the state extractor part of Fig.2, the current region is first warped to 224*224 pixels and then processed by a pre-trained CNN following the VGG-16 architecture [34].
- A represents a finite set of actions allowing the agent to interact with the environment, namely, the way to change the bounding box, where $a^{(k)} \in A$ is the action the agent performs at search step k. We define 6 spatial actions illustrated in Fig.4, aiming to change the detected bounding boxes. The action set is composed of 5 scaling actions (top left scaling, top right scaling, bottom left scaling, bottom right scaling, centre scaling), and one terminate action to stop the search process. For example, the centre scaling with a scaling factor α is described in Eq.(1), which changes $\{x_1, y_1, x_2, y_2\}$ to $\{x'_1, y'_1, x'_2, y'_2\}$:

$$\begin{aligned} x'_1 &= x_1 + 0.5 \times \alpha \times (x_2 - x_1), \\ y'_1 &= y_1 + 0.5 \times \alpha \times (y_2 - y_1), \\ x'_2 &= x_2 - 0.5 \times \alpha \times (x_2 - x_1), \\ y'_2 &= y_2 - 0.5 \times \alpha \times (y_2 - y_1). \end{aligned} \quad (1)$$

  Actually, the scaling factor α and the action type influence the precision and the speed of localization, which is evaluated in the experimental section.
- $R(s, a)$ is a function of state-action pairs, where $r^{(k)} = R(s^{(k)}, a^{(k)})$ indicates the feedback from the environment after taking action $a^{(k)}$ at current state $s^{(k)}$, reflecting how well the next detected bounding box covers the target activity.
- π is a distribution over actions given states, where $\pi(a^{(k)}|s^{(k)}) \sim p(a^{(k)}|s^{(k)})$ denotes the probability of choosing action $a^{(k)}$ at state $s^{(k)}$. The policy fully defines the behavior of the localization agent.

As shown in Fig.3(b), starting from the initial state (*e*.g., the whole region of each frame), a search trajectory $s^{(1)}$, $a^{(1)}$, $r^{(1)}, \ldots, s^{(K)}$, $a^{(K)}$, $r^{(K)}$ is obtained according to the policy π, namely, the Markov decision process $\langle S, A, R; \pi \rangle$. Specifically, given the current state $s^{(k)}$, the agent chooses the action $a^{(k)}$ with the maximal probability of $\pi(a^{(k)}|s^{(k)})$; then, the agent correspondingly updates the next state $s^{(k+1)}$ and thus receives a scalar reward $r^{(k)}$. The goal is to learn an optimal policy to maximize the total reward of the search trajectory $r^{(1)}+r^{(2)}+\ldots+r^{(K)}$ during the whole search procedure with K steps. This problem can be simplified as choosing an optimal action at each search step to maximize the associated cumulative future reward, which reflects a long-term effect of the current choice. To achieve this aim, the action-value function (Q-function) is defined as the expected total future discounted reward under policy π in Eq.(2), which measures how performing action a at state s is beneficial in the long term.

$$Q^{\pi}(s, a) = E[\sum_{k=d}^{K} \gamma^{k-d} R(s^{(k)}, a^{(k)}) | s = s^{(d)}, a = a^{(d)}, \pi] \quad (2)$$

where γ is a discount factor that gradually decreases the influence of the latter state on the current state. In practice, the Q-function is approximated by the proposed DWRLQN with the capacity of mapping the state set to the action set.

Authorized licensed use limited to: The Libraries at Rensselaer Polytechnic Institute. Downloaded on January 13,2023 at 19:29:15 UTC from IEEE Xplore. Restrictions apply.

By maximizing the Q-function, it provides us with the optimal policy,

$$a^* = \arg\max_{a \in A} \pi(a|s) = \arg\max_{a \in A} Q^{\pi}(s, a) \tag{3}$$

In the following sections, we will detail the reward functions and training strategies for frames with and without annotations in DWRLQN, respectively.

B. Reward Function With Weak Knowledge

In the deep reinforcement learning, the reward function directly guides the optimization procedure of the agent, since the model is adjusted according to the feedback of each search step and the model's objective function is defined depending on the reward function. For the frames with detailed annotation of the bounding box, we utilize the ground truth to define DWRLQN's reward. Meanwhile, for those frames without annotation, the reward is calculated to evaluate each move depending on weak knowledge, which is treated as a weak reward.

1) Reward for the Frame With Annotations: The reward function $R(s_t, a_t)$ is defined as the improvement of the localization accuracy after a state transition. The intersection-over-union (IoU) between the detected bounding box bb_t and the ground truth bg_t of the current frame is employed to measure the localization accuracy. The reward function can be divided into two groups: terminate reward R_T in Eq.(4) and non-terminate reward R_N in Eq.(5):

$$R_T(s_t, a_t) = \lambda_2 sign(IoU(bb_t, bg_t) - \tau) \tag{4}$$

$$R_N(s_t, a_t) = \lambda_1 sign(IoU(bb'_t, bg_t) - IoU(bb_t, bg_t)) \tag{5}$$

where $R_T : S \times A \to \mathbb{R}$ and $R_N : S \times A \to \mathbb{R}$ are both scalar reward functions. Intuitively, the non-terminate reward represents that there will be a positive feedback $\lambda_1 = 1$ if the IoU improves when the agent performs action a_t at state s_t, which correspondingly changes the bounding box from bb_t to bb'_t to move closer to the target activity; otherwise, the feedback is negative. Since the bounding box is not changed anymore at terminate state, the terminate reward has a different definition, where the agent will obtain a positive feedback $\lambda_2 = 3$ when the IoU is above the given threshold τ at the terminate state s_t, which means that the agent successfully locates the target activity.

2) Weak Reward for the Frame Without Annotations: For those frames without annotations, we do not have the ground truth to evaluate the improvement at each search step. However, some prior knowledge and underlying constraints exist between neighboring frames, which can be considered as weak annotations. In order to complement incomplete annotations, we take advantage of several weak constraints to calculate the weak reward as follows:

Smoothness constraint: It requires the activity proposals to be a smooth path, since human activity is a relatively gradual process. Therefore, the location in the previous frame can be treated as a weak annotation for the current frame. The IoU between the optimal detection of current frame bb_t^* and the detection of previous frame bb_{t-1} should be higher than other detections bb_t as shown in Eq.(6).

$$IoU(bb_t^*, bb_{t-1}) > IoU(bb_t, bb_{t-1}) \tag{6}$$

Consistent-appearance constraint: It requires the activity proposal to correspond to a path with a consistent appearance; thus, it is more likely to track the same actor, where the appearance is represented by the state in our framework. As shown in Eq.(7), the appearance of the optimal detection of current frame $f_s(I(bb_t^*))$ should be more similar to that of previous detections than others. Since activity is a dynamic process and its appearance is constantly changing, we use the average appearance of previous detections $\bar{f}_s(\mathbf{I}(bb_{1:t-1}))$ by averaging the states of detections from 1 to $t-1$ frames.

$$\|f_s(I(bb_t^*)) - \bar{f}_s(\mathbf{I}(bb_{1:t-1}))\|_2^2 < \|f_s(I(bb_t)) - \bar{f}_s(\mathbf{I}(bb_{1:t-1}))\|_2^2 \tag{7}$$

Increasing classification-confidence constraint: It suggests that the classification confidence could be improved by precise localization and detection-aware features. Intuitively, we are more confident about the activity category if a more accurate activity location is detected. When employing the observation of the optimal current detection bb_t^* instead of other inaccurate detections bb_t, the classification confidence score of the video from 1 to t frames should increase as shown in Eq.(8). The classification confidence score can be obtained by any activity recognition model $f_c()$, whose input is the observation of the detections $\mathbf{I}(bb_1, bb_2, \ldots, bb_t)$ and output is the probability of activity, and a pre-trained Two-Stream Inflated 3D ConvNet [5] is used in this paper.

$$f_c(\mathbf{I}(bb_1, bb_2, \ldots, bb_t^*)) > f_c(\mathbf{I}(bb_1, bb_2, \ldots, bb_t)) \tag{8}$$

Thus, the weak reward function is defined by Eq.(9), which is a sum of smoothness reward $R_{smo}(s_t, a_t)$ in Eq.(10), consistent-appearance reward $R_{app}(s_t, a_t)$ in Eq.(12) and increasing classification-confidence reward $R_{class}(s_t, a_t)$ in Eq.(11). This combination can make the weak constraint stronger and provide more supervised information to guide the training of the model, since each weak constraint is too weak to independently and accurately locate human activity.

$$R(s_t, a_t) = R_{smo}(s_t, a_t) + R_{app}(s_t, a_t) + R_{class}(s_t, a_t) \tag{9}$$

where these individual rewards based on the corresponding weak constraints are calculated by:

$$R_{smo} = \begin{cases} \lambda_1 sign(IoU(bb'_t, bb_{t-1}) - IoU(bb_t, bb_{t-1})) \\ \lambda_2 sign(IoU(bb_t, bb_{t-1}) - \tau_{smo}) \end{cases} \tag{10}$$

$$R_{class} = \begin{cases} \lambda_1 sign(f_c(\mathbf{I}(bb_1, \ldots, bb'_t)) - f_c(\mathbf{I}(bb_1, \ldots, bb_t))) \\ \lambda_2 sign(f_c(\mathbf{I}(bb_1, \ldots, bb_t)) - \tau_{class}) \end{cases} \tag{11}$$

$$R_{app} = \begin{cases} \lambda_1 sign(\|f_s(I(bb_t)) - \bar{f}_s(\mathbf{I}(bb_{1:t-1}))\|_2^2 \\ -\|f_s(I(bb'_t)) - \bar{f}_s(\mathbf{I}(bb_{1:t-1}))\|_2^2) \\ \lambda_2 sign(\tau_{app} - \|f_s(I(bb_t)) - \bar{f}_s(\mathbf{I}(bb_{1:t-1}))\|_2^2) \end{cases} \tag{12}$$

Authorized licensed use limited to: The Libraries at Rensselaer Polytechnic Institute. Downloaded on January 13,2023 at 19:29:15 UTC from IEEE Xplore. Restrictions apply.

Similar to the reward for annotated frames, each individual reward contains the non-terminate reward and the terminate reward. In short, if the detected bounding box satisfies these weak constraints, the localization agent will receive a positive feedback, and vice versa. In turn, the localization agent encourages achieving the positive increasing classification-confidence reward, consistent-appearance reward and smoothness reward between the consecutive moves from bb_t to bb'_t in order to fully satisfy these weak constraints. In addition, such a definition involves an underlying ordinal constraint, which guarantees the search moves to the target activity step-by-step.

C. Training the Agent With Weak Constraints

We utilize reinforcement learning to learn the optimal decision policy that maximizes the Q-function at each search step and finds the next better location until accurately locating the target activity. However, it is difficult to solve by the traditional reinforcement learning method, due to the high-dimensional continuous state space and the model-free environment. Therefore, a deep Q-network is applied as the Q-function approximator to estimate the optimal value for each state-action pair, *i.e.,* $Q(s, a; \theta)$, which is parameterized by θ. This Q-network consists of two fully connected layers of 1024 neurons each and an output layer with 6 action decisions as shown in Fig. 2.

The training procedure is demonstrated in Algorithm 1. We need to run the agent N episodes for each video clip and for each frame to learn the model's parameters. At each search step, we choose the next better location until reaching the terminate action or maximum search step. Specifically, the agent starts from the whole image $\{1, 1, W, H\}$ for each frame and then takes a sequence of pre-defined actions to update the detected bounding box and optimize the DWRLQN. In the Q-learning algorithm, we adopt ε-greedy for the behaviour of the agent during training. That is, the agent will select a random action with probability ε and take the action depending on the already learned policy with probability $1 - \varepsilon$. It is actually an exploration and exploitation process, where the exploration helps the agent avoid becoming stuck in a local optimum and the exploitation makes the best decision given the current model. The replay memory is also incorporated to store experiences and then randomly sample a mini-batch $\{(s_i, a_i, r_i, s'_i)\}$ to update the DWRLQN. Each (s_i, a_i, r_i, s'_i) is a transition sample from state s_i to s'_i after taking action a_i and obtaining reward r_i.

During DWRLQN training, two cases exist: 1) For a frame with annotations, the agent will terminate search when the IoU between the detected bounding box and the ground truth is above the threshold τ. The reward is calculated depending on the ground truth and the loss is defined by the Bellman equation [24], [25], as shown in Eq.(13):

$$L = \sum_i (R(s_i, a_i) + \gamma \max_{a'} Q(s'_i, a'_i; \theta)) - Q(s_i, a_i; \theta). \quad (13)$$

2) For a frame without annotations, the agent will reach the terminate state when all the IoU scores, appearance scores and classification scores are above their corresponding thresholds $\tau_{smo}, \tau_{app}, \tau_{class}$, respectively. Instead of the ground truth, the weak reward defined in Eq.(9) for unannotated frames depends on the smoothness constraint in Eq.(6), consistent-appearance constraint in Eq.(7) and increasing classification-confidence constraint in Eq.(8). The weak loss defined in Eq.(16) for unannotated frames is calculated based on the action-consistency constraint and minimum entropy constraint:

Action-consistency constraint: The localization agent tends to select similar actions for similar states. If states of bb_i and bb_j are more similar than those of bb_i and bb_k, it is more likely to select the same action for bb_i and bb_j, as shown in Eq.(14).

$$if\, ||f_s(I(bb_i)) - f_s(I(bb_j))||_2^2 < ||f_s(I(bb_i)) - f_s(I(bb_k))||_2^2, \quad p(a_i = a_j|s_i) > p(a_i = a_k|s_i) \quad (14)$$

Minimum entropy constraint: It is a common constraint for unsupervised learning, which requires the model's output probability to be less uncertain by minimizing the entropy $H(a)$, as shown in Eq.(15).

$$minH(a) = min(-\sum_k p(a_i = k|s_i) \log p(a_i = k|s_i)) \quad (15)$$

$$L = \sum_i (R(s_i, a_i) + \gamma \max_{a'} Q(s'_i, a'_i; \theta)) - Q(s_i, a_i; \theta) + \sum_{i,j \neq i} \frac{||Q(s_i, a_i; \theta) - Q(s_j, a_j; \theta)||_2^2}{||s_i - s_j||_2^2} - \sum_{i,k} p_i(a_k) \log p_i(a_k), \quad (16)$$

Algorithm 1 Training Procedure for the DWRLQN

1: Initialize the replay memory D with $\{\}$;
2: Initialize the parameter of DWRLQN by θ;
3: **for** each episode $n \in [1, N]$, each video sample $i \in [1, M]$ and each frame $t \in [1, T]$ **do**
4: Initialize the bounding box $bb_t = \{1, 1, W, H\}$ and extract the state $s_t = f_s(I(bb_t))$;
5: **While** until taking the terminate action or reaching the maximum search step **do**
6: Select action a_t with ε-greedy and correspondingly change bounding box to bb'_t;
7: **if** the ground-truth bounding box gb_t is annotated
8: Calculate the reward r_t as Eq.(5) or Eq.(4);
9: Calculate the loss as Eq.(13);
10: **else**
11: Calculate the weak reward r_t as Eq.(9);
12: Calculate the weak loss as Eq.(16);
13: **end if**
14: Store transition (s_t, a_t, r_t, s'_t) into experiences D;
15: Sample random mini-batch (s_i, a_i, r_i, s'_i) from D, update the DWRLQN by Q-learning algorithm [24];
16: Update the currently detected bounding box $bb_t = bb'_t$ and state $s_t = s'_t = f_s(I(bb'_t))$;
17: **end while**
18: **end for**
19: **return** θ^*;

Authorized licensed use limited to: The Libraries at Rensselaer Polytechnic Institute. Downloaded on January 13,2023 at 19:29:15 UTC from IEEE Xplore. Restrictions apply.

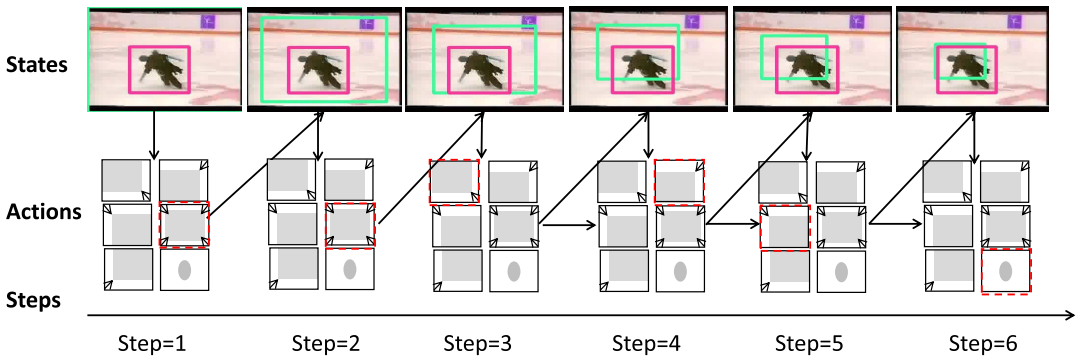


Fig. 5. Illustration of the locating process by DWRLQN, where the red bounding box is the ground truth and the green one is the detected bounding box. A sequence of actions is taken by the optimized agent to detect 'IceDancing' step-by-step. The algorithm attends regions and decides how to transform the bounding box to progressively locate the target activity. At the first search step, the agent starts from the whole frame and reaches the target activity by successively taking 'centre scaling', 'centre scaling', 'top left scaling', 'bottom left scaling', 'bottom right scaling', and 'terminate action'. It can be found that at each search step, the agent changes the detected bounding box to gradually approach the target activity until locating it at the sixth step.

In addition to the first Bellman equation term, the last two terms are derived from the action-consistency constraint and minimum entropy constraint. In the minimum entropy constraint term, the output probability of the $i-th$ sample for the $k-th$ action $p_i(a_k) = p(a_i = k|s_i)$ is computed by adding a softmax layer.

D. Locating Activity With the Optimized Agent

Once the agent is trained using the procedure described above, the testing is set up in a similar manner to that of the training runs, with the exception that the agent does not receive rewards and the model is not updated. During testing, the localization agent takes a sequence of actions to transform the bounding box following the optimal policy, where the action with the highest predicted Q-value is selected at each search step. Thus, it forms a search path from the whole frame to the region of the target activity. After reaching the terminate action, we stop the search process and consider the last region in the optimal search path as the final detection proposal. Fixing the number of search steps is an alternative to terminating the search. Consider the 'IceDancing' as an example in Fig.5. Starting from the whole frame, the agent reaches the target by successively taking 'centre scaling', 'centre scaling', 'top left scaling', 'bottom left scaling', 'bottom right scaling', and 'terminate action'. At each search step, the agent changes the detected bounding box so that it gradually approaches the target activity until locating it at the sixth step. In addition to efficiency, the policy is appealing in that it mimics the human attention mechanism and follows human perception procedure.

The Q-value is designed for proposal extraction rather than for activity classification because it just estimates the improvement of each change instead of the discriminative score; thus, we train an external classification network to recognize these activity proposals. The state-of-the-art human recognition model is adopted [5], namely Two-Stream Inflated 3D ConvNet (I3D), which is pre-trained on Kinetics. We replace its last layer by a new classification FC layer of 24 neurons, 10 neurons or 12 neurons (consisting of 24 UCF-101 detection classes, 10 UCF-Sports detection classes or 12 sub-JHMDB detection classes) during fine-tuning. Note that we do not focus on the recognition task and any other recognition models can be utilized here.

IV. Experiments

We evaluate the localization performance of our DWRLQN on the UCF-Sports [29], sub-JHMDB [16], [45], ActivityNet [2] and UCF-101 [37]. In this section, we first introduce the datasets, evaluation metrics and implementation details. Then, we analyse our method comprehensively, consisting of evaluating each weak constraint, evaluating DWRLQN under sparse spatial supervision and evaluating different action selections. Finally, we compare DWRLQN with the state-of-the-art methods and show some qualitative results.

A. Datasets and Evaluation

UCF-Sports [29]: This dataset is used for activity localization to detect spatial locations of activities in realistic scenes. Videos in UCF-Sports are already segmented into short clips and bounding-box annotations are provided for all frames as well. It includes 150 clips from various sport events with 10 categories of sport activities: diving, golf, kicking, lifting, horse riding, running, skate boarding, swing bench, swing side and walking.

Sub-JHMDB [16], [45]: The JHMDB is fully annotated with body joints. Its aim is to evaluate activity recognition performance with thoroughly human annotated data. The sub-JHMDB is a subset of JHMDB where all the joints are inside the frame, and it contains 316 clips with 12 activity categories: Catch, ClimbStairs, Golf, Jump, Kick, Pick, Pullup, Push, Run, ShootBall, Swing and Walk. It shows that this subset is much more challenging than the complete dataset [16] for activity recognition.

UCF-101 [37]: This large dataset is collected from YouTube for activity recognition with more than 13000 videos and 101 categories. For the activity localization task, there are detailed spatial location annotations for a subset that contains 3207 video sequences with 24 categories of activities: Basketball, BasketballDunk, Biking, CliffDiving, CricketBowling, Diving, Fencing, FloorGymnastics, GolfSwing, HorseRiding,

Authorized licensed use limited to: The Libraries at Rensselaer Polytechnic Institute. Downloaded on January 13,2023 at 19:29:15 UTC from IEEE Xplore. Restrictions apply.

TABLE I

ACTIVITY LOCALIZATION RESULTS UTILIZING DIFFERENT WEAK CONSTRAINTS ON UCF-SPORTS AND SUB-JHMDB WHEN ONLY ANNOTATING THE FIRST FRAME PER VIDEO, WHERE 'CONSTRAINT A' REFERS TO THE SMOOTHNESS CONSTRAINT; 'CONSTRAINT B' MEANS THE INCREASING CLASSIFICATION-CONFIDENCE CONSTRAINT; 'CONSTRAINT C' DENOTES THE CONSISTENT-APPEARANCE CONSTRAINT; 'CONSTRAINT D' REPRESENTS THE ACTION-CONSISTENCY CONSTRAINT; AND 'CONSTRAINT E' IS THE MINIMUM ENTROPY CONSTRAINT

Methods / mAP@IoU	UCF-Sports				sub-JHMDB			
	0.05	0.1	0.2	0.3	0.05	0.1	0.2	0.3
First frame	0.6680	0.5797	0.4073	0.2684	0.4889	0.4714	0.3947	0.2482
First frame + Constraint A	0.7202	0.6752	0.5927	0.4488	0.5850	0.5587	0.4597	0.3026
First frame + Constraint A,B	0.6721	0.6359	0.5626	0.4549	0.5472	0.5403	0.4776	0.3037
First frame + Constraint A,B,C	0.7787	0.7251	0.6103	0.4828	0.5861	0.5715	0.5005	0.3298
First frame + Constraint A,B,C,D	0.7323	0.6873	0.6406	0.5099	0.5607	0.5510	0.5144	0.3336
First frame + Constraint A,B,C,D,E	0.7510	0.7365	0.6750	0.5112	0.6069	0.5982	0.5430	0.3774

IceDancing, LongJump, PoleVault, RopeClimbing, SalsaSpin, SkateBoarding, Skiing, Skijet, SoccerJuggling, Surfing, TennisSwing, TrampolineJumping, VolleyballSpiking and WalkingWithDog. In contrast to UCF-Sports and Sub-JHMDB, videos in UCF-101 are relatively untrimmed, which makes it more realistic and challenging for both the spatial and temporal localization tasks.

ActivityNet-1.3 [2]: is a large dataset for activity temporal localization, including more than 648 hours of untrimmed videos. There are 19228 videos a total from 200 activity categories. The distribution among training, validation and testing is approximately 50%, 25%, and 25% respectively.

Evaluation Metrics: A localization is considered as correct if its intersection over union (IoU) with the ground truth is above a threshold δ. In this paper, a detection is accepted correct if both the predicted activity label and the predicted location match the ground truth. To fully evaluate our DWRLQN, we consider the IoU threshold of [0.05,0.1,0.2,0.3] for spatial localization on UCF-Sports, sub-JHMDB and UCF-101, and [0.5,0.75] for temporal localization on ActivityNet and UCF-101. By default, the reported metric is the mean average precision (mAP) at the IoU threshold $\delta = 0.2$, which comprehensively represents the relationship between recall and precision. In addition, we report the receiver operating characteristic curve (ROC) as calculated in previous works, which is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings. Compared with the ROC [7], the mAP is a more suitable metric for localization task, since it is not impacted by the set of negative detections. For instance, if we add many easy negative samples, *i.e.*, negatives that are ranked after all positives, the ROC will change significantly while the mAP remains the same.

Implementation Details: For the UCF-Sports, Sub-JHMDB, UCF-101 and ActivityNet, we employ the standard training and test split given in [16], [29], [37], [45] and [2]. We train the DWRLQN using reinforcement learning with a ε-greedy policy for 15 epochs, each completed after performing an episode for all training video sequences. During the ε-greedy training, ε is annealed linearly from 1 to 0.1 over the first 10 epochs, which is called the exploration process and allows the agent to utilize its own learned search policy progressively. Then, ε is fixed to 0.1 in the last 5 epochs, and the agent updates the DWRLQN parameters from experiences produced by its own decisions, which is considered an exploitation process. The parameters are optimized by the Adam optimizer [18] with a learning rate of 1e-6, and we also use dropout regularization [38] to avoid overfitting. For the training process, the discount factor γ in Eq.(13) is set to 0.9, and we run each episode with a maximum of 100 search steps. We utilize an experience pool with size $|D| = 1000$, and the mini-batch size is set to 64. By default, the scaling factor α of the action in Eq.(1) is set to 1/8, which will be evaluated further. Meanwhile, the threshold τ of the terminate reward in Eq.(4) is set to 0.5, and τ_{smo}, τ_{app}, τ_{class} in Eq.(10)(11)(12) are auto-adapted and are equal to their corresponding scores calculated on the whole region of the current frame.

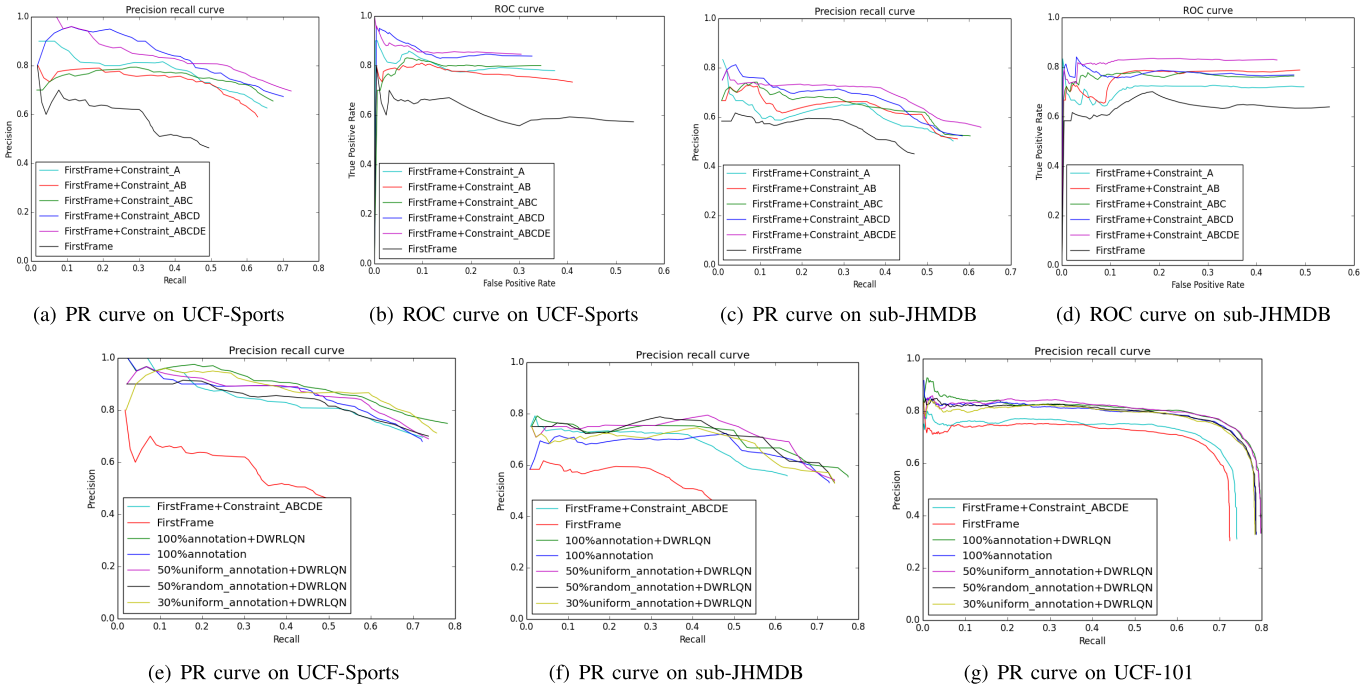
B. Evaluation of DWRLQN for Activity Localization

The main strength of our approach lies in integrating some weak constraints into DQN for weak activity localization. To comprehensively verify this advantage, we evaluate DWRLQN for localization task with respect to three aspects: effectiveness of each weak constraint, performance of sparse spatial supervision and impact of action selection.

1) Evaluation of Each Weak Constraint for Activity Localization: To evaluate the effectiveness of each weak constraint for activity localization, we compare their mAPs on UCF-Sports and sub-JHMDB. As shown in Table I, only the annotation of the first frame is available, and then different weak constraints are gradually added. Our DWRLQN combines all five weak constraints together, namely, 'First frame + Constraint A,B,C,D,E' is equivalent to 'First frame + DWRLQN'. For the default IoU = 0.2, we also analyse the ROC curves and precision recall curves for models trained with different weak constraints, as shown in Fig.6(a)-6(d). These two kinds of curves are both plotted by ranking all proposals generated by the localization agent.

From these comparisons, the following points are indicated: 1. These weak constraints are indeed helpful for localization and can complement incomplete annotations and improve performance. By integrating these weak constraints, we achieve a significant improvement on both UCF-Sports and sub-JHMDB. It can be seen that at IoU=0.3, the mAP increases from 26.84% to 51.12% on UCF-Sports and from 24.82% to 37.74% on sub-JHMDB. Additionally, it is worth noting that the results increase significantly at high IoU, but increase

Authorized licensed use limited to: The Libraries at Rensselaer Polytechnic Institute. Downloaded on January 13,2023 at 19:29:15 UTC from IEEE Xplore. Restrictions apply.



(a) PR curve on UCF-Sports (b) ROC curve on UCF-Sports (c) PR curve on sub-JHMDB (d) ROC curve on sub-JHMDB

(e) PR curve on UCF-Sports (f) PR curve on sub-JHMDB (g) PR curve on UCF-101

Fig. 6. The ROC curves and precision recall curves for activity localization on UCF-Sports, sub-JHMDB and UCF-101. The first four curves compare models trained with different constraints, while the last three compare models trained with different percentages of annotations.

slightly at low IoU. In more detail, the values of mAPs are almost constant in all cases at IoU=0.05, with a variation of less than 8%, while the improvement is almost 25% at IoU=0.3 on the UCF-Sports dataset. That is, the annotation of the first frame only provides a coarse region of position for the target activity, and these weak constraints contribute to refine it to obtain a more accurate spatial boundary. 2. The smoothness constraint is the fundamental one and other constraints are utilized to fine-tune its result, where the previous detection can be considered a weak annotation for the current frame. In addition to the smoothness constraint, the consistent-appearance constraint also provides strong evidence, which helps the localization agent track the same actor. Overall, after adding the smoothness constraint, we obtain 18.04% and 5.44% improvements on the two datasets, as well as 2.79% and 2.61% improvements with the consistent-appearance constraint. Meanwhile, other three constraints are also beneficial for improving the performance with approximately 1% gains. Specifically, the increasing classification-confidence constraint makes the extracted proposals more discriminative, and the action-consistency constraint and minimum entropy constraint help the localization agent make decisions with more confidence and consistency. In addition to 'First frame + Constraint A', we evaluate 'First frame + Constraint B' and 'First frame + Constraint C' for activity localization. Experimental results show that the two constraints lead to termination very quickly, since they are too weak to independently locate human activity. In addition, the action-consistency constraint and minimum entropy constraint cannot be used individually for frames without annotations because they only provide a way to redefine the loss function instead of the reward function. 3. The model's performance steadily increases by integrating more weak constrains and combining all of the weak constrains achieves the highest precision. The reason is that integrating more weak constraints can provide more supervised information for training, and different weak constraints can complement each other, which makes the weak constraint stronger.

2) Evaluation of DWRLQN Under Sparse Spatial Supervision: To measure the effectiveness of DWRLQN under sparse spatial supervision, we evaluate the model with different percentages of annotations in Table II. We measure the impact of varying the number of annotated frames from only 1 to all per video. Furthermore, we also give a comparison between the model with and without weak constraints. Among them, '100% annotation' is treated as a baseline, which means that the model trained under full supervision, where the reward is computed by Eq.(5) and Eq.(4), and the loss is computed by Eq.(13). The '50% annotation + DWRLQN' denotes that the model is trained by half of the data with the ground truth and half of the data only with weak constraints, where the reward and loss are calculated by Eq.(5), Eq.(4) and Eq.(13) for frames with annotations and by Eq.(9) and Eq.(16) for frames without annotations. The '100% annotation + DWRLQN' denotes that the model is trained by complete annotations and weak constraints, which integrates the reward and loss with the weak reward and weak loss. For a more fair comparison, we utilize two sampling strategies to annotate frames: random sampling and uniform sampling, which is less sensitive to the change of the frame rate.

In summary, it is beneficial to combine weak constraints for activity localization under sparse spatial supervision, which improves performance significantly, *e.g.*, the results of 'First frame + DWRLQN' are nearly 24.28%, 12.92% and 5.09%

Authorized licensed use limited to: The Libraries at Rensselaer Polytechnic Institute. Downloaded on January 13,2023 at 19:29:15 UTC from IEEE Xplore. Restrictions apply.

TABLE II
ACTIVITY LOCALIZATION RESULTS UTILIZING DIFFERENT PERCENTAGES OF ANNOTATIONS ON UCF-SPORTS, SUB-JHMDB AND UCF-101

Methods / mAP@IoU	UCF-Sports			sub-JHMDB			UCF-101		
	0.1	0.2	0.3	0.1	0.2	0.3	0.1	0.2	0.3
First frame	0.5797	0.4073	0.2684	0.4714	0.3947	0.2482	0.7223	0.5890	0.4144
First frame + DWRLQN	0.7365	0.6750	0.5112	0.5982	0.5430	0.3774	0.7559	0.6211	0.4653
100%annotation	0.7252	0.6863	0.6212	0.6607	0.6174	0.4443	0.7534	0.6680	0.5064
30%annotation (uniform) + DWRLQN	0.7344	0.7176	0.6160	0.6818	0.6371	0.4706	0.7652	0.6719	0.5151
50%annotation (random) + DWRLQN	0.7292	0.7011	0.6376	0.6748	0.6565	0.4943	0.7889	0.6740	0.5080
50%annotation (uniform) + DWRLQN	0.7418	0.7059	0.6332	0.6974	0.6677	0.5028	0.7804	0.6859	0.5241
100%annotation + DWRLQN	0.7861	0.7437	0.6608	0.6910	0.6732	0.5175	0.7794	0.6853	0.5256

higher than those of 'First frame'. Similarly, '100% annotation + DWRLQN' outperforms '100% annotation' by 3.96%, 7.32% and 1.92% on UCF-Sports, sub-JHMDB and UCF-101, respectively. Furthermore, compared with the full supervision, the drop of mAP is very small even for the worst case where only 1 frame is annotated per video by complementing these weak constraints. It can be seen that the difference between the performance of 'First frame + DWRLQN' and '100% annotation' is only 6.69% and 4.11% on sub-JHMDB and UCF-101, respectively. In another instance, the mAP of '50% annotation + DWRLQN' is very close to that of DWRLQN with complete annotations. That is, we can achieve an acceptable or a comparable localization performance only requiring annotation of the first frame or half of the frames for a video sequence. Obviously, when implemented without weak constraints, the localization precision increases as more data frames are annotated, *e*.g., the precision of '100% annotation' is significantly higher than that of 'First frame'. It is worth noting that combining weak constraints increases the result significantly only for the case with sparse spatial annotations and increases slightly for that with complete annotations. This effect occurs mainly because the complete annotations already potentially contain this prior knowledge and weak constraints. More interestingly, the result demonstrates that DWRLQN trained with half annotations and weak constraints even outperforms the model with full supervision by an average of 1.20%, 5.85% and 1.77% on the three datasets. Last, since the uniform sampling can preserve the maximum temporal relationship, its localization precision is better than that of random sampling's. To illustrate this result further, we plot the PR curves in Figs.6(e)-6(g), which also confirm above conclusions.

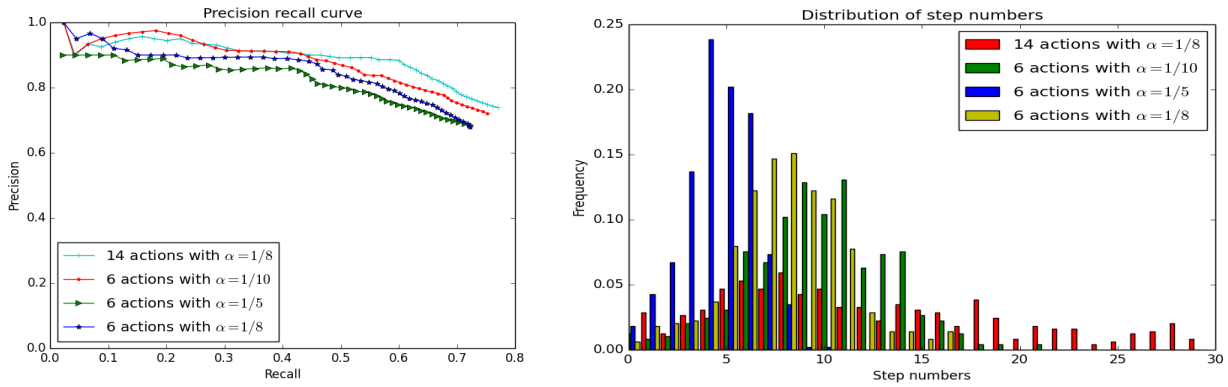
3) Impact of Action Selection: The scaling factor α and the action type have an influence on both the precision and speed of localization. Therefore, we measure the impact of these factors on the final localization performance by running DWRLQN with different action selections on UCF-Sports. As illustrated in Fig.7, the mAP measures the model's localization performance, while the average number of search steps decides the speed of localization. Specifically, utilizing 6 actions is the simplest way to transform the bounding box to cover any location of an image, while the following additional 8 actions result in a total of 14 actions: moving left/right, moving up/down, becoming shorter/longer horizontally and becoming shorter/longer vertically.

The results show the following: 1. With a decreasing value of the scaling factor, the mAP increases and the number of required search steps also increases. 2. Compared with 6 actions, utilizing 14 actions achieves a more accurate localization boundary but requires many more steps to locate the target activity. In Fig.7(b), we just show the distribution ranging from 1 to 30 steps. Actually, the distribution of '14 actions with $\alpha = 1/8$' has a long tail, with 82.24% of detections requiring fewer than 30 steps, because utilizing 14 actions or a small scaling factor makes the search process slower and more refined. To balance the localization precision with the convergence speed, we adopt 6 actions (with $\alpha = 1/8$) by default in this paper, which results in satisfactory localization performance with fewer search steps.

C. Comparison With the State-of-the-Art Methods

We continue the evaluation for localization task in a comparison with the state-of-the-art methods in Table III. We report our results with two settings: one for training on sparse spatial annotations (last two lines), and another for training on complete annotations (third row from the bottom). We first compare our approach with [23] and [48], which are both sparse spatial supervised methods for activity localization. In the case where there is only one annotated frame for each video, we obtain a 4.71% performance gain over [48] by integrating prior knowledge and weak constraints. With weaker supervision, our model substantially outperforms [23] by 27.31%, *i.e.,* annotation of 1 point per frame *vs* annotation of 1 bounding box (2 points) per video. Then, we compare our model with some fully supervised approaches, including the currently best detection methods based on faster R-CNN [28], [30]. Both in full supervision and sparse supervision, our DWRLQN already outperforms the state-of-the-art methods on UCF-Sports and sub-JHMDB, and it is on par with [30] on UCF-101. More remarkable, the localization mAP of DWRLQN trained with incomplete annotations is even higher than that of most fully supervised methods [28], [36], [45]. In particular, we achieve better performance with much fewer proposals, *e.g.* approximately 10 (ours) *vs* 10K [52], and approximately 10 (ours) *vs* 256 [28], [30]. The results clearly demonstrate the superior ability of DWRLQN in extracting proposals both effectively and efficiently as well as the benefit of explicitly integrating weak constraints for sparse spatial supervised learning.

Authorized licensed use limited to: The Libraries at Rensselaer Polytechnic Institute. Downloaded on January 13,2023 at 19:29:15 UTC from IEEE Xplore. Restrictions apply.



(a) PR curve for DWRLQN with different action selections. (b) Distribution of the required number of steps for DWRLQN with different action selections.

Fig. 7. The precision recall curve and the distribution of the required number of steps for DWRLQN with different action selections on UCF-Sports. The action selection is in terms of the scaling factor and the action type.

TABLE III
COMPARISON WITH THE STATE-OF-THE-ART METHODS FOR ACTIVITY LOCALIZATION

Methods/IoU	Proposal Number	Supervision way	UCF-Sports	sub-JHMDB	UCF-101	
			0.2	0.2	0.05	0.2
Yu *et al.* [52]	10K	Full supervision	-	-	49.9%	42.8%
Wang *et al.* [45]	-	Full supervision	47%	36%	-	-
Gemert *et al.* [9]	1449/6706	Full supervision	54.6%	-	-	34.5%
Soomro *et al.* [36]	200-300	Full supervision	55%	42%	-	-
Peng *et al.* [28]	256	Full supervision	-	-	54.46%	42.27%
Saha *et al.* [30]	256	Full supervision	-	-	79.12%	66.75%
Gkioxari *et al.* [11]	2K	Full supervision	68.1%	36.2%	-	-
Weinzaepfel *et al.* [47]	256	Full supervision	-	63.1%	54.28%	46.77%
Mettes *et al.* [23]	1449/6706	Sparse: per point per frame	54.5%	-	-	34.8%
Weinzaepfel *et al.* [48]	256	Sparse: per frame per video	-	63.9%	70.0%	57.4%
100%annotation + DWRLQN	Approximately 10	Full supervision	**74.37%**	**67.32%**	**81.88%**	68.53%
First frame + DWRLQN	Approximately 10	Sparse: per frame per video	67.50%	54.30%	81.78%	62.11%
50%annotation + DWRLQN	Approximately 10	Sparse: half of frames per video	70.59%	66.77%	81.44%	**68.59%**

D. Temporal Localization Evaluation

Our model can also be used for temporal localization to find the time interval of an activity that consists of the start frame and the end frame by simply replacing the spatial tuple with the temporal tuple (s, a, r). Specifically, a temporal pooling layer is used to generate the temporal state by averaging the feature of each frame within the time interval; the action set for temporal localization is composed of 3 scaling actions (left scaling, right scaling, and centre scaling), 2 translation actions (right shifting and left shifting) and one terminate action; and the temporal reward is calculated in a similar manner to that of the spatial reward by replacing the bounding box with the time interval in Eq.(4) and Eq.(5).

In Table IV, we report the mAP of the temporal localization obtained by our method on UCF-101 and ActivityNet. For temporal localization, our method outperforms [12], [31], [49] on ActivityNet at both IoU = 0.5 and IoU = 0.75. Compared with [22], which is a typical two-stream model, we obtain a higher mAP at IoU = 0.5 and comparable performance with a similar proposal number at IoU = 0.75.

TABLE IV
TEMPORAL LOCALIZATION RESULTS OF OUR METHOD

Methods/IoU	UCF-101		
	mAP@0.5	mAP@0.75	Proposal Num
Our method	71.85%	57.58%	3.23
Methods/IoU	**ActivityNet**		
	mAP@0.5	mAP@0.75	Proposal Num
Wang *et al.* [46]	42.28%	3.76%	-
SCC [12]	40.0%	17.90%	100
SSN [49]	39.12%	23.48%	56
CDC [31]	43.83%	25.88%	11.2
Lin@5 [22]	42.57%	28.26%	5
Lin@100 [22]	44.39%	**29.65%**	100
Our method	**45.85%**	28.08%	4.67

E. Qualitative Analysis

We illustrate some qualitative results to show how the DWRLQN model refines the bounding boxes in Fig.8. It is observed that for most cases (as seen in the top five rows), the agent successfully zooms towards the target activity and completes the search process in a few steps. For example, even for small instances of targets (as shown in the first and third rows), the agent also accurately transforms the bounding boxes to approach the target activities with just five or six steps. However, for some cases, where the target has a small size as well as a similar appearance to the background, the agent can be confused and terminate at a wrong location (as seen in the last row). Furthermore, we illustrate the distributions of the

Authorized licensed use limited to: The Libraries at Rensselaer Polytechnic Institute. Downloaded on January 13,2023 at 19:29:15 UTC from IEEE Xplore. Restrictions apply.

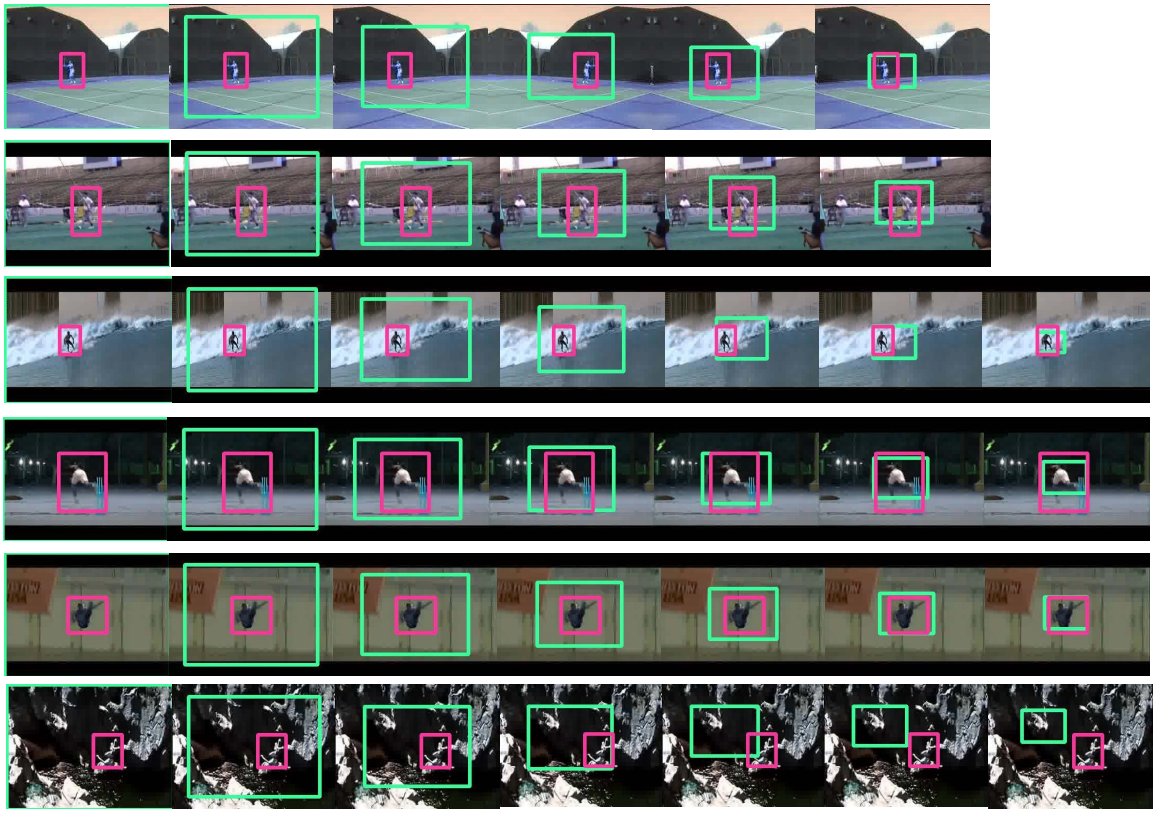


Fig. 8. Visualizations of the search process by the localization agent. Except for the last line, the agent locates all the target activities successfully. The red bounding box is the ground truth and the green one is our detection by DWRLQN.

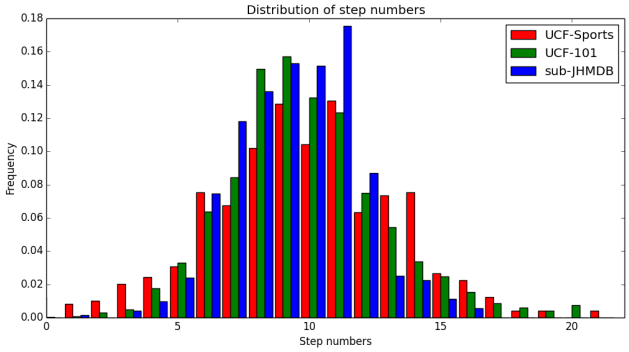


Fig. 9. The distributions of step numbers necessary to locate target activity on UCF-Sports, sub-JHMDB and UCF-101. Almost all detections require fewer than 20 steps to be obtained, with an average of approximately 10 steps.

TABLE V

THE DISTRIBUTIONS OF SELECTED ACTIONS DURING THE LOCATING PROCESS ON UCF-SPORTS, SUB-JHMDB AND UCF-101

Action	1	2	3	4	5	6
UCF-Sports	0.173	0.192	0.143	0.191	0.208	0.093
sub-JHMDB	0.151	0.144	0.172	0.249	0.188	0.097
UCF-101	0.174	0.134	0.214	0.218	0.167	0.093

required step numbers and selected actions during the locating process in Fig.9 and Table V. It is worth noting that almost all detections require fewer than 20 steps to be obtained, with an average of 10.69, 10.31 and 10.71 for UCF-Sports, sub-JHMDB and UCF-101, respectively.

V. CONCLUSION

In this paper, we propose a unified deep Q-network with weak reward and weak loss (DWRLQN) for human activity localization under sparse spatial supervision. This approach integrates some weak knowledge and weak constraints into the reward function and the loss function to complement incomplete annotations. Following the human perception procedure, we learn a coarse-to-fine searching strategy. More specifically, a localization agent is trained to locate activities gradually and progressively, with the ability to decide where to focus the attention next via pre-defined actions. By casting the activity localization problem as a Markov decision process and searching for the target activity in terms of the learned policy, we obtain a more accurate spatial boundary without significantly increasing the number of proposals. By taking full advantage of the weak information involved in an activity video sequence, we achieve comparable and even better performance than fully supervised methods.

REFERENCES

[1] M. Bellver, X. Giro-i-Nieto, F. Marques, and J. Torres, "Hierarchical object detection with deep reinforcement learning," 2016, *arXiv:1611.03718*. [Online]. Available: https://arxiv.org/abs/1611.03718

[2] F. C. Heilbron, V. Escorcia, B. Ghanem, and J. C. Niebles, "ActivityNet: A large-scale video benchmark for human activity understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 961–970.

[3] J. C. Caicedo and S. Lazebnik, "Active object localization with deep reinforcement learning," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 2488–2496.

Authorized licensed use limited to: The Libraries at Rensselaer Polytechnic Institute. Downloaded on January 13,2023 at 19:29:15 UTC from IEEE Xplore. Restrictions apply.

[4] Q. Cao, L. Lin, Y. Shi, X. Liang, and G. Li, "Attention-aware face hallucination via deep reinforcement learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 690–698.

[5] J. Carreira and A. Zisserman, "Quo vadis, action recognition? A new model and the kinetics dataset," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 6299–6308.

[6] J. Choi, J. Kwon, and K. M. Lee, "Real-time visual tracking by deep reinforced decision making," 2017, *arXiv:1702.06291*. [Online]. Available: https://arxiv.org/abs/1702.06291

[7] M. Everingham, L. V. Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The Pascal visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, 2010.

[8] A. Gaidon, Z. Harchaoui, and C. Schmid, "Temporal localization of actions with actoms," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 11, pp. 2782–2795, Nov. 2013.

[9] J. C. Van Gemert, M. Jain, E. Gati, and C. G. M. Snoek, "APT: Action localization proposals from dense trajectories," in *Proc. BMVC*, 2015, p. 4.

[10] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 1440–1448.

[11] G. Gkioxari and J. Malik, "Finding action tubes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 759–768.

[12] F. C. Heilbron, W. Barrios, V. Escorcia, and B. Ghanem, "SCC: Semantic context cascade for efficient action detection," in *Proc. CVPR*, Jul. 2017, pp. 3175–3184.

[13] F. C. Heilbron, J. C. Niebles, and B. Ghanem, "Fast temporal activity proposals for efficient detection of human actions in untrimmed videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 1914–1923.

[14] M. Hoai, L. Torresani, F. D. L. Torre, and C. Rother, "Learning discriminative localization from weakly labeled data," *Pattern Recognit.*, vol. 47, no. 3, pp. 1523–1534, 2014.

[15] M. Jain, J. Van Gemert, and H. Jégou, P. Bouthemy, and C. G. M. Snoek, "Action localization with Tubelets from motion," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 740–747.

[16] H. Jhuang, J. Gall, S. Zuffi, C. Schmid, and M. J. Black, "Towards understanding action recognition," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 3192–3199.

[17] Z. Jie, X. Liang, J. Feng, X. Jin, W. Lu, and S. Yan, "Tree-structured reinforcement learning for sequential object localization," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 127–135.

[18] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," Dec. 2014, *arXiv:1412.6980*. [Online]. Available: https://arxiv.org/abs/1412.6980

[19] X. Kong, B. Xin, Y. Wang, and G. Hua, "Collaborative deep reinforcement learning for joint object search," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 1695–1704.

[20] T. Lan, Y. Wang, and G. Mori, "Discriminative figure-centric models for joint action localization and recognition," in *Proc. Int. Conf. Comput. Vis.*, Nov. 2011, pp. 2003–2010.

[21] X. Liang, L. Lee, and E. P. Xing, "Deep variation-structured reinforcement learning for visual relationship and attribute detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 848–857.

[22] T. Lin, X. Zhao, and Z. Shou, "Temporal convolution based action proposal: Submission to Activitynet 2017," 2017, *arXiv:1707.06750*. [Online]. Available: https://arxiv.org/abs/1707.06750

[23] P. Mettes, J. C. Van Gemert, and C. G. M. Snoek, "Spot on: Action localization from pointly-supervised proposals," in *Proc. Eur. Conf. Comput. Vis.*, 2016.

[24] V. Mnih *et al.*, "Playing Atari with deep reinforcement learning," 2013, *arXiv:1312.5602*. [Online]. Available: https://arxiv.org/abs/1312.5602

[25] V. Mnih *et al.*, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, pp. 529–533, Feb. 2015.

[26] E. A. Mosabbeb, R. Cabral, F. De la Torre, and M. Fathy, "Multi-label discriminative weakly-supervised human activity recognition and localization," in *Proc. Asian Conf. Comput. Vis.*, 2014.

[27] A. Nair *et al.*, "Massively parallel methods for deep reinforcement learning," 2015, *arXiv:1507.04296*. [Online]. Available: https://arxiv.org/abs/1507.04296

[28] X. Peng and C. Schmid, "Multi-region two-stream R-CNN for action detection," in *Proc. Eur. Conf. Comput. Vis.*, 2016.

[29] M. D. Rodriguez, J. Ahmed, and M. Shah, "Action MACH a spatio-temporal Maximum Average Correlation Height filter for action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–8.

[30] S. Saha, G. Singh, M. Sapienza, P. H. Torr, and F. Cuzzolin, "Deep learning for detecting multiple space-time action tubes in videos," 2016, *arXiv:1608.01529*. [Online]. Available: https://arxiv.org/abs/1608.01529

[31] Z. Shou, J. Chan, A. Zareian, K. Miyazawa, and S.-F. Chang, "CDC: Convolutional-de-convolutional networks for precise temporal action localization in untrimmed videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5734–5743.

[32] Z. Shou, D. Wang, and S.-F. Chang, "Temporal action localization in untrimmed videos via multi-stage CNNs," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1049–1058.

[33] D. Silver *et al.*, "Mastering the game of go without human knowledge," *Nature*, vol. 550, no. 7676, pp. 354–359, 2017.

[34] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: https://arxiv.org/abs/1409.1556

[35] P. Siva and T. Xiang, "Weakly supervised action detection," in *Proc. BMVC*, vol. 2, 2011, p. 6.

[36] K. Soomro, H. Idrees, and M. Shah, "Action localization in videos through context walk," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 3280–3288.

[37] K. Soomro, A. R. Zamir, and M. Shah, "UCF101: A Dataset of 101 human actions classes from videos in the wild," 2012, *arXiv:1212.0402*. [Online]. Available: https://arxiv.org/abs/1212.0402

[38] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.

[39] C. Sun, S. Shetty, R. Sukthankar, and R. Nevatia, "Temporal localization of fine-grained actions in videos by domain transfer from Web images," in *Proc. 23rd Int. Conf. Multimedia*, 2015, pp. 371–380.

[40] R. S. Sutton, *Reinforcement Learning: An Introduction*, vol. 135. Cambridge, MA, USA: MIT Press, 1998.

[41] Y. Tian, R. Sukthankar, and M. Shah, "Spatiotemporal deformable part models for action detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 2642–2649.

[42] D. Tran and J. Yuan, "MAX-margin structured output regression for Spatio-temporal action localization," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 350–358.

[43] H. Van Hasselt, A. Guez, and D. Silver, "Deep reinforcement learning with double q-learning," in *Proc. AAAI*, 2016, pp. 2094–2100.

[44] L. Wang, Y. Qiao, and X. Tang, "Action recognition and detection by combining motion and appearance features," *THUMOS Action Recognit. Challenge*, vol. 1, no. 2, p. 2, 2014.

[45] L. Wang, Y. Qiao, and X. Tang, "Video action detection with relational dynamic-poselets," in *Proc. Eur. Conf. Comput. Vis.*, 2014.

[46] R. Wang and D. Tao, "Uts at activitynet 2016," in *Proc. AcitivityNet Large Scale Activity Recognit. Challenge*, 2016, p. 8.

[47] P. Weinzaepfel, Z. Harchaoui, and C. Schmid, "Learning to track for spatio-temporal action localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 3164–3172.

[48] P. Weinzaepfel, X. Martin, and C. Schmid, "Human action localization with sparse spatial supervision," 2017, *arXiv:1605.05197*. [Online]. Available: https://arxiv.org/abs/1605.05197

[49] Y. Xiong, Y. Zhao, L. Wang, D. Lin, and X. Tang, "A pursuit of temporal accuracy in general activity detection," 2017, *arXiv:1703.02716*. [Online]. Available: https://arxiv.org/abs/1703.02716

[50] J. Yang and J. Yuan, "Common action discovery and localization in unconstrained videos," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 2176–2185.

[51] S. Yeung, O. Russakovsky, G. Mori, and L. Fei-Fei, "End-to-end learning of action detection from frame glimpses in videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 2678–2687.

[52] G. Yu and J. Yuan, "Fast action proposals for human action detection and search," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 1302–1311.

[53] J. Yuan, Z. Liu, and Y. Wu, "Discriminative subvolume search for efficient action detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 2442–2449.

[54] J. Yuan, Z. Lin, and Y. Wu, "Discriminative video pattern search for efficient action detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 9, pp. 1728–1743, Sep. 2011.

[55] S. Yun, J. Choi, Y. Yoo, K. Yun, and Y. C. Jin, "Action-decision networks for visual tracking with deep reinforcement learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 2711–2720.

Authorized licensed use limited to: The Libraries at Rensselaer Polytechnic Institute. Downloaded on January 13,2023 at 19:29:15 UTC from IEEE Xplore. Restrictions apply.

[56] D. Zhang, H. Maei, X. Wang, and Y.-F. Wang, "Deep reinforcement learning for visual object tracking in videos," 2017, *arXiv:1701.08936*. [Online]. Available: https://arxiv.org/abs/1701.08936

Wanru Xu received the B.S. degree in biomedical engineering and the Ph.D. degree in signal and information processing from Beijing Jiaotong University, Beijing, China, in 2011 and 2017, respectively. She is currently a Postdoctoral Researcher with the Institute of Information Science, School of Computer and Information Technology, Institute of Information Science, Beijing Jiaotong University, Beijing. Her current research interests include computer vision, machine learning, and pattern recognition.

Zhenjiang Miao (M'11) received the B.E. degree from Tsinghua University, Beijing, China, in 1987, and the M.E. and Ph.D. degrees from Northern Jiaotong University, Beijing, in 1990 and 1994, respectively. From 1995 to 1998, he was a Postdoctoral Fellow with École Nationale Supérieure d'Electrotechnique, d'Electronique, d'Informatique, d'Hydraulique et des Télécommunications, Institut National Polytechnique de Toulouse, Toulouse, France, and was a Researcher with the Institute National de la Recherche Agronomique, Sophia Antipolis, France. From 1998 to 2004, he was with the Institute of Information Technology, National Research Council Canada, Nortel Networks, Ottawa, Canada. He joined Beijing Jiaotong University, Beijing, in 2004. He is currently a Professor and the Director of the Media Computing Center, Institute of Information Science, Beijing Jiaotong University, and the Director of the Institute for Digital Culture Research, Center for Ethnic and Folk Literature and Art Development, Ministry of Culture, China. His current research interests include image and video processing, multimedia processing, and intelligent human-machine interaction.

Jian Yu received the B.S. and M.S. degrees in mathematics and the Ph.D. degree in applied mathematics from Peking University, Beijing, China, in 1991, 1994, and 2000, respectively. He is currently a Professor with Beijing Jiaotong University, Beijing, and the Director of the Beijing Key Laboratory of Traffic Data Analysis and Mining. His current research interests include machine learning, image processing, and pattern recognition.

Qiang Ji (F'15) received the Ph.D. degree from the University of Washington. From January 2009 to August 2010, he served as a Program Director for the National Science Foundation, managing NSF's machine learning and computer vision programs. Prior to joining RPI in 2001, he was an Assistant Professor with the Department of Computer Science, University of Nevada, Reno. He also held research and visiting positions in the Beckman Institute, University of Illinois at Urbana-Champaign, the Robotics Institute, Carnegie Mellon University, and the US Air Force Research Laboratory. He is currently a Professor with the Department of Electrical, Computer, and Systems engineering, RPI. He is a fellow of IAPR.

Authorized licensed use limited to: The Libraries at Rensselaer Polytechnic Institute. Downloaded on January 13,2023 at 19:29:15 UTC from IEEE Xplore. Restrictions apply.