IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received September 22, 2020, accepted October 1, 2020, date of publication October 6, 2020, date of current version October 14, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3028910

Surgical Tools Detection Based on Training Sample Adaptation in Laparoscopic Videos

GUANGYAO WANG AND SHENGSHENG WANG, (Member, IEEE)
College of Computer Science and Technology, Jilin University, Changchun 130012, China
Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University, Changchun 130012, China

Corresponding author: Shengsheng Wang (wss@jlu.edu.cn)

This work was supported in part by the Science & Technology Development Project of Jilin Province, China under Grant 20190302117GX and Grant 20180101334JC, and in part by the Innovation Capacity Construction Project of Jilin Province Development and Reform Commission under Grant 2019C053-3.

ABSTRACT The performance of object detection methods plays an important role in the recognition of surgical tools, and is a key link in the automated evaluation of surgical skills. In this paper, we propose a novel framework for one-stage object detection based on a sample adaptive process controlled by reinforcement learning, which can maintain the speed advantage while maintaining higher accuracy than two-stage object detection methods. We use m2cai16-tool-locations and AJU-Set, two datasets covering seven surgical tools with spatial information collected from hospital gallbladder surgery videos to evaluate and verify the effectiveness of our proposed framework. The experiments show that our proposed framework can make the one-stage object detection method achieve 70.1% and 77.3% accuracy on m2cai16-tool-locations and AJU-Set, respectively. We further validated the effectiveness of our proposed framework by analyzing the usage patterns, motion trajectories, and mobile values of surgical tools.

INDEX TERMS Laparoscopic surgery, reinforcement learning, object detection.

I. INTRODUCTION

Surgery, as an important part of clinical medicine, plays a key role in solving human diseases. However, due to an imbalance in the level of social and economic development among regions [1], a considerable number of people cannot receive high-quality surgical treatment. In a state of lack of medical conditions, patients suffer trauma and complications due to low-quality surgical treatment, leading to a series of serious sequelae and even death. In response to this problem, the traditional model [2] in the medical field relies on assessment from senior experts to guide surgeons who need to communicate and learn, but this is limited by the impact of individual subjectivity and time-consuming processes.

To solve the abovementioned problems, in recent years, researchers have taken advantage of the rapid development of image processing technology to carry out automated assessment of surgical skills. The academic community, with the assistance of surgeons, analyzes videos recorded during operations to provide learners with a more objective, standardized and automated evaluation of surgical skills. The identification and positioning of surgical tools during surgery is the basis for automated evaluation of surgical skills and, can be achieved with the support of object detection technology.

Object detection is currently divided into two broad categories, anchor-based and anchor-free detectors. Anchor-based detectors are divided into two types: one-stage [3]–[5] and two-stage [6]–[10] detectors due to different processing methods in the preprocessing stage. Anchor-free detectors are divided into two types, keypoint-based [11]–[13] and center-based [14]–[16] detectors, due to the different positional relationship between the predicted points and the object.From a formal point of view, anchor-free object detection methods are better than anchor-based methods because they eliminate the predefined design of an anchor, but this is not the case. Recent research [17] shows that the essential difference among object detection methods lies in the strategy defined for training positive and negative samples, rather than whether to use anchors.Therefore,a factor that has an important influence on object detection is the reasonable division method of positive and negative training samples. Based on this theoretical discovery, in this paper, we propose a novel one-stage object detection framework based on a sample adaptive process controlled by reinforcement learning,

The associate editor coordinating the review of this manuscript and approving it for publication was Kok-Lim Alvin Yau.

that is used to detect surgical tools quickly and accurately. Compared with current object detection methods, our proposed framework is more targeted towards the detection of surgical tools with complex backgrounds and small training sample sizes. With the support of the proposed framework, sample adaptation allows an object detection model to set thresholds based on a sample's own attributes to more reasonably distinguish between positive and negative training samples, and with reinforcement learning control [18],it can manipulate deformations in a negative sample bounding box to reach the positive sample standard. Therefore, with the help of the proposed framework, an object detection method can focus more on the surgical tools under complex background conditions, to achieve higher accuracy while maintaining the advantage of one-stage object detection speed.

The main contributions of our work are as follows.:

- (1) For the first time, we use reinforcement learning control to optimize sample adaptation, a novel definition strategy for training positive and negative samples.
- (2) We first propose a one-stage object detection framework based on sample adaptation for the task of surgical tool detection.
- (3) The one-stage object detection supported by our proposed framework based on reinforcement learning control sample adaptation achieves better performance than other object detection methods on the cholecystectomy surgery datasets m2cai16-tool-locations and AJU-Set [19].

II. RELATED WORK

A. LAPAROSCOPIC CHOLECYSTECTOMY

With the rapid development of electronic information technology in various fields, the method of recording surgical procedures with micro lenses has been increasingly widely adopted. The early purpose of this recording method was to facilitate a surgeon returning to the operation process in reflecting and summarizing the operation links, and has laid the foundation for subsequent research on the automated analysis of surgery. With the development of machine vision, research on the automatic analysis of surgery also bears obvious traces of the times. Early traditional surgical automation analysis research [20]–[22] is used for stage analysis, and is completed by many statistical models that rely on manual design features, such as conditional random fields [23], [24], Bayes classifiers [25], hidden Markov models [21], [26].

With the advent of the milestone technology of deep convolutional neural networks (CNNs) [6], which are a milestone technology, the traditional manual design feature model has been replaced by the automatic description of features obtained by CNNs, and has led to impressive results [27]–[29] in the analysis and research of surgical automation based on CNNs. However, the application of most of the current models belongs to the frame-level tool presence detection in the M2CAI 2016 tool presence detection challenge, which is essentially a surgical training task that is different from true surgery. Compared with true surgery, surgical training only focuses on specific tasks in an operation, which makes it unable to reflect the unpredictable conditions of smoke and lens fogging and anatomical deformation that may occur in an actual surgical environment. Only the method in [30], which truly achieves the technical evaluation of the true surgical level in a complete environment, expands the spatial information for the detection of the M2CAI 2016 tool to obtain a new dataset and uses faster regions with CNN features(R-CNN) [8] as the object detection method for surgical tools detection.

B. ANCHOR-BASED VS ANCHOR-FREE MODELS

Affected by the idea of traditional classic object detection algorithms, object detection methods based on deep convolution network technology also retain the concept of anchors. As a landmark object detection model in the introduction of deep convolutional networks, R-CNN [6] greatly surpassed the performance of previous related models based on traditional methods. It is for this reason that subsequent object detection based on deep convolutional network has been deeply affected by R-CNN and two-stage object detection [7]–[10] methods were developed that still occupy an important position in this field. The methods are called two-stage object detection methods because candidate boxes are generated for images that need to be recognized first, and then detector is performed on these candidate boxes to identify the category and position. The two-stage object detection methods have achieved very impressive results in accuracy, but in practical applications, object detection requires higher execution efficiency to ensure real-time performance. In response to this problem, the academic community has proposed one-stage object detection methods [3]–[5], which combine the two parts of a two-stage methods into one. Single shot multibox detector(SSD) [4] pioneered the use of multiscale layers to directly predict objects, ensuring high efficiency while greatly improving accuracy. Since then, the academic community has put forward much work to promote its performance in different aspects [5], [31]–[34].

The object detection methods based on an anchor depend on the design of the properties of the anchor box in advance, and this design has a great influence on the effect of the object detection methods. To avoid this problem, in recent years, the academic community has proposed anchor-free object detection methods. The idea of anchor-free object detection methods is to associate objects with specific points, that is, to detect objects by predicting some points. Object detection methods that are anchor-free are divided into two types due to the different positions of these observation points on the object. One object detection method type contains observation points that are mainly distributed around the detected object, which is called keypoint-based object detection [11]–[13]. Another object detection method type contains observation points that are mainly distributed in the center of the detected object and is called center-based object detection [14]–[16].

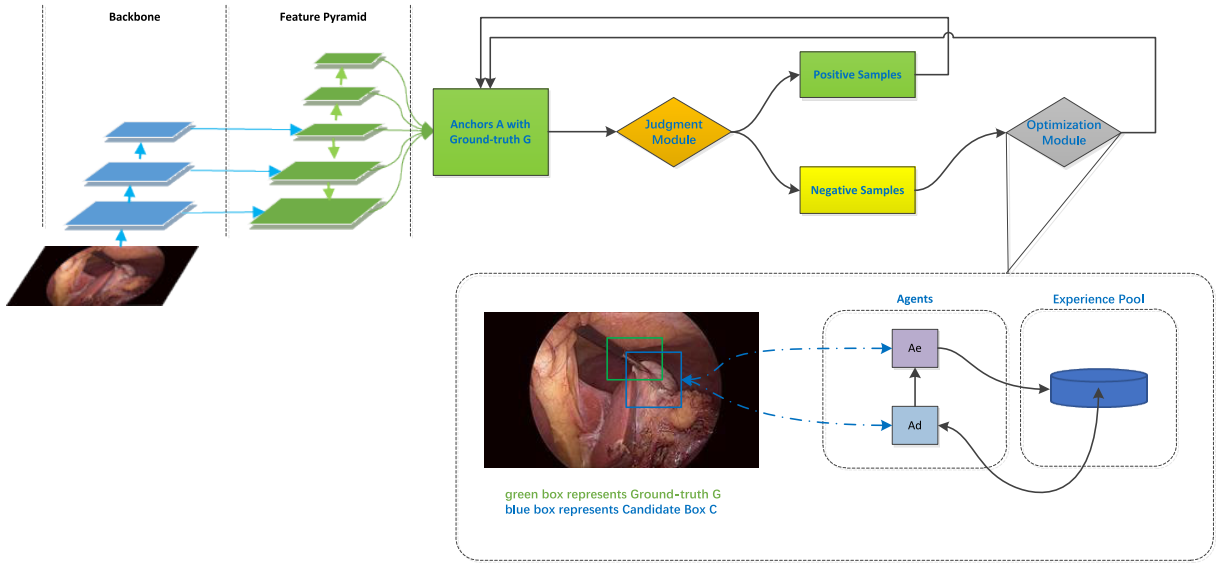

FIGURE 1. Architecture of the reinforcement training sample adaptation. After the image passes through the basic network ResNet-101 and the feature pyramid structure, several candidate boxes are generated to train the object detection model. These training samples are first sent to the judgment module separate positive training samples *P* and negative training samples *N*. Then the negative samples are sent to the optimization module, and the agent performs deformation control under the reinforcement learning framework to reach the positive sample standard.

For the method in [17], similar anchor-based and anchor-free methods were selected for an in-depth analysis of the essential differences. The conclusion of the final analysis shows that the real difference between the two method types is not the anchor boxes but the definition of positive and negative training samples.

C. REINFORCEMENT LEARNING

Reinforcement learning is an important part of the field of artificial intelligence, and is a concern of the academic community. With the successful application of deep convolutional network technology in various fields, deep reinforcement learning [18] has came into being. Due to the good performance of deep convolutional networks for feature expression, reinforcement learning can be applied to high-dimensional problems that involve images [35] and videos [36].

In the field of object detection, the method in [35] links deep reinforcement learning with object detection for the first time, providing another research idea for an object detection subject. The difference between object detection methods based on reinforcement learning and current mainstream methods [4], [5], [8] is that the former treats object detection as a sequential decision problem, thus introducing a reinforcement learning framework to solve it, and the latter treats object detection as a regression problem.

Deep reinforcement learning has achieved very attractive results in the field of machine vision, especially in video games [18], [37], [38],which have surpassed human level. However, to obtain satisfactory results the models require longer training time. This defect restricts the object recognition algorithm based on reinforcement learning. Fortunately, the distributed reinforcement learning framework [39] solves this problem very well. Only raising the CPU core can greatly improve the efficiency of model operations without increasing the GPU requirements.

III. METHODOLOGY

According to the conclusion drawn by the sample adaptive method analysis [17], we know that the definition of positive and negative training samples has a substantial impact on object detection. Based on this theoretical discovery, we propose a new framework for defining positive and negative training samples based on RetinaNet [5]. An overview of the framework is shown in Fig. 1. Our proposed framework includes two modules, a judgment module and an optimization module. The principle of the judgment module is to determine the threshold of the intersection over union(IoU) according to the candidate box information from the five layers of the different scales output by RetinaNet, instead of using the prior knowledge to determine the fixed threshold. Its role is to differentiate the training samples adaptively, that is, to automatically distinguish between positive and negative training samples based on the statistical characteristics of the samples themselves. The idea of the optimization module is to use the reinforcement learning framework [18], [35] to deform the negative sample candidate box to reach the standard of positive samples. Its purpose is to increase the proportion of positive samples within the sample.

A. JUDGMENT MODULE

The judgment module draws on the idea of the method in [17] and classifies the candidate boxes generated in RetinaNet [5] into positive and negative samples.Algorithm 1 describes how the judgment module works after an image is input into the model. For each object in the input image, all we have to do is collect the candidate boxes for this object. As described in Line 2,for L layers output by the Feature Pyramid Networks(FPN) [40] in RetinaNet, each layer generates several candidate boxes corresponding to the ground truth of an object. According to the L_2 distance between the center points of the candidate box and the ground-truth, we select k candidate boxes as training samples in Line 3. Therefore, the ground-truth of each object in the image will correspond to $L * k$ training samples.To distinguish the positive and negative of these training samples, we first calculate the IoU value between the candidate boxes and the ground-truth, then we calculate the mean and variance of these IoUs in Line 6, and finally we use the sum of these two parts as the threshold to judge the training sample. Then, we add constraints to exclude candidate boxes whose center is outside the ground-truth. Finally, we obtain positive training samples P and negative training samples N.

B. OPTIMIZATION MODULE

The optimization module operates the candidate box in the negative sample N, and utilizes the agent under the reinforcement learning framework to perform a series of deformation operations to reach the the positive sample standard.

Action. The agent deforms the negative sample candidate boxes through a series of actions to reach the positive sample standard. This series of actions includes *horizontal movement, vertical movement, zoom in and out, and stop*.

State. State s is composed of two parts, feature vector o and history vector h, in a tuple (o, h). Feature vector o represents the content of the observed area, and the starting position is the negative sample candidate box. History vector h represents the record of actions adopted by the agent.

Reward Function. Reward function r represents the feedback obtained after the agent takes action. This feedback can reflect the quantified distance between the observed areas, that is, the negative sample candidate box, which is deformed by the agent using action a and the ground-truth. The quantization distance used here is IoU, which is the relative position relationship between observed area b and ground-truth g. When the agent adopts action a, the negative sample candidate box is changed from initial b to b', that is, after the agent interacts with the environment, state s changes to s'. At this time, the center of the observed area is p, the center after deformation is p', and the center of the ground-truth is p_g. Therefore, the reward function is as follows:

$$r_a(s, s') = sign(IoU(b, g) - IoU(b', g)) + \lambda(d_2(p, p_g) - d_2(p', p_g)) \quad (1)$$

Algorithm 1 Reinforcement Training Sample Adaptation

Input: image I.
Output: positive samples P, negative samples N.

1 **Initialize:**
2 ground-truth boxes on the image $g \in G$
3 feature pyramid levels l
4 anchor boxes on the i_{th} pyramid level $A_i \in A$
5 candidate positive samples of the $g : C_g \leftarrow \emptyset$
6 action $a \in (Up,\ down,\ left,\ right,\ zoom\ out,\ zoom\ in,\ stop)$
7 **for** each $g \in G$ **do**
8 **for** each level $i \in [1, l]$ **do**
9 In A_i, select the k anchors that are closest to the center point of g based on the L_2 distance $\rightarrow S_i$
10 $C_g = C_g \cup S_i$
11 **end for**
12 IoU threshold for g: $T_g = mean\left(IoU(C_g, g)\right) + std(IoU\left(C_g, g\right))$
13 **for** each $c \in C_g$ **do**
14 **if** IoU$(c, g) \geq T_g$ and center of c in g **then**
15 $P = P \cup c$
16 **end if**
17 **end for**
18 **end for**
19 $N = A - P$
20 **for** each $c \in N$ **do**
21 agent adopts a series of a to act on c according to the feedback r until it satisfies Equation 2
22 $P = P \cup c$
23 **end for**

The feedback of the reward function is divided into two parts, one is the difference in the IoU change between the observed area and the ground-truth, and the other is the difference in the change in L_2 distance between the center of the observed area and the ground-truth center, where λ, as the coefficient, balances the two parts. Relying on the feedback of the reward function, the agent can judge the pros and cons of the action in a certain state, and trigger the stop action when the target state is reached. Here, we set the target state as follows:

$$r_t(s, s') = \begin{cases} 1, & if\ IoU(b, g) \geq t_g and\ p\ in\ GT \\ -1, & otherwise \end{cases} \quad (2)$$

$IoU(b, g) \geq T_g$ shows that the condition that triggers the agent to adopt the termination action needs to satisfy IoU threshold T_g in the determination module, and it also needs to simultaneously satisfy the condition in which the center point of the deformed candidate box is in the ground-truth(GT).

Inspired by the Ape-X architecture in the distributed reinforcement learning method [39], we divide the agents in the optimization module into two categories, exploratory agent (A_e) and development agent (A_d). Here, the role of an exploratory agent is to participate in the operation of the

candidate boxes deformation and input the obtained experience feedback into public experience pool B, while a development agent directly uses these experiences to update its priorities. We set the exploratory agent to regularly update itself with the latest network parameters from the development agent, and their numbers are divided into 6 and 2. We use the Q function, which is based on the Bellman equation, to evaluate the performance of the agents in the optimization module. The optimal value in the Q function is Q^*, and its formula is as follows:

$$Q^*(s,a) = E_{s'}[R(s,a) + \gamma \max_{a'} Q^*(s',a')|s,a] \quad (3)$$

In Equation 3, s represents the state and a represents the action. According to the deep reinforcement learning algorithm [18], we minimize the loss function through the *i-th* iteration to learn the Q function of the candidate action. The formula is as follows:

$$L_i = (R(s,a) + \gamma \max_{a'} Q^*(s',a';\theta_i^-) - Q(s,a;\theta_i))^2 \quad (4)$$

In Equation 4, $R(s,a)$ represents reward r obtained by the agent after deforming the candidate box in the s state. Six exploration agents participate in the deformation operation of the candidate boxes simultaneously and add the interactive results $(s, a, R(s,a))$ to the experience pool B. The two development agents further utilize the experience in the processing pool and share them with the exploration agent. This kind of distributed reinforcement learning structure with multiagent division of labor and coordination ensures a more efficient and concurrent execution of the algorithm, which overcomes the shortcomings of the long training time and low efficiency of the reinforcement learning algorithm.

C. DISCUSSION

Our work is inspired by ATSS and object detection methods based on reinforcement learning.In this section, we compare the differences between them and our work. (1) The judgment module in our work draws from the solution method in ATSS,which automatically determines the IoU threshold according to the anchor's own attributes as the basis for the agent to take action in the reinforcement learning process in the optimization module. In contrast to ATSS, which only automatically adjusts the threshold based on anchor attributes, our work also uses reinforcement learning control strategies to change the anchor shape to increase the IoU value between it and the ground-truth.For object detection methods, anchors with higher IoU values indicate that the candidate observation area is closer to the ground- truth of the detected object, which can lead to higher quality detection. (2) Our work draws on the object detection method based on reinforcement learning, but unlike previous methods, we use a distributed reinforcement learning architecture. The single agent-based reinforcement learning algorithm, DQN, requires training time and does not converge easily due to its long and unstable nature, which affects its scalability and practicality as an object detection method.In response to these problems, we propose an object detection idea based on distributed reinforcement learning architecture. Unlike previous solutions based on a single agent, we use multiagents (A_e and A_d) with a different division of labor in our work. A_e is responsible for changing the anchor shape, generating a series of IoU values between the candidate observation area and the ground- truth of the detected object, and then adding it to experience pool B. A_d is responsible for extracting the operations from the experience pool that have a significant improvement effect on verification, updating the control strategy parameters, and sharing them with A_e. The concurrent execution of multiagents in a distributed architecture greatly improves the efficiency of reinforcement learning, reduces training time and is more stable.

IV. EXPERIMENTS

A. EXPERIMENTAL SETUP AND IMPLEMENTATION

We perform experiments on the m2cai16-tool-locations dataset [30] and the private AJU-Set dataset. Both datasets are composed of images extracted from video frames, which contain seven different types of surgical tools. Since there have been few previous studies on surgical tools, which results in limited available data, we have adopted a variety of data enhancement tricks [41] to expand the dataset. All models in this article have been trained on two NVIDIA Geforce RTX 2070 GPUs and Core i7 9700k CPU, and our solution can achieve real-time processing speed, thus providing excellent recognition performance.

B. DATASETS

1) m2cai16-TOOL-LOCATIONS

The m2cai16-tool-locations dataset comes from the surgeon's expansion of the spatial location information of surgical tools in m2cai16-tool [29]. The m2cai16-tool dataset, which contains 15 laparoscopic surgery videos recorded at 25 fps, obtains 12541 test samples and 23287 training samples through label processing at 1 frame per second. For the task of automatically evaluating the use of surgical tools, m2cai16-tool, which only contains information on whether surgical tools are stored, is not sufficient.To meet the needs of the task, with the assistance of a surgeon, we selected 2532 frames from the m2cai16-tool dataset for spatial information annotation. Following the classic partitioning strategy, we divide the data set into a training set, a test set, and a validation set according to the proportions of 50%, 30%, and 20%, respectively.

2) AJU-SET

Considering the complex background of the surgical environment, a single dataset may reduce the performance of the object detection model. To solve this problem, we obtained 20 laparoscopic cholecystectomy surgery videos with the assistance of the Second Hospital of Jilin University to form a new dataset, AJU-Set, which is shown in Fig. 2. AJU-Set, which has the same recording rate and label rate as

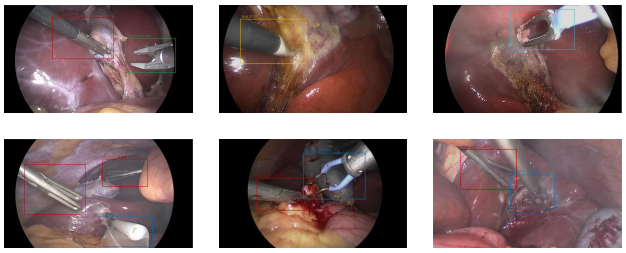

FIGURE 2. The detection situation for surgical tools in different actual scenes.

TABLE 1. Number of annotations for seven surgical tools in two databases.

Tool	m2cai16-tool-locations	AJU-Set
Grasper	923	882
Bipolar	350	483
Hook	308	607
Scissors	400	532
Clipper	400	554
Irrigator	485	480
Specimen Bag	275	412
Total	3141	3952
Number of Frames	2532	3164

m2cai16-tool-locations, contains 3164 labeled frames, and maintains the same dataset division ratio. With the assistance of a professional surgeon, all surgical tools in the video have been accurately labeled.

The above two datasets cover seven types of surgical tools such as grippers, bipolars, hooks, scissors, clippers, irrigators, and specimen bags. The related category number distribution information and some sample displays are shown in Table 1 and Fig. 3.

C. BASELINE METHODS

To verify the effectiveness of our method, we set ATSS [17] as a benchmark in the experiment, which is shown in Table 3. For the first time, we adopt an object detection method based on reinforcement learning for sample adaptation in the detection and analysis of surgical tools. The two-stage object detection method has a higher accuracy than the one-stage object detection method because the candidate box generation stage mines richer object context information.To verify the effectiveness of the sample adaptive method, we apply the method proposed by ATSS and our own method to the one-stage object detection method and compared it with two-stage object detection. The comparison results show that the one-stage object detection model under the optimization of the sample adaptive method maintains the previous speed and has higher accuracy than the two-stage object detection model. For the ATSS, which is also sample adaptive, and the method we proposed, in terms of the effect of applying the one-stage object detection model, the method we proposed has better results than ATSS.

Thanks to the better performance of the object detection model, we can use this as a basis for identifying positioning and trajectory tracking of surgical tools, thus laying the foundation for the analysis of surgical behavior quality. As shown in Table 2, we can observe the detection of surgical tools optimized by our sample adaptive method with respect to the surgical tools in the two datasets. From the table, we can see that the two surgical tools with higher detection accuracy are the clipper and hook. This might be because these two tools require a better angle to operate during surgery. In addition, from the table, we can observe that the detection accuracy for the two surgical tools, bipolar and irrigator, is lower. This could be attributed to the poor observation angle due to the difference in function and the more complicated background of surgery at this stage.

D. ABLATION STUDY

To verify the effectiveness of the modules in our proposed sample adaptation method, we set up several method variants to verify on the datasets. For a fair comparison, we compared different variants under the same conditions. Table 4, includes only the judgment module, the judgment module joint optimization module, and the multiple iteration judgment module joint optimization module. We use JM to represent the judgment module. The judgment module determines the threshold for judging the positive and negative training samples according to the mean and variance calculation of the IoU value between the candidate box and ground-truth in each feature layer. Correspondingly, OM represents the optimization module. The optimization module is based on the optimization control of the detection behavior under the reinforcement learning framework. Its purpose is to use the reinforcement learning agent to deform the candidate frame of the negative training sample to reach the positive sample standard. From Table 4, we can see that when only the JM is used, the performance of the object detection method is not as good as that when the JM and OM are used in combination. In the case where the number of iterations is $n \geq 20$, the object detection method has better performance, which leads to a decrease in overall framework performance due to a substantial increase in computing resource consumption and a decrease in speed.

E. SURGICAL SKILLS ASSESSMENT

In the following, by applying the proposed framework to the two datasets, we analyze the spatial and temporal information of surgical tools to evaluate the skill level of surgeons. To achieve this process, we propose an object detection method based on the framework of reinforcement learning in order to control sample adaptation to automatically detect and evaluate the use status of surgical tools, and complete the evaluation of the surgical process by using the surgical tools' usage patterns, motion trajectories, and mobile values.

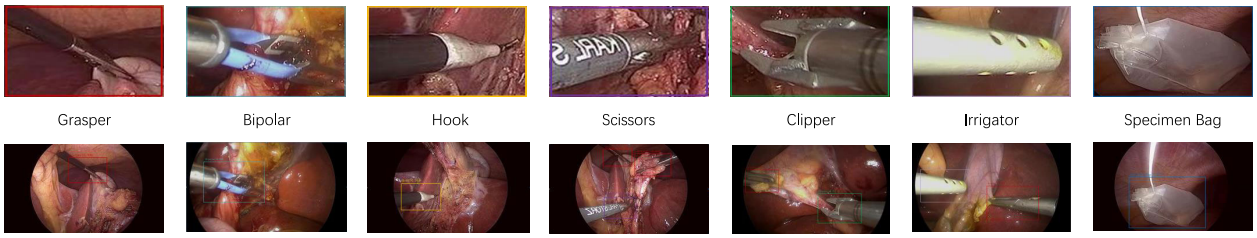

FIGURE 3. Top: Seven surgical tools are wrapped by different colored bounding boxes. **Bottom:** Recognition and positioning of surgical tools in different samples.

TABLE 2. The performance of the object detection method based on the reinforcement training sample adaptation framework on the m2cai16-tool-locations and AJU-Set datasets.

Dataset	Grasper	Bipolar	Hook	Scissors	Clipper	Irrigator	Specimen Bag	mAP
m2cai16-tool-locations	54.7	69.9	87.3	74.4	84.7	42.1	77.6	70.1
AJU-Set	74.6	65.9	93.9	76.6	92.4	55.8	81.9	77.3

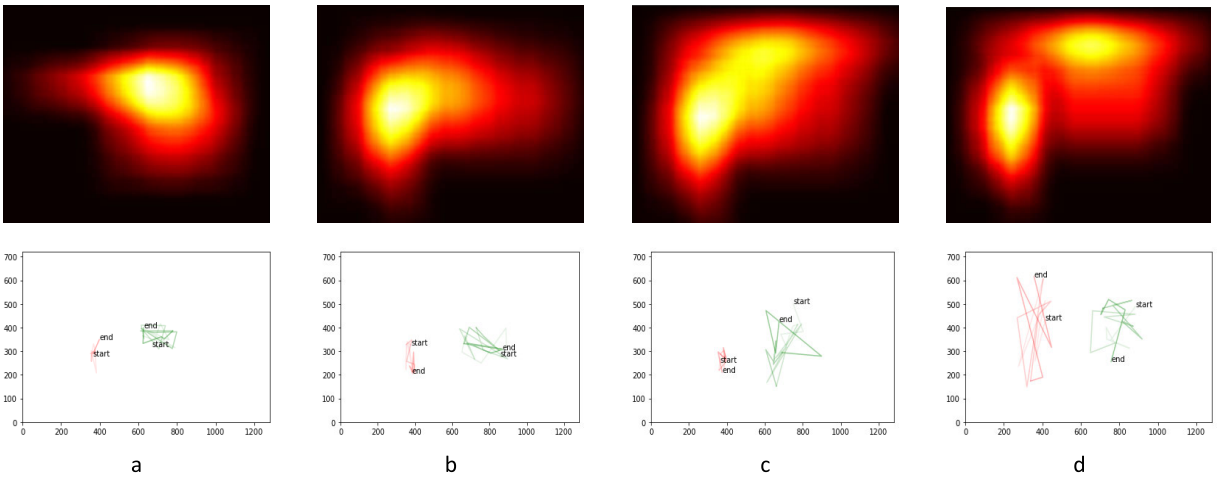

FIGURE 4. a-d correspond to video 1-video 4. The top row: the position of the surgical tool in the image generates a heat map illustrating the distribution range of the tool. The bottom row: the movement trajectories of a clipper and grasper during the shearing stage. The level of surgical skills can be intuitively reflected through the figure.

We extracted four test videos from the AJU-Set to evaluate the surgical skill level. As shown at the top of Fig. 4, we generated a heat map to represent the range of motion of surgical tools by detecting the bounding box derived from the surgical tools. Through medical practice and experience, high-level surgical operations are carried out frequently and accurately in a specific area, showing the higher mobile value of the operation. From the observations in Fig. 4, it can be seen that heat map a corresponding to video 1 has the smallest range, reflecting the doctor's proficient skills and level during surgery.

The separation of the triangle of the gallbladder is a critical operation in cholecystectomy, and in related to biliary tract injuries and complications. Since the surgical operation at this stage performs subtle operations in a short time, we chose to observe and study the two key surgical tools, clipper and grasper. Through the information shown at the bottom of Fig. 4, we can observe that the movement trajectories of the clipper and grasper in video 1 and video 2 converge in a specific range, which shows that the two surgical tools show good cooperation during the operation. Correspondingly, the estimated range of movement for the two surgical tools in video 3 and video 4 is larger and irregular, showing that the surgical skills are not proficient.

In addition to the heat map and motion trajectory chart mentioned above, we also counted the usage time of the surgical tools to analyze the doctor's skill proficiency during the surgery. In the histogram in Fig. 5, we can observe that the bipolar images of video 3 and video 4 have been used longer, which indicates that more hemostasis operations are required during the surgery and that the surgical skills are not proficient.

TABLE 3. The performance of the detection methods on the m2cai16-tool-locations and AJU-Set datasets, boldface represents the best result.

Method	Backbone	m2cai16-tool-locations		AJU-Set	
		mAP (%)	FPS	mAP (%)	FPS
anchor-based two-stage:					
R-CNN [6]	ResNet-101	37.1	0.04	40.9	0.03
Faster R-CNN [8]	ResNet-101	62.3	11.5	65.5	11.2
C-Mask RCNN [9]	ResNet-101	63.8	18.4	70.4	18.1
Cascade R-CNN [10]	ResNet-101	65.1	23.1	71.9	22.6
Revisiting RCNN [42]	ResNet-101+ResNet-152	65.5	20.8	72.4	20.3
SNIP [43]	DPN-98	69.5	18.7	76.7	18.3
anchor-based one-stage:					
DSSD513 [44]	ResNet-101	50.5	13.8	55.8	13.4
RefineDet512 [33]	ResNet-101	55.4	25.4	61.1	24.6
RetinaNet [5]	ResNet-101	59.5	**32.3**	65.7	**31.3**
anchor-free center-based:					
GA-RPN [45]	ResNet-50	60.5	20.7	66.8	20.3
FoveaBox [14]	ResNeXt-101	63.9	25.8	70.7	25.3
FSAF [16]	ResNeXt-64x4d-101	65.2	9.7	72.1	9.5
FCOS [15]	ResNeXt-64x4d-101	65.7	12.4	72.5	12.1
with ATSS:					
DSSD513 + ATSS	ResNet-101	67.7	13.8	74.7	13.2
RefineDet512 +ATSS	ResNet-101	67.8	25.2	74.9	24.6
RetinaNet + ATSS	ResNet-101	68.1	32.1	74.8	31.2
GA-RPN + ATSS	ResNet-50	67.9	20.4	74.7	20.3
FoveaBox + ATSS	ResNeXt-101	67.8	25.8	74.9	25.1
FSAF + ATSS	ResNeXt-64x4d-101	67.6	9.7	74.7	9.3
FCOS + ATSS	ResNeXt-64x4d-101	68.1	12.4	75.1	12.1
with Ours:					
DSSD513 + Ours	ResNet-101	69.7	12.4	76.9	12.1
RefineDet512 + Ours	ResNet-101	69.9	22.8	76.8	22.1
RetinaNet + Ours	ResNet-101	**70.1**	29.1	**77.3**	28.2
GA-RPN + Ours	ResNet-50	69.7	18.6	76.8	18.2
FoveaBox + Ours	ResNeXt-101	69.8	23.2	77.2	22.7
FSAF + Ours	ResNeXt-64x4d-101	69.9	8.7	77.1	8.5
FCOS + Ours	ResNeXt-64x4d-101	**70.1**	11.1	**77.3**	10.9

TABLE 4. Comparison of the ablation study of our proposed framework on the m2cai16-tool-locations and AJU-Set datasets.

Method	Backbone	mAP (%) (m2cai16-tool-locations)	mAP (%) (AJU-Set)
RetinaNet + JM	ResNet-101	68.2	74.6
RetinaNet + JM+OM	ResNet-101	70.1	77.3
RetinaNet +n(JM+OM)	ResNet-101	**70.7**	**78.1**

To prove the effectiveness of our proposed framework, we invited four surgical experts to conduct an evaluation. They agreed that the surgical skills demonstrated in video 1 are the best, and that both video 1 and video 2 have better surgical skills than videos 3 and 4, which confirms the effectiveness of our evaluation method.

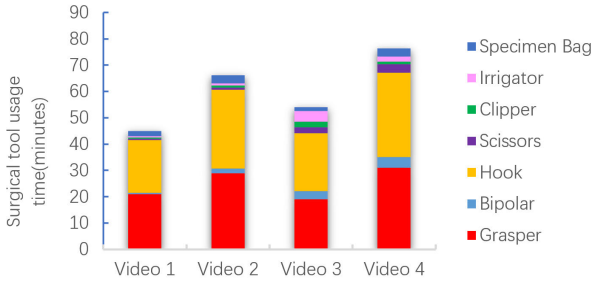

FIGURE 5. **The usage times of different surgical tools in the four videos reflect the level and quality of surgical skills.**

V. CONCLUSION

In this paper, we propose a novel framework for one-stage object detection based on a sample adaptive process controlled by reinforcement learning. Different from the traditional method of choosing fixed thresholds for defining strategies, our adaptive framework sets the thresholds according to the statistical characteristics of the samples themselves. In addition, our proposed method also uses flexible control of the reinforcement learning framework to optimize the negative sample candidate boxes to increase the proportion of positive training samples, and thus improves the accuracy of the object detection model for object detection. For the m2cai16-tool-locations and AJU-Set datasets with fewer training samples for surgical instrument detection, our sample adaptive method allows the one-stage object detection algorithms to perform better than the two-stage object detection while maintaining high speed. Accurate surgical instrument detection is helpful in analyzing the operation behavior pattern, movement trajectory and movement value of a instrument during the surgical operation process, and provides powerful assistance in summarizing and improving a doctor's surgical skills and professional communication. For future work, we hope that we can continue to improve the accuracy and real-time nature of the object detection model to achieve the function of on-site online learning and assisted guidance of surgery.

REFERENCES

[1] B. C. Alkire, N. P. Raykar, M. G. Shrime, T. G. Weiser, S. W. Bickler, J. A. Rose, C. T. Nutt, S. L. M. Greenberg, M. Kotagal, J. N. Riesel, M. Esquivel, T. Uribe-Leitz, G. Molina, N. Roy, J. G. Meara, and P. E. Farmer, "Global access to surgical care: A modelling study," *Lancet Global Health*, vol. 3, no. 6, pp. e316–e323, Jun. 2015.

[2] M. A. Healey, S. R. Shackford, T. M. Osler, F. B. Rogers, and E. Burns, "Complications in surgical patients," *Arch. Surg.*, vol. 137, no. 5, pp. 611–618, May 2002.

[3] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 7263–7271.

[4] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis.*, Springer, 2016, pp. 21–37.

[5] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2980–2988.

[6] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 580–587.

[7] J. Dai, Y. Li, K. He, and J. Sun, "R-fcn: Object detection via region-based fully convolutional networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 379–387.

[8] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 91–99.

[9] Z. Chen, S. Huang, and D. Tao, "Context refinement for object detection," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 71–86.

[10] Z. Cai and N. Vasconcelos, "Cascade R-CNN: Delving into high quality object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6154–6162.

[11] H. Law and J. Deng, "CornerNet: Detecting objects as paired keypoints," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 734–750.

[12] K. Duan, S. Bai, L. Xie, H. Qi, Q. Huang, and Q. Tian, "CenterNet: Keypoint triplets for object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6569–6578.

[13] X. Zhou, J. Zhuo, and P. Krahenbuhl, "Bottom-up object detection by grouping extreme and center points," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 850–859.

[14] T. Kong, F. Sun, H. Liu, Y. Jiang, L. Li, and J. Shi, "FoveaBox: Beyound anchor-based object detection," *IEEE Trans. Image Process.*, vol. 29, pp. 7389–7398, 2020.

[15] Z. Tian, C. Shen, H. Chen, and T. He, "FCOS: Fully convolutional one-stage object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9627–9636.

[16] C. Zhu, Y. He, and M. Savvides, "Feature selective anchor-free module for single-shot object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 840–849.

[17] S. Zhang, C. Chi, Y. Yao, Z. Lei, and S. Z. Li, "Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9759–9768.

[18] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, and D. Hassabis, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, Feb. 2015.

[19] B. Zhang, S. Wang, L. Dong, and P. Chen, "Surgical tools detection based on modulated anchoring network in laparoscopic videos," *IEEE Access*, vol. 8, pp. 23748–23758, 2020.

[20] A. Agustinos and S. Voros, "2d/3d real-time tracking of surgical instruments based on endoscopic image processing," in *Computer-Assisted Robotic Endoscopy*. Springer, 2015, pp. 90–100.

[21] F. Lalys, L. Riffaud, D. Bouget, and P. Jannin, "A framework for the recognition of high-level surgical tasks from video images for cataract surgeries," *IEEE Trans. Biomed. Eng.*, vol. 59, no. 4, pp. 966–976, Apr. 2012.

[22] M. J. Primus, K. Schoeffmann, and L. Boszormenyi, "Temporal segmentation of laparoscopic videos into surgical phases," in *Proc. 14th Int. Workshop Content-Based Multimedia Indexing (CBMI)*, Jun. 2016, pp. 1–6.

[23] L. Tao, L. Zappella, G. D. Hager, and R. Vidal, "Surgical gesture segmentation and recognition," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent*. Springer, 2013, pp. 339–346.

[24] K. Charrière, G. Quellec, M. Lamard, D. Martiano, G. Cazuguel, G. Coatrieux, and B. Cochener, "Real-time analysis of cataract surgery videos using statistical models," *Multimedia Tools Appl.*, vol. 76, no. 21, pp. 22473–22491, Nov. 2017.

[25] H. Lin, I. Shafran, D. Yuh, and G. Hager, "Towards automatic skill evaluation: Detection and segmentation of robot-assisted surgical motions," *Comput. Aided Surg.*, vol. 11, no. 5, pp. 220–230, Sep. 2006.

[26] N. Padoy, T. Blum, H. Feussner, M. Berger, and N. Navab, "On-line recognition of surgical activity for monitoring in the operating room," in *Proc. AAAI*, 2008, pp. 1718–1724.

[27] M. Sahu, A. Mukhopadhyay, A. Szengel, and S. Zachow, "Tool and phase recognition using contextual CNN features," 2016, *arXiv:1610.08854*. [Online]. Available: http://arxiv.org/abs/1610.08854

[28] A. P. Twinanda, D. Mutter, J. Marescaux, M. de Mathelin, and N. Padoy, "Single- and multi-task architectures for tool presence detection challenge at M2CAI 2016," 2016, *arXiv:1610.08851*. [Online]. Available: http://arxiv.org/abs/1610.08851

[29] A. P. Twinanda, S. Shehata, D. Mutter, J. Marescaux, M. de Mathelin, and N. Padoy, "EndoNet: A deep architecture for recognition tasks on laparoscopic videos," *IEEE Trans. Med. Imag.*, vol. 36, no. 1, pp. 86–97, Jan. 2017.
[30] A. Jin, S. Yeung, J. Jopling, J. Krause, D. Azagury, A. Milstein, and L. Fei-Fei, "Tool detection and operative skill assessment in surgical videos using region-based convolutional neural networks," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2018, pp. 691–699.
[31] P. Zhou, B. Ni, C. Geng, J. Hu, and Y. Xu, "Scale-transferrable object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 528–537.
[32] Z. Shen, Z. Liu, J. Li, Y.-G. Jiang, Y. Chen, and X. Xue, "DSOD: Learning deeply supervised object detectors from scratch," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1919–1927.
[33] S. Zhang, L. Wen, X. Bian, Z. Lei, and S. Z. Li, "Single-shot refinement neural network for object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4203–4212.
[34] J. Nie, R. M. Anwer, H. Cholakkal, F. S. Khan, Y. Pang, and L. Shao, "Enriched feature guided refinement network for object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9537–9546.
[35] J. C. Caicedo and S. Lazebnik, "Active object localization with deep reinforcement learning," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 2488–2496.
[36] S. Yun, J. Choi, Y. Yoo, K. Yun, and J. Y. Choi, "Action-decision networks for visual tracking with deep reinforcement learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2711–2720.
[37] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. van den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, S. Dieleman, D. Grewe, J. Nham, N. Kalchbrenner, I. Sutskever, T. Lillicrap, M. Leach, K. Kavukcuoglu, T. Graepel, and D. Hassabis, "Mastering the game of go with deep neural networks and tree search," *Nature*, vol. 529, no. 7587, pp. 484–489, Jan. 2016.
[38] O. Vinyals *et al.*, "Grandmaster level in StarCraft II using multi-agent reinforcement learning," *Nature*, vol. 575, no. 7782, pp. 350–354, Nov. 2019.
[39] D. Horgan, J. Quan, D. Budden, G. Barth-Maron, M. Hessel, H. van Hasselt, and D. Silver, "Distributed prioritized experience replay," 2018, *arXiv:1803.00933*. [Online]. Available: http://arxiv.org/abs/1803.00933
[40] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2117–2125.
[41] E. D. Cubuk, B. Zoph, J. Shlens, and Q. V. Le, "Randaugment: Practical automated data augmentation with a reduced search space," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2020, pp. 702–703.
[42] B. Cheng, Y. Wei, H. Shi, R. Feris, J. Xiong, and T. Huang, "Revisiting RCNN: On awakening the classification power of faster RCNN," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 453–468.
[43] B. Singh and L. S. Davis, "An analysis of scale invariance in object detection–SNIP," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3578–3587.
[44] C.-Y. Fu, W. Liu, A. Ranga, A. Tyagi, and A. C. Berg, "DSSD : Deconvolutional single shot detector," 2017, *arXiv:1701.06659*. [Online]. Available: http://arxiv.org/abs/1701.06659
[45] J. Wang, K. Chen, S. Yang, C. C. Loy, and D. Lin, "Region proposal by guided anchoring," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2965–2974.

GUANGYAO WANG received the B.S. degree from the College of Computer Science and Technology, Changchun University of Science and Technology, in 2009, and the M.S. degree from the College of Computer Science and Technology, Jilin University, in 2013, where he is currently pursuing the Ph.D. degree. His current research interests include deep learning, reinforcement learning, and image processing.

SHENGSHENG WANG (Member, IEEE) received the B.S., M.S., and Ph.D. degrees in computer science from Jilin University, in 1997, 2000, and 2003, respectively. He is currently a Professor with the College of Computer Science and Technology, Jilin University. His current research interests include the areas of computer vision, deep learning, and data mining.