IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received January 6, 2020, accepted January 23, 2020, date of publication January 28, 2020, date of current version February 6, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2969885

Surgical Tools Detection Based on Modulated Anchoring Network in Laparoscopic Videos

BEIBEI ZHANG[1,2], SHENGSHENG WANG[1,2], LIYAN DONG[1,2], AND PENG CHEN[3]

[1]College of Computer Science and Technology, Jilin University, Changchun 130012, China
[2]Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University, Changchun 130012, China
[3]The Second Hospital of Jilin University, Changchun 130041, China

Corresponding author: Liyan Dong (dongly@jlu.edu.cn)

This work was supported in part by the Science and 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 Minimally invasive surgery like laparoscopic surgery is an active research area of clinical practice for less pain and a faster recovery rate. Detection of surgical tools with more accurate spatial locations in surgical videos not only helps to ensure patient safety by reducing the incidence of complications but also makes a difference to assess the surgeon performance. In this paper, we propose a novel Modulated Anchoring Network for detection of laparoscopic surgery tools based on Faster R-CNN, which inherits the merits of two-stage approaches while also maintains high efficiency of comparable speed as state-of-the-art one-stage methods. Since objects like surgical instruments with a wide aspect ratio are difficult to recognize, we develop a novel training scheme named as modulated anchoring to explicitly predict arbitrary anchor shapes of objects of interest. For taking the relationship of different tools into consideration, it is useful to embed the relation module in our network. We evaluate our method using an existing dataset (m2cai16-tool-locations) and a new private dataset (AJU-Set), both collected from cholecystectomy surgical videos in hospital, covering information of seven surgical tools with spatial bounds. We show that our detector yields excellent detection accuracy of 69.6% and 76.5% over the introduced datasets superior to other recently used architectures. We further verify the efficiency of our method by analyzing the usage patterns of tools, the economy of the movement, and the dexterity of operations to assess surgical quality.

INDEX TERMS Laparoscopic surgery, tool detection, convolutional neural network, operational quality assessment.

I. INTRODUCTION

Vision-based laparoscopic surgery, a representative minimally invasive surgery, has attracted increasing attention during the deep learning era. Unlike conventional incision procedure, it is performed through a small hole (incision) using a variety of surgical tools with the help of the endoscopic camera, and especially, the challenge of ablation and suturing with tissue corresponds to specialized surgical instruments in medical processes. Moreover, lacking personalized, objective feedback on surgical skills and quality stands out as one of the crucial problems behind laparoscopic surgery for surgeons.

To address this problem, analysis of the surgical video recorded by an endoscopic camera has been gradually developed in recent years, which could assess operative skills efficiently and objectively as well as is significant for other potential applications in future clinical practice. However, traditional assessment of surgeon video is manual with great time-consume and effort of experts. Thus, we propose a detection network to locate surgical tools accurately for automated surgical video analysis in real-time to assess performance faster and better. The assessment contains some relevant tasks, such as analysis of tool usage patterns, the economy of the movement, and the motion scope, which could substantiate our method remarkable and impressive for surgical tools detection.

Present works of frame-level surgical instrument detection in the video mostly come from the 2016 M2CAI Tool Presence Detection Challenge (a satellite event of MICCAI 2016 in Athens), providing a dataset called m2cai16-tool

The associate editor coordinating the review of this manuscript and approving it for publication was Hugo Proenca.

with binary annotations [1]. The set can only be used to detect whether surgical tools exist or not in videos, but not enough to get their specific locations. Therefore, in our work datasets with object coordinate annotations are needed for the instrument detection tasks. To our best knowledge, m2cai16-tool-locations dataset [2], extending the m2cai16-tool dataset, is the only public dataset of medical tools with spatial bounds. In addition, we propose a new dataset collected from several private videos of laparoscopic surgery performed at The Second Hospital of Jilin University, which contains locations of instruments as well and can provide a choice to verify effectiveness of our network architecture more comprehensively.

The current object detectors can be divided into two categories: one-stage and two-stage. The former does not have the stage of generating region proposals and produces the class probability as well as object coordinates directly. The latter first generates candidate boxes and then performs classification and regression. No matter which method we choose, one-stage detector or two-stage detector, anchors are all regard as core issues with great importance for object detection. Most existing schemes use the sliding window to generate anchors (i.e., fixed reference boxes), which predefine anchors with fixed scales and aspect ratios consistently for every spatial location. However, different from sparse schemes generating a small number of sparsely distributed boxes, the above dense anchoring schemes producing lots of anchors do not show outstanding performance in practice with a few defects, especially for targets of arbitrary shapes. First, it is hard to find a suitable set of predefined aspect ratios to match objects from different datasets, which are unevenly distributed on the image. Meanwhile, the design of this work depends to a large extent on the ability of the researchers themselves, which makes it easy to cause errors. Second, a dense anchoring scheme always corresponds to too many anchors distributed in the background area, and at the cost of time and efficiency. Third, recognition of objects with a large disparity in length and width, such as surgical instruments, has great obstacles for the usual anchor-based methods.

Therefore, to solve the difficulties above, we develop a sparse anchoring scheme, called modulated anchoring, which is a well-behaved method for the detection of laparoscopic surgical instruments. Both motivated by the idea of unevenly distributed anchors, the most similar approach to our work is Guided Anchoring Region Proposal Network (GA-RPN) [3] that decouples the process of generating anchors into two phases, including location prediction and shape prediction. However, there is a significant difference between them that our method presents a new feature adaption module, namely, modulated feature module. Getting the idea from the success of recent deformable network [4] about concentrating more on areas of interest, our module not only learns offsets during the process of convolutional deformation but also adds a modulation term to adjust the feature amplitude, performed to adapt feature for complying with the consistent anchor design guideline. In addition, considering the cooperative use among some surgical instruments, relation modules are inserted into our network. Different from original work [5], our relation block extends the field to laparoscopic surgery, taking tools relationship into account, and it turns out that the accuracy can be further improved. In principle, we construct a Modulated Anchoring Network, exploiting semantic features to perform tool detection. The related experimental results could prove that the proposed framework is outstanding in practice and the real-time requirements can also be well satisfied.

The main contributions of our work are as follows. (1) We formulate a novel framework Modulated Anchoring Network, leveraging semantic features to detect non-uniformly distributed surgical tools of arbitrary anchor shapes effectively and efficiently. (2) We propose a modulated feature module that covers the modulation mechanism to enhance modeling capacity incorporating the anchor shape information into the feature map. In addition, to take cooperative use of several surgical tools into consideration, the relation module is embedded in our existing network. (3) We present a new dataset AJU-Set with spatial locations of instruments, for laparoscopic surgery video understanding. (4) The experimental results demonstrate that our framework achieves remarkable performance in the instrument detection tasks. Furthermore, accurate recognition contributes to post-surgical quality assessment.

II. RELATED WORK

A. LAPAROSCOPIC SURGERY

With the widespread use of devices to record surgical procedures in minimally invasive surgery, automated analysis of surgical tools in videos has become a popular research area, mainly involving classification, segmentation, tracking, detection, and other directions. Unlike the earlier methods [6]–[11], rely on various handcrafted features, the existing approaches mainly use deep learning to extract more high-level features for surgical workflow recognition and tool detection. The traditional analysis of surgical phases is based on a number of statistical models, involving Conditional Random Fields [12]–[15], Hidden Markov Models [7], [16], [17], Hidden semi-Markov Models [18], [19], Linear Dynamical Systems [20] and so on. Recently several approaches [1], [21], [22], apply a structure of convolutional neural network (CNN) and recurrent neural network (RNN) to recognize surgical phases, which works effectively and becomes one of the mainstream structures in this field. Most present detection methods are frame-level tool presence detection from the M2CAI 2016 Tool Presence Detection Challenge. These methods [23]–[25], represented by the victory method [23] regard the task as the problem of image classification to judge the possibility of the existence of tools in the frame-level without exploiting temporal information. To take long-term relationships between continuous video frames into consideration, some works such as GCNs [26] rely on Graph Convolutional Networks to learn better features as well as achieve better performance.

Several previous strategies [27], [28] perform surgical tool localization in specific tasks of robot-assisted surgery videos. And most of the existing robot-assisted surgeries are specific surgical training tasks. But there are limitations to using robotic arms in practice because of their relatively high cost, and there may be differences between specific training tasks and complete surgery. Laparoscopic surgery mentioned elsewhere in the paper refers to complete surgery rather than surgical training tasks unless otherwise specified. In fact, what really locates the surgical tools in complete laparoscopic surgical videos using the CNN is the method [2], which proposes a new dataset containing spatial bounds of tools from full-length surgical operations. Additionally, the network only relies on Faster R-CNN [29] to detect instruments and evaluate surgical skills, inspiring us to improve the detection accuracy by combining and adjusting some modular structures built on the basic framework to locate instruments and assess surgical performance more efficiently.

B. SEQUENTIAL OBJECT DETECTORS

Nowadays, the classical object detection methods based on CNN include representative structures of one-stage [30], [31] and two-stage [29], [32], [33]. However, a common flaw in these methods is that there is a fixed receptive field, which is problematic in predicting multi-scale targets. Most of the state-of-the-art approaches focus on making the network structure have more abundant features, such as Scale-Transferrable Detection Network (STDN) [34], Single-Shot Refinement Neural Network (RefineDet) [35] and other methods to simulate the image pyramid, so as to better combine the semantic features of high and low layers. Nevertheless, a large number of experiments show that these methods are more suitable for the one-stage method, and tend to lead to the challenge of category imbalance with the problem that the candidate box size is fixed and cannot detect the object whose aspect ratio changes greatly. Now several methods [36]–[38] are moving towards the direction of improving the region proposal network (RPN). For example, Beyond Anchor-based Object Detector (FoveaBox) [37] directly predicts the probability that it belongs to a certain kind of object and the offset relative to a certain border for each point on the feature map. There exists difficulty of above anchor-free approaches that it is hard for them to deal with some complex situations within a single stage, lacking anchors and anchor-based refinement. Hence, we present a new anchoring scheme, modulated anchoring, to promote the performance of object detection.

C. RELATION MODELING

Since attention modules have been successfully applied to the field of natural language processing recently, driven by this tendency, more and more methods in the field of biological image have added these modules and obtained significant improvements. Before this, amounts of recent works employ attention mechanisms to perform sequence modeling. As one of the above, LSTM [39] incorporates contextual information into a detection network to model object relations, but like other algorithms, it remains exist constraint of sequential computation and complexity of training. Besides, attention mechanisms are also applied to human scenarios in a few cases, whereas with the cost of introducing additional annotations. Taking motivation from these prior works, Hu *et al*. [5] develop a relation module to model relations between objects with the advantages of plug and play and no additional supervision. So following this idea, relation modules are flexibly inserted into our architecture, aim to consider collaboration between surgical instruments. Moreover, since our network focuses on unedited and complete surgical operations in the laparoscopic surgery, there are bound to be things (e.g., changing anatomy, lens fogging, etc.) that block recognition of the surgical tools. To alleviate this situation, getting the idea from deformable convolution [4], [40], our work introduces a deformable module into the backbone network. Intuitively, this module is seen as a special attention mechanism to enhance the ability to focus on the relevant image areas, with which the integrated post-operative assessment will be facilitated in the long run.

III. DATASET

To the best of our knowledge, there are limited datasets of automated surgical instrument analysis in frame-level for public use. Particularly, most of existing datasets focus on the presence detection of surgical tools derived from the Challenges of Cholec80 [1] and M2CAI 2016 Tool Presence Detection about cholecystectomy surgeries, provided by the University Hospital of Strasbourg/IRCAD in France. The Cholec80 dataset contains 80 surgical videos, labeled with the phase (at 25 fps) and tool presence annotations (at 1 fps). Another more popular dataset, m2cai16-tool, including 15 videos of laparoscopic procedures with 23287 training samples and 12541 testing samples, which are recorded at 25 fps and labeled to 1 fps for processing. In fact, in addition to some proprietary robot-assisted surgical tool datasets [27], the only publicly available and comprehensive datasets we know having specific locations of the tools are m2cai16-tool-locations by Jin *et al*. [2], that utilizes the m2cai16-tool dataset with assistance of surgeon to generate 2532 frames labels containing coordinates of spatial bounds of instruments. Following the original partitioning strategy, we adopt this dataset in our work by dividing it into training set, test set, and validation set in proportion with 50%, 30%, and 20%. The detailed number of seven surgical instruments covered in the dataset and the actual samples are shown in Table 1 and top of Fig. 1 respectively.

Considering that the surgical process is affected by the complex external environment, it is possible that the observed results with only one dataset lack sufficient representativeness and comprehensiveness. For this reason, we gather and construct a new dataset AJU-Set with more annotations of spatial bounds of instruments, which is collected from 20 laparoscopic cholecystectomy surgeries videos at The Second Hospital of Jilin University. Additionally, this dataset

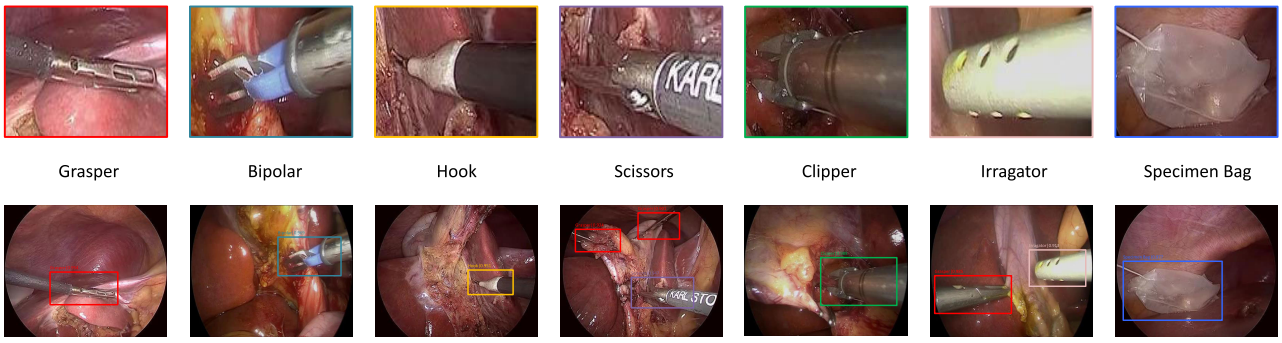


FIGURE 1. Top: different color bounding boxes correspond to seven surgical tools. Bottom: several samples of spatial detection results.

TABLE 1. Number of annotated instances for each tool on both datasets.

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

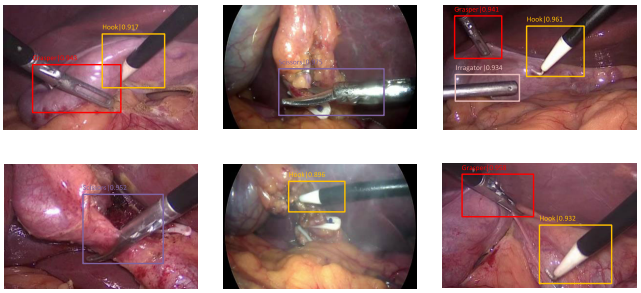


FIGURE 2. The detection results of some samples under the influence of complex external environment.

consists of 3164 labeled frames, having the same recording rate as well as labeling rate as m2cai16-tool-locations and dividing the set by the same scale. Each tool is annotated in video with the assistance of three professional surgeons for accuracy. In total, there are seven kinds of surgical tools covering grasper, bipolar, hook, scissors, clipper, irrigator, and specimen bag. The number of distributions in each category and some examples are separately shown in Table 1 and Fig. 2.

IV. METHODOLOGY

The traditional methods of generating anchors provide excessive region proposals with a fixed size and position, resulting in a large number of anchors containing background areas outside the target and unrealistically to find appropriate ratios of the anchor by manually predefined parameters to match objects of various sizes. In this paper, we develop a novel framework for recognition of laparoscopic surgery instruments named as Modulated Anchoring Network based on Faster R-CNN [29], which consists of a new anchoring scheme modulated anchoring and the relation module. Refer to the overall network structure shown in Fig. 3. The modulated anchoring network is formed by three parts, i.e., anchor location prediction branch, shape prediction branch, and modulated feature module. The anchor location prediction branch aims to yield a probability map to show where the object center may exist. The anchor shape prediction branch predicts the most likely shape of the object at the corresponding position based on the positional possibility of the center point mentioned above. Besides, to follow the consistency criterion, that is, feature of the anchor should match its shape, we propose a core component modulated feature module to integrate the shape information of the anchor into the feature map directly to meet our needs. Then we subtly embed the relation module in the existing network structure to perform joint reasoning on related objects. Below we will describe the overall network framework in detail.

A. ANCHOR LOCATION PREDICTION

Suppose that the detection box of the object on image I is represented as a 4-tuple of (x, y, w, h), where (x, y), w and h denote the center, width, and height of the anchor, respectively. Next, the anchor generation block for location and shape prediction can be viewed as the following conditional distribution:

$$p(x, y, w, h|I) = p(x, y|I)p(w, h|x, y, I) \tag{1}$$

It indicates that the anchor generation process is decoupled into two stages of location and shape prediction. After the possible position of the target is determined, the shape can get an appropriate result according to the location. Following Fig. 3, the feature map f_I with a size of $W \times H$ is generated through the backbone network and input to the modulated anchoring scheme. Then a probability map having the same size with the feature map and whose each entry's value $p\,(i, j|f_I)$ represents the possibility that the center of the object exists at the corresponding position is derived from the anchor location prediction branch. Specifically,

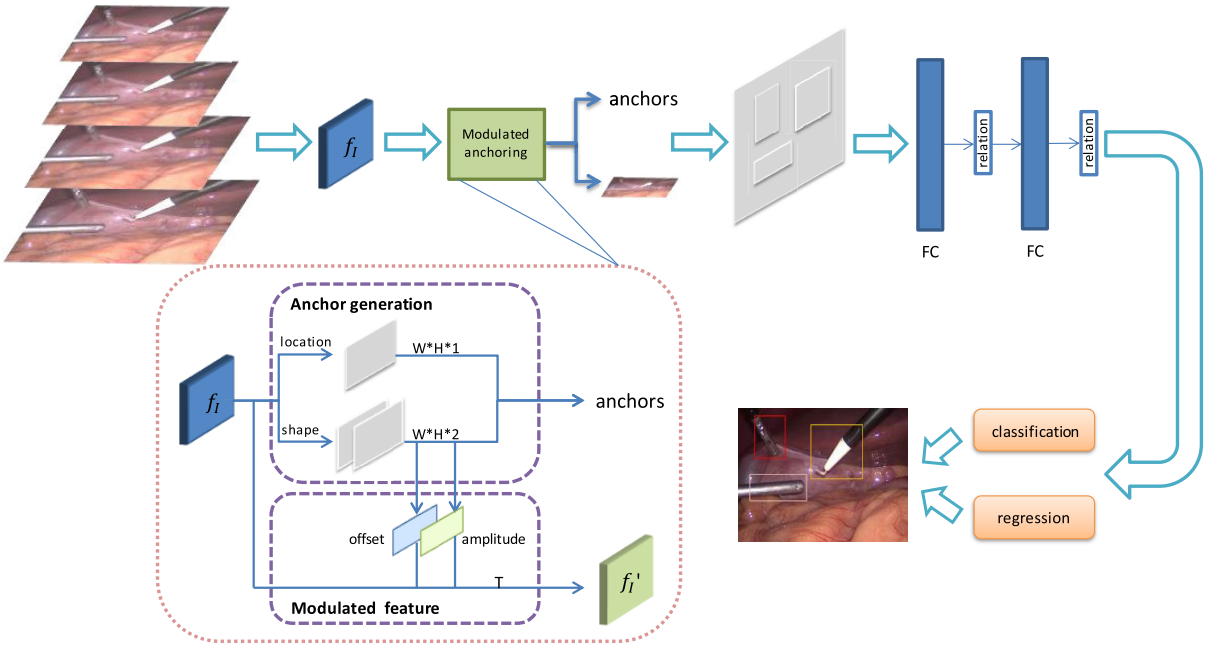


FIGURE 3. Architecture of Modulated Anchoring Network. The multi-level features generated by the feature pyramid are input into modulated anchoring, then the anchor generation module and the modulated feature module with a modulation mechanism output anchor location, shape, and adaptive feature to the subsequent network. Besides, the relation modules are inserted into the last two fully connected layers of the network to further enhance the detection effect.

through comprehensive consideration of speed and accuracy, we adopt a 1×1 convolution on the feature map f_I in this subnetwork and then convert it into a possible value by the sigmoid function. Further, according to the generated map of probability scores, the redundant candidates are filtered out by a predefined suitable threshold ε while ensuring that the recall remains stable to select an active region in which the object may exist for the next subnet.

B. ANCHOR SHAPE PREDICTION

Different from the traditional regression method, our anchor shape prediction sub-network adheres to the criterion of anchor center and feature alignment, predicting the optimal anchor shape (w, h) without changing the given position of object center getting from the position prediction branch. In order to obtain the optimal width and height of the box, there is a matter of fact that it is problematic and complicated to directly predict the values of w and h, for these two values vary widely. To that end, we present an intermediate mapping transformation to reduce the output range by referring to GA-RPN [3]. That is, the shape prediction branch generates the feature map of two channels applying the convolution of 1×1, which outputs the intermediate variables dw and dh respectively, and then maps to w and h through a transformation layer according to Eq.(2), where s represents the stride, and σ is the proportion coefficient with a value of 8 in our experiments.

$$w = \sigma \cdot s \cdot e^{dw}, \quad h = \sigma \cdot s \cdot e^{dh} \tag{2}$$

In total, the anchor shape prediction branch tends to work out the best w and h values by solving two tasks, including selecting the appropriate ground-truth bounding box that matches the anchor and maximizing the IoU between them to yield desired results. Concretely, we assume that the 4-tuple form of prediction box P_{wh} regarding shape and position is (x_0, y_0, w, h), and the 4-tuple corresponding to a ground truth bounding box G is $\left(x_g, y_g, w_g, h_g\right)$, then define the following formula for IoU.

$$\text{mIoU}\left(P_{wh}, G\right) = \max_{w>0, h>0} \text{IoU}\left(P_{wh}, G\right) \tag{3}$$

In theory, the IoU is calculated separately between the prediction boxes of various possible shapes corresponding to each position and all the ground truth bounding boxes, then it is natural to determine the maximum as $\text{mIoU}\left(P_{wh}, G\right)$. Whereas, whether it is traversing all likely locations as well as arbitrary shapes relative to anchor, or the repeated calculation of the formula about IoU is a challenging problem. To overcome this dilemma, practice indicates that efficiency and accuracy can be compromised by sampling several scales and aspect ratios to simulate enumerate arbitrary shapes of bounding boxes. In fact, in the experiment, we sampled 10 sets of data (w, h) to achieve a simplified and approximate effect.

It is worth noting that this method produces one box per position, which greatly reduces the number of anchors, and is suitable for the scene where the target is not dense in the image, such as the detection of surgical instruments. Furthermore, the experimental part also validates the superior efficiency of this shape prediction design.

C. MODULATED FEATURE MODULE

For the traditional anchor generation methods, different positions in the same layer of convolution correspond to the same scope of receptive field generating uniform anchors with a fixed scale and aspect ratio. However, our newly proposed anchor scheme provides non-uniform distribution boxes with learnable shapes, which inevitably induces inconsistencies between the variable anchors and features. In general, the position of feature corresponds to a large anchor with a large receptive field, and a small anchor should match a small region. Furthermore, such condition of violating the consistent anchor design principle may also plague subsequent classification and regression processes. Hence, we propose a modulated feature module to address the problem, that is, to directly integrate the shape information of the anchor into the feature, so that the newly obtained feature map can be adapted to the shape of anchor for each position, and the transformation of feature is computed as

$$F_i' = T\left(F_i, W_i, H_i\right) \tag{4}$$

where T is a 3×3 deformable convolution to modify the original feature map, F_i is the feature corresponding to the original i-th position and (W_i, H_i) is the anchor shape with respect to this location.

Concretely, motivated by deformable network [4], since different spatial locations in the receptive field contribute differently to the result of object detection, our new feature adaption module not only learns offsets during the process of convolutional deformation, but also adds a modulation term to adjust the feature amplitude to more flexibly manipulate spatial support regions. In the convolution kernel with i sampling points, w_i and s_i respectively represent the weight and the predefined offset at the i-th position. $x(s)$ and $y(s)$ are features of the position s corresponding to the input feature map x and the output feature map y. The formula of modulated deformable convolution can be expressed as the following form:

$$y(s) = \sum_{i=1}^{K} w_i \cdot x\left(s + s_i + \Delta s_i\right) \cdot \Delta m_i \tag{5}$$

where Δs_i and Δm_i are respectively the offset and modulation amplitude of deformable convolution for position i on the feature map. It has no definite limit to the Δs_i, and the Δm_i ranges from 0 to 1. Specifically, there is a key difference from the previous deformable network [16] that our scheme predicts offset and modulation amplitude based on the shape of the anchor rather than on the original feature map. As shown in Fig. 3, the shape prediction branch outputs offset and amplitude fields respectively through separate convolution layers, and then we apply feasible deformable convolution with the original feature map to obtain the adapted feature map f_I', which matches anchor greatly as well as facilitates subsequent network process.

D. RELATION MODULE

Due to the fact that there is collaboration between different surgical instruments in some scenarios, we apply Relation Network [5] in the proposed framework to enhance the performance of recognition. Specifically, the relation module derives comprehensive features by integrating the object's regular image features and multiple relation features associated with other targets, keeping the input and output dimensions unchanged, so it is easy to embed into the network structure. As illustrated in Fig. 3, the relation module is added to the two fully connected layers in sequence, followed by subsequent classification and regression. Experimental results show that the introduction of relational components can further improve recognition accuracy.

E. TRAINING AND INFERENCE

1) BACKBONE NETWORK

Our detector is based on Faster R-CNN [29] framework using the backbone of ResNet-101 [41] with FPN [42], which inputs video frames of laparoscopic surgery and outputs detection results of several surgical tools with bounding boxes labeling. Additionally, the output of detection prompts us to perform automatic surgical video analysis, which provides insights into future work through a comprehensive assessment of surgical skills.

2) TRAINING OBJECTIVE

The joint training objective for our structure is to minimize a multi-task loss:

$$L = \lambda_1 L_{loc} + \lambda_2 L_{shape} + L_{cls} + L_{reg} \tag{6}$$

Here L_{cls} and L_{reg} are conventional classification loss and regression loss. Moreover, L_{loc} is the loss corresponding to the anchor location prediction branch, and Focal Loss [43] is adopted. L_{shape} is the loss of shape prediction branch, which selects the modified version of bounded iou loss [44] to optimize only the width and height of anchor. Both λ_1 and λ_2 are trade-off coefficients for these two branches.

3) TRAINING DETAILS

It is experiential that leveraging high-quality proposals as input contributes to training a more efficient and accurate detector. Concretely, there is a vital premise for using high-quality proposals that the distribution of training samples and region proposals are consistent. Hence, a more rigorous screening criterion of the proposal is performed in our formulation by improving the IoU thresholds for positive and negative samples and generating fewer candidate anchors than the RPN-based approach.

F. DISCUSSION

In this section we compare the differences between Cascade R-CNN [45] and our work. (1) Both the Cascade R-CNN and our method are based on Faster R-CNN as the backbone network. Cascade R-CNN is a multi-stage object detection

TABLE 2. Results for two datasets of m2cai16-tool-locations and AJU-Set.

Dataset	Grasper	Bipolar	Hook	Scissors	Clipper	Irrigator	Specimen Bag	mAP
Mtl	54.1	69.5	86.8	73.9	84.2	41.6	77.1	69.6
AJU-Set	73.2	65.7	93.1	75.8	91.6	54.7	81.3	76.5

structure proposed to address the problem that the detection performance decreases when the IoU threshold increases. Different from the complex repetitive structure of Cascade R-CNN, our structure is more concise and clear, by adding function modules on the backbone network to generate adaptive anchors to achieve superior detection accuracy. (2) In order to produce high-quality proposals, the Cascade R-CNN consists of a series of detectors with rising IoU thresholds, where the output of the previous detector is used as input to train the next higher-quality detector. Compared with the above scheme, the design of our structure determines that the generated bounding boxes have higher quality, so it is easier to select fewer but high-quality proposals by setting relatively large IoU thresholds for positive and negative samples. (3) Different from the Cascade R-CNN based on the RPN to produce bounding boxes with fixed aspect ratios, a vital highlight of our method is to generate anchors with learnable shapes, which is more suitable for the scene of nonuniform distribution of objects like surgical tool recognition.

V. EXPERIMENTS

A. EXPERIMENT SETUP AND IMPLEMENTATION

We evaluate our detector on the m2cai16-tool-locations dataset [2] and our private dataset AJU-Set, which both include a number of images for video frames containing seven different kinds of surgical instruments. Since there have been few previous studies on the recognition of surgical instruments with limited data, we expand the above datasets by means of random horizontal flipping. For the evaluation of the following experiments, the standard mean average precision (mAP) and average precision (AP) are considered as the performance evaluation criteria for calculation.

To better balance the location and shape prediction branches, parameters λ_1 and λ_2 were set to 1 and 0.1, respectively. Using our laparoscopic surgery dataset, we take a stochastic gradient descent approach to fine-tune our network for 60K iterations and train our structure with a mini-batch size of 50. We start the learning rate at 0.001 and decay it by a factor of 10 every 10K iterations. All models in this paper are trained on NVIDIA Geforce RTX 2080 Ti GPUs, and our scheme achieves a real-time processing speed, which presents superior identification performance.

B. BASELINE METHODS

We evaluate our approach to the localization of surgical instruments in laparoscopic surgery separately on the m2cai16-tool-locations [2] and AJU-Set datasets. So far as we know, we are the first to use a comprehensive network model to detect surgical instruments in complete laparoscopic surgeries, which presents a good start for future medical analysis and research. As shown in Table 2, our method gives the detection results for seven surgical tools on the two datasets. Mtl denotes the dataset of m2cai16-tool-locations. Affected by various environmental factors such as illumination and lens sharpness, the two datasets obtain different detection performances respectively. From the second row, it can be seen that both the clipper and hook achieve higher detection accuracy in this dataset. The possible reason is that clipper usually has good visibility when operated, which makes it easy to recognize from the scene, and the hook is used relatively frequently in laparoscopic surgery as well as has a unique shape that contributes to distinguishing. Moreover, from our observation of the third row, these two tools bipolar and irrigator they get lower precision, reasonable explanations are that they all have irregular shapes similar to most medical tools and lack of enough data to learn from. Overall, our approach shows excellent performance for the task of laparoscopic surgery tool detection. Fig. 1 depicts the example frames of tool identity results on the m2cai16-tool-locations dataset. The detection performance of several samples from the AJU-Set involving adverse environmental factors such as occlusion, illumination, and smog refer to Fig. 2. It shows the remarkable recognition ability of our scheme.

In addition, to further validate our method, we compare it with existing one-stage and two-stage representative structures on the two datasets in Table 3. Generally, baseline methods of one-stage, such as SSD, RefineDet, and STDN are relatively less accurate than the two-stage methods like Faster R-CNN, Cascade R-CNN, and our method. We infer the reason that the two-stage approaches provide more accurate region proposals for predicting the variety of surgical tools. Moreover, it is exciting that our modulated anchoring network outperforms the previous scheme [2] with 6.5% and is 6.1% higher than the anchor-free method [37] on the m2cai16-tool-locations dataset, and also achieves better precision in AJU-Set. In summary, it demonstrates that our scheme performs well in the detection task of surgical instruments.

C. ABLATION STUDY

In order to verify the effectiveness of each component in Modulated Anchoring Network, we design several variants and evaluate them on our datasets, shown in the lower half of Table 3. Particularly, these variant models use the same settings for a fair comparison. Here Basic Network means the backbone network, that is, the remaining part after removing the anchor generation module, modulated feature module, and relation module from our framework. Note that AG

TABLE 3. Comparison with the state-of-the-art detection methods and results of ablation study on m2cai16-tool-locations and AJU-Set datasets.

Method	Backbone	mAP(%)	
		Mtl	AJU-Set
Faster R-CNN [29]	ResNet-101	62.3	65.5
SSD300 [30]	VGG-16	51.4	49.6
RefineDet [35]	VGG-16	62.8	67.3
Cascade R-CNN [45]	ResNet-101	65.1	71.9
FoveaBox [37]	ResNet-101	63.5	64.7
STDN300 [34]	DenseNet-169	62.1	64.2
Basic Network	ResNet-101	61.7	65.8
Basic + AG	ResNet-101	63.9	68.3
Basic + AG + DL	ResNet-101	63.0	68.1
Basic + AG + FA	ResNet-101	65.2	70.8
Basic + AG + MFA	ResNet-101	67.5	73.2
Modulated Network	ResNet-101	**69.6**	**76.5**

means anchor generation module and DL denotes applying a deformable convolution layer directly after the anchor generation module. FA is the feature adaption module, where shape prediction branch outputs offset field without modulation terms. MFA represents a modulated feature module, that is, the shape prediction branch predicts the offset as well as the modulation term for each position, and then obtains adaptive feature maps by applying a deformable convolution with offset and modulation scalar on the original feature maps.

Basic + AG indicates that we increase anchor generation module on the backbone network to predict the learnable shape and location of the anchor, proving the effectiveness of this module with the gain of 2.2% and 2.5% on the two datasets. It can be seen from the ablation experiment that the most obvious improvement is after adding the modulated feature module, which leads to the gain of 3.6% and 4.9%. Specifically, the outstanding performance is not only attributed to the use of a feasible deformable convolutional layer but also because we can better integrate anchor information into the feature by predicting offset and modulation amplitude of the deformable convolution through shape prediction branch. If the original feature maps with deformable convolution or adjusting the feature only by offset are directly used for prediction, which correspond to Basic + AG + DL and Basic + AG + FA in the table, it is obvious that detection performances are worse than using MFA. In addition, we compare the results of the last two rows in Table 3 and find that the relation module further improves the mAP by 2.1% and 3.3%, indicating that the module fused rich feature information around facilitates detection.

D. QUALITY ASSESSMENT OF SURGERY

Effective postoperative feedback can reduce the risk of complications in patients and provide paradigms for other young researchers. So the assessment of surgical techniques is a very significant stage but the existing assessment methods evaluated manually by experts are subjective as well as time-consuming. For this reason, we utilize our network structure to automatically evaluate surgical skills by tracking surgical tools and analyzing the patterns, the range of trajectories, the mobility economy of tools employed, following the medical assessment criteria for laparoscopic surgery.

We use four testing videos from the AJU-Set to assess surgical performance. As shown in the top of Fig. 4, the range of movement is studied utilizing heat maps generated by the position of bounding boxes of the surgical tools. Medical experience suggests that skilled surgical procedures should be operated more accurately in areas that are more concentrated in a certain range, which also presents the economy of tool movement. Through observation, it is obvious that the heat map corresponding to video 3 is the one with the best surgical performance among the four videos, showing the proficiency and high efficiency of doctor's operation.

Separating the gallbladder triangle is a very vital step in the surgical procedure, which can lead to biliary injury and complications if the surgeon has a slight deviation during the operation. So in order to further intuitively evaluate the mobility economy of tools, we select the gallbladder shearing stage from above step in the testing videos to generate trajectory maps of tool movement, presented in the bottom part of Fig. 4. Specifically, the surgeon places clips using the clipper to clamp the gallbladder artery and the cystic duct with the grasper holding the gallbladder properly, after that cut cystic duct with scissors. Due to the short time and small moving distance of the scissors, here we only study the clipper and grasper these two error-prone tool operations. The trajectory maps clearly show that the two surgical tools in the shearing stage of video 3 and video 4 are manipulated more smoothly and accurately, and compared with the surgical tools in the first two videos, there are no frequent movement, showing an excellent economy of motion. This condition may be due to the fact that the surgeons in the latter two videos are more

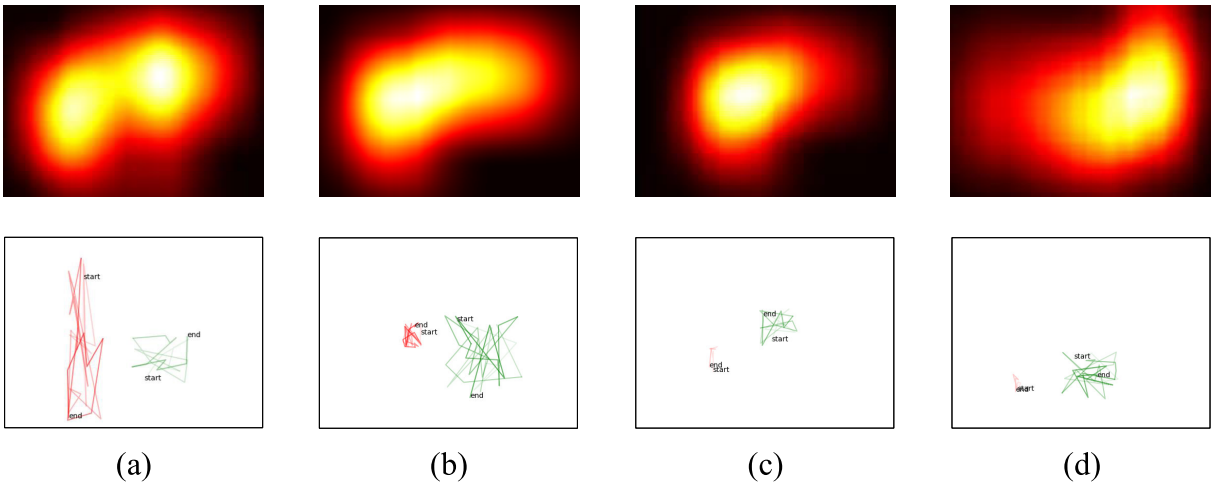


FIGURE 4. (a) to (d) correspond to video 1 to video 4, respectively. The top part is the heat maps generated by the positions of the surgical instruments, and the bottom part is the movement trajectories of tools during the shearing stage. These figures can be used to visually reflect the execution skills and efficiency of the surgical procedures.

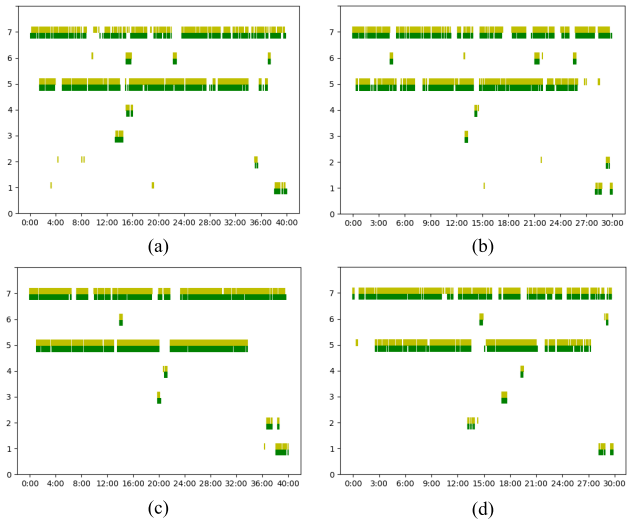


FIGURE 5. The yellow and green bands are derived from the prediction box and the ground-truth box, respectively. In the tool usage timelines, (a) to (d) correspond to video 1 to video 4. 1-7 represent the seven tools of Specimen Bag, Irrigator, Clipper, Scissors, Hook, Bipolar, and Grasper. Compared with other videos, video 3 indicates excellent hand dexterity and smooth operation.

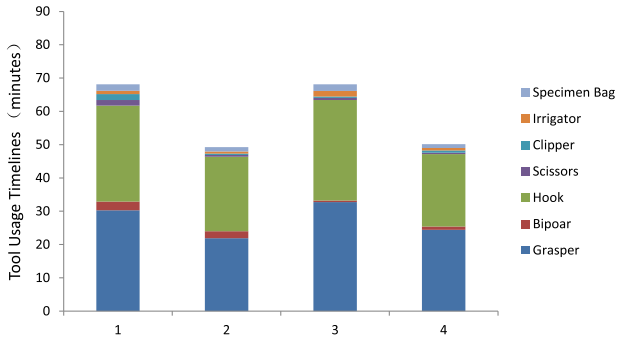


FIGURE 6. The total use time of each tools in the four videos reflects the skill and proficiency of the surgeon.

experienced, the position of the clips, the direction as well as the force of controlling the grasper are more appropriate. The above conjecture also gets confirmation during the expert's manual assessment of the surgical video.

In order to study the usage patterns of the instruments, usage timelines in the testing videos are generated to quantitatively assess surgical skills. From Fig. 5, we can see that the tools used in video 1 are switched relatively frequently and the time intervals are slightly longer. A reasonable explanation is that the surgical operation is somewhat unskilled, some details of the operation part are not done well, so additional operations are needed to remedy. For example, improper traction causes additional bleeding to affect subsequent work, requiring bipolar for rapid hemostasis. From the timelines, we can also draw conclusions that compared with the operation of video 1, video 3 is a smooth execution, more dexterous and referable. Besides, through observation, we can see that the overall change trend of the yellow band corresponding to the prediction box and the green band corresponding to the ground-truth box are consistent, and by analyzing the tool switching frequency and the time interval, the two color bands can reach a consistent conclusion of surgical evaluation, which also indicates that inaccurate detection of individual images does not affect the overall assessment result and the pattern chart is a reliable method for surgical evaluation.

As shown in Fig. 6, in addition to the above pattern chart, we also generate bar graphs corresponding to the total timelines of each tools used in the four videos to observe the skill and proficiency of the operation of the surgeon. For instance, we can clearly see that compared to other videos, the longer time the bipolar appears in video 1 and video 2 indicates that more bleeding is caused by the unskilled tissue handling and the bipolar has to be used many times to stop the bleeding. The above conclusion is consistent with

the pattern chart, which further verifies the effectiveness of the method in Fig. 5.

Moreover, aim to verify the reliability of all the above methods of evaluating surgical performance, we also ask four experts to manually assess the performance of the four testing videos from medical perspectives, and they all agree that video 3 is the best performing surgical procedure, which is consistent with our evaluation results.

VI. CONCLUSION

In this paper, we present a novel Modulated Anchoring Network based on Faster R-CNN, which consists of an anchor generation mechanism and the relation module. Unlike traditional schemes to provide anchors with fixed aspect ratios, our structure employs semantic information to generate adaptable shape anchors to detect surgical instruments that appear in laparoscopic surgery videos more flexibly and accurately. The modulated feature module with a modulation mechanism is proposed to expand the scope of deformable convolution to incorporate the anchor information into the feature map, and the relation module is embedded in our network to consider relationships of different tools. For this specific tool spatial detection task, our framework achieves excellent detection accuracy of 69.6% and 76.5% mAP on the m2cai16-tool-locations and AJU-Set datasets respectively, which is 4.5% and 4.6% higher than the current state-of-the-art method Cascade R-CNN, and also performs better than other comparable methods. Moreover, accurate detection results can further contribute to analyzing the tool usage patterns, trajectory range, and economy of motion from medical perspectives to assess the quality of surgery comprehensively. For the future work, we expect that a reliable medical evaluation system in three-dimensional space can be established to better analyze laparoscopic surgery and provide learning reference for inexperienced beginners.

REFERENCES

[1] 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.

[2] 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.

[3] 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.

[4] X. Zhu, H. Hu, S. Lin, and J. Dai, "Deformable convnets V2: More deformable, better results," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 9308–9316.

[5] H. Hu, J. Gu, Z. Zhang, J. Dai, and Y. Wei, "Relation networks for object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3588–3597.

[6] A. Agustinos and S. Voros, "2D/3D real-time tracking of surgical instruments based on endoscopic image processing," in *Proc. 2nd Int. Workshop Comput.-Assist. Robot. Endoscopy*, Vol. 9515. New York, NY, USA: Springer-Verlag, 2015, pp. 90–100.

[7] 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.

[8] I. Laina *et al.*, "Concurrent segmentation and localization for tracking of surgical instruments," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2017, pp. 664–672.

[9] 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.

[10] A. Zia, Y. Sharma, V. Bettadapura, E. L. Sarin, T. Ploetz, M. A. Clements, and I. Essa, "Automated video-based assessment of surgical skills for training and evaluation in medical schools," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 11, no. 9, pp. 1623–1636, Sep. 2016.

[11] M. Sahu, D. Moerman, P. Mewes, P. Mountney, and G. Rose, "Instrument state recognition and tracking for effective control of robotized laparoscopic systems," *Int. J. Mech. Eng. Robot. Res.*, vol. 5, no. 1, p. 33, 2016.

[12] L. Tao *et al.*, "Surgical gesture segmentation and recognition," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Berlin, Germany: Springer, 2013, pp. 339–346.

[13] G. Quellec, M. Lamard, B. Cochener, and G. Cazuguel, "Real-time segmentation and recognition of surgical tasks in cataract surgery videos," *IEEE Trans. Med. Imag.*, vol. 33, no. 12, pp. 2352–2360, Dec. 2014.

[14] C. Lea, G. D. Hager, and R. Vidal, "An improved model for segmentation and recognition of fine-grained activities with application to surgical training tasks," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, Jan. 2015, pp. 1123–1129.

[15] 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.

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

[17] R. Cadène, T. Robert, N. Thome, and M. Cord, "M2cai workflow challenge: Convolutional neural networks with time smoothing and hidden Markov model for video frames classification," 2016, *arXiv:1610.05541*. [Online]. Available: https://arxiv.org/abs/1610.05541

[18] O. Dergachyova, D. Bouget, A. Huaulmé, X. Morandi, and P. Jannin, "Automatic data-driven real-time segmentation and recognition of surgical workflow," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 11, no. 6, pp. 1081–1089, Jun. 2016.

[19] D. Tran, R. Sakurai, H. Yamazoe, and J.-H. Lee, "Phase segmentation methods for an automatic surgical workflflow analysis. *Int. J. Biomed. Imag.*, vol. 2017, 2017, Art. no. 1985796.

[20] L. Zappella, B. Béjar,, G. Hager, and R. Vidal, "Surgical gesture classification from video and Kinematic data," *Med. Image Anal.*, vol. 17, no. 7, pp. 732–745, Oct. 2013.

[21] Y. Jin, Q. Dou, H. Chen, L. Yu, and P.-A. Heng, "EndoRCN: recurrent convolutional networks for recognition of surgical workflow in cholecystectomy procedure video," Chin. Univ., Hong Kong, Tech. Rep., Oct. 2016.

[22] S. Bodenstedt, M. Wagner, D. Katić, P. Mietkowski, B. Mayer, H. Kenngott, B. Müller-Stich, R. Dillmann, and S. Speidel, "Unsupervised temporal context learning using convolutional neural networks for laparoscopic workflow analysis," 2017, *arXiv:1702.03684*. [Online]. Available: https://arxiv.org/abs/1702.03684

[23] S. Wang, A. Raju, and J. Huang, "Deep learning based multi-label classification for surgical tool presence detection in laparoscopic videos," in *Proc. IEEE 14th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2017, pp. 620–623.

[24] A. Raju S. Wang, and J. Huang, "'M2CAI surgical tool detection challenge report,'" in *Proc. Workshop Challenges Modeling Monit. Comput. Assist. Intervent. (M2CAI)*, Athens, Greece, 2016. [Online]. Available: http://camma.u-strasbg.fr/m2cai2016/reports/Raju-Tool.pdf

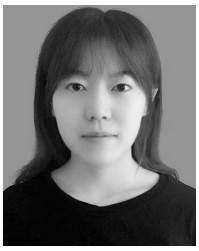
[25] M. Sahu, A. Mukhopadhyay, A. Szengel, and S. Zachow, "Tool and phase recognition using contextual CNN features," 2016, *arXiv:1610.08854*. [Online]. Available: https://arxiv.org/abs/1610.08854

[26] S. Wang *et al.*, "Graph convolutional nets for tool presence detection in surgical videos," in *Proc. Int. Conf. Inf. Process. Med. Imag.* Cham, Switzerland: Springer, 2019, pp. 467–478.

[27] D. Sarikaya, J. J. Corso, and K. A. Guru, "Detection and localization of robotic tools in robot-assisted surgery videos using deep neural networks for region proposal and detection," *IEEE Trans. Med. Imag.*, vol. 36, no. 7, pp. 1542–1549, Jul. 2017.

[28] B. Choi, K. Jo, S. Choi, and J. Choi, "Surgical-tools detection based on Convolutional Neural Network in laparoscopic robot-assisted surgery," in *Proc. 39th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2017, pp. 1756–1759.

[29] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.
[30] W. Liu *et al.*, "SSD: Single shot multibox detector," in *Proc. Eur. Conf. computer Vis.* Cham, Switzerland: Springer, 2016, pp. 21–37.
[31] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 779–788.
[32] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1904–1916, Sep. 2015.
[33] 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.
[34] 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.
[35] 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.
[36] Z. Jie, X. Liang, J. Feng, W. F. Lu, E. H. F. Tay, and S. Yan, "Scale-aware pixelwise object proposal networks," *IEEE Trans. Image Process.*, vol. 25, no. 10, pp. 4525–4539, Oct. 2016.
[37] T. Kong, F. Sun, H. Liu, Y. Jiang, and J. Shi, "Foveabox: Beyond anchor-based object detector," 2019, *arXiv:1904.03797*. [Online]. Available: https://arxiv.org/abs/1904.03797
[38] Z. Yang, S. Liu, H. Hu, L. Wang, and S. Lin, "Reppoints: Point set representation for object detection," 2019, *arXiv:1904.11490*. [Online]. Available: https://arxiv.org/abs/1904.11490
[39] J. Li, Y. Wei, X. Liang, J. Dong, T. Xu, J. Feng, and S. Yan, "Attentive contexts for object detection," *IEEE Trans. Multimedia*, vol. 19, no. 5, pp. 944–954, May 2017.
[40] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei, "Deformable convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 764–773.
[41] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[42] 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.
[43] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 2980–2988.
[44] L. Tychsen-Smith and L. Petersson, "Improving object localization with fitness NMS and bounded IoU loss," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6877–6885.
[45] 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.



BEIBEI ZHANG received the B.S. degree from the College of Computer Science and Technology, Jilin University, in 2018, where she is currently pursuing the M.E. degree. Her current research interests include deep learning, object detection, and medical image processing.

SHENGSHENG WANG 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 are in the areas of computer vision, deep learning, and data mining.

LIYAN DONG received the Ph.D. degree from Jilin University, in 2007. He is currently a Professor with the College of Computer Science and Technology, Jilin University. His research interests include database theory, data mining, distributed database and application, embedded database, and information retrieval. He has published more than 60 articles in domestic and foreign science and technology journals and international conferences.

PENG CHEN was born in 1976. She received the Ph.D. degree from Jilin University. She is currently a Professor with The Second Hospital of Jilin University.