IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received November 9, 2020, accepted December 15, 2020, date of publication December 21, 2020,
date of current version December 31, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3046258

Real-Time Surgical Tool Detection in Minimally Invasive Surgery Based on Attention-Guided Convolutional Neural Network

PAN SHI[1], ZIJIAN ZHAO[1], (Member, IEEE), SANYUAN HU[2], AND FALIANG CHANG[1]
[1]School of Control Science and Engineering, Shandong University, Jinan 250061, China
[2]Department of General Surgery, First Affiliated Hospital of Shandong First Medical University, Jinan 250014, China

Corresponding author: Zijian Zhao (zhaozijian@sdu.edu.cn)

This work was supported by the National Key Research and Development Program of China under Grant 2019YFB1311300.

ABSTRACT To enhance surgeons' efficiency and safety of patients, minimally invasive surgery (MIS) is widely used in a variety of clinical surgeries. Real-time surgical tool detection plays an important role in MIS. However, most methods of surgical tool detection may not achieve a good trade-off between detection speed and accuracy. We propose a real-time attention-guided convolutional neural network (CNN) for frame-by-frame detection of surgical tools in MIS videos, which comprises a coarse (CDM) and a refined (RDM) detection modules. The CDM is used to coarsely regress the parameters of locations to get the refined anchors and perform binary classification, which determines whether the anchor is a tool or background. The RDM subtly incorporates the attention mechanism to generate accurate detection results utilizing the refined anchors from CDM. Finally, a light-head module for more efficient surgical tool detection is proposed. The proposed method is compared to eight state-of-the-art detection algorithms using two public (EndoVis Challenge and ATLAS Dione) datasets and a new dataset we introduced (Cholec80-locations), which extends the Cholec80 dataset with spatial annotations of surgical tools. Our approach runs in real-time at 55.5 FPS and achieves 100, 94.05, and 91.65% mAP for the above three datasets, respectively. Our method achieves accurate, fast, and robust detection results by end-to-end training in MIS videos. The results demonstrate the effectiveness and superiority of our method over the eight state-of-the-art methods.

INDEX TERMS Attention mechanism, convolutional neural network, light-head module, real-time, surgical tool detection.

I. INTRODUCTION

Minimally invasive surgery (MIS) has attracted increasing attention in recent years, because it overcomes the major drawbacks of open surgery and provides surgeons with sufficient information only through small incisions [1]. Robot-assisted surgery (RAS) and laparoscopic surgery, two representative minimally invasive surgery, are widely used in a variety of clinical surgeries and aimed to improve surgeons' ability and ensure the safety of patients [2]. However, the indirect observation and operation method of MIS weakens surgeons' hand-eye coordination ability, which may affect the cognitive perception of visual data by surgeons during the operation process. The surgeons need to obtain additional information to monitor the movement of surgical tools in the patient's body, which hinders the translation of MIS globally [3].

To address these problems, surgical tool detection (STD) is widely used in recent years, which can provide accurate position estimation of two- or three-dimensional surgical tools by considering tool identification and positioning based on visual data. It also has many potential applications, such as accurate tracking [4]–[6], pose estimation [7]–[9], optimization of surgical scheduling, real-time reminders of surgery, and integrated post-surgical assessment [10]–[12]. They can further be used to warn clinical doctors of possible complications and to provide accurate real-time navigation for surgeons. Hence, detecting the surgical tools with high accuracy and speed in MIS is the focus of this study.

A series of research has been performed on STD [13]–[15]. The conventional STD is based on traditional shallow machine learning methods, most of them rely on handcrafted

The associate editor coordinating the review of this manuscript and approving it for publication was Navanietha Krishnaraj Rathinam.

features, such as color, gradient, shape, or color. Although such approaches are practical, they utilize markers that require modifications to surgical tools design and interfere with the surgical workflow. Therefore, the deep learning methods based on convolutional neural networks (CNN) for STD have gradually become a trend, ensuring a smooth operation and having no modification to surgical tools [16]. Nowadays, the CNN-based methods of STD can be subdivided into single- and two-stage ones. Choi *et al.* [17] proposed a single-stage method, which improved the architecture of you only look once (Yolo) [18] and regarded the detection task as a regression problem, which can directly predict the coordinate values of the boundary box in real time but has no outstanding accuracy. The two-stage methods proposed by Sarikaya *et al.* [19] and Jin *et al.* [20] can achieve high detection accuracy by improving the algorithm of faster regions with convolutional neural network (Faster R-CNN) [21], but fail to detect surgical tools on a real-time scale (operating at less than 10 FPS). Therefore, a common deficiency of these methods is a huge imbalance between accuracy and speed, which hinders the subsequent practical application.

To solve this problem, many novel methods have been proposed for surgical tool detection. Zhang *et al.* [22] formulated a novel framework called ''modulated anchoring network,'' leveraging semantic features to effectively detect non-uniformly distributed surgical tools of arbitrary anchor shapes. This method achieved good detection accuracy of surgical tools, but Zhang *et al.* [22] did not disclose the detection speed. Zhao *et al.* [23] proposed using a cascade CNN to perform STD, which consisted of an hourglass network and a modified visual geometry group (VGG) network. This algorithm achieved better detection performance in terms of detection accuracy and speed, but it failed to achieve end-to-end training. Later, Liu *et al.* [24] proposed an anchor-free CNN architecture by using a lightweight stacked hourglass network, which modelled the surgical tool as a single point: the center point of its bounding-box. This method eliminated the need to design a set of anchor boxes, and achieved end-to-end training.

However, none of these studies paid attention to the dependency between channels and the importance of different channel features. As far as we know, there are various adverse environmental factors in MIS videos (such as motion blurring, high deformation, tools with occlusion, tools' overlap, and missing parts), which may affect the detection results of surgical tools. Recent studies [25]–[27] indicated that the attention mechanism can be added to the convolution neural network structure, which could help the network capture the key regions more effectively by modeling the dependency between channels, and further improve the detection accuracy of objects.

Therefore, to overcome the challenges above and inspired by RefineDet [28], we propose a real-time attention-guided CNN for frame-by-frame detection of surgical tools in MIS videos, which subtly combines the attention mechanism and light-head modules. The former helps the network adaptively fuse more context information and enhance the model's ability to focus on the relevant image areas, with which the subsequent regression and classification will be facilitated. The latter reduces the number of parameters and the computational complexity, which is used to accelerate the network's detection speed. We evaluated our method on three surgical datasets, including the EndoVis Challenge dataset [8], ATLAS Dione dataset [19], and a new dataset we introduced, Cholec80-locations. The experimental results prove that the proposed method has an excellent performance in terms of accuracy and speed and surpasses eight state-of-the-art detection algorithms. The main contributions of this work are as follows:

(1) We proposed a single-stage CNN with attention mechanism for real-time STD in MIS videos, incorporating the squeeze-and-excitation network (SENet) [29] to promote the network learning of the most useful feature representations and improve the detection accuracy of surgical tools.

(2) We designed a lightweight detection head module, called light-head, which integrates the lightweight idea (depth-wise separable convolution [30]) in our architecture. The novel module enabled the network directly output coordinates and classification scores of surgical tools using only 1×1 convolutions. It reduced the number of parameters and the computational complexity of the network, which can boost the proposed STD method's speed. The experimental results on three surgical datasets demonstrate the effectiveness and superiority of our method over state-of-the-art methods.

(3) We introduced a new dataset, Cholec80-locations, which extended the Cholec80 dataset [31] with the coordinates annotations of bounding boxes of tools, for STD in MIS. It is considered an important reference for researchers in the STD field.

The rest of this paper is organized as follows. We first present our proposed method, then introduce two public datasets and a new dataset we established, Cholec80-locations. After that we describe the experiments and results, next provide a comparative discussion of our results and future directions for further improvement. Finally, we draw the main conclusions.

II. METHODOLOGY

Inspired by works [29], [30], we designed a lightweight attention-guided CNN that inherits the advantages of the single- and two-stage detection methods and works more accurately and efficiently than RefineDet [28]. The overall framework is shown in Figure 1. The proposed approach performed the STD via a coarse detection module and a refined detection module. The method achieved end-to-end training by using the multi-task loss function. The distance intersection-over-union non-maximum suppression (DIoU-NMS) was proposed to post-process the tools detection results.

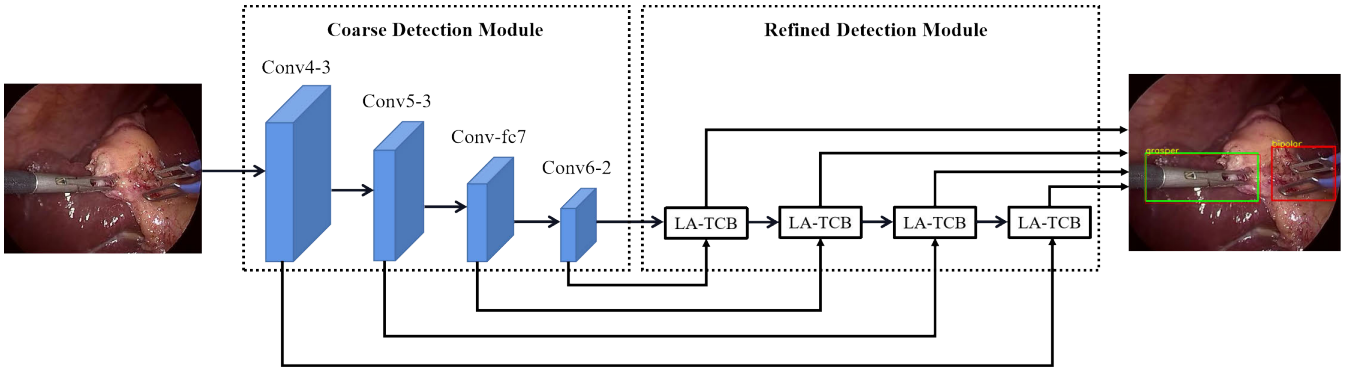


FIGURE 1. The overall framework of our proposed method. The network, including a coarse detection module and a refined detection module, is trained jointly in an end-to-end fashion with the multi-task loss function.

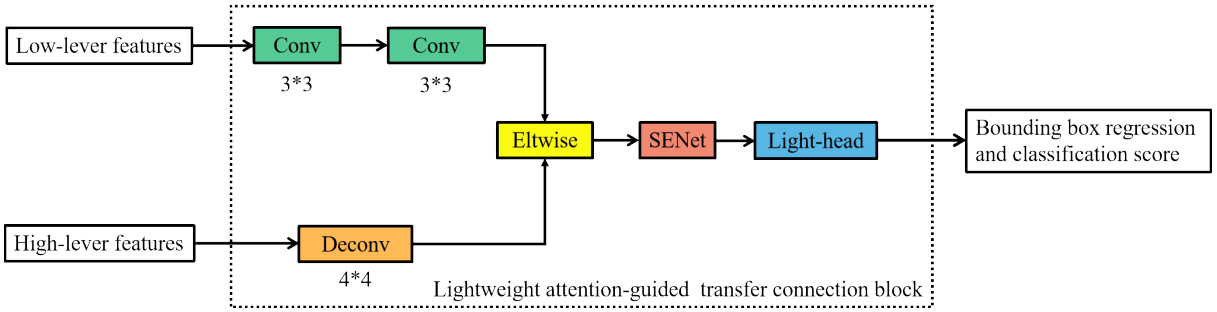


FIGURE 2. The structure of the lightweight attention-guided transfer connection block, incorporating the SENet and light-head module.

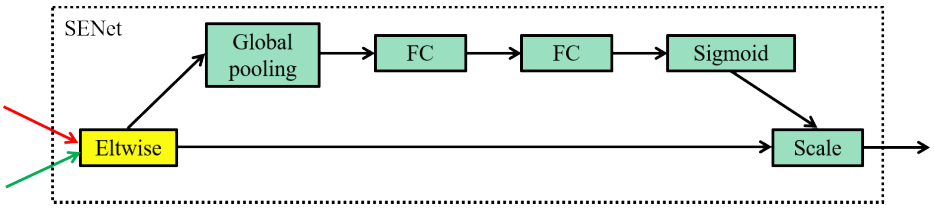


FIGURE 3. The overview of the SENet module.

A. COARSE DETECTION MODULE

Like RefineDet [28], our architecture's backbone network is VGG-16, which is pre-trained on the ImageNet [32]. The two fully connected layers (fc6, fc7) of VGG-16 are converted to two convolution layers (Conv-fc6, Conv-fc7), and two extra convolution layers (Conv6-1, Conv6-2) are added after VGG-16. We also use the features from Conv4-3, Conv5-3, Conv-fc7, and Conv6-2 for multi-scale prediction (see Figure 1). The coarse detection module (CDM) is aimed to coarsely regress the parameters of locations to get the refined anchors, which can provide better initialization for refined detection module (RDM); at the same time, it focuses on a binary classification task that decides whether the anchor is a tool or background, and filters out a large number of negative anchors to address the imbalance problem of positive and negative samples.

B. REFINED DETECTION MODULE

The refined detection module (RDM) consists of four lightweight attention-guided transfer connection blocks (LA-TCBs), designed to adaptively transfer the features of low- and high-lever layers from the CDM, and further utilize the refined anchors from CDM to generate accurate locations and classification scores of surgical tools. The architecture of the LA-TCB is illustrated in Figure 2.

We subtly incorporated the SENet [29] module in the LA-TCB without introducing too many extra computational parameters. Thus, the network could fuse more context information and enhance the network's ability to focus on the relevant image areas, facilitating the subsequent regression and classification. The SENet module, as shown in Figure 3, comprises a global average pooling layer, two full connection layers, and a sigmoid activation function. It analyzes the

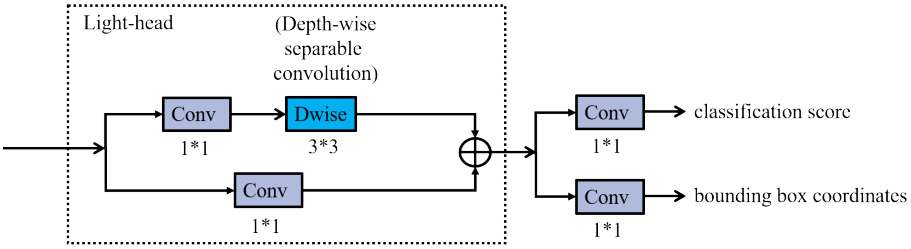


FIGURE 4. The overview of the light-head module.

relationship between channels and enables the network to pay more attention to the most informative channel features and suppress the unimportant channel features automatically. Therefore, with the guidance of the attention mechanism, our method's detection accuracy is expected to be further improved.

Besides, we designed a lightweight detection head based on the depth-wise separable convolution [30], namely the light-head, which replaced the original 3×3 standard convolutions. The light-head module further fused the features through a mixture of two ways. The first one was a 1×1 convolution, while the second one combined a 1×1 convolution and a 3×3 depth-wise separable convolution. More details can be found in Figure 4. More importantly, the light-head module enabled our network to directly output the coordinates and classification scores of surgical tools using only 1×1 convolutions, which reduced the number of parameters and the network's computational complexity so that the detection speed could be improved greatly.

C. LOSS FUNCTION AND DIoU-NMS

As a kind of single-stage detector, our method also inherited the advantages of the two-stage detection methods by using a two-step cascaded regression and classification strategy, which could detect surgical tools more accurately and efficiently in MIS videos.

As was earlier mentioned, the loss of our method contained two parts: the coarse losses in the CDM and the refined losses in the RDM. For the CDM, we coarsely regressed the parameters of locations to get refined anchors; simultaneously, each anchor was assigned a binary class label (tool or background). Then, in the RDM, we utilized the refined anchors from the CDM to generate accurate coordinates and classification scores of surgical tools. Therefore, we defined the loss function as:

$$L(\{p_i\},\{X_i\},\{c_i\},\{B_i\}) = l_{cls}^{cdm}(\{p_i\}) + l_{loc}^{cdm}(\{X_i\}) + l_{cls}^{rdm}(\{c_i\}) + l_{loc}^{rdm}(\{B_i\}), \quad (1)$$

$$l_{cls}^{cdm}(\{p_i\}) = \frac{1}{N_{cdm}}\left(\sum_i [l_i^* \geq 1] * C(p_i, l_i^*)\right), \quad (2)$$

$$l_{loc}^{cdm}(\{X_i\}) = \sum_i [l_i^* \geq 1] * L_{r1}(X_i, B_i^*), \quad (3)$$

$$l_{cls}^{rdm}(\{c_i\}) = \frac{1}{N_{rdm}}\left(\sum_i [l_i^* \geq 1] * C(c_i, l_i^*)\right), \quad (4)$$

$$l_{loc}^{rdm}(\{B_i\}) = \sum_i [l_i^* \geq 1] * L_{r2}(B_i, B_i^*). \quad (5)$$

In Eqs.(1-5), i denotes the anchor index in a mini-batch; p_i and X_i are the predicted class confidence (of the anchor i being a tool) and localization coordinates of the anchor i in the CRM, respectively; c_i and B_i are the predicted surgical tool category and coordinates of the bounding box in the RDM, respectively; N_{cdm} and N_{rdm} are the numbers of positive anchors in the CRM and RDM, respectively; l_i^* is the class label of ground truth, and B_i^* is the coordinate of ground truth localization. In the CRM, $C(*)$ is the cross-entropy loss over two classes (tool or background), the regression loss L_{r1} is the smooth L1 loss. In the RDM, $C(*)$ is the cross-entropy loss over multiple classes. The regression loss L_{r2} is the distance intersection-over-union (DIoU) loss [33], which considers the overlap area and the central distance of the bounding boxes, and achieves better regression. It can be formulated as:

$$L_{r2}(B_i, B_i^*) = 1 - IoU(B_i, B_i^*) + \frac{\rho^2(b, b^*)}{c^2}, \quad (6)$$

where b and b^* are the central points of B_i and B_i^*, respectively, $\rho(*)$ is the Euclidean distance, while c denotes the diagonal length of the smallest enclosing box of B_i and B_i^*.

Meanwhile, the DIoU-NMS [33] was employed as the post-processing approach to produce the final detection results. It can be formally defined as:

$$s_i = \begin{cases} s_i & DIoU(M, B_i) \leq \varepsilon \\ 0 & DIoU(M, B_i) \geq \varepsilon, \end{cases} \quad (7)$$

where s_i is the confidence of classification, and M is the predicted box with the highest confidence. This means that redundant detection boxes will be removed as long as its overlap area with M greater than the threshold ε.

III. DATASET

There are two public datasets for STD, including the EndoVis Challenge dataset [8] and ATLAS Dione dataset [19]. The former included 1083 frames from ex-vivo video sequences of interventions. Each frame was labeled with coordinates of tools' spatial bounds, and the resolution was 720×576. This dataset was manually re-labeled and split into 810 and

FIGURE 5. STD challenges presented by the ATLAS Dione dataset [19].

FIGURE 6. Seven tools in the Cholec80-locations dataset (left to right): grasper, bipolar, hook, scissors, clipper, irrigator, and specimen bag.

283 frames for the model training and testing, respectively. The ATLAS Dione dataset [19] contained 99 video clips of 10 surgeons from the Roswell Park Cancer Institute (Buffalo, USA) performing six different surgical tasks on the da Vinci Surgical System (dVSS). Each frame was labeled with coordinates of spatial bounds of tools, and the resolution was 854 × 480. Similarly, all video clips were split into 90 video clips (20491 frames) and nine video clips (1976 frames) for training and testing, respectively. Noteworthy is that the ATLAS Dione dataset featured camera movement and zoom, free movement of surgeons, a wider range of expertise levels, background objects with high deformation, and annotations including tools with occlusion, changed position. Therefore, the tool detection tasks had some challenges, such as motion blurring, tools with occlusion, overlap, and missing tool parts, as shown in Figure 5.

Noteworthy is that the public datasets for STD in MIS are limited. The ATLAS Dione dataset was a phantom setting, while the EndoVis Challenge dataset was from ex-vivo video sequences. Particularly, Jin *et al.* [20] established the m2cai16-tool-locations, which extended the m2cai16-tool dataset with the coordinates of spatial bounding boxes around surgical tools. This dataset consisted of 2532 labled frames, which were selected from among the 23,000 total frames. In m2cai16-tool-locations, most frames contained just one tool, while a few with two or three tools. However, there were various adverse environmental factors in MIS videos (such as motion blurring, high deformation, tools with occlusion, tools' overlap, and missing parts), this dataset only included part of the challenges. In order to get sufficiently representative and comprehensive detection results, we need a more challenging dataset for the detection of surgical tools.

The Cholec80 dataset [31], as well as m2cai16-tool, focused on the presence detection of surgical tools. The Cholec80 dataset [31] contained 80 cholecystectomy surgical videos performed by 13 surgeons at the University Hospital of Strasbourg in France. The videos were captured at 25 FPS and downsampled to 1 FPS for processing. Each frame was labeled with the tool presence annotations, without their spatial annotations. Therefore, based on the Cholec80 dataset, we collected and introduced a new, more challenging dataset, Cholec80-locations, that labeled 4011 frames with spatial annotations of surgical tools. Followed the data annotation standard in the ATLAS Dione dataset [19], we manually annotated the bounding boxes with the supervision of an expert MIS surgeon. It is worth noting that it's not the entire tool's body that was annotated but simply the tips. This is because only the tip of the tool is visible in most MIS videos, and most of their handles are similar. In particular, for surgical tools without handles, such as specimen bag, we labeled its entire body. Additionally, the 4011 frames were selected from the total frames, and the resolution of each frame is 854 × 480. Since each frame could contain one or more tools, this dataset was split into 3289 and 722 frames for training and testing, respectively. In total, the Cholec80-locations dataset contained seven kinds of surgical tools: grasper, bipolar, hook, scissors, clipper, irrigator, and specimen bag. The actual samples are shown in Figure 6. The specific annotated statistical data in Cholec80-locations and the complete Cholec80 dataset are listed in Table 1.

IV. EXPERIMENTS AND RESULTS

A. EXPERIMENTAL SETUP AND IMPLEMENTATION

We resized the input image frames to 320 × 320 pixels before training and performed data augmentation, including optical transformation and geometric transformation. The former included a random adjustment of brightness and contrast; the latter included random expanding, cropping the original training frames, and random flipping of frames horizontally. Most of the operations were random processes to ensure as much data richness as possible. We fine-tuned the network using stochastic gradient descent (SGD) and trained with

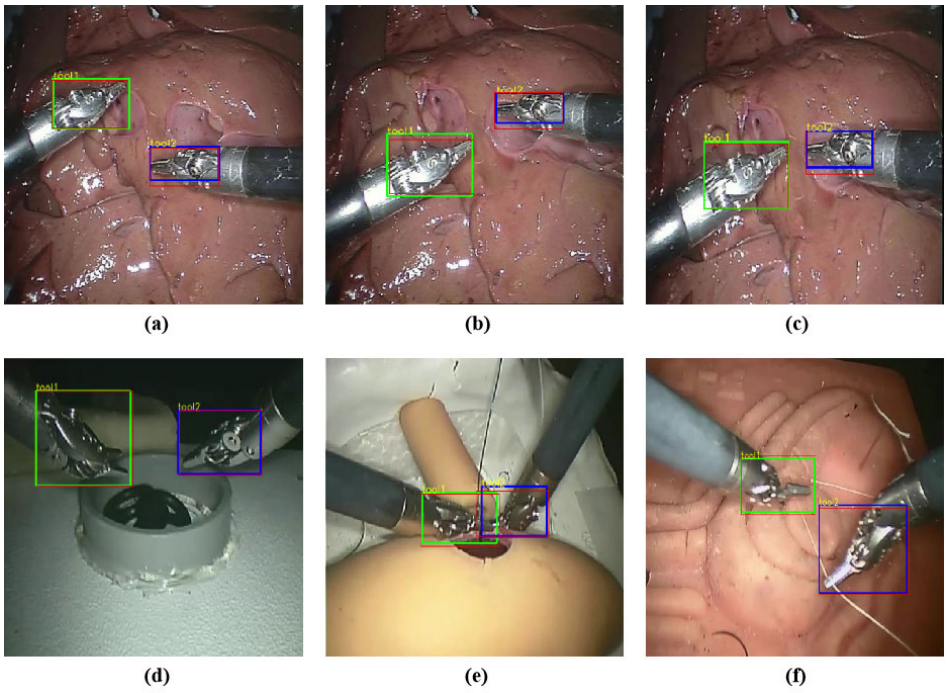


FIGURE 7. Detection results for two public datasets. The first rows (a-c) are from the Endovis Challenge dataset [8], and the second rows (d-f) from the ATLAS Dione dataset [19].

TABLE 1. Dataset statistics. (Row1) Number of frames for each tool in the complete Cholec80 dataset. (Row2) Number of annotated frames for each tool in the Cholec80-locations dataset.

Tools	Cholec80	Cholec80-locations
Grasper	102588	2880
Bipolar	8876	579
Hook	103106	1263
Scissors	3254	388
Clipper	5986	400
Irrigator	9814	485
Specimen bag	11462	476
Total	245086	6471

a mini-batch size of 16. All layers were initialized with a learning rate of 5×10^{-5}, a momentum of 0.9, and a weight decay of 5×10^{-5}. The learning rate was decreased by a factor of 10 at 100 and 150 epochs. We trained the whole network in an end-to-end fashion for 200 epochs. To compare with other state-of-the-art methods fairly, all experiments were conducted on PyTorch 1.0, Ubuntu 18.04 LTS operating system using an NVIDIA GeForce GTX TITAN X GPU accelerator. Our method was found to have a considerable inference speed, achieving real-time STD in MIS videos.

B. DETECTION RESULTS

We extensively validated our method on the three surgical datasets mentioned above. Figures 7 and 8 show our model's detection examples on the two public datasets and the Cholec80-locations dataset, respectively. The boxes with thin and thick lines represent the ground-truth and our detection results, respectively. These detection examples proved that our method could accurately detect the locations and sizes of surgical tools in MIS. And it can be noted that our proposed network could distinguish different surgical tools with a similar form in a frame. To the best of our knowledge, this was the first attempt to use a lightweight attention-guided CNN model for detecting surgical tools in MIS videos, with which the subsequent tasks (such as the surgical reports generation, the operation process optimization) could be facilitated in the long run.

For the quantitative evaluation of STD performance in MIS, we adopted the widely used two metrics: mean average precision (mAP) and frames per second (FPS). The former represents the mean of all classes' average precision, which is the most commonly used index to evaluate the quality of the detection model. As known, the average precision (AP) is the average of the ratio of the correctly detected surgical tools to the total number. We followed the definition in the Pascal VOC dataset [34], the detection was considered correct if the bounding box intersection over union (IoU) between the detected tool and the ground truth exceeded 0.5. The latter refers to the mean number of detected frames per second during the whole detection process, which can evaluate the speed of STD methods. Under the two evaluation metrics, we separately compared our proposed method's results to those of eight state-of-the-art detection methods on the three MIS datasets introduced above. The eight state-of-the-art detection methods included three two-stage methods (Faster RCNN [21], RelationNet [35] and Cascade CNN [23]) and five single-stage methods (RetinaNet [36], Yolov3 [37],

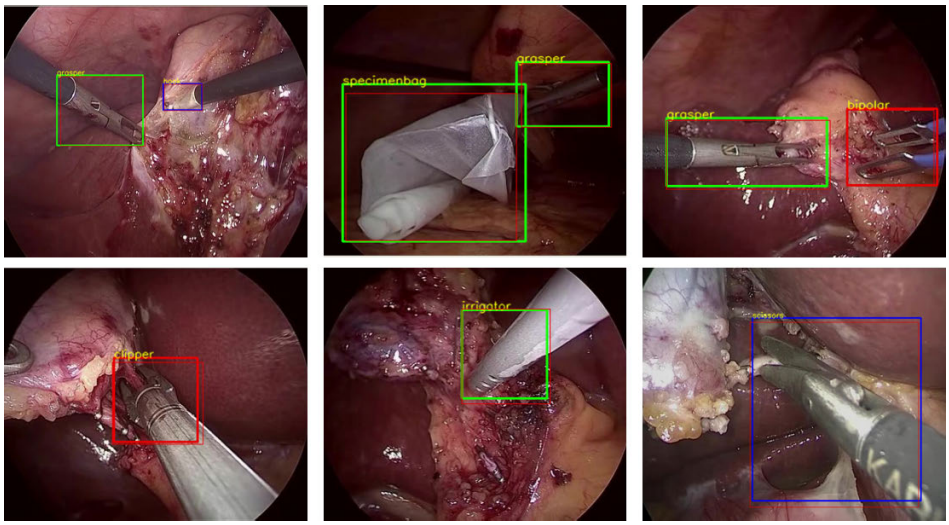


FIGURE 8. Detection results for the Cholec80-locations dataset.

TABLE 2. STD results of all methods on two public datasets. The mAP1 and mAP2 refer to the detection mAP on the EndoVis Challenge [8] and ATLAS Dione [19] datasets, respectively.

	Method	Backbone	mAP1(%)	mAP2(%)	FPS
Two-stage	Faster RCNN	VGG-16	100	90.36	16.1
	RelationNet	VGG-16	100	92.71	19.3
	Cascade CNN	VGG-16	100	91.60	43.5
Single-stage	RetinaNet	ResNet-50	99.99	89.39	12.2
	Yolov3	DarkNet-53	99.07	90.92	29.4
	HRNet	HRNet-W32	100	91.57	32.3
	CenterNet	Hourglass	100	98.50	37.0
	RefineDet	VGG-16	100	91.62	47.6
	This study	VGG-16	100	94.05	55.5

HRNet [38], CenterNet [24] and RefineDet [28]). The detection results for all methods on the two public datasets and the Cholec80-locations dataset are summarized in Tables 2 and 3, respectively.

In Table 2, the detection mAP of all methods on the EndoVis Challenge and ATLAS Dione datasets are represented by mAP1 and mAP2, respectively. Our method achieved a different detection mAP on the two public datasets because the ATLAS Dione dataset is more challenging due to various disturbing factors, such as different surgical tasks, motion blurring, high deformation, tools with occlusion, tools' overlap, and missing tool parts. With the detection mAP of 100 and 94.05% on the EndoVis Challenge and ATLAS Dione datasets, respectively, our method surpassed all methods except CenterNet [24]. Our method had the highest speed (55.5 FPS) on the two public datasets, which could satisfy the real-time requirement of STD in MIS.

As shown in Table 3, we also provided the average precision (AP) per tool of all methods on the Cholec80-locations dataset to further prove our method's feasibility. As Table 3 shows, our method achieved the mAP of 91.65% on the new dataset, which was slightly lower than CenterNet [24], but outperformed all other detection methods. The AP values of all surgical tools, except grasper and irrigator, exceeded 90%. Moreover, the hook generally had a higher detection AP in all of the methods, whereas our method achieved the AP of 99.33%. Usually, the hook had good visibility and high discrimination, making it easy to distinguish from other tools. However, it can be seen that both the grasper and irrigator usually get lower AP. The possible reason is that their appearance is similar to some other surgical tools, and the shape of the two kinds of tools is irregular, which are not being considered in our detection method. This issue needs to be explored and mitigated in future work.

C. ABLATION STUDY

In order to show the advantages of each component in our method, we designed several variants and evaluated them on the three datasets previously mentioned. As shown in Table 4, the mAP1, mAP2 and mAP3 refer to the detection mAP on the EndoVis Challenge [8], ATLAS Dione [19] and Cholec80-locations datasets, respectively. The Basic Network here refers to the baseline architecture after removing the attention mechanism module and replacing the light-head module with original 3×3 standard convolutions in our network. Here AG and LH denote the attention mechanism and light-head module, respectively.

Basic + AG indicates that we incorporated the attention mechanism module (SENet) in the baseline architecture to promote the network learning of the most useful feature representations. It can be seen from Table 4 that the attention mechanism module in our network can effectively improve the detection accuracy. This module improved the mAP of 2.66 and 2.92%, respectively, on the ATLAS Dione and Cholec80-locations datasets. Basic + LH indicates that we replaced the original 3×3 standard convolution with

TABLE 3. STD results of all methods on the Cholec80-locations dataset.

	Method	Grasper	Bipolar	Hook	Scissors	Clipper	Irrigator	Specimen bag	mAP(%)	FPS
Two-stage	Faster RCNN	81.28	86.68	92.69	87.90	83.54	80.33	89.63	86.01	16.1
	RelationNet	85.69	90.41	99.30	90.12	88.77	89.39	90.03	90.53	19.3
	Cascade CNN	88.58	90.98	92.80	88.56	91.80	90.11	89.86	90.38	43.5
Single-stage	RetinaNet	85.41	90.36	90.84	90.58	90.05	87.42	89.98	89.23	12.2
	Yolov3	81.31	86.03	92.30	87.35	87.50	85.97	88.74	87.03	29.4
	HRNet	89.94	90.90	98.09	90.67	90.19	88.98	89.69	91.21	32.3
	CenterNet	90.95	92.36	99.58	92.65	91.94	90.86	91.84	92.88	37.0
	RefineDet	85.38	90.88	90.82	90.61	90.56	87.54	90.05	89.41	47.6
	This study	89.88	90.52	99.33	90.78	90.19	89.62	91.25	91.65	55.5

TABLE 4. Results of ablation study on the above three datasets. The mAP1, mAP2 and mAP3 refer to the detection mAP on the EndoVis Challenge [8], ATLAS Dione [19] and Cholec80-locations datasets, respectively.

Method	mAP1(%)	mAP2(%)	mAP3(%)	FPS
Basic Network	100	91.62	89.41	47.6
Basic+AG	100	94.28	92.33	43.5
Basic+LH	100	91.16	87.94	57.2
Basic+AG+LH	100	94.05	91.65	55.5

light-head module in the baseline architecture to reduce the number of parameters and the computational complexity of the network. Although the detection accuracy was reduced slightly, the speed was increased by nearly 10 FPS, indicating the effectiveness of light-head module. Basic+AG+LH denotes the attention-guided CNN we proposed, which incorporated the attention mechanism module and light-head module. Compared to Basic Network, the speed was increased by 7.9 FPS, and accuracy was improved by 2.43 and 2.24%, respectively, on the ATLAS Dione and Cholec80-locations datasets. It achieved the best trade-off between detection speed and accuracy, proving the effectiveness of our method.

V. DISCUSSION

As shown in Tables 2 and 3, in the eight state-of-the-art detection approaches, RefineDet [28] achieved a better trade-off between the detection speed and accuracy because of its two-step cascaded regression and classification strategy. Our method inherited the above advantages and further improved the detection performance by using the attention mechanism and light-head modules specially designed to detect the surgical tool in MIS. In terms of accuracy, our method surpassed all methods except CenterNet [24], but had the highest speed (55.5 FPS) that could satisfy the real-time requirement of STD in MIS. Compared to Basic Network in Table 4, our method's accuracy was improved by 2.43 and 2.24%, respectively, on the ATLAS Dione and Cholec80-locations datasets, which was largely attributed to the introduction of the attention mechanism. With the guidance of the attention mechanism, our network considered the dependencies on each channel, paid more attention to the most informative channel features, and suppressed the unimportant channel features. Meanwhile, the detection speed was increased by 7.9 FPS, indicating that our proposed light-head module considerably improved the network's computational efficiency.

In summary, all of these results demonstrate that our framework had a remarkable ability in terms of accuracy and speed, and outperformed all the state-of-the-art detection algorithms mentioned above by a large margin. The accurate and real-time detection results also demonstrate our method has the potential to be further applied to other video analysis tasks of MIS, and has good generalization ability of STD in various datasets with different challenges, such as motion blurring, high deformation, tools with occlusion, tools' overlap, and missing parts, and so on.

However, there is still a lot of room for improvement. For example, one can use the recurrent neural network (RNN) or its variants to extract the MIS videos' long-term temporal information and model the temporal dependency between the frames to get better results. Moreover, there are limited datasets with location annotations of tools, the weakly supervised [39], [40] or self-supervised methods [41]–[43] can be used to reduce the dependency on spatially annotated data. To the best of our knowledge, the Cholec80 dataset [31] contains various real-world environments, so we will continue to label the Cholec80 dataset with the coordinates of bounding boxes of surgical tools to address the lack of public datasets and further promote the development of STD in MIS.

VI. CONCLUSION

This paper proposed a single-stage CNN architecture with an attention mechanism for the effective frame-by-frame detection of surgical tools in MIS videos: the CDM and the RDM. The CDM was trained to coarsely regress the locations' parameters to get the refined anchors, and perform binary classification that would decide whether the anchor is a tool or background. The RDM subtly incorporated the attention mechanism to generate accurate detection results utilizing the refined anchors from CDM. Then, a light-head module was designed to reduce the number of parameters and the computational complexity. Besides, we collected and introduced a new dataset, Cholec80-locations, to address the lack of public STD datasets. The experimental results show that our method ran in real-time at 55.5FPS and achieved a superior

detection accuracy even with various disturbing factors (such as different surgical tasks, motion blurring, high deformation, tools with occlusion, tools' overlap, and missing parts), which outperformed most of the state-of-the-art algorithms on STD in MIS videos. In the future, we will focus on exploring and designing more effective CNN with the attention mechanism to improve the detection accuracy and speed.

REFERENCES

[1] D. Bouget, M. Allan, D. Stoyanov, and P. Jannin, "Vision-based and marker-less surgical tool detection and tracking: A review of the literature," *Med. Image Anal.*, vol. 35, pp. 633–654, Jan. 2017.

[2] J. Ryu, J. Choi, and H. C. Kim, "Endoscopic vision-based tracking of multiple surgical instruments during robot-assisted surgery," *Artif. Organs*, vol. 37, no. 1, pp. 107–112, Jan. 2013.

[3] 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.

[4] Z. Chen, Z. Zhao, and X. Cheng, "Surgical instruments tracking based on deep learning with lines detection and spatio-temporal context," in *Proc. Chin. Automat. Congr. (CAC)*, Oct. 2017, pp. 2711–2714.

[5] X. Du, M. Allan, A. Dore, S. Ourselin, D. Hawkes, J. D. Kelly, and D. Stoyanov, "Combined 2D and 3D tracking of surgical instruments for minimally invasive and robotic-assisted surgery," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 11, no. 6, pp. 1109–1119, Apr. 2016.

[6] Z. Zhao, Z. Chen, S. Voros, and X. Cheng, "Real-time tracking of surgical instruments based on spatio-temporal context and deep learning," *Comput. Assist. Surg.*, vol. 24, no. 1, pp. 20–29, Oct. 2019.

[7] E. Colleoni, S. Moccia, X. Du, E. De Momi, and D. Stoyanov, "Deep learning based robotic tool detection and articulation estimation with spatio-temporal layers," *IEEE Robot. Autom. Lett.*, vol. 4, no. 3, pp. 2714–2721, Jul. 2019.

[8] T. Kurmann, P. M. Neila, X. Du, P. Fua, D. Stoyanov, S. Wolf, and R. Sznitman, "Simultaneous recognition and pose estimation of instruments in minimally invasive surgery," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, Sep. 2017, pp. 505–513.

[9] X. Du, T. Kurmann, P.-L. Chang, M. Allan, S. Ourselin, R. Sznitman, J. D. Kelly, and D. Stoyanov, "Articulated multi-instrument 2-D pose estimation using fully convolutional networks," *IEEE Trans. Med. Imag.*, vol. 37, no. 5, pp. 1276–1287, May 2018.

[10] Y. Jin, Q. Dou, H. Chen, L. Yu, J. Qin, C.-W. Fu, and P.-A. Heng, "SV-RCNet: Workflow recognition from surgical videos using recurrent convolutional network," *IEEE Trans. Med. Imag.*, vol. 37, no. 5, pp. 1114–1126, May 2018.

[11] H. Al Hajj, M. Lamard, P.-H. Conze, B. Cochener, and G. Quellec, "Monitoring tool usage in surgery videos using boosted convolutional and recurrent neural networks," *Med. Image Anal.*, vol. 47, pp. 203–218, Jul. 2018.

[12] H. Nakawala, R. Bianchi, L. E. Pescatori, O. De Cobelli, G. Ferrigno, and E. De Momi, "'Deep-Onto' network for surgical workflow and context recognition," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 14, no. 4, pp. 685–696, Apr. 2019.

[13] K. Mishra, R. Sathish, and D. Sheet, "Learning latent temporal connectionism of deep residual visual abstractions for identifying surgical tools in laparoscopy procedures," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR) Workshops*, Jul. 2017, pp. 2233–2240.

[14] H. A. Hajj, M. Lamard, K. Charrière, B. Cochener, and G. Quellec, "Surgical tool detection in cataract surgery videos through multi-image fusion inside a convolutional neural network," in *Proc. 39th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2017, pp. 2002–2005.

[15] S. Wang, Z. Xu, C. Yan, and J. Huang, "Graph convolutional nets for tool presence detection in surgical videos," in *Proc. Int. Conf. Inf. Process. Med. Imag. (IPMI)*, vol. 11492, Jun. 2019, pp. 467–478.

[16] M. Sahu, A. Mukhopadhyay, A. Szengel, and S. Zachow, "Addressing multi-label imbalance problem of surgical tool detection using CNN," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 12, no. 6, pp. 1013–1020, Jun. 2017.

[17] 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.

[18] J. Redmon, S. K. Divvala, R. B. 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.

[19] 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.

[20] 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.

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

[22] 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, Jan. 2020.

[23] Z. Zhao, T. Cai, F. Chang, and X. Cheng, "Real-time surgical instrument detection in robot-assisted surgery using a convolutional neural network cascade," *Healthcare Technol. Lett.*, vol. 6, no. 6, pp. 275–279, Dec. 2019.

[24] Y. Liu, Z. Zhao, F. Chang, and S. Hu, "An anchor-free convolutional neural network for real-time surgical tool detection in robot-assisted surgery," *IEEE Access*, vol. 8, pp. 78193–78201, May 2020.

[25] X. Hu, L. Yu, H. Chen, J. Qin, and P. Heng, "AGNet: Attention-guided network for surgical tool presence detection," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, vol. 10553, Sep. 2017, pp. 186–194.

[26] K. Song, H. Yang, and Z. Yin, "Multi-scale attention deep neural network for fast accurate object detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 10, pp. 2972–2985, Oct. 2019.

[27] W. Li, K. Liu, L. Zhang, and F. Cheng, "Object detection based on an adaptive attention mechanism," *Sci. Rep.*, vol. 10, no. 1, Jul. 2020.

[28] 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.

[29] J. Hu, L. Shen, and G. Sun, "Squeeze-and-Excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.

[30] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1800–1807.

[31] 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.

[32] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Apr. 2015.

[33] Z. Zheng, P. Wang, W. Liu, J. Li, R. Ye, and D. Ren, "Distance-IoU loss: Faster and better learning for bounding box regression," in *Proc. AAAI Conf. Artif. Intell.*, Feb. 2020, pp. 12993–13000.

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

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

[36] T. Lin, P. Goyal, R. B. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 2, pp. 318–327, Feb. 2020.

[37] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, *arXiv:1804.02767*. [Online]. Available: http://arxiv.org/abs/1804.02767

[38] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5693–5703.

[39] C. I. Nwoye, D. Mutter, J. Marescaux, and N. Padoy, "Weakly supervised convolutional LSTM approach for tool tracking in laparoscopic videos," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 14, no. 6, pp. 1059–1067, Jun. 2019.

[40] A. Vardazaryan, D. Mutter, J. Marescaux, and N. Padoy, "Weakly-supervised learning for tool localization in Laparoscopic videos," 2018, *arXiv:1806.05573*. [Online]. Available: http://arxiv.org/abs/1806.05573

[41] 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: http://arxiv.org/abs/1702.03684
[42] G. Yengera, D. Mutter, J. Marescaux, and N. Padoy, "Less is more: Surgical phase recognition with less annotations through self-supervised pretraining of CNN-LSTM networks," 2018, *arXiv:1805.08569*. [Online]. Available: http://arxiv.org/abs/1805.08569
[43] I. Funke, A. Jenke, S. Torge Mees, J. Weitz, S. Speidel, and S. Bodenstedt, "Temporal coherence-based self-supervised learning for Laparoscopic workflow analysis," 2018, *arXiv:1806.06811*. [Online]. Available: http://arxiv.org/abs/1806.06811

PAN SHI received the B.S. degree from the School of Control Science and Engineering, China University of Petroleum, Qingdao, Shandong, China, in 2019. She is currently pursuing the M.S. degree in control science and engineering with Shandong University, Jinan, Shandong.

She is the author of two articles. Her current research interests include deep learning, pattern recognition, and computer vision.

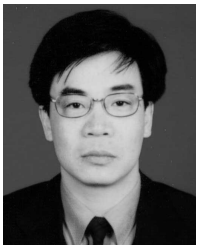
ZIJIAN ZHAO (Member, IEEE) received the M.S. degree in electrical engineering from Shandong University, in 2005, and the Ph.D. degree in image processing and pattern recognition from Shanghai Jiao Tong University, in 2009.

From 2009 to 2010, he was a Postdoctoral Researcher with the University of Oulu, Finland. He was a Research Engineer with TIMC-IMAG, University Joseph Fourier, France, in 2010. He joined Shandong University, as an Associate Professor, in July 2012. He is the author of more than 30 articles. His research interests include computer vision, robot vision, and computer assisted surgery.

SANYUAN HU graduated from Shandong Medical University, in July 1987.

He entered the Second Affiliated Hospital of Shandong Medical University, in 1987. He successively served as a Resident, a Chief Physician, a Deputy Chief Physician, and a Chief Physician. He was the Director of surgery and general surgery of the Qilu Hospital of Shandong University, in 2003. He was the Director of endoscopic diagnosis and treatment technology training base of the Ministry of Health of Qilu Hospital of Shandong University, in 2008. He was appointed as the Vice President of the Qilu Hospital of Shandong University, in 2011. He was hired as the Mount Tai Scholar Distinguished Professor of Shandong, in 2012. He was the President of the Shandong Qianfoshan Hospital (probation period is one year), in 2019. In the field of laparoscopic research, Prof. Sanyuan Hu led the team to win the first prize for scientific and technological progress in Shandong and nine other scientific research awards at provincial and ministerial levels. He published more than 30 SCI papers and applied for two invention patents. He has published 16 monographs, translated works, and five audio-visual teaching materials.

Mr. Hu is a standing or editorial board member of 20 magazines. He was received the title of Shandong's Medical Technical Expert, in 2006, the Highest Award for Endoscopic Medicine by the Chinese Medical Association, in 2005, 2006, and 2008, the Endoscopic Award, in 2008, the fifth Honorary Award for Humanities Medicine, in 2008, and the Honorary Title of the Outstanding Graduate Instructor of Shandong University, in 2009. He is the Editor-in-Chief of the *Journal of Laparoscopic Surgery* and the *Journal of Clinical Practical Surgery*. He is the Deputy Editor-in-Chief of the *China Journal of Endoscopy* and the *China Journal of Modern Medicine*. He was an Outstanding Academic Leader in Shandong's Health System, in 2005, and a Young and Middle-Aged Key Scientific and Technological Talent. Prof. Huang Zhiqiang, an Academician of the Chinese Academy of Engineering, praised him as one of the pioneers in developing laparoscopic surgery in China.



FALIANG CHANG received the B.S. and M.S. degrees from the Automation Department, Shandong University of Technology, Jinan, Shandong, in 1986 and 1989, respectively, and the Ph.D. degree in engineering from Shandong University, Jinan, in 2005.

He began teaching with the Department of Automation, Shandong University of Technology, in 1989, where he was promoted to a Lecturer, in 1992, and an Associate Professor, in 1996. He was a Professor with Shandong University, in 2000. He was a Visiting Scholar with the School of Engineering, State University of Michigan, USA, in 2007. He is the person in charge of the subject of pattern recognition and intelligent systems, the Director of the Engineering System Control Laboratory, Shandong Provincial Key Laboratory, a Member of the Shandong Provincial Informatization Expert Group, and the Deputy Director of the Automation Technology Committee of the Shandong Automation Institute. He is the author of more than 70 articles. He holds seven patents. His research interests include pattern recognition, computer vision, biometric recognition and authentication, and so on.