IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received March 15, 2020, accepted April 18, 2020, date of publication April 23, 2020, date of current version May 8, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2989807

An Anchor-Free Convolutional Neural Network for Real-Time Surgical Tool Detection in Robot-Assisted Surgery

YUYING LIU[1], ZIJIAN ZHAO[1], (Member, IEEE), FALIANG CHANG[1], AND SANYUAN HU[2]
[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 Robot-assisted surgery (RAS), a type of minimally invasive surgery, is used in a variety of clinical surgeries because it has a faster recovery rate and causes less pain. Automatic video analysis of RAS is an active research area, where precise surgical tool detection in real time is an important step. However, most deep learning methods currently employed for surgical tool detection are based on anchor boxes, which results in low detection speeds. In this paper, we propose an anchor-free convolutional neural network (CNN) architecture, a novel frame-by-frame method using a compact stacked hourglass network, which models the surgical tool as a single point: the center point of its bounding box. Our detector eliminates the need to design a set of anchor boxes, and is end-to-end differentiable, simpler, more accurate, and more efficient than anchor-box-based detectors. We believe our method is the first to incorporate the anchor-free idea for surgical tool detection in RAS videos. Experimental results show that our method achieves 98.5% mAP and 100% mAP at 37.0 fps on the ATLAS Dione and Endovis Challenge datasets, respectively, and truly realizes real-time surgical tool detection in RAS videos.

INDEX TERMS Anchor-free, center point, RAS, single-stage, stacked hourglass network, and surgical tool detection.

I. INTRODUCTION

Robot-assisted surgery (RAS) is the latest development in minimally invasive surgical technology. Robotic surgical tools make it easy to perform complex motion tasks during surgery by transforming the surgeon's real-time hand movements and forces acting on the tissue into small-scale movements [1]. Despite its advantages in minimally invasive surgery, the RAS system still has problems, such as a narrow field of view, narrow operating space, and insufficient tactile feedback, which may cause holes in organs and tissues during an operation [2]. Surgical tool detection can help solve these problems by providing the trajectory of a tool to realize surgical navigation. Also, to have real-time information on the motions of a surgical tool can help model poses for real-time automated surgical video analysis [3]–[6], which assists surgeons with automatic report generation, optimized scheduling, and offline video indexing for educational purposes [7]. Hence, in this study, we focus on real-time surgical tool detection in videos.

Many methods have been proposed for surgical tool detection. Image-based methods are becoming more popular, as they rely purely on equipment already in the operating theatre [8]. Deep convolutional neural network (CNN) has been merged into various RAS medical image-based tasks, such as surgical tool detection [9]–[13], tracking [14]–[18], pose estimation [19]–[21], and segmentation [22]–[25]. Single- and two-stage detectors are generally used to detect surgical tools. Two-stage detectors [1], [3] apply a region proposal network (RPN) to generate region proposals before being passed to a final classification and bounding box refinement network; single-stage detectors [13], [26] place anchor boxes densely over an image and generate final box predictions by scoring anchor boxes and refining their coordinates through regression. Both single- and two-stage detectors use anchor boxes extensively, but single-stage detectors are more

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

competitive and efficient than two-stage detectors. However, anchor boxes have two drawbacks [27]. They introduce many hyperparameters that require fine design, and they create a huge imbalance between positive and negative anchor boxes that slows down training. Methods using anchor boxes usually detect surgical tools with high accuracy, but cannot detect them in real time (handling at less than 20 fps). The method proposed by Jin *et al.* [5] and the surgical tool detection method proposed by Twinanda *et al.* [9] only detect the presence of the tool and cannot output the location of the surgical tool. Zhao *et al.* [13] presented a CNN-cascaded surgical tool detection method, which can not achieve end-to-end training and needs to design the output heatmaps carefully. Compared with the work of Zhao *et al.* [13], our method can not only achieve end-to-end training, but also innovatively use a more efficient and compact CNN backbone, which has an accuracy rate that exceeds their work at comparable speeds.

In view of the deficiencies of the various methods mentioned above and the inspiration of CenterNet [28], we propose a single-stage approach to detect surgical tools without anchor boxes. We introduce a compact stacked hourglass network [29] to detect the surgical tool as the center point of its bounding box. We evaluated the performance of the proposed method on the publicly available ATLAS Dione dataset [1] and the EndoVis Challenge dataset [21], and our approach performed better than three state-of-art detection methods with regard to detection accuracy and speed.

Our main contributions are summarized as follows:

(1) We propose an anchor-free CNN architecture for real-time surgical tool detection in RAS. We integrate the lightweight idea (fire module and depthwise separable convolution [30]–[32]) in our architecture so that the accuracy is basically not reduced and the speed of detection of surgical tools is faster.

(2) Our approach distributes the "anchor" based only on location rather than box overlap [33]. Each of our objects has only one positive "anchor," so no NMS is needed, and only local peaks in the keypoint heatmap must be extracted to achieve points to bounding boxes.

(3) We extensively evaluate our proposed surgical tool detection approach on the ATLAS Dione and EndoVis Challenge datasets. For greater accuracy, we manually relabeled the EndoVis Challenge dataset. Our approach demonstrates superior performance over state-of-the-art approaches.

The rest of this paper is organized as follows. Section II introduces our approach, including the network architecture and the loss function for learning. Section III elaborates on the experiments and results. We discuss the effectiveness of our approach and future directions for improvement in Section IV. Finally, our conclusions are drawn in Section V.

II. METHODOLOGY

A. NETWORK ARCHITECTURE

Inspired by [30]–[32], we designed a lightweight hourglass backbone that works better than CenterNet [28]. The new network consists of two hourglass modules, and the residual modules in the traditional hourglass backbone are replaced with the more effective fire modules [30]–[32] to predict the heatmap at the center point of all instances of the surgical tools. Additional details can be found in Figure 1. As we can see, the fire module first uses a 1×1 kernel to squeeze the input channels, which reduces the parameters to accelerate our network. Then, it passes through a mixture of 1×1 and 3×3 kernels to feed the results. To accelerate the training of the network structure, we replace the original 3×3 standard convolution with a 3×3 depthwise separable convolution, as shown in the orange block (Dwise) in Figure 1. Peaks in the heatmap correspond to tool centers [34]. Image features at each peak predict the surgical tool bounding box's height and weight (Figure 2). Inference is performed by a single network forward-pass, without non-maximal suppression (NMS) [35] for post-processing. In general, the depthwise separable convolution splits the ordinary convolution into deep convolution and point-by-point convolution. The advantage of depthwise separable convolution is that the number of parameters and the computational complexity can be greatly reduced with less loss of precision.

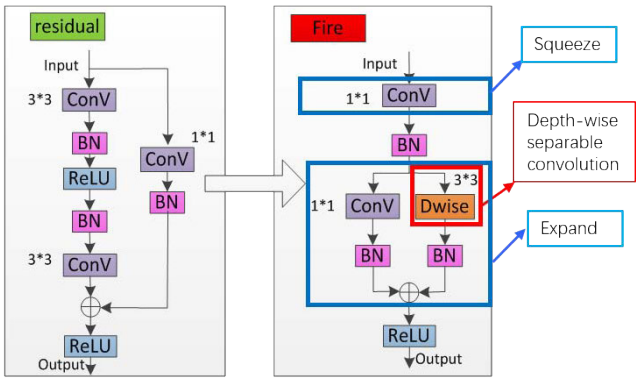


FIGURE 1. The fire module replaces the residual module.

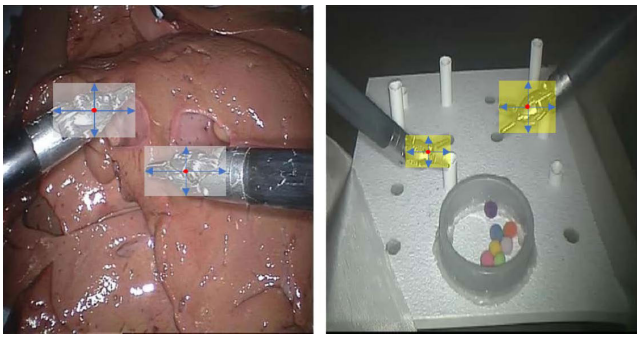


FIGURE 2. We model the surgical tool as the center point of its bounding box. From the keypoint features of the center, the bounding box size and other attributes of the surgical tool can be inferred.

In our architecture (Figure 3), we use a 7×7 convolution module and a residual module to reduce the input image size (512×512) by a factor of four, followed by

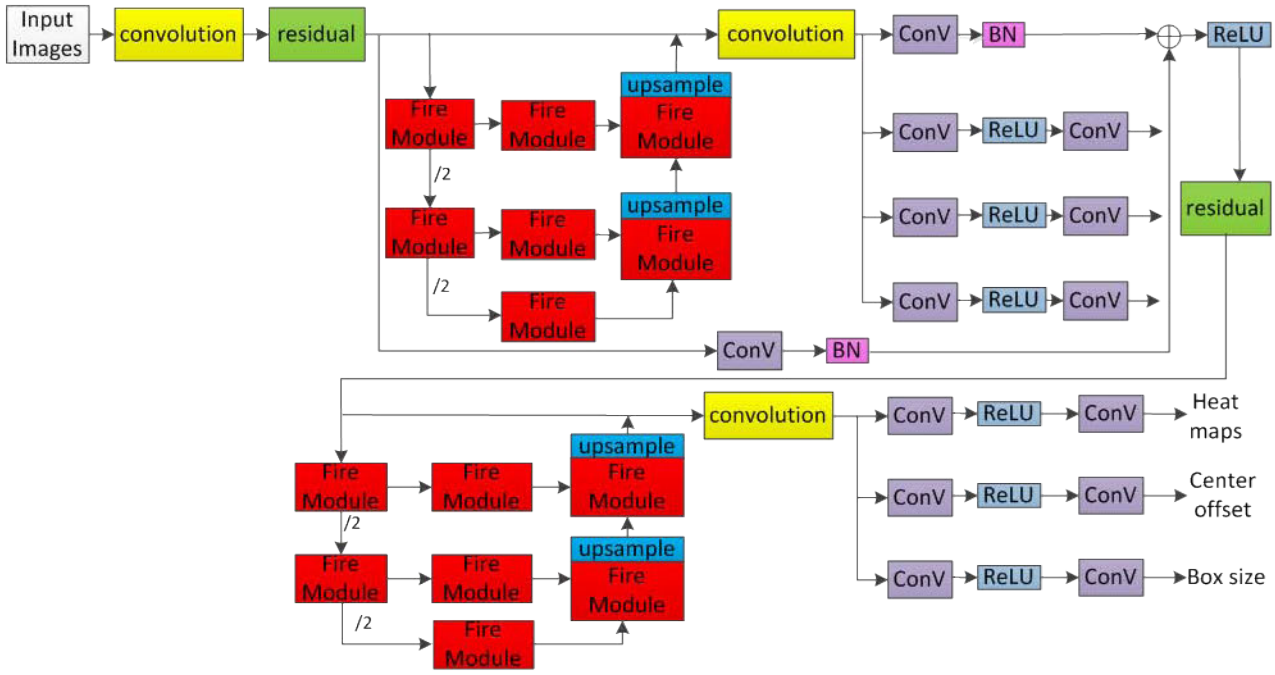


FIGURE 3. Architecture of proposed network.

two hourglass modules. We modified the architecture of the hourglass modules. Each is a symmetric 2-layer downsample and upsample CNN with skip connections, each consisting of a fire module. A fire module followed by nearest neighbor upsampling is applied to upsample the features. There is a fire module in the middle of each hourglass module. We do not use max pooling, but simply use stride 2 to reduce the feature resolution. We increase the number of feature channels along the way (384,512) and reduce feature resolutions two times. We also adopt a 1×1 Conv-BN module to both the input and output of the first hourglass module as intermediate supervision. Inference is performed by a single CNN forward pass, without NMS for post-processing. The features of the stacked lightweight hourglass backbone are then passed through a separate 3×3 convolution, ReLU, and another 1×1 convolution.

B. LOSS FUNCTION FOR LEARNING

We denote an input video frame of width W and height H by $I \epsilon R^{\frac{W}{R} \times \frac{H}{R} \times 3}$. Then, we leverage the lightweight stacked hourglass network to predict the keypoint heatmap $\hat{Y} \epsilon [0,1]^{\frac{W}{R} \times \frac{H}{R} \times C}$, local offset $\hat{O} \epsilon \Re^{\frac{W}{R} \times \frac{H}{R} \times 2}$, and size $\hat{S} \epsilon \Re^{\frac{W}{R} \times \frac{H}{R} \times 2}$, where R is the output stride and C is the number of surgical tool classes. We predict the heatmap at the center point of all instances of the surgical tools. Peaks in the heatmap correspond to object centers. Image features at each peak predict the surgical tool bounding box's height and weight. We train our network following Zhou *et al.* [28]. Focal loss [36] mainly solves the problem of severe imbalance of positive and negative samples in single-stage surgical tool detection. The focal loss function reduces the weight of a large number of simple negative samples in training, which can also be interpreted as a kind of difficult sample mining. The training objective is a penalty-reduced pixel-wise logistic regression with modified focal loss:

$$L_k = \frac{-1}{N} \sum_{xyc} \begin{cases} (1-\hat{Y}_{xyc})^{\alpha} \log(\hat{Y}_{xyc}) & if\ Y_{xyc} = 1 \\ (1-Y_{xyc})^{\beta} (\hat{Y}_{xyc})^{\alpha} & \\ * \log(1-\hat{Y}_{xyc}) & otherwise, \end{cases} \tag{1}$$

where $\alpha = 2$ and $\beta = 4$ [27] are hyperparameters of the focal loss, N is the number of keypoints in image I, Y_{xyc} is a Gaussian kernel, and at the center point $Y_{xyc} = 1$, the diffusion of Y_{xyc} around the center point slowly decreases from 1 to 0. $\hat{Y}_{xyc} = 1$ corresponds to a detected keypoint, while $\hat{Y}_{xyc} = 0$ is the background. The offset is trained with an L1 loss:

$$L_o = \frac{1}{N} \sum_{p} |\hat{O}_{\tilde{p}} - (\frac{p}{R} - \tilde{p})|, \tag{2}$$

where p is ground truth of the keypoint, and $\tilde{p} = \lfloor \frac{p}{R} \rfloor$ is a low-resolution equivalent. We use an L1 loss at the center point:

$$L_s = \frac{1}{N} \sum_{k=1}^{N} |\hat{S}_{p_k} - s_k|, \tag{3}$$

where $s_k = (x_2^{(k)} - x_1^{(k)}, y_2^{(k)} - y_1^{(k)})$ is the object size, $p_k = (\frac{(x_1^{(k)} + x_2^{(k)})}{2}, \frac{(y_1^{(k)} + y_2^{(k)})}{2})$ is the center point location, and $\hat{S}_{p_k}$ is a single size prediction for all tools. We let $(x_1^{(k)}, y_1^{(k)}, x_2^{(k)}, y_2^{(k)})$

be the bounding box of object k with category c_k. The overall training objective is

$$L_{det} = L_k + \lambda_s L_s + \lambda_o L_o. \quad (4)$$

We set $\lambda_{size} = 0.1$ and $\lambda_o = 1$ in our experiments. From center points $\hat{p} = \{(\hat{x}_i, \hat{y}_i)\}_{i=1}^{n}$ to bounding boxes:

$$(\hat{x}_i + \delta\hat{x}_i - \hat{w}_i/2, \hat{y}_i + \delta\hat{y}_i - \hat{h}_i/2, \hat{x}_i + \delta\hat{x}_i + \hat{w}_i/2, \hat{y}_i + \delta\hat{y}_i + \hat{h}_i/2), \quad (5)$$

where $(\delta\hat{x}_i, \delta\hat{y}_i) = \hat{O}_{\hat{x}_i,\hat{y}_i}$ is the offset prediction and $(\hat{w}_i, \hat{h}_i) = \hat{S}_{\hat{x}_i,\hat{y}_i}$ is the size prediction. No output needs post-processing.

The entire model regresses only four values (x, y, w, h) in addition to the surgical tool attributes: the center point (x, y) and (w, h) of the bounding box. First, the heatmap of the picture is obtained through the backbone network, and then the ground truth keypoints are distributed to the heatmap through the Gaussian kernel function $Y_{xyc} = exp(-\frac{(x-\tilde{p}_x)^2+(y-\tilde{p}_y)^2}{2\sigma_p^2})$, where σ_p is a surgical tool size-adaptive standard deviation [27]. According to the peaks on the feature map, 100 peaks that are greater than or equal to 8-connected neighbors values around are selected as the central keypoints for preliminary prediction. Then it is necessary to predict the offset of the center keypoint $\hat{O}_{\hat{x}_i,\hat{y}_i}$ (because there will be deviation after scaling the extracted feature scale). Next, we can predict the size of bounding box $\hat{S}_{\hat{x}_i,\hat{y}_i}$. Finally, we can predict the coordinates of the bounding box by Equation 5. In summary, the procedures of training our network are performed as Algorithm 1 with the steps.

Algorithm 1 Steps of Training Our CNN Network

Input: input images I with labels;
Output: output (x,y,w,h);
Preliminary;
for $I \epsilon R^{\frac{W}{R} \times \frac{H}{R} \times 3}$ **do**
  splatting all ground truth keypoints onto a heatmap by using a Gaussian kernel;
  if then
    using keypoints estimator to predict all center points;
  end
end
Objects as Points;
repeat
  calculating the loss on training via Equation 4;
  updating parameters via back propagation;
until *the iteration satisfies the second stop condition*;
From points to bounding boxes;
for $i=1,2,\cdots,n$ **do**
  calculating the bounding box coordinates via Equation 5;
end
return the center point (x, y) and (w, h) of the bounding box.

III. EXPERIMENTS AND RESULTS

A. DATASET

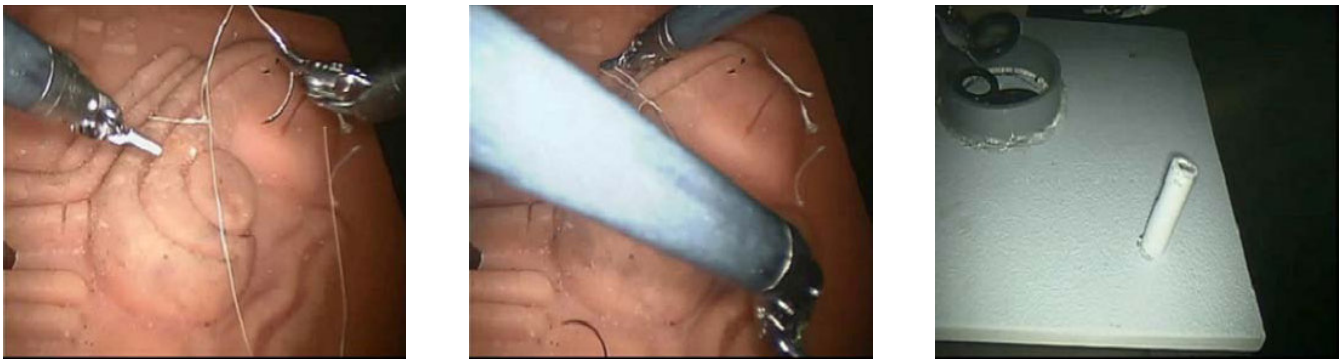
We used the ATLAS Dione dataset [1], consisting of 99 action video clips of ten surgeons from the Roswell Park Cancer Institute (RPCI) (Buffalo, NY) performing six surgical tasks (subject study) on the da Vinci Surgical System (dVSS). The resolution of each frame is 854 × 480 with the surgical tool annotations. Despite being a phantom setting, the ATLAS Dione dataset is challenging, as it has camera movement and zoom, free movement of surgeons, a wide range of expertise levels, background objects with high deformation, and annotations including tools with occlusion, change in pose, and articulation. Figure 4 shows some disturbing factors of the ATLAS Dione dataset. To train our model, we divided the entire set of video clips into two subparts: 90 video clips (20491 frames) for training and the leftover nine video clips (1976 frames) for testing. To validate the extensibility of our architecture, we evaluated our approach on the MICCAI'15 EndosVis Challenge dataset [21], which includes 1083 frames from ex-vivo video sequences of interventions. The resolution of each frame is 720 × 576 with the surgical tool annotations. For greater accuracy, we relabeled the dataset manually. This dataset was separated into a training set (984 frames) and test set (109 frames). The ATLAS Dione dataset is more challenging than the EndosVis Challenge dataset because there are more disturbing factors, such as motion blurring, fast movement, and background changes.

B. EXPERIMENTAL SETTINGS

We implemented the lightweight hourglass networks on the Ubuntu 18.04 LTS operating system using the PyTorch 1.0 framework based on Python 3.6, CUDA 10.1, and CUDNN 7.4. The Titan Xp GPU was used as an accelerator for training. We fixed the input and image resolution to 512 × 512 and 128 × 128, respectively. Before training, we used random scaling, flipping, cropping, and color jittering as data augmentation. The learning rate was initialized at $3.125e^{-5}$ for all layers, and decreased by a factor of 10 at 90 and 120 interations. We trained the networks for 140 epochs. To guarantee the fairness of comparison, we downloaded code and pre-trained models to test run time for each model on the same machine. As for the ATLAS Dione dataset, training on a TITAN GPU, our method uses half of the time required by CenterNet.

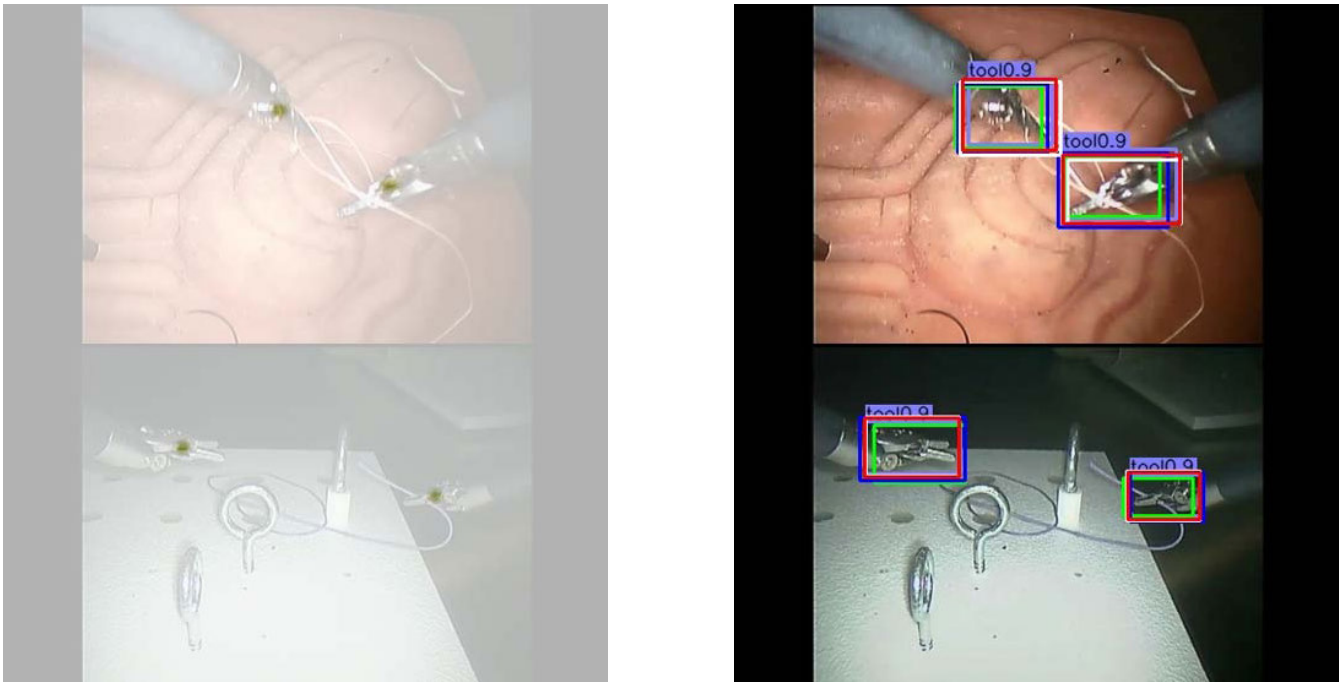
C. RESULTS

We elaborate the surgical tool detection outputs of our method in the video frames, in Figure 5 and Figure 6, where column (a) shows the outputs of the heatmaps of prediction, and the green area overlaid on the original image in column (a) consists of heatmaps of the center point of the surgical tool, and column (b) shows the prediction bounding boxes of different methods. The bounding boxes are the locations and sizes of surgical tools. We also show bounding boxes detected



(a) (b) (c)

FIGURE 4. Disturbing factors of ATLAS Dione dataset. (a) motion blurring; (b) high deformation; (c) annotations including tools with occlusion.

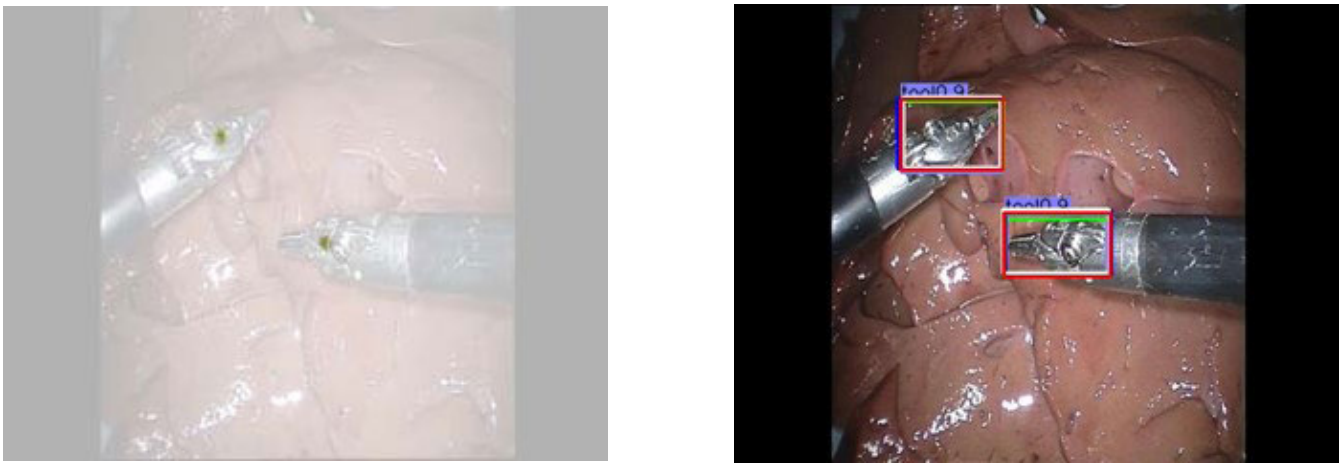


(a) (b)

FIGURE 5. Outputs of proposed network (based on ATLAS Dione dataset). (a) The green area overlaid on the original image consists of heatmaps of the center point of the surgical tool; (b) The prediction bounding boxes of different methods. As shown in the example frames: our method is in purple, Faster RCNN in green, Yolov3 in blue, CenterNet in white, and the ground truth is in red.

by three other state-of-the-art methods as a comparison. Our method is in purple (the probabilities are indicated in the top-left corners of bounding boxes), Faster RCNN is in green, Yolov3 is in blue, CenterNet is in white, and the ground truth is in red. To eliminate the need for multiple anchor boxes [37], our surgical tool detector uses a larger output resolution (output stride of 4) compared to many object detectors (output stride of 16) [38], [39]. To demonstrate the effective generalization capability of our backbone, we performed extensive experiments with five backbones: ResNet-18, ResNet101 [39], DLA-34 [40], Hourglass-104 [29], and ours(lightweight Hourglass). We also modified both ResNets and DLA-34 employing deformable convolution layers and leveraged the Hourglass network [28], [41]. For the DLA-34 and ResNet backbones, the learning rate, learning rate dropped, and training epochs were set the same as our backbone in Section 3.2. For Hourglass-104, we complied with ExtremeNet [42] and used batch size 8 and learning rate $3.125e^{-5}$ for 50 epochs with $10\times$ learning rate dropped at the 40th epoch. After training for 140 epochs, all backbones could converge.

Speed and accuracy tradeoffs for different backbones on the ATLAS Dione and EndosVis Challenge datasets are displayed in Table 1, respectively, from which we can observe the performance of these backbones. We present the mean average precision (mAP) at intersection over union (IoU) threshold 0.5 (this threshold is given by referring to Pascal VOC [43] dataset: if the IoU of the predicted bounding



(a) (b)

FIGURE 6. Outputs of proposed network (based on EndoVis dataset). (a) The green area overlaid on the original image consists of heatmaps of the center point of the surgical tool; (b) The prediction bounding boxes of different methods. As shown in the example frames: our method is in purple, Faster RCNN in green, Yolov3 in blue, CenterNet in white, and the ground truth is in red.

TABLE 1. Speed and accuracy tradeoff for different backbones. The mAP1 and the mAP2 represent the detection mAP on the ATLAS Dione and EndoVis Challenge datasets, respectively.

Backbone	mAP1	mAP2	Time(seconds)	fps
DLA-34 [40]	98.5%	100%	0.033	30.3
ResNet-101 [39]	98.5%	100%	0.030	33.3
ResNet-18 [39]	98.4%	100%	0.014	71.4
Hourglass-104 [29]	98.5%	100%	0.060	16.7
ours(lightweight Hourglass)	98.5%	100%	0.027	37.0

box and the ground truth were greater than 0.5, then we considered the surgical tools to be successfully detected in a frame.). IoU is the ratio of the intersection and union of the prediction bounding box and ground truth, and is also referred to as the Jaccard index. We set different thresholds (0.5, 0.75, 0.95), comprehensively compare the experimental results of different backbones, and found that our backbone is the best in the balance of speed and accuracy. We also notice that the performance growth rate tends to be slower with the increase of ResNet deep, and our lightweight hourglass backbone works better than the Hourglass-104 backbone. The superior performance on both the ATLAS Dione and EndosVis Challenge datasets verifies the extensibility of our approach.

To prove the value of our tools detection method, we compared our method to three state-of-the-art detection methods on the ATLAS Dione and EndosVis Challenge datasets. We selected two anchor-based methods, Faster RCNN [44] and Yolov3 (Darknet-53) [45], and one anchor-free method, CenterNet (Hourglass-104) [28]. As described in Table 2, the mAP1 and the mAP2 represent the detection mAP on the ATLAS Dione and EndoVis Challenge dataset, respectively. Our method achieved a mAP of 98.5% for the ATLAS Dione dataset, and a mAP of 100% for the surgical tool detection of the EndoVis Challenge dataset. We compared the speed of our method with those of the other three state-of-the-art detection methods on two datasets, as shown in Table2. Our method had real-time performance at a speed of 0.027 seconds (over 20 fps), which demonstrates its potential for online surgical tool detection.

To more comprehensively reveal the advantages of our method, we also evaluated our method by the distance evaluation method. If the distance between the center of the predicted bounding box and the center of the ground-truth bounding box is less than the threshold in the image coordinates, then the surgical tool is considered to have been correctly detected. The experimental results are shown in Figures 7 and 8. CenterNet and our method achieved competitive performance on the ATLAS Dione dataset at the cost of lower 2× detection speed. Our method shows the best performance on the Endovis Challenge dataset.

IV. DISCUSSION

Automatically detecting tool location from videos plays a important role of the development of the RAS. Based on Table 2, we can see that our method is more accurate than the other three methods. In particular, experiments on the ATLAS Dione dataset demonstrate the superior performance of our method, which exceeds Fast-Rcnn and Yolov3 by a

TABLE 2. Comparison of speed and accuracy of different methods. The mAP1 and the mAP2 represent the detection mAP on the ATLAS Dione and EndoVis Challenge datasets, respectively.

Method	mAP1	mAP2	Time(seconds)	fps
Faster RCNN [43]	90.31%	100%	0.062	16.1
Yolov3 [44]	90.97%	99.07%	0.034	29.4
CenterNet [28]	98.50%	100%	0.060	16.7
Ours	98.50%	100%	0.027	37.0

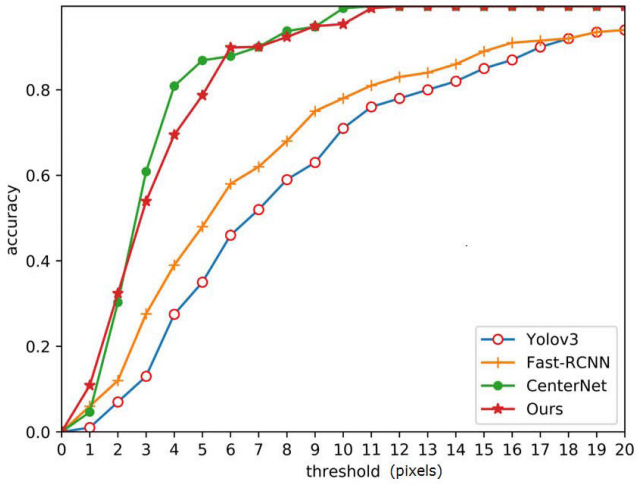


FIGURE 7. Detection accuracy of surgical tool based on distance evaluation (ATLAS Dione dataset).

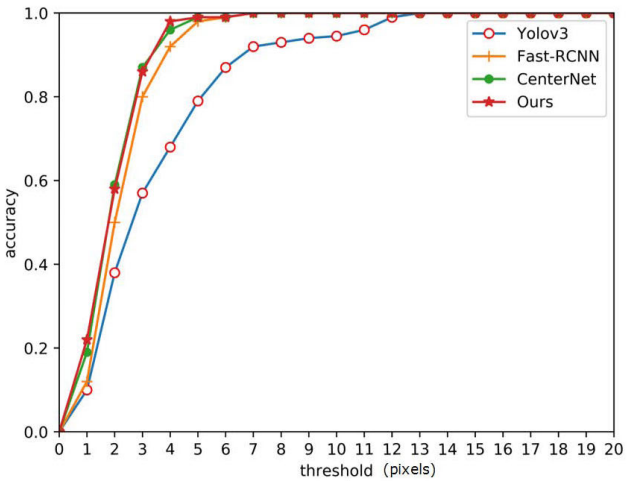


FIGURE 8. Detection accuracy of surgical tool based on distance evaluation (EndosVis Challenge dataset).

significant margin. On the other hand, our approach shows superior speed to Faster RCNN and CenterNet, is competitive with Yolov3, and is two times faster than CenterNet. The considerable improvement of CenterNet (based on Hourglass-104) is largely attributed to the replacement of the residual modules in the hourglass backbone with the more effective fire modules and the utilization of depthwise separable convolution.

Our method achieved good results, but there are potential limitations. For example, if the center points of two surgical tools just overlap, our method can only predict one of them. The lack of large datasets (with tool annotations), the need to improve the speed, and the high training costs are other limitations of our study. Based on the above considerations, the following ideas should be investigated. With regard to the lack of datasets, our future work will pay more attention to extending the detection of weakly supervised surgical tools. To increase the speed, we will try to leverage temporal information (using a long short-term memory network to extract temporal information) for the surgical tool detection task. We hope to employ time information to realize the detection task of surgical tools with a faster speed and greater accuracy.

V. CONCLUSION

We introduced an anchor-free CNN architecture and a frame-by-frame method using a lightweight stacked hourglass network to predict the heatmap at the center point of a surgical tool for real-time surgical tool detection in robot-assisted surgery. Peaks in the heatmap correspond to tool centers. Image features at each peak predict a tool's bounding box size. Our detector eliminates the need to design a set of anchor boxes, and is end-to-end differentiable, simpler, more accurate, and more efficient than corresponding anchor box-based detectors. We believe our method is the first to incorporate the anchor-free idea for surgical tool detection in RAS videos. Our method has achieved good accuracy and speed to realize real-time surgical tool detection in RAS videos.

REFERENCES

[1] 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.

[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] 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," 2018, *arXiv:1802.08774*. [Online]. Available: http://arxiv.org/abs/1802.08774

[4] 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.

[5] Y. Jin, H. Li, Q. Dou, H. Chen, J. Qin, C.-W. Fu, and P.-A. Heng, "Multi-task recurrent convolutional network with correlation loss for surgical video analysis," 2019, *arXiv:1907.06099*. [Online]. Available: http://arxiv.org/abs/1907.06099

[6] G. Yengera, D. Mutter, J. Marescaux, and N. Padoy, "Less is more: Surgical phase recognition with less annotations through self-supervised pre-training of CNN-LSTM networks," 2018, *arXiv:1805.08569*. [Online]. Available: http://arxiv.org/abs/1805.08569

[7] O. Zisimopoulos, E. Flouty, I. Luengo, P. Giataganas, J. Nehme, A. Chow, and D. Stoyanov, "Deepphase: Surgical phase recognition in CATARACTS videos," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, Granada, Spain, Sep. 2018, pp. 265–272.

[8] 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.

[9] 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.

[10] 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. Workshops (CVPRW)*, Honolulu, HI, USA, Jul. 2017, pp. 2233–2240.

[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] 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
[13] 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.
[14] 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, Jun. 2016.
[15] 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, pp. 33–38, 2016.
[16] Z. Zhao, S. Voros, Z. Chen, and X. Cheng, "Surgical tool tracking based on two CNNs: From coarse to fine," *J. Eng.*, vol. 2019, no. 14, pp. 467–472, Feb. 2019.
[17] 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.
[18] 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.
[19] N. Rieke, D. J. Tan, F. Tombari, J. P. Vizcaino, C. A. D. S. Filippo, A. Eslami, and N. Navab, "Real-time online adaption for robust instrument tracking and pose estimation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2016, pp. 422–430.
[20] 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.*, 2017, pp. 505–513.
[21] 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.
[22] L. C. Garcia-Peraza-Herrera, W. Li, L. Fidon, C. Gruijthuijsen, A. Devreker, G. Attilakos, J. Deprest, E. van der Poorten, D. Stoyanov, T. Vercauteren, and S. Ourselin, "ToolNet: Holistically-nested real-time segmentation of robotic surgical tools," 2017, *arXiv:1706.08126*. [Online]. Available: http://arxiv.org/abs/1706.08126
[23] F. Qin, Y. Li, Y.-H. Su, D. Xu, and B. Hannaford, "Surgical instrument segmentation for endoscopic vision with data fusion of CNN prediction and kinematic pose," in *Proc. Int. Conf. Robot. Autom. (ICRA)*, May 2019, pp. 9821–9827.
[24] C. Gao, M. Unberath, R. Taylor, and M. Armand, "Localizing dexterous surgical tools in X-ray for image-based navigation," 2019, *arXiv:1901.06672*. [Online]. Available: http://arxiv.org/abs/1901.06672
[25] I. Laina, N. Rieke, C. Rupprecht, J. Page Vizcaíno, A. Eslami, F. Tombari, and N. Navab, "Concurrent segmentation and localization for tracking of surgical instruments," 2017, *arXiv:1703.10701*. [Online]. Available: http://arxiv.org/abs/1703.10701
[26] 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)*, Jeju Island, South Korea, Jul. 2017, pp. 1756–1759.
[27] H. Law and J. Deng, "Cornernet: Detecting objects as paired keypoints," in *Proc. Eur. Conf. Comput. Vis.*, 2019, pp. 765–781.
[28] X. Zhou, D. Wang, and P. Krähenbühl, "Objects as points," 2019, *arXiv:1904.07850*. [Online]. Available: http://arxiv.org/abs/1904.07850
[29] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," 2016, *arXiv:1603.06937*. [Online]. Available: http://arxiv.org/abs/1603.06937
[30] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "MobileNets: Efficient convolutional neural networks for mobile vision applications," 2017, *arXiv:1704.04861*. [Online]. Available: http://arxiv.org/abs/1704.04861
[31] F. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, "SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5MB model size," 2017, *arXiv:1602.07360*. [Online]. Available: https://arxiv.org/abs/1602.07360
[32] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 1800–1807.
[33] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Santiago, Chile, Dec. 2015, pp. 1440–1448.
[34] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2D pose estimation using part affinity fields," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 1302–1310.
[35] N. Bodla, B. Singh, R. Chellappa, and L. S. Davis, "Soft-NMS–improving object detection with one line of code," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Venice, Italy, Oct. 2017, pp. 5562–5570.
[36] T. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Venice, Italy, Oct. 2017, pp. 2999–3007.
[37] B. Singh and L. S. Davis, "An analysis of scale invariance in object detection-SNIP," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit., (CVPR)*, Salt Lake City, UT, USA, Jun. 2018, pp. 3578–3587.
[38] K. He, G. Gkioxari, P. Dollar, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Venice, Italy, Oct. 2017, pp. 2980–2988.
[39] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 770–778.
[40] F. Yu, D. Wang, E. Shelhamer, and T. Darrell, "Deep layer aggregation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 2403–2412.
[41] 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)*, Venice, Italy, Oct. 2017, pp. 764–773.
[42] X. Zhou, J. Zhuo, and P. Krahenbuhl, "Bottom-up object detection by grouping extreme and center points," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA, Jun. 2019, pp. 850–859.
[43] 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.
[44] S. Ren, K. He, R. B. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, Montreal, QC, Canada, Dec. 2015, pp. 91–99.
[45] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, *arXiv:1804.02767*. [Online]. Available: http://arxiv.org/abs/1804.02767

YUYING LIU received the B.S. degree from the Automation Department, Henan University of Technology, Zhengzhou, Henan, in 2018. She is currently pursuing the M.S. degree in control science and engineering with Shandong University, Jinan, Shandong, China.

She is the author of two articles. Her research interests include 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.

He was a Postdoctoral Researcher with the University of Oulu, Finland, from 2009 to 2010. In 2010, he was a Research Engineer with TIMC-IMAG, University Joseph Fourier, France. In July 2012, he joined Shandong University as an Associate Professor. He is the author of more than 20 articles. His research interests include computer vision, robot vision, and computer assisted surgery.

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 has been a Professor with Shandong University, in 2000. In 2007, he was a Visiting Scholar with the State University of Michigan's School of Engineering, USA. He is the person in charge of the subject of pattern recognition and intelligent systems, the Director of the Engineering System Control Laboratory of the 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, and biometric recognition and authentication.

SANYUAN HU received the degree from Shandong Medical University, in July 1987.

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

Prof. Hu was an Outstanding Academic Leader in Shandong's health system, in 2005, the young and middle-aged key scientific and technological talent, and was awarded the title of Shandong's Medical Technical Expert, in 2006. He was awarded 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 Outstanding Graduate Instructor of Shandong University, in 2009. He is the Academician of the Chinese Academy of Engineering, praised him as one of the pioneers in developing laparoscopic surgery in China. He is also the Editor-in-Chief of the *Journal of Laparoscopic Surgery* and the *Journal of Clinical Practical Surgery*, the Deputy Editor-in-Chief of the *China Journal of Endoscopy* and the *China Journal of Modern Medicine*, and standing or editorial board member of 20 magazines.