IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 3 February 2023, accepted 19 February 2023, date of publication 28 February 2023, date of current version 28 March 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3250400

RESEARCH ARTICLE

Anchor-Free Feature Aggregation Network for Instrument Detection in Endoscopic Surgery

GUANZHI DING[1], XIUSHUN ZHAO[2], CAI PENG[2], LI LI[3], JING GUO[2,4], DEPEI LI[5], AND XIAOBING JIANG[5]

[1]School of Computer Science and Technology, Guangdong University of Technology, Guangzhou, Guangdong 510006, China
[2]School of Automation, Guangdong University of Technology, Guangzhou, Guangdong 510006, China
[3]Baiyun Power Group Co Ltd, Guangzhou, Guangdong 510460, China
[4]Smart Medical Innovation Technology Center, Guangdong University of Technology, Guangzhou, Guangdong 510006, China
[5]Department of Neurosurgery/ Neuro-onco logy, Sun Yat-sen University Cancer Center, State Key Laboratory of Oncology in South China, Collaborative Innovation Center for Cancer Medicine, Guangzhou, Guangdong 510060, China

Corresponding authors: Jing Guo (toguojing@gmail.com) and Xiaobing Jiang (jiangxiaob1@sysucc.org.cn)

This work was supported by Guangdong Basic and Applied Basic Research Foundation (grant number 2022A1515012430 to Xiaobing Jiang).

This work involved human subjects or animals in its research. The author(s) confirm(s) that all human/animal subject research procedures and protocols are exempt from review board approval.

ABSTRACT Endoscopic endonasal approach has been widely used for removing various sellae tumors including pituitary adenomas, meningiomas, etc. While, performing these surgeries in such a narrow space with different instruments remains a challenge for surgeons, due to the limited field of view, varying illumination, and occlusion of instruments during the operation. Thus, a proper surgical instrument detection method that can provide classification and location information of the operated surgical instrument is critical for surgeons to understand the surgical scenarios and enhance the safety of the clinical operation. To this end, we propose an anchor-free feature aggregation network (AFA-Net) to improve the detection precision of surgical instruments from the endoscopic operation view field. The proposed method utilizes the improved feature pyramid network (FPN) with the depthwise separable convolution and a weighted feature aggregation module to enhance the feature information of the operated surgical instruments. Based on the anchor-free method, a weighted heatmap aggregation module is used to detect surgical instruments. Experimental studies on a public dataset Cholec80 and an intraoperative dataset from a local hospital are conducted, and the detection performance is assessed by the mean precision (AP) and average recall (AR). From both datasets and comparisons, the proposed method achieves 74.1% AP, 67.0% AR and 73.6% AP, 66.7% AR, respectively, which show significant advantages over five mainstream methods in terms of detection performance.

INDEX TERMS Endoscopic surgery, surgical instrument detection, convolutional neural network, feature pyramid network, anchor-free detection.

I. INTRODUCTION

Endoscopic endonasal approach (EEA) represents a milestone route to the ventral skull base [1]. Recently, with the significant advances of endoscopic techniques and related instruments, EEA has become the preferred approach for the majority of the sellae lesions, including pituitary adenomas, craniopharyngiomas and clivus chordomas [2], [3]. In contrast to the transcranial and transsphenoidal microscopic approaches, EEA has considerable advantages, such as a superior close-up view of the skull base, multiple angles working field, as well as an increased panoramic vision inside the surgical area [3]. However, some limitations of the endoscopic approach can also be noticed [4]. As the operative space is quite narrow, visual occlusion frequently occurs [5], [6], which may lead to unnecessary trauma to adjunct organs and tissues, such as the artery and optic nerve, etc.

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

In recent years, deep learning methods have been utilized with promising results in image analysis tasks, such as recognition [7], [8], detection [9], [10], [11], segmentation [12], [13], [14], [15], [16], and pose estimation [17]. Many research works on surgical instrument detection have been observed to be applied for identifying and locating surgical instruments intraoperatively, thus achieving surgical instrument tracking [5], [6], [18], surgical skill assessment [19] and surgical instrument monitoring [20]. The detection task consisting of classification and localization of instruments provides valuable information to surgeons in complex surgical scenarios, such as discovering the locational relationship between instruments and tissues, which benefits the operation of various surgical instruments in surgery [21], [22]. For example, instrument detection can remind surgeons to stop the surgery and adjust the locations of surgical instruments which are out of the visible surface or blocked. However, many challenges still need to be addressed to achieve high-precision instrument detection in endoscopic surgery, such as the blood on the instrument surface, as well as the collision of the operated surgical instruments during the surgery, let along the poor feedback on endoscopic images during the surgery which might be covered by the blur, shadow, and even reflections, further decreases the detection precision [13], [20]. Therefore, designing the effective method of instrument detection in endoscopic images to improve the precision and the safety of endoscopic surgery has become a crucial research topic [9], [20], [23], [24], [25].

To address the challenges of instrument detection in the complex environment of endoscopic surgery, there are two main strategies: anchor-based and anchor-free methods. The anchor-based methods include two-stage detectors and one-stage detectors, in which the two-stage detectors generate region proposals and then predict the classification and location of objects on the candidate regions [26], [27], [28]. The one-stage detectors directly produce classification and location through regression, such as the YOLO series [29], [30], [31], [32], which has faster detection speed than two-stage detectors. However, the anchor-based methods always require many parameters calculation to generate the anchor boxes, which makes them not applicable for surgical activities as the fast reflection of surgical state and the precise adjustment of instruments should be given to avoid any possible trauma [19]. Therefore, the anchor-free methods have received increasing attention [23], [33], [34], which is beneficial to improve both the precision and efficiency of object detection and is more competitive and efficient than the anchor-based methods in the development of endoscopic surgical instrument detection from clinical scenarios.

In this paper, motivated to improve the detection precision of surgical instruments in endoscopic surgery, we proposed an anchor-free feature aggregation network (AFA-Net) inspired by CenterNet. Specifically, our AFA-Net utilizes an improved feature pyramid network with depthwise separable convolution in the process of lateral connections to generate multi-scale features, which provides more subtle feature information and enhances the ability of object recognition. Then a feature aggregation module with a top-down weighted feature fusion pathway is utilized to further refine edge features, which reduces the loss of feature information during the feature fusion process. Moreover, our AFA-Net enhances the ability to locate the center point of object in the heatmap by an aggregation module with the top-down weighted heatmap fusion pathway, thus achieving accurate instrument detection in endoscopic surgical scenarios.

The main contributions of this paper are summarized as follows:

1) We propose an improved feature pyramid network and a weighted feature aggregation module to extract multi-scale features and enhance the feature representation.
2) We design a weighted heatmap aggregation module to enhance the ability to locate the center point of object and achieve accurate instrument location.
3) Experiments are conducted on a public dataset Cholec80 [7] and an endoscopic transsphenoidal dataset provided by Sun Yat-sen University Cancer Center. The proposed AFA-Net achieves superior performance in comparison to the five mainstream methods in terms of detection.

The rest of this paper is organized as follows: Section II presents the related works. Section III describes the proposed method including network architecture and loss function. Section IV provides the experimental results and discusses the effectiveness of our method. And the conclusion of the paper is shown in Section V.

II. RELATED WORKS

In this section, a review of surgical instrument detection, anchor-free detection, and multi-scale feature fusion is presented.

A. SURGICAL INSTRUMENT DETECTION

Most traditional methods of instrument detection typically have relied on low-level visual features, such as color and shape, for a simple computer vision task of color segmentation or thresholding [35], [36]. As deep learning approaches using CNNs have been increasing in popularity, many methods are proposed to achieve surgical instrument detection with the extraction of high-level features [7], [19], [20], [37], [38]. Andru et al. [7] proposed EndoNet which first adopts CNN trained with labeled surgical images to achieve instrument detection and recognition from the video of endoscopic surgery. Following this work, more neural network architectures are introduced in instrument detection tasks from surgical scenarios, including an attention-guided CNN for real-time instrument detection in minimally invasive surgery [9], a multi-level feature aggregation network for multi-instrument identification [20], an arrow object bounding box network based on YOLO for identification and localization of instrument tips [38], an anchor-free CNN for instrument detection in robot-assisted surgery [23].

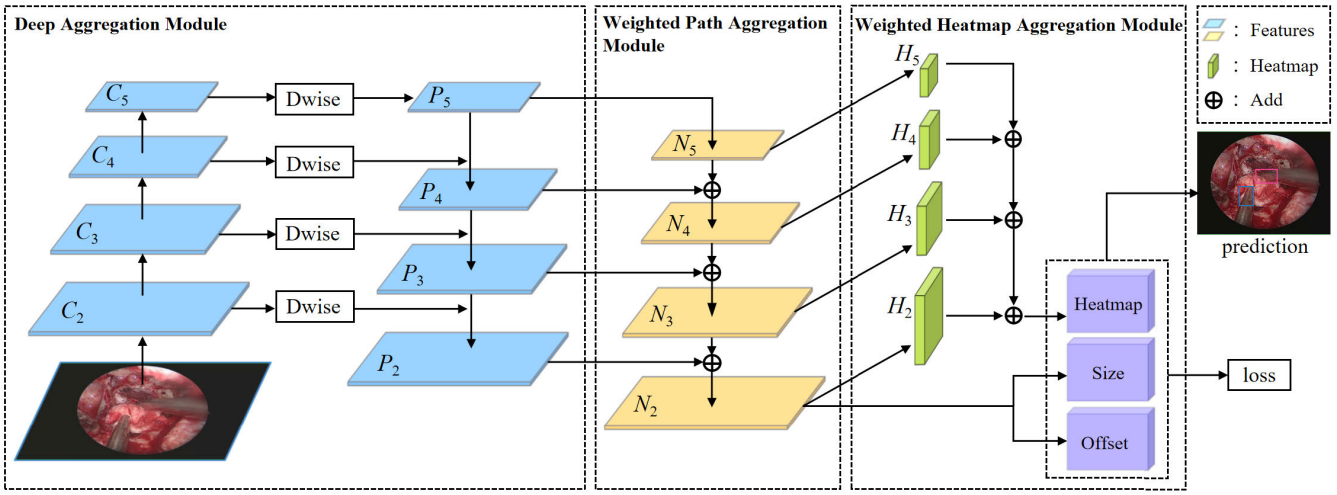

FIGURE 1. The overall framework of the anchor-free feature aggregation network for endoscopic surgical instrument detection. The three modules deep aggregation module (DAM), weighted path aggregation module (WPAM), and weighted heatmap aggregation module (WHAM) in the dotted line rectangle correspond to each part of our network and finally output the prediction and loss.

Furthermore, weakly supervised methods for surgical instrument detection are proposed to utilize the images without annotations of bounding boxes, which extends the approach of network training [39].

B. ANCHOR-FREE DETECTION

Different from the anchor-based method using anchor boxes to detect objects, the anchor-free method interprets the detection task as the prediction of the different types or numbers of keypoint, such as edge-based detection [40], [41] and center-based detection [42], [43], [44], [45]. In CornerNet [40], Law et al. detect the object bounding box as the keypoints in the top-left corner and the bottom-right corner and then combine them together. The reason for choosing corner points on edges is that the corner ones are more conducive for training when compared with the center ones. However, only predicting the corner location of an object somehow might fail to make full use of the information in the frame, which could easily lead to false object detection. Therefore, CenterNet [43] is proposed to directly detect the central area and size information of an object, which is conducive to accelerating prediction. In addition, the method of keypoint triplets is introduced to combine two corners and the center points, which further improves both precision and recall of detection [43].

C. MULTI-SCALE FEATURE FUSION

To solve the multi-scale issue of objects, Lin et al. [46] utilize the feature maps of different scales to form a feature pyramid network (FPN), and then adopt a top-down architecture to achieve multi-scale feature fusion. After that, most studies of feature fusion focus on improving FPN [27], [47], [48], [49], [50], [51]. The traditional feature pyramid network is prone to semantic dilution when directly fusing multi-scale features. Therefore, Wang et al. [47] propose a novel interconnected feature pyramid network (IFPN), which

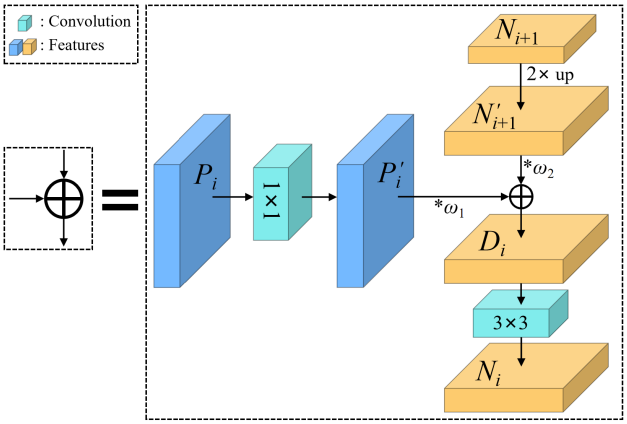

FIGURE 2. Illustration of the building block of weighted path aggregation module (WPAM). ω_1 and ω_2 are the weights obtained from the gaussian kernel [50].

selects attention features through the attention mechanism. Liang et al. propose a twin feature pyramid network (TFPN) to create a double pyramid structure fusion feature, which can effectively reduce the increase of noise [49]. Aiming to alleviate potential feature dislocation and loss of details, the weighted feature pyramid network (WFPN) proposed by Li et al. [50] fuses the feature maps in a weighted way. In our study, WFPN directly inspires our fusion method of features and heatmaps to detect multi-scale objects in endoscopic images.

III. METHODS

A. NETWORK OVERVIEW

The overall framework of the proposed network is illustrated in Figure 1, which consists of three modules as follows: deep aggregation module (DAM), weighted path aggregation module (WPAM), and weighted heatmap aggregation module (WHAM). The original images are the input of the proposed

network. In the DAM, the improved FPN with depthwise separable convolution is introduced to extract more subtle multi-scale features of images. Then WPAM is proposed to fuse and further enhance the extracted features information through a top-down weighted feature fusion pathway. Inspired by the anchor-free method of CenterNet, WHAM uses features to generate the heatmap, size and offset of objects. Then the heatmaps are fused through a top-down weighted heatmap fusion pathway to adjust the peaks in the heatmap. The fused heatmap, size, and offset are combined to calculate the loss and output the prediction.

B. DEEP AGGREGATION MODULE

Feature pyramid network [46] consists of a bottom-up pathway for feature extraction and a top-down pathway for upsampling feature maps. Between them, lateral connections are used to merge feature maps with high-resolution information and feature maps with high-level semantic information from top to bottom. The traditional FPN [46] fuses low-level and high-level feature information, while it is easy to lose the feature information of objects. Therefore, the proposed deep aggregation module to improve FPN, as shown in Figure 1, introduces depthwise separable convolution (Dwise) [52] into FPN to provide more subtle details of instruments for the feature fusion process.

As shown in Figure 1, the bottom-up pathway of FPN is composed of reference feature maps computed from each layer of the pyramid network with a scaling step of 2. We use ResNet-101 as the backbone, and select the outputs of the last residual block from the conv2, conv3, conv4, and conv5 layers as referenced feature maps, denoted as $\{C_2, C_3, C_4, C_5\}$. The depthwise separable convolution consists of depthwise convolution and pointwise convolution [52]. Different from FPN, in the lateral connections, we replace the 1×1 convolution layers after $\{C_2, C_3, C_4, C_5\}$ with 1×1 depthwise separable convolution layers. The advantage is that convolution operation is performed independently on each channel of feature maps, which is conducive to extracting subtle features and has less computational complexity. Then we upsample the feature maps at the higher pyramid level by a factor of 2 to obtain new feature maps with the same spatial size as the corresponding feature maps at the lower pyramid level. Then, along the top-down pathway direction, feature maps are merged by element-wise addition. Finally, Each merged feature map is processed by a 3×3 convolutional layer to obtain the final feature maps, denoted as $\{P_2, P_3, P_4, P_5\}$.

C. WEIGHTED PATH AGGREGATION MODULE

Different methods have been proposed to enhance feature information. For example, the path aggregation network enhances the entire feature hierarchy through bottom-up pathway enhancement [27]. The weighted feature pyramid network adopts a Gaussian kernel function to assign different weights to different image feature information [50]. Inspired by the above methods, we propose a weighted path aggregation module to adjust and enhance the edge features of the object in the feature map after DAM, as shown in Figure 1.

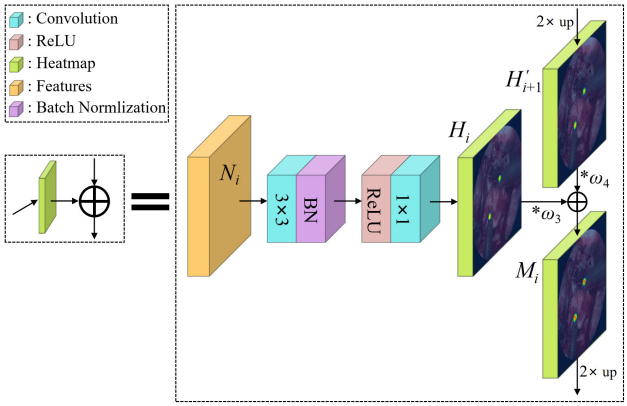

FIGURE 3. Illustration of the building block of weighted heatmap aggregation module (WHAM). To more intuitively display the changes in the process of heatmap fusion, the heatmaps are demonstrated in the figure instead of the heatmap features. ω_3 and ω_4 are the weights obtained from the gaussian kernel [50].

Typically, the feature map at the higher level reflects the overall structure of objects, while the feature map at the lower level is more fine-grained with detailed information on the edge. The direct fusion of the top-down feature of the FPN might result in the loss of the edge feature of object in feature map at the lower level. We use an additional top-down pathway to perform the weighted fusion of feature maps at different levels, which is conducive to adjusting the previously obtained features and further refining the edge information [50]. Specifically, feature map P_5 passes through a 1×1 convolutional layer for adjusting the feature channels to obtain feature map N_5. As shown in Figure 2, in each building block, the feature map P_i passes through a 1×1 convolutional layer to obtain P'_i, and the feature map N_{i+1} is upsampled by a factor of 2 to obtain N'_{i+1}. Following the weight assignment method in [50], we use the difference between P'_4 and N'_5 to calculate the weight value based on a two-dimensional gaussian kernel function as below,

$$\text{gauss}(\sigma, r, c) = e^{-\frac{\left(r-\frac{H-1}{2}\right)^2+\left(c-\frac{W-1}{2}\right)^2}{2\sigma^2}} \tag{1}$$

where H and W represent the height and width of the feature map respectively, while r and c are the coordinates of each point, and σ is the width parameter of the function and controls the radial range of the function. The reason for using P'_4 and N'_5 to calculate weights is that there is less difference in the object feature of the feature map at the top level [50]. We assign the larger weights to the feature map containing more information to calculate the weights ω_1 and ω_2. In each lateral connection, each element of the feature maps P'_i and N'_{i+1} is multiplied by the weights ω_1 and ω_2, respectively. Then the weighted P'_i and N'_{i+1} are merged by element-wise addition to obtain D_i. As shown in Eq. 2, the feature maps $\{P'_4, P'_3, P'_2\}$ and $\{N'_4, N'_3, N'_2\}$ of different levels are merged from top to bottom to obtain new feature maps $\{D_4, D_3, D_2\}$.

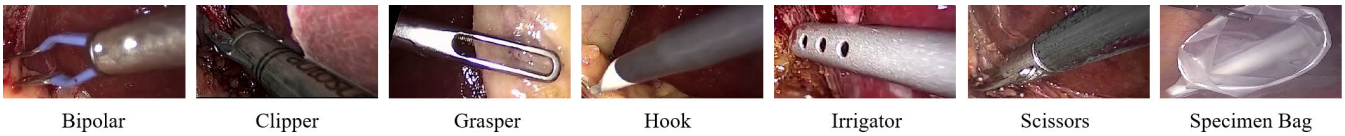

FIGURE 4. The seven surgical instruments used in the cholec80-sub-tool-locations dataset.

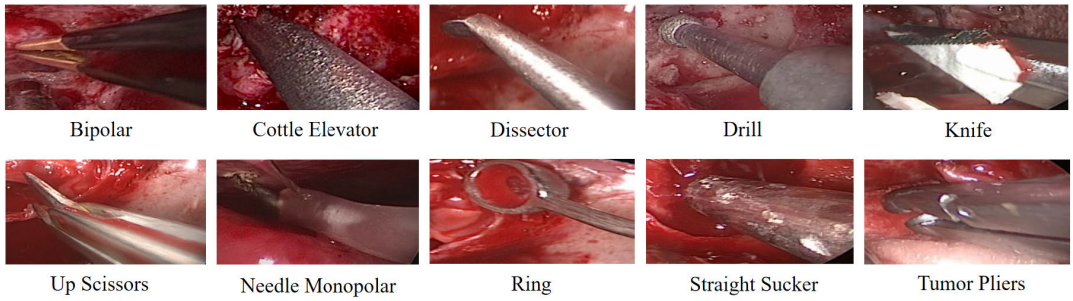

FIGURE 5. The ten surgical instruments used in the sun21-tool-locations dataset.

Then $\{D_4, D_3, D_2\}$ is processed by 3×3 convolutional layers to obtain the final feature maps, denoted as $\{N_4, N_3, N_2\}$, and finally we obtain $\{N_5, N_4, N_3, N_2\}$.

$$\{D_4, D_3, D_2\} = \omega_1 * \left\{P'_4, P'_3, P'_2\right\} + \omega_2 * \left\{N'_4, N'_3, N'_2\right\} \tag{2}$$

D. WEIGHTED HEATMAP AGGREGATION MODULE

The anchor-free method of CenterNet [43] adopts heatmap regression to obtain the center point to represent the object, and then uses the features around the center point to predict the offset and size of the object for locating object. Therefore, the accurate prediction of the center point location plays a critical role in the detection of instruments. We propose the weighted heatmap aggregation module to enhance the ability to locate the center point of object through the weighted fusion of heatmap.

$$M_i = \omega_3 * H_i + \omega_4 * H'_{i+1} \tag{3}$$

In CenterNet, peaks in the heatmap are recognized as the object centers [43]. The feature map N_2 has the more subtle feature of the instrument, which is conducive to generating the accurate heatmap of object center. We add a top-down weighted heatmap fusion pathway to adjust the peaks in the heatmap to further improve the prediction precision of object center. As shown in Figure 3, in each building block, feature map N_i is processed through a 3×3 convolutional layer, a batch normalization layer, a ReLU activation layer and a 1×1 convolutional layer to generate the heatmap H_i. The heatmap at the higher level is upsampled by a factor of 2 to obtain H'_{i+1}. Similarly, inspired by the weight assignment method in [50], we use the difference between heatmap H_4 and the heatmap H'_5 to calculate weight value based on Eq. 1. And then we assign the larger weights to the heatmap containing higher peak information to calculate the weights ω_3 and ω_4 from the H_4 and the H'_5. As shown in Eq. 3, in each lateral connection, each element of the heatmap H_i and H'_{i+1} is multiplied by the weights ω_3 and ω_4, respectively. Then the weighted H_i and H'_{i+1} are merged by element-wise addition to obtain M_i. After the top-down weighted heatmap fusion process, the final merged heatmap M_2 is used to predict the center points of objects. N_2 is processed through a 3×3 convolutional layer, a batch normalization layer, a ReLU activation layer, and a 1×1 convolutional layer to generate the offset and size of objects. Finally, following the design in CenterNet, the heatmap, size, and offset are combined to predict the classifications and the location of the instruments.

TABLE 1. The number of annotated instances for each instrument on STL dataset and CSTL dataset.

STL		CSTL	
Instrument	Number	Instrument	Number
Bipolar	525	Grasper	3568
Cottle Elevator	346	Bipolar	662
Dissector	320	Hook	1761
Drill	856	Scissors	348
Knife	257	Clipper	619
Needle Monopolar	406	Irrigator	776
Ring	465	Specimen Bag	520
Straight Sucker	2660	-	-
Tumor Pliers	614	-	-
Up Scissors	331	-	-
Total	6780	Total	9519

E. LOSS FUNCTION

In this paper, we adopt the same loss function as CenterNet [43], which can be divided into three parts:

1) L_k is the loss of center point prediction in the heatmap;
2) L_{size} is the size prediction loss of object;
3) L_{off} is the offset prediction loss of center point.

As shown in Figure 1, the final heatmap is used to predict center points of object. The positive and negative samples in the heatmap are balanced with a penalty-reduced pixel-wise

logistic regression with the focal loss [43]:

$$L_k = \frac{-1}{N} \sum_{xyc} \begin{cases} \left(1 - \hat{Y}_{xyc}\right)^{\alpha} \log\left(\hat{Y}_{xyc}\right) & \text{if } Y_{xyc} = 1 \\ \left(1 - Y_{xyc}\right)^{\beta} \left(\hat{Y}_{xyc}\right)^{\alpha} & \\ \log\left(1 - \hat{Y}_{xyc}\right) & \text{otherwise} \end{cases} \tag{4}$$

where $\hat{Y}_{xyc}$ is the prediction of the keypoints, and Y_{xyc} is the ground truth of the keypoints. We set the hyperparameters $\alpha = 2$ and $\beta = 4$, following the work in [43], and N is the number of keypoints in the images.

In our experiments, both L_{off} and L_{size} are defined by L1 loss. Specifically, we define L_{off} as below,

$$L_{\text{off}} = \frac{1}{N} \sum_{p} \left| \widehat{O}_{\tilde{P}} - \left(\frac{P}{R} - \widetilde{P} \right) \right| \tag{5}$$

where $\widehat{O}_{\tilde{P}}$ is the prediction center point offset of object k, P is the center point coordinates, R is the down-sampling factor, $\widetilde{P}$ is the center point coordinates acquired after taking the scale $\frac{P}{R}$, and the $\left(\frac{P}{R} - \widetilde{P}\right)$ is the ground truth of center point offset. On the other hand, L_{size} is defined as below,

$$L_{size} = \frac{1}{N} \sum_{k=1}^{N} \left| \hat{S}_{p_k} - s_k \right| \tag{6}$$

where $\hat{S}_{p_k}$ is the predicted size of the object, and s_k is the ground truth. Accordingly, the total loss function is

$$L = L_k + \lambda_{\text{size}} L_{\text{size}} + \lambda_{off} L_{off} \tag{7}$$

where we set coefficients $\lambda_{\text{size}} = 0.1$ and $\lambda_{off} = 1$, following the work in [43].

IV. EXPERIMENTAL RESULTS

In this section, extensive experimental studies are conducted to verify the performance of the proposed AFA-Net with the comparison of five highly cited mainstream methods (RetinaNet, Yolov4, Yolov5, Faster-RCNN and CenterNet) in terms of detection, meanwhile to evaluate the effectiveness of the proposed AFA-Net through ablation studies.

A. DATASETS

To evaluate the performance of the proposed method, proper datasets are necessary with clinical scenarios. In our experimental studies, two datasets are chosen to analyze surgical instrument detection: (1) a dataset of endoscopic transsphenoidal pituitary adenoma resection named Sun21 from the Sun Yat-sen University Cancer Center, in which 21 videos recorded at 30 fps of the surgery from 2020 to 2021 are provided; (2) a dataset of endoscopic cholecystectomy procedures named Cholec80 [7], in which 80 videos recorded at 25 fps of the surgery are provided.

In the Sun21 dataset, we frame the first 10 videos at 30 fps and select 4136 images, named sun21-tool-locations (STL). We annotate 10 classifications of instruments, namely bipolar, cottle elevator, dissector, drill, knife, needle monopolar, ring, straight sucker, tumor pliers, and up scissors, as shown in Figure 5. While, in the Cholec80 dataset, we frame the first 15 videos at 25fps and select 5199 images, named cholec80-sub-tool-locations (CSTL). We annotate 7 classifications of instruments, namely grasper, bipolar, hook, scissors, clipper, irrigator, and specimen bag, as shown in Figure 4.

Both datasets are annotated with bounding boxes containing spatial coordinates under the assistance of an experienced surgeon, and the number of annotated instances of each surgical instrument is shown in Table 1. Following the work in [19], for instruments with handles, the tips are annotated, while for the specimen bag without handles, the entire body is annotated. Each dataset is randomly divided into the training set, testing set, and validation set with a ratio of 6:2:2, which are respectively used for training, tuning, and optimizing the performance of the network.

B. IMPLEMENTATION

After the configuration of the datasets, the PyTorch framework is used for the implementation of the proposed AFA-Net and the five mainstream networks including RetinaNet [53], Yolov4 [31], Yolov5 [32], Faster-RCNN [54] and CenterNet [43]. As part of data preprocessing, the input image size is fixed to 512×512 and the datasets are extended with data augmentation methods, including random cropping, flipping, and color dithering. During training, the backbone networks of all methods are pre-trained on ImageNet [55]. The batch size is set to 32, and AdamW [56] is used as the optimizer with an initial learning rate of 1×10^{-4} with a weight decay of 5×10^{-4}. We use the warmup learning strategy to adjust the learning rate, with a minimum learning rate of 1×10^{-8}, and the learning rate starts to decay with a multiplication factor of 0.1 at the 60th and 70th epochs. All methods in the experimental studies are trained on an NVIDIA Geforce RTX 3090 GPU with 300 epochs and verified to obtain the results, respectively. For evaluation metrics, we use the mean average precision (AP) and average recall (AR), which are used as performance evaluation criteria in object detection, to quantify the detection results of all methods.

C. COMPARATIVE STUDY

The results of the experimental studies are provided in Table 2, where the $AP_{0.5}$ and $AP_{0.75}$ are APs calculated at IoU = 0.5 and 0.75, respectively. From IoU = 0.5 to 0.95, the AP and AR values are calculated every 0.05, and then the mean value is calculated to obtain $AP_{0.5:0.95}$ and AR. Among the five mainstream methods, CenterNet has good performance on extensive evaluation indicators, which reflects the advantage of anchor-free method in the task of surgical instrument detection. Our proposed AFA-Net achieves the SOTA against the compared methods on APs and AR. Specifically, for the $AP_{0.5:0.95}$, we achieve 74.1% and 73.6% on STL and CSTL datasets respectively, which are 3.6% and 3.7% higher than those of CenterNet. For AR, our AFA-Net achieves 4.2% and 3.6% gains on the two datasets respectively, indicating the higher credibility of our

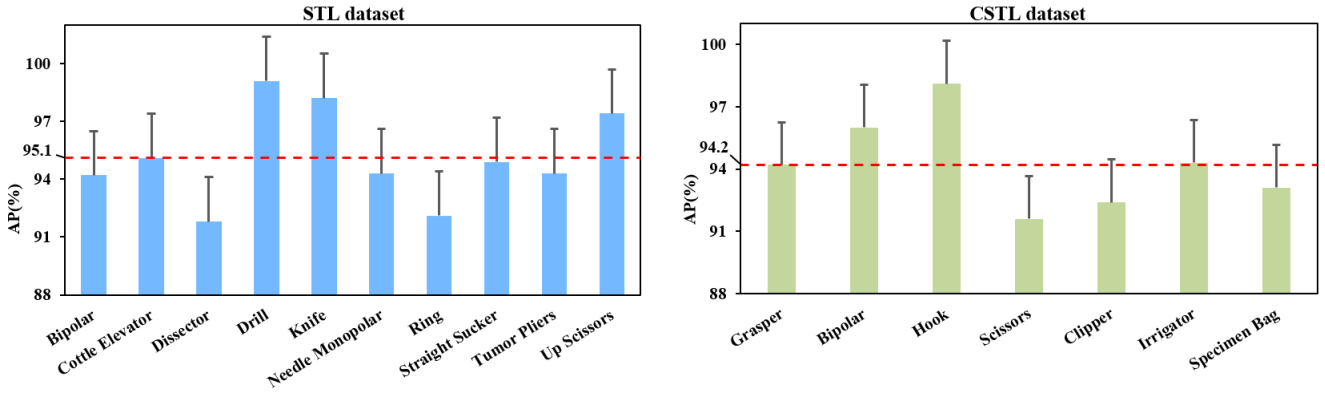

FIGURE 6. The AP for each instrument on STL dataset and CSTL dataset when IoU = 0.5, respectively. Red dotted line represents the average value of all APs, and the vertical error bar represents the positive standard.

TABLE 2. Results(%) of various methods on STL dataset and CSTL dataset.

		STL					CSTL				
Method	Backbone	$AP_{0.5}$	$AP_{0.75}$	$AP_{0.5:0.95}$	AR	FPS	$AP_{0.5}$	$AP_{0.75}$	$AP_{0.5:0.95}$	AR	FPS
RetinaNet [53]	ResNet-50	89.2	71.2	65.1	58.7	20.4	88.3	70.1	64.1	57.2	22.0
Yolov4 [31]	DarkNet-53	89.8	72.1	66.6	60.8	32.5	88.7	71.9	64.8	58.8	33.8
Yolov5 [32]	DarkNet-53	90.9	74.4	68.6	60.5	56.6	89.8	71.8	65.9	60.9	59.8
Faster-RCNN [54]	ResNet-101	91.0	74.6	69.1	62.7	16.1	90.3	74.4	68.3	63.0	17.0
CenterNet [43]	ResNet-101	91.5	75.3	70.5	62.8	32.8	91.1	74.9	69.9	63.1	32.7
AFA-Net(Ours)	ResNet-101	95.1	80.8	74.1	67.0	31.0	94.2	79.6	73.6	66.7	30.2

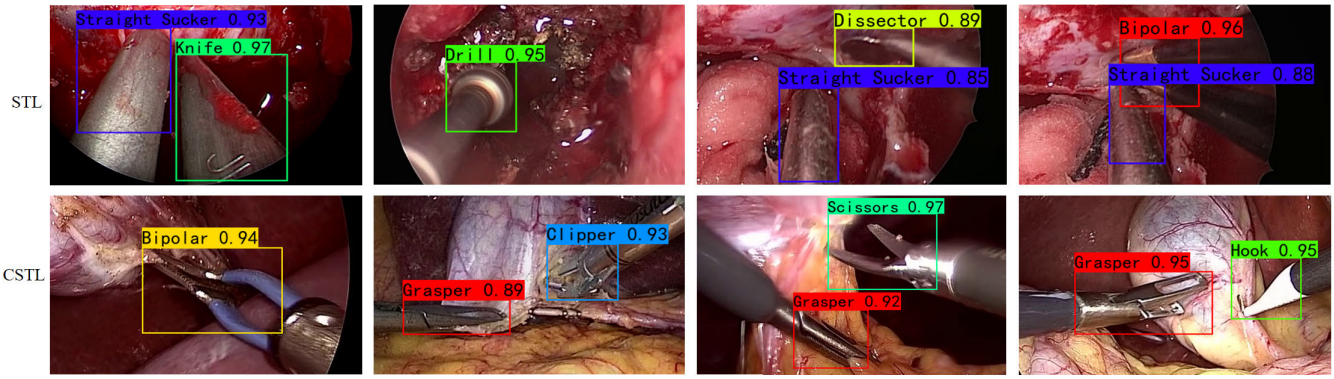

FIGURE 7. The detection results of samples on STL dataset and CSTL dataset. Bounding boxes in different colors display the surgical instrument names and scores on different datasets.

TABLE 3. Results(%) of the AFA-Net with different modules on STL dataset and CSTL dataset.

					STL				CSTL			
Neural Nets	FPN	DAM	WPAM	WHAM	$AP_{0.5}$	$AP_{0.75}$	$AP_{0.5:0.95}$	AR	$AP_{0.5}$	$AP_{0.75}$	$AP_{0.5:0.95}$	AR
Net-A (Base)	-	-	-	-	91.5	75.3	70.5	62.8	91.1	74.9	69.9	63.1
Net-B	✓	-	-	-	92.1	75.8	71.1	62.8	91.6	75.1	70.0	64.3
Net-C	-	✓	-	-	94.5	76.1	72.0	63.2	93.8	75.4	70.2	64.4
Net-D	-	✓	✓	-	94.8	77.3	72.2	66.9	94.0	77.2	71.3	66.5
Net-E (AFA-Net)	-	✓	✓	✓	95.1	80.8	74.1	67.0	94.2	79.6	73.6	66.7

method. Furthermore, the FPS of our AFA-Net is close to CenterNet, which reaches the desired effect and indicates the high computational efficiency of our method. It is notable that the images on CSTL dataset largely differ from those on STL dataset, since they contain different classifications of instruments in different surgical scenarios. AFA-Net performs fairly well on both datasets, which indicates its versatility for instrument detection in diverse surgeries.

Under the parameter IoU = 0.5, the AP of each surgical instrument obtained by our method on the STL and CSTL datasets is shown in Figure 6, with positive standard deviations of 2.3 and 2.1, respectively. The results show that the detection performance with different surgical instruments is close to each other, and high-precision results are obtained, which verifies the stability of our method. Notably, although the drill and hook do not have the maximum number of annotations, they obtain the highest AP. The possible reason is that during the procedure, the surgeon ensures that it is clearly visible during use, while the tip feature of the drill and hook is more pronounced relative to other instruments. The illustration of instrument detection by our method is shown in Figure 7, where we can observe that the prediction bounding boxes of the instruments are highlighted by different colors.

To sum up, our proposed method has achieved desired detection precision, when compared with all the other baselines. Our method can effectively improve the precision of instrument detection and has stable performance and versatility in surgical instrument detection under the complex clinical environment.

D. ABLATION STUDY

To better illustrate the advantage of each module in AFA-Net, the CenterNet is selected as the based neural network, with gradually added feature pyramid networks (FPN), deep aggregation module (DAM), weighted path aggregation module (WPAM), weighted heatmap aggregation module (WHAM) to obtain different neural networks, as shown in Table 3, Net-E represents the proposed AFA-Net. All the neural networks with the same hyperparameter and experimental setting-ups on STL and CSTL datasets are evaluated, respectively.

The obtained results are shown in Table 3, it is noticed that the detection performance of the network is effectively improved with added different modules. By comparing Net-A with Net-B, we observe no significant improvement in terms of APs and AR when using FPN, which might be the loss of feature information in the fusion procedures of FPN. Compared with Net-B, the improvement effect of Net-C is better, which indicates the effectiveness of depthwise separable convolution in DAM. Compared with Net-A, Net-D achieves 4.1% and 3.4% gains of AR on STL and CSTL datasets respectively, which demonstrates the importance of subtle feature information. Furthermore, Net-E achieves the highest APs and AR among all neural networks and performs well on both datasets, which shows the effectiveness of the weighted heatmap aggregation in WHAM and reveals the advantage of the proposed AFA-Net in the surgical instrument detection task.

V. CONCLUSION

In this paper, the anchor-free feature aggregation network (AFA-Net) is proposed for endoscopic surgical instrument detection. The proposed method combines FPN with the depthwise separable convolution to extract features and then adds a weighted feature aggregation module to further enhance the feature information of instruments. Based on the anchor-free method, a weighted heatmap aggregation module is used to detect the surgical instruments, thus can provide the surgeons with accurate information of the surgical instruments, then improving the safety of endoscopic surgery in which inaccurate assessment of the operated surgical instrument may cause unexpected damage. Extensive experimental results with two different endoscopic surgery datasets show that AFA-Net can significantly improve the detection precision of surgical instruments. In future work, more efficient neural networks to improve the detection precision of surgical instruments and further evaluation of the proposed AFA-Net with more datasets will be conducted.

REFERENCES

[1] A. B. Kassam, C. Snyderman, P. Gardner, R. Carrau, and R. Spiro, "The expanded endonasal approach: A fully endoscopic transnasal approach and resection of the odontoid process: Technical case report," *Operative Neurosurg.*, vol. 57, p. E213, Jul. 2005.

[2] R. Martinez-Perez, L. C. Requena, R. L. Carrau, and D. M. Prevedello, "Modern endoscopic skull base neurosurgery," *J. Neuro-Oncol.*, vol. 151, no. 3, pp. 461–475, 2021.

[3] E. W. Wang, P. A. Gardner, and A. M. Zanation, "International consensus statement on endoscopic skull-base surgery: Executive summary," *Int. Forum Allergy Rhinol.*, vol. 9, no. S3, pp. S127–S144, Jul. 2019.

[4] M. Koutourousiou, J. C. Fernandez-Miranda, E. W. Wang, C. H. Snyderman, and P. A. Gardner, "The limits of transsellar/transtuberculum surgery for craniopharyngioma," *J. Neurosurg. Sci.*, vol. 62, no. 3, pp. 301–309, May 2018.

[5] 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.

[6] 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.

[7] 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. 2016.

[8] X. Li, P. Zhao, M. Wu, Z. Chen, and L. Zhang, "Deep learning for human activity recognition," *Neurocomputing*, vol. 444, pp. 214–216, Jul. 2021.

[9] P. Shi, Z. Zhao, S. Hu, and F. Chang, "Real-time surgical tool detection in minimally invasive surgery based on attention-guided convolutional neural network," *IEEE Access*, vol. 8, pp. 228853–228862, 2020.

[10] Y. Liu, X.-Y. Zhang, J.-W. Bian, L. Zhang, and M.-M. Cheng, "SAMNet: Stereoscopically attentive multi-scale network for lightweight salient object detection," *IEEE Trans. Image Process.*, vol. 30, pp. 3804–3814, 2021.

[11] J. Liu, L. Bai, Y. Xia, T. Huang, B. Zhu, and Q.-L. Han, "GNN-PMB: A simple but effective online 3D multi-object tracker without bells and whistles," *IEEE Trans. Intell. Vehicles*, early access, Oct. 27, 2022, doi: 10.1109/TIV.2022.3217490.

[12] Y. Liu, P.-T. Jiang, V. Petrosyan, S.-J. Li, J. Bian, L. Zhang, and M.-M. Cheng, "Del: Deep embedding learning for efficient image segmentation," in *Proc. IJCAI*, vol. 864, 2018, p. 870.

[13] C. Doignon, P. Graebling, and M. de Mathelin, "Real-time segmentation of surgical instruments inside the abdominal cavity using a joint hue saturation color feature," *Real-Time Imag.*, vol. 11, nos. 5–6, pp. 429–442, Oct. 2005.

[14] W. Tan, P. Liu, X. Li, S. Xu, Y. Chen, and J. Yang, "Segmentation of lung airways based on deep learning methods," *IET Image Process.*, vol. 16, no. 5, pp. 1444–1456, Apr. 2022.

[15] J. Liu, W. Xiong, L. Bai, Y. Xia, T. Huang, W. Ouyang, and B. Zhu, "Deep instance segmentation with automotive radar detection points," *IEEE Trans. Intell. Vehicles*, vol. 8, no. 1, pp. 84–94, Jan. 2023.

[16] W. Xiong, J. Liu, Y. Xia, T. Huang, B. Zhu, and W. Xiang, "Contrastive learning for automotive mmWave radar detection points based instance segmentation," in *Proc. IEEE 25th Int. Conf. Intell. Transp. Syst. (ITSC)*, Oct. 2022, pp. 1255–1261.
[17] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2016, pp. 483–499.
[18] L. Bouarfa, O. Akman, A. Schneider, P. P. Jonker, and J. Dankelman, "In-vivo real-time tracking of surgical instruments in endoscopic video," *Minimally Invasive Therapy Allied Technol.*, vol. 21, no. 3, pp. 129–134, May 2012.
[19] 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.
[20] Y. Chu, X. Yang, H. Li, D. Ai, Y. Ding, J. Fan, H. Song, and J. Yang, "Multi-level feature aggregation network for instrument identification of endoscopic images," *Phys. Med. Biol.*, vol. 65, no. 16, Aug. 2020, Art. no. 165004.
[21] 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.* Springer, 2019, pp. 467–478.
[22] M. Attia, M. Hossny, S. Nahavandi, and H. Asadi, "Surgical tool segmentation using a hybrid deep CNN-RNN auto encoder–decoder," in *Proc. IEEE Int. Conf. Syst., Man, Cybern. (SMC)*, Oct. 2017, pp. 3373–3378.
[23] 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, 2020.
[24] C. Yang, Z. Zhao, and S. Hu, "Image-based laparoscopic tool detection and tracking using convolutional neural networks: A review of the literature," *Comput. Assist. Surg.*, vol. 25, no. 1, pp. 15–28, Jan. 2020.
[25] L. Yu, P. Wang, Y. Yan, Y. Xia, and W. Cao, "MASSD: Multi-scale attention single shot detector for surgical instruments," *Comput. Biol. Med.*, vol. 123, Aug. 2020, Art. no. 103867.
[26] L. Du, R. Zhang, and X. Wang, "Overview of two-stage object detection algorithms," *J. Phys., Conf. Ser.*, vol. 1544, no. 1, May 2020, Art. no. 012033.
[27] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path aggregation network for instance segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8759–8768.
[28] 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.
[29] 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.
[30] M. J. Shafiee, B. Chywl, F. Li, and A. Wong, "Fast YOLO: A fast you only look once system for real-time embedded object detection in video," 2017, *arXiv:1709.05943*.
[31] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "YOLOv4: Optimal speed and accuracy of object detection," 2020, *arXiv:2004.10934*.
[32] G. Jocher, A. Stoken, J. Borovec, S. Christopher, and L. C. Laughing, "Ultralytics/YOLOv5: V4.0-nn. SiLU() activations, weights & biases logging, PyTorch hub integration," Zenodo, Tech. Rep., 2021.
[33] Y. Guo, F. Chen, Q. Cheng, J. Wu, B. Wang, Y. Wu, and W. Zhao, "Fully convolutional one-stage circular object detector on medical images," in *Proc. 4th Int. Conf. Adv. Image Process.*, Nov. 2020, pp. 21–26.
[34] D. Wang, N. Zhang, X. Sun, P. Zhang, C. Zhang, Y. Cao, and B. Liu, "AFP-Net: Realtime anchor-free polyp detection in colonoscopy," in *Proc. IEEE 31st Int. Conf. Tools Artif. Intell. (ICTAI)*, Nov. 2019, pp. 636–643.
[35] C. Doignon, F. Nageotte, and M. De Mathelin, "Detection of grey regions in color images: Application to the segmentation of a surgical instrument in robotized laparoscopy," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, vol. 4, Sep./Oct. 2004, pp. 3394–3399.
[36] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. Van Der Laak, B. Van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," *Med. Image Anal.*, vol. 42, pp. 60–88, Dec. 2017.
[37] 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.
[38] J. Peng, Q. Chen, L. Kang, H. Jie, and Y. Han, "Autonomous recognition of multiple surgical instruments tips based on arrow OBB-YOLO network," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–13, 2022.
[39] Y. Xue, S. Liu, Y. Li, P. Wang, and X. Qian, "A new weakly supervised strategy for surgical tool detection," *Knowl.-Based Syst.*, vol. 239, Mar. 2022, Art. no. 107860.
[40] H. Law and J. Deng, "CornerNet: Detecting objects as paired keypoints," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 734–750.
[41] H. Law, Y. Teng, O. Russakovsky, and J. Deng, "CornerNet-lite: Efficient keypoint based object detection," 2019, *arXiv:1904.08900*.
[42] Z. Tian, C. Shen, H. Chen, and T. He, "FCOS: A simple and strong anchor-free object detector," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 4, pp. 1922–1933, Apr. 2022.
[43] X. Zhou, D. Wang, and P. Krähenbühl, "Objects as points," 2019, *arXiv:1904.07850*.
[44] K. Duan, S. Bai, L. Xie, H. Qi, Q. Huang, and Q. Tian, "CenterNet: Keypoint triplets for object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6569–6578.
[45] K. Duan, S. Bai, L. Xie, H. Qi, Q. Huang, and Q. Tian, "CenterNet++ for object detection," 2022, *arXiv:2204.08394*.
[46] T.-Y. Lin, P. Dollár, 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.
[47] Q. Wang, L. Zhou, Y. Yao, Y. Wang, J. Li, and W. Yang, "An interconnected feature pyramid networks for object detection," *J. Vis. Commun. Image Represent.*, vol. 79, Aug. 2021, Art. no. 103260.
[48] C. Wang and C. Zhong, "Adaptive feature pyramid networks for object detection," *IEEE Access*, vol. 9, pp. 107024–107032, 2021.
[49] Y. Liang, W. Changjian, L. Fangzhao, P. Yuxing, L. Qin, Y. Yuan, and H. Zhen, "TFPN: Twin feature pyramid networks for object detection," in *Proc. IEEE 31st Int. Conf. Tools Artif. Intell. (ICTAI)*, Nov. 2019, pp. 1702–1707.
[50] X. Li, T. Lai, S. Wang, Q. Chen, C. Yang, R. Chen, J. Lin, and F. Zheng, "Weighted feature pyramid networks for object detection," in *Proc. IEEE Int. Conf Parallel Distrib. Process. Appl., Big Data Cloud Comput., Sustain. Comput. Commun., Social Comput. Netw. (ISPA/BDCloud/SocialCom/SustainCom)*, Dec. 2019, pp. 1500–1504.
[51] Y.-H. Wu, Y. Liu, L. Zhang, W. Gao, and M.-M. Cheng, "Regularized densely-connected pyramid network for salient instance segmentation," *IEEE Trans. Image Process.*, vol. 30, pp. 3897–3907, 2021.
[52] 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*.
[53] 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. (ICCV)*, Oct. 2017, pp. 2980–2988.
[54] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 28, 2015, pp. 1–9.
[55] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.
[56] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," 2017, *arXiv:1711.05101*.

GUANZHI DING is currently pursuing the B.S. degree with the College of Computer Science, Guangdong University of Technology, Guangzhou, China. His current research interests include computer vision, deep learning, and medical image processing.

XIUSHUN ZHAO received the B.S. degree from Henan University, Kaifeng, China, in 2021. He is currently pursuing the M.S. degree with the College of Automation, Guangdong University of Technology, Guangzhou, China. His current research interests include deep learning, object detection, and medical imaging.

CAI PENG received the B.S. degree from the Hunan Institute of Technology, Hengyang, China, in 2020. She is currently pursuing the M.S. degree with the College of Automation, Guangdong University of Technology, Guangzhou, China. Her current research interests include deep learning, object detection, and medical imaging.

LI LI received the B.Eng. degree in automation from the Wuhan Huaxia University of Technology, China, in 2017, and the M.Eng. degree in control engineering from the Guangdong University of Technology, Guangzhou, China, in 2020. She is currently a Software Engineer with Baiyun Power Group Company Ltd., Guangzhou. Her current research interests include deep learning and image processing.

JING GUO received the B.S. and M.S. degrees from the Guangdong University of Technology, in 2009 and 2012, respectively, and the Ph.D. degree from LIRMM, CNRS-University of Montpelier, France, in 2016. He was a Research Fellow with the National University of Singapore (NUS), from 2016 to 2018. He is an Associate Professor with the Guangdong University of Technology. His current research interests include robotic control and learning, haptic bilateral teleoperation, and surgical robotics. He has served as a Guest Editor for IEEE ROBOTICS AND AUTOMATION LETTERS and *Frontiers in Robotics and AI*.

DEPEI LI received the Ph.D. degree in neurosurgery from Southern Medical University, China. He is currently an attending Neurosurgeon with the Department of Neurosurgery, Sun Yat-sen University (SYSU) Cancer Center, Guangzhou, China. He had published several articles in international academic journals, including *Journal of Neurosurgery*, *Journal of Neuro-Oncology*, and *International Journal of Clinical Oncology*. His current research interests include endoscopic surgery for intracranial tumors and translational research in brain tumors.

XIAOBING JIANG received the Ph.D. degree from Sun Yat-sen University. He is currently an Associate Professor with the Department of Neurosurgical Oncology, Sun Yat-sen University (SYSU) Cancer Center, Guangzhou, China. He had published 18 articles in international journals, including *Molecular Cancer*, *Oncogene*, and *The Journal of Clinical Endocrinology and Metabolism* as the first author or the corresponding author. His current research interests include transnasal endoscopic surgery to resect sellar tumors and research of artificial intelligence in the diagnosis and treatment of pituitary tumors.