Zhao *et al.*
Health Information Science and Systems (2023) 11:58
https://doi.org/10.1007/s13755-023-00260-9

Health Information Science and Systems

RESEARCH

CLAD-Net: cross-layer aggregation attention network for real-time endoscopic instrument detection

Xiushun Zhao[1], Jing Guo[1], Zhaoshui He[1], Xiaobing Jiang[2], Haifang Lou[3*] and Depei Li[2*]

Abstract

As medical treatments continue to advance rapidly, minimally invasive surgery (MIS) has found extensive applications across various clinical procedures. Accurate identification of medical instruments plays a vital role in comprehending surgical situations and facilitating endoscopic image-guided surgical procedures. However, the endoscopic instrument detection poses a great challenge owing to the narrow operating space, with various interfering factors (e.g. smoke, blood, body fluids) and inevitable issues (e.g. mirror reflection, visual obstruction, illumination variation) in the surgery. To promote surgical efficiency and safety in MIS, this paper proposes a cross-layer aggregated attention detection network (CLAD-Net) for accurate and real-time detection of endoscopic instruments in complex surgical scenarios. We propose a cross-layer aggregation attention module to enhance the fusion of features and raise the effectiveness of lateral propagation of feature information. We propose a composite attention mechanism (CAM) to extract contextual information at different scales and model the importance of each channel in the feature map, mitigate the information loss due to feature fusion, and effectively solve the problem of inconsistent target size and low contrast in complex contexts. Moreover, the proposed feature refinement module (RM) enhances the network's ability to extract target edge and detail information by adaptively adjusting the feature weights to fuse different layers of features. The performance of CLAD-Net was evaluated using a public laparoscopic dataset Cholec80 and another set of neuroendoscopic dataset from Sun Yat-sen University Cancer Center. From both datasets and comparisons, CLAD-Net achieves the $AP_{0.5}$ of 98.9% and 98.6%, respectively, that is better than advanced detection networks. A video for the real-time detection is presented in the following link: https://github.com/A0268/video-demo.

Keywords: Cross-layer feature aggregation, Composite attention mechanism, Refinement module, Surgical instrument detection

Introduction

Compared to traditional surgeries, the MIS is more advantageous, allowing for less trauma, less bleeding, faster recovery and lower post-operative complication rates [1, 2]. Robot-assisted surgery [3] and endoscopic surgery [4] are two representative MIS that are widely used in various clinical procedures, which improves the efficiency of typical surgeries and meanwhile benefits the safety of patients. The endoscopic surgery requires a series of operations by surgical instruments at the target part in a patient's body, where an endoscope transmits images of the surgical process to a screen in front of the surgeon. This surgery is often performed by highly skillful surgeons, plus the superb cooperation with the clinicians. The surgeon can not directly touch the tissue during the operational procedure, which makes surgery more complex. In addition, endoscopic surgery faces many challenges in the narrow operating space, such as visual obstruction, mirror reflection and illumination variation, which may cause accidental damages to surrounding tissues by surgical instruments

*Correspondence: louhf-2005@163.com; lidp@sysucc.org.cn
[2] Department of Neurosurgery, Sun Yat-Sen University Cancer Center, Guangzhou 510006, China
[3] Department of Gastroenterology, The First Affiliated Hospital of Zhejiang Chinese Medical University (Zhejiang Provincial Hospital of Chinese Medicine), Hangzhou 310006, China
Full list of author information is available at the end of the article

and harm the safety of patients. To overcome the aforementioned difficulties, ensuring precise identification and detection of surgical instruments during endoscopic surgery becomes crucial in providing real-time visual understanding and improved perspectives for surgeons during the operative procedure. Hence the endoscopic instrument detection gains massive attention in recent research and practical surgeries [5, 6].

Traditional approaches for surgical instrument detection typically extract simple features such as colour, gradient, texture from key points or regions [7], simplifying the problem as the color recognition or threshold segmentation task to detect and classify instruments [8]. Some approaches also rely on manual markers or barcode markers for detection [9]. However, these approaches often have poor detection results and generalization performance. The emergence of deep learning models has provided a new paradigm for the surgical instrument detection, which is becoming increasingly popular because it does not require any modification of surgical instruments and performs well in identifying the location and class. The most research on endoscopic instrument detection is currently focused on designing efficient neural networks [10], which can be divided into two categories, (1) the two-stage detection networks such as RetinaNet [11] and Faster R-CNN [12], which generate the region proposal and then regress to the target with high accuracy; (2) the single-stage detection network such as YOLO [13] and SSD [14], which directly predict the location and category of the target with high efficiency.

Currently, the convolutional neural network has significantly improved the detection accuracy of single-stage detection networks, even surpassing that of two-stage detection networks [15]. Meanwhile, the evolved single-stage networks maintain the characteristics of real-time and high efficiency, therefore becoming a mainstream method of object detection [16, 17]. However, the endoscopic instrument detection remains a great challenge due to the difficult to rapidly distinguish multiple instruments in a narrow operating space, interference by the smoke and body fluids produced during the surgery and other inevitable factors.

In this paper, we propose a cross-layer aggregated attention detection network (CLAD-Net) for endoscopic instrumentation detection, which is evaluated on three surgical instrument detection datasets, i.e., the Cholec80-sub dataset, Sun21 dataset and ATLAS Dione dataset. The experimental findings demonstrate that the CLAD-Net outperforms ten advanced methods in terms of both detection accuracy and efficiency. The key contributions of this research are outlined below:

(1) We propose a cross-layer aggregation attention module to fuse global contextual information, enhance the effectiveness of lateral propagation of feature information, and improve the network's ability to regress to target boundaries.
(2) To solve the problems of inconsistent target size and low contrast in endoscopy, we proposed composite attention mechanism. It extracts context information at different scales through adaptive attention branch and captures long-distance dependencies, effectively solving the problem of target size inconsistency. In addition, it uses multi-scale attention branch to model the importance of each channel in the feature map, reducing the information loss caused by feature fusion, effectively distinguishing foreground and background areas, and solving the problem of low contrast.
(3) The proposed feature refinement module effectively enhances the network's ability to extract target edge and detail information by adaptively adjusting feature weights to fuse features at different levels to achieve refined operation on input features.

The rest of this article follows. Section "Related work" describes the related works. Section "Methods" introduces CLAD-Net in detail, including the CAM for mitigating information loss and solving the problem of inconsistent target size and low contrast in complex backgrounds, the RM for fusing features at different levels. Section "Experiments" discusses the results of comparative and ablation experiments. Finally, the work done in this paper is summarized in Section "Conclusion".

Related work

This section provides an overview of surgical instrument detection, feature pyramid network, and attention mechanisms.

Surgical instrument detection

Conventional approaches for instrument detection predominantly relied on simple features including color, gradient, and texture [7, 8]. The optical tracking, kinematic template matching [18], radio frequency identification (RFID) tracking [19] and image-based detection methods [20] are also studied. Nowadays, deep learning based methods are becoming increasingly popular. One important reason is that there is no need to modify the surgical instrument to provide positioning information. Xue et al. [21] introduced a novel framework for instrument detection, utilizing pseudo-bounding box regression to generate target bounding boxes, but the accuracy of the detection achieved by this framework is relatively low. Namazi et al. [22] presented a multi-label classifier

that incorporates contextual information to identify the presence of surgical instruments in individual frames of a laparoscopic video, but it does not provide precise localization of the instruments within the image. Yang et al. [23] introduced a multiscale fusion network based on transformer models, which effectively segments surgical instruments from endoscopic images and yields promising outcomes. However, the current deep learning based surgical instrument detection methods still have a gap to handle complex surgical scenarios with various interfering factors, which are not effective enough to meet the practical requirements of clinical surgeries.

Feature pyramid network

Feature Pyramid Network (FPN) [24] is a network structure for target detection and image segmentation tasks. FPN fuses feature maps from different levels through horizontal connection to achieve the purpose of multi-scale information transmission, so as to solve the problem of inconsistent target sizes in complex background. With further research, several approaches have been proposed to improve the detail loss in different stages of feature fusion. Spatial pyramid pooling can provide rich contextual information and multi-scale features, cross-layer feature fusion can alleviate the detail loss in the fusion process by combining features at different levels to obtain a richer feature representation. Wang et al. [25] proposed an adaptive FPN that fuses feature contexts by adaptively upsampling operations, aiming to obtain better semantic information, and predict the coordinate offsets of a series of relevant sampling points for each target. Li et al. [26] utilized a cross-layer FPN that incorporates direct cross-layer communication, this dynamic aggregation of multi-scale features enhances the FPN's capability for detecting salient objects. While these methods attempt to retrieve missing information prior to feature aggregation, they do not effectively address feature misalignment and detail loss during fusion, and the lateral propagation of feature information is poor.

Attention mechanisms

The attention mechanism is a way to imitate the human brain's attention to useful information in target objects and is used to improve the performance of algorithms in visual models, including squeeze-and-excitation attention (SE) [27], CBAM [28], Coordinate attention (CA) [29] and so on. At this stage, various attention-based deep learning networks have achieved good performance in tasks such as object classification, object detection and semantic segmentation [30–32]. Ni et al. [33] introduced a surgical instrument segmentation network with an attention mechanism. The attention module enables the network to prioritize key regions and consequently enhance the accuracy of segmentation. Liu et al. [34] proposed a dual-attention context-guided (DACG) module for extracting rich contextual information in the target region to realize the segmentation accuracy of the network for small targets in complex contexts. Li et al. [35] designed a network with SE module to extract image features to recognize surgical stages. As research progressed, several variants of the attention mechanism emerged. Among them, Self-Attention [36] is a common variant that captures global dependencies in sequence models, Multi-Head Attention [37] can learn multiple representations of different models' attentional preferences through a multi-head mechanism. However, the endoscopic instrument detection with the attention mechanism has not been fully investigated so far.

Methods

This section describes the cross-layer aggregation attention module and its main components, including the CAM for capturing contextual information, mitigating information loss due to feature fusion, and modeling the importance of each channel, as well as the RM for performing refinement operations on the input features by fusing features from different layers and adaptively adjusting the feature weights.

Cross-layer aggregation attention module

In the multi-scale feature fusion network, the shallow feature maps generated by the shallow network possess more texture features of target objects, containing rich information of details. In contrast, the deep feature maps generated by the deep network extract more semantic information through larger perceptual fields. Most of the existing detection networks transfer different levels of features by fusing multi-scale features to improve the perception of different sized targets and enhance the expressive ability. Deep features contain abundant semantic information, making them well-suited for detecting larger target objects. On the other hand, shallow features contain more detailed information, making them more suitable for detecting smaller target objects. In other words, the contributions of feature maps from different layers are disparate to the detection of target objects with diverse sizes. However, multi-scale feature fusion requires multiple sampling operations, which will lead to the loss of information of the higher-level features. Moreover, there are semantic gaps between different feature layers, and direct fusion will ignore the mapping relationship between them and reduce the multiscale representation capability. To address the above limitation, we propose a cross-layer aggregation attention module to wisely fuse heterogeneous feature maps.

The cross-layer aggregation attention module is designed to further stimulate the detection network to involve shallow features, while fully integrating the semantic information extracted from layers of the backbone, which enhance the network in terms of regression to the target boundary. Specifically, it adds lateral propagation between input and output nodes of the same size, which effectively integrates shallow features, such as details, edges and contours information, into the deeper network, which renders the regression to the target boundary more accurate. In addition, we combine depthwise separable convolution in the cross-layer aggregation attention module to extract deeper features, which benefit the exploration of the detailed information and thus gain more accurate recognition and location. As seen from Fig. 1, we first pass the feature maps C3, C4, C5 extracted from the backbone via depth-separable convolutional transfer, and then fuse them with the corresponding feature maps P3, P4, P5 via RM to obtain rich contextual information as well as refined features. The next step involves upsampling the feature maps N2, N3, N4 to match the spatial dimensions of C3, C4, C5. Subsequently, these feature maps are concatenated along the bottom-up pathway, and the CAM is employed to address the problem of information loss during fusion, as well as the problem of inconsistent target size and low contrast in complex backgrounds. The final feature maps N3, N4, N5 are obtained as inputs to the subsequent detection head section for predicting the location information and category of the target.

Composite attention mechanism

In endoscopic surgery, due to the difference in distance and angle between the instrument and endoscope, the imaging effect of the instrument will change greatly. The CAM is introduced to extract contextual information and long-distance dependencies, model the correlation between each channel, and solve the problem of inconsistent target size and low contrast due to the movement of instruments or endoscopes. CAM includes AAB and MSAB. AAB extracts context information at different scales through adaptive pooling layers to better capture long-distance dependencies and improve the network's detection capabilities for slender-shaped targets (e.g., Straight Sucker and Irrigator), thus effectively solving the problem of inconsistent target sizes. By modeling the correlation between each channel, MSAB integrates global and local feature information to effectively distinguish the target from the background and solve the problem of low contrast. Next, we will introduce them one by one.

Adaptive attention branching

AAB establishes distance relationships between different positions in the sequence through adaptive pooling layers to more comprehensively consider contextual information and capture long-term correlations, reduce information loss in the feature fusion process, and solve the problem of inconsistent target sizes. As illustrated in Fig. 2, AAB first obtains 4 contextual features of different scales through the adaptive pooling layer, then adjusts the number of channels through 1×1 convolution respectively, and upsamples them to the input feature map size. The four extracted contextual features are then concatenated and passed through a 1×1 convolutional layer, a ReLU activation layer, a BatchNorm layer, a 3×3 convolutional layer, and a Sigmoid activation layer in turn to generate the corresponding weights f_a for each feature map. Finally, the f_a is used to guide the importance of the

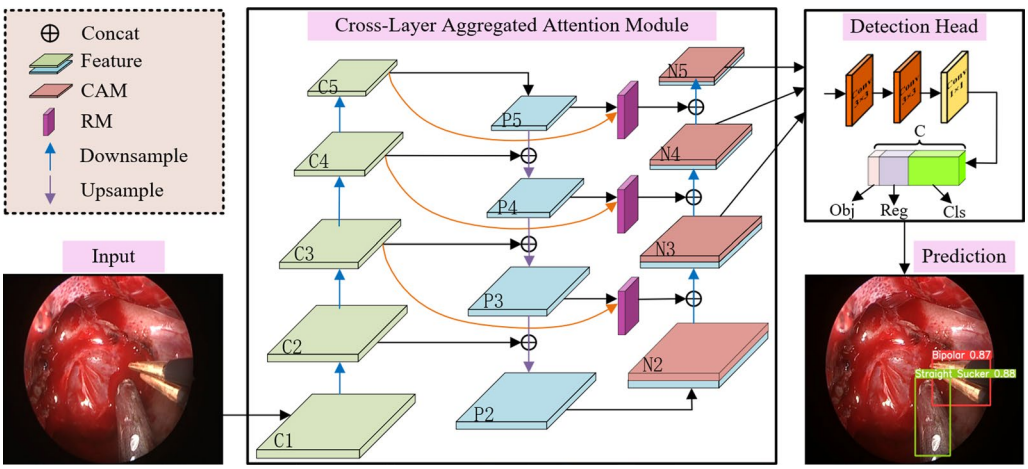

Fig. 1 The architecture of CLAD-Net, where Cls is the category prediction, Reg is the center point and width-height coordinate prediction, and Obj is the confidence prediction

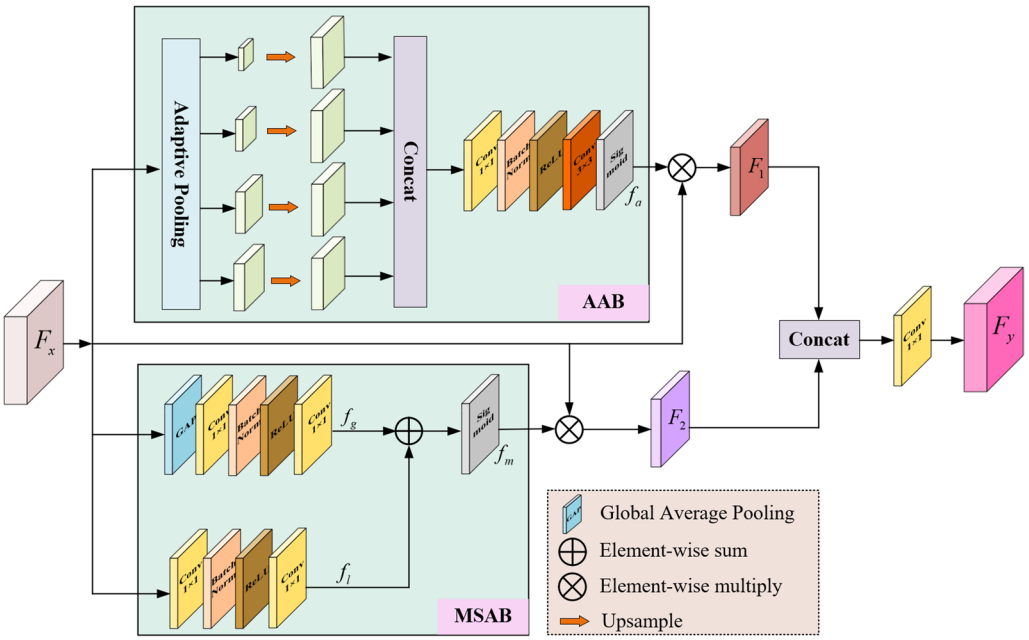

Fig. 2 The structure of CAM, including adaptive attention branching (AAB) and multi-scale attention branching (MSAB)

channels in the input feature to obtain the feature map F_1 with rich contextual information. AAB mitifies the loss of information from feature fusion, and further strengthens the feature extraction capability for surgical instruments.

Multi-scale attention branching

To effectively differentiate the detection target from the complex background, it is essential to extract the target features and enhance their discriminative properties. For this purpose, a MSCA is introduced in CAM, which can be used to distinguish the importance of feature information at different locations, and its structure is shown in Fig. 2. MSCA consists of global channel attention (GCA) and local channel attention (LCA). During the training process, GCA allows the network to capture the significance of each channel in the feature map, highlights the useful information and suppresses the redundant features, which can effectively solve the problems of blurring, occlusion and low contrast of the image. It first performs a global average pooling (GAP) operation on the input feature map, integrates the global information of the feature map, and turns the input features into a vector of channel dimensions, and each element in the vector corresponds to a channel in the input feature map, which has a global receptive field. Subsequently, the number of parameters and complexity of the model are reduced through 1×1 convolution, and then the nonlinear relationship between channels is extracted through BatchNorm and ReLU. Finally, the number of channels of the output feature map is restored through 1×1 convolution, and the weight matrix f_g of each channel is obtained.

LCA first reduces the number of channels of the input feature map by 1×1 convolution and models the correlation between channels, then enhances the correlation between channels by BatchNorm and ReLU, and recovers the number of channels of the feature map by 1×1 convolution, and outputs the weight matrices f_l of the elements at different positions on the same channel. Finally, the weights f_g and f_l are summed up and passed through the Sigmoid function to obtain the final multiscale attention weight f_m.

The GCA is focused on the input feature map by the convolution kernel of size $H \times W$ that integrates the global channel information, while the LCA can be regarded as pooling on the input features by the convolution kernel of size 1×1, which aims to model the importance between different channels on each pixel to avoid the small-scale targets being neglected due to the interference of noise information. The overall computational process of CAM can expressed as:

$$F_y = Conv_1(Concat[f_a \cdot F_x, f_m \cdot F_x]). \tag{1}$$

where *Concat* denotes the feature spliced along the channel direction, the $Conv_1$ is the 1×1 convolution layer.

Refinement module

Our proposal involves the use of a RM to optimize and enhance the low-level features obtained from the

backbone, as well as the high-level features extracted from the feature extraction network. As shown in Fig. 3, the RM first employs a GAP to capture the global context information, encodes each channel of the input features F_a and F_b to obtain the direction-aware information, and guides the feature learning with the direction-aware information. The information is then transformed using 1×1 convolution to obtain the direction-aware feature map. Then, the expressive ability of the network is enhanced by a Non-Linear activation function, and two attention weights ω_1 and ω_2 are generated by the 1×1 convolution and Sigmoid activation function, respectively, then ω_1 and ω_2 are summed for weight integration, and the integrated weights are multiplied by the input features F_a to obtain F_{ax}. At the same time, the attention weights $1 - \omega$ are multiplied with F_b to get F_{bx}, and finally F_{ax} is concatenated with F_{bx} to get the final refinement result F_x. This design can refine the output features at each stage of the context path, fusing features of different levels or resolutions and equalizing the proportion of information carried by the input features, enhancing the network's ability to extract information about the edges and details of the target. The overall computational process of RM can expressed as:

$$F_x = Concat[(\omega \cdot F_a), ((1 - \omega) \cdot F_b)]. \quad (2)$$

Experiments

This section first introduces the datasets, implementation details, loss function, and evaluation metrics. Next, the performance of CLAD-Net is compared with state-of-the-art object detection networks to evaluate its effectiveness. Additionally, a set of ablation experiments are conducted to evaluate the impact of key modules within CLAD-Net.

Datasets and implementation

In our experimental studies, we use two endoscopic surgical instrument datasets to validate the performance of CLAD-Net, as described below.

An endoscopic pituitary adenoma resection dataset Sun21 from the Center for Cancer Control, Sun Yat-sen University, which provides 21 surgical videos recorded from 2020 to 2021. The first 10 videos of the dataset were annotated at 30 FPS to obtain 4136 images and annotate 10 instruments.

An endoscopic cholecystectomy procedures dataset Cholec80 [38], which provides 80 surgical videos recorded at 25 FPS. The initial 15 videos from the dataset were annotated, resulting in 5199 images. The annotations include 7 different surgical instruments, this dataset was named Cholec80-sub. The names of each surgical instrument and the number of annotated instances can be found in Table 1, and some examples of the two datasets are shown in Fig. 4.

We perform annotation under the guidance of a surgeon, and the annotation rule is that if the instrument head is visible in the current frame, the visible part of the surgical instrument is surrounded by the smallest rectangular box in the current frame. For instruments with handles we annotate only the head, whereas for specimen bags we annotate the entire body. Both datasets are framed in chronological order, with complete chronological information. For both datasets, we partitioned the data into three sets: training, testing, and validation, using an 8:1:1 ratio.

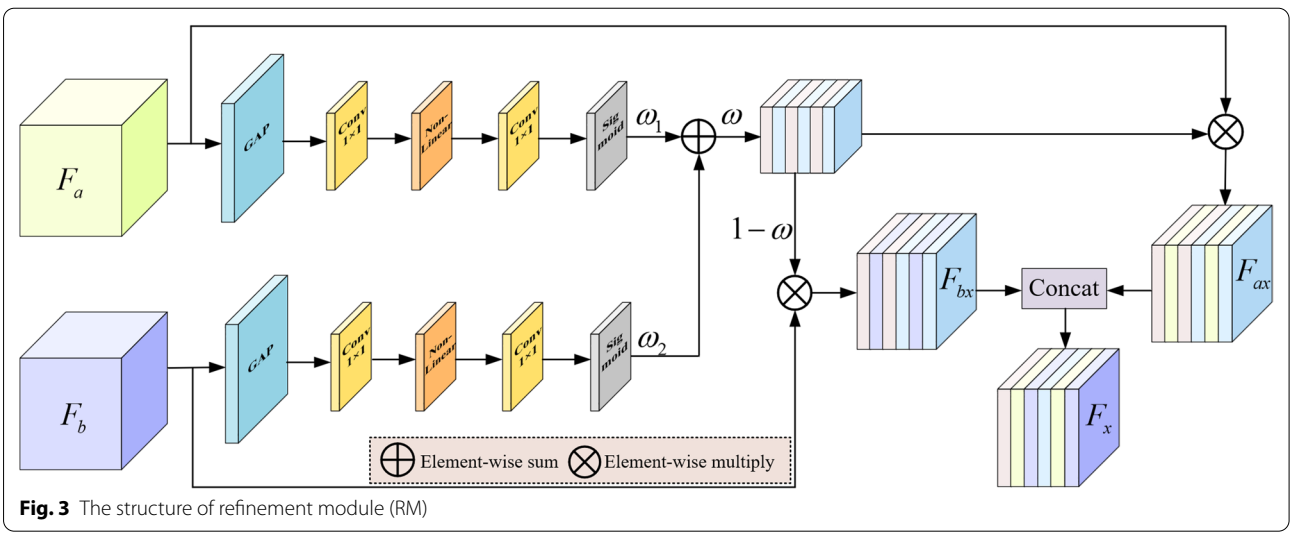

Fig. 3 The structure of refinement module (RM)

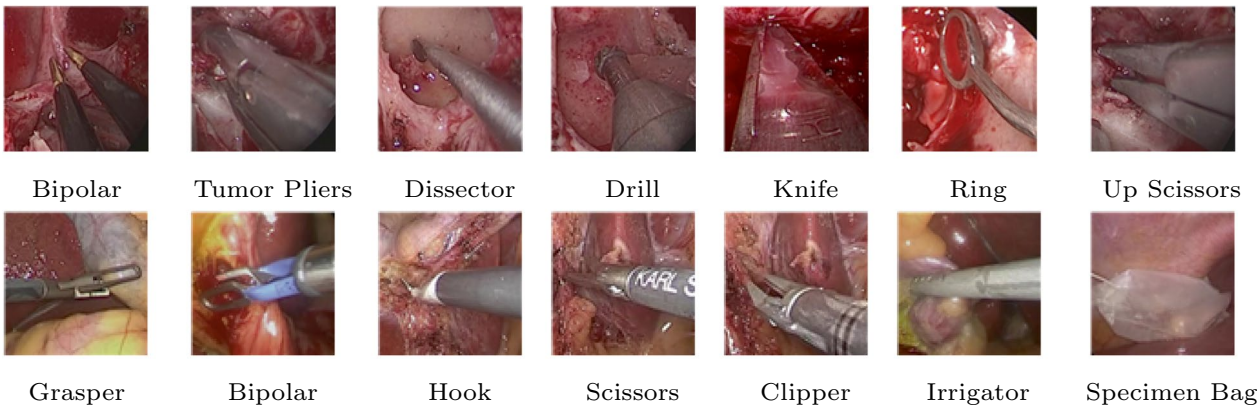

Fig. 4 Visualization of datasets (Upper: the Sun21 dataset; Lower: the Cholec80-sub dataset). Except the ones in the figure, the Sun21 dataset also contains Straight Sucker, Cottle Elevator and Needle Monopolar

Table 1 Details of the Sun21 dataset and the Cholec80-sub dataset

Sun21		Cholec80-sub	
Instruments	**Numbers**	**Instruments**	**Numbers**
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

We experiment with python program and PyTorch 1.7.1. Meanwhile, we use NVIDIA GeForce GTX 3070Ti GPU, with CUDA version 11.1. Regarding the dataset processing part, we set both the training and test image sizes to 640 × 640 and perform mosaic data enhancement (MDE) on the training images. MDE selects four images for random scaling, cropping, flipping and stitching, which can expand the original dataset to prevent the occurrence of overfitting. It also mitigates the impact of the data category imbalance on detection results. During the training phase, we trained a total of 150 epochs. Using the SGD optimizer, the initial learning rate is 0.01 and the batch size is 16.

Loss function and evaluation indicators

The loss function consists of localization loss $Loss_{Reg}$, confidence loss $Loss_{Obj}$ and classification loss $Loss_{Cls}$. The loss is expressed as below:

$$Loss = \alpha Loss_{Reg} + \beta Loss_{Obj} + \gamma Loss_{Cls}. \tag{3}$$

where the loss weights α, β, γ are set to 0.05, 1.0, 0.5, respectively. $Loss_{Obj}$ and $Loss_{Cls}$ are calculated by cross-entropy (CE) loss, $Loss_{Reg}$ is calculated by CIoU loss [39].

The formula for CIoU loss is as follows:

$$v = \frac{4}{\pi^2}\left(\arctan\frac{\omega^{gt}}{h^{gt}} - \arctan\frac{\omega}{h}\right)^2, \tag{4}$$

$$\alpha = \frac{v}{(1 - IoU) + v}, \tag{5}$$

$$Loss_{CIoU} = 1 - IoU + \frac{\rho^2(b, b^{gt})}{c^2} + \alpha v. \tag{6}$$

where b represents the central point of the predicted box, and b^{gt} represents the central point of the ground truth box, w^{gt} and w represent the width of the ground truth box and the prediction box. Similarly, h^{gt} and h represent the height of the ground truth box and the prediction box. ρ represents the distance between the center points of the two boxes, α is an adjustable hyperparameter, and v is used to calculate the difference in aspect ratio.

This article uses $AP_{0.5}$, $AP_{0.5:0.95}$, Recall and FPS as measurement indicators. FPS is an abbreviation for Frames Per Second, which refers to the number of images detected per second. It is commonly used to measure the speed of a network. Recall represents the proportion of correct predictions to prediction samples. The definition of Recall is as follows:

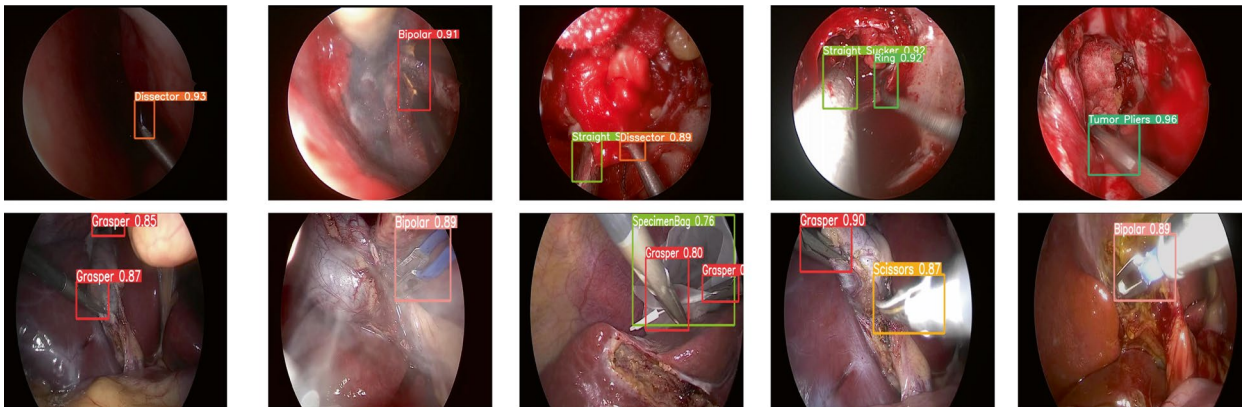

Fig. 5 Visualization of endoscopic instrument detection (Upper: results for the Sun21 dataset; Lower: results for the Cholec80-sub dataset)

$$Recall = \frac{TP}{TP + FN}, \tag{7}$$

where TP denotes the count of correctly predicted instruments, while FN represents the count of incorrect predictions.

The defining equation of mAP is as follows:

$$AP = \int_0^1 P(R)dR, \tag{8}$$

$$mAP = \frac{\sum_{i=1}^{N} AP_i}{N}. \tag{9}$$

where N is the number of categories of instruments, $AP_{0.5}$ refers to the average precision of all categories when the accuracy evaluation IoU threshold is set to 0.5. On the other hand, $AP_{0.5:0.95}$ represents the average precision calculated by varying the IoU threshold in increments of 0.05 from 0.5 to 0.95.

Comparative study

We validate our approach (CLAD-Net) on the two surgical datasets described in Sect. 4.1 and compare it with existing advanced networks, including Faster R-CNN [12], RetinaNet [11], SSD [14], CenterNet [40], EfficientDet [41], DETR [42], YOLOv5 [43], YOLOX [44], YOLOv6 [45], RT-DETR [46]. The experimental results are presented in Table 2, where we highlight the best results for each metric and dataset using the **Bold** formatting.

The results show that on the Cholec80-sub dataset, CLAD-Net's $AP_{0.5}$ and $AP_{0.5:0.95}$ are 98.9% and 70.2% respectively, while on the Sun21 dataset they are 98.6% and 67.0% respectively, exceeding the other 10 baselines.

Table 2 Results of state-of-the-art detection methods for three datasets

Method	Cholec80-sub			Sun21			ATLAS dione			Parameter(M)
	$AP_{0.5}$(%)	$AP_{0.5:0.95}$(%)	$FPS(s^{-1})$	$AP_{0.5}$(%)	$AP_{0.5:0.95}$(%)	$FPS(s^{-1})$	$AP_{0.5}$(%)	$AP_{0.5:0.95}$(%)	$FPS(s^{-1})$	
Faster R-CNN	96.4	52.0	15.9	94.5	52.6	16.0	95.5	75.8	13.4	124.8
RetinaNet	91.6	55.3	20.4	91.2	54.6	22.0	93.2	73.2	18.1	86.9
SSD	93.5	51.9	64.1	91.1	50.6	53.2	94.2	71.6	63.0	26.3
CenterNet	96.8	58.8	31.0	95.2	57.0	28.6	97.9	81.0	29.7	32.6
EfficientDet	95.3	62.9	29.6	93.4	61.7	24.5	94.5	81.7	36.4	20.7
DETR	97.6	66.3	28.0	96.9	64.1	25.2	97.7	84.2	33.0	36.7
YOLOv5	98.1	68.2	43.2	97.8	65.9	44.8	99.1	86.8	47.5	21.2
YOLOX	98.3	68.2	46.0	97.4	64.6	37.7	98.9	87.1	49.2	9.1
YOLOv6	97.8	68.4	64.2	98.0	66.3	58.2	**99.5**	88.0	67.6	16.3
RT-DETR	97.5	69.0	27.4	97.7	66.3	33.2	99.1	87.6	34.5	32.8
CLAD-Net(Ours)	**98.9**	**70.2**	**68.5**	**98.6**	**67.0**	**58.7**	**99.5**	**88.2**	**71.2**	**7.5**

Bold values indicate the best results for each indicator on different models

The detection speed is 68.5 FPS and 58.7 FPS on two datasets, which meets the real-time requirements for surgical instrument detection in endoscopy. This verifies that the CLAD-Net can fuse global contextual information and refinement features, raise the effectiveness of lateral propagation of feature information and improve the network's ability to regress to target boundaries. It is worth noting that the Cholec80-sub dataset and Sun21 dataset reflect different surgical scenarios and contain different types of surgical instruments. CLAD-Net achieves good results on both datasets, which verifies its versatility in different surgical scenarios. The instrument detection in challenging situations by CLAD-Net is visualized in Fig. 5, including illumination variation, smoke, visual obstruction, mirror reflection and blurriness. We observe that the diverse instruments are well distinguished by boxes with different colors. The real-time detection is displayed in the following link: https://github.com/A0268/video-demo.

We analyze the different networks in Table 2 in terms of inference speed. For the two-stage network Faster R-CNN and RetinaNet, candidate frames need to be generated first during prediction, and then these candidate frames are classified and positioned. This method has higher model complexity and leads to slower inference speed. For single-stage networks, SSD uses lightweight MobileNetv2 as the backbone network, which has faster detection speed. CenterNet uses ResNet-50 with a large number of parameters as the backbone feature extraction network, resulting in a relatively low FPS. In the EfficientDet network, we chose EfficientDet-D4, which has a good balance between accuracy and speed, as the base model. However, because its number of parameters is almost three times that of CLAD-Net, its FPS is much slower. DETR and RT-DETR use the transformer architecture, which requires a large number of multi-head self-attention mechanisms when processing image data, resulting in relatively high computational complexity and therefore low FPS. In order to pursue higher accuracy, we used the YOLOv5m model for training. From Table 2, we can see that its AP value is very close to CLAD-Net, but its large number of parameters results in lower FPS. Compared with CLAD-Net, YOLOX has greater model complexity, including deeper network layers and more parameters, which means that YOLOX requires more computing resources for image reasoning. CLAD-Net uses the lightweight CSPdarkernet53 as the backbone feature extraction network, and employs RM and CAM in the cross-layer aggregation attention module to extract features more efficiently, so it has the fastest FPS.

To verify the generalization performance of CLAD-Net, we performed experiments on the publicly accessible dataset ATLAS Dione [47]. The ATLAS Dione dataset comes from performing six different simulated surgical tasks on the DaVinci surgical system and has 22,467 annotated images. We use the same configuration as Sect. 4.1 for training, and the experimental results are shown in Table 2. It can be seen that CLAD-Net still has the best performance. Figure 6 shows the results of CLAD-Net's detection at different task stages on the ATLAS Dione dataset.

In order to verify the reliability of the model and reduce the evaluation bias caused by different single divisions of the dataset, we used ten-fold cross-validation to test the performance of CLAD-Net on the Sun21 dataset. The ten-fold cross-validation randomly divided the dataset into 10 mutually disjoint subsets S_1, S_2,..., S_{10}, each subset contains 413 images, and a total of 10 experiments are performed. In the jth experiment, S_j is selected as the test set, and the other remaining subsets are used as the training set. The weights trained by the training set are tested on the test set, and the results of ten sets of experiments are averaged as the final evaluation index. We select YOLOv6, which performs best on the Sun21 dataset except CLAD-Net, as a comparison model. The experimental results are shown in Table 3. It can be seen from the experimental results that CLAD-Net is 0.65% and 0.73% higher than YOLOv6 in $AP_{0.5}$ and $AP_{0.5:0.95}$ indicators respectively.

To sum up, the proposed CLAD-Net can detect surgical instruments in various complex environments and achieve desirable detection accuracy. In addition, compared with the existing advanced detection networks, our detection network effectively improves the accuracy of

Fig. 6 Detection results for different task stages on the ATLAS Dione dataset

Table 3 Ten-fold cross-validation comparison experiments of CLAD-Net and YOLOv6 on Sun21 dataset

Experiment	CLAD-Net		YOLOv6	
	$AP_{0.5}$ (%)	$AP_{0.5:0.95}$ (%)	$AP_{0.5}$ (%)	$AP_{0.5:0.95}$ (%)
1	98.5	66.8	97.9	65.8
2	98.5	67.4	98.1	66.0
3	98.6	66.7	97.7	66.4
4	97.9	66.5	97.2	65.9
5	98.2	66.9	97.5	65.8
6	98.8	66.7	98.3	66.7
7	98.5	67.2	97.4	66.5
8	98.3	67.0	97.8	66.5
9	98.4	66.9	98.0	65.9
10	98.6	67.1	97.9	66.4
Mean	**98.43**	**66.92**	97.78	66.19

Bold values indicate the best results for each indicator on different models

the endoscopic instrument detection, with reliable performance and good versatility.

Ablation study

To validate the components proposed in this paper, we designed six networks and evaluated them on the Cholec80-sub and Sun21 dataset to determine the effectiveness of each component. Among them, Baseline is the removal of CAM and RM in CLAD-Net. Baseline-A refers to the introduction of AAB in Baseline, Baseline-M refers to the introduction of MSAB in Baseline, Baseline-C refers to the introduction of CAM in Baseline, Baseline-R refers to the introduction of RM in Baseline, CLAD-Net refers to the introduction of CAM and RM in Baseline. For the training of each detection network, we use the same experimental configuration as mentioned in Sect. 4.1. The results are shown in Table 4. Figure 7 illustrates the $AP_{0.5}$ and $AP_{0.5:0.95}$ curves for CLAD-Net, Baseline-C, Baseline-R, and Baseline on the Cholec80-sub and Sun21 datasets.

As indicated in Table 4, the network's $AP_{0.5}$, $AP_{0.5:0.95}$, and Recall on both datasets are boosted after introducing AAB and MSAB in the cross-layer aggregation attention module. As shown by the experimental results of Baseline-C, the enhancement effect is more obvious after the introduction of CAM, which improves $AP_{0.5}$, $AP_{0.5:0.95}$, and Recall on the Cholec80-sub dataset by 1.0%, 1.2%, and 0.6%, respectively, and on the Sun21 dataset by 1.3%, 1.2%, and 1.5%, respectively. This demonstrates that CAM effectively solves the problems of target occlusion and low contrast in complex scenarios by capturing contextual information through AAB, mitigating information loss due to feature fusion, and modeling the importance of each channel in the feature map using MSAB. Comparing the experimental results of CLAD-Net and Baseline-C, it can be seen that the model after the introduction of RM improves the $AP_{0.5}$, $AP_{0.5:0.95}$, and Recall on Cholec80-sub dataset by 0.3%, 1.5%, and 0.8%, respectively, and on the Sun21 dataset by 0.4%, 1.5%, and 0.5%, respectively. This demonstrates that RM effectively enhances the network's ability to extract target edge and detail information by adaptively adjusting feature weights to fuse features at different levels to achieve refined operations on input features.

Discussion

In endoscopic surgery, accurate detection of the position and status of surgical instruments in real time can help surgeons better observe and perceive the surgical process and prevent accidents. Table 2 shows that CLAD-Net is superior to other SOTA models and meets the real-time requirements for surgical instrument detection in MIS. Table 4 shows the effectiveness of our proposed CAM and RM. Nevertheless, there are still many thorny issues that need to be addressed in endoscopic instrument detection. For example, scab on surgical instrument is a common phenomenon during endoscopic surgery, which will hinder the operation, increase the patient's risk of infection, and interfere with the doctor's field of vision. Our current method is not effective in detecting surgical

Table 4 Ablation study results (%) of different components on Cholec80-sub dataset and Sun21 dataset

Method	AAB	MSAB	CAM	RM	Cholec80-sub			Sun21		
					$AP_{0.5}$	$AP_{0.5:0.95}$	Recall	$AP_{0.5}$	$AP_{0.5:0.95}$	Recall
Baseline					97.6	67.5	96.6	96.9	64.3	95.6
Baseline-A	✓				98.1	68.2	97.1	97.5	65.4	96.7
Baseline-M		✓			97.9	68.2	97.4	98.0	65.3	97.1
Baseline-C			✓		98.6	68.7	97.2	98.2	66.5	97.1
Baseline-R				✓	98.6	68.3	96.9	97.9	65.8	96.5
CLAD-Net			✓	✓	**98.9**	**70.2**	**98.0**	**98.6**	**67.0**	**97.6**

Bold values indicate the best results for each indicator on different models

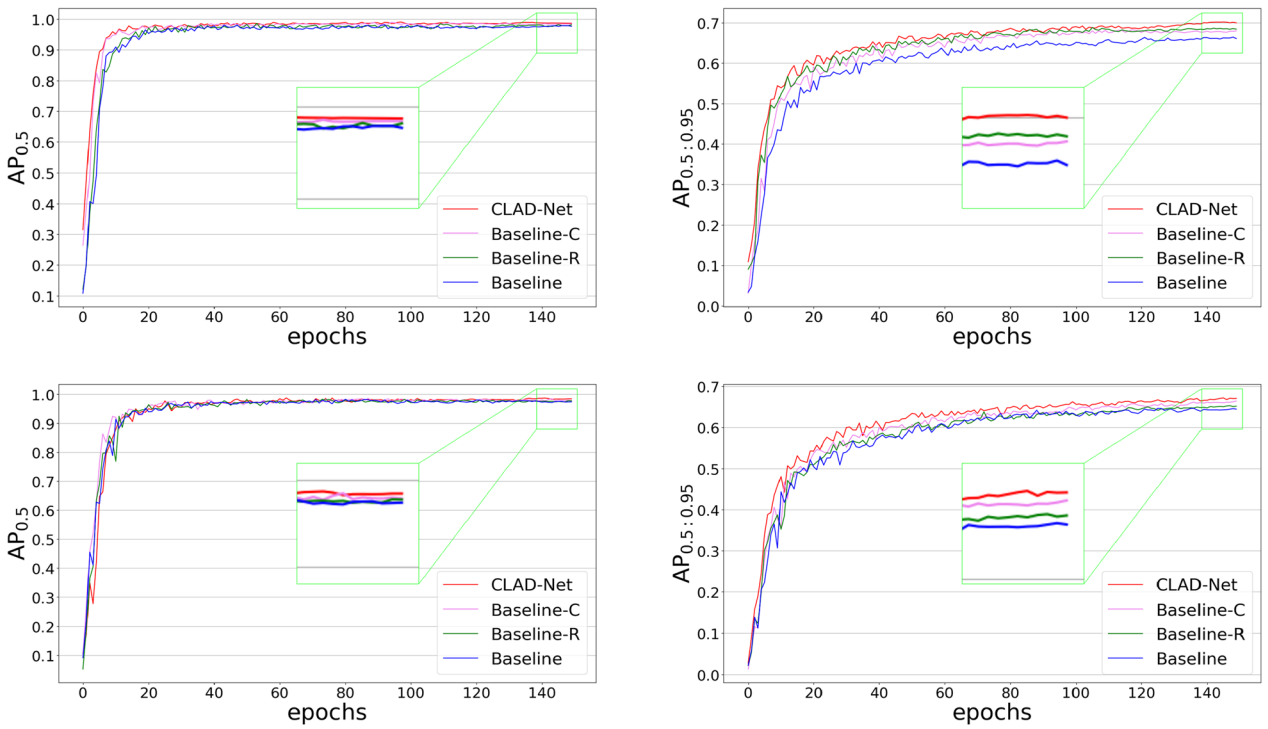

Fig. 7 $AP_{0.5}$ and $AP_{0.5:0.95}$ curves on the above models (Upper: results for the Cholec80-sub dataset; Lower: results for the Sun21 dataset)

instrument scab. In further research, we will explore more deeply the performance of CLAD-Net on bipolar forceps scab detection and conduct a more detailed analysis of other challenges it may face in various scenarios. This will help provide a more comprehensive and accurate assessment and guide future improvement efforts.

In practical applications, how to reduce algorithm complexity and deploy it under limited hardware resources is also an issue worth considering. Next, we will study how to reduce the computational complexity and memory usage of the algorithm from the perspective of network structure optimization. Specifically, a shallower cross-layer aggregated attention module can be used to reduce the amount of parameters and computational complexity, or the number of channels in CAM and RM can be reduced and the number of convolution kernels in the convolutional layer can be reduced to achieve lightweight design [48]. For scalability, we will test the effect of the model in laparoscopic surgery and cystoendoscopic surgery, and make some improvements to the model based on actual conditions to adapt to different surgical scenarios.

In addition, since our datasets has complete chronological information, it can be considered to combine CLAD-Net and Long Short Term Memory network (LSTM) [49], and use LSTM to extract the chronological information and context dependence of images. Since the datasets with location annotations are limited and manual data annotation requires a lot of effort, semi-supervised methods [50] can be considered to train the model to reduce the dependence on location annotation data.

Conclusion

This paper proposes a cross-layer aggregated attention detection network (CLAD-Net) for accurate and efficient detection of endoscopic instruments in complex surgical scenarios. First, fuse global contextual information through the cross-layer aggregation to raise the effectiveness of lateral propagation of feature information and enhance the perception of different-sized targets. Secondly, CAM is used to extract contextual information at different scales and model the importance of each channel in the feature map to reduce the information loss caused by feature fusion and effectively solve the problems of inconsistent target sizes and low contrast in complex backgrounds. Finally, the RM is used to fuse different levels of features, and the refinement operation of weighting the input features is achieved by adaptively adjusting the feature weights, which enhances the ability to extract edge and detail information. The experimental results show that CLAD-Net achieves the best results in terms of detection accuracy and efficiency compared with existing advanced methods. In the future, we plan to further

evaluate the CLAD-Net on more endoscopic datasets and deploy it in practical use.

Acknowledgements
This work is supported by the Fundamental and Applied Basic Research Program of Guangdong Province (Grant No. 2023A1515030179)

Declarations

Conflict of interest
The authors declare no conflicts of interest.

Author details
[1]School of Automation, Guangdong University of Technology, Guangzhou 510006, China. [2]Department of Neurosurgery, Sun Yat-Sen University Cancer Center, Guangzhou 510006, China. [3]Department of Gastroenterology, The First Affiliated Hospital of Zhejiang Chinese Medical University (Zhejiang Provincial Hospital of Chinese Medicine), Hangzhou 310006, China.

Received: 18 September 2023 Accepted: 5 November 2023
Published: 27 November 2023

References

1. Omisore OM, Han S, Xiong J, Li H, Li Z, Wang L. A review on flexible robotic systems for minimally invasive surgery. IEEE Trans Syst Man Cybern Syst. 2020;52(1):631–44.
2. Tonutti M, Elson DS, Yang G-Z, Darzi AW, Sodergren MH. The role of technology in minimally invasive surgery: state of the art, recent developments and future directions. Postgrad Med J. 2017;93(1097):159–67.
3. Casas-Yrurzum S, Gimeno J, Casanova-Salas P, García-Pereira I, Olmo E, Salvador A, Guijarro R, Zaragoza C, Fernández M. A new mixed reality tool for training in minimally invasive robotic-assisted surgery. Health Inform Sci Syst. 2023;11(1):34.
4. Kim M, Kim H-S, Oh SW, Adsul NM, Singh R, Kashlan ON, Noh JH, Jang IT, Oh SH. Evolution of spinal endoscopic surgery. Neurospine. 2019;16(1):6–14.
5. Chu Y, Yang X, Li H, Ai D, Ding Y, Fan J, Song H, Yang J. Multi-level feature aggregation network for instrument identification of endoscopic images. Phys Med Biol. 2020;65(16): 165004.
6. Lam K, Lo FP-W, An Y, Darzi A, Kinross JM, Purkayastha S, Lo B. Deep learning for instrument detection and assessment of operative skill in surgical videos. IEEE Trans Med Robot Bion. 2022;4(4):1068–71.
7. Fuente López E, García ÁM, Del Blanco LS, Marinero JCF, Turiel JP. Automatic gauze tracking in laparoscopic surgery using image texture analysis. Comput Methods Programs Biomed. 2020;190:105378.
8. Cartucho J, Wang C, Huang B, Elson SD, Darzi A, Giannarou S. An enhanced marker pattern that achieves improved accuracy in surgical tool tracking. Comput Methods Biomech Biomed Eng. 2022;10(4):400–8.
9. Kranzfelder M, Schneider A, Fiolka A, Schwan E, Gillen S, Wilhelm D, Schirren R, Reiser S, Jensen B, Feussner H. Real-time instrument detection in minimally invasive surgery using radiofrequency identification technology. J Surgical Res. 2013;185(2):704–10.
10. Liu Y, Zhao Z, Shi P, Li F. Towards surgical tools detection and operative skill assessment based on deep learning. IEEE Trans Med Robot Bion. 2022;4(1):62–71.
11. Lin T-Y, Goyal P, Girshick R, He K, Dollár P. Focal loss for dense object detection. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017:2980–2988.
12. Ren S, He K, Girshick R, Sun J. Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Trans Pattern Anal Mach Intell. 2017;39(6):1137–49.
13. Redmon J, Divvala S, Girshick R, Farhadi A. You only look once: Unified, real-time object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016:779–788.
14. Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu C-Y, Berg AC. SSD: Single shot multibox detector. In: European conference on computer vision, 2016:21–37.
15. Liu Y, Zhang C, Wu W, Zhang B, Zhou F. MiniYOLO: a lightweight object detection algorithm that realizes the trade-off between model size and detection accuracy. Int J Intell Syst. 2022;37(12):12135–51.
16. Peng J, Chen Q, Kang L, Jie H, Han Y. Autonomous recognition of multiple surgical instruments tips based on arrow obb-yolo network. IEEE Trans Instrum Meas. 2022;71:1–13.
17. Sarki R, Ahmed K, Wang H, Zhang Y. Automated detection of mild and multi-class diabetic eye diseases using deep learning. Health Inform Sci Syst. 2020;8(1):32.
18. Qin F, Li Y, Su Y-H, Xu D, Hannaford B. Surgical instrument segmentation for endoscopic vision with data fusion of cnn prediction and kinematic pose. In: 2019 international conference on robotics and automation (ICRA), 2019:9821–9827.
19. Yamashita K, Kusuda K, Ito Y, Komino M, Tanaka K, Kurokawa S, Ameya M, Eba D, Masamune K, Muragaki Y, et al. Evaluation of surgical instruments with radiofrequency identification tags in the operating room. Surgical Innov. 2018;25(4):374–9.
20. Yang C, Zhao Z, Hu S. Image-based laparoscopic tool detection and tracking using convolutional neural networks: a review of the literature. Comput Assist Surg. 2020;25(1):15–28.
21. Xue Y, Liu S, Li Y, Wang P, Qian X. A new weakly supervised strategy for surgical tool detection. Knowl-Based Syst. 2022;239: 107860.
22. Namazi B, Sankaranarayanan G, Devarajan V. A contextual detector of surgical tools in laparoscopic videos using deep learning. Surg Endosc. 2021;8:1–10.
23. Yang L, Gu Y, Bian G, Liu Y. TMF-Net: a transformer-based multiscale fusion network for surgical instrument segmentation from endoscopic images. IEEE Trans Instrum Meas. 2023;72:1–15.
24. Lin T-Y, Dollár P, Girshick R, He K, Hariharan B, Belongie S. Feature pyramid networks for object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017:2117–2125.
25. Wang C, Zhong C. Adaptive feature pyramid networks for object detection. IEEE Access. 2021;9:107024–32.
26. Li Z, Lang C, Liew JH, Li Y, Hou Q, Feng J. Cross-layer feature pyramid network for salient object detection. IEEE Trans Image Process. 2021;30:4587–98.
27. Hu J, Shen L, Sun G. Squeeze-and-Excitation Networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018:7132–7141.
28. Woo S, Park J, Lee J-Y, Kweon IS. CBAM: Convolutional block attention module. In: Proceedings of the European conference on computer vision (ECCV), 2018:3–19.
29. Hou Q, Zhou D, Feng J. Coordinate attention for efficient mobile network design. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021:13713–13722.
30. Peng Y, Xu Y, Wang M, Zhang H, Xie J. The nnU-Net based method for automatic segmenting fetal brain tissues. Health Inform Sci Syst. 2023;11(1):17.
31. Wang H, Cao P, Yang J, Zaiane O. MCA-UNet: multi-scale cross co-attentional u-net for automatic medical image segmentation. Health Inform Sci Syst. 2023;11(1):10.
32. Lin Z, He Z, Yao R, Wang X, Liu T, Deng Y, Xie S. Deep dual attention network for precise diagnosis of Covid-19 from chest ct images. In: IEEE Transactions on Artificial Intelligence, 2022:1–11.
33. Ni Z-L, Bian G-B, Xie X-L, Hou Z-G, Zhou X-H, Zhou Y-J. RASNet: Segmentation for tracking surgical instruments in surgical videos using refined attention segmentation network. In: 2019 41st annual international conference of the IEEE engineering in medicine and biology society (EMBC), 2019:5735–5738.
34. Liu T, He Z, Lin Z, Cao G-Z, Su W, Xie S. An adaptive image segmentation network for surface defect detection. In: IEEE Transactions on Neural Networks and Learning Systems, 2022:1–14.
35. Li Y, Li Y, He W, Shi W, Wang T, Li Y. SE-OHFM: A surgical phase recognition network with se attention module. In: 2021 International Conference on Electronic Information Engineering and Computer Science (EIECS), 2021:608–611.
36. Shaw P, Uszkoreit J, Vaswani A. Self-Attention with relative position representations. arXiv preprint arXiv:1803.02155 2018.
37. Xu Y, Huang H, Feng C, Hu Y. A supervised multi-head self-attention network for nested named entity recognition. Proc AAAI Conf Artif Intell. 2021;35:14185–93.

38. Twinanda AP, Shehata S, Mutter D, Marescaux J, De Mathelin M, Padoy N. EndoNet: a deep architecture for recognition tasks on laparoscopic videos. IEEE Trans Med Imaging. 2016;36(1):86–97.
39. Zheng Z, Wang P, Liu W, Li J, Ye R, Ren D. Distance-IoU loss: Faster and better learning for bounding box regression. In: Proceedings of the AAAI conference on artificial intelligence, 2020:12993–13000.
40. Zhou X, Wang D, Krähenbühl P. Objects as points. arXiv preprint arXiv:1904.07850 2019.
41. Tan M, Pang R, Le QV. Efficientdet: Scalable and efficient object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020:10781–10790.
42. Carion N, Massa F, Synnaeve G, Usunier N, Kirillov A, Zagoruyko S. End-to-end object detection with transformers. In: European conference on computer vision, 2020:213–229.
43. ultralytics: yolov5. https://github.com/ultralytics/yolov5
44. Ge Z, Liu S,Wang F, Li Z, Sun J. YOLOX: Exceeding yolo series in 2021. arXiv preprint arXiv:2107.08430 2021.
45. Li C, Li L, Jiang H, Weng K, Geng Y, Li L, Ke Z, Li Q, Cheng M, Nie W et al. YOLOv6: A single-stage object detection framework for industrial applications. arXiv preprint arXiv:2209.02976 2022.
46. Lv W, Xu S, Zhao Y, Wang G, Wei J, Cui C, Du Y, Dang Q, Liu Y. Detrs beat yolos on real-time object detection. arXiv preprint arXiv:2304.08069 2023.
47. Sarikaya D, Corso JJ, Guru KA. Detection and localization of robotic tools in robot-assisted surgery videos using deep neural networks for region proposal and detection. IEEE Trans Med Imaging. 2017;36(7):1542–9.
48. Shi M, Shen J, Yi Q, Weng J, Huang Z, Luo A, Zhou Y. LMFFNet: a well-balanced lightweight network for fast and accurate semantic segmentation. IEEE Trans Neural Netw Learn Syst. 2023;34(6):3205–19.
49. Shi X, Chen Z, Wang H, Yeung D-Y, Wong W-K, Woo W-c. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. Advances in Neural Information Processing Systems. 2015;28.
50. Xu H, Xie H, Tan Q, Zhang Y. Meta semi-supervised medical image segmentation with label hierarchy. Health Inform Sci Syst. 2023;11(1):26.