IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received December 19, 2021, accepted February 23, 2022, date of publication March 8, 2022, date of current version March 24, 2022.

*Digital Object Identifier 10.1109/ACCESS.2022.3157330*

# Automatic Abdominal Hernia Mesh Detection Based on YOLOM

**SIQI CHEN[1,2], JINLI XU[1], JINHUA YU[1], (Member, IEEE), JUN WU[2], AND GUOHUI ZHOU[1]**
[1]Center for Biomedical Engineering, School of Information Science and Technology, Fudan University, Shanghai 200433, China
[2]Department of Electronic Engineering, Yunnan University, Kunming 650500, China

Corresponding authors: Jun Wu (wujun@ynu.edu.cn) and Guohui Zhou (zhough@fudan.edu.cn)

This work was supported in part by the Natural Science Foundation of China under Grant 61661050 and Grant 61561050; and in part by the Priority Union Foundation of Yunnan Provincial Science and Technology Department, Kunming Medical University, under Grant 2018FE001(-143).

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by Medical Ethics Committee, School of Medicine, Yunnan University, and performed in line with the relevant provisions of medical ethics.

**ABSTRACT** As a new 3-D ultrasound imaging method, an automated breast ultrasound (ABUS) has been widely used in breast abnormality examinations. Because of its excellent 3D visualization, ABUS is also well suited to the detection of an abdominal wall hernia mesh. Due to the inherent low signal-to-noise ratio of ultrasound imaging and the large amount of data generated during ABUS scanning, mesh detection based on subjective observation is extremely time-consuming and prone to missed detection. Therefore, we proposed a novel abdominal hernia wall mesh detection method based on the you only look once version 3 (YOLOv3) method named the YOLOv3 for mesh (YOLOM) method to detect abdominal wall hernia mesh to speed up the ABUS reading process. To make a YOLOM method with a good detection efficiency, we utilized a lightweight cross stage partial attention network (CSPA-Net) as the backbone and applied a feature enhancement network (FEP-Net) to boost the mesh detection accuracy. An improved loss function with completed intersection-over-union (CIoU) and the Swish activation function were also employed to optimize the proposed YOLOM method. We designed ablation study to verify the validity of the proposed method. The average mesh detection precision reached 98.36%, which was 12.51% and 2.35% higher than that of the YOLOv3 and you only look once version 4 (YOLOv4) methods, respectively. The experimental results and comparisons demonstrated that the proposed YOLOM detector is efficient for abdominal wall hernia mesh detection.

**INDEX TERMS** Abdominal wall hernia mesh, automated 3-D ultrasound, YOLO, attention mechanism.

## I. INTRODUCTION

An abdominal wall hernia is one of the most common complications of abdominal surgery. According to clinical data statistics, the probability of an abdominal wall hernia occurring is 2% - 11%. If infection occurs, the probability of occurrence will increase to 23% [1]–[3]. Abdominal wall hernias do not heal on their own and gradually expand outward from the infected area; therefore, prompt surgery is the only treatment option [4], [5]. Mesh repair surgery has become the standard procedure for repairing abdominal wall hernias worldwide. However, a variety of mesh-related complications, such as mesh infection, migration, hematoma and intestinal adhesion are possible [6], [7]. Therefore, the correct preoperative detection of an abdominal wall hernia mesh can help surgeons adjust the surgical plan to predict the difficulty of an abdominal wall hernia mesh surgery and reduce the incidence of related complications or the removal of a previous mesh.

With the rapid development of mesh materials, meshes have become increasingly lightweight. Lightweight (LW) mesh is the first choice for abdominal hernia surgery. However, because the detection range of a handheld ultrasound (HHUS) probe is relatively narrow, it is impossible to completely and reliably detect and identify a LW mesh at the same time. It is difficult to detect meshes in either the axial or sagittal plane using 2-D ultrasound due to the light and thin characteristics of the LW mesh. As a new 3-D ultrasound

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

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

imaging method, ABUS provides more comprehensive diagnostic information through the coronal plane [8], [9] and has been a concern of sonographers and scholars [10]–[14]. Compared to narrowing the HHUS detection range, the ABUS scanning range has been greatly improved, but in practical applications, ABUS may need to scan repeatedly to obtain an image of a large area, which leads to the amount of data generated by ABUS for each patient being tremendous. This causes the following two problems in the detection of a LW mesh: 1) It is time-consuming and labor-intensive to manually examine the ABUS ultrasound images. 2) The detection accuracy is heavily dependent on the experience of sonographers, which easily leads to missed diagnoses and misdiagnosis. Therefore, imaging studies have become important for the detection and identification of abdominal wall hernia meshes, which can provide effective guidance for follow-up surgery and treatment.

Two-stage detectors, such as a region-based convolutional neural network (R-CNN) [15], Fast R-CNN [16] and Faster R-CNN [17], are performed in two stages. In the first stage, a region proposal network is used to process images and generate box proposals where objects may exist. In the second stage, these box proposals are used as features from the intermediate feature maps. Then, these features are fed to the final layers to localize and classify the objects of each box proposal. However, two-stage detectors usually use more proposal regions, which help to obtain local optimal solutions and improve detection accuracy at the cost of longer computational time. In contrast, single-stage detectors, such as the You Only Look Once (YOLO) [18]–[21] and Single Shot Multibox Detector (SSD) [22] algorithms, are usually faster and yield less desirable results than two-stage detectors. The YOLO series is a typical single-stage detection method. Compared with the two-stage algorithms, the YOLO series algorithm does not require the region proposals stage, and directly predicts the category probability and location information of the target. It transforms an object detection problem into a regression problem, which can completely achieve end-to-end detection. However, the limitations of the YOLOv3 [20] algorithm for object detection on ultrasound images are as follows: (1) The performance in detecting small objects is often not good when the image noise is large, the resolution is low and the background is complex. (2) The YOLOv3 method fuses all low-level features directly with high-level features. In fact, not all the low-level detail features are beneficial to detection. (3) The aspect ratio of the bounding box is different from the ground truth, which is not conducive to postoperative evaluation of the implanted mesh. Recently, the YOLOv4 [21] algorithm, which uses the cross-stage partial network as the backbone network to extract the feature and applies a large number of data augmentation techniques, has received widespread attention. However, the number of parameters and the module storage size of the YOLOv4 algorithm are much larger than those of the YOLOv3 method, which increases deployment costs and reduces training and reasoning speed.

To solve these problems, we proposed a novel real-time detector for abdominal wall hernia mesh based on the YOLOv3 algorithm named the YOLOM method to improve the detection efficiency. Our main contributions are as follows:

1) In the feature extraction process, we introduce a cross-stage partial network to design a more lightweight feature extraction network, which enhances the feature extraction capability of the backbone, improves the detection effect for small targets, and reduces the amount of calculation and model storage size. At the same time, we introduce a channel attention mechanism SE-Net to make the network pay more attention to the features of the mesh target during the feature extraction process.

2) Multiscale spatial pyramid pooling (SPP) is added to the convolutional layer at the end of the proposed backbone. The SPP block performs pooling operations on the input feature map at different scales and connects the three pooled feature maps and the input feature map to increase the receptive field of the feature map in such a way that the YOLOM method can detect the object more comprehensively.

3) To ensure the consistency of the bounding box aspect ratio and accelerate the network convergence, we introduce Complete-IoU (CIoU) to optimize the YOLOv3 loss function. The shape of the postoperative mesh is an important guide for doctors to evaluate the condition of the implanted mesh and the recovery of the hernia area.

The rest of this paper is organized as follows. The related work is introduced in the second part. The third part presents the overview of the proposed YOLOM detection method, including the motivation and the structure of the YOLOM method. In part four, experimental results and discussions are given, where the billion floating point operations per second (BFLOP/S), the model size, mean average precision (mAP), the different subscripts of mAP represent mAP calculated under different IoU conditions. The real-time performance of frames per second (FPS) are compared. Furthermore, the comparison results between the YOLOM and other state-of-the-art detection algorithms are given in part four. Finally, conclusions are drawn in part five.

## II. RELATED WORK

### A. MESH DETECTION DATASET

An ABUS database was collected from three types of experiments: gelatin phantom experiments, animal (porcine peritoneal) ex vivo experiments and patient experiments. There were three sets of gelatin phantom experiments, five sets of animal ex vivo experiments, and data from 97 patients. Signed informed consent from patients was waived because it was a retrospective study. The study was approved by the Ethics Research Committee of our institute. All the images were taken from the ABUS scanner. The built-in linear array probe model is 14L5BV, the frequency range is 5 MHz to 15 MHz, and the maximum scanning depth is 6 cm. Each scan produces 318 frames of axial images, 730 frames of sagittal images and 573 frames of coronal images. As shown in Fig. 1,

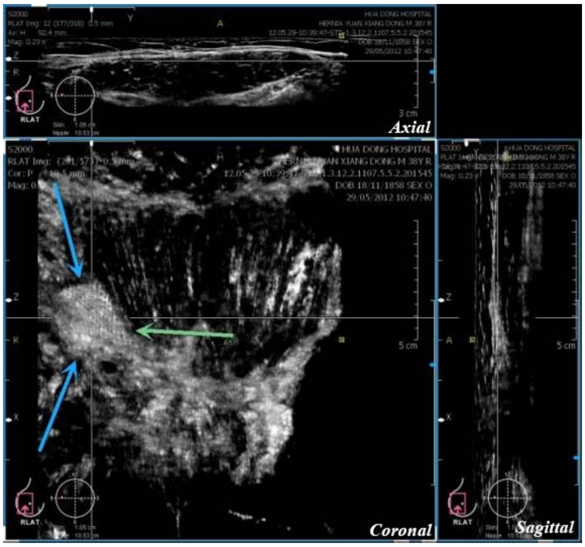


**FIGURE 1.** An ABUS image of the incision site with LW mesh in a 45-year-old male patient.

3-D ultrasound shows the side of the LW mesh in the axial and sagittal planes. Because the LW mesh is very thin (the thickness of the LW mesh is only 0.5 mm), doctors usually choose the coronal plane to detect and evaluate the mesh in clinical diagnosis. In this paper, we also used multilayer coronal images for mesh recognition and detection.

### B. THE INTRODUCTION OF YOLOV3

The YOLOv3 method is composed of the newly designed backbone network Darknet-53 and a multiscale detection network, as shown in the Appendix part A.

The main idea of Darknet-53 is to use five continuous downsampling modules to change the size of the resolution of the input image from 416 × 416 to 13 × 13. To solve the gradient problem caused by network deepening, the residual module in ResNet is introduced in each downsampling module. To preserve more image information, the YOLOv3 method uses a convolution stride of 2 instead of the traditional max-pooling layer. Each down-sampling module contains a different number of stacked residual units. The method of multi-scale prediction based on feature pyramid networks (FPNs) is used to predict the multi-size targets in the YOLOv3 algorithm, especially for small object detection. Darknet-53 generates three feature maps of different sizes. The prediction results of YOLOv3 are the relative positional relationship between the anchors and the bounding boxes. Anchor is a set of boxes with only width and height parameters obtained by k-means clustering on the ground truth (GT) of the dataset. We can convert the prediction results of YOLO into bounding boxes through (1) to (4).

$$b_x = \sigma(t_x) + c_x \tag{1}$$

$$b_y = \sigma(t_y) + c_y \tag{2}$$

$$b_w = p_w e^{t_w} \tag{3}$$

$$b_h = p_h e^{t_h} \tag{4}$$

where $t_x, t_y$ are the offset of the coordinates, and $t_w, t_h$ are the ratio coefficients of the width and height of the bounding boxes relative to the anchors, respectively. $p_w, p_h$ are the weight and height of the anchors. We can divide the image into $N \times N$ grid cells according to the size of the feature map, and $c_x, c_y$ are the coordinates of the upper left corner of the grid cell where the feature map is located. $\sigma$ is the sigmoid activation function, which can normalize the coordinate offset between 0 and 1.

### C. SQUEEZE AND EXCITATION NETWORK

For ultrasound images, to disregard invalid information in images more efficiently and guide where the network should pay attention, the selection of a suitable attention mechanism in the CNNs is important. Jie *et al.* designed a lightweight gating mechanism named Squeeze-and-Excitation (SE) network to improve the expression ability of the whole network [23], which established the relationship between channels using an efficient fully connected layer. The structure of the SE module is shown in Fig. 2, which can be roughly divided into three stages: squeeze, excitation and combination. And we put the details in Appendix part B.

### D. SWISH ACTIVATION FUNCTION

The choice of activation functions in deep networks has a significant effect on the training dynamics and detection performance [24]. In 2017, Ramachandran P *et al.* proposed the Swish activation function to speed up network convergence and improve classification accuracy. In this study, we used Swish in the residual module. Swish is defined as (5),

$$f(x) = x(1 + exp(-\beta x)^{-1}) \tag{5}$$

which means that the result is obtained by multiplying the input value with the sigmoid activation function. $\beta$ is either a trainable parameter or a constant parameter. The characteristics of a good activation function should generally be smooth and robust to negative values. ReLU function can solve the gradient disappearance problem when the input $x > 0$, but it sets all the values of the input as negative numbers to 0. If an outlier appears, the biases of the neural network are likely to become very large, making subsequent normal inputs all become negative numbers, which will make the parameters no longer update. However, as an activation function between linear function and ReLU function, Swish can effectively alleviate this problem. If $\beta = 0$, Swish becomes the linear function $f(x) = x/2$, and as $\beta \to 0$, the sigmoid component approaches as a function from 0 to 1, so Swish becomes like the ReLU function.

## III. PROPOSED METHOD

### A. YOLOM NETWORK

The entire YOLOM detector architecture is shown in Fig. 3. The detector mainly consists of three modules: feature extraction, feature enhancement and multiscale detection. First, we replaced the traditional Darknet-53 network with a cross-stage partial attention network (CSPA-Net). To extract deeper

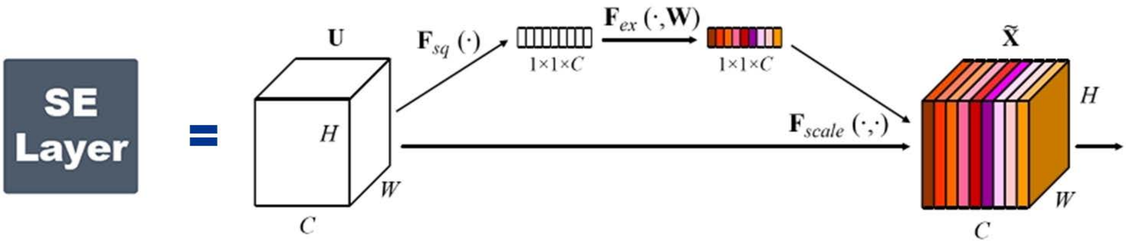


**FIGURE 2.** The structure of SE module.

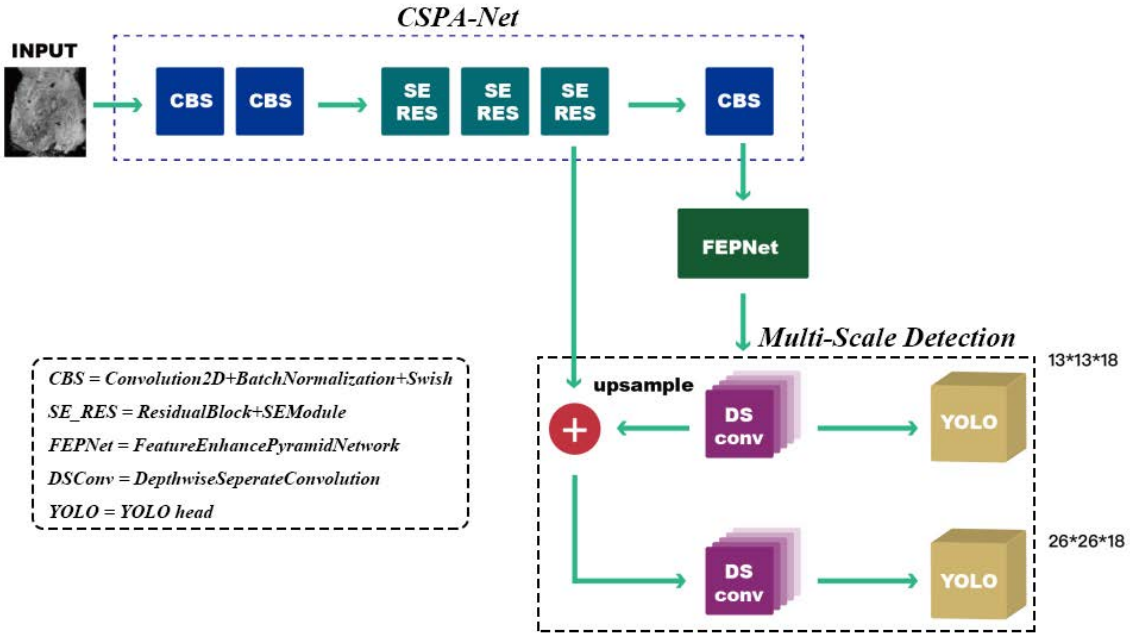

CSPA-Net
INPUT
CBS
CBS
SE RES
SE RES
SE RES
CBS
FEPNet
Multi-Scale Detection
13*13*18
upsample
DS conv
YOLO
26*26*18
DS conv
YOLO
CBS = Convolution2D+BatchNormalization+Swish
SE_RES = ResidualBlock+SEModule
FEPNet = FeatureEnhancePyramidNetwork
DSConv = DepthwiseSeperateConvolution
YOLO = YOLO head

**FIGURE 3.** The entire architecture of YOLOM.

semantic information, we designed a feature-enhanced pyramid network (FEP-Net). The images are input to the CSPA-Net to extract two feature maps with different sizes. Then, in the feature enhancement stage, the feature map obtained from the last residual block in the feature extraction module is input to the FEP-Net module to obtain a more efficient feature. In the multiscale detection stage, the two feature maps with different sizes obtained from the residual blocks in the feature extraction module are up-sampled and concatenated to obtain feature maps with different receptive field sizes. The sizes of the two feature maps are $13 \times 13$ and $26 \times 26$, respectively. Finally, we used the YOLO head of the YOLOv3 method to generate the bounding boxes. In this paper, the number of category is 1, and the channel number of each feature map is 18. Therefore, the YOLOM method generated 2,382 proposals on each abdominal wall hernia mesh image, while 8,265 proposals were reduced compared to the YOLOv3 method.

### B. CSPA-NET: FEATURE EXTRACTION NETWORK

CNNs often face the problem of too much calculation in the training and detection processes, which directly leads to model training and detection slowdown. Cross-stage partial network (CSP-Net) [25] can restrict the variability of the gradients by integrating feature maps from the beginning and the end of a network stage, which reduces computations by 20% with equivalent or even superior accuracy. And it is easy to implement and general enough to cope with architectures based on ResNet. To reduce the amount of calculation while keeping or even improving accuracy, we proposed a new backbone based on a CSPNet. We divide the input in the channel dimension into two: one part is extracted through a residual network (ResNet), and then the two parts of the feature map are feature-fused to achieve a lightweight backbone. Due to the heavy computation and high complexity of Darknet-53, a novel lightweight backbone network CSPA-Net is proposed. CSPA-Net is composed

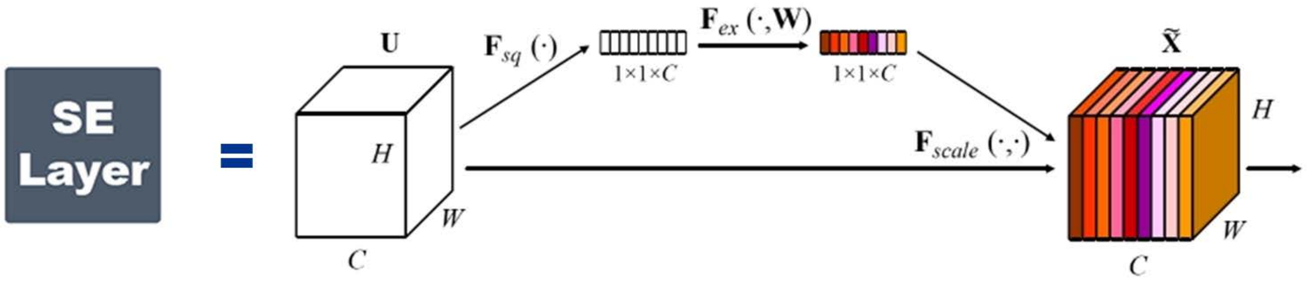


**FIGURE 4.** The structure of SE_Res module.

of CBS modules and SE_Res modules. A CBS module is a 3 × 3 convolution kernel, batch normalization and a Swish activation function. A SE_Res module is shown in Fig. 4. At the shortcut connection, we divide the input in the channel dimension into two, and half goes through the downsampling operation of the CBS modules, followed by feature fusion through concatenation. The shortcut operation does not increase any parameters or computational complexity. The SE module restricts the interdependence between the channels before the feature fusion process of the shortcut and adaptively recorrects the corresponding strength of the features between the channels through the global loss function of the network. Here, the max-pooling window size is 2.

### C. FEP-NET:FEATURE ENHANCED PYRAMID NETWORK

The YOLOv3 algorithm utilizes the global features of different convolutional layers of the network but does not make full use of the multiscale local region features of the convolutional layer. SPPNet [26] is able to fuse the receptive fields of different sizes and improve the scale invariance of the network, so that the detector has better robustness to mesh targets of different sizes. To effectively make use of the local region features of the backbone, we proposed a feature-enhanced pyramid network (FEP-Net) based on a SPPNet to fuse the multiscale local and global features, as shown in Fig. 5. Here, the multiscale SPP block is composed of three max-pooling layers, and the size of the pooling window can be computed from (6).

$$Size_p = \frac{Size_f}{n_i} \tag{6}$$

where $Size_p$ represents the size of the pooling windows, and $Size_f$ represents the size of the feature maps, and $n_i = 1, 2, 3$. Due to the large amount of convolution operation parameters of the deep neural network, the inference speed of the neural network are reduced. Therefore, we introduced the depthwise separable convolution (DSconv) [27] into this module. And we can obtain the pooling windows from (6) are 1, 5, 9, 13, respectively. The strides of the pooling windows are all 1, and the input feature maps are padded with 0 to ensure that the output feature maps after pooling are the same size as the input.

A DSconv can greatly decrease the number of parameters. Hence, the computation time and model size are reduced. We can factorize a normal convolution into a DSconv. A DSconv includes depthwise and pointwise layers based on the dotted box in Fig. 5. The former carries out a single-channel convolution operation on the input, but such an operation does not make use of the spatial information between the channels of the input feature map, so the latter carries out a convolution operation on the former results in the depth direction to ensure that the number of layers of the network can be deepened and the performance of the network can be improved while reducing the amount of convolution operation computation.

To make the network have a better detection effect on mesh targets of different sizes, we fused feature maps of different scales. We used five DBL modules to enhance the feature map output by FEP-Net, and used up-sampling to change its size from 13 × 13 to 26 × 26, and concatenated it to the output of the last SE_Res block. The result also went through five DBL modules that were used for feature fusion, and finally, we obtained two feature maps with sizes of 13 × 13 × 18 and 26 × 26 × 18.

### D. IMPROVING YOLOV3 LOSS FUNCTION WITH CIoU

The YOLOv3 loss function is a linear sum of three parts: the coordinate loss, classification loss and confidence loss. The loss function can be denoted by (7)

$$Loss = Loss_{coord} + Loss_{conf} + Loss_{class} \tag{7}$$

where the $Loss_{coord}$ denotes the coordinate loss, the confidence loss is presented by $Loss_{conf}$, and $Loss_{class}$ is calculates the classifying loss. The $Loss_{coord}$ of the YOLOv3 method regards (w,h) and (x,y) in (1) as independent variables for loss calculation. In fact, there is a certain spatial constraint relationship between the center point coordinates and the width and height between the bounding box and the GT. Using a traditional IoU to improve the loss function will cause the loss function to not be a derivative and make the network training unable to converge if the bounding box and the GT do not stack or if the bounding box includes the GT. To overcome these disadvantages, we introduced intersection

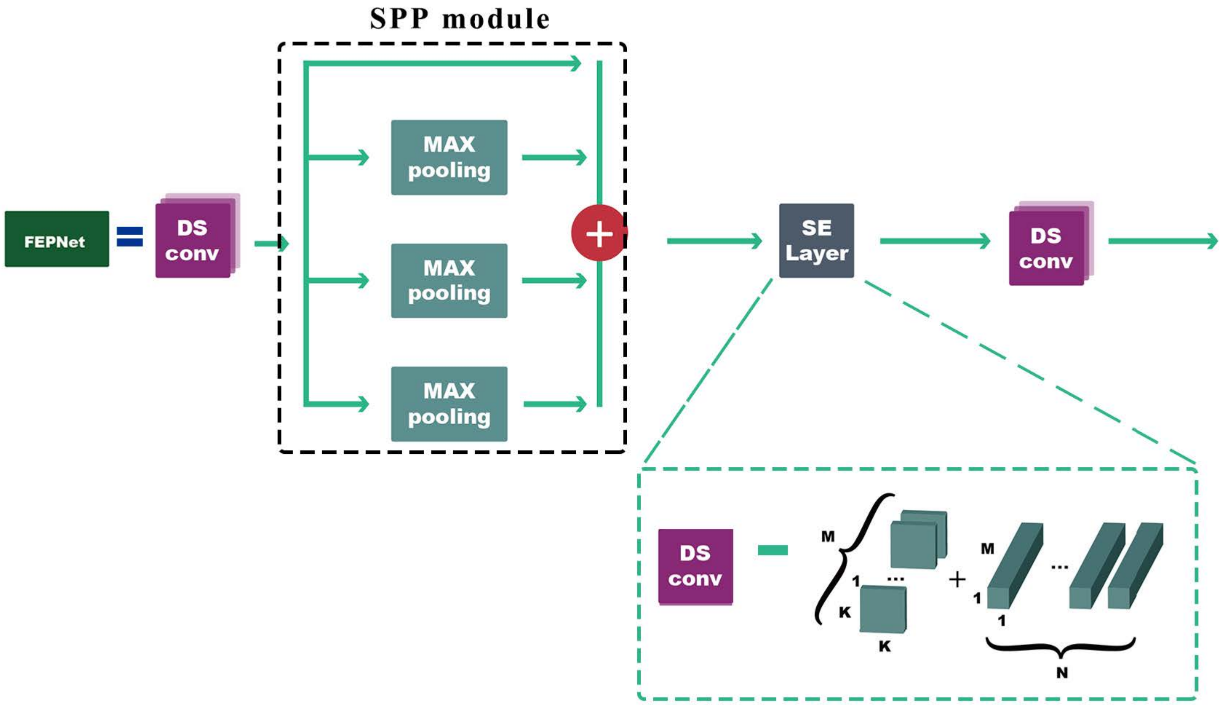


**FIGURE 5.** The structre architecture of FEP-Net.

over union (CIoU) [28] by considering three geometric measures, overlap area, central point distance and aspect ratio, which better describe the regression of the bounding box. Therefore, in our research, CIoU is utilized to improve and modify $Loss_{coord}$. As shown in the Appendix part C, the purple, gray and yellow rectangles represent the bounding box, the GT and the smallest enclosing box covering two boxes, respectively. The improved $Loss_{coord}$ using CIoU is as in (8),

$$\begin{aligned} Loss_{CIoU} &= 1 - IoU + \mathcal{R}_{CIoU} \\ IoU &= \frac{BoudingBox \cap GroundTruth}{BoudingBox \cup GroundTruth} \\ \mathcal{R}_{CIoU} &= d/C^2 + \alpha v \\ v &= \frac{4}{\pi^2}(\arctan\frac{w^{gt}}{h^{gt}} - \arctan\frac{w}{h})^2 \\ \alpha &= \frac{v}{(1 - IoU + v)} \end{aligned} \tag{8}$$

where $C$ is the diagonal length of the smallest enclosing box covering two boxes, and $d = \rho(b, b^{tg})$ is the Euclidean distance between the central points of two boxes. $IoU$ is the intersection-over-union between the bounding box and the GT, which constrains the overlapping area between the bounding box and the GT, $R_{CIoU}$ uses $d$ and $C$ to address the problem of the loss possibly not being able to update the gradient when the bounding box and the GT are not stacked, $\alpha$ is the scale factor, $v$ ensures the consistency of the aspect ratio by calculating the diagonal slopes of the bounding box and the GT.

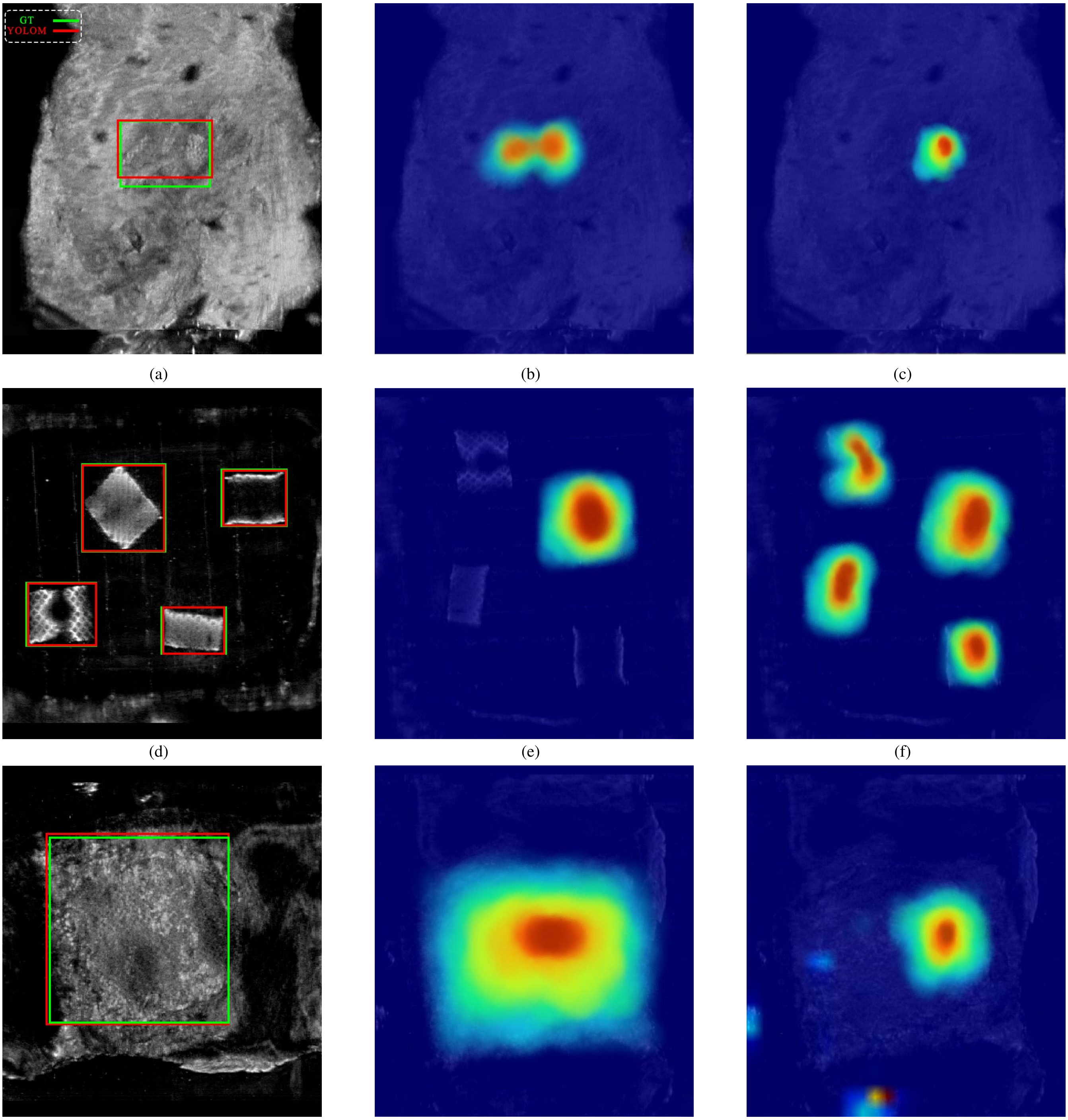
## IV. EXPERIMENTAL ANALYSIS

In this paper, all experiments are conducted on a Windows 10 (64-bit) Dell workstation with 64 GB of memory an Intel(R) Xeon(R) E5-2650 V3 2.30 GHz CPU and an NVIDIA Titan XP GPU with 12.0 GB video memory. The deep learning framework was PyTorch, and the map and precision-recall curve were used to evaluate the proposed method. Based on a priori knowledge of the abdominal thickness range, with the removal of a large number of frames unlikely to contain mesh, we collected 2100 original coronal images, divided the images into a model building set and an independent testing set using a ratio of 9:1, and further divided the model building set into a training set and a validation set using a ratio of 9:1. To ensure that the trained model has certain generalization, we adapted data augmentation techniques including rotation, flipping, scaling, and random cropping in the training set to obtain 13,608 images for model training. In addition, Adam was used for gradient optimization, and the initial learning rate was set to 0.0001.

### A. ALGORITHM COMPLEXITY

Operational efficiency is critical to the implementation of an algorithm, so we compare the algorithm complexity in terms of time complexity and space complexity, and the formulas are shown in (9),

$$\begin{aligned} M &= (X - K + 2 * Padding)/Stride + 1 \\ Time &= O(\sum_{l=1}^{D} M_l^2 \cdot K_l^2 \cdot C_{l-1} \cdot C_l) \end{aligned}$$



(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

**FIGURE 6.** **The detection and visualization results. (a), (d), (g) are the detection results of YOLOM method, (b), (e), (h) are the Grad-CAM visualization results obtained through the 13 × 13 feature map, (c), (f), (i) are the Grad-CAM visualization results obtained through the 26 × 26 feature map.**

$$Space = o\left(\sum_{l=1}^{D} K_l^2 \cdot C_{l-1} \cdot C_l + \sum_{l-1}^{D} M^2 \cdot C_l\right) \tag{9}$$

where $M$ is the width or height of the output feature map of the kernel, $K$ is the width or height of each kernel, respectively, $D$ represents the depth of the network, $C_l$ represents the number of channel of each kernels. Time complexity is the number of operations of a model, and we use floating-point operations (FLOPs) and multiply accumulated operations (MACCs) for evaluation. Time complexity determines the time for model training and detection, and too much time complexity will seriously affect the model training and detection speed. Space complexity refers to the number of parameters of a model, and we use the number of parameters,

**TABLE 1.** Compare the YOLOM algorithm complexity with other SOTA.

| Model | FLOPs | MACCs | Parameters | Model Size | MemR+W |
|---|---|---|---|---|---|
| Faster R-CNN | 126.41G | 253.14G | 136,689,024 | 253.14G | 2145MB |
| SSD | 34.92G | 69.34G | 136,689,024 | 337.91MB | 556MB |
| YOLOv3 | 33.05G | 65.98G | 61,949,149 | 443.61MB | 1100MB |
| YOLOv4 | 29.95G | 59.85MB | 64,040,001 | 606.54MB | 1120MB |
| YOLOM | 7.12G | 14.23G | 18,962,884 | 98.04MB | 231MB |

model size and memory read and write (MemR+W) for evaluation. Networks with higher spatial complexity have a large number of parameters, and a large amount of data is required to train the network. However, a real dataset is usually not too large, which makes the model prone to overfitting. Table 1 shows the algorithmic complexity comparison between the YOLOM and other state-of-the-art (SOTA) methods. The algorithm complexity of the two-stage detector Faster R-CNN far exceeds that of the one-stage detectors in both time and space. For the one-stage detector, our proposed method reduces the time complexity by 80% and greatly reduces the space complexity.

### B. VISUALIZATION OF the MULTISCALE FEATURE MAPS

To illustrate the effectiveness of the multiscale feature maps, we visualize the classification information of the feature maps in two layers by Gradient-weighted Class Activation Mapping (Grad-CAM) [29]. As shown in Fig. 6, the first column shows the detection results and the ground truth of meshes (the red boxes denote the detection results, and the green boxes denote the ground truth), and the last two columns show the locations of clusters with sizes of 13 and 26 our method finds in the feature maps. It is clear that the multiscale features can detect meshes of different sizes better. The 13 × 13 feature map has large perspective field so that it can detect large meshes more efficiently and the 26 × 26 feature map can detect tiny meshes well. Finally, we use a nonmaximum suppression (NMS) algorithm to filter the detection results at different scales to obtain the final mesh detection results. The step of NMS is to select the bounding box with the highest confidence, and then calculate the IoU in pairs with other bounding boxes, filter out the bounding boxes with IoU greater than the threshold, and iterate until only the last bounding box is left.

### C. ABLATION STUDY

In order to verify the effectiveness of each module, we conducted an ablation study. In Table 2, the effects of CSPA-Net, FEP-Net, CIoU, and the Swish activation function on the YOLOM method are mentioned. According to this table, when the YOLOv3 backbone is replaced with CSPA-Net, the mean average precision (mAP) of the YOLOM method increases from 85.85% to 89.59%, which fully shows that the introduction of the cross-stage partial network and SE module into the backbone can make the network better learn the characteristics of a mesh. The mAP increased from 89.59% to 95.47% when we used FEP-Net in our method. This proves that FEP-Net can expand the receptive field of the feature map and is effective for multiscale detection. Moreover, the mAP increased by 1.87% when the YOLOM method uses CIoU to improve the loss function. CIoU loss takes three geometric properties into account, the overlap area, central point distance and aspect ratio, and leads to faster convergence and better performance. The Swish activation function improved the mAP by 1.02%, and we can conclude that using Swish instead of ReLU can improve the detection accuracy without changing any network structure. The final mAP obtained on the mesh dataset is 98.36%, which is 12.51% higher than the original YOLOv3 method. As a result, the components we used increase the mAP.

**TABLE 2.** The ablation experiments analyses of YOLOM.

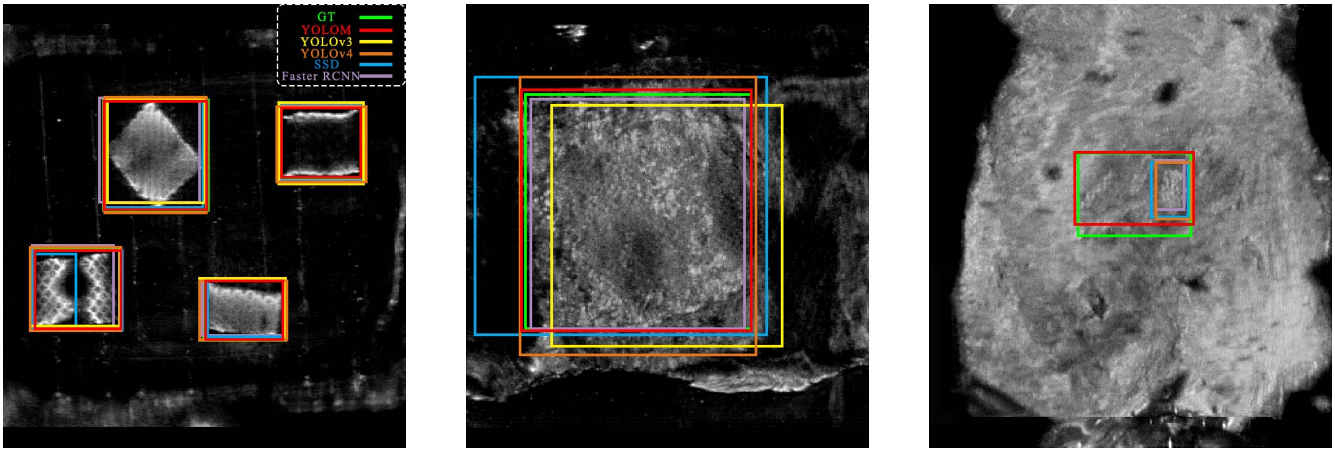
| Backbone YOLOv3 | CSPA-Net | FEP-Net | CIoU | Swish | $mAP_{50}$ |
|---|---|---|---|---|---|
| ✓ | | | | | 85.85 |
| | ✓ | | | | 89.59 |
| | ✓ | ✓ | | | 95.74 |
| | ✓ | ✓ | ✓ | | 97.34 |
| | ✓ | ✓ | ✓ | ✓ | 98.36 |

### D. DETECTION SPEED

In Table 3, the detection time of the proposed YOLOM detector and other SOTA detector algorithms is shown. According to this table, the average detection time of the YOLOM detector on an Nvidia TITAN Xp is 21.4 ms for each test image. In addition, because of the low parameters of the proposed detector, it is capable of running on a CPU, and its detection time is also very short. However, the detection time of the original YOLOv3 method on the same GPU is 50.5 ms. The two-stage detector Faster R-CNN has the longest detection time. Therefore, when using GPU for detection, YOLOM's detection speed is approximately three times faster than YOLOV3, which meets the requirements of real-time ultrasound image detection.

### E. VISUAL RESULTS

The detection results of the YOLOM and other SOTA detectors for three test images are visualized in Fig.7. According to this figure, it can be seen that the proposed YOLOM detector has a perfect ability to detect mesh objects using ABUS. As seen in Fig. 7(a), the SSD recognizes the more



(a) (b) (c)

**FIGURE 7.** The visualization detection result of the proposed YOLOM and other SOTA detectors for three test images.(a) is a gelatin phantom experiments image; (b) is an animal (porcine peritoneal) ex vivo experiments image; and (c) is a patient experiments image.

**TABLE 3.** Comparing the detection time of the proposed YOLOM with other SOTA detector algorithms.

| Model | Time[$ms$] | FPS |
|---|---|---|
| Faster R-CNN | 384.7 | 2.6 |
| SSD | 58.5 | 17.1 |
| YOLOv3 | 50.5 | 19.8 |
| YOLOv4 | 76.9 | 13.0 |
| YOLOM | 21.4 | 46.6 |

curved mesh target as two targets. As shown in Fig. 7(c), the YOLOv3 method did not completely detect the mesh target, and the SSD detected the mesh target with a clear mesh structure as a single mesh target, while the YOLOv4 method only detected the mesh with a clear mesh structure. At the same time, it is not difficult to find that our method is more similar to the ground truth shape and position and has a better detection effect on small target objects with subtle differences between the foreground and background.

### *F. COMPARASION WITH OTHER SOTA METHODS*

The performance comparison between the YOLOM detector we proposed and the SOTA detectors is shown in Table 4. All experiments in this paper are conducted on the same test dataset. It can be seen from the table that the method proposed in this paper has a perfect detection effect for the mesh dataset, and can fully ensure the high efficiency and practicability of clinical auxiliary diagnosis. Compared with the original YOLOv3 algorithm and the latest YOLOv4 algorithm, the $mAP_{50}$ of our proposed method is 12.51 and 2.35 higher than them, respectively, and at the same time greatly improves the detection speed.

The Precision-Recall curve (PRC) is an important indicator for evaluating the object detection model. PRC is a curve representing precision and recall rate under different confidence thresholds. In the dataset, we define the label as mesh as True (T), and the label as mesh as False (F); in the prediction result, the confidence higher than the threshold is defined as the correct classification as positive (P), otherwise

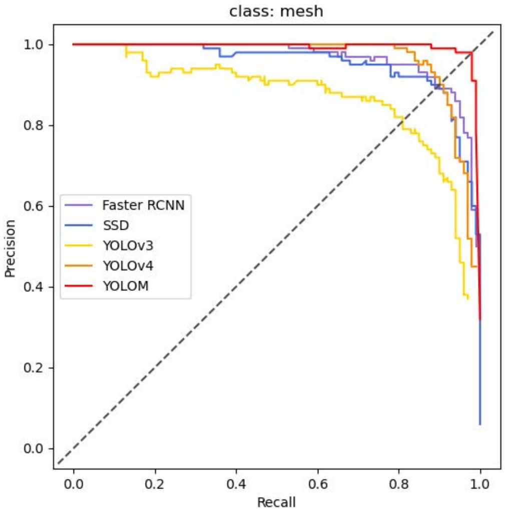
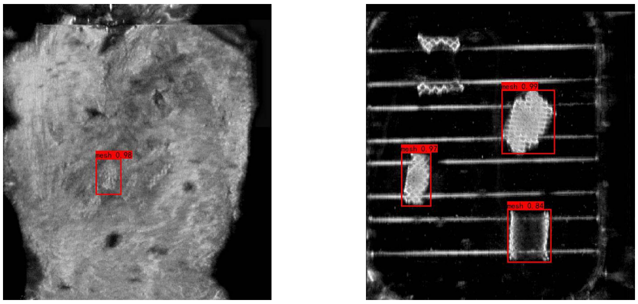


**FIGURE 8.** The Precision-Recall curve of the proposed YOLOM with other SOTA detectors.



(a) (b)

**FIGURE 9.** Failure cases of YOLOM detection. (a) is a mesh implanted in the abdominal cavity close to the fascia; (b) is an in vitro images.

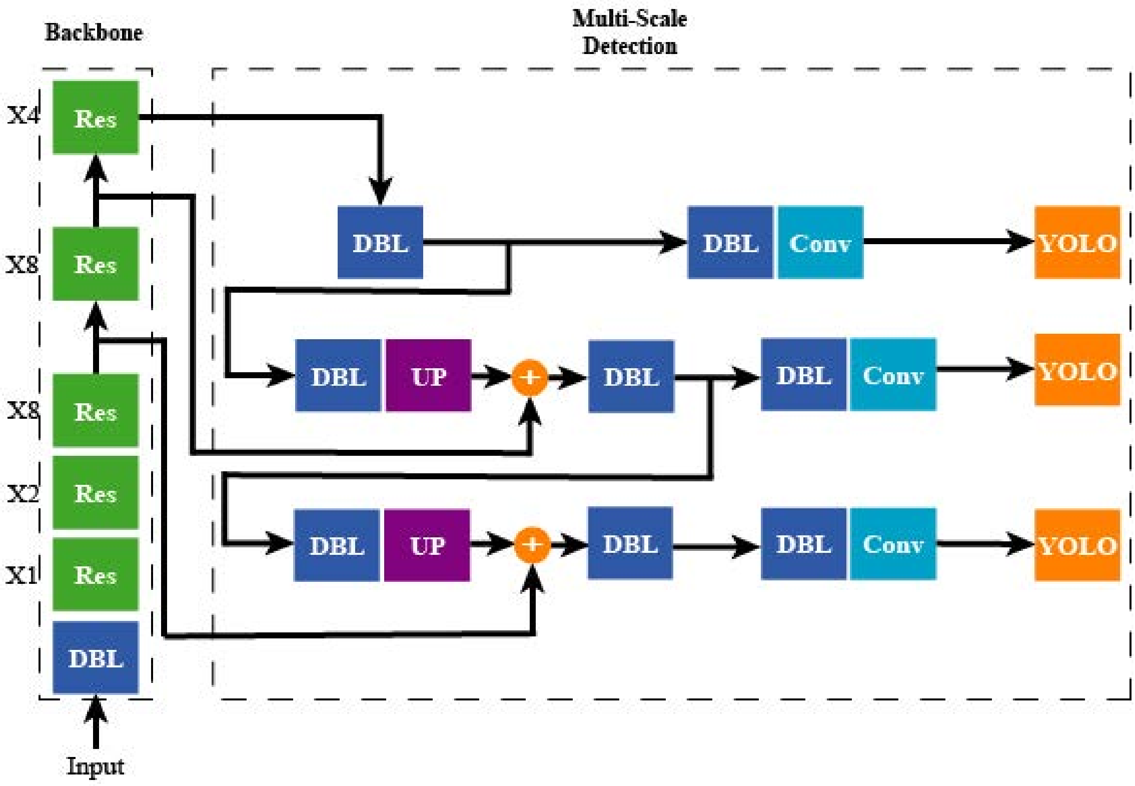


**FIGURE 10.** The entire structure of YOLOv3.

it is negative (N). As show in (10), precision represents the percentage of samples with a label of true among the samples predicted to be positive. And recall means the percentage of samples predicted to be positive among the samples with a label of true. The mesh PRC of are illustrated in Fig. 8. According to this figure, the detection precision is high for most test images. From the PRC, we can see that the YOLOM detection of mesh objects has the best effect because it encompasses all the SOTA algorithm PRC. The performance of the Faster R-CNN, SSD and YOLOv4 methods are almost the same, but it can be seen from their intersection with the black dotted line that the performance of the YOLOv4 method is slightly higher than the other two. The detection performance of the YOLOv3 method for mesh targets is not as good as the other SOTA algorithms.

$$\begin{aligned} Precision &= \frac{TP}{TP+FN} \\ Recall &= \frac{TP}{TP+FP} \end{aligned} \tag{10}$$

### G. LIMITATIONS OF THE PROPOSED METHOD

After testing, it was found that the detection effect of a small part of the test images is not ideal. We enumerate them to analyze the limitations of the YOLOM method, as shown in Fig. 9. In the first column of the image, due to the sharp deformation of the mesh, YOLOM only detects part of the mesh target. The second column of the image has a large degree of mesh object curvature that was not detected. From these two failure cases, we can infer that the reason for the detection failure are as follows: when a mesh target is completely perpendicular to the coronal plane, the target may not be detected accurately. Therefore, in practical applications, doctors should pay attention to the angle of the mesh scan and perform multi-angle scans if necessary.

**TABLE 4.** Results comparison.

| Model | Backbone | $mAP_{30}$ | $mAP_{50}$ | $mAP_{70}$ | FPS |
|---|---|---|---|---|---|
| Faster R-CNN | VGG-16 | 97.13 | 96.04 | 94.68 | 2.6 |
| SSD | VGG-16 | 95.85 | 95.25 | 90.43 | 17.1 |
| YOLOv3 | Darknet-53 | 86.70 | 85.85 | 92.40 | 19.8 |
| YOLOv4 | CSPDarknet-53 | 96.94 | 96.01 | 92.40 | 13.0 |
| YOLOM | CSPA-Net | 99.21 | 98.36 | 94.63 | 46.6 |

## V. CONCLUSION

In this research, a mesh detection method based on the YOLOv3 method is proposed that utilizes CSPA-Net, FEP-Net, the Swish activation function and CIoU. The results of the experiments and comparisons demonstrated that the proposed YOLOM detector was more efficient than other existing methods for abdominal wall hernia mesh detection in ABUS images. In this study, the backbone we used could efficiently reduce the number of parameters of the YOLOM detector.

Since the calculation amount is only one-eleventh of the original method, we can use a mediocre GPU for training. In addition, the proposed YOLOM method is a flexible detector because its backbone can be changed from CSPA-Net to other backbones, such as MobileNet or EfficientNet, for different datasets without programming difficulties. Due to the high importance of activation functions and their direct impact on models, the proposed method employs the Swish activation function. The results of the experiments show that Swish improves the efficiency compared with other functions such as LeakyReLU. In addition, in this study, for bounding box regression and improving the loss function, the CIoU method was applied. This method directly minimizes the normalized distance between central points of two bounding boxes, which leads to much faster convergence than other methods such as IoU. Moreover, we found that the coronal mesh texture of an abdominal wall hernia mesh was particularly effective. Automated 3-D ultrasound can offer significant evidence for clinical diagnosis and surgical repair procedures and is a promising detection method for abdominal wall hernia mesh imaging.

## APPENDIX

### A. THE STRUCTURE OF YOLOV3 METHOD

The whole structure of YOLOv3 method is shown in Fig. 10.

### B. SQUEEZE AND EXCITATION NETWORK

In the squeeze stage, the feature map is compressed into a $(1 \times 1 \times N)$ tensor by a global average pooling layer. N represents the global information of each channel. Feature extraction for each channel as in (11):

$$z_N = F_{sq}(\mu_N) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} \mu_N(i, j) \tag{11}$$

where $\mu_N$ is the the feature map of the $N^{th}$ channel, H and W are the height and width of the feature map, respectively.

In the excitation stage, to obtain the weight of each channel, the correlation of the channels is established through two fully connected layers, as shown in (12):

$$s = F_{ex}(z, W) = \sigma(W_2 \delta(W_1 z)) \tag{12}$$

where z is the result of the squeeze stage, $W_1$ and $W_2$ are the fully connected layers, respectively. The number of channel for $W_1$ is $\frac{N}{r}$ and the number of channel for $W_2$ is N. $r$ is a scaling factor to reduce the amount of parameters. $\sigma$ and $\delta$ are the sigmoid and ReLU activation functions, respectively.

In the combination stage, the channel feature is merged with the original feature map, as shown in (13):

$$\widetilde{X} = F_{com}(\mu_N, s_N) = \mu_N \times s_N \tag{13}$$

where $s_N$ is the weight for each channel.

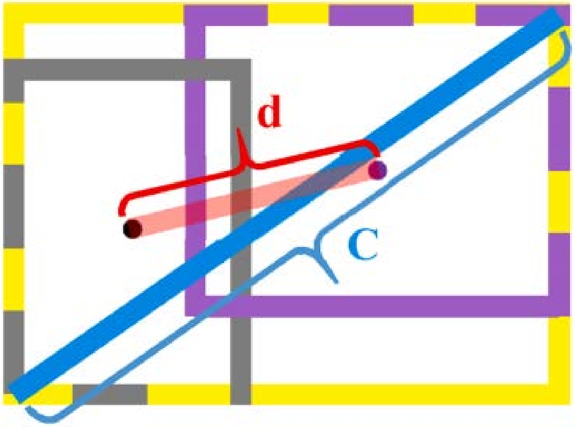


**FIGURE 11.** CIoU between bounding box and GT, where the normalized distance between central points can be directly minimized. $C$ is the diagonal length of the smallest enclosing box covering two boxes and $d = \rho(b, b^{gt})$ is the Euclidean distance between the central points of two boxes.

### C. CIOU

The detailed description of CIoU as shown in Fig. 11.

## REFERENCES

[1] T. E. Bucknall, P. J. Cox, and H. Ellis, "Burst abdomen and incisional hernia: A prospective study of 1129 major laparotomies," *Brit. Med. J.*, vol. 284, no. 6320, p. 931, 1982.

[2] R. Ladurner, Q. Linhuber, F. Hohenbleicher, P. N. Khalil, and T. Mussack, "[surgical treatment of incisional hernias]," *Mmw Fortschritte Der Medizin*, vol. 152, no. 45, p. 41, 2010.

[3] M. Mudge and L. E. Hughes, "Incisional hernia: A 10 year prospective study of incidence and attitudes," *Brit. J. Surgery*, vol. 72, no. 1, pp. 70–71, Dec. 2005.

[4] J. A. Halm, J. W. A. Burger, and J. Jeekel, "Incisional abdominal hernia: The open mesh repair," *Langenbeck's Arch. Surgery*, vol. 389, no. 4, p. 313, Aug. 2004.

[5] J. W. A. Burger, R. W. Luijendijk, W. C. J. Hop, J. A. Halm, E. G. G. Verdaasdonk, and J. Jeekel, "Long-term follow-up of a randomized controlled trial of suture versus mesh repair of incisional hernia," *Ann. Surg.*, vol. 240, no. 4, pp. 578–585, 2004.

[6] M. E. Falagas and S. K. Kasiakou, "Mesh-related infections after hernia repair surgery," *Clin. Microbiol. Infection*, vol. 11, no. 1, pp. 3–8, 2005.

[7] G. Girish, E. M. Caoili, A. Pandya, Q. Dong, M. G. Franz, Y. Morag, E. J. Higgins, J. M. Rubin, and D. A. Jamadar, "Usefulness of the twinkling artifact in identifying implanted mesh after inguinal hernia repair," *J. Ultrasound Med.*, vol. 30, no. 8, pp. 1059–1065, Aug. 2011.

[8] T. Tan, H. Huisman, B. Platel, A. Grivegnee, R. Mus, and N. Karssemeijer, "Classification of breast lesions in automated 3D breast ultrasound," *Proc. SPIE*, vol. 7963, Mar. 2011, Art. no. 79630X.

[9] L. Xi, J. Wang, H. Feng, J. Fu, and A. Li, "Analysis of eighty-one cases with breast lesions using automated breast volume scanner and comparison with handheld ultrasound," *Eur. J. Radiol.*, vol. 81, no. 5, pp. 873–878, 2012.

[10] Y. W. Kim, S. K. Kim, H. J. Youn, E. J. Choi, and S. H. Jung, "The clinical utility of automated breast volume scanner: A pilot study of 139 cases," *J. Breast Cancer*, vol. 16, no. 3, pp. 329–334, 2013.

[11] X. Diao, Y. Chen, Z. Qiu, Y. Pang, and L. Chen, "Diagnostic value of an automated breast volume scanner for abdominal hernias," *J. Ultrasound Med.*, vol. 33, no. 1, pp. 39–46, 2014.

[12] J. Wu, Y. Wang, J. Yu, X. Shi, J. Zhang, Y. Chen, and Y. Pang, "Intelligent speckle reducing anisotropic diffusion algorithm for automated 3-D ultrasound images," *J. Opt. Soc. Amer. A, Opt. Image Sci.*, vol. 32, no. 2, p. 248, 2015.

[13] J. Wu, Y. Wang, J. Yu, Y. Chen, and Z. Qiu, "Identification of implanted mesh after incisional hernia repair using an automated breast volume scanner," *J. Ultrasound Med.*, vol. 34, no. 6, pp. 1071–1081, 2015.

[14] J. Yang, H. Li, J. Wu, L. Sun, and L. Chen, ''Pore texture analysis in automated 3D breast ultrasound images for implanted lightweight hernia mesh identification: A preliminary study,'' *Biomed. Eng. OnLine*, vol. 20, p. 23, 2021.

[15] R. Girshick, J. Donahue, T. Darrell, and J. Malik, ''Rich feature hierarchies for accurate object detection and semantic segmentation,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 580–587.

[16] R. Girshick, ''Fast R-CNN,'' in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 1440–1448.

[17] P. Christiansen, N. L. Nielsen, A. K. Steen, N. R. Jørgensen, and H. Karstoft, ''DeepAnomaly: Combining background subtraction and deep learning for detecting obstacles and anomalies in an agricultural field,'' *Sensors*, vol. 16, no. 11, p. 1904, 2016.

[18] 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.

[19] J. Redmon and A. Farhadi, ''YOLO9000: Better, faster, stronger,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 6517–6525.

[20] J. Redmon and A. Farhadi, ''YOLOv3: An incremental improvement,'' 2018, *arXiv:1804.02767*.

[21] A. Bochkovskiy, C. Y. Wang, and H. Liao, ''Yolov4: Optimal speed and accuracy of object detection,'' 2020.

[22] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Y. Fu, and A. C. Berg, ''SSD: Single shot multibox detector,'' in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 1–37.

[23] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, ''Squeeze-and-excitation networks,'' *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 8, pp. 2011–2023, Aug. 2020.

[24] P. Ramachandran, B. Zoph, and Q. V. Le, ''Searching for activation functions,'' 2017, *arXiv:1710.05941*.

[25] C. Y. Wang, H. Liao, Y. H. Wu, P. Y. Chen, and I. H. Yeh, ''CSPNet: A new backbone that can enhance learning capability of CNN,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 390–391.

[26] K. He, X. Zhang, S. Ren, and J. Sun, ''Spatial pyramid pooling in deep convolutional networks for visual recognition,'' *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1904–1916, Sep. 2015.

[27] F. Chollet, ''Xception: Deep learning with depthwise separable convolutions,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 1251–1258.

[28] Z. Zheng, P. Wang, W. Liu, J. Li, R. Ye, and D. Ren, ''Distance-IoU loss: Faster and better learning for bounding box regression,'' 2019, *arXiv:1911.08287*.

[29] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, ''Grad-cam: Visual explanations from deep networks via gradient-based localization,'' *Int. J. Comput. Vis.*, vol. 128, no. 2, pp. 336–359, Oct. 2019.

**SIQI CHEN** received the B.Sc. degree in electronic and information engineering from Yunnan University, China, in 2020. He is currently pursuing the master's degree with the Department of Electronic Engineering, Fudan University. His current research interests include biomedical image processing and deep learning applications.

**JINLI XU** received the B.Sc. degree in communication engineering from Xiamen University, China, in 2018. He is currently pursuing the master's degree with the Department of Electronic Engineering, Fudan University. His current research interests include biomedical image processing and deep learning applications.

**JINHUA YU** (Member, IEEE) received the Ph.D. degree in electronic engineering from Fudan University, Shanghai, China, in 2008. From 2008 to 2010, she was a Post-Doctoral Fellow with the Department of Bioengineering, University of Missouri, Columbia, MO. She is currently a Full Professor with the Electronic Engineering Department, Fudan University. Her current research interests include ultrasound imaging and medical signal analysis.

**JUN WU** received the Ph.D. degree in biomedical engineering from Fudan University, Shanghai, China, in 2015. He is an Associate Professor with the Department of Electronic and Information Engineering, Yunnan University. His current research interests include medical image processing and deep learning applications.

**GUOHUI ZHOU** received the Ph.D. degree in biomedical engineering from Fudan University, Shanghai, China, in 2012. He is a Lecturer with the Department of Electronic Engineering, Fudan University. His current research interests include medical image processing and cardiac electrophysiology.