



UoloNet: based on multi-tasking enhanced small target medical segmentation model

Kejia Zhang¹ · Lan Zhang¹ · Haiwei Pan¹

Published online: 3 February 2024
© The Author(s) 2024

Abstract

In recent years, UNET (Ronneberger et al. 2015) and its derivative models have been widely used in medical image segmentation with more superficial structures and excellent segmentation results. Due to the lack of modeling for the overall characteristics of the target, the division tasks of minor marks will produce some discrete noise points, resulting in a decline in model accuracy and application effects. We propose a multi-tasking medical image analysis model UoloNet, a YOLO-based (Redmon et al. 2016; Shafiee et al. 2017) object detection branch is added based on UNET. The shared learning of the two tasks through semantic segmentation and object detection has promoted the model's mastery of the overall characteristics of the target. In the reasoning stage, merging the two functions of object detection and semantic segmentation can effectively remove discrete noise points in the division and enhance the accuracy of semantic segmentation. In the future, the object detection task will be the problem of excessive convergence of semantic segmentation tasks. The model uses CIOU (Zheng et al. 2020) losses instead of IOU losses in YOLO, which further improves the model's overall accuracy. The effectiveness of the proposed model is verified both in the MRI dataset SEHPI, which we posted and in the public dataset LITS (Christ 2017).

Keywords Multi-task enhancement · UNET · YOLO · Medical object detection · Liver dataset

Kejia Zhang provided the overall direction and relevant preliminary theories for the thesis. Lan Zhang successfully implemented the holistic approach, completing experimental validation. Haiwei Pan has provided the corresponding dataset.

✉ Lan Zhang
zhanglan2015@hrbeu.edu.cn

Kejia Zhang
kejiazhang@hrbeu.edu.cn

Haiwei Pan
panhaiwei@hrbeu.edu.cn

¹ Computer Science and Technology, Harbin Engineering University, Nantong Street, Harbin, Heilongjiang, China

1 Introduction

Encoder-decoder networks have become a popular approach to semantic segmentation. It completes the small data concentration image of small data concentration images with a simple and beautiful structure. Among them (Ronneberger et al. 2015; Badrinarayanan et al. 2017; Chen et al. 2018), the encoding process uses the pooling layer to reduce the location information and extract abstract characteristics gradually. The decoding process gradually restores the location information and finally realizes the generating image Mask. Because of its excellent performance, it has also been widely used in other fields, especially in medical image segmentation, the most prominent of which is the UNET framework. The UNET framework is widely used with a more straightforward structure and excellent segmentation results. At the same time, the proposal of different UNET variants has also enriched the UNET family (Zhou et al. 2018; Çiçek et al. 2016; Milletari et al. 2016; Huang et al. 2020). They further improve segmentation accuracy from various angles, such as pruning operations, attention mechanisms, and masks. In addition, the UNET structure combines with other basic models, such as the object detection model, to achieve the result of simultaneous implementation of multiple tasks. However, the traditional combined object detection model is large, the end-to-end output cannot be achieved, and the training is complicated. In addition, more is needed for small-scale datasets such as medical images. This effect may be pronounced for small-scale medical datasets.

Different from the more significant target segmentation, the appearance of discrete dot noise is irreversible for the segmentation of minor targets. A mechanism involves reducing the discrete point noise generated during prediction. This limitation on clinical management ultimately affects the etiological diagnosis, treatment, and accuracy of surveillance data. At the same time, clinically diseased doctors in the same area generally recognize tumors of a category. Nonetheless, it will lead to predicting multiple labels in the same area because the target segmentation in the semantic segmentation stage is based on pixels. This measure will significantly reduce model prediction accuracy and cause problems for subsequent clinical treatment.

The existing multi-tasking division models are principally based on the combination of the anchor (Wang et al. 2017) model and UNET. This method is straightforward; only two parts must be connected in series. Nevertheless, in the traditional object detection model, such as the RCNN (Girshick et al. 2014; Ren et al. 2015a; He et al. 2017a) series, computations are enormous, and image feature extractions are separate parameter fixing and training networks. It takes work to achieve end-to-end training. At the same time, the duplication of the feature extraction structure of the sub-sampling in the subsequent UNET will cause a loss of computing resources. So the paper wants to simplify the model of multi-tasking. Because the characteristic extraction layer of YOLO (Redmon et al. 2016) is comparable to the UNET encoder structure, the paper proposes combining the two in the process of coding and downsampling achieves feature extraction in target recognition.

The YOLO series model proposes a significant improvement in the prediction speed of the model. At the same time, the YOLO model is a relatively lightweight feature extraction model, and the coding structure is in the codec structure. Therefore, this paper considers the advantages of combining the YOLO model and the encoder in the UNET structure to add a object detection branch to the traditional UNET. Get a new model to achieve a lightweight multi-tasking model and achieve target segmentation and target recognition in the field of medical imaging.

In the traditional target recognition framework, IOU usually uses the target box to compare. The shortcomings of IOU regression, nonetheless, occur. Classic IOUs cannot express the distance between the two boxes when the two boxes are not intersecting, the length and width ratio of the appearance, and the length and width ratio of the center point. IOU can lead to problems such as bottlenecks and poor results in the training process. To alleviate the above problems and to improve the convergence, the paper recommends comparing the target box with CIOU (Zheng et al. 2020) instead of conventional IOU. The CIOU boundary box loss function can continuously move the prediction box to the actual frame by iteration. It furthermore ensures that the width and height of the actual box and the high vertical and horizontal ratio of the actual frame are as close as possible, which accelerates the regression convergence speed of the prediction box. The stable convergence of this improvement to a large extent has facilitated the convergence rate and improved the accuracy of model prediction.

Our multi-task approach solves the problem of directly dividing tumors and solves the training instability caused by too many negative samples during training. We use the target positioning task to enhance the segmentation task's performance. During the final experiment, we also improved the degree of gain of different parameters to the model convergence. We increased some small tracks, reducing the time of multi-task model training as much as possible. Unlike existing methods, our segmented objects are targets of small labels, and the model can train a network end-to-end to get the positioning and mark of markers simultaneously. The training time is short, and the model convergence is fast. The main contribution of the paper is as follows:

- For tumor target recognition and segmentation, the paper recommends a multi-tasking target recognition division based on the UNET model. The network can gain MASK segmentation and achieve target recognition of minor target tumors.
- For target recognition and division tasks, the paper recommends an identification and segmentation training LOSS function so that the model can end-end training. At the same time, the problem of slow convergence in the target segmentation task is used to replace conventional IOU and experimentally verify its effectiveness.
- For the noise generated during the MASK prediction process, the paper uses the target box generated by the model to propose a prediction process. Resolve discrete noise generated during the MASK prediction process and further enhance the results of tumor prediction.

2 Related works

Due to the rapid development of deep learning, many CNN-based methods have achieved incredible success in semantic segmentation. They can be applied to driving area segmentation tasks and produce pixel-level results. FCN (Long et al. 2015) initially proposed a fully convolutional network, which integrates different scale feature maps. Nevertheless, since the details of wild fixation. It is easy to lose the details and needs more space consistency. Using the de-pooling layer to sample the feature diagram, SEGNET (Badrinarayanan et al. 2017) retains the integrity of the details, but the low-resolution feature diagram will disregard the closer pixel information. Deconvnet (Noh et al. 2015), filter at different positions can supplement details and capture shape information, but it is problematic to process details. Deeplab uses a random airport with hollow convolutional connection conditions,

but the prediction results retrieved are only 1/8 of the original input. PSPNET (Zhao et al. 2016) recommends the pyramid module to aggregate background information but uses different pyramid pool modules with higher requirements for details. Edgenet (Plastiras et al. 2019) combines edge detection with division tasks that can be driven by the drive and obtain more accurate segmentation results without affecting reasoning speed. GCN (Deferrard et al. 2016) proposes a coding structure with large-scale convolution kernels, but more complicated parameters have been calculated.

The above method is not optimal for the target division of small details, and the requirements for data sets and models during the calculation process need to be lowered. For the semantic division of small targets, UNET model switched the characteristics in the encoder to the onset of sampling characteristics of each stage and adopted a skip method to allow a decoder to learn the correlation of encoder loss. Good segmentation performance has gradually become the mainstream of medical image segmentation. 3DUET (Korez et al. 2016) uses 3D volumes as the input three-dimensional UNET due to improved GPU storage and computing power. The three-dimensional U network structure is also the main trunk of the exterior segmentation subnet. Although 3DUNET has excellent performance, it often requires streaming titles to be used. Specifically, we cut the three-dimensional volume into small pieces in order, input them into the network independently, get the output, and then suture the output together to get the final segmentation result. Since most tumor images are in the background, this strategy wastes computing resources.

Currently, the mainstream object detection algorithm can be divided into two stages and one phase method. The production of regional suggestions has gone through several development stages (Girshick et al. 2014; Girshick 2015; Ren et al. 2015b; He et al. 2017b). FasterR-CNN shows excellent detection and positioning capabilities. Nevertheless, this method will predominantly fix a part of the network extraction network, and the model cannot conduct end-to-end training. At the same time, the model parameters are significant, and it is problematic for datasets with a small amount of data in medical images. SSD series (Liu et al. 2015; Yi et al. 2019) and YOLO (Redmon et al. 2016; Shafiee et al. 2017) series algorithms are milestones in a single-stage method. This algorithm can simultaneously carry out border boxes and target classification. YOLO (Redmon et al. 2016) splits the image as the $S \times S$ grid instead of utilizing the RPN network to extract regional suggestions, considerably speeding up the detection speed. YOLO9000 (Redmon and Farhadi 2017) introduced an anchoring mechanism to upgrade the recall rate of detection.

Multi-tasking learning aims to learn better representations through shared information between multiple tasks. In particular, multi-tasking learning methods based on CNN can also achieve convolutional sharing of network structures. Maskr-CNN (He et al. 2017a) adding a branch of the R-CNN essentially combines instance segmentation and object detection tasks. These two tasks can improve one another's performance. CFUN (Xu et al. 2018) constructs a dual task that combines FasterR-CNN and U-NET networks to execute the heart segmentation and recognition. DLT-NET (Qian et al. 2019) inherits the structure of the encoder-decoder and constructs the context tension between the submission decoder to share the specified information between the tasks. Covering R-CNN (Girshick et al. 2014) network can be identified, classified, and divided simultaneously. Some people have already attempted to apply Maskr-CNN to medical images. CaraNet (Lou et al. 2022) proposes a Context Axial Reserve Attention Network for small medical objects to improve the segmentation performance of small objects compared with several recent state-of-the-art models. This work (Liu et al. 2022) develops a multi-level structural loss by integrating the region, boundary, and pixel-wise information to supervise feature fusion and precise segmentation. Rfa and

Fga (2019) method focuses on solving the efficiency problem in the segmentation problem, using excellent pre-processing methods, and the model uses the atlas-based and clustering schemes method. This work (Ngo et al. 2020) proposes a way to accurately locate tumors using expansion convolution and multitasking, but this method needs to be more effective at identifying tumor boundaries. HSN (Chen et al. 2019) proposes a hybrid segmentation network based on a convolutional neural network to segment SCLC from computed tomography (CT) images automatically and a hybrid features fusion module to fuse the 2D and 3D features effectively and to train these two CNNs jointly.

3 Method

3.1 Definition

R represents the real number set. First of all, let us define dataset $\mathcal{D}$. For each piece of data $(S, M, B) \in \mathcal{D}$, $S \in \mathbb{R}^{w \times h}$ is a piece of medical image, where $w \times h$ is the size of each slice; $M \in \mathbb{C}^{w \times h}$ is the label of semantic segmentation tasks (MASK), where $\mathcal{C} = \{0, 1, \dots, n\}$ represents the category in the label; $B \in 2^{\mathbb{R}^4 \times \mathcal{C}}$ is the label we use M to draw object detection tasks. For each element $b = (x_c, y_c, l_w, l_h, c)$, $(x_c, y_c) \in \mathbb{R}^2$ is the outer rectangular center point coordinates generated by M on each target, $(l_w, l_h) \in \mathbb{R}^2$ is the size of the rectangle, and $c \in \mathcal{C}$ is its category tab. We need to do a dual task learning, that is, using the same model to do the semantic segmentation and object detection, to gain the role of the two tasks to promote each other, and eventually enhance the accuracy of the semantic segmentation of the model.

The UoloNet proposed includes a shared encoder, a object detection module, and a Mask generating module, that is, the decoder in Fig. 1. We will now review the details of each part.

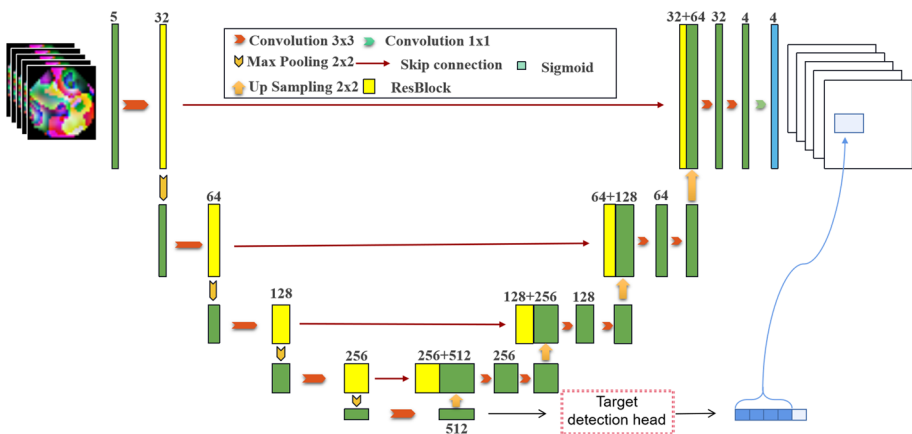


Fig. 1 UoloNet structure

3.2 Encoder-decoder struction

In order to better represent the essential three-dimensional, unique signs in medical images, the input of the model will be made of multiple continuous slices $S \in \mathbb{R}^{w \times h}$ to enter $X \in \mathbb{R}^{(2d+1) \times w \times h}$; The mask and box corresponding to the $d + 1$ slicing in multiple continuous slices are labeled. The edge part uses a method of mirror increase. In the previous work, we verified that the effect was best when $d = 2$.

The main component of Ulone has shared encoders. The encoder consists of repetition applications of residual blocks; each block contains two convolutional layers, $3 \times 3 \times 3$ cores. After each convolutional layer, it is activated by group nativeness and RELU. In each down-shift, the resolution of the input feature diagram is used to use the convolution with a rod of 2. In the first layer, the number of channels is set to 32 and doubles after each downgrade to predict the time complexity of each layer. We performed a total of four sampling operations in the encoder, given an input image X and output to H . The shared encoder module primarily implements the feature extraction of the image. The H output contains the mask information and target recognition information of the picture. It has multiple semantics and is not affected by specific tasks. This allows target recognition and semantic segmentation to share an encoder. Symmetrically, the decoder is sampled to enhance the resolution and progressively decrease the number of channels by halving its channels. In each step, the feature diagram of the upper sample first follows the low-level feature diagram corresponding to the encoder and then refreshes it through the residual block. After four samples of the feature diagram, you get: $\hat{M}$, $\hat{M} \in \mathbb{R}^{w \times h \times n}$; is a pre-segment feature diagram, and the number of channels is set to n be consistent).

3.3 Object detection module

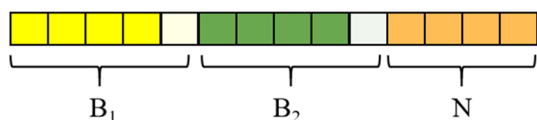
We desire to unify the separation component of object detection to connect a single neural network. Our network uses features from the entire image to predict each bounding box. It also predicts Boundingbox, all targets of all boundaries. This means that we can conduct network reasoning on all objects in the image in the broad range and collect our semantic division modules. The YOLO architecture makes end-to-end training and real-time speed possible while maintaining high average accuracy.

3.3.1 Object coding

Similar to YOLO's target coding, we coded the goal to a tensor of $g \times g \times (k \times 5 + n)$, input model and predictive value computation loss. Similarly, our picture is split into a grid of 7×7 , that is, $g = 7$. At the same time, 2 boxes in each area are generated, $k = 2$. n is the number of labels.

In Fig. 2, the number of categories in the paper is 4, the number of grids is 7, and the number of target box $k = 2$, so the target coding is a tensor with a target coding of $7 \times 7 \times (2 \times 5 + 4) = 7 \times 7 \times 14$. Extend the entire picture to get the final code of the target.

Fig. 2 Coding structure



In addition, it is different from traditional IOU when calculating the confidence in the calculation code.

3.3.2 Object detection head

The object detection head is shown in Fig. 3.

It is stacked by two Convblock, two linear layers and LeakRelu. The two convolutional modules are composed of convolutional pooling and LeakRelu. The purpose of using two full connection layers is to prohibit the use of a pure full connection layer directly flatten. In this way, we get the output of the object detection header as follows:

$$O = TDH(H), O \in \mathbb{R}^{g \times g \times (k \times 5 + n)} \quad (1)$$

3.3.3 CIoU in the object detection

In the experiment, we found that the IOU used in the YOLO algorithm cannot specify the distance when the target box and the prediction box do not intersect. It took a long time to recognize the small target. So we use CIoU to replace the IOU algorithm to upgrade the accuracy and convergence rate of algorithms. The ciou formula is shown below:

$$CIOU(\hat{b}, b) = IOU(\hat{b}, b) - \frac{dist^2(\hat{b}, b)}{diag^2(\hat{b}, b)} - p(\hat{b}, b)q(\hat{b}, b) \quad (2)$$

b is the target box marked, $\hat{b}$ is the prediction target box for the model output. $dist(\hat{b}, b)$ means the European-style distance between $\hat{b}$ and b center points $diag(\hat{b}, b)$ means that the diagonal distance of the minimum closing area of $\hat{b}$ and b can be included at the same time.

$$p(\hat{b}, b) = \frac{q(\hat{b}, b)}{1 - IOU(\hat{b}, b) + q(\hat{b}, b)} \quad (3)$$

$$q(\hat{b}, b) = \frac{4}{\pi^2} \left(\arctan\left(\frac{l_w}{l_h}\right) - \arctan\left(\frac{\hat{l}_w}{\hat{l}_h}\right) \right)^2 \quad (4)$$

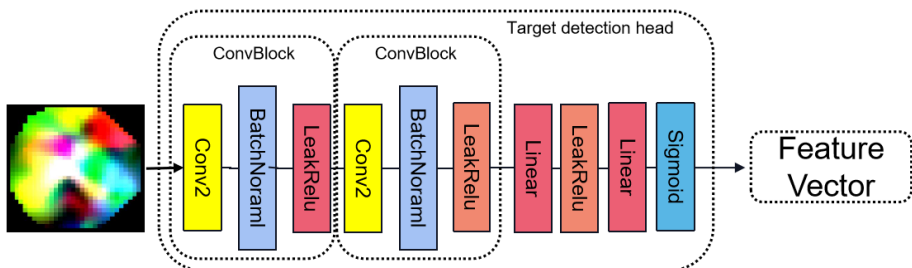


Fig. 3 Object detection head

3.4 Loss function

3.4.1 Semantic loss

For simplicity, we consider the division of an organ and related tumors as two dual-segmentation tasks, and combined with DICE loss and binary cross entropy as the target of each task. The loss function is expressed as:

$$\begin{aligned} \mathcal{L}_{seg}(\hat{\mathbf{M}}, \mathbf{M}) = & 1 - \frac{2 \sum_{i,j} \hat{\mathbf{M}}[i,j] \mathbf{M}[i,j]}{\sum_{i,j} (\hat{\mathbf{M}}[i,j] + \mathbf{M}[i,j] + \epsilon)} \\ & - \sum_{i,j} \left(\mathbf{M}[i,j] \log \hat{\mathbf{M}}[i,j] + (1 - \mathbf{M}[i,j]) \log (1 - \hat{\mathbf{M}}[i,j]) \right) \end{aligned} \quad (5)$$

Among them, $\hat{\mathbf{M}}[i,j]$ and $\mathbf{M}[i,j]$ indicate the segmentation prediction and vector of the pixels of (i, j) , and add ϵ as a smooth factor.

3.4.2 Object detection loss

Similar to the YOLO V1 network, each grid unit can predict multiple boundary boxes. To estimate the loss of true positive. We believe that one of the boxes is responsible for the goal. To this end, select the box with the highest IOU with the GT. When the center point of a real object falls into a cell, the cell is responsible for recognizing the object. The border of the maximal IOU with the real object is set to a positive sample. The category of this area is the category of the real object. The confidence of the frame is 1. Except for the borders that are given a positive sample above, the remaining borders are negative samples. The negative sample has no category loss and the position of the border position, only the confidence loss, the true value is 0.

The model uses the errors between the predicted value and the GT to calculate the loss (MSE). The loss function includes: Localization loss coordinates, classification loss, confidence loss, objectness of the box.

$$\begin{aligned} \mathcal{L}_{det} = & \lambda_c \sum_{i=0}^{s^2} \sum_{j=0}^2 \mathbf{1}_{[j=j_{max}]} \left[(x_i - \hat{x}_i)^2 + (y_i - \hat{y}_i)^2 \right] \\ & + \lambda_c \sum_{i=0}^{s^2} \sum_{j=0}^2 \mathbf{1}_{[j=j_{max}]} \left[\left(\sqrt{w_i} - \sqrt{\hat{w}_i} \right)^2 + \left(\sqrt{h_i} - \sqrt{\hat{h}_i} \right)^2 \right] \\ & + \sum_{i=0}^{s^2} \sum_{j=0}^2 \mathbf{1}_{[j=j_{max}]} (c_i - \hat{c}_i)^2 \\ & + \lambda_n \sum_{i=0}^{s^2} \sum_{j=0}^2 \mathbf{1}_{[j \neq j_{max}]} (c_i - \hat{c}_i)^2 \\ & + \sum_{i=0}^{s^2} \mathbf{1}_{[j=j_{max}]} \sum_{n \in \text{classes}} (p_i(n) - \hat{p}_i(n))^2 \end{aligned} \quad (6)$$

Among them, λ_c and λ_n are the weight coefficients of losses without the location of the object, and 5 and 0.5 are taken in similar YOLO training, respectively.

3.4.3 Overall loss definition

Combined with the loss function of the two tasks, the final loss is defined as:

$$\mathcal{L} = \mathcal{L}_{seg} + \alpha \mathcal{L}_{det} \quad (7)$$

Among them, $\alpha = 0.1$ is a hyperparameter.

3.5 Prediction process

In the traditional UNET model, the final prediction result mask will have many discrete noise points. In the process of large targets, this small noise point gives the model output result less. However, in the identification and division of small labels, this type of false positive is unforgivable. To solve this problem, we proposed the model's accuracy for object detection during the prediction process. The primary purpose is to use target recognition results to tell the model mask segmentation which specific area should be focused on, thereby reducing discrete false positives and improving accuracy.

The principle of the prediction process is illustrated in Fig. 4.

Volume produces two results after the model $\hat{M}$ and $O \in \mathbb{R}^{g \times g \times (k \times 5 + n)}$. The point in each grid corresponds to a $k \times 5 + n$ dimension vector. Similar to Yolo algorithms, select box coordinates with credibility of more than 0.5. After the NMS algorithm generates predictive results:

$$\hat{B}, \{x_c, y_c, w, h\} \in \mathbb{R}^{1 \times 4} \quad (8)$$

During the calculation process, the label is expanded to the tensor of the entire graph size: $\hat{B} \in \mathbb{R}^{w \times h}$, and the pixels in the border box are set 1, and the rest are 0. The scope B and

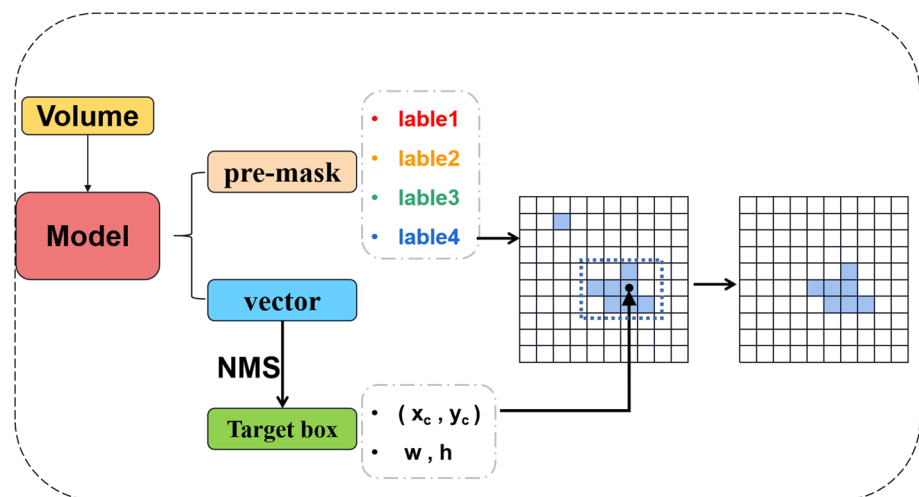


Fig. 4 Prediction process

MASK for the target box to get the final mask for the Hadamard operation. This process will erase separate dots and further improve the accuracy of model predictions. We define the range of the target box and the label range predicted by Maks. Mix is used as $Mix > 0.5$ as a positive superposition; otherwise, it is a negative superposition.

4 Experiment

4.1 Dataset

The dataset in the experiment is selected by the liver tumor segmentation of the previously accumulated liver tumors. Meanwhile, four medical experts used specialist software to label them. Thus, a large-scale enhanced liver and gallbladder image (SEHPI) small label dataset is formed. In addition, this dataset is a liver-specific contrast agent capable of dynamically enhancing MR images in the liver and gallbladder. Unlike MR images of non-liver-specific contrast agents, hepatobiliary images of liver-specific contrast agents can assist in the screening and diagnosing hepatic nodules, and it is simpler to outline the lesion boundaries. In the labeling process, tumors are differentiated and classified. They are liver cancer, metastases, cysts, and hemangioma (in Fig. 5).

Our dataset is segmented and classified into common liver diseases on liver and gallbladder MR phase images, which can help physicians diagnose and diagnose liver diseases. Each medical section is a single-channel gray image, and the range of each pixel value is $[0, 5000]$. This dataset contains a total of 400 labeled cases. Also, we compare with the LITS standard dataset. Compared to LITS, a common liver segmentation dataset, our dataset has a substantial data volume, more precise label separation, and is more focused on tumor classification.

4.2 Evaluation index

For input X , suppose its corresponding label is Y , and the model output Y' . The number of channels Y and Y' are both q (the number of tag categories). We use the following indicators to evaluate the effect of the model.

4.2.1 Dice

For the evaluation criteria in the segmentation process, the Dice Similarity Coefficient (DSC) is mainly used. It is a measure of ensemble similarity, which is usually used to calculate the similarity of two samples, $\delta = 1e - 5$:

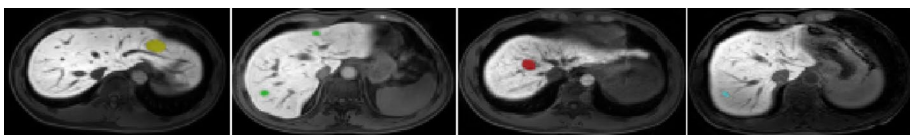


Fig. 5 SEHPI. Label1, red, liver cancer; Label2, blue, metastatic tumor; Label3, yellow, cyst; Label4, green, hemangioma. (Color figure online)

$$Dice = \sum_{m=1}^q \frac{2Y[m] * Y'[m] + \delta}{Y[m] + Y'[m] + \delta} \quad (9)$$

4.2.2 PPV

Positive predicted value (PPV) reflects the ability of the classifier or model to correctly predict the accuracy of positive samples, that is, how many of the predicted positive samples are real positive samples. The higher value, the better performance.

$$PPV = \sum_{m=1}^q \frac{2Y[m] * Y'[m] + \delta}{Y'[m] + \delta} \quad (10)$$

4.2.3 Sensitivity

Sensitivity reflects the ability of the classifier or model to correctly predict the full degree of positive samples, and increases the prediction of positive samples as positive samples. It reflects the proportion of positive samples predicted as positive samples in the total positive samples. The higher value, the better performance.

$$Sensitivity = \sum_{m=1}^q \frac{2Y[m] * Y'[m] + \delta}{Y[m] + \delta} \quad (11)$$

4.2.4 Hausdorff 95

The Dice coefficient is more sensitive to the filling inside the mask, and the hausdorff distance is more sensitive to the boundary of the segmentation: B, B' are the masks of ground truth and the model's output. $\|\cdot\|_2$ represents the L_2 distance between b and b'

$$dist(B, B') = \max_{b \in B} \min_{b' \in B'} \|b - b'\|_2 \quad (12)$$

$$d_H(B, B') = \max\{dist(B, B'), dist(B', B)\} \quad (13)$$

We use Hausdorff 95 to eliminate the influence of a tiny subset of outliers.

$$d_{95\%HD} = d_H * 95\% \quad (14)$$

4.3 Training

During the training process, 40 of the 400 examples were selected as a test set, and 40% of the training set was selected as a value set. Conditions for model stop training: When the best dice concentrated in VAL are not improved after 20 times, the model stops training. Since the VAL collection does not participate in training, it can ensure that the training is the best result without fitting. The size of the batch is 10. The initial learning rate is 0.0003, and the model is optimized using the SGD method.

4.3.1 Experiment on SEHPI dataset

On the dataset we collected by ourselves, we compared the four groups of experiments to compare our performance results in the division of MASK. In ULONET, the output of the last layer consists of four foreground probability diagrams. The prospects have a higher probability (> 0.5) than the background and are considered part of the segmentation. After the final scope determination, the final MASK is formed. During the training process, 40 of the 400 examples were selected as a test set, and 40% of the training set was selected as a value set. Model stop training conditions: When the best DICE result of val_data has not improved after 20 epochs, the model stops training. Because the evaluation indicators do not participate in training, it can ensure that the best results are reached without fitting. The size of the batch is 10. The momentum is 0.9. The initial learning rate is 0.0003. The result details can be seen in (Table 1).

The experimental results of UoloNet should be outstanding for the other four results. However, there are two indicators in PPV and Hausdorff, although not the best, but also compared with the best results. Also, the visible results of different models, as shown in Fig. 7.

At the same time, we visualized the process of adding a bounding box Fig. 6. It can be observed that our model can avoid discrete noise points as much as possible and improve the accuracy of the model. It can be observed that the discrete points pointed out by the blue arrows in Fig. 6a and c can be eliminated through the prediction process. In Fig. 6c, the points the green arrows point can be retained.

4.3.2 Experiment on LITS dataset

Since our model is labeled for small scales, we only object the tumor label part when compared to the LITS public dataset, which is more targeted. The performance of the different models is as follows Table 2. It can be seen that UOLONET performs well on public datasets.

4.3.3 CIOU and IOU comparison experiments

The target loss will have a training bottleneck at the beginning of the training with large values. Total_loss is affected by target_loss and mask_loss, which converge slowly at the beginning of training. After a while, it will decrease promptly and eventually stabilize. Looking at the convergence of the conventional IOU model in Fig. 8. B, one can see that the training bottleneck is cracked at 130 epochs, and it rapidly decreases, but it quickly enters the next bottleneck. The training bottleneck is crossed again at 178 epoch, after which the model tends to converge. This can be found in the model using CIOU, which was able to quickly break through the training bottleneck at 92 epochs without re-entering the training bottleneck. The objective loss in 110 epoch tends to converge during the subsequent training process. The convergence is even better. At the same time, we quantitatively compare the effect of using CIU and IOU on the experimental results, as shown in the following table:

As can be observed from Table 3, the model predictions using IOU not only use the concepts of the CIOU model but also have considerably higher FFCE than the CIOU

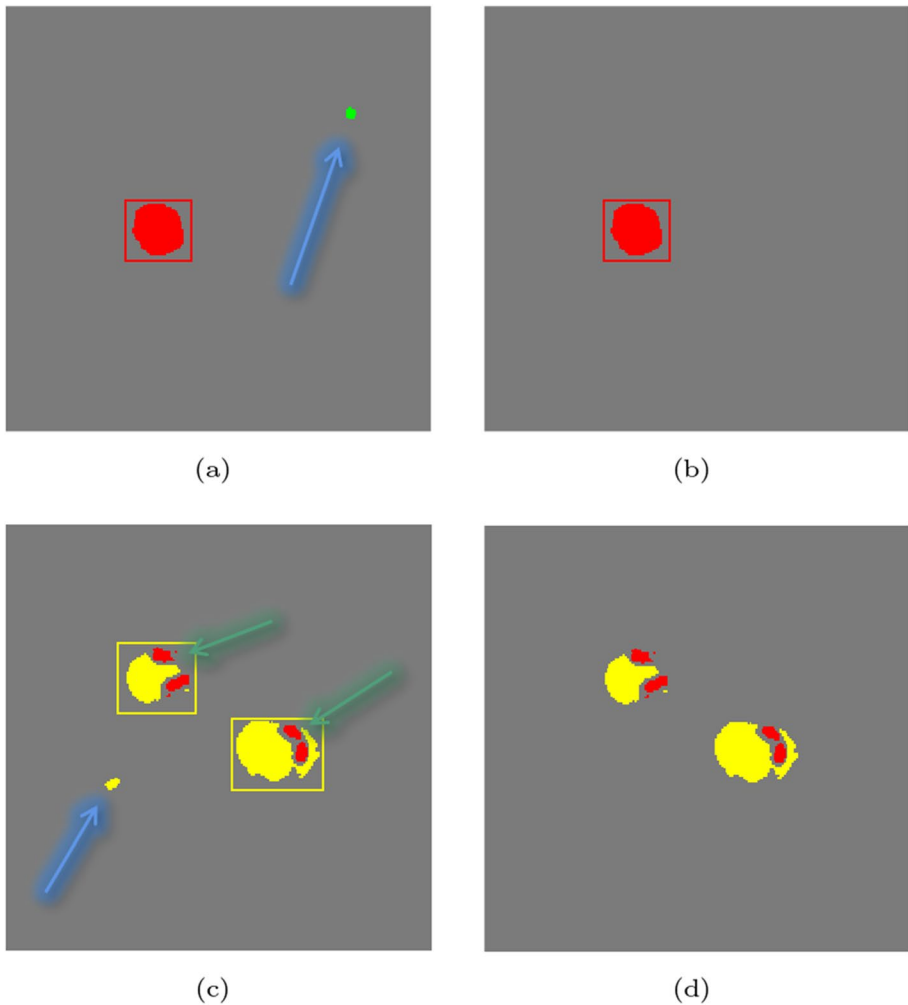


Fig. 6 Prediction process

model. This can demonstrate that CIOU can considerably enhance the convergence and accuracy of the model.

4.3.4 Hyperparameter selection

The α proposed in the last overall loss calculation is important for hyperthermia. It provides the ratio of object recognition and semantic segmentation in the loss. To obtain this relation, we performed a series of experiments with different values of α . In the experiments, the mean value of the dice values and the first fastest convergence of the epoch values for the four labels are selected for evaluation. It can be seen from Table 4 that the effect of the $\alpha = 0.1$ model is the best, and the fastest convergence is the fastest.

Table 1 SEHPI dataset experimental results

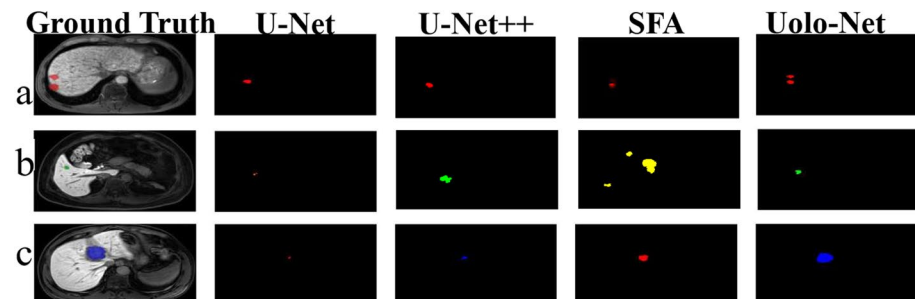
	Dice				Sensitivity			
	Label1	Label2	Label3	Label4	Label1	Label2	Label3	Label4
U-Net (Ronneberger et al. 2015)	0.8605	0.8761	0.9168	0.9207	0.9783	0.9821	0.9903	0.9671
U-Net++ (Zhou et al. 2018)	0.8684	0.8965	0.8576	0.9156	0.9506	0.9834	0.9791	0.9673
U-Net3+ (Huang et al. 2020)	0.8619	0.8965	0.9134	0.8423	0.9733	0.9776	0.9876	0.9627
nnUnet (Isensee et al. 2021)	0.8912	0.8931	0.8729	0.8933	0.9714	0.9812	0.9890	0.9693
SFA (Fang et al. 2019)	0.8618	0.8821	0.8488	0.8903	0.9701	0.9654	0.9702	0.9613
MF (Ngo et al. 2020)	0.8734	0.8834	0.9023	0.9233	0.9607	0.9841	0.9870	0.9601
UoloNet	0.9001	0.9052	0.9232	0.9346	0.9793	0.9845	0.9918	0.9711
	PPV				Hausdorff 95			
	Label1	Label2	Label3	Label4	Label1	Label2	Label3	Label4
U-Net	0.8756	0.8886	0.9228	0.9441	0.6218	0.4616	0.2589	0.2081
U-Net++	0.9104	0.929	0.8754	0.9468	0.5916	0.3113	0.4470	0.3167
U-Net3+	0.8828	0.9154	0.9227	0.8614	0.5696	0.3645	0.2640	0.5364
SFA	0.8895	0.8932	0.9394	0.9592	0.5431	0.4234	0.3319	0.3438
MF	0.9012	0.9199	0.9298	0.9125	0.5024	0.3714	0.3233	0.2869
nnUnet	0.8911	0.9233	0.9391	0.9741	0.4236	0.3321	0.2069	0.2091
UoloNet	0.9192	0.9331	0.9499	0.9748	0.4601	0.3523	0.1617	0.1994

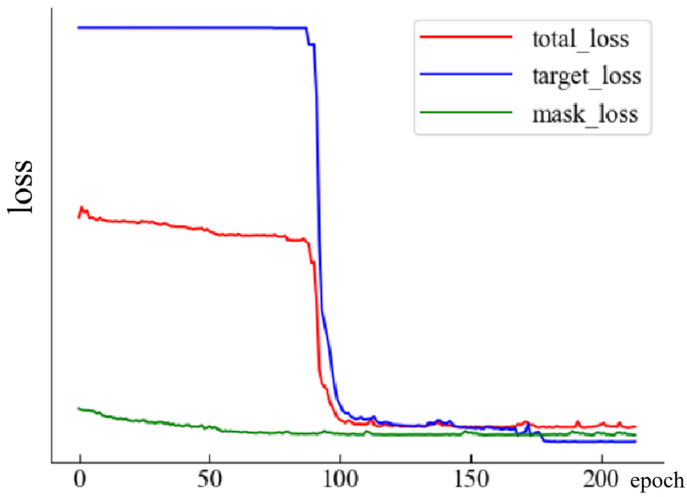
The bold text represents the optimal performance of various methods on different datasets

Table 2 LiTS tumor dataset experimental results

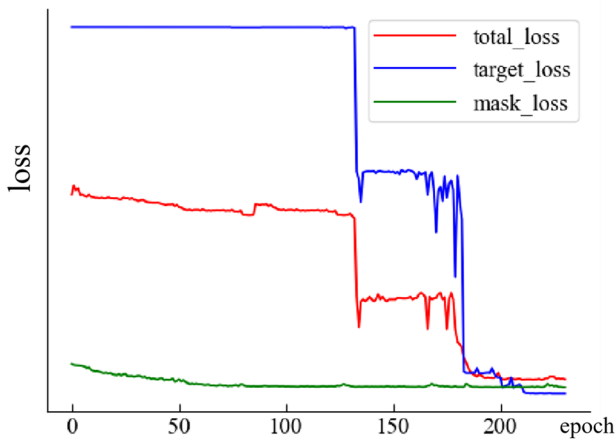
	Dice	PPV	Sensitivity	Hausdorff95
Unet (Ronneberger et al. 2015)	0.3813	0.5984	0.6300	4.9384
Unet++ (Zhou et al. 2018)	0.3992	0.5724	0.7233	3.9812
Unet3+ (Huang et al. 2020)	0.4013	0.5432	0.8092	4.3324
nnUnet (Isensee et al. 2021)	0.4501	0.6012	0.8201	3.1742
SFA (Fang et al. 2019)	0.3901	0.5824	0.7624	3.2105
MF (Ngo et al. 2020)	0.4213	0.6055	0.8123	3.2733
UoloNet	0.4624	0.6067	0.8214	2.8541

The bold text represents the optimal performance of various methods on different datasets

**Fig. 7** Visualization of experimental results



(a) IOU Loss



(b) CIOU Loss

Fig. 8 The detailed structure**Table 3** CIOU and IOU comparison experiments

(Mean)	Dice	PPV	Sensitivity	Hausdorff95	FFCE
IOU-model	0.9288	0.96789	0.9403	0.2833	178
CIOU-model	0.9315	0.9846	0.9421	0.2728	92

The bold text represents the optimal performance of various methods on different datasets

Table 4 Hyperparameter α selection

$\alpha =$	10	1	0.1	0.001
Dice	0.9084	0.9012	0.9113	0.9105
FFCE	125	98	92	101

The bold text represents the optimal performance of various methods on different datasets

5 Conclusion

In order to resolve the problem of medical image segmentation with small-scale labels, we propose UoloNet. This architecture adopts a dual task mode, and further upgrades the accuracy of target segmentation mask results by utilizing the results of object detection. The iou module in the traditional yolo structure is enhanced, and ciou is used to enhance the convergence speed of the model. Aiming at the noise generated in the mask prediction process, a prediction process is recommended utilizing target box. Resolve the discrete point noise generated in the mask prediction process, and further enhance the tumor prediction results. Experiments demonstrate that UoloNet has excellent performance in small-scale label segmentation.

Acknowledgements This work was supported by National Natural Science Foundation of China (Grant No.62072135), Fundamental Research Funds for the Central Universities (Grant No.3072020CF0602) and Natural Science Foundation of Ningxia Hui Autonomous Region (Grant No. 2022AAC03346).

Author contributions KZ and LZ wrote the main manuscript text and HP prepared datasets which is our collection. All authors reviewed the manuscript.

Declarations

Competing interests The authors declare no competing interests.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Badrinarayanan V, Kendall A, Cipolla R (2017) Segnet: a deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans Pattern Anal Mach Intell* 39(12):2481–2495
- Chen LC, Papandreou G, Kokkinos I, Murphy K, Yuille AL (2018) Deeplab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFS. *IEEE Trans Pattern Anal Mach Intell* 40(4):834–848
- Chen W, Wei H, Peng S, Sun J, Qiao X, Liu B (2019) Hsn: hybrid segmentation network for small cell lung cancer segmentation. *IEEE Access* 7:75591–75603. <https://doi.org/10.1109/ACCESS.2019.2921434>
- Christ P (2017) Lits—liver tumor segmentation challenge (lits17)

- Çiçek Ö, Abdulkadir A, Lienkamp SS, Brox T, Ronneberger O (2016) 3d u-net: learning dense volumetric segmentation from sparse annotation. In: International conference on medical image computing and computer-assisted intervention. Springer, New York, pp 424–432
- Defferrard M, Bresson X, Vandergheynst P (2016) Convolutional neural networks on graphs with fast localized spectral filtering
- Fang Y, Chen C, Yuan Y, Tong KY (2019) Selective feature aggregation network with area-boundary constraints for polyp segmentation. In: Medical image computing and computer assisted intervention–MICCAI 2019: 22nd international conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part I 22, Springer, New York, pp 302–310
- Girshick R (2015) Fast r-cnn. In: Proceedings of the IEEE international conference on computer vision (ICCV)
- Girshick R, Donahue J, Darrell T, Malik J (2014) Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), pp 580–587
- He K, Gkioxari G, Dollár P, Girshick R (2017a) Mask r-cnn. *IEEE Trans Pattern Anal Mach Intell*
- He K, Gkioxari G, Dollár P, Girshick R (2017b) Mask r-cnn. In: Proceedings of the IEEE international conference on computer vision (ICCV)
- Huang H, Lin L, Tong R, Hu H, Zhang, Q, Iwamoto Y, Han X, Chen, YW, Wu, J (2020) Unet 3+: a full-scale connected unet for medical image segmentation. In: ICASSP 2020-2020 IEEE international conference on acoustics, speech and signal processing (ICASSP), pp 1055–1059. IEEE
- Isensee F, Jaeger PF, Kohl SA, Petersen J, Maier-Hein KH (2021) nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nat Methods* 18(2):203–211
- Korez R, Likar B, Pernu F, Vrtovec T (2016) Model-based segmentation of vertebral bodies from MR images with 3d CNNs. In: International conference on medical image computing and computer-assisted intervention
- Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu CY, Berg AC (2015) Ssd: single shot multibox detector
- Liu Y, Duan Y, Zeng T (2022) Learning multi-level structural information for small organ segmentation. *Signal Process* 193:108418. <https://doi.org/10.1016/j.sigpro.2021.108418>
- Long J, Shelhamer E, Darrell T (2015) Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3431–3440
- Lou A, Guan S, Ko H, Loew M (2022) CaraNet: context axial reverse attention network for segmentation of small medical objects
- Milletari F, Navab N, Ahmadi SA (2016) V-net: fully convolutional neural networks for volumetric medical image segmentation. In: 2016 fourth international conference on 3D vision (3DV), pp 565–571. IEEE
- Ngo DK, Tran MT, Kim SH, Yang HJ, Lee GS (2020) Multi-task learning for small brain tumor segmentation from MRI. *Appl Sci* 10(21):7790
- Noh H, Hong S, Han B (2015) Learning deconvolution network for semantic segmentation. In: Proceedings of the IEEE international conference on computer vision (ICCV)
- Plastiras G, Kyrkou C, Theodoridis T (2019) Edgenet: balancing accuracy and performance for edge-based convolutional neural network object detectors. In: International conference on distributed smart cameras
- Qian Y, Dolan JM, Yang M (2019) Dlt-net: joint detection of drivable areas, lane lines, and traffic objects. *IEEE Trans Intell Transp Syst* 21(11):4670–4679
- Redmon J, Farhadi A (2017) Yolo9000: better, faster, stronger. In: IEEE, pp 6517–6525
- Redmon J, Divvala S, Girshick R, Farhadi A (2016) You only look once: unified, real-time object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)
- Ren S, He K, Girshick R, Sun J (2015a) Faster r-cnn: towards real-time object detection with region proposal networks. *Adv Neural Inf Process Syst* 28:91–99
- Ren S, He K, Girshick R, Sun J (2015b) Faster r-cnn: towards real-time object detection with region proposal networks. In: Cortes C, Lawrence N, Lee D, Sugiyama M, Garnett R (eds) *Advances in neural information processing systems*, vol 28. Curran Associates Inc, New York
- Rfa B, Fga C (2019) Towards an efficient segmentation of small rodents brain: a short critical review. *J Neurosci Methods* 323:82–89
- Ronneberger O, Fischer P, Brox T (2015) U-net: convolutional networks for biomedical image segmentation. In: International conference on medical image computing and computer-assisted intervention
- Shafiee MJ, Chywl B, Li F, Wong A (2017) Fast yolo: a fast you only look once system for real-time embedded object detection in video. *J Comput Vis Imaging Syst* 3(1)

- Wang X, Shrivastava A, Gupta A (2017) A-fast-rcnn: hard positive generation via adversary for object detection. In: 2017 IEEE conference on computer vision and pattern recognition (CVPR)
- Xu Z, Wu Z, Feng J (2018) Cfun: combining faster r-cnn and u-net network for efficient whole heart segmentation
- Yi J, Wu P, Metaxas DN (2019) Assd: attentive single shot multibox detector. *Comput Vis Image Underst* 189:102827. <https://doi.org/10.1016/j.cviu.2019.102827>
- Zhao H, Shi J, Qi X, Wang X, Jia J (2016) Pyramid scene parsing network. In: IEEE computer society
- Zheng Z, Wang P, Ren D, Liu W, Ye R, Hu Q, Zuo W (2020) Enhancing geometric factors in model learning and inference for object detection and instance segmentation
- Zhou Z, Siddiquee MMR, Tajbakhsh N, Liang J (2018) Unet++: a nested u-net architecture for medical image segmentation. In: *Deep learning in medical image analysis and multimodal learning for clinical decision support*. Springer, New York, pp 3–11

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.