Weakly Supervised Liver Tumor Segmentation Based on Anchor Box and Adversarial Complementary Learning

Mengyao Fan, Haofeng Liu, Zhenyu Zhu, Changzhe Jiao, Shuiping Gou

Weakly Supervised Liver Tumor Segmentation Based on Anchor Box and Adversarial Complementary Learning

Mengyao Fan, Haofeng Liu, Zhenyu Zhu, Changzhe Jiao✉, and Shuiping Gou

School of Artificial Intelligence, Xidian University,
Xi'an 710071, Shaanxi, China
cjiao@xidian.edu.cn

Abstract. Segmentation of liver tumors plays an important role in the subsequent treatment of liver cancer. At present, the mainstream method is the fully supervised method based on deep learning, which requires medical experts to manually label a large number of pixel level labels for training, resulting in high time and labor cost. In this article, we focus on using bounding boxes as weak label to complete the segmentation task. It can be roughly divided into two steps. The first step is to use region mining technology to obtain pixel level labels from the bounding box. The second step uses pixel level labels to train the semantic segmentation network to obtain segmentation results. In the whole task, the quality of pixel level labels obtained from bounding boxes plays an important role in the performance of segmentation results. Therefore, our goal is to generate high-quality pixel level labels. Aiming at the problem that the current region mining method based on classification network is inaccurate and incomplete in object location, we use the Adversarial Complementary Learning module to make the network pay attention to more complete objects. We conduct analysis to validate the proposed method and show that our approach performs is comparable to that of the fully supervised method.

Keywords: Weakly supervised learning · Tumor segmentation · Bounding box annotation.

1 Introduction

Liver cancer is one of the most common and highest mortality cancers in the world. According to the World Cancer Report 2020, liver cancer ranks fifth in incidence and second in mortality [1].

A liver tumor segmentation mask obtained from a medical image such as computed tomography(CT) provides important delineation information of liver tumor, which is of great significance for subsequent diagnosis and treatment.

Recently, semantic segmentation networks based on deep learning have achieved excellent performance in the field of medical image segmentation [2]. However,

the fully supervised semantic segmentation network based on full pixel annotation needs a large number of pixel level labels to train, which leads to huge labor and time costs, especially in the field of medicine.

Therefore, Many scholars have focused on how to use weak label to complete the segmentation task. Figure 1 shows pixel level label and common weakly supervised label. In Fig 1, (a), (c), (d) and (e) are weakly typed annotations, while (b) is a complete pixel level annotation.

Image-level labels are the easiest to obtain. But it only indicates whether an object is present in the image, and provides no other information about the object's location or profile. Therefore, image-level labels make the problem very challenging and performance is limited.

Another common type of weakly label is the bounding box, which limits the object to a rectangular area and specifies the background area (outside the box). WSSL[3] first uses Dense Conditional Random Field(CRF) [4] to generate pixel-level labels and then carries out iterative training. Youngmin Oh et al, [5] proposed the background aware pooling method, which calculates the cosine similarity between the background features outside the box and the features inside the box, so as to obtain different weights for different positions inside the box and enhance the distinction between the background and the foreground inside the box. This process makes the generated label more accurate.

The region mining technique is usually based on the classification network to obtain the location and shape information of a specific class objects. At present, the mainstream region mining method for weakly supervised semantic segmentation is using Class Activation Map(CAM) [6] to obtain localization map, which is used to generate pixel level labels through refinement algorithms such as Dense CRF [4]. However, CAM often only highlights the salient area of the object, resulting in inaccurate generated pixel labels.

In this work, we focused on the segmentation of liver tumors using boxes. In order to make up for the shortcomings of CAM [6], we used an Adversarial Complementary Learning(ACoL) [7] module to mine the non-salient object regions.

The accuracy of classification and the completeness of acquisition object can be balanced by setting appropriate salient threshold. If the salient threshold is too high, only a very small part of the area will be shielded, which is not obvious

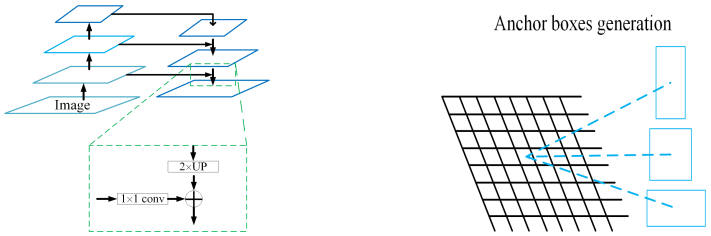
(a) image (b) fully pixel (c) box (d) scribble (e) point

Fig. 1. Examples of fully supervised mask annotation and weakly supervised box annotation

to the mining of the remaining area of the object. If the salient threshold is too low, the background may be identified as the foreground.

2 Approach

Our approach mainly consists of two stages. First, we trained a classification network based on Adversarial Complementary Learning [7] to obtain pixel level labels using box as positive and negative samples. The second is to train the semantic segmentation network with pixel level labels.



(a) (b)

Fig. 2. (a):Feature Pyramid Network. (b): Region Proposal Network anchor boxes generation process.

2.1 Anchor Boxes generation

The first problem we need to solve is how to obtain Bounding boxes from the image. It is obvious that the positive samples are the Ground truth bounding boxes containing tumors labeled by us. For the generation of negative samples, our solution strategy is as follows. More specifically, we adopt the anchor generation process of Faster RCNN with Feature Pyramid Network(FPN) [8] structure, which collects the features of different scale boxes on different feature layers through feature fusion.

Figure 2 shows the generation process of anchor boxes and FPN.

2.2 Adversarial Complementary Learning

For mining complete tumor from bounding boxes, we use Adversarial Complementary Learning(ACoL) [7] strategy, which is a CAM-based variant.

It is necessary to review CAM [6], which is one of the most basic mining techniques. After the feature of the picture is extracted by the feature extractor, followed by a Global Average Pooling(GAP) and a fully connected layer with length C(number of the classes). It is assumed that the feature map of the last

C channels is $S \in R^{W \times H \times C}$ and the weight of the fully connected layer is W. The localization map M_c of class c is calculated as follows:

$$M_c = \sum_k S_k \cdot W_k^c \tag{1}$$

Where S_k is the $k-th$ channel of S and W_k^c is the weight of class c for the $k-th$ channel.

The main idea of ACoL [7] consists of two classification branches A and B, which mine different regions through a complementary operation. A and B have the same structure. Where, the input feature map of branch B is guided by the tumor localization map of A to shield the salient regions (see Fig3). The tumor localization map M_A of A is subjected to RELU operation and normalized by min-max to obtain $\overline{M}_A$. The erasing operation is performed by setting a salient threshold δ. Specifically, if $\overline{M}_A$ in the position (i, j) is greater than δ, multiply the feature at (i, j) by 0 and finally send the feature after erasing the salient area to B to mine the non-salient regions.

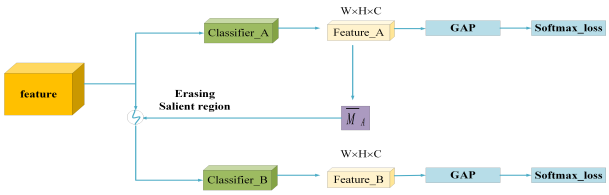


Fig. 3. ACoL architecture

2.3 Application

So far, the two main modules used to generate pixel-level labels from the bounding box have been introduced. Next, we use these two modules to build region mining model based anchor and ACoL [7], as shown in Fig 4. Negative Bounding boxes are defined as anchors completely non-overlapping with Ground truth bounding boxes.

It should be noted that we do not use all anchors generated on all feature layers, because this may lead to the imbalance of positive and negative samples.

Therefore, The sampling strategy is to first calculate the Euclidean distance between the center point of each anchor and the center point of its closest Ground truth bounding box. According to the distance, give priority to the negative samples with short distance. We select the top 10% anchor boxes of each layer according to the distance, and then randomly select N anchor boxes from the selected anchor boxes. Each classifier consists of two convolution layers, in which the output channel of the first convolution layer is 512 and the second is 2 (foreground and background). RELU operation is performed between two convolution layers.

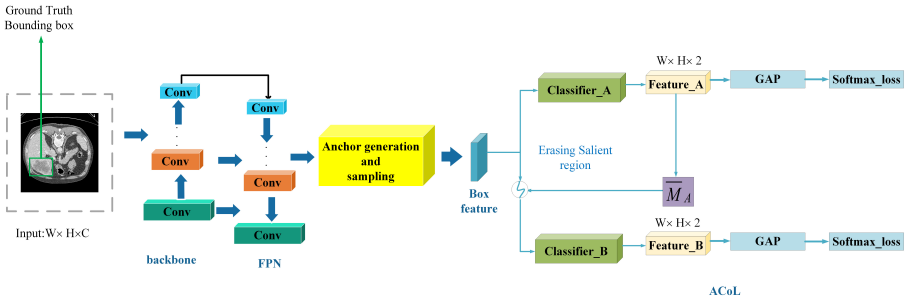
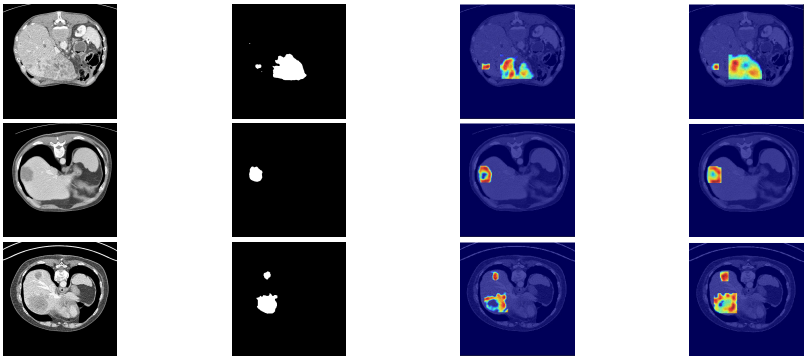


Fig. 4. The structure of regional mining method based on ACoL and anchor



(a) Input (b) Ground truth (c) CAM (d) ACoL

Fig. 5. Visualization of localization map generated by using CAM and ACoL

2.4 Pseudo mask generation

In order to generate pseudo pixel level labels. For a Ground truth bounding box, the tumor localization map M_A of A and the tumor localization map M_B of B are generated according to the pipeline of CAM [6]. M_A and M_B are subjected to RELU operation and normalized by min-max to obtain $\overline{M}_A$ and $\overline{M}_B$.

The fused object localization map $\overline{M}_{fuse}$ is defined as the element-wise maximum result over $\overline{M}_A$ and $\overline{M}_B$. The $\overline{M}_{fuse}$ is adjusted to the size of the Ground truth bounding box through bilinear interpolation. The ACoL [7] fused tumor localization map and tumor localization map of CAM are shown in Fig 5. We use Dense CRF [4] to estimate pixel level labels from bounding box localization maps. The unary term of Dense CRF for tumor class is set $\overline{M}_{fuse}$. The unary term of the background class is set $(1 - \overline{M}_{fuse})$.

Considering the influence of noise in the Ground truth bounding boxes, we adopt a method similar to [9], ignoring the background in the Ground truth bounding boxes(when training the segmentation model, the background area in the Ground truth bounding boxes will not be calculated in the loss).

3 Experiments

In this section, We will describe the experimental details and the environment. We use pytoch deep learning framework to build the proposed network model, and use NVIDIA 2080Ti GPU for training and verification.

3.1 Datasets and evaluated metric

In this paper, public dataset Liver Tumor Segmentation Challenge (LiTS-ISBI2017) is used as the research.

We use SGD optimizer with momentum of 0.9 and weight decay of 0.0005 train the region mining model based on classification network for 50 epochs, and the batch size was set to 8. The learning rate is initialized to 0.001.

The image size is 512×512. We select 3762 images from LiTS-ISBI2017 as the training set and 1669 images as the test set. The Dice coefficient is used as an evaluation metric.

3.2 Classification Network and Hyperparameter settings

ResNet50 is used as backbone. In ResNet [10], the layers with the same output feature size are classified into the same stage. The output of the last residual block of each stageC_i are used for features fusion and generate anchor.

The fusion results of $\{C_2, C_3, C_4, C_5\}$ according to FPN strategy are called $\{P_2, P_3, P_4, P_5\}$. Follow the settings of [8], the anchor area on P_2 is set to 32^2, P_3 to 64^2, P_4 to 128^2 and P_5 to 256^2 pixels, and use 3 aspect ratios $\{1{:}1, 1{:}2, 2{:}1\}$. For a feature map with size $W \times H$, a total of $W \times H \times 3$ anchors are generated.

The salient threshold δ mentioned in Section 2.2 is set to 0.6, and the number of negative samples N mentioned in Section 2.3 is set to 256. A box with height h and width w(on the input image) is assigned to P_k to obtain features, where k is calculated according to the following formula:

$$k = \lfloor 4 + \log_2^{(\sqrt{wh}/224)} \rfloor \tag{2}$$

3.3 Segmentation network and test results

We use the pixel level labels to train the semantic segmentation network. For the segmentation model, we choose Deeplab-v3[11] with ResNet-50 architecture as the backnone model, and use dice loss proposed in [12] for training. Dice loss can well solve the problem of imbalance between foreground and background in image segmentation. DeepLabV3 is trained for 50 epochs using the SGD optimizer with momentum of 0.9 and weight decay of 0.0001. Batch size is set to 6. The initial learning rate is 0.005. The learning rate adjustment strategy of DeepLabV3 is $(initial\_learning\_rate) \times (1 - \frac{iter}{max_{iter}})^{0.9}$.

The segmentation results on the test set are shown in Table 1. We define a naive baseline that treats all pixels in the boxes as foreground. The comparison of baseline, CAM and ACoL is shown in Table 2

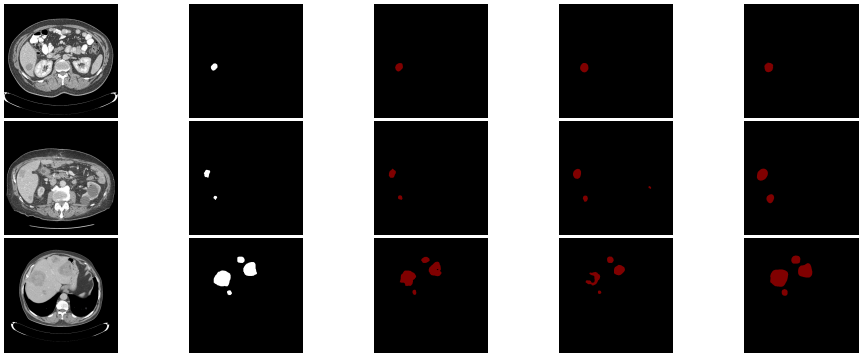
Table 1. Comparison of segmentation performance between based on ACoL method and fully supervised and weakly supervised methods on test set.

Methods	Annotation type	Dice
U-Net[13]	Pixel-level	0.702
DeepLab-V3[11]	Pixel-level	0.711
SDI_{box}[9]	Box-level	0.626
WSSL[3]	Box-level	0.632
Our	Box-level	0.658

Table 2. Comparison of baseline, CAM and ACoL segmentation results on test set.

Methods	Annotation type	Dice
Baseline	Box-level	0.563
CAM	Box-level	0.617
Our	Box-level	0.658

Finally, in Figure 6 we show some segmentation examples.



(a) Image **(b)** Ground truth **(c)** DeepLabv3[11] **(d)** WSSL[3] **(e)** Our

Fig. 6. Examples of predicted semantic masks

4 Conclusions

A weakly supervised liver tumor segmentation method based on box labeling is proposed with the help of ACoL [7] region mining strategy. The final results on the test set showed that the proposed method is comparable to the fully supervised method, which proved the effectiveness of the proposed method in liver tumor segmentation.

References

1. Christopher Wild, Elisabete Weiderpass, and Bernard W Stewart. *World cancer report: cancer research for cancer prevention*. IARC Press, 2020.
2. Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.
3. George Papandreou, Liang-Chieh Chen, Kevin P Murphy, and Alan L Yuille. Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation. In *Proceedings of the IEEE international conference on computer vision*, pages 1742–1750, 2015.
4. Philipp Krähenbühl and Vladlen Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. *Advances in neural information processing systems*, 24, 2011.
5. Youngmin Oh, Beomjun Kim, and Bumsub Ham. Background-aware pooling and noise-aware loss for weakly-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6913–6922, 2021.
6. Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016.
7. Xiaolin Zhang, Yunchao Wei, Jiashi Feng, Yi Yang, and Thomas S Huang. Adversarial complementary learning for weakly supervised object localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1325–1334, 2018.
8. Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017.
9. Anna Khoreva, Rodrigo Benenson, Jan Hosang, Matthias Hein, and Bernt Schiele. Simple does it: Weakly supervised instance and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 876–885, 2017.
10. Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
11. L-CCGP Florian and Schroff Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. In *Conference on Computer Vision and Pattern Recognition (CVPR). IEEE/CVF*, 2017.
12. Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, pages 565–571. IEEE, 2016.
13. Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.