Pattern Analysis and Applications (2024) 27:35
https://doi.org/10.1007/s10044-024-01251-6

THEORETICAL ADVANCES

A weakly supervised end-to-end framework for semantic segmentation of cancerous area in whole slide image

Yanbo Feng[1] · **Adel Hafiane**[1] · **Hélène Laurent**[1]

Received: 8 February 2023 / Accepted: 4 December 2023

Abstract
The segmentation of pathological image is an indispensable content in the cancerous diagnosis and grading, which is provided to doctors for the location and quantitative analysis of pathologically altered tissue. However, pathological whole slide image (WSI) generally has gigapixel size and huge region-level objective to be segmented. Extracting patches from WSI can address the limitation of computer memory, but the integrity of target is hence affected. Moreover, supervised learning methods require manually annotated labels for training, which is laborious and time-consuming. Thus, we studied a novel weakly supervised learning (WSL)-based end-to-end framework for semantic segmentation of cancerous area in WSI. The proposed framework is based on the block-level segmentation of convolutional neural network (CNN), while CNN is required to integrate the global average pooling layer and single fully connected layer as WSL-CNN. Class activation map and dense conditional random field (DenseCRF) are adapted to realize pixel-level segmentation of the cancerous area in patch, which is incorporated into the classification process of WSL-CNN. The hierarchically double use of DenseCRF effectively improves the precision of semantic segmentation. A region-based annotation method and a flexible method of constructing training dataset are proposed to reduce the workload of annotation. Experiments show that the block-level segmentation of CNNs has better performance than the pixel-level segmentation of fully convolutional networks, ResNet50 is the best one that achieves F1 score of 0.87426, Jaccard score of 0.78079, Recall of 0.94251 and Precision of 0.82182. The proposed framework can effectively refine the block-level prediction as semantic segmentation without pixel-level label. The precision of all tested CNNs get improved in the experiments, with WSL-ResNet50 achieving F1 score of 0.90630, Jaccard score of 0.83230, Recall of 0.92051 and Precision of 0.89789. We propose a complete end-to-end framework, including the specific structure of neural network, the construction of training dataset, the prediction method using neural network and the post-processing. CNN-like architectures can be widely transplanted into this framework to realize semantic segmentation, solving the problem of insufficient label of large-scale medical image to a certain extent.

Keywords Liver cancer segmentation · Deep learning · Weakly supervised learning · Image processing · DenseCRF · Machine learning

1 Introduction

Hepatocellular carcinoma (HCC) currently is the fifth most common malignancy [1], which is commonly found with predisposing factors such as cirrhosis, chronic liver diseases, prior infection with hepatitis B or C virus, and ingestion of aflatoxins [2]. Meanwhile, HCC is listed as the second leading cause of cancer-related death worldwide, its incidence is increasing globally [3]. With the development of biomedical technology, the clinical methods of HCC screening, therapy and surveillance are enriched. Some of them require the assessment of cancerous area in tissue sample, for instance, the proportion of tumor cells in genetic testing,

✉ Yanbo Feng
yanbo.feng@insa-cvl.fr

Adel Hafiane
adel.hafiane@insa-cvl.fr

Hélène Laurent
helene.laurent@insa-cvl.fr

[1] INSA CVL, University of Orléans, PRISME, EA 4229, 88 Boulevard Lahitolle, 18022 Bourges, France

Published online: 02 April 2024

the response rates for chemoradiotherapy, etc., where the accurate segmentation of the cancerous area in whole slide image (WSI) is the key work.

Biopsy diagnosis is the golden standard in pathology [4], the interpretation of pathological cytological specimens is typically performed under microscope [5]. Despite the importance of this task, manual segmentation is laborious, time-consuming and can suffer from subjectivity among pathologists [6]. With the development of digital pathology, a large amount of histological slides are digitalized to produce high-resolution images [7]. In particular, these images serve as an enabling platform for the application of computer vision. However, the complexity of digital pathology slides is higher than many imaging modalities because of its large size (a resolution of 10 gigapixels is common) and diversity of the composition of human tissue. In addition, some man-made factors are introduced in the process of preparation of pathological specimens, for example, dyeing shade, shape and thickness. Figure 1 presents examples of pathological images, the images on the right are cropped from the WSI presented on the left at the original resolution. These images show the huge size of WSI, the distinction of pathomorphology in tumor and normal tissues, and the complexity of interlaced foreground and background in cancer area.

With the development of artificial intelligence in the field of computer vision, the application of deep learning in medical image processing has received much attention in recent years due to its super ability. Over the past decade, deep learning based algorithms have gotten dominating place in many challenges [8], they have become an important tool for assisting medical experts to interpret and analyze medical images. When the input image is a WSI, it is a great challenge for the computer memory to feed the entire image into neural network, particularly the cascaded network architecture can dramatically enlarge the memory footprint. Correspondingly, the insufficient pixel-level annotation for WSI is another problem, since marking every pixel on gigapixel image is labor-intensive and time-consuming. Moreover, only pathologists are able to accurately give the label of pathological image. To address the difficulty of obtaining full ground-truth required in supervised learning, weakly supervised learning (WSL) [9] is proposed, which is a comprehensive concept that covers many approaches. Multiple Instance Learning (MIL) [10] is a typical representative of weakly supervised learning method, it achieves the classification of WSI using slide-level label, which is equivalent to produce a diagnostic report about the existence of cancerous cell. In practice, WSI is generally divided into a set of patches, deep learning model processes each patch separately, the slide-level label acts as a coarse-grained label for all tiles which is easily accessible from diagnostic report. Several applications of MIL have been found in the cancerous diagnosis of WSI [11–13], they are based on an assumption that there must be at least one patch cropped from a cancerous WSI that has cancerous cells, while there is none patch cropped from a normal WSI that contains cancerous cells. At the same time, the localization of the region-of-interest (ROI) of WSI is realized by choosing the tiles with high predicted score [14–16]. However, when it comes to the semantic segmentation of cancerous regions within WSI, the slide-level label is too coarse to train MIL method [17]. In that case, the fully supervised learning methods show the better performance [18–20]. Whereas the other weakly supervised learning methods are able to achieve semantic

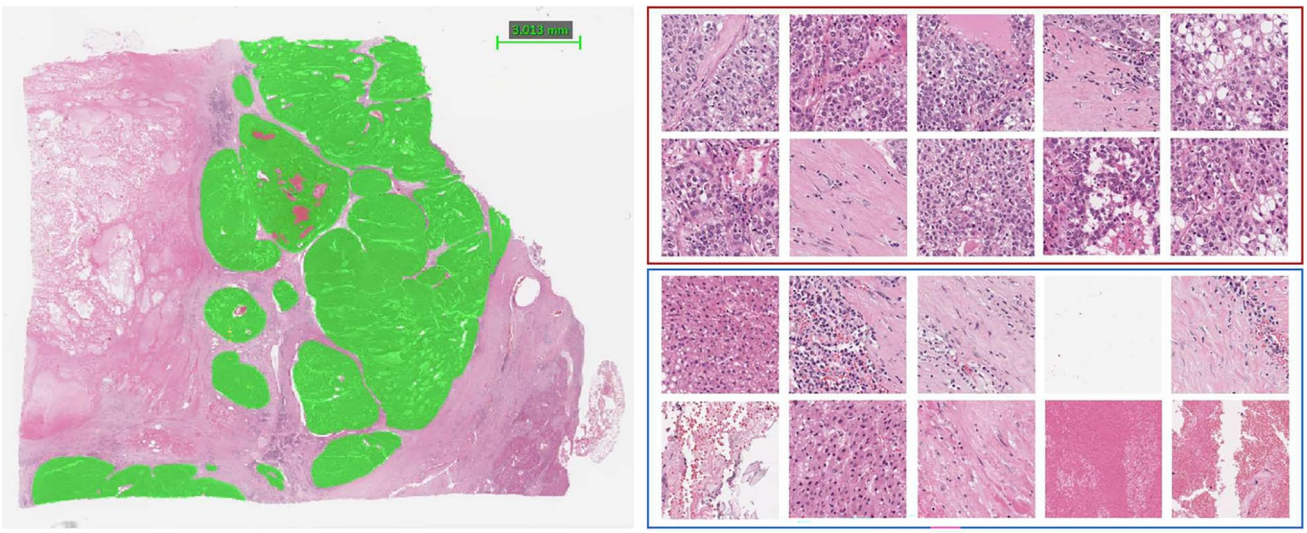

Fig. 1 A WSI at left with the size of 47,807 × 39,864 pixels, the green mask indicates the cancer area. Cropped patches with the size of 448 × 448 pixels are shown at right, blocks within blue rectangle are normal areas, blocks within red rectangle contain tumor areas

segmentation, the label needs to be refined but not at the pixel-level, meaning there is still a hierarchy of differences between the expected result and the label, for instance, image-level label [21], scribbles [22], points [23] and bounding boxes [24].

Based on the observation of WSI, it can be inferred that it is not necessary to do pixel-level prediction for all pixels in the task of segmenting region-level target. As illustrated in Fig. 2, it is more appropriate to take the images in (a), (b) and (c) columns into a task of classification, and then give them a block-level label, whereas the images in (d) column need to be segmented in pixel-level. This is a common situation that the non-target area takes a large part of WSI, it will waste the computation to process it at pixel-level. Moreover, the cropped patches, which are similar to those in Fig. 2a where all pixels belong to cancerous class, provide the neural network very limited receptive field since they are a part of whole target in WSI and lose the boundary information between target and non-target region.

Thus block-level segmentation based on convolutional neural network (CNN) of classification is another approach. It is usually seen in the task of target detection [25, 26] in which the patch is given with the predicted probability of CNN. It can be easily transformed into segmentation. More explicitly, each cropped patch of WSI is classified by CNN, then all pixels in that patch will be assigned with the predicted block-level label. Finally, all predicted patches are combined to the original size WSI. In this research, we do the comparison between block-level segmentation using CNNs and semantic segmentation using fully convolutional networks (FCNs). The five outstanding models of CNNs, VGG16 [27], ResNet50 [28], DenseNet121 [29], Inception V3 [30] and Xception [8], are adopted to do this task. At the same time, U-Net [31], DeconvNet [32], SegNet [33], Pyramid scene parsing network (PSPNet) [34], DeepLabV3 [35], Attention U-Net [36] and Global Convolutional Network (GCN) [37], are employed to do semantic segmentation of WSI. By quantifying these segmentation results, it is found that block-level method is more suited to the task of region-level segmentation, although the predicted results of CNNs are relatively coarse since it can not segment pixels inside patch.

Block-level segmentation can result in a large number of incorrectly predicted pixels since it treats all pixels in a cropped patch indiscriminately, which is troublesome when the proportion of target in a patch is small. As illustrated in the first row of Fig. 3, the patches are classified as cancerous category. Number 0 and 1 are used to represent the normal and cancerous category respectively, thus all pixels in each patch presented in the first row of Fig. 3 will be assigned with 1. Therefore, the non-target region which is the black area shown in ground truth will be wrongly predicted. To solve the problem of coarse block-level result, we propose in this study a simple and effective method employing class activation map (CAM) in weakly supervised learning [38] and full connected CRF [39]. In order to obtain CAM, we

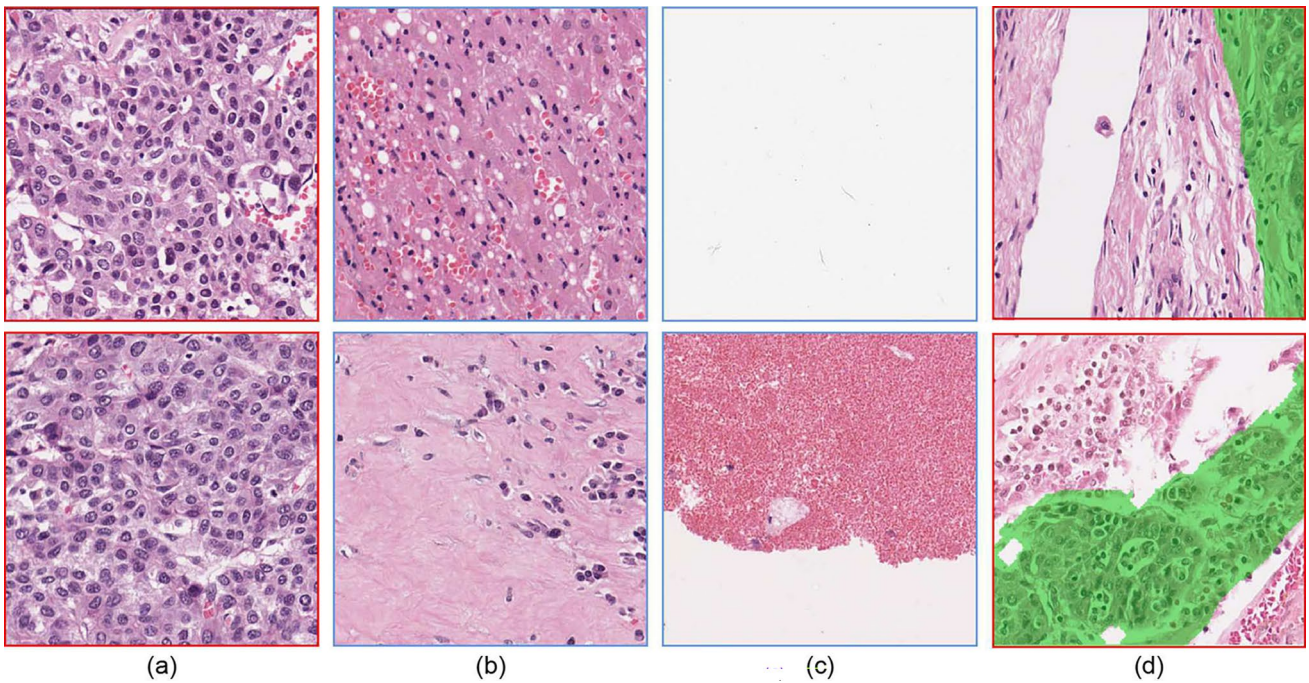

Fig. 2 Examples of patches cropped from WSI with the size of 448 × 448 pixels. **a** whole of patch is cancer area. **b** whole of patch is normal tissue. **c** patches contain background or background and normal tissue. **d** patches contain mixed areas where cancer area is indicated by green mask

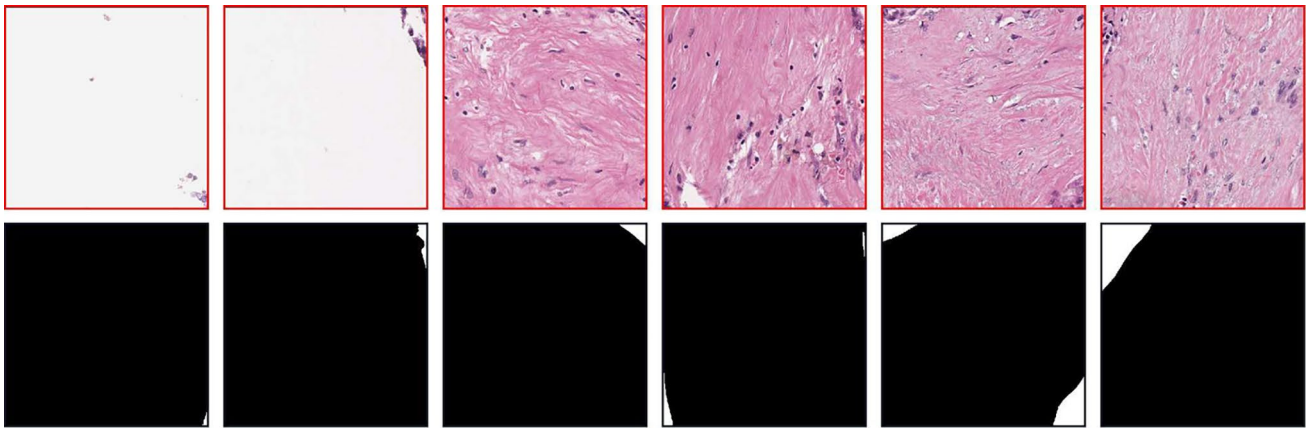

Fig. 3 The images with red rectangle in first range are examples of cropped patches which are predicted as positive, the images in second range are the corresponding ground truth, where white pixels indicates cancer area

have requirement for the architecture of neural network, one global average pooling (GAP) layer and one fully connected layer are used to modify the five CNN models as WSL-CNNs. For achieving better segmentation of target, we further study the use of full connected CRF. In addition, it may be noted that previously mentioned block-level methods [25, 26] still use the pixel-level annotation, and another method proposed by L. Chan et al. [40] uses patch-level annotation while the workload of annotating such a large dataset [41] is still heavy. In this study, we further study the potential of block-level segmentation to reduce the work pressure of annotation, thus a more flexible region-based marking method and the corresponding method of constructing dataset are proposed to simplify the work of labeling. Based on the above introduction, this study explores the possibility of using this method as a general framework for using more CNNs.

This method realizes the refinement of block-level segmentation into pixel-level segmentation without using pixel-level label. It makes CNNs more powerful without affecting the original classification function of CNNs. From the quantified results of experiments, our framework achieves the best performance compared with FCNs and CNNs. The main contributions of this paper are as follows:

- A universally applicable end-to-end solution is provided to employ CNNs to directly output pixel-level segmentation of region-level objective in whole slide image (WSI).
- Weakly supervised learning and DenseCRF were adopted to segment the patches with cancerous area in pixel-level. A hierarchically double use of DenseCRF effectively improves the precision of segmentation.
- A region-based marking method is proposed to effectively reduce the workload of annotation, it also includes a flexible method of constructing the training dataset and the criteria of classification for CNNs.

2 Materials and method

2.1 Dataset

The dataset used in this research comes from the 2019 MICCAI PAIP Challenge [42]. The original dataset contains 50 WSIs and ground truths for training, 10 WSIs and ground truths for validation, 40 WSIs for test. All WSIs are stained by hematoxylin and eosin, and scanned by Aperio AT2 at x20 power. In this research, we compared CNNs for block-level segmentation (essentially equivalent to classification), FCNs for semantic segmentation and the refined block-level segmentation. Two datasets are constituted based on the WSIs with ground truths provided for the training part. The first dataset is for CNNs: patches with the size of 448×448 pixels are cropped from WSIs, and their labels are patch-level. The second dataset is for FCNs: the annotation is then pixel-level. Here, a modification was implemented on the ground truth for better training of FCNs. The original ground truth has two classes, which are cancer area and the other area. It was found that the cancer area is composed of heterogeneous cellular components. For scattered tumor cells, there is large background inside or around tissue, which is able to mislead the FCNs during training. Thus the threshold of RGB value (235, 210, 235) provided by the challenge was used to make a third label which represents background.

2.2 Data augmentation

A large dataset is crucial for the performance of the deep learning model. Neural networks are expected to be robust in a variety of conditions, such as different orientation, location, scale, brightness etc. However, the datasets acquired in digital pathology are small and taken in a limited set of conditions. Data augmentation is then necessary, based on the principle that the images created must be related to

original images and as realistic as possible. According to the morphological characteristics of tissue sections and possible situations in the procedure of making WSI, data augmentation techniques used here are flip, rotation, crop and translation. They are systematically applied to all training datasets in this paper.

2.3 Segmentation framework

Figure 4 shows an overview of the proposed procedure for cancer region segmentation in histopathology images.

Firstly, because of the large size of WSI, it is impossible to input the whole WSI into network, thus the cropping is a necessary step. Secondly, the cropped patches are inputted into WSL-CNN to do classification. If a patch is classified as negative, which means there is no cancer area, a zero-value map will be generated. Thirdly, the patch predicted as positive is further processed, as its heatmap is calculated and normalized. Fourthly, the normalized heatmap and the corresponding patch are inputted into DenseCRF to generate a preliminary binary segmentation. Then this result and the corresponding patch are inputted into DenseCRF again to generate second segmentation. Finally, since the location of each cropped patch in WSI is known, all segmentation are assembled back to the size of WSI. A post-processing step is lastly applied to obtain the final result.

2.4 The architecture of WSL-CNN

In this research, WSL-CNNs are used to do classification of each cropped patch firstly, and then the feature maps of positive patches are considered to do segmentation. Figure 5 shows the detailed architecture of WSL-VGG16, because there are some major changes compared with the original VGG16.

All of the WSL-CNNs employ one fully connected layer as output layer to predict the label of input image, and GAP layer as the last pooling layer connected to the fully connected layer. The output layers of WSL-CNNs are composed of two units which correspond to the two classes in the task of classification, except WSL-DenseNet121 which has three units. Since it is observed in the experiments that, when

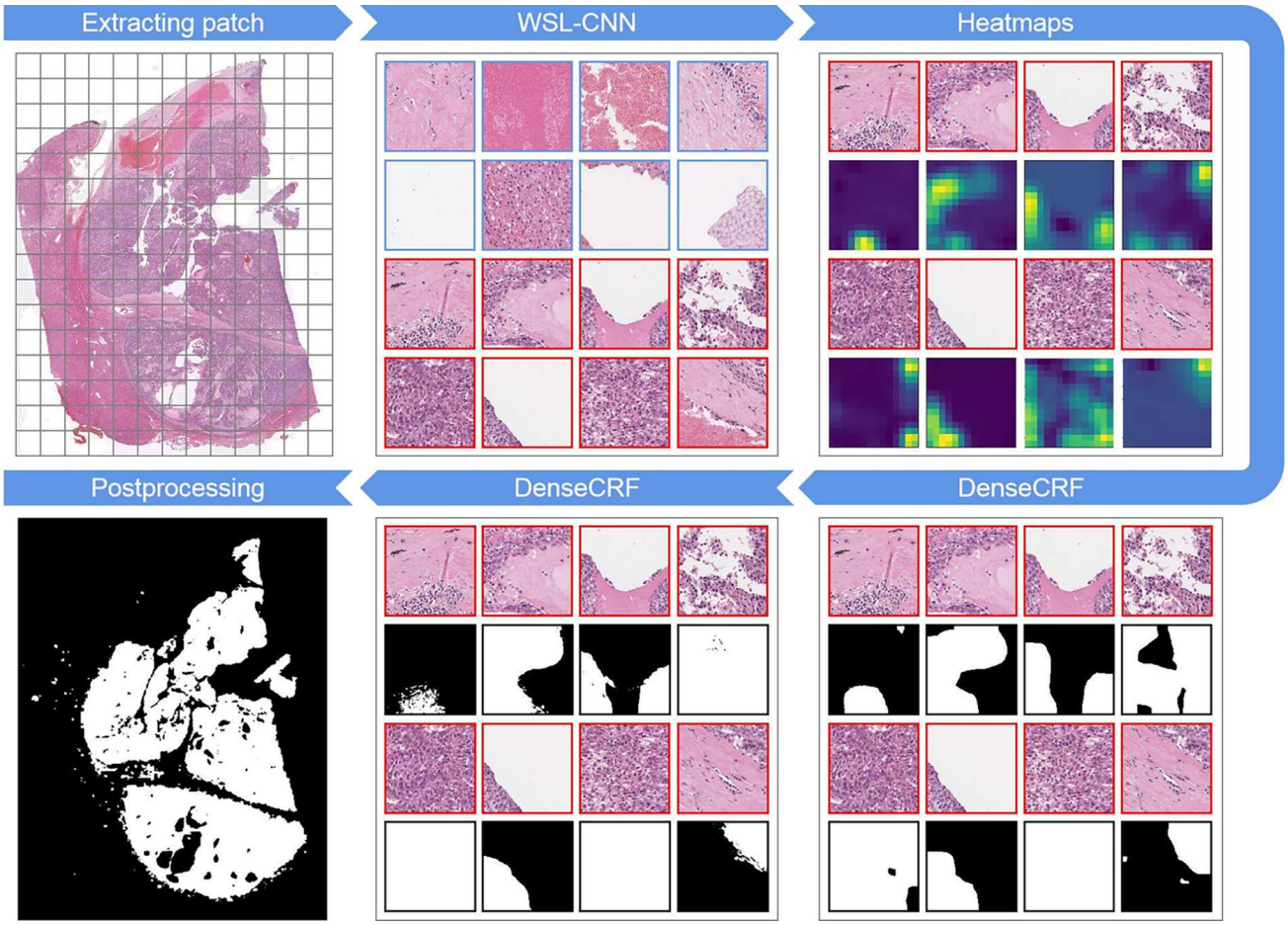

Fig. 4 Illustrations of the proposed framework for cancer area segmentation in histology images

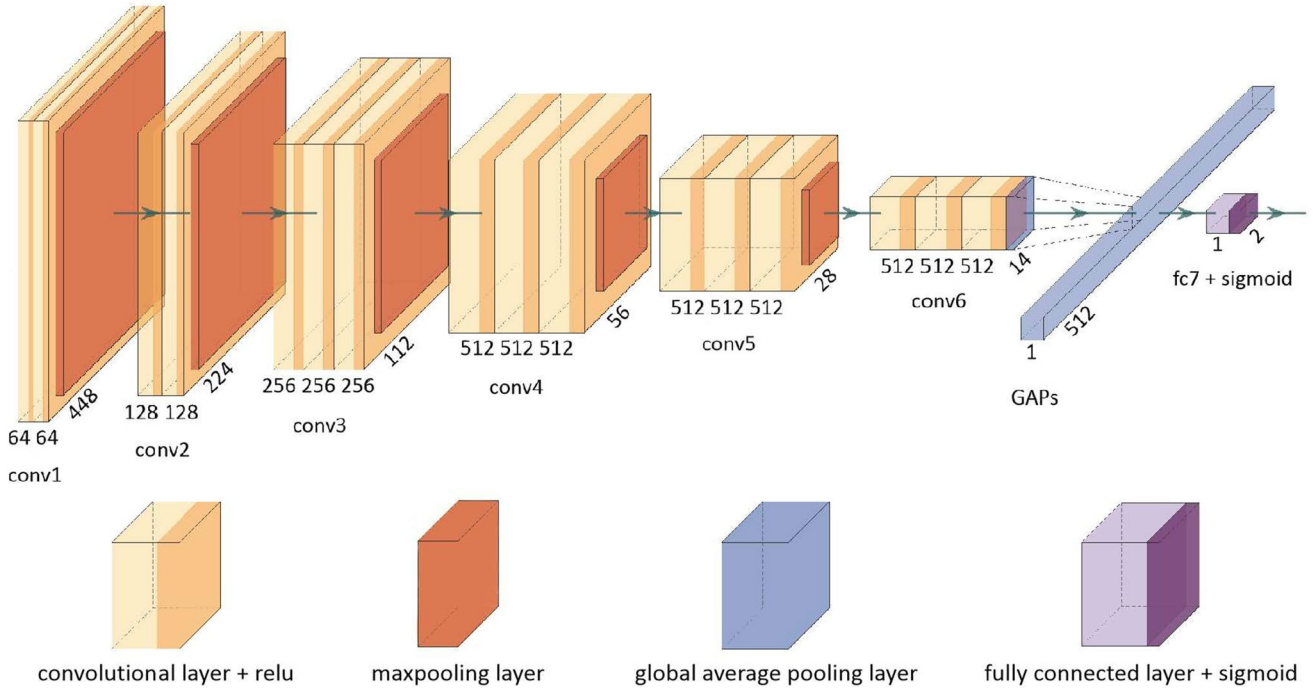

Fig. 5 Architecture of the WSL-VGG16. This network is built based on VGG16, the original fully connected layers are removed, the block of conv6 composed of three convolutional layers is added, the next layers are GAP layer and fully connected layer with two elements using sigmoid as activation

WSL-DenseNet121 has two output units, it cannot distinguish background patch and cancerous patch, thus another unit is added to represent background independently. The sigmoid layer is used as activation layer after the fully connected layer. The bias is not used in the weight of fully connected layer, since the effects of bias are systemic that it is equivalent to shifting sigmoid, which provides a differentiated standard for the generation of heatmaps. The abandon of bias brings positive effects on the acquisition of heatmaps, and further optimizes the segmentation results.

2.5 Training strategy

One of objective in this research is to effectively reduce the workload of annotation, since the huge size of WSI and pixel-level annotation are so challenging for pathologist. As it is well-know that deep learning is a data-driven procedure, large amounts of input data and corresponding label data are required for training neural networks in supervised learning. Weakly supervised learning employed in this research is also a kind of supervised learning, its significance lies in the use of block-level labels to obtain pixel-level segmentation results. Meanwhile, the WSI cannot be fully inputted into network because of the limitation hardware, the neural network is then trained by the patches cropped from WSI which make up the training dataset. In this research, a region-based annotation, a method of constructing training dataset and a standard of classification are proposed and adapted to weakly supervised learning for simplifying annotation.

As shown in Fig. 6, the cancer areas and normal tissue areas are annotated by red and blue rectangles respectively. This is the region-based annotation: the pathologist does not

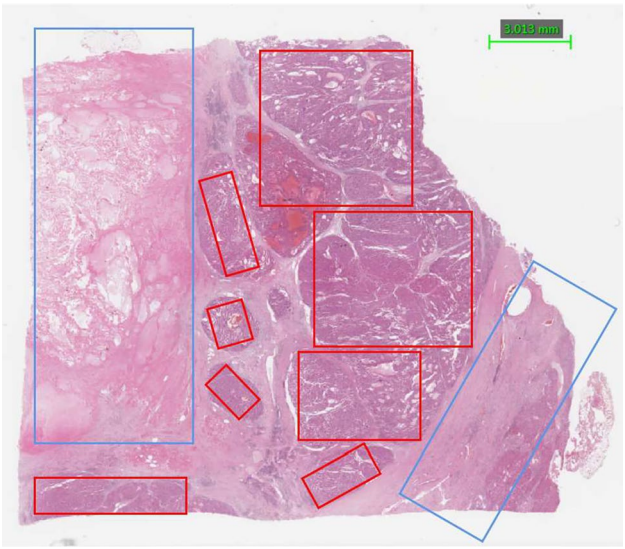

Fig. 6 Illustration of region-based annotation. The cancer areas are annotated in red rectangle, the normal areas and background are annotated in blue rectangle

need to label each pixel anymore. Then the dataset used in training step is constructed by cropping the patches from red and blue rectangles. A very important principle is to ensure the purity of the cropped patches of normal tissue, this can be realized by cropping patches inner the blue rectangles. The cropped patches of tumor tissue can exceed the red rectangles as long as they contain the contents of the red rectangles. Thus the principle of classification is that as long as the cropped image contains an area of cancer, regardless of the size of the area, it will be classified as cancer image (class-1). Cropped patches of pure normal tissue or background will be classified as normal image (class-0). This standard of classification makes network sensitive to the region of interest (ROI). At the same time, the high sensitivity of the network to ROI also makes the network provide more accurate feature maps, which will be put to profit afterwards.

2.6 Class activation map

Class activation map (CAM) [38] is a kind of heatmap obtained from CNN for particular objective. It indicates the ROIs captured by CNN automatically which are used as critical factors for classification. CAM is used in many WSL-based objective detection [43–45] and semantic segmentation [46–48]. In this research, CAM is obtained from the feature maps outputted by the layers before GAP layer. Using the spatial averages outputted by GAP layer to represent feature maps is able to emphasize the overall contribution of the feature maps to the final decision. The trained weights after GAP is image-level, thus they can be migrated directly to the feature maps. The CAM can be calculated as shown in Eq. 1.

$$CAM_c(x,y) = \sum_{i=1}^{N} W_c^i F_i(x,y) \tag{1}$$

where $CAM_c(x,y)$ is the pixel at position (x, y) on the CAM of class c (since this study is a two-category task, c can be 0 or 1 which means class-0 or class-1); N is the number of the channels of feature maps; $F_i(x,y)$ is the pixel at position (x, y) on the i_{th} feature map; W_c^i is the weight of class c in the fully connected layer corresponding to i_{th} feature map.

2.7 Dense conditional random field (DenseCRF)

DenseCRF [39] has been widely used in semantic segmentation [49]. It uses the probability map outputted by FCN to calculate a finer segmentation result [50, 51], compensating for the lack of correlation of each predicted pixel in FCN. The CAM is able to indicate the area of ROI through the pixel values, the location of ROI can be identified preliminarily by obtaining the position of the maximum pixel value, while exact boundary of ROI cannot be determined. Two heatmaps, which are complementary, are used in this work, they can be taken as probability maps after normalization. In order to obtain the pixel-level segmentation, DenseCRF is employed to process the CAMs and their corresponding input image.

In the model of DenseCRF, pairwise connectivities are established between any two pixels, since each pixel is considered to be related to the other pixels in the image. A conditional random field is modeled by Gibbs distribution, the energy function of DenseCRF is described as:

$$E(\mathrm{x}) = \sum_i \psi_u(x_i) + \sum_{i<j} \psi_p(x_i, x_j) \tag{2}$$

where $\psi_u(x_i)$ is the unary potential function, it takes into account the assignment x_i of each individual pixel and learns gradually from data during the solving process, generally it is initialized by probability map of each pixel. $\psi_p(x_i, x_j)$ is the pairwise potential function, it calculates the potential between any two pixels x_i and x_j in image, considering the whole image. Thus, the energy function of DenseCRF sums over the unary and pairwise potential of all pixels together.

After building the DenseCRF model, a method of mean-field approximation is employed to optimize the model iteratively until convergence, the objective of optimization is described as:

$$\hat{x} = \underset{x}{argmax}\, P(\boldsymbol{x}), \quad where \quad P(\boldsymbol{x}) = exp(-E(\boldsymbol{x})) \tag{3}$$

It can be inferred that the objective is to find the most likely assignment $\hat{x}$ for each pixel, which is also the final segmentation result. In this research, the CAMs are equivalent to the probability maps assigned to each pixel after normalization. Thus, they are used in unary potential to calculate preliminary result at first time. Different from heatmap, the preliminary result is a binary label map. Then this preliminary result is inputted into unary potential again as a hard label to get the final result.

2.8 Evaluation protocol

Four criteria were used to measure the performance of networks, including F1 score (*F1*), Jaccard score (*Jacc*), Recall and Precision. Let A be the set of predicted pixels, B the set of pixels from ground truth, $|.|$ a set cardinal, TP the true positives, FP the false positives and FN the false negatives. The four criteria are calculated based on pixels:

$$Precision = \frac{TP}{TP + FP} \tag{4}$$

$$Recall = \frac{TP}{TP + FN} \tag{5}$$

$$F1 = 2 \cdot \frac{precision \cdot recall}{precision + recall} \tag{6}$$

$$Jacc(A, B) = \frac{|A \cap B|}{|A| + |B| - |A \cap B|} \tag{7}$$

Generally, precision and recall are used to evaluate the effect of dichotomy model. But when these two indicators collide, it's hard to compare them between models. F1 score is defined as the harmonic average of precision and recall, it ranges from 0 to 1, better result corresponding to larger score. Jaccard score is used to compare the similarity and difference between finite sample sets, which measures the overlap between the ground truth and the prediction. The larger the jaccard score is, the higher the similarity is.

3 Experiments and results

3.1 Block-level segmentation

Block-level segmentation is achieved in classification task, five CNNs are tested using the validation data provided by the 2019 MICCAI PAIP Challenge. The WSIs are firstly tiled, and the patch-level label is inferred from the ground truths which have the same size of WSIs. Table 1 presents the quantified result of classification, the compared ground truth is patch-level. ResNet50 gets the most comprehensive performance, VGG16 and DenseNet121 achieve second and third performance, whereas Inception V3 and Xception that both use the structure of inception, are not sensitive to this task. Table 2 shows the evaluated result of block-level segmentation, the original pixel-level ground truths are used for evaluation. As might be expected, it can be observed that the results of metrics have fallen: because all pixels in a patch will be labeled as positive once CNNs classify that patch as positive, thus there is a large amount of FP pixels in block-level predictions.

3.2 CAM

Here, five CNN-models with different architectures are tested, leading to a different number of feature maps and weights used to calculate the CAM. The feature maps outputted by the layer before the GAP layer are used to calculate CAMs for each class. The ReLU layers in WSL-VGG16, WSL-ResNet, DenseNet50, Xception output 512, 2048, 1024, 2048 feature maps respectively, the concatenate layer of Inception V3 outputs 2048 feature maps. The number of weights N for each CAM corresponds to the number of extracted feature maps. Figure 7 presents 30 feature maps from ResNest50, which are continuously extracted from the 2048 available ones, the corresponding original input image is Fig. 8a.

Each feature map dispersedly highlights different part of image, and the spatial orientation is consistent with the input image, which facilitates the spatial mapping from feature map to input image. Generally, the emphasized points in feature maps show the characteristics of clustering, with some maps focusing on cancerous areas and others on normal areas. One phenomenon is observed that, when ROI takes

Table 1 The evaluation results (S: score; R: rank) of five CNNs in the task of classification

	F1 score		Jaccard score		Recall		Precision	
	S	R	S	R	S	R	S	R
VGG16	0.89133	2	0.81204	2	0.89323	3	0.91065	2
ResNet50	**0.90977**	**1**	**0.83716**	**1**	**0.91002**	**1**	**0.91370**	**1**
DenseNet121	0.85756	3	0.75706	3	0.90931	2	0.82935	4
Inception V3	0.69805	4	0.56674	4	0.78675	4	0.65829	5
Xception	0.66314	5	0.54731	5	0.61375	5	0.87484	3

Bold values indicate the best results of evaluations

Table 2 The evaluation results (S: score; R: rank) of the block-level segmentation of CNNs

	F1 score		Jaccard score		Recall		Precision	
	S	R	S	R	S	R	S	R
VGG16	0.85501	2	0.75761	2	0.92265	3	**0.82414**	**1**
ResNet50	**0.87426**	**1**	**0.78079**	**1**	**0.94251**	**1**	0.82182	2
DenseNet121	0.80954	3	0.69055	3	0.92775	2	0.74094	4
Inception v3	0.65915	5	0.51941	5	0.80669	4	0.58693	5
Xception	0.75779	4	0.63303	4	0.77221	5	0.80102	3

Bold values indicate the best results of evaluations

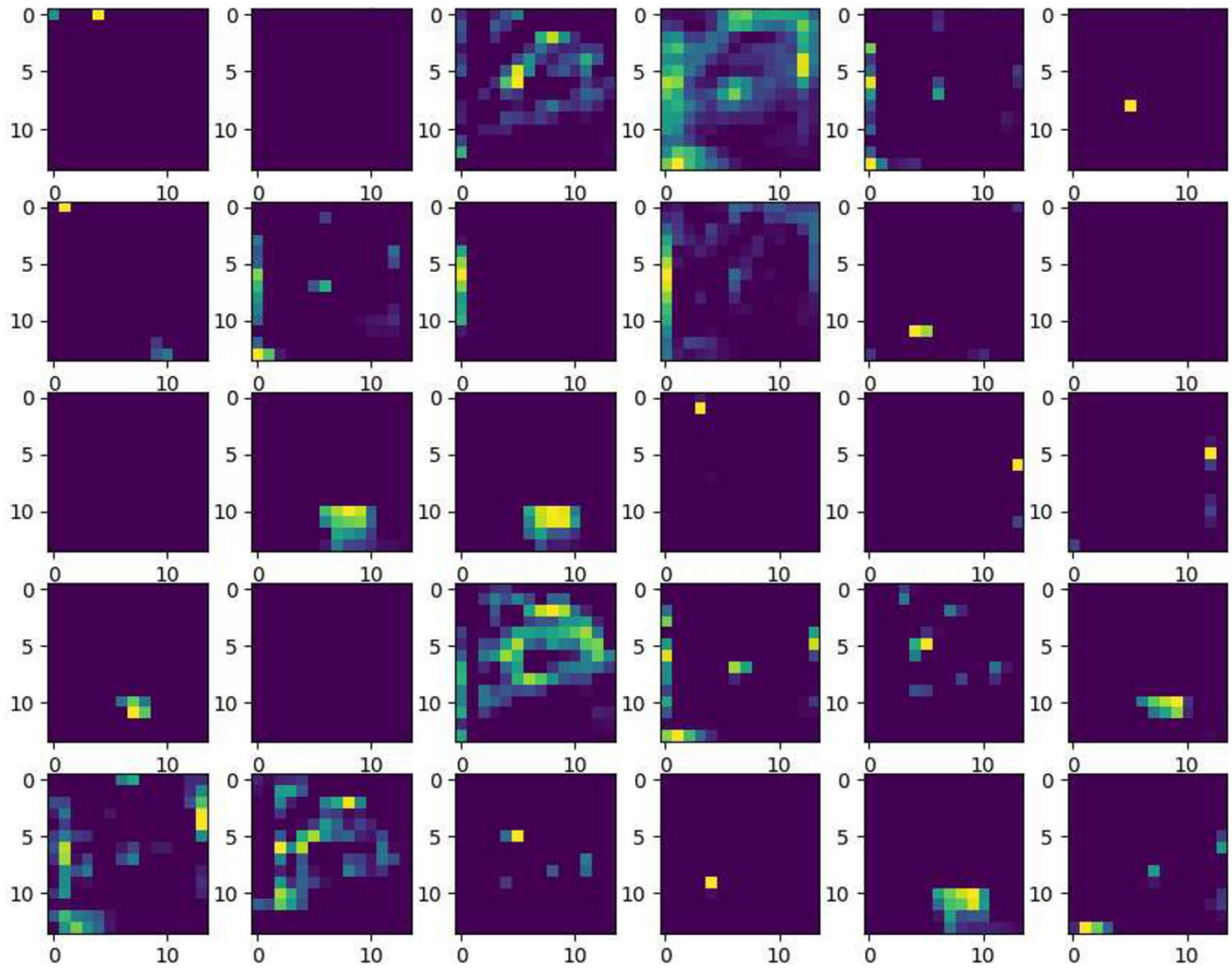

Fig. 7 Illustration of feature maps from WSL-ResNet50 for cancerous area segmentation (the corresponding original input image is presented in Fig. 8**a**)

a large percentage of image, normally less feature maps are able to activate the whole area of ROI. Based on Eq. 1, the CAMs are calculated, an example of the obtained results is presented in Fig. 8.

In this research, two CAMs are derived from the feature maps, which correspond to class-0 and class-1, they are the heatmaps of the distribution of ROI. By comparing the ROI indicated in Fig. 8 a, the CAM in Fig. 8b highlights background and normal tissue, and the CAM in Fig. 8c effectively highlights cancer area. It should be noted that the ability of finding the ROIs of corresponding classes is learned on network's own, no information about the location of ROI is given to the network during the training. Both heatmaps have relatively strict symmetry in the position of ROIs, and they have the same numerical effect, which can be observed from the histograms in Fig. 8b and c. The distribution is roughly same, the difference lies in the change of numerical signs. This effect is partly due to the abandonment of bias in the full connection layer of neural network in this study, so that the network has a unified threshold for judging two classes.

3.3 DenseCRF for segmentation

Heatmaps are obtained from the WSL-CNNs when the cropped patch is classified as positive during the procedure of classification. Although heatmaps are able to point out the location of cancerous area, the boundaries of ROI are not clear, thus heatmaps need to be binarized to achieve segmentation. In this research, DenseCRF is utilized twice to process the heatmaps and obtain the final result of segmentation. Firstly, the heatmaps are resized to the size of input image; secondly, heatmaps need to be normalized as probability distribution maps, then both of the heatmap and input image are inputted in DenseCRF in the first round.

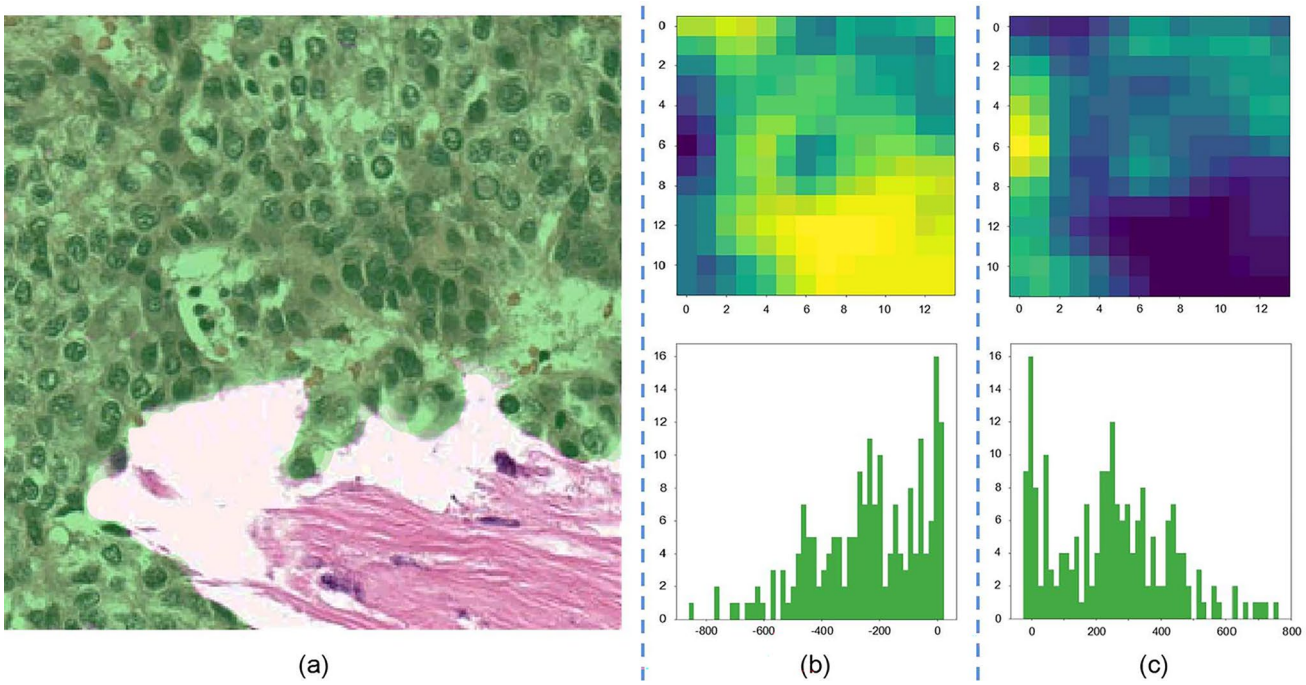

Fig. 8 Example of heatmaps obtained from ResNet50. **a** is a cropped patch, the green mask indicates the cancerous region. **b** presents the heatmap and its histogram for normal region. **c** presents the heatmap and its histogram for cancerous region

Figure 9e presents some examples of binary maps obtained after this first round.

Figure 9e illustrates that the results of firstly using DenseCRF are not accurate, the images are roughly segmented and the results only indicate the general areas of ROI. This preliminary segmentation is then used in the second round as a hard labeling, it is incorporated into DenseCRF with the corresponding input image to generate the final segmentation result. The results in Fig. 9 f are the final results, they are more refined, and the boundaries are clearer and more accurate. By comparing with the ground truths, the final results show more details, which are at the pixel level.

3.4 Post-processing

Information about the location and area of ROI is extracted from neural network itself, which has low accuracy. And cancerous region is a comprehensive area, it contains cancer cell, background, even some normal cells. Thus, it can be observed that the final results of DenseCRF have incorrectly predicted pixels, as shown in Fig. 10. Particularly for the images where all pixels are cancerous class, since their heatmaps do not highlight whole of ROI uniformly but only a few points. Consequently, DenseCRF emphasizes more the areas with high probability, resulting in holes in the segmentation result.

In this study two methods are utilized to remove those false pixels, one is morphological operation, another one is median filter. For both methods, the configuration of the shape and size of kernel is directly related to the final result. We tested discs, ellipses and squares with different sizes, as illustrated in Table 3. Taking into account the result of evaluation and the computing time comprehensively, a square structural element with size of 200×200 pixels is used to sequentially perform close and open operations for the morphological operation. A median filter with the kernel size of 209 is used here after comparing different kernel sizes of 101, 151, 175, 201, 209, 225, 251 and 301.

In order to evaluate the effect of denoising more clearly, the validation data processed by the five WSL-networks are quantitatively evaluated before and after post-processing, as illustrated in Fig. 11, both methods effectively remove unsatisfactory predicted pixels to a certain extent. In the processing of median filter, the false pixels are treated as salt &pepper noise. Comparatively, median filter tends to uniform areas even removing true positive pixels while morphological operations tend to uniform areas even removing true negative pixels.

It can be observed that morphological operations optimize the predicted results of all five networks in terms of F1 score, Jaccard score and Recall score, and the effect of optimization is better than that of median filter. Compared with before denoising, morphological operations globally reduce the Precision score. Median filter can optimize part of the results, in particular the Precision score, but it does not achieve ideal optimization effect on the whole.

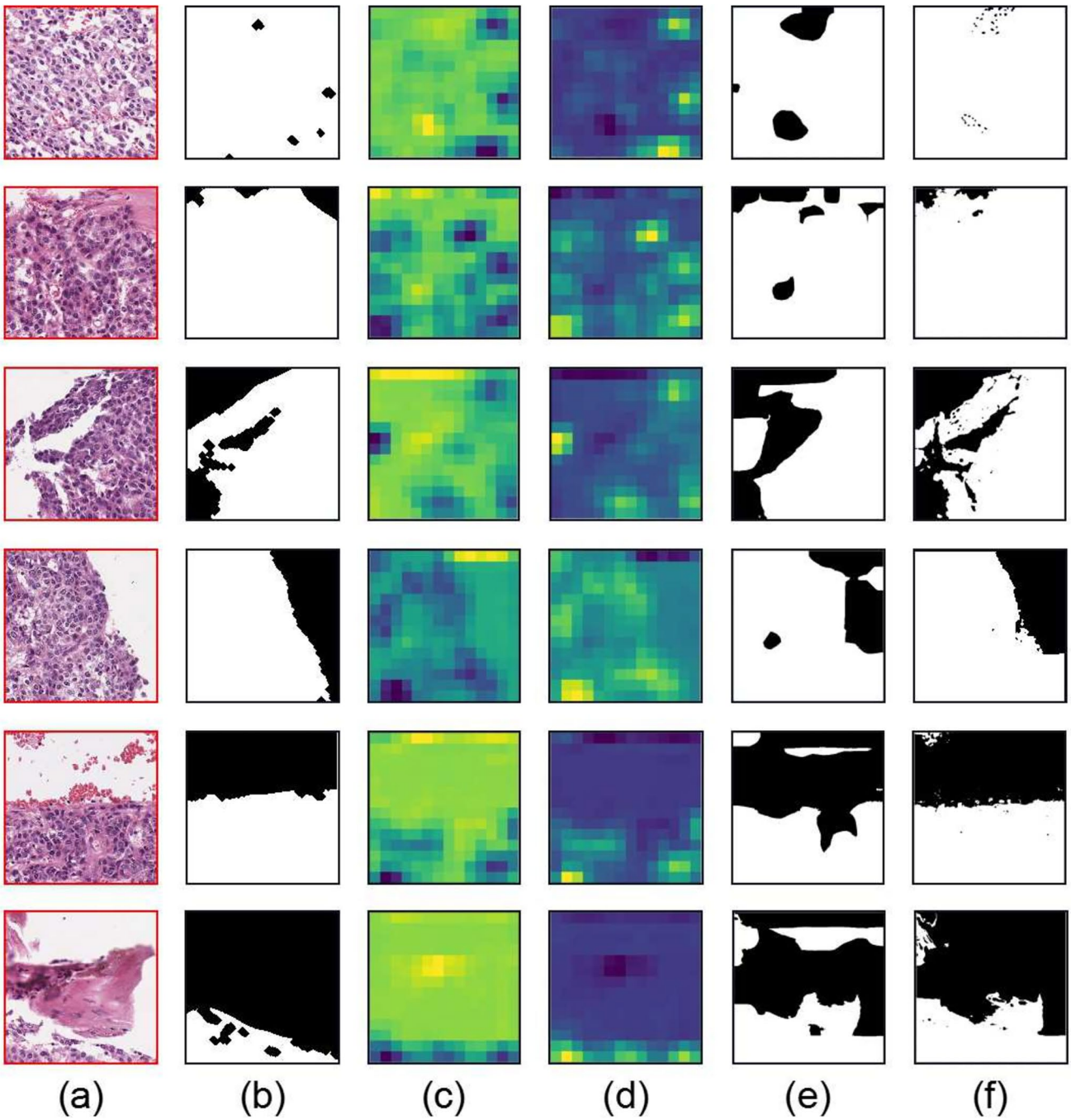

Fig. 9 Illustration of segmentation results obtained after use of DenseCRF. **a** examples of positive cropped patches. **b** ground truths. **c** heatmaps for class-0. **d** heatmaps for class-1. **e** results of DenseCRF at the first round. **f** results of DenseCRF at the second round

3.5 Pixel-level evaluation

In Table 4, the evaluation of semantic segmentation results of FCNs and WSL-CNNs are listed. Generally WSL-CNNs have better performance than FCNs. Although WSL-Xception gets the relatively worst evaluation result in WSL-CNNs, it is better than Deeplab V3 which is best one in FCNs in terms of F1 score, Jaccard score and Precision score. To complete the comparative study, we also present in this article the results achieved by a multi-scale method we proposed in [52], that employs model ensembling in the pyramidal layers of WSI. This multi-scale method yields good results, its Jaccard score is the best. Each network, whether WSL-CNN or FCN, has a different performance in this mission. Deeplab V3 is a very strict network, since it gets the best Precision score while the Recall score is

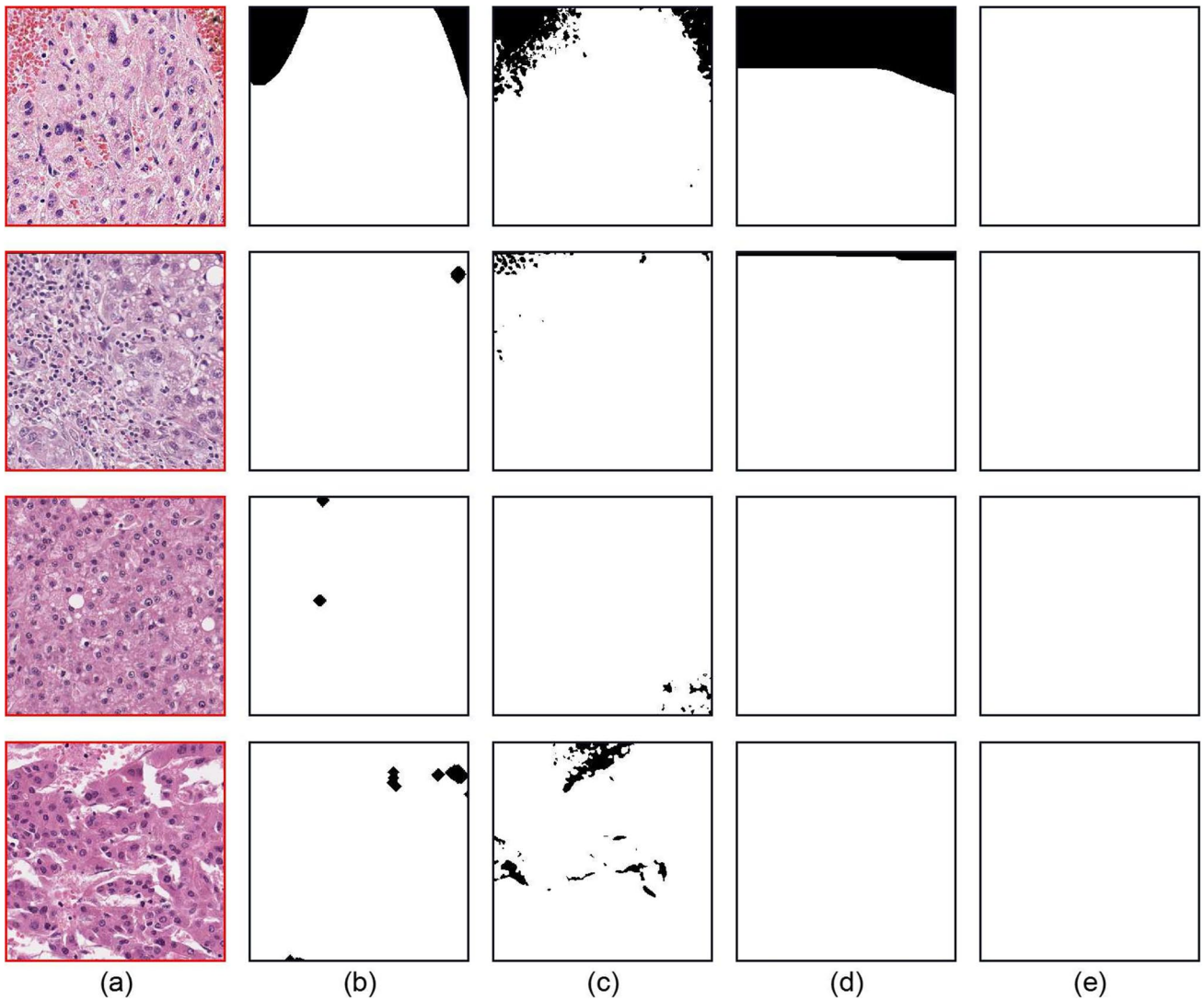

Fig. 10 Illustration of final segmentation results obtained after post-processing. **a** examples of positive cropped patches. **b** the corresponding ground truths. **c** the segmentation results obtained after applying dual-use DenseCRF. **d** the final results after median filtering. **e** the final results after morphological operations

Table 3 Comparison of the different configurations of structural element for morphological operation

Structures	F1 score	Jaccard score	Recall	Precision	Time(s)
Disc(r=50)	0.88215	0.79379	0.88716	**0.88450**	**493**
Disc(r=100)	0.88687	0.80111	0.89866	0.88211	1947
Disc(r=200)	**0.89621**	**0.81564**	**0.91893**	0.88005	7619
Ellipse(r_1=50, r_2=37)	0.88129	0.79251	0.88480	**0.88527**	359
Ellipse(r_1=100, r_2=75)	0.88575	0.79936	0.89594	0.88262	1471
Ellipse(r_1=200, r_2=150)	**0.89421**	**0.81261**	0.91433	0.88076	5735
Ellipse(r_1=37, r_2=50)	0.88135	0.79259	0.88498	0.88520	**354**
Ellipse(r_1=75, r_2=100)	0.88577	0.79938	0.89609	0.88250	1492
Ellipse(r_1=150, r_2=200)	0.89418	0.81254	**0.91435**	0.88068	5753
Square(l=50)	0.88275	0.79469	0.88920	**0.88358**	**35**
Square(l=100)	0.88787	0.80263	0.90163	0.88109	80
Square(l=200)	**0.89707**	**0.81684**	**0.92351**	0.87712	185

Bold values indicate the best results of evaluations

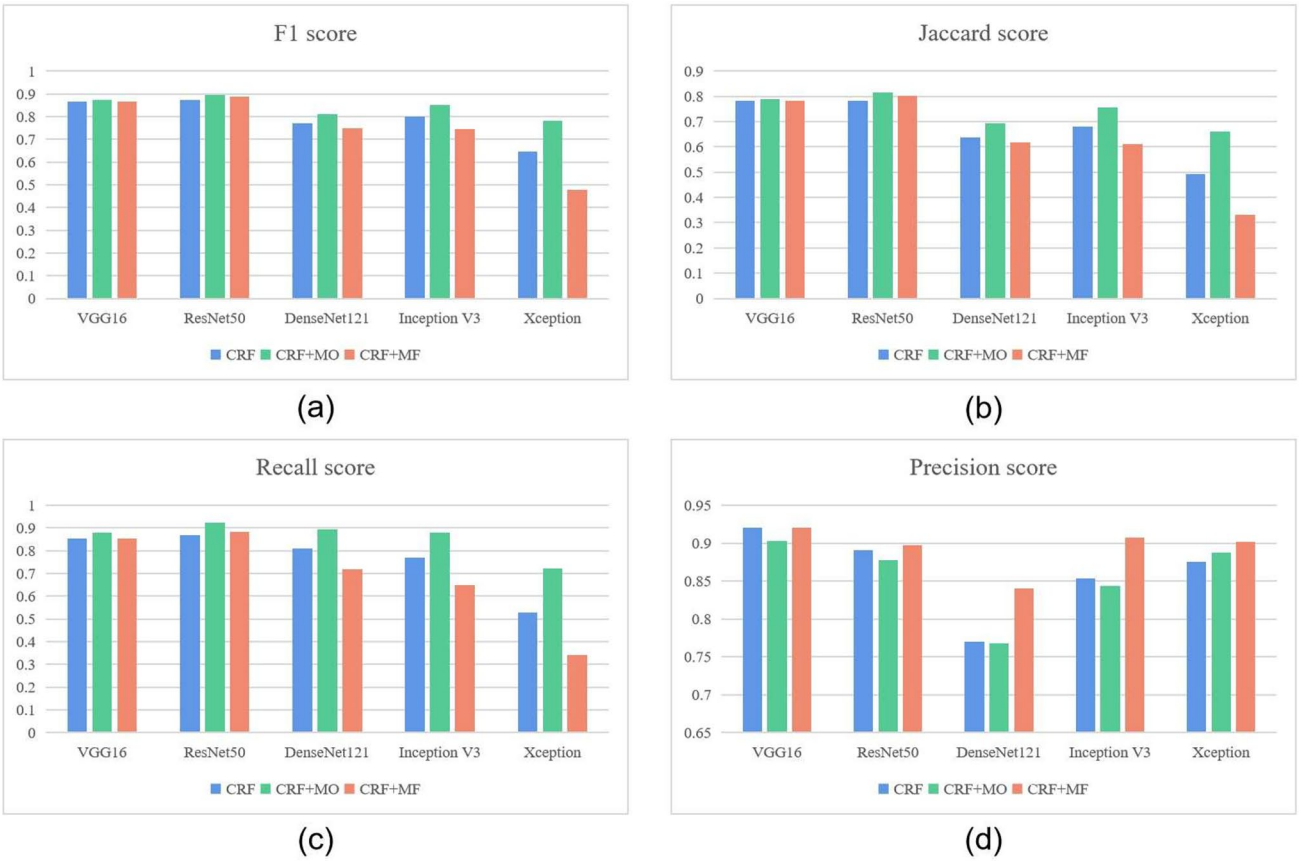

Fig. 11 Result of evaluation before and after post-processing. CRF represents the result without post-processing, CRF+MO represents the result processed by morphological operations, CRF+MF represents the result processed by median filter

Table 4 The pixel-level evaluation results (S: score; R: rank) of FCNs, multi-scale approach and WSL-CNNs

	F1 score		Jaccard score		Recall		Precision	
	S	R	S	R	S	R	S	R
U-Net	0.59823	12	0.49797	12	0.71865	9	0.60760	13
Deeplab V3	0.71474	7	0.63704	7	0.65738	11	**0.95927**	**1**
GCN	0.68666	9	0.59288	8	0.67319	10	0.87934	6
SegNet	0.70203	8	0.57608	9	0.86319	6	0.68189	11
DeconvNet	0.60947	11	0.50260	11	0.83853	7	0.62060	12
PSPNet	0.61164	10	0.52152	10	0.61608	12	0.72499	10
Atten-UNet	0.54082	13	0.39791	13	0.50147	13	0.73039	9
Multi-scale method	0.89200	2	**0.84493**	**1**	0.90781	2	0.88838	4
WSL-VGG16	0.87375	3	0.78964	3	0.87774	5	0.90281	2
WSL-ResNet50	**0.90630**	**1**	0.83230	2	**0.92051**	**1**	0.89789	3
WSL-DenseNet121	0.82016	5	0.70421	5	0.89207	3	0.78464	8
WSL-Inception V3	0.85307	4	0.75575	4	0.88036	4	0.84362	7
WSL-Xception	0.78232	6	0.65977	6	0.72046	8	0.88746	5

Bold values indicate the best results of evaluations

low, which means that it only predicts the most likely pixels. WSL-ResNet50 shows an overall good track record. It should be also noticed by comparing Tables 2 and 4 that the results of CNNs get significant improvement after the refinement of WSL. In particular, WSL-Inception V3 and WSL-Xception are greatly improved compared to the other three

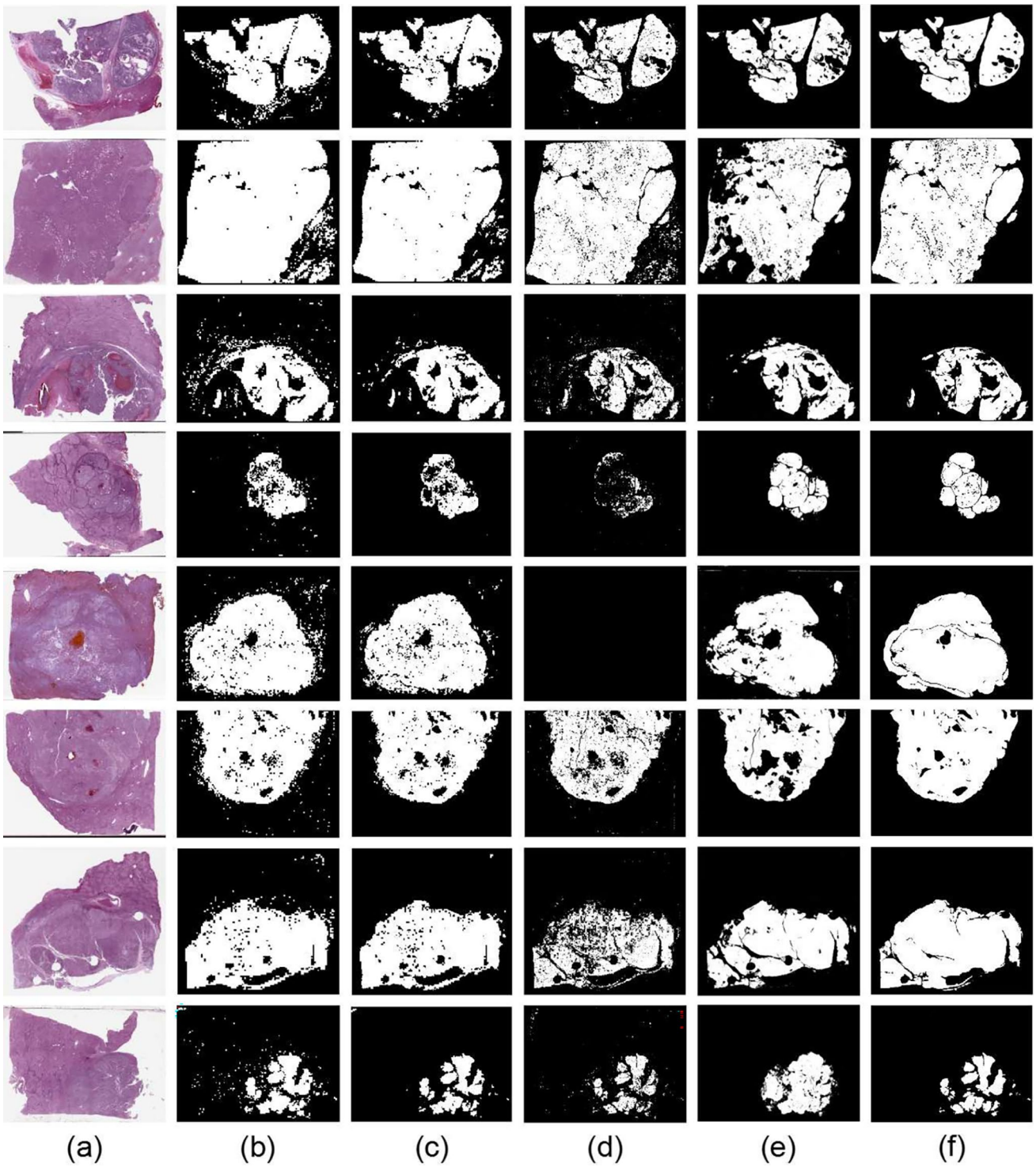

Fig. 12 Illustration of segmentation performances. **a** original WSIs. **b** block-level segmentation results of ResNet50. **c** semantic segmentation results obtained through WSL-ResNet50. **d** semantic segmentation results predicted by Deeplab V3. **e** semantic segmentation results of multi-scale method. **f** ground truths

networks. To illustrate the segmentation performances, the predicted results of ResNet50, WSL-ResNet50, Deeplab V3 and multi-scale method are selected to be shown in Fig. 12.

3.6 Time consumption

The time consumption in this task is mainly about two aspects, one is the annotation of ground truth, another one is computation. In this research, the WSL-CNNs are based on block-level segmentation, the patch-level labels are given to WSL-CNNs. Therefore compared with the annotation in semantic segmentation which is pixel-level the workload is reduced dramatically. In addition, WSL-CNNs are able to accept incorrect label during training. Especially, as a consequence of the cropping step, a small part of patches has relatively low percentage of ROI. This sort of images will be difficult for the pathologist to label, since it provides limited information. Based on the result of experiment, a small number of wrong labels is not able to mislead the WSL-CNNs. From an engineering perspective, training the network does not require WSIs fully annotated, it can be achieved by providing sufficient typical areas for each categories. Hence, it is sufficient to have a part of WSIs annotated at region-level and a part of WSIs annotated at pixel-level. This method can reduce the workload of annotation to a certain extent.

The comparison of computation between different methods is illustrated in Fig. 13, with the hardware settings that CPU: Intel(R) Xeon(R) W-2123 @ 3.60GHz, GPU: Quadro 4000, and RAM 32 G. Since FCNs and multi-scale method process images in same way, bars are used to represent the running time. However, WSL-CNNs firstly classify the patches, if a patch is classified as cancerous, then it will be further segmented, thus time is measured separately for processing the images of cancer and non-cancer. Figure 13 illustrates that when WSL-CNNs process the non-cancerous images, which is equivalent to classify the input images, the consumed time is generally less than the time spent by FCNs. However, when they process the images that contain ROI, the acquisition of CAMs and DenseCRF significantly increases the time consumption. As illustrated in Fig. 14, the time consumption of WSL-ResNet50 is strongly related to the amount of cancerous patches, while the time consumption of other two methods is proportional to the number of total patches. Concerning the time requirement of the multi-scale method, it is significantly increased due to the need to predict and fuse multi-layer images. After comparing various methods, the

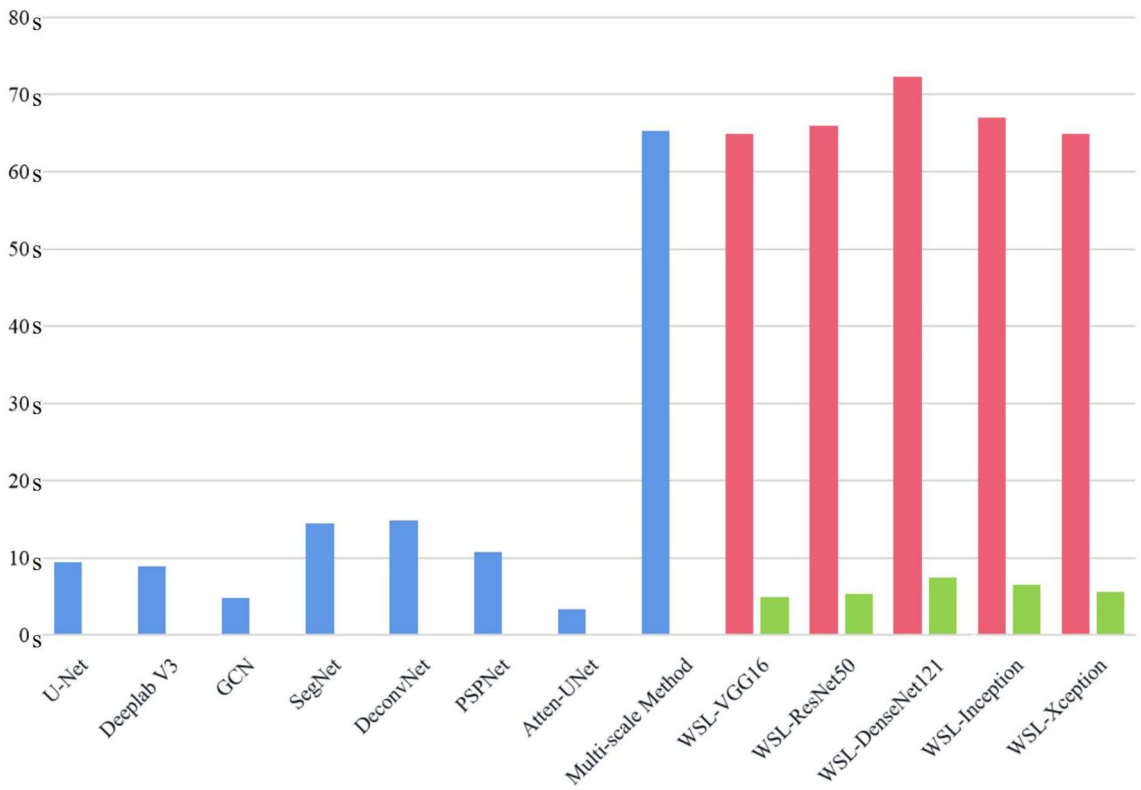

Fig. 13 The computation time of different methods. The bars represent the time (s: second) that each network takes to process 100 images with the size of 448 × 448. For WSL-CNNs, the red bars indicate that the images to be processed are classified as class-1, the green bars indicate that the images to be processed are classified as class-0

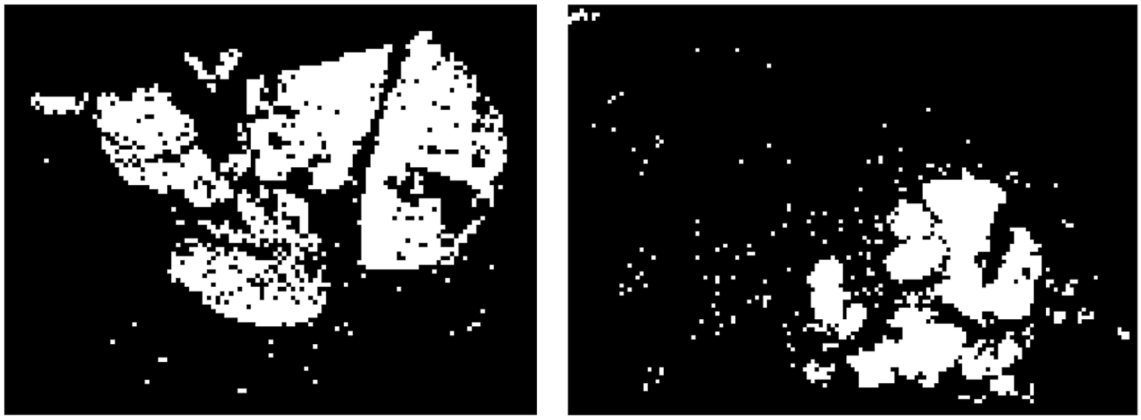

Fig. 14 Two examples for comparing the computation time, the size of patch is 448 × 448 pixels. **a** there are 7134 class-0 patches and 4119 class-1 patches. Multi-scale method takes 7347 s, U-Net takes 1057 s and WSL-ResNet50 takes 3092 s. **b** there are 12986 class-0 patches and 1743 class-1 patches. Multi-scale method takes 9617 s, U-Net takes 1325 s and WSL-ResNet50 takes 1823 s

efficiency of WSL-CNNs is higher in terms of time consumption than that of the multi-scale method. The comparison of computation time efficiency between WSL-CNNs and FCNs is mostly in favor of FCNS, even if this depends on the target size contained in WSI, but WSL-CNNs are better than FCNs in terms of segmentation performance.

4 Discussion

4.1 Comparison with other methods

In general, using WSL method to achieve semantic segmentation is divided into two steps, the first step is to use a WSL method to create or enrich the pixel-level annotation; the second step is to use the generated annotation to train a fully supervised model, and then realize semantic segmentation [53, 54]. Notably, our proposed framework is an end-to-end way which directly outputs the pixel-level segmentation result, it is simple but has competitive performance. Specifically, it combines fully supervised learning and weakly supervised learning in the task of classification, this method is based on the experience from practical engineering that large part of cropped patches have only single category, there is no boundary of the object to be segmented in most images, thus the neural networks for classification are preferred. However, the patches that contain boundary need to be segmented at pixel-level to satisfy the semantic segmentation, this is achieved by using the semantic information implied in the network and the textural feature of image. Overall, our proposed framework directly outputs semantic segmentation by combining the results of these two parts. In terms of semantic segmentation, the framework of this study does not need pixel-level annotation, only patch-level, which naturally reduces the workload of annotation. A corresponding region-based annotation method is also proposed for rapidly building a dataset, which is also successfully used in this study.

4.2 The performance of CNNs and FCNs

In this study, 5 CNNs are employed for the block-level segmentation, 5 WSL-CNNs and 7 FCNs are utilized for pixel-level segmentation. All results are evaluated in pixel-level, as illustrated in Tables 2 and 4. By comparing these two tables, several interesting points can be observed.

Firstly, the networks of Inception V3 and Xception have relatively bad performance compared with the other three CNNs. Based on the results of Table 2, the Inception V3 predicts the pixels of target with the Recall score of 80.6%, but it makes more incorrect pixels with the Precision score of 58.6%. Xception has higher Precision score reaching 80.1%, but it only successfully predicts 77.2% of the pixels of target. These are two completely opposite situations. Inception V3 is relatively active, it can easily predict tumor, but most of them are wrong. Xception is conservative and the prediction of tumor is cautious. Our proposed WSL method significantly improves the WSL-Inception V3 and WSL-Xception over the other WSL-CNNs by increasing precision score.

Secondly, the performance of FCNs is generally worse than CNNs and WSL-CNNs. Convolutional layer construct the whole structure of FCN, it can be trained end-to-end, and directly outputs semantic segmentation result. Generally the result of FCN is more elaborate than the block-level segmentation of CNN. However, the comparison between the pixel-level evaluation of CNNs and FCNs in Table 2 and Table 4 shows that the block-level segmentation achieves better performance.

In Table 2, VGG16, ResNet50 and DenseNet121 have very good recall score, it means that they can correctly predict most of the tumor region. This is very important that missing a tumor has a higher cost than a false positive in clinical diagnosis. The precision scores of VGG16 and ResNet50 in Table 2 are pretty good. For FCNs, Deeplab V3 and GCN show even better precision score in Table 4 but the recall score is low which means that they miss positive predictions.

One reason could be that the FCNs independently treat each pixel that needs to be predicted, they do not take into account the relationship between this pixel and the surrounding pixels. This defect is further amplified in the pathological images of liver cancer since pathological image contains a very large number of cells and surrounding tissues. These elements are repeated over and over again especially in the nucleus. The diagnosis of cancer area is a comprehensive judgment, it could contain tissue that does not have the hallmarks of cancer. On the opposite CNNs predict the label of input image at block-level, which allows CNNs to consider the features of input image globally, while it could assign the wrong label to pixels with the predicted patch-level label. It can be observed from Table 4 that the precision scores of WSL-CNNs are better than CNNs, which means that the incorrect pixel-level labels inherited from block-level label have been corrected. Although there is a loss in recall score, better comprehensive results are obtained in F1 score and Jaccard score.

4.3 The underlying limitation of full convolutional networks in the segmentation of area

From the result of quantitative evaluation, WSL-CNNs have interesting performance compared with other networks, while the performance of FCNs is disappointing. As networks with the ability to segment objects more delicately, they have more faults. One reason could be that the objects in this task are region. As can be observed in Fig. 9, the objective is composed of heterogeneous components, including different kinds of cells, blood vessels, stromal tissue and even background. In the area where cells are sparse, some region between cells is filled with normal tissues and background but is also annotated as positive. As known that more precise the annotation is, better performance FCNs have, the phenomenon of inaccurate annotation constitutes a challenge to FCNs.

In addition, the boundary of ROI does not have obvious regularity. It is the edge of a cluster of cancer cell, but it does not have naturally biological characteristics. On the contrary, the edge of ROI is closely relative to the procedure of labeling WSI, which is random and irregular. Another problem could be that the cropping operation makes the ROI of WSI lose its integrity. A part of cropped patches has 100 percent of ROI, these cases lose the boundary of ROI. Generally FCNs label out the areas around cancer cells where are thin tissue and background. This result could be understood as that FCNs take the cells as objects. On the opposite, these potential problems have little impact on CNNs, since CNNs are intended to do classification of image, what they need to focus on is the presence of cancer area, rather than which pixel belongs to cancer area.

5 Conclusion

In this research, we have proposed a weakly supervised end-to-end framework which can directly generate semantic segmentation of the cancerous area in WSI. The architectures of CNNs are modified with the global average pooling layer and single fully connected layer. Then weakly supervised learning CNNs are trained to detect if there exists ROI, even if the ROI is an extremely small area. Based on the block-level segmentation, the feature maps are derived from the WSL-CNNs for the classified cancerous patches, then they are used to calculate the heatmaps of class-0 and class-1 based on the equation of class activation map. Heatmaps are processed twice by DenseCRF after normalization. Finally, the segmentation results of patches are assembled back to the original size of WSI, the final segmentation result is obtained after the post-processing of morphological operation. We applied this method with five CNNS-like architectures, at the same time, seven FCNs-like architecture are applied to the same task. The results of CNNs, WSL-CNNs and FCNs are compared, the evaluation shows that the block-level segmentation conducted by CNNs is better than pixel-level segmentation conducted by FCNs when the objective is region-level in pathological image. Our method effectively realizes the transformation of CNN from block-level segmentation to pixel-level, achieving significant improvement.

Acknowledgements The authors gratelfully acknowledge financial support from China Scholarship Council. Deidentified pathology images and annotations used in this research were prepared and provided by the Seoul National University Hospital by a grant of the Korea Health Technology R &D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (grant number: HI18C0316).

Data availability The datasets analyzed during the current study are publicly available. These datasets were derived from the following public domain resources: http://www.wisepaip.org/paip.

Declarations

Conflict of interest The authors have no relevant financial or non-financial interests to disclose. The authors declared that they have no conflicts of interest to this work. We declare that we do not have any commercial or associative interest that represents a conflict of interest in connection with the work submitted.

References

1. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, Bray F (2021) Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 71(3):209–249
2. WHO (2017) Global hepatitis report 2017
3. Llovet JM, Castet F, Heikenwalder M, Maini MK, Mazzaferro V, Pinato DJ, Pikarsky E, Zhu AX, Finn RS (2022) Immunotherapies for hepatocellular carcinoma. Nat Rev Clin Oncol 19(3):151–172
4. Salamat, Shahriar M (2010) Robbins and Cotran: pathologic basis of disease. J Neuropathol Exp Neurol 69(2):214–214
5. Evered A, Dudding N (2011) Accuracy and perceptions of virtual microscopy compared with glass slide microscopy in cervical cytology. Cytopathology 22(2):82–87
6. Graham S, Chen H, Gamper J, Dou Q, Heng PA, Snead D, Tsang Y, Rajpoot N (2018) Mild-net: minimal information loss dilated network for gland instance segmentation in colon histology images. Med Image Anal. https://doi.org/10.1016/j.media.2018.12.001
7. Janowczyk A, Madabhushi A (2016) Deep learning for digital pathology image analysis: a comprehensive tutorial with selected use cases. J Pathol Inform 7:29. https://doi.org/10.4103/2153-3539.186902
8. Chollet F (2017) Xception: Deep learning with depthwise separable convolutions. In: 2017 IEEE Conference on computer vision and pattern recognition (CVPR), pp 1800–1807. https://doi.org/10.1109/CVPR.2017.195
9. Zhou Z-H (2018) A brief introduction to weakly supervised learning. Natl Sci Rev 5(1):44–53
10. Dietterich TG, Lathrop RH, Lozano-Pérez T (1997) Solving the multiple instance problem with axis-parallel rectangles. Artif Intell 89(1–2):31–71
11. Carbonneau M-A, Cheplygina V, Granger E, Gagnon G (2018) Multiple instance learning: a survey of problem characteristics and applications. Pattern Recogn 77:329–353
12. Cheplygina V, de Bruijne M, Pluim JP (2019) Not-so-supervised: a survey of semi-supervised, multi-instance, and transfer learning in medical image analysis. Med Image Anal 54:280–296
13. Wang X, Yan Y, Tang P, Bai X, Liu W (2018) Revisiting multiple instance neural networks. Pattern Recogn 74:15–24
14. Campanella G, Hanna MG, Geneslaw L, Miraflor A, Werneck Krauss Silva V, Busam KJ, Brogi E, Reuter VE, Klimstra DS, Fuchs TJ (2019) Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. Nat Med 25(8):1301–1309
15. Courtiol P, Tramel EW, Sanselme M, Wainrib G (2018) Classification and disease localization in histopathology using only global labels: a weakly-supervised approach. arXiv preprint arXiv:1802.02212
16. Kanavati F, Toyokawa G, Momosaki S, Rambeau M, Kozuma Y, Shoji F, Yamazaki K, Takeo S, Iizuka O, Tsuneki M (2020) Weakly-supervised learning for lung carcinoma classification using deep learning. Sci Rep 10(1):1–11
17. Frénay B, Verleysen M (2013) Classification in the presence of label noise: a survey. IEEE Trans Neural Netw Learn Syst 25(5):845–869
18. Bulten W, Bándi P, Hoven J, Loo R, Lotz J, Weiss N, Laak J, Ginneken B, Hulsbergen-van de Kaa C, Litjens G (2019) Epithelium segmentation using deep learning in h &e-stained prostate specimens with immunohistochemistry as reference standard. Sci Rep 9(1):1–10
19. Kumar N, Verma R, Anand D, Zhou Y, Onder OF, Tsougenis E, Chen H, Heng P-A, Li J, Hu Z et al (2019) A multi-organ nucleus segmentation challenge. IEEE Trans Med Imaging 39(5):1380–1391
20. Liu J, Xu B, Zheng C, Gong Y, Garibaldi J, Soria D, Green A, Ellis IO, Zou W, Qiu G (2018) An end-to-end deep learning histochemical scoring system for breast cancer TMA. IEEE Trans Med Imaging 38(2):617–628
21. Wei Y, Feng J, Liang X, Cheng MM, Zhao Y, Yan S (2017) Object region mining with adversarial erasing: A simple classification to semantic segmentation approach. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1568–1576
22. Lin D, Dai J, Jia J, He K, Sun J (2016) Scribblesup: Scribble-supervised convolutional networks for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3159–3167
23. Bearman A, Russakovsky O, Ferrari V, Fei-Fei L (2016) What's the point: semantic segmentation with point supervision. In: European conference on computer vision, pp 549–565. Springer
24. Dai J, He K, Sun J (2015) Boxsup: exploiting bounding boxes to supervise convolutional networks for semantic segmentation. In: Proceedings of the IEEE international conference on computer vision, pp 1635–1643
25. Cruz-Roa A, Basavanhally A, González F, Gilmore H, Feldman M, Ganesan S, Shih N, Tomaszewski J, Madabhushi A (2014) Automatic detection of invasive ductal carcinoma in whole slide images with convolutional neural networks. In: Medical imaging 2014: digital pathology, vol. 9041, p 904103. SPIE
26. Liu Y, Gadepalli K, Norouzi M, Dahl GE, Kohlberger T, Boyko A, Venugopalan S, Timofeev A, Nelson PQ, Corrado GS et al. (2017) Detecting cancer metastases on gigapixel pathology images. arXiv preprint arXiv:1703.02442
27. Simonyan K, Zisserman A (2014) Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv e-prints, 1409–1556 arXiv:1409.1556 [cs.CV]
28. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: 2016 IEEE Conference on computer vision and pattern recognition (CVPR), pp 770–778. https://doi.org/10.1109/CVPR.2016.90
29. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ (2017) Densely connected convolutional networks. In: 2017 IEEE Conference on computer vision and pattern recognition (CVPR), pp 2261–2269. https://doi.org/10.1109/CVPR.2017.243
30. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z (2016) Rethinking the inception architecture for computer vision. In: 2016 IEEE Conference on computer vision and pattern recognition (CVPR), pp 2818–2826. https://doi.org/10.1109/CVPR.2016.308
31. Ronneberger O, Fischer P, Brox T (2015) U-net: convolutional networks for biomedical image segmentation. In: Navab N, Hornegger J, Wells WM, Frangi AF (eds) Medical image computing and computer-assisted intervention—MICCAI 2015. Springer, Cham, pp 234–241
32. Noh H, Hong S, Han B (2015) Learning deconvolution network for semantic segmentation. In: Proceedings of the IEEE international conference on computer vision (ICCV)
33. 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. https://doi.org/10.1109/TPAMI.2016.2644615
34. Zhao H, Shi J, Qi X, Wang X, Jia J (2017) Pyramid scene parsing network. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)
35. 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

36. Oktay O, Schlemper J, Folgoc LL, Lee MJ, Heinrich M, Misawa K, Mori K, McDonagh SG, Hammerla N, Kainz B, Glocker B, Rueckert D (2018) Attention u-net: learning where to look for the pancreas. ArXiv **abs/1804.03999**
37. Peng C, Zhang X, Yu G, Luo G, Sun J (2017) Large kernel matters—improve semantic segmentation by global convolutional network. In: 2017 IEEE Conference on computer vision and pattern recognition (CVPR), pp 1743–1751
38. Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A (2016) Learning deep features for discriminative localization. In: 2016 IEEE Conference on computer vision and pattern recognition (CVPR), pp 2921–2929. https://doi.org/10.1109/CVPR.2016.319
39. Krähenbühl P, Koltun V (2012) Efficient inference in fully connected crfs with gaussian edge potentials. CoRR **abs/1210.5644**. arXiv:1210.5644
40. Chan L, Hosseini M, Rowsell C, Plataniotis K, Damaskinos S (2019) Histosegnet: semantic segmentation of histological tissue type in whole slide images. In: 2019 IEEE/CVF International conference on computer vision (ICCV), pp 10661–10670. https://doi.org/10.1109/ICCV.2019.01076
41. Hosseini, MS, Chan L, Tse G, Tang M, Deng J, Norouzi S, Rowsell C, Plataniotis KN, Damaskinos S (2019) Atlas of digital pathology: A generalized hierarchical histological tissue type-annotated database for deep learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 11747–11756
42. https://paip2019.grand-challenge.org/Dataset/
43. Choe J, Park JH, Shim H (2018) Improved techniques for weakly-supervised object localization. CoRR **abs/1802.07888**arXiv:1802.07888
44. Choe J, Shim H (2019) Attention-based dropout layer for weakly supervised object localization. In: 2019 IEEE/CVF Conference on computer vision and pattern recognition (CVPR), pp 2214–2223. https://doi.org/10.1109/CVPR.2019.00232
45. Durand T, Mordan T, Thome N, Cord M (2017) Wildcat: Weakly supervised learning of deep convnets for image classification, pointwise localization and segmentation. In: 2017 IEEE Conference on computer vision and pattern recognition (CVPR), pp 5957–5966. https://doi.org/10.1109/CVPR.2017.631
46. Hong S, Yeo D, Kwak S, Lee H, Han B (2017) Weakly supervised semantic segmentation using web-crawled videos. CoRR **abs/1701.00352**. arXiv:1701.00352
47. Kolesnikov A, Lampert CH (2016) Seed, expand and constrain: three principles for weakly-supervised image segmentation. In: Leibe B, Matas J, Sebe N, Welling M (eds) Computer vision—ECCV 2016. Springer, Cham, pp 695–711
48. Oh SJ, Benenson R, Khoreva A, Akata Z, Fritz M, Schiele B (2017) Exploiting saliency for object segmentation from image level labels. In: 2017 IEEE Conference on computer vision and pattern recognition (CVPR), pp 5038–5047. https://doi.org/10.1109/CVPR.2017.535
49. Chandra S, Kokkinos I (2016) Fast, exact and multi-scale inference for semantic image segmentation with deep gaussian crfs. In: Leibe B, Matas J, Sebe N, Welling M (eds) Computer vision—ECCV 2016. Springer, Cham, pp 402–418
50. Chen LC, Papandreou G, Kokkinos I, Murphy K, Yuille A (2016) Deeplab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE Transactions on pattern analysis and machine intelligence PP. https://doi.org/10.1109/TPAMI.2017.2699184
51. Fu H, Xu Y, Lin S, Kee Wong DW, Liu J (2016) Deepvessel: retinal vessel segmentation via deep learning and conditional random field. In: Ourselin S, Joskowicz L, Sabuncu MR, Unal G, Wells W (eds) Medical image computing and computer-assisted intervention—MICCAI 2016. Springer, Cham, pp 132–139
52. Feng Y, Hafiane A, Laurent H (2021) A deep learning based multiscale approach to segment the areas of interest in whole slide images. Comput Med Imaging Graph 90:101923. https://doi.org/10.1016/j.compmedimag.2021.101923
53. Xu G, Song Z, Sun Z, Ku C, Yang Z, Liu C, Wang S, Ma J, Xu W (2019) Camel: a weakly supervised learning framework for histopathology image segmentation. In: Proceedings of the IEEE/CVF international conference on computer vision (ICCV)
54. Ahn J, Kwak S (2018) Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)