MACD R-CNN: An Abnormal Cell Nucleus Detection Method

BAOYAN MA, JIAN ZHANG, FENG CAO, AND YONGJUN HE
Department of Computer Science, Harbin University of Science and Technology, Harbin 150080, China
Corresponding author: Yongjun He (heyongjun@hrbust.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61673142, and in part by the Natural Science Foundation of Hei Longjiang Province of China under Grant F2017013.

ABSTRACT The detection of abnormal cell nuclei is a key technique of the cytopathic automatic screening system, which directly determines the performance of the system. Although the Mask R-CNN which combines target detection and semantic segmentation has achieved good performance in general target detection tasks, the performance in abnormal cell detection is still unsatisfactory. To solve this problem, we design a new deep neural network for abnormal cell detection based on the Mask R-CNN, named mask abnormal cell detection R-CNN (MACD R-CNN). First, in the classification branch of Mask R-CNN, it generates the same size of feature maps from different size of RoIs as the input. The nuclei in this part of the feature maps will be deformed to varying degrees. We design a fixed proposal module to generate fixed-sized feature maps of nuclei, which allows the new information of nucleus is used for classification. Then we use the attention mechanism to merge the original RoI and Fixed RoI features. Finally, we increase the depth of the convolution layer to further improve the accuracy of cell classification. Experiments show that the MACD R-CNN can effectively improve the performance of abnormal cell detection.

INDEX TERMS Cervical cancer, deep learning, instance segmentation, attention mechanism.

I. INTRODUCTION
Cervical cancer is the second killer of women [1], which seriously threatens women’s lives. Early detection and early treatment are effective ways to deal with this problem. Current diagnosis of cervical cancer mainly relies on manual screening by doctors, that is, observing shape, color, and area of cervical cells by naked eyes to determine whether there are cancer cells. In a diagnosis, a doctor needs to search cancer cells from hundreds of thousands of cells under a microscope. The workload is large and the accuracy is low. In addition, this method can cause misdiagnosis and missed diagnosis inevitably, which brings huge losses to patients. In recent years, with the rapid development of artificial intelligence [2] technology, computer automatic screening has gradually become a practical application, which has provided strong support for the diagnosis of cervical cancer [3].

The detection of abnormal cells is a key technique of the automatic screening system, and it is the prerequisite for the computer to automatically accurately find abnormal cells. Traditional methods first use image segmentation methods such as the watershed [4], k-means [5], etc., to segment the nuclei in images, then extract features such as optical density [6], texture [7], and morphology [8] of the nuclei, and finally use KNN [9], SVM [10] and other methods for classification. These methods adopts features designed artificially and can achieve good performance in general target detection tasks. However, their performance is degraded substantially in abnormal cell detection, because there are a lot of unfavorable factors such as differences in illumination, uneven staining, overlapping cells, and trash impurities etc. In recent years, the methods based on deep learning [11], such as the target detection [12]–[15] and instance segmentation [16]–[18], can automatically extract features and locate cell nuclei, as well as achieving good performance in detecting abnormal cells.

Among these methods, the R-CNN [19] is the first one to use deep learning for object detection. The R-CNN follows the procedure of traditional object detection. That is, the category-independent region proposals are generated, then a fixed-length feature vector from each region is extracted via a convolutional neural network, and finally the linear SVMs are used to classify objects. The difference is that, it replaces traditional features (such as SIFT [20], HOG [21] features,
etc.) with features extracted by deep convolutional networks. In order to improve the efficiency of the R-CNN, the Fast R-CNN [22] uses the RoI Pooling, and unifies image classification and bounding box regression to improve the efficiency. In addition, it replaces the SVM with softmax, employs the multi-task loss function, which effectively improved the accuracy and efficiency of target detection. However, the extraction box in the Fast R-CNN is the same to that in the R-CNN, i.e., they both adopt the selective search algorithm which cannot run on GPUs. In contrast, the Faster R-CNN [23] provides a Region Proposal Network (RPN) to find regions of targets and can run the entire target detection network on the GPU to improve the efficiency. Based on the Faster R-CNN, Kaiming He et al proposed the Mask R-CNN by combining mask detection and semantic segmentation [24]. This network uses the ResNet-FPN [24] as backbone and adds a full convolution network to predict the edges of cells or nuclei. In addition, the Mask R-CNN replace RoI Pooling with RoI Align [24], making the segmentation results more accurate. The above mentioned methods are two-stage networks, which are more suitable for detecting small targets than one-stage networks.

The Faster R-CNN [23] has been successfully applied in cell detection tasks. Liu et al. [25] trained a Faster R-CNN model using a self-built Circulating Tumor Cell (CTC) dataset to detect tumor cells. In order to improve the detection efficiency, the detected image area is reduced, and the number of anchor boxes is increased. Qiu et al. [26] used the Faster R-CNN to detect red blood cells. This method first finds the cell location with the Faster R-CNN, and then uses the ResNet for classification. Kaushik et al. [27] adopted the Faster R-CNN and Yolo [28] to detect breast cancer cells and verified the effectiveness of target detection. Liang et al. [29] introduced Faster R-CNN to detect cervical cancer, and used a contrast detection training strategy to improve accuracy. Because of many categories and less training data, its accuracy rate is low to 26.3% and the recall rate is 35.7%, which cannot meet the needs of practical applications. Because the Mask R-CNN [24] has achieved great success in instance segmentation, it has been applied to medical image segmentation and semantic segmentation [24]. This network uses the ResNet-FPN [24] as backbone and adds a full convolution network to predict the edges of cells or nuclei. In addition, the Mask R-CNN replace RoI Pooling with RoI Align [24], making the segmentation results more accurate. The above mentioned methods are two-stage networks, which are more suitable for detecting small targets than one-stage networks.

There are not only various cells such as epithelium, lymph, and neutrophils but also various impurities and trash in cell images. Abnormal cells are formed by mutations in epithelial cells. If our task is to classify various types of cells and trash impurities, the RoI Align [24] in the Mask R-CNN can keep high classification accuracy. However, in the task of abnormal cell detection, the goal is to distinguish abnormal epithelial cells from other cells, including normal epithelial cells. On the one hand, abnormal epithelium has differences in shape and texture compared with normal epithelial cells. On the other hand, there is a significant difference in nuclear size, that is, an abnormal nuclear is usually twice as large as a normal epithelial cells [35]. The Mask R-CNN scales the original area into a fixed-size feature map by the RoI Align after generating the proposed boxes, which helps to classify various types of cells and trash impurities. However, this processing also changes the cell size and nuclear texture, which reduces the discrimination between abnormal and normal cells. In addition, the background around the nucleus is cytoplasm, and partial retention also helps to improve the differentiation between normal and abnormal epithelial cells. The human brain can selectively process the input image information by paying different degrees of attention to different image information. And human can learn where to focus on an image from previously observed images. Based on this principle, researchers introduced attention mechanisms into deep learning [36]. The attention mechanisms can be realized in the spatial domain [37], channel domain [38], layer domain [39], and mixed domain [40]. Inspired by [36] and [38], this paper introduces an attention mechanism, and proposes an abnormal cell detection network, which is called mask abnormal cell detection R-CNN (MACD R-CNN).

The main contributions of this work are as follows. Firstly, in order to preserve the scale of nuclei and local cytoplasm, we propose a fixed proposal module (FPM) to extract fixed-size RoIs which extract cervical cell feature maps with the size kept. Secondly, we applied feature map addition and attention mechanism to combine the feature maps generated by the original RoIs with the feature maps generated by the fixed size of RoIs. A module called SE-FPM is design based on the Squeeze and Extension (SE) [38]. This allows the classification branch of this network to not only notice the characteristics of the texture and optical density of the nucleus, but also learn morphological features such as the size of the nucleus. Finally we increase the convolution layer to improve the ability to extract nuclear features.

II. METHOD
A. MASK R-CNN

The Mask R-CNN is a two-stage network, of which structure is shown in Fig. 1. In the first stage the feature maps are generated by extracting image features through the backbone. The feature map is then input into a Region Proposal Network (RPN) to generate proposals. There are two branches in the RPN, one of which is used to predict the bounding box, and the other uses softmax to predict whether there are target objects in the box. In the second stage, each proposed feature map is first extracted through the RoI Align, and then the three branches are used for classification, bounding box regression, and mask prediction, respectively.
The classification and bounding box regression share the parameters of convolutional layers and fully connected layers.

**B. RoI ALIGN**

In the RoI Pooling [23] the rounding operation is used when transforming regions of interest from feature maps of different sizes into fixed small feature maps. The RoI Align improved the RoI Pooling by using floating point numbers instead of rounding operations, and bilinear interpolation in calculating the RoI results. In a common classification task, the captured objects are large or small on the image. After the size of the RoI is adjusted by the RoI Align, the detected objects are scaled to the same size, which can improve accuracy. Under the microscope, the cell images are captured in a fixed position, so the cell size is an important feature to determine whether a cell is abnormal. The RoI Align can improve the accuracy in the prediction of nuclear edges and the prediction of nuclear boxes, but the prediction results is not used for the cell classification branch.

**C. FIXED PROPOSAL MODULE**

Professional pathologists mark cells as abnormal ones by observing the size, morphology, color, granularity of staining, granularity of chromatin, and color and morphology around the nuclei in an image. The RoI Align can obtain RoI features
of different sizes, and then scales their width and height to a fixed size as the input of the next layer. The RoI Align method can accurately locate and segment nuclei, but the prediction results is not used for the cell classification branch. We propose the FPM to keep the same scale of feature maps and extract local features around the nuclei. The comparison of the original RoI and the fixed size of RoI are shown in Fig. 2. The small nucleus framed on the left is normal, and the large nucleus framed on the right is abnormal. In the input feature maps of classification branch, the difference between the normal and abnormal nuclei becomes smaller, so the normal nucleus looks like an abnormal one. After dealing with the FPM, normal nucleus and abnormal nucleus can be clearly distinguished. The size of the generated feature maps is a hyperparameter.

In the Mask R-CNN, the regional proposal network first outputs multiple proposal boxes. Each proposal box can be determined by a tuple \((r, c, h, w)\), where \((r, c)\) and \(h, w\) represent the upper left corner coordinate, height and width of the proposal box, respectively. In the FPM \((r, c, h, w)\) is converted to \((r', c', h', w')\) by:

\[
\begin{align*}
    r' &= r + \frac{h - h'}{2} \\
    c' &= c + \frac{w - w'}{2}
\end{align*}
\]

where \((r', c')\) is the coordinate of the upper left corner after conversion, and \(h'\) and \(w'\) are the width and height of a fixed size. Then the regional feature map is transformed into a fixed size feature map. \(h'\) and \(w'\) are also hyperparameters.

The MACD R-CNN does not change the mask prediction branch architecture of the Mask R-CNN and the effect of mask predictions. The implementation of the MACD R-CNN is described as follows (as shown in Fig. 3). First the ResNet-50-FPN is used as the backbone, and then the feature maps extracted by the proposed boxes are used for mask prediction, bounding box regression and category prediction. The FPM is used to generate the fixed size of RoIs. At last the features generated by the fixed size of RoIs and the original RoIs are combined and input to the classification branch. The merging method is described in the next section.

D. FEATURE MERGING

Both the fixed size of RoIs and the original RoIs have their own advantages. The original RoIs is good for classifying all kinds of cells and impurities, and the fixed size of RoIs is good for distinguishing normal cells from abnormal cells. In order to use the features obtained by the two parts, this article uses feature addition and attention mechanism to merge features. The process of adding convolutional layers after feature merging will be introduced in section F.

1) ADDITION METHOD

The common methods to merge the RoI features generated by the fixed size of RoIs and the original RoIs include connection, addition, and multiplication. The connection merging doubles the channel dimension of the feature map, which increases the dimension too much. However, the addition method does not increase the feature map channels. The
multiplication method will be used in the next section. The structure of the addition merging is shown in Fig. 4. The feature maps of $7 \times 7 \times 256$ are first generated by the fixed size of RoIs and original RoIs, respectively, and then added to form feature maps of $7 \times 7 \times 256$, which are input into two fully connected layers with a size of 1024. Finally, the softmax function is used for classification. In all hidden layers, the ReLU [41] is used as the activation function.

2) ATTENTION MECHANISM METHOD
The channel attention mechanism is used to merge these two RoI features. It can assign weights to both the fixed size of RoIs and the original RoIs, so that the network model automatically learns nuclear features with larger channel weights. The implementation of attention mechanism is shown in Fig. 5. Firstly, the RoI features generated by the original RoIs are input to the Squeeze and Extension [38] block, and then the output is multiplied with the RoI features generated by the fixed size of RoIs.

As shown in Fig. 2, the size of bounding box obtained by the original RoI is the same to that of the nuclear bounding box, and it is closer to the true size of each nucleus than the size of bounding box obtained by the fixed size of RoI. Therefore, the RoI features generated by the original RoIs are more suitable as input to attention. At the same time, the RoI features generated by the fixed size of RoIs can also learn morphological features such as the area and roundness of the nucleus at the same level.

The structure of the SE-FPM is shown in Fig. 5. It first adds a global pooling layer after the RoI Align. And then a fully connected layer with a ReLU activation function and a fully connected layer with Sigmoid activation function are added sequentially. Finally, the output of the sigmoid function is multiplied with the RoI features generated by the fixed size of RoIs. Where $k$ is the number of input channels and $l$ is a hyperparameter.

E. THE SIZE OF RoI
The RoI size used by the FPM is a key parameter. The size of the nucleus varies within a range. It is necessary to determine a proper RoI size, which can contain a complete nucleus and at the same time does not include irrelevant other objects. Although a RoI with a size big enough, such as $200 \times 200$, can cover a cell completely, it may also cover other nuclei, neutrophils and impurities. As shown in Fig. 6, the RoI in the left image is relatively small, and that in the right image is relatively large. They generate corresponding frames centered on normal cells and abnormal cells respectively. The RoIs in the right image contains both normal and abnormal cells. A cell has two labels in different RoIs, which is harmful for classification. In order to obtain a suitable RoI size, the RoI size $H \times W$ generated by the FPM is determined through experiments. Firstly, the RoI size is changed from $200 \times 200$ to $100 \times 100$ in steps of 50 and from $100 \times 100$ to $50 \times 50$ in steps of 10, and then the performance is compared. The size with the highest recall of abnormal cells will be chosen as the best RoI size for the FPM. Because the size of abnormal nuclei is generally larger than $50 \times 50$, so we set $50 \times 50$ as the minimum size.

F. ADDING CONVOLUTIONAL LAYERS
Adding convolution layers to the classification branch needs to separate it from the bounding box regression branch. The convolutional layers are added as shown in Fig. 7. The RoI feature size for the fixed size of RoIs is $28 \times 28 \times 256$. The convolution kernel size of the last convolution layer is $7 \times 7$, and the size of other convolution kernels are all $3 \times 3$. After the convolutional layers, Batch Normalization [42] and ReLU are added. The classification branch which uses the original RoIs to add the convolution layers is shown in Fig. 8. The convolutional layers are the same as that in Fig. 7. The bounding box regression branch and classification branch share this part of the parameters. The size of feature maps generated by RoI is also $28 \times 28 \times 64$.

On the classification branch, the convolutional layer is placed behind the fusion module. Two methods are used here. The first is to use the Add operation to merge and then add convolution layers. As shown in Fig. 9, the RoI features generated by the fixed size of RoIs and the RoI features generated by the original RoIs are first added, and then the result is input into a convolution layer. The convolution kernel size of the last convolution layer is $7 \times 7$, and the other convolution kernel sizes are all $3 \times 3$. The ReLU and batch normalization are used in hidden layers. The sizes of RoI features generated

![FIGURE 5. The structure of SE-FPM.](image-url)
FIGURE 6. RoI in different sizes.

FIGURE 7. Classification branch using the fixed size of RoIs to add convolution layers.

FIGURE 8. Classification and bounding box regression branches using the original RoI to add convolution layers.

FIGURE 9. Adding convolution layers after merging with addition method.

III. EXPERIMENTS AND RESULTS

We implement the MACD R-CNN based on the open source of the Mask R-CNN [43]. In our experiments, the proper RoI size of the FPM is determined first. Then the methods of adding convolution layer, different activation functions are evaluated. Finally, the comparison of the proposed methods and other methods are made.

A. DATA PREPARATION

Two datasets are used to evaluate the proposed method. The first one is the Herlev dataset [34]. This dataset consists of 917 images, and each image contains only one cell. There are seven classes: superficial squamous epithelial, intermediate squamous epithelial, mild squamous non-keratinizing dysplasia, moderate squamous non-keratinizing dysplasia, severe squamous non-keratinizing dysplasia, and squamous cell carcinoma in situ intermediate. We use two classes only, the Herlev dataset is divided into normal and abnormal cells: the first three classes as normal cells with 242 cells and the last four classes as abnormal cells with 675 cells. We validate the methods with 5-fold cross validation.

The second one is collected from Harbin Maria Obstetrics and Gynecology Hospital (called HMOG dataset). An automatic pathological scanner was used to acquire images. This equipment consists of an automatic optical microscope, an industrial camera, a controller. It can control the electric platform to automatically move, focus and control the camera to grab images. The size of images is 2048 × 2048. In this paper, the objects on the image are labeled by professional pathologists into two categories, i.e., abnormal or normal cells.

The size of the input images for the network model is 512 × 512, so the images of 2048 × 2048 for training and test are cut into images of 512 × 512. The amount of normal cervical cells is large while that of abnormal cervical cell is small. To make the data of the two classes balance, we split the data into two subsets. The first subset contains 8914 images, with no abnormality or a small
number of abnormal cells, are used to train the localization branch; the second contains 2953 images with more than one abnormal cell in each, are used to fine-tune the classifier branch. They are called Data A and Data B.

As is shown in Fig. 10, the abnormal cells are marked by red dotted rectangle. The abnormal cells are not enough, and the ratio of normal cells to abnormal cells is about 20:1. In order to make the data balance to some extent, we increased the label number of each abnormal cell by 20 times. The test set contains 261 images with abnormal cells.

B. TRAINING
The number of parameters in the network model is very large. However, the abnormal cells of the HMOG dataset are not enough. Therefore, we used the following strategies to train the localization branch:

1) Initialize model parameters using the ImageNet pre-trained model [44].
2) Train the model with Data A.
3) Fine-tune the model with Data B.

In step 2), the head is trained first, and then all layers are trained. In step 3), all layers are directly trained.

C. ABNORMAL CELL DETECTION
To increase the number of abnormal cells, the data are augmented [45] by rotation (90, 180, 270 degrees), left-to-right and up-and-down flipping, Gaussian blurring. In this way, one image can generate 6 extra images. The methods chosen for comparison are as follows:

1) Mask R-CNN: The original Mask R-CNN [24].
2) Mask R-CNN-FPM: Mask R-CNN replacing the original RoIs with the fixed size of RoIs.
3) MACD R-CNN-Add: MACD R-CNN using the Add method to merge the RoI features.
4) MACD R-CNN-SE-FPM: MACD R-CNN merging the original RoI features and the fixed size of RoI features with attention mechanism.

The MACD R-CNN and Mask R-CNN use the same loss function. In the MACD R-CNN, the mask prediction branch and the bounding box regression branch remain the same to that in the Mask R-CNN. The result of the MACD R-CNN-SE-FPM is shown in Figure 11, where abnormal cells are in the red box and normal cells are in the green box. The bounding box and mask of the normal nuclei and abnormal nuclei locate cells accurately. However, the masks of abnormal cells are relatively large, because the number of abnormal cells is small. The classification branch will be evaluated in depth in the next few sections.

D. CHOOSING ROI SIZE
The RoI size of FPM is an important parameter which directly determines system performance. The statistics of the HMOG training dataset show that the size of the abnormal nucleus is around 60 × 60. And we use 60 × 60 plus 10 × 10, which is 70 × 70 as the default RoI size. In order to find a reasonable size, we use the Mask R-CNN-FPM with the RoI size varied to compare the performance. The results are shown in Table 1. The used measures are average recall (AR), average precision (AP), mean average precision (mAP), and Macro F1 score. And “ab” represents abnormal, “nor” represents normal. The RoI sizes used in the experiments are 200 × 200, 150 × 150, 100 × 100, 90 × 90, 80 × 80, 70 × 70, 60 × 60, and 50 × 50, respectively.
The AR_{ab} varies with the size of RoI. When the size of RoI is 70 × 70, the AR_{ab} is up to 88.89%, and when the size of the RoI is 200 × 200, the AR_{ab} is 66.67%. In the experimental data, the sizes of the abnormal nucleus are between 60 × 60 and 70 × 70. When the RoI size is bigger than 70 × 70 or smaller than 60 × 60, the performance will degrade. Therefore, the size of 70 × 70 is the proper one which is used in the following experiments.

### E. ADDING CONVOLUTIONAL LAYERS

In this section, the method of adding convolutional layers is evaluated on HMOG dataset. The comparison methods include the Mask R-CNN, Mask R-CNN-FPM, MACD R-CNN-Add, MACD R-CNN-SE-FPM, and methods of adding convolution layers to these four models. The backbone is ResNet-50-FPN. The results of adding convolutional layers are shown in Table 2. The Original row indicates the results without adding a convolution layer, and the Add Conv row indicates the results of adding a convolution layer.

| Method               | RoI Size | AR_{ab} | AP_{ab} | AR_{IoU} | AP_{IoU} | mAP | Macro F1 score |
|----------------------|----------|---------|---------|----------|----------|-----|----------------|
| Mask R-CNN-FPM       | 200 × 200| 66.67   | 57.81   | 97.61    | 98.35    | 78.08| 79.95         |
| Mask R-CNN-FPM       | 150 × 150| 68.58   | 59.47   | 97.70    | 98.44    | 78.96| 80.88         |
| Mask R-CNN-FPM       | 100 × 100| 73.56   | 62.79   | 97.94    | 98.69    | 81.24| 83.32         |
| Mask R-CNN-FPM       | 90 × 90  | 77.39   | 67.11   | 98.13    | 98.88    | 82.99| 85.19         |
| Mask R-CNN-FPM       | 80 × 80  | 82.76   | 71.76   | 98.40    | 99.14    | 85.45| 87.82         |
| Mask R-CNN-FPM       | 70 × 70  | 88.89   | 77.08   | 98.70    | 99.45    | 88.26| 90.82         |
| Mask R-CNN-FPM       | 60 × 60  | 86.21   | 74.75   | 98.57    | 99.32    | 87.83| 89.52         |
| Mask R-CNN-FPM       | 50 × 50  | 81.23   | 70.43   | 98.32    | 99.07    | 84.75| 87.07         |

We can see that the AR_{ab} of the Mask R-CNN is 80.46%. The Mask R-CNN-FPM obtains an AR_{ab} of 88.89%, with an improvement of 8.43% over the Mask R-CNN. The AR_{ab} of these two methods after adding the convolution layer are 81.61% and 89.27%, respectively. The performance of the Mask R-CNN-FPM is much better than that of the Mask R-CNN. When the MACD R-CNN-Add without adding the convolutional layer, the AR_{ab} is 91.57%, indicating that the convolutional layer can make the network learn more features. The MACD R-CNN-SE-FPM obtains an AR_{ab} of 93.10%, which is better than the MACD R-CNN-Add, indicating that the attention mechanism in SE-FPM is effective. The results show that adding the convolution layers can improve the performance.

### F. SELECTING ACTIVATION FUNCTION

The activation function used in the Mask R-CNN is ReLU. It is reported that another activation function Mish [46] is
better than the ReLU. It uses a small number of negative values to make the neural network obtain a better gradient flow, instead of directly setting some features to zero. The ReLU randomly sets the nuclear features to zero, which is equivalent to directly discarding these features. Mish randomly adds negative weights to some nuclear features and makes feature weights excessively smoother. The Mish activation function can learn more information, so we replaced the ReLU with the Mish activation function, and did a comparative experiment to evaluate the impact of the Mish activation function on the network model on HMOG.

Table 3 shows the comparison of the methods with Mish and ReLU. The RoI size is set as 70 × 70. The Mask R-CNN, Mask R-CNN-FPM, MACD R-CNN-Add, and MACD R-CNN-SE-FPM are compared. The backbone is ResNet50-FPN. All methods have improved their performance after using the Mish activation function.

### G. COMPARED WITH OTHER MODELS

In this section, we compared the MACD R-CNN with other target detection methods, including the SSD [47] and YOLO-v3 [48], Faster R-CNN [29], Mask R-CNN+VGG-like [32], and the traditional method, including the KNN [9] and SVM [10]. We also compared the MACD R-CNN with the U-Net [49] and Dense U-Net [49]. And the measures for the segmentation task are the prediction, recall and Zijdenbos similarity index (ZSI). The hidden layer activation functions in the Mask R-CNN and MACD R-CNN are ReLU. The backbone used by the Mask R-CNN is ResNet-50 and ResNet-50-FPN.

Table 4 shows that the AR_{ab} of the MACD R-CNN is higher than those of the YOLOv3, SSD, Faster R-CNN, Mask R-CNN and Mask R-CNN-VGG-like. The Mask R-CNN and Faster R-CNN achieve a higher AR_{ab} than the YOLOv3 and SSD. The Faster R-CNN, Mask R-CNN and MACD R-CNN methods are two-phase networks, and the SSD and YOLOv3 methods are one-phase networks. It is seen that the two-stage networks perform better than one-stage networks in distinguishing abnormal cells from normal cells. Deep learning methods perform better than traditional KNN and SVM methods, mainly because the features extracted manually cannot fully express the complex and high similarity of nuclear images. Deep learning methods can automatically extract more discriminative features. The AR_{ab} of the MACD R-CNN is much higher than those of other methods. The reason is that it adds the FPM to learn morphological features such as the same level of nucleus size.

Table 5 shows that the ZSI_{nor} of U-Net, Dense U-Net and MACD R-CNN are relatively high, i.e., 85.73%, 89.58% and 91.48%, respectively. The ZSI_{ab} of U-Net, Dense U-Net and MACD R-CNN are relatively low, i.e., 57.48%, 61.32% and 63.06%, respectively. The ZSI_{ab} of these methods is below 65%. This is because abnormal cells account for a small proportion in the entire image.

### H. EXPERIMENTS AND RESULTS ON HERLEV

In this section, we compared the MACD R-CNN with other methods on Herlev dataset (binary classification). The statistics of the Herlev training dataset show that the biggest size of the abnormal nucleus is around 90 × 90. We use 90 × 90 plus 10 × 10, which is 100 × 100 as the default RoI size. The hidden layer activation functions in the MACD R-CNN are ReLU. The backbone used in the Mask R-CNN and MACD R-CNN is the ResNet-50-FPN. The measures for...
### TABLE 4. The classification results of different models on HMOG dataset.

| Method             | Backbone   | AR<sub>IoU</sub> | AP<sub>IoU</sub> | AR<sub>area</sub> | AP<sub>area</sub> | mAP | Macro F1 score |
|--------------------|------------|------------------|------------------|-------------------|------------------|-----|----------------|
| KNN                | -          | 72.80            | 63.12            | 97.91             | 98.65            | 80.89| 82.95          |
| SVM                | -          | 74.33            | 64.45            | 97.98             | 98.73            | 81.59| 83.70          |
| SSD-300            | ResNet-50  | 76.63            | 66.45            | 98.10             | 98.84            | 82.65| 84.82          |
| YOLOv3-416         | ResNet-50  | 78.16            | 67.77            | 98.17             | 98.92            | 83.35| 85.57          |
| YOLOv3-416         | Darknet-53 | 78.54            | 68.11            | 98.19             | 98.94            | 83.53| 85.76          |
| Mask R-CNN         | ResNet-50  | 79.69            | 69.10            | 98.25             | 98.99            | 84.05| 86.32          |
| Faster R-CNN       | ResNet-50  | 80.08            | 69.44            | 98.26             | 99.01            | 84.23| 86.51          |
| Mask R-CNN         | ResNet-50-FPN | 80.46       | 69.77            | 98.28             | 99.03            | 84.40| 86.69          |
| Mask R-CNN+VGG-like| ResNet-50  | 80.84            | 70.10            | 98.30             | 99.05            | 84.58| 86.88          |
| Mask R-CNN-FPM     | ResNet-50-FPN | 88.89        | 77.08            | 98.70             | 99.45            | 88.26| 90.82          |
| MACD R-CNN-Add     | ResNet-50-FPN | 91.57        | 79.40            | 98.83             | 99.58            | 89.49| 92.13          |
| MACD R-CNN-SE-FPM  | ResNet-50-FPN | 93.10        | 80.73            | 98.91             | 99.66            | 90.19| 92.88          |

### TABLE 5. The segmentation results of different models on HMOG dataset.

| Method             | ZSI<sub>IoU</sub> | P<sub>IoU</sub> | R<sub>area</sub> | ZSI<sub>area</sub> | P<sub>area</sub> | R<sub>area</sub> |
|--------------------|-------------------|-----------------|-----------------|-------------------|-----------------|-----------------|
| U-Net              | 57.48             | 56.91           | 63.74           | 85.73             | 84.27           | 92.41           |
| Dense U-Net        | 61.32             | 58.26           | 64.53           | 89.58             | 89.14           | 93.86           |
| Mask R-CNN         | 62.09             | 60.18           | 61.29           | 90.62             | 89.36           | 90.52           |
| Mask R-CNN-FPM     | 62.17             | 61.53           | 63.59           | 91.26             | 89.92           | 91.04           |
| MACD R-CNN-Add     | 62.84             | 61.72           | 64.21           | 91.34             | 90.03           | 91.26           |
| MACD R-CNN-SE-FPM  | 63.06             | 62.14           | 64.79           | 91.48             | 90.72           | 92.83           |

### TABLE 6. The classification results of different models on Herlev dataset for 2-class problem.

| Method               | Sensitivity | Specificity | Accuracy | F1 score |
|----------------------|-------------|-------------|----------|----------|
| KNN [9]              | 97.62       | 100.00      | 98.31    | NA       |
| SVM [10]             | 97.95       | 87.50       | 95.20    | 90.52    |
| Mask R-CNN+VGG-like [32] | 96.70     | 98.60       | 98.10    | 96.50    |
| DeepPap [2]         | 98.20       | 98.30       | 98.30    | 98.80    |
| Mask R-CNN-FPM      | 97.79       | 99.17       | 98.36    | 98.87    |
| MACD R-CNN-Add      | 98.52       | 99.24       | 98.59    | 98.93    |
| MACD R-CNN-SE-FPM   | 99.07       | 99.39       | 99.14    | 99.28    |

### TABLE 7. The segmentation results of different models on Herlev dataset for 2-class problem.

| Method               | ZSI | Prediction | Recall |
|----------------------|-----|------------|--------|
| U-Net [49]          | 87.90 | 85.70 | 94.10 |
| Dense U-Net [49]    | 91.00 | 89.30 | 95.60 |
| Mask R-CNN+VGG-like [32] | 91.00 | 92.00 | 91.00 |
| Mask R-CNN-FPM      | 91.83 | 92.16 | 91.39 |
| MACD R-CNN-Add      | 92.07 | 92.58 | 92.24 |
| MACD R-CNN-SE-FPM   | 92.14 | 92.63 | 94.51 |

### IV. DISCUSSION

In section III-E, III-F and III-G we have discussed the impact of several methods on the detection of abnormal cells with HMOG dataset. One of the most influential is the attention mechanism method. The attention mechanism method mainly imitates human beings’ process of observing and processing images. When looking for abnormal cells, human beings also pay attention to the location of the cells and then observe the local characteristics of the cells. And professional pathologists will compare its size, color, shape and other characteristics with other nuclei to verify whether it is an abnormal cell. The MACD R-CNN-SE-FPM proposed by us is to imitate this process, and we merge the size information of the cell nucleus into the model through the attention mode of the SE block. This makes its detection effect greatly improved. We can also merge the box prediction results and mask prediction results into the classification branch, which will be a new solution.
It can be seen from Table 4 that the classification of abnormal cells is a difficult task. The AR$_{ab}$ of SVM is 74.33%, the AR$_{ab}$ of Faster R-CNN is 80.08%. And the method we proposed can still achieve an AR$_{ab}$ of 93.10% and an AP$_{ab}$ of 80.73%. According to Table 5, the segmentation effect of MACD R-CNN can also achieve the effect of other methods, but the abnormal cell nuclei segmentation is less effective than normal cell nuclei. This is mainly due to the lack of abnormal cell data. It is seen from Table 6 that the F1 score of DeepPaP is 98.80%, and the F1 score of MACD R-CNN-SE-FPM is 99.28%. According to Table 7, the ZSI of Dense U-Net is 91.00%. The ZSI of MACD R-CNN-SE-FPM is 92.14%. The proposed method can also perform better on other dataset.

The model parameters of the KNN, SVM, Faster R-CNN, SSD, and Yolov3 are smaller and the inference time is less. The parameters of the Mask R-CNN are more than that of the Faster R-CNN, and the inference time is slightly longer. Since the MACD R-CNN introduces an attention mechanism, its inference time will be longer than that of other methods.

V. CONCLUSION

The morphology of cervical nucleus is important to distinguish normal cervical cells from abnormal cervical cells. Because the original RoI features are scaled from different sizes to fixed sizes in the classification branch of Mask R-CNN, and the results of nuclear edges prediction and nuclear boxes prediction are not used in the classification branch. Our proposed FPM maintains the same scale of cervical nucleus, which can effectively extract and detect the features of abnormal cervical cells. In addition, the attention mechanism is used to combine the original RoI features and the fixed size of RoI features, which enhances the discrimination in classification. The convolution layer can further improve the classification accuracy of the network model. Experiments show the MACD R-CNN method achieved a significant improvement over the Mask R-CNN.

REFERENCES

[1] F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, and A. Jemal, “Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries,” CA: A Cancer J. Clinicians, vol. 68, no. 6, pp. 394–424, Nov. 2018.

[2] L. Zhang, L. Lu, I. Nogues, R. M. Summers, S. Liu, and J. Yao, “DeepPap: Deep convolutional networks for cervical cell classification,” IEEE J. Biomed. Health Informat., vol. 21, no. 6, pp. 1633–1643, Nov. 2017.

[3] M. Wu, C. Yan, H. Liu, Q. Liu, and Y. Yin, “Automatic classification of cervical cancer from cytological images by using convolutional neural network,” Biosci. Rep., vol. 38, no. 6, Dec. 2018, Art. no. BSRR20181769.

[4] E. A. Mohammed, M. M. A. Mohamed, C. Naugler, and B. H. Far, “Chronic lymphocytic leukemia cell segmentation from microscopic blood images using watershed algorithm and optimal thresholding,” in Proc. 26th IEEE Can. Conf. Elect. Comput. Eng. (CCECE), May 2013, pp. 1–5.

[5] C. Zhang, X. Xiao, X. Li, Y.-J. Chen, W. Zhen, J. Chang, C. Zheng, and Z. Liu, “White blood cell segmentation by Color-Space-Based K-Means clustering,” Sensors, vol. 14, no. 9, pp. 16128–16147, Sep. 2014.

[6] M. E. Mattie, L. Staib, E. Stratmann, H. D. Tagare, J. Duncan, and P. L. Miller, “PathMaster: Content-based cell image retrieval using automated feature extraction,” J. Amer. Med. Inform. Assoc., vol. 7, no. 4, pp. 404–415, Jul. 2000.

[7] C. Fatichah, M. L. Tangel, M. R. Widianto, and F. Dong, “Parameter optimization of local fuzzy patterns based on fuzzy contrast measure for white blood cell texture feature extraction,” J. Adv. Comput. Intell. Inform., vol. 16, no. 3, pp. 412–419, May 2012.

[8] S. Chen, M. Zhao, G. Wu, C. Yao, and J. Zhang, “Recent advances in morphological cell image analysis,” Comput. Math. Methods Med., vol. 2012, pp. 1–10, 2012.

[9] P. R. Paul, M. K. Bhownik, and D. Bhattacharjee, “Automated cervical cancer detection using Pap smear images,” in Soft Computing for Problem Solving, New Delhi, India: Springer, 2015, pp. 267–278.

[10] K. Bora, M. Chowdhury, L. B. Mahanta, M. K. Kundu, and A. K. Das, “Automated classification of pap smear images to detect cervical dysplasia,” Comput. Methods Programs Biomed., vol. 138, pp. 31–47, Jan. 2017.

[11] G. Litiens, T. Kooi, B. E. Bejnordi, A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. W. M. van der Laak, B. van Ginneken, and C. I. Sánchez, “A survey on deep learning in medical image analysis,” Med. Image Anal., vol. 42, pp. 60–88, Dec. 2017.

[12] I. Sevo and A. Avramovic, “Convolutional neural network based automatic object detection on aerial images,” IEEE Geosci. Remote Sens. Lett., vol. 13, no. 5, pp. 740–744, May 2016.

[13] J. Hung and A. Carpenter, “Applying faster R-CNN for object detection on malaria images,” Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR) Workshops, Jul. 2017, pp. 566–561.

[14] W. Xie, J. A. Noble, and A. Zisserman, “Microscopy cell counting and detection with fully convolutional regression networks,” Comput. Methods Biomech. Biomed. Eng., Imag. Vis., vol. 6, no. 3, pp. 283–292, May 2018.

[15] Y. Xue, G. Bigras, J. Hugh, and N. Ray, “Training convolutional neural networks and compressed sensing End-to-End for microscopy cell detection,” IEEE Trans. Med. Imag., vol. 38, no. 11, pp. 2632–2641, Nov. 2019.

[16] F. A. Guerrero-Pena, P. D. Marrero Fernandez, T. Ing Ren, M. Yui, E. Rothenberg, and A. Cunha, “Multiclass weighted loss for instance segmentation of clustered cells,” in Proc. 25th IEEE Int. Conf. Image Process. (ICIP), Oct. 2018, pp. 2451–2455.

[17] Y. Yi, P. Wu, D. J. Hoeppner, and D. Metaxas, “Pixel-wise neural cell instance segmentation,” in Proc. IEEE 15th Int. Symp. Biomed. Imag. (ISBI), Apr. 2018, pp. 373–377.

[18] V. Kulikov and V. Lempitsky, “Instance segmentation of biological images using harmonic embeddings,” 2019, arXiv:1904.05257. [Online]. Available: http://arxiv.org/abs/1904.05257.

[19] R. Girshick, J. Donahue, T. Darrell, and J. Malik, “Rich feature hierarchies for accurate object detection and semantic segmentation,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2014, pp. 580–587.

[20] D. G. Lowe, “Distinctive image features from scale-invariant keypoints,” Int. J. Comput. Vis., vol. 60, no. 2, pp. 91–110, Nov. 2004.

[21] N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection,” in Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR), vol. 1, Jun. 2005, pp. 886–893.

[22] R. Girshick, “Fast R-CNN,” in Proc. IEEE Int. Conf. Comput. Vis., 2015, pp. 1440–1448.

[23] S. Ren, K. He, and R. Girshick, “Faster R-CNN: Towards real-time object detection with region proposal networks,” in Proc. Adv. Neural Inf. Process. Syst., 2015, pp. 91–99.

[24] K. He, G. Gkioxari, and P. Dollár, “Mask R-CNN,” in Proc. IEEE Int. Conf. Comput. Vis., Oct. 2017, pp. 2961–2969.

[25] Y. Liu, H. Si, Y. Chen, and Y. Chen, “Faster R-CNN based robust circulating tumor cells detection with improved sensitivity,” in Proc. 2nd Int. Conf. Big Data Technol. (ICBDT), 2019, pp. 252–256.

[26] W. Qiu, J. Guo, X. Li, M. Xu, M. Zhang, N. Guo, and Q. Li, “Multi-label detection and classification of red blood cells in microscopic images,” 2019, arXiv:1910.02672. [Online]. Available: http://arxiv.org/abs/1910.02672.

[27] S. Kaushik, S. V. Raghavan, and B. Sivaselvan, “A study of deep learning methods for mitotic cell detection towards breast cancer diagnosis,” in Proc. Int. Conf. Adv. Comput. Data Sci. Singapore: Springer, 2019, pp. 254–263.

[28] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You only look once: Unified, real-time object detection,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 779–788.

[29] Y. Liang, Z. Tang, M. Yan, J. Chen, Q. Liu, and Y. Xiang, “Comparison-based convolutional neural networks for cervical Cell/Clumps detection in the limited data scenario,” 2018, arXiv:1810.05952. [Online]. Available: http://arxiv.org/abs/1810.05952.
[30] J. W. Johnson, “Adapting mask-RCNN for automatic nucleus segmentation,” 2018, arXiv:1805.00500. [Online]. Available: http://arxiv.org/abs/1805.00500

[31] X. Xie, Y. Li, and M. Zhang, “Robust segmentation of nucleus in histopathology images via mask R-CNN,” in Proc. Int. MICCAI Brainlesion Workshop, Cham, Switzerland: Springer, 2018, pp. 428–436.

[32] Kurniaansih, K. H. S. Allehaibi, L. E. Nugroho, Widyawan, and T. Mantoro, “Segmentation and classification of cervical cells using deep learning,” IEEE Access, vol. 7, pp. 116925–116941, 2019.

[33] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” 2014, arXiv:1409.1556. [Online]. Available: http://arxiv.org/abs/1409.1556

[34] J. Jantzen, J. Norup, and G. Dounias, “Pap-smear benchmark data for pattern classification,” in Proc. Nature Inspired Smart Inf. Syst. (NiSIS), Albufeira, Portugal, 2005, pp. 1–9.

[35] E. S. Cibas, T. J. Browne, and B. M. H. Mantel, “Enlarged squamous cell nuclei in cervical cytologic specimens from perimenopausal women (‘PM cells’) A cause of ASC overdiagnosis,” Amer. J. Clin. Pathol., vol. 124, no. 1, pp. 58–61, 2005.

[36] V. Mnih, N. Heess, and A. Graves, “Recurrent models of visual attention,” in Proc. Adv. Neural Inf. Process. Syst., 2014, pp. 2204–2212.

[37] J. Hu, L. Shen, and G. Sun, “Squeeze-and-Excitation networks,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 7132–7141.

[38] T. Xiao, Y. Xu, K. Yang, J. Zhang, Y. Peng, and Z. Zhang, “The application of two-level attention models in deep convolutional neural network for fine-grained image classification,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2015, pp. 842–850.

[39] H. Li, P. Xiong, J. An, and L. Wang, “Pyramid attention network for semantic segmentation,” 2018, arXiv:1805.10180. [Online]. Available: http://arxiv.org/abs/1805.10180

[40] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang, “Residual attention network for image classification,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 3156–3164.

[41] V. Nair and G. E. Hinton, “Rectified linear units improve restricted Boltzmann machines,” in Proc. 27th Int. Conf. Mach. Learn. (ICML), 2010, pp. 1–8.

[42] J. Lei Ba, J. Ryan Kiros, and G. E. Hinton, “Layer normalization,” 2016, arXiv:1607.06450. [Online]. Available: http://arxiv.org/abs/1607.06450

[43] W. Abdulla. (2017). “Robust segmentation of nucleus in histopathology images via mask R-CNN,” in Proc. Int. MICCAI Brainlesion Workshop, Cham, Switzerland: Springer, 2018, pp. 428–436.

[44] V. Nair and G. E. Hinton, “Rectified linear units improve restricted Boltzmann machines,” in Proc. 27th Int. Conf. Mach. Learn. (ICML), 2010, pp. 1–8.

[45] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang, “Residual attention network for image classification,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 3156–3164.

[46] V. Mnih, N. Heess, and A. Graves, “Recurrent models of visual attention,” in Proc. Adv. Neural Inf. Process. Syst., 2014, pp. 2204–2212.

[47] J. Hu, L. Shen, and G. Sun, “Squeeze-and-Excitation networks,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 7132–7141.

[48] T. Xiao, Y. Xu, K. Yang, J. Zhang, Y. Peng, and Z. Zhang, “The application of two-level attention models in deep convolutional neural network for fine-grained image classification,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2015, pp. 842–850.

[49] H. Li, P. Xiong, J. An, and L. Wang, “Pyramid attention network for semantic segmentation,” 2018, arXiv:1805.10180. [Online]. Available: http://arxiv.org/abs/1805.10180

[50] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang, “Residual attention network for image classification,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 3156–3164.

[51] V. Nair and G. E. Hinton, “Rectified linear units improve restricted Boltzmann machines,” in Proc. 27th Int. Conf. Mach. Learn. (ICML), 2010, pp. 1–8.

[52] J. Lei Ba, J. Ryan Kiros, and G. E. Hinton, “Layer normalization,” 2016, arXiv:1607.06450. [Online]. Available: http://arxiv.org/abs/1607.06450

[53] W. Abdulla. (2017). “Robust segmentation of nucleus in histopathology images via mask R-CNN,” in Proc. Int. MICCAI Brainlesion Workshop, Cham, Switzerland: Springer, 2018, pp. 428–436.