Mathematical Biosciences and Engineering

MBE, 20 (2): 2364–2381.
DOI: 10.3934/mbe.2023111
Received: 26 September 2022
Revised: 31 October 2022
Accepted: 09 November 2022
Published: 18 November 2022

http://www.aimspress.com/journal/MBE

Research article

Cervical cell extraction network based on optimized yolo

Nengkai Wu, Dongyao Jia*, Chuanwang Zhang and Ziqi Li

Beijing Jiaotong University, School of Electronics and Information Engineering, No. 3 Shangyuancun Haidian District, Beijing, China, 100044

* **Correspondence:** Email: dongyaojia1974@163.com.

Abstract: Early screening for cervical cancer is a common form of cancer prevention. In the microscopic images of cervical cells, the number of abnormal cells is small, and some abnormal cells are heavily stacked. How to solve the segmentation of highly overlapping cells and realize the identification of single cells from overlapping cells is still a heavy task. Therefore, this paper proposes an object detection algorithm of Cell_yolo to effectively and accurately segment overlapping cells. Cell_yolo adopts a simplified network structure and improves the maximum pooling operation, so that the information of the image is preserved to the greatest extent during the model pooling process. Aiming at the characteristics of many overlapping cells in cervical cell images, a non-maximum suppression method of center distance is proposed to prevent the overlapping cell detection frame from being deleted by mistake. At the same time, the loss function is improved and the focus loss function is added to alleviate the imbalance of positive and negative samples in the training process. Experiments are conducted on a private dataset (BJTUCELL). Experiments have verified that the Cell_yolo model has the advantages of low computational complexity and high detection accuracy, and it is superior to common network models such as YOLOv4 and Faster_RCNN.

Keywords: cervical cells; yolo; overlapping cell; deep learning; object detection

1. Introduction

As one of the diseases that seriously endanger women's health, cervical cancer has become the second most deadly malignant tumor. According to the World Cancer Research Fund International, there were around 14.1 million cases of cancer in 2012. The number of cancer cases is expected to reach about 24 million by 2035 [1]. Especially in developing countries, the incidence of cervical cancer is higher, and the age of onset is earlier. In response to the growing situation, in 2020, the World Health Organization launched a major initiative to eliminate cervical cancer worldwide [2]. However, the HPV vaccination rate of women in the world is currently low, so the early screening and diagnosis of cervical cancer can reduce the incidence of cervical cancer and play a crucial role in the prevention of the disease. Machine vision assisted detection has become a key technology. The main steps are image segmentation, feature extraction and selection and image classification. Among them, good image segmentation can be used to extract cell structure information (shape, color and number) [3,4], which is also the most critical step to complete cell detection.

The segmentation of cervical cell images is a major difficulty in the medical field. Overlapping cells, impurity interference and complex backgrounds in the images all make segmentation difficult. Currently, cell segmentation in the field of medical images is mainly divided into two categories: segmentation algorithms based on image processing and deep learning methods. When using traditional classification algorithms, it is usually necessary to apply various features such as texture, shape and color of cervical cells. Harandi N.M. et al. [5–7] took the shape and color of cytoplasm and nucleus as the basis for the segmentation of cervical cancer cells. The image segmentation and classification of simple cervical cancer cells were acheieved. Feature selection has become a key factor affecting the detection accuracy. Chankong T. et al. [8] tested the Herlev and LCH datasets with a classifier with 9 cell features, and the classification accuracy was more than 93%. In order to obtain various features, Lee H. et al. [9,10] tried to extract the edge features of cells in different ways. The segmentation of overlapping cells has been realized preliminarily. Jung C. et al. [11] proposed an unsupervised Bayesian classification method for separating overlapping nuclei. Combined with the prior knowledge about the regular shape of cluster nuclei, the overlapping cells are segmented. Diniz, D. N. [12] et al. used eight traditional classification algorithms such as decision tree, random forest and K-NN and successfully realized the classification of Herlev and CRIC datasets. However, the performance of each classifier is different in the testing process, and a single classifier is often unable to handle complex datasets. Problems such as low detection accuracy and poor sensitivity also appeared in the testing process, and the model's migration ability was also poor. At the same time, it is difficult for traditional image processing methods to find a suitable set of parameters such that these segmentation methods can simultaneously segment cell images containing multiple complex situations, and they can only provide accurate segmentation for images with certain specific cell patterns.

Relatively speaking, detection algorithms based on deep learning have better feature extraction capabilities, and they can quickly and effectively extract the target object in microscopic images. Long J. et al. [13] designed a complete convolutional network to achieve efficient and accurate segmentation. Badrinarayanan V. et al. [14] improved the deep full convolutional neural network structure. On the premise of achieving good segmentation effect, the computational memory and precision are balanced. With the development and improvement of deep learning technology, it has been applied more widely in the medical field. More and more people try to use deep learning algorithm to segment medical images and achieve good segmentation effect [15–17]. Ronneberger O. et al. [18] proposed a network and training strategy that achieved good results in the 2015 ISBI cell Tracking Challenge. At the same time, the pathological cells in microscopic images can be extracted quickly and effectively. Although models such as Convolutional Neural Network (CNNs) and Fully Convolutional Networks (FCNs) have good segmentation effects, they are not ideal when dealing with highly overlapping targets. The reason is that both CNNs and FCNs are discriminative models, both based on pixel classification to achieve image segmentation. At the same time, deep learning methods inevitably suffer from insufficient training data and network overfitting. For example, for Hep-2 segmentation detection, Li Y and Shen L [19] pre-trained their network on a large data set, I3A. However, due to the problem of overfitting, the performance of MIVIA on a small sample data set was reduced.

In response to the above problems, we have studied and improved the YOLOv4 model [20,21]. Compared with other detection models, YOLOv4 has better performance in both detection accuracy and detection speed. The existing YOLO series researchers have greatly improved its network structure, but the current method is still difficult to use to accurately identify single cells from highly overlapping cells. In order to obtain a target detection model that meets this requirement, this study adopts an improved YOLOv4 model, namely, the Cell_Yolo model. In order to build an efficient Cell_Yolo detection network, this paper makes the following contributions:

1. Improve the preliminary image feature extraction of the Cell_Yolo network based on the YOLOv4 network structure, and use multi-volume structures to enhance the feature extraction capability of the network. The design of the residual network is used to slow down the problem that the gradient disappears with the deepening of the network. The two valid feature layers obtained in the backbone extraction network were fused using the FPN structure.

2. To perform cluster analysis of the target frame of the cell data set by improved non-supervised clustering algorithms, set reasonable proven frame size and quantity.

3. On the basis of the original network loss function, add the Focus_pooling operation. This avoids the loss of local information in the extraction characteristics, and the boundary information of the overlapping cells remains.

4. Center distance NMS is used to process the model. This improves the detection accuracy of overlapping cells.

The rest of the paper is structured as follows. The second part introduces the improved network model. The third part introduces the details of the experimental design. The fourth part is the summary.

2. Network model optimization

2.1. YOLO network structure and improvement

As mentioned above, the Two-Stage object detection algorithm can obtain candidate regions in advance, and it can fully learn the characteristics of the target, with high detection and positioning accuracy. However, this algorithm has complex network structure, a large amount of computation and slow detection speed, so it is not suitable for high real-time application scenarios.

The One-Stage object detection algorithm is simple in structure, can process the input image directly, has high detection accuracy and fast detection speed and can realize real-time detection. This algorithm can meet some real-time on-line detection application scenarios, such as real-time surface defect detection, real-time fire detection, real-time aerial operation detection and so on. However, the One-Stage algorithm has low detection accuracy for small targets and multi-target objects. Especially in complex scenes, the detection accuracy cannot meet the requirements. Some people put forward the concept of multi-scale fusion, which has achieved better results [22]. Since this scenario uses the object detection algorithm to do two-class cell recognition (that is, the network model only needs to distinguish the cell from the background, does not need to classify the cell type specifically and needs the algorithm to detect as fast as possible), it draws on the well-known

YOLOv4 [23] network. The YOLOv4 network is one of the most outstanding network models in target identification tasks. Using the YOLOv4 network for cell classification, YOLOv4 has obvious advantages in speed and accuracy. The basic structure of YOLOv4 is shown in Figure 1, in which the box numbers represent the characteristic image size. The main feature extraction layer uses CSPDarknet53, with SPP added to the middle for multi-pooling, which processes the feature map of the upper output using four different scales of maximum pooling. Maximum pooled core sizes are 13*13, 9*9, 5*5 and 1*1 (1*1 is unprocessed), which greatly increases the field of perception, isolates the most significant contextual features and hardly increases the run time of YOLOv4. The middle structure of the network is a simple two-way feature fusion PANet structure [24]. An important means in the object detection algorithm is to improve the FPN (feature pyramid). PANet is the first model to propose the second from bottom to top fusion. PANet is based on the FPN of Faster RCNN and simply adds the fusion path from bottom to top. A single feature map cannot effectively represent objects of different scales. Using multi-feature charts to represent objects of different scales can significantly improve the performance of object detection, especially for small targets, by fusing high-level and low-level features. Multiscale prediction is used in YOLOv4, which improves the detection results for different scales.

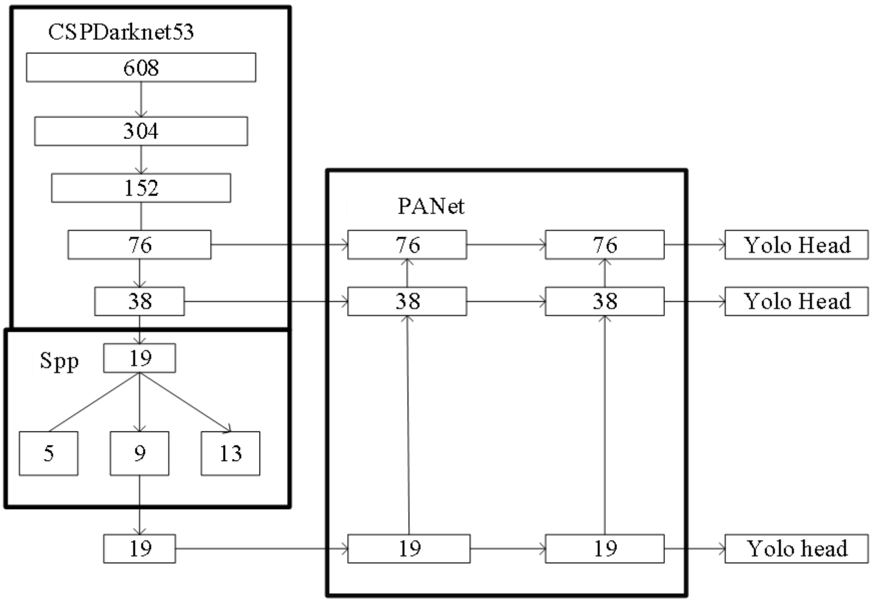


Figure 1. The structure of YOLOv4.

New activation function Mish function is used in feature extraction of YOLOv4, as shown in (1).

$$Mish = x * tanh\big(ln(1 + e^{x})\big) \tag{1}$$

Its function image is shown in Figure 2.

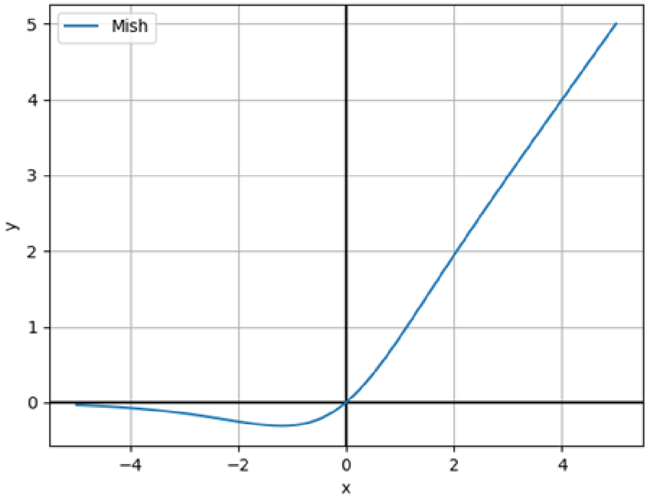


Figure 2. Mish function image.

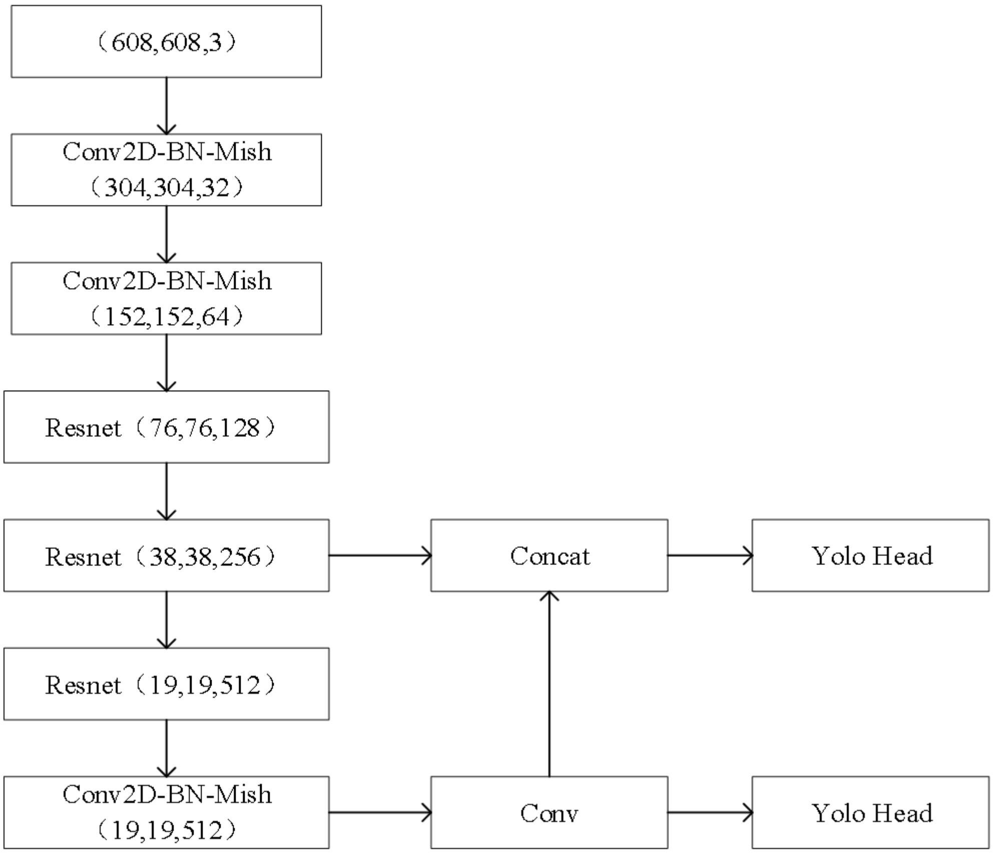


Figure 3. The structure of Cell_yolo.

From the function image, it can be seen that the Mish function has a lower bound and no upper bound, which can effectively avoid the gradient descent to speed up the training process, and the attributes with the lower bound can help to achieve strong regularization effect. The Mish function is not completely truncated when the value is negative, allowing a smaller negative gradient to flow in, ensuring the flow of information. At the same time, the Mish function has infinite continuity and smoothness, has strong generalization ability and optimization ability and can improve the quality of training results to a certain extent.

This paper presents a target recognition network Cell_Yolo that optimizes the network structure while maintaining the detection accuracy. The structure of Cell_Yolo is shown in Figure 3. The backbone of Cell_Yolo extracts the YOLOv4 design from the network and simplifies and optimizes its structure, as shown in Figure 3. Conv2D-BN-Mish represents the structure blocks of convolution, batch standardization, Mish activation function, and Resnet represents the structure blocks of residual network. Preliminary image feature extraction using multiconvolution structure enhances the feature extraction capability of the network, while the design of a residual network alleviates the problem of gradient disappearance as the network deepens. In the middle of Cell_Yolo, the FPN structure fuses the two valid feature layers obtained from the trunk extraction network from the bottom to the top. The FPN convolutes the valid feature layers of the last scale, then samples them up, stacks and convolutes them with the valid feature layers of the previous scale to fuse the high-level feature information with the low-level feature information.

In the final prediction section, the original YOLOv4 network has three detection heads, which are designed to predict targets of different scales. Large objects are predicted using small-scale feature layers, while small objects are less predicted using large feature layers. A large number of cervical cancer cell datasets have been observed, the cell size differences are not large, and the small targets in the background are impurities and other types of unrelated cells. The design reduces the output of a detection head, focusing on accurate identification of the location of cervical cells, while reducing the focus on unrelated cells and impurities. Cell_Yolo only sets two feature detection heads, each responsible for detecting targets of different scales. Three different prior boxes are preset in the feature map of each detection head to generate prediction boxes for predicting target objects. Cell_Yolo sets six prior boxes for both detection heads. In the process of network training, the model parameters need to be changed iteratively, and the information of the preset box should be changed according to the real-time model parameters, so that the preset box fits the target box as much as possible. Whether the preset box is selected correctly or not is directly related to the convergence speed of the network in the training process, but it also has an impact on the final detection accuracy. In this paper, an improved K-means clustering algorithm is used to cluster tag data to determine the optimal size of the preset box. The original K-means algorithm flow is described as follows:

Step 1: Randomly select K sample points as initial cluster centers in a given dataset.

Step 2: Calculate the distance between each sample and K cluster centers in the dataset and assign the sample points to the nearest cluster centers so that each cluster center can form a cluster.

Step 3: For each cluster, the centroid of all samples in the cluster is calculated and selected as the new cluster center.

Step 4: Repeat step 2 and step 3 until the cluster center location does not change.

Obviously, the selection of initial cluster centers in the algorithm is random. Random cluster centers can cause a lot of time waste and classification error. To overcome these shortcomings, the improved K-means algorithm is described as follows:

Step 1: Randomly select a sample point in a given dataset as the initial cluster center;

Step 2: Calculate the distance between each sample point and the current nearest cluster center, expressed as $D(x)$, and then calculate the probability that each sample will become the next cluster center with (2). The next cluster center is determined by the probability value.

$$P(x) = \frac{D(x)^2}{\sum_{x \in X} D(x)^2} \tag{2}$$

Step 3: Repeat step 2 until K cluster centers are selected, and repeat steps 2 to 4 of the K-means algorithm.

The improved K-means algorithm mainly improves the selection of the initial cluster centers. From the algorithm steps, the improved K-means ensures that the initial cluster centers are not too close and that the distances between cluster centers are as far as possible, thus ensuring the final cluster results are more accurate.

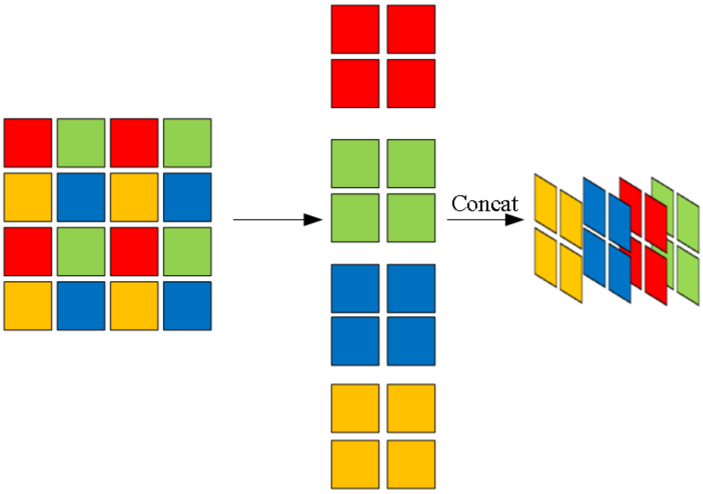


Figure 4. The process of Focus_pooling.

For the pooling operation in the network, the Focus_pooling operation is improved in Cell_Yolo. Although the original Maxpooling maximum pooling operation can reduce the size of the feature map to extract features, it also means that local information will be lost. The boundary characteristics of overlapping cells are extremely important, and excessive use of pooling operation will adversely affect the final result, so the Focus_pooling operation is added to the design. The operation flow chart of Focus_pooling is shown in Figure 4. For input pictures, take a value every other pixel value, so that the original picture becomes four pictures. Four pictures contain all the information of the original picture, that is, the W and H dimensions of the original picture are reduced by two times, but the channel number dimension is expanded by four times, using concat stitching. The resulting pictures are quadrupled compared to the original number of picture channels, and then the resulting pictures are convoluted, resulting in a double down-sampling feature map with no loss of information. The original input picture is 640*640*3, which first becomes four 320*320*3 slices, then uses concat stitching to form a 12*320*320 feature map, and then convolutes to 32*320*320. Focus_pooling collects w, h information on the channel without loss of information when sampling images, and then uses convolution to extract features, which makes feature extraction more efficient. In this structure, three operations using Focus_pooling are designed to replace Maxpooling. First, Focus_pooling is performed once when the picture enters the backbone extraction network, and then Focus_pooling is added before two effective feature layers are obtained for feature fusion. The addition of Focus_pooling can stabilize the transmission of feature information and improve the feature extraction ability of network model.

2.2. Center distance NMS

Non maximum suppression (NMS) is generally used to exclude non maximum elements. It is mainly used to solve the problem of repeated detection of the same target in object detection. Generally speaking, when the analytical model is output to the target box, there will be many target boxes, in which many duplicate boxes are located to the same target. NMS is used to remove these duplicate boxes and obtain the real target box. Yolo sets NMS to solve the problem of repeated detection of a target. It relies on the classifier to obtain multiple candidate boxes and the probability values of the categories in the candidate boxes. The flow of the algorithm is as follows:

Step 1: Sort the scores of all boxes and select the box with the highest score.

Step 2: Traverse all the remaining boxes, and calculate the IOU of each box and the box with the highest score. If this IOU is greater than a certain threshold, it will be considered as duplicate detection, and the current box will be deleted.

Step 3: Continue to select a box with the highest score. After step 2, the box with the highest score is usually the label box of another object. After selection, repeat the above process.

There are several disadvantages in traditional NMS. The first is that the selection of threshold greatly depends on experience. Whether the threshold is appropriate or not has a great impact on the box screening. In practice, it is difficult to select an appropriate threshold to ensure better accuracy and recall. Secondly, the NMS method is too rough. If the IOU of the current box and the box with the highest score is greater than the threshold, it will be deleted directly. However, in practice, two objects may be very close, resulting in too large detection frame IOU, and a target frame will be deleted by mistake. This situation is particularly prominent in images with more overlapping cells, as shown in Figure 5.

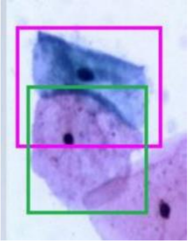


Figure 5. NMS in overlapping cells.

In Figure 5, the two frames belong to two targets, but their IOU is large. One may be deleted by mistake after NMS algorithm. Due to the complexity of cervical cell image, there are many duplicate cells and overlapping cells and impurities in the image, so the traditional NMS is not suitable for the application scenario of this project.

In fact, the main reason why NMS mistakenly deletes boxes of other categories is that NMS only considers IOU. NMS believes that a too large IOU means that an object is repeatedly detected by multiple frames, but this situation can be effectively alleviated when considering the center distance of the two frames. Therefore, on the basis of the NMS algorithm, the center distance index of the box is added to judge whether the current box is deleted, as shown in (3) and (4).

$$s_i = \begin{cases} s_i, IoU - R_{DIoU}(M, B_i) < \varepsilon \\ 0, IoU - R_{DIoU}(M, B_i) \geq \varepsilon \end{cases} \tag{3}$$

$$R_{DIoU} = \frac{\rho^2(b, b^{gt})}{c^2} \tag{4}$$

Where R_{DIoU} is the distance between the center points of two boxes, b^{gt} represents the box with the highest score, b represents the currently selected box, ρ^2 represents the distance, and c represents the diagonal length of the smallest box containing two boxes. The biggest difference between the center distance NMS and NMS is that the center distance NMS believes that the boxes with two distant center points may be located on different objects and should not be deleted. The problem of box deletion of overlapping cells can be effectively alleviated through the center distance NMS.

2.3. Definition and introduction of loss function

Generally, IoU will be used to measure the gap between the prediction box and the real box as (5).

$$IoU = \frac{A \cap B}{A \cup B} \tag{5}$$

where A is the area of the prediction box area, and B is the area of the label box area. The IoU of A and B is the intersection of the area of the prediction box area and the area of the real box area divided by its union.

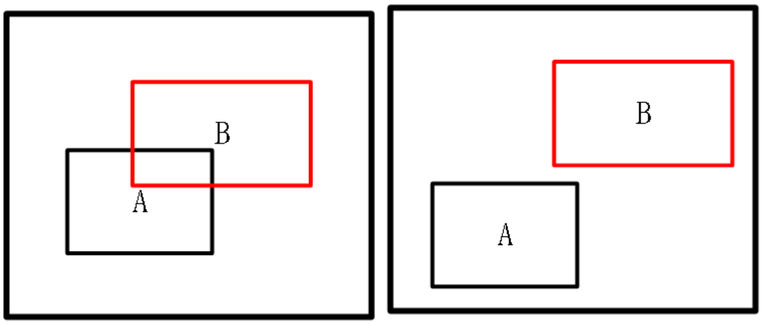


Figure 6. Calculation of IoU.

As can be seen from (5), it is clear that IoU measures the overlap rate between the prediction box and the real box. For two objects with the same IoU, their alignment cannot be represented. If the prediction box does not overlap the real box, the IoU is always zero, and as shown in the right figure in Figure 6, the regression loss optimization of the border is not equivalent to the optimization of the IoU. Therefore, IoU as a loss function can result in a severe deviation of the prediction border.

This paper uses $DIoU$ as the loss function of the border. $DIoU$ can directly minimize the distance between two targets, and $DIoU$ adds a penalty on top of IoU to measure the center point distance between the prediction box and the real box. In the process of minimizing the distance between the center point of the bounding box, the bounding box converges faster, and $DIoU$ can alleviate the problem that the prediction box contains all the real boxes, as shown in (6).

$$DIoU = IoU - \frac{\rho^2(b,b^{gt})}{c^2} \tag{6}$$

where b and b^{gt} represent the center points of the prediction box and the marker box, respectively, ρ represents the Euclidean distance of the center points of the two boxes, c represents the diagonal distance of the minimum closure area that can contain both the prediction box and the label box.

As shown in Figure 7, if the IoU index is used, the same IoU loss will be obtained, but in fact the result of the two returning to the real box is different. Because $DIoU$ adds a penalty of center distance between frames, this problem can be solved effectively.

YOLO series of algorithms belong to one-stage single-stage object detection algorithms, and the data needs to be labeled manually. On the one hand, for a small number of target objects and a large number of backgrounds in an image, the difference between positive and negative samples is large. On the other hand, because YOLO algorithm presets a lot of prior boxes, input an image and divide it into N*N grids, each grid will produce a lot of prediction boxes, but the real number of targets is certain. Only a few of the many prior boxes contain target objects, which results in simple and difficult sample problems. Background is a distinguishable sample. Too many distinguishable samples will cause the overall learning direction of the model to deviate, making the downward direction of the loss function unexpected, resulting in invalid learning. That is, after training, the network can easily distinguish the background area without the target object and cannot distinguish the specific target object. Reducing the weight of easy-to-classify samples makes the model more focused on hard-to-classify samples in training.

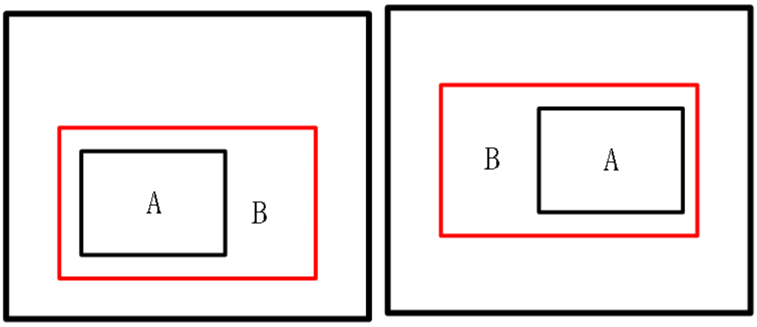


Figure 7. Complete inclusion problem.

Aiming at the above problems, this paper improves the design of the loss function of the Cell_yolo model. The improved loss function is composed of three parts, including the coordinate position prediction loss of the target box $loss_{box}^{DIoU}$, the confidence loss of the target box $loss_{obj}$ and the category loss $loss_{cls}^{Focal\ loss}$. The loss function for Cell_yolo is (7).

$$Loss = loss_{box}^{DIoU} + loss_{obj} + loss_{cls}^{Focal\ loss} \tag{7}$$

where $loss_{box}^{DIoU}$ is the loss calculated by mean square deviation function, and $loss_{obj}$ and

$loss_{cls}^{Focal\ loss}$ are the loss calculated by cross-entropy function.

For more standard detection, the improved loss function based on Focal loss is used in the model of this paper. Focal loss starts from this idea and modifies the original cross-entropy function on the basis of two-class cross-entropy function formulas such as (8):

$$L = -ylogy' - (1-y)log(1-y') = \begin{cases} -logy' & ,y = 1 \\ -\log(1-y') & ,y = 0 \end{cases} \tag{8}$$

The output of the activation function is a y' in the range of 0 to 1. The objective of the neural network is to minimize the loss function. In (8), the smaller the loss function is, the larger the probability value of the output of the activation function is for positive samples, indicating that the network model is more certain that the sample is positive. Similarly, for negative samples, the smaller the probability value of the activation function output is, the higher the probability that the model considers the sample to be negative. Focal loss is calculated as (9):

$$L_{fl=}\begin{cases} -(1-y')^{\gamma}logy' & ,y = 1 \\ -y'^{\gamma}\log(1-y') & ,y = 0 \end{cases} \tag{9}$$

Focal loss improvements seem small but can be very helpful in training. Focal loss adds a factor γ to the original. When $\gamma > 0$, the loss of easily classified samples is reduced, which makes the model more focused on difficult samples. Assuming that the probability of activation function output for a positive sample is 0.9 with $\gamma = 2$, the $(1-y')^{\gamma}$ value will be small, and the loss function value will be small. Conversely, if the probability value of the positive sample output through the activation function is 0.2, this means that although it is a positive sample, it is more likely that the model will consider the negative sample, which is a difficult sample and should be trained more. At this time, the $(1-y')^{\gamma}$ value is very large, resulting in the value of the loss function being too large, and the model will increase the learning intensity and pay more attention to the difficult samples. Similarly, for negative samples, the output probability value of 0.1 is much smaller than the loss value of 0.8. This slight change in the loss function reduces the impact of simple samples, strengthens the training of difficult samples and shifts the focus of network learning to difficult features. This solves the problem of simple and difficult samples. In addition, to solve the problem of balancing positive and negative samples, the balance coefficient α is added, and the final formula is as (10).

$$L_{fl=}\begin{cases} -\alpha(1-y')^{\gamma}logy' & ,y = 1 \\ -(1-\alpha)y'^{\gamma}\log(1-y') & ,y = 0 \end{cases} \tag{10}$$

In summary, the improved Cell_yolo loss function consists of three parts: One is the loss of confidence, which continues the cross-entropy loss function in YOLOv4. The second is the location loss, which uses the $DIoU$ loss function. The third is category loss, which uses an improved loss function based on Focal loss.

3. Region of interest recognition experiment

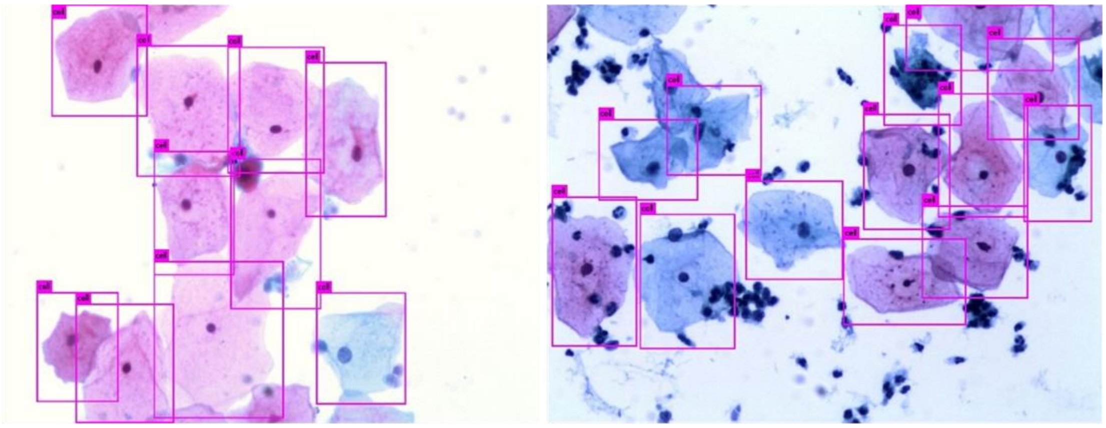
This chapter compares the Cell_Yolo network proposed in this article with the main target recognition networks YOLOv4, YOLOv4_tiny and Faster-RCNN. Faster-RCNN is an improvement of RCNN network and Faster-RCNN network, and it is a two-stage detection algorithm. The first phase of Faster-RCNN mainly generates target recommendation boxes, and the second phase adjusts and classifies target recommendation boxes. Its main improvement is to use RPN (Region box regression) instead of the original Selective Search method to generate target recommendation boxes and to achieve end-to-end object detection. Faster-RCNN uses a set of neural networks for object detection, in which parameters can be shared, which greatly improves the speed of the two-stage detection algorithm. Compared with single-stage detection methods, the recognition error rate is low, but the detection speed is poor. YOLOv4 has been described above and is not covered here. YOLOv4_tiny is a simplification of YOLOv4. Compared with YOLOv4, $Mish$ activation function is not used in feature extraction, and only one feature pyramid is used in feature fusion. The most significant feature of YOLOv4_tiny is its fast speed, but due to the simplification of feature extraction and feature fusion structure, the model is not accurate enough to detect small objects and two near objects.

Cell_Yolo is trained to recognize a single cell in the entire cell image and to give the location of each cell in the overlapping cell image. In this section, two index evaluation algorithms, Intersection over Union and Frame Per Second, are introduced. Since the Cell_Yolo used in this study does not judge the cell type, but identifies and locates the cell, the mAP (mean Average Precision) commonly used in target recognition detection is not compared. Intersection and union ratio (IoU) is a concept used to measure target positioning accuracy in object detection. The intersection-merge ratio represents the degree of coincidence between the prediction box and the real box. Mathematically, the intersection-merge ratio refers to the ratio between the intersection and the union of the two. This is described in Chapter 3 and is not repeated here. Since multiple targets can be identified, and multiple target frames can be generated in the image, using only one frame of IoU does not prove the superiority of the algorithm, so the $mIoU$ index is introduced for judgment. The $mIoU$ is the average intersection-union ratio, which is based on the actual labeled data. For example, (11) calculates the IoU of each prediction box separately from the true box and divides the number of boxes by the sum.

$$mIoU = \frac{\sum_{n=1}^{k} IoU_n}{k} \tag{11}$$

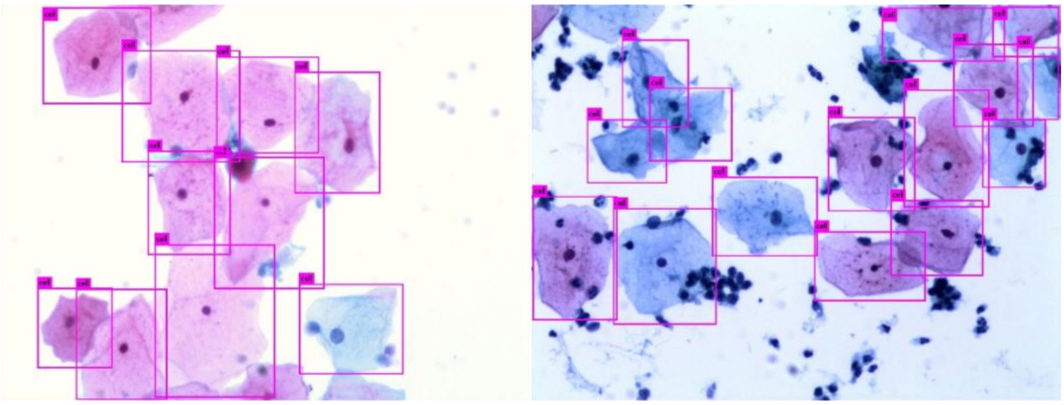
where k represents the number of actual boxes in the labeled data, and IoU_n is the IoU for calculating the nth box. This avoids the high $mIoU$ caused by the precise recognition of individual boxes in the algorithm. FPS detection frames per second is a common measure of object detection algorithms. This experiment uses FPS to evaluate the efficiency of the algorithm. When comparing FPS values of different algorithms, it is necessary to ensure that each algorithm needs the same hardware environment. The higher the FPS value is, the higher the efficiency of the algorithm, the better the performance of the algorithm in terms of efficiency. This paper compares the Cell_yolo network with the main target recognition networks YOLOv4, YOLOv4_tiny and Faster_RCNN.

Figure 8 selects two typical cell images of cervical cells, with the green box as a rectangular box indicating the location of the cells. Figure a) is a highly overlapping cell image labeled. Each cell in the field of view overlaps to some extent, and the boundary contrast at the overlap is poor. Figure b) is a labeled background complex cell image, with a small number of overlapping cells in the field of view, and excessive surrounding impurities and other unrelated cells make the background extremely complex. The principle of labeling data is to label cells that are intact, not to consider cells at the periphery of the field of view and to require that only one complete cell exists in a target frame.



a) Overlapping cell label data b) Complex background label data

Figure 8. Label data.

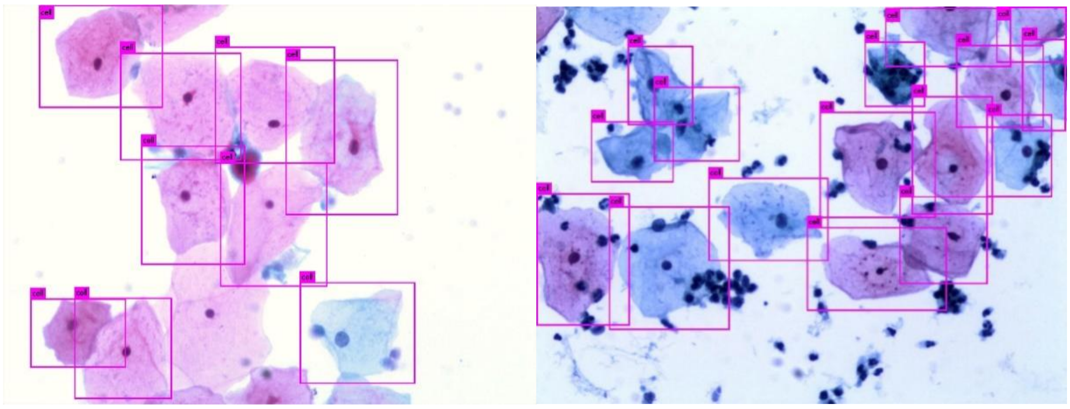


a) Overlapping cell recognition result b) Complex background recognition result

Figure 9. Cell_Yolo results.

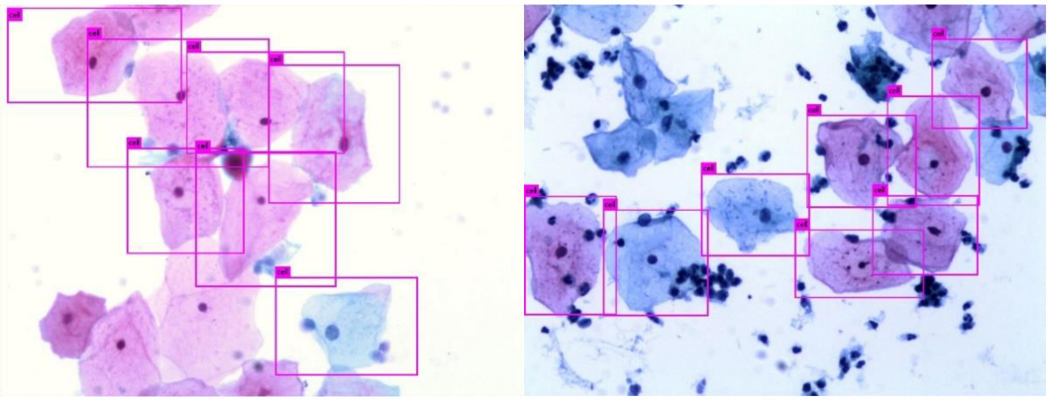
Figure 9 shows the recognition performance of the Cell_Yolo algorithm proposed in this paper. Figure a) Cell_Yolo has a strong ability to recognize overlapping cells, and each target frame does not destroy cell integrity. The model can accomplish overlapping cell recognition tasks. Figure b) The results of identifying overlapping cells in complex background show that the model has good generalization ability, few impurities can be identified as cells, the model can accurately identify the cell location under the interference of many impurities, and the model has good anti-interference ability. In terms of visual effect, Cell_Yolo can complete the identification and labeling of complex background and overlapping cell images.

Figure 10 shows the recognition result of YOLOv4. YOLOv4 is qualified in overlapping cell images and complex background images. It can recognize overlapping cells accurately and has strong anti-interference ability. Figure b) shows that the impurities surrounding the cells do not adversely affect the final accurate identification.



a) Overlapping cell recognition result   b) Complex background recognition result

Figure 10. YOLOv4 results.

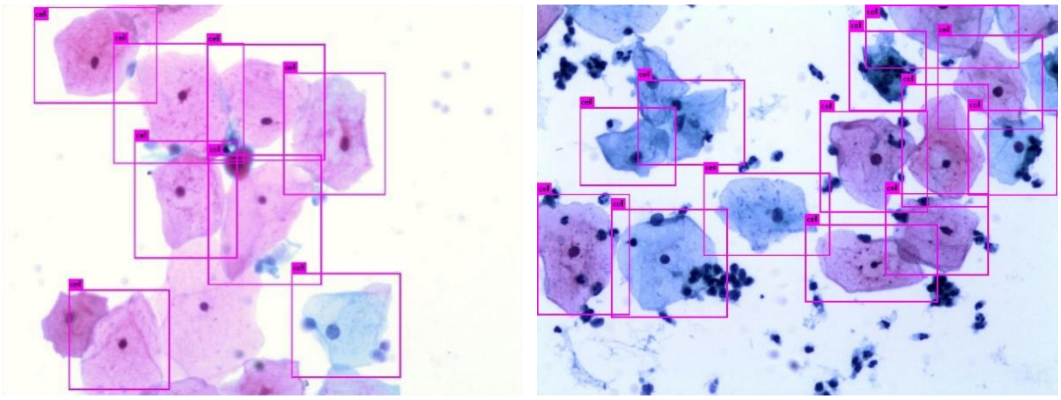


a) Overlapping cell recognition result   b) Complex background recognition result

Figure 11. YOLOv4_tiny results.

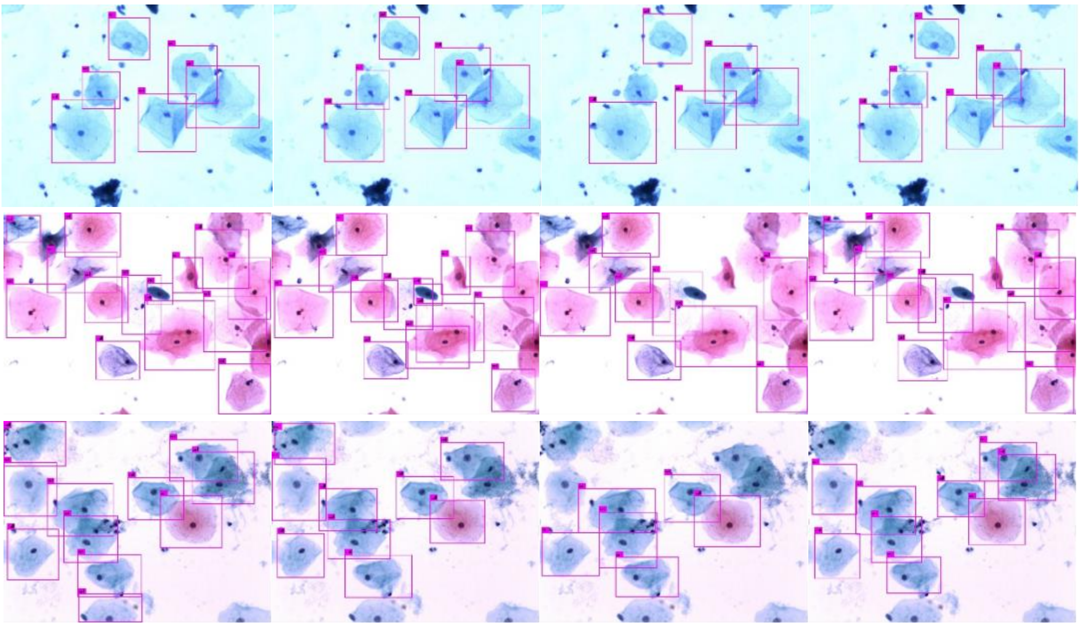
Figure 11 shows the recognition effect of YOLOv4_tiny. From the result diagram, overlapping cells are not recognized in both images. The model only has strong recognition ability for single cells, but in reality, the images of overlapping cells and complex background account for the majority. Compared with Cell_Yolo and yolov4, the recognition effect of YOLOv4_tiny model has a large gap. Compared with yolov4, YOLOv4_tiny model is greatly simplified, and the parameters of YOLOv4_tiny are ten times less than yolov4. The simplification of model and parameters leads to the inaccuracy of detection. The results in the Figure show that the recognition ability of YOLOv4_tiny on overlapping cells is poor, which makes it difficult to meet the actual needs.

The recognition effect of Faster-RCNN is shown in Figure 12. Faster-RCNN performs well in overlapping cell images and images with high background complexity, and cells can be recognized correctly. However, the regression position of the target frame is inaccurate, and some cells fail to be calibrated completely.



a) Overlapping cell recognition result b) Complex background recognition result

Figure 12. Faster-RCNN results.



a) Cell_Yolo b) YOLOv4 c) YOLOv4_tiny d) Faster_RCNN

Figure 13. Comparison of experimental data.

Figure 13 shows more experimental results, from which it can be seen that each network can better achieve the segmentation of single cells, but there are certain differences in the processing of overlapping cells and cell images under complex background. By comparison, it is obvious that the segmentation effect of YOLOv4_tiny model in special scenes is the most unsatisfactory. It has the obvious possibility of missing detection. YOLOv4 and Faster_RCNN models also achieve good results in the segmentation process, but the regression position of the target box is not ideal. The proposed model in this paper has achieved better results in various scenarios, and it is more accurate in detection accuracy and target box labeling.

The above results provide a visual comparison of the actual recognition images of each network. Table 1 shows the performance of each network model in the evaluation index $mIoU$ and FPS.

Table 1. Comparison of cell recognition effects.

	Cell_Yolo	YOLOv4	YOLOv4_tiny	Faster_RCNN
$mIoU$	0.905	0.907	0.629	0.910
FPS	67	50	90	15

From the above experimental results, it can be seen that Cell_Yolo performs well in both $mIoU$ and FPS parameters. Although YOLOv4_tiny performs best in FPS metrics, it performs worst in accuracy, which confirms the image results above. YOLOv4 is slightly higher than Cell_Yolo in $mIoU$ index, but its frame rate differs greatly from Cell_Yolo. Similarly, Faster_RCNN networks perform best in $mIoU$ metrics but are inefficient. Operating efficiency is an important reference index in this project. The whole segmented network is in two stages, where one stage completes the identification of single-cell regions of interest, and the second stage is the segmentation of cell images. Therefore, select the one with higher operating efficiency as far as possible in this phase. Cell_Yolo maintains high precision while operating efficiently. From the experimental results, Cell_Yolo is designed to meet the requirements, and its operating efficiency is about 34% higher than YOLOv4.

4. Conclusions

The ultimate goal of the method proposed in this paper is to target highly overlapping regions between cells in cervical cell segmentation, and the poor contrast of the overlapping boundaries of cells makes cervical cancer image segmentation difficult. This paper takes cervical squamous cells as the research object, simplifies the network structure of YOLOv4, adopts the improved maximum pooling method to maximize the transmission of image feature information in the neural network and proposes a center distance for the problem of mistaken deletion of overlapping cell frames. Regarding the NMS algorithm, the whole algorithm greatly improves the detection rate while ensuring the detection accuracy. Simplified network architecture also facilitates model training and practical application. In the subsequent research process, we will continue to collect and find more data sets to optimize the network. It will also be tested in segmentation of other cancer cells.

Acknowledgments

We acknowledge that this research is supported by the Beijing Jiaotong University (grant 2022YJS020, W21ZZ200030 and W19L00130).

Conflict of interest

The authors declare there is no conflict of interest.

References

1. R. Elakkiya, K. S. S. Teja, L. J. Deborah, C. Bisogni, C. Medaglia, Imaging based cervical cancer diagnostics using small object detection-generative adversarial networks, *Mult. Tools Appl.*, **81** (2022), 191–207. https://doi.org/10.1007/s11042-021-10627-3

2. J. C. Davies-Oliveira, M. A. Smith, S. Grover, K. Canfell, E. J. Crosbie, Eliminating cervical cancer: progress and challenges for high-income countries, *Clin. Oncol.*, **33** (2021), 550–559. https://doi.org/10.1016/j.clon .2021.06.013
3. M. E. Plissiti, C. Nikou, A Review of Automated Techniques for Cervical Cell Image Analysis and Classification, *Springer Netherlands*, **4** (2013), 1–18. https://doi.org/10.1007/978-94-007-4270-3_1
4. M. E. Plissiti, C. Nikou, Overlapping Cell Nuclei Segmentation Using a Spatially Adaptive Active Physical Model, *IEEE Transactions on Image Processing A Publication of the IEEE Signal Processing Society*, **21** (2012), 4568–580. https://doi.org/10.1109/TIP.2012.2206041
5. N. M. Harandi, S. Sadri, N. A. Moghaddam, R. Amirfattahi, An Automated Method for Segmentation of Epithelial Cervical Cells in Images of ThinPrep, *J. Med. Syst.*, **34** (2010), 1043–1058. https://doi.org/10.1007/s10916-009-9323-4
6. A. Genctav, S. Aksoy, S. Onder, Unsupervised segmentation and classification of cervical cell images, *Pattern Recogn.*, **45** (2012), 4151–4168. https://doi.org/10.1016/j.patcog.2012.05.006
7. A. Kale, S. Aksoy, Segmentation of cervical cell images, *2010 20th International Conference on Pattern Recognition, IEEE*, (2010), 2399–2402. https://doi.org/10.1109/ICPR.2010.587
8. T. Chankong, N. Theera-Umpon, S. Auephanwiriyakul, Automatic cervical cell segmentation and classification in Pap smears, *Computer Meth. Progr. Biomed.*, **113** (2014), 539–556. https://doi.org/10.1016/j.cmpb.2013.12.012
9. H. Lee, J. Kim, Segmentation of overlapping cervical cells in microscopic images with superpixel partitioning and cell-wise contour Refinement, *29th IEEE Conference on Computer Vision and Pattern Recognition*, (2016), 1367–1373. https://doi.org/10.1109 /CVPRW.2016.172
10. B. Dong, L. Jia, Y. Wang, J. Li, G. Yang, An improved watershed algorithm based on k-medoids in cervical cancer images, *2019 IEEE International Conferences on Ubiquitous Computing & Communications (IUCC) and Data Science and Computational Intelligence (DSCI) and Smart Computing, Networking and Services (SmartCNS), IEEE*, (2019), 190–195. https://doi.org/10.1109/IUCC/DSCI/SmartCNS.2019.00060
11. C. Jung, C. Kim, S. W. Chae, S. Oh, Unsupervised Segmentation of Overlapped Nuclei Using Bayesian Classification, *IEEE Transact. Biomed. Eng.*, **57** (2010), 2825–2832. https://doi.org/10.1109/tbme.2010.2060486
12. D. N. Diniz, M. T. Rezende, A. G. C. Bianchi, C. M. Carneiro, D. M. Ushizima, F. N. S. de Medeiros, M. J. F. Souza, A hierarchical feature-based methodology to perform cervical cancer classification, *Appl. Sciences-Basel*, **11** (2021), 4091. https://doi.org/10.3390/app11094091
13. J. Long, E. Shelhamer, T. Darrell, Fully convolutional networks for semantic segmentation, *The IEEE conference on computer vision and pattern recognition*, **39** (2017), 640–651. https://doi.org/10.1109/TPAMI.2016.2572683
14. V. Badrinarayanan, A. Kendall, R. Cipolla, SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation, *IEEE Transact. Pattern Anal. Mach. Intell.*, **39** (2017), 2481–2495. https://doi.org/10.1109/TPAMI.2016.2644615
15. E. Giacomello, D. Loiacono, L. Mainardi, Brain MRI Tumor Segmentation with Adversarial Networks, *2020 International Joint Conference on Neural Networks (IJCNN), IEEE*, (2020), 1–8. https://doi.org/10.1109/IJCNN48605.2020.9207220

16. D. Yang, D. Xu, S. K. Zhou, B. Georgescu, M. Q. Chen, D. Comaniciu, Automatic liver segmentation using an adversarial image-to-image network, *International Conference on Medical Image Computing and Computer-Assisted Intervention*, (2017), 507–515. https://doi.org/10.1007/978-3-319-66179-7_58
17. R. Krithiga, P. Geetha, Breast cancer detection, segmentation and classification on histopathology images analysis: A systematic review, *Arch. Comput. Methods Eng.*, **28** (2021), 2607–2619. https://doi.org/10.1007/s11831-020-09470-w
18. O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, *International Conference on Medical image computing and computer-assisted intervention. Springer. Cham.*, (2015), 234–241. https://doi.org/10.48550/arXiv.1505.04597
19. X. Y. Li, L. L. Shen, cC-GAN: A robust transfer-learning framework for HEp-2 specimen image segmentation, *IEEE Access*, (2018),14048–14058. https://doi.org/10.1109/access.2018.2808938
20. Y. Nambu, T. Mariya, S. Shinkai, M. Umemoto, H. Asanuma, I. Sato, et al., A screening assistance system for cervical cytology of squamous cell atypia based on a two-step combined CNN algorithm with label smoothing, *Cancer Med.*, **11** (2022), 520–529. https://doi.org/10.1002/cam4.4460
21. Z. Q. Xing, X. Chen, F. Q. Pang, DD-YOLO: An object detection method combining knowledge distillation and Differentiable Architecture Search, *IET Computer Vision*, **16** (2022), 418–430. https://doi.org/10.1049/cvi2.12097
22. L. C. Jiao, F. Zhang, F. Liu, S. Y. Yang, L. L. Li, Z. X. Feng, et al., A survey of deep learning-based object detection, *IEEE Access*, **7** (2019), 12883–128868. https://doi.org/10.1109/ACCESS.2019.2939201
23. A. Bochkovskiy, C. Y. Wang, H. Liao, YOLOv4: Optimal speed and accuracy of object detection, *Computer Sci.*, (2020). https://doi.org/10.48550/arXiv.2004.10934
24. S. Liu, L. Qi, H. F. Qin, J. P. Shi, J. Y. Jia, Path aggregation network for instance segmentation, *31st IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),* (2018), 8759–8768. https://doi.org/10.1109/CVPR.2018.00913