A Multi-scale Pyramid of Fully Convolutional Networks for Automatic Cell Detection

Jiang Gu[1]*, Yichen Zhu[2]*, Bohong Yang[1], Jingkai Jia[3], Juanjuan Wang[4], Jian Yang[5], Wenqiang Zhang[1]**
[1]Shanghai Key Lab of Intelligent Information Processing, Fudan University, Shanghai, China
[2]Department of Statistical Sciences, University of Toronto, Toronto, Canada
[3]Department of Electrical and Computer Engineering, University of Toronto, Toronto, Canada
[4]Shanghai Intelligent Medical Tech Co.,Ltd, Shanghai, China
[5]Changhai Hospital of Shanghai, Shanghai, China
*Co-first author
**Corresponding author: wqzhang@fudan.edu.cn

***Abstract*—Cell detection from the histology images is a crucial prerequisite step to narrow down the interested areas for important following analyses. While detecting cell manually is a time-consuming and labor-intensive work, automatic cell detection is in demand. However, automatic cell detection is a challenging task due to the complexity of histology images, including the variations in cell size, shape and texture. Previous methods favour handcrafted features, which lack the robustness to different cell types, while convolutional neural network based methods often suffer from the problem of maintaining spatial information and semantic information. We propose a novel network structure to leverage the balance between spatial information and semantic information on cell detection. We also design a variant of focal loss to deal with the severe imbalance between nuclei pixels and background pixels, which is a common problem in cell detection that may greatly affect the results from network. Our proposed method is evaluated on a typical colorectal histology image dataset and the result from our method outperforms the previous state-of-the-art method (F1 score 0.834) by a large margin, achieving a new state-of-the-art (F1 score 0.889).**

***Index Terms*—cell detection, histopathology, convolutional neural network**

I. Introduction

Cell detection is considered to be a prerequisite for histology image analysis. Many quantitative analyses, such as cell morphology analysis, including the computation of cell shape, size, location and other imagenomics, are based on the results of cell detection [1]. Therefore, the results of cell detection will greatly affect the quality of the following analyses. Due to the complex texture of histology images and the large amount of cells in histology images, it is a time-consuming and labor-intensive work to perform manual cell detection and the results of manual cell detection are largely influenced by the subjectivity of the observers. Therefore lots of efforts have been made to design an effective automatic method for cell detection.

Most of the early automatic cell detection methods are based on handcrafted features. These methods use image processing methods such as edge detection, to extract features from histology images, and then use machine learning techniques including random forest and support vector machine (SVM) to detect the cells [2]–[5]. However, handcrafted features are not capable of the complexity of histology images. Taking colorectal histology images as an example, as shown in Fig.1, there are four types of cells in a single image. These cells differ in size, shape and texture, making it difficult for handcrafted feature based methods to establish generalized features which are suitable for all cell types at different gland stages.

Recently, with the development of deep learning techniques, researchers have tried to apply convolutional neural networks (CNN) for cell detection. One common problem is how to leverage the balance between spatial information and semantic information. The former is essential for finding the accurate locations of cells and the latter is necessary for telling whether a pixel is cell or not. Without either spatial information or semantic information, the model would fail to detect the cells.

In this paper, we propose a novel yet effective encoder-decoder based network structure for cell detection. Inspired by Yu et al. [6], we use dilated convolution in our network structure to integrate more spatial information which helps detecting cells from complicated background pixels. To strike the severe imbalance between cell pixels and background pixels, we propose a variant focal loss which enforces the network to focus on cell pixels as well as the hard negative cells during the training phase. To demonstrate the effectiveness of our network, we evaluate our proposed model on a typical cell detection dataset. We show that our method outperforms the previous state-of-the-art model by a large margin, setting a new state-of-the-art.

In summary, our contributions are:

- We proposed a network that employs dilated convolution to our encoder module and add a pyramid-based decoder module to fully capture the spatial localization information, resulting in a stronger cell detector.
- A variant of focal loss is introduced to relieve the class imbalance issue.
- Our proposed model outperforms previous state-of-the-art cell detection model by a large margin on a typical colorectal histology image dataset and achieves a new state-of-the-art.

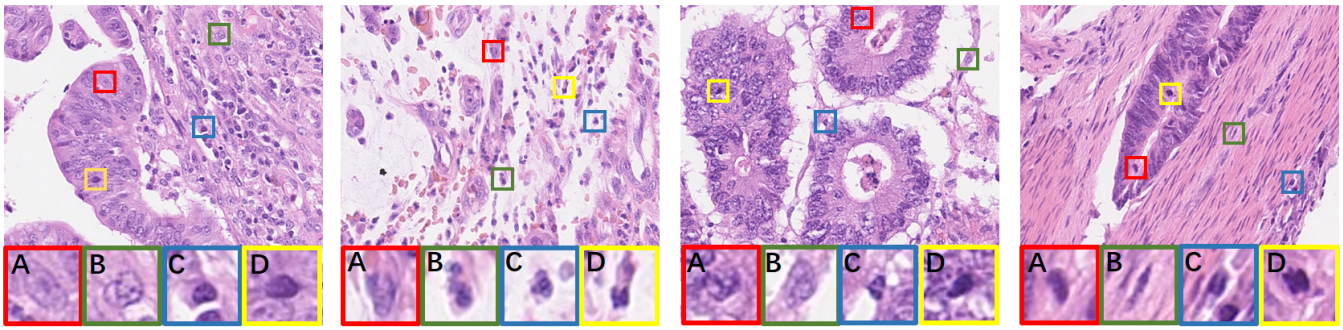


Fig. 1. Examples for different cell types in histology images. There are four types of nuclei in these images. A. epithelial nuclei, B. fibroblast nuclei, C. inflammatory nuclei, D. miscellaneous nuclei. The nuclei differ in size, shape and texture, bringing difficulties for cell detection.

The rest of this paper is organized as follows: In Section II, we introduce some related works. In Section III, we give a detailed introduction about our network structure and the variant focal loss. In Section IV, we compare our experiment results with other cell detection methods and have an ablation study on our network.

II. Related Work

In this section, we make a brief introduction to the previous work on cell detection and other related techniques that are utilized in our work.

Cell detection: Early researches on automatic cell detection are mainly based on handcrafted features, together with machine learning methods. Al-Kofahi et al. [2] used a graph-cuts-based binarization to automatically extract foreground from the image, and then performed nuclear seed points detection by using Laplacian-of-Gaussian filtering. The seed points were used to perform segmentation of the nuclei which was refined by another graph-cuts-based algorithm. Veta et al. [3] applied morphological grayscale reconstruction to preprocess the images, and then used marker-controlled watershed segmentation to find the regions of interest on whole slide images. Mualla et al. [4] applied SIFT to extract keypoints which are invariant to cell size and shape from the image, and used random forest as a background-cell classifier to discard the background keypoints. Then, a profile was established to tell whether two keypoints belong to a same cell. Finally, hierarchical clustering was used to combine the results of keypoint learning and profile learning, finding the location of the nuclei. Cosatto et al. [5] used color unmixing to separate the histology images into two channels, and then used DoG filter and Hough transform to perform cell detection. Kuse et al. [7] proposed a local isotropic phase symmetry measure to identify image pixels that may be high in local symmetry for beta cell detection. However, due to the complexity of histology images, the handcrafted features which integrate low-level features of the images do not have enough representation ability. The lack of rich semantic information makes them sensitive to cell morphology changes and lack robustness for cell detection.

Recently, with the development of convolutional neural networks, deep learning methods have shown their advantages in automatic cell detection. Xue et al. [8] employed random projections to encode the output pixel space to a compressed vector of fixed dimension and then used a CNN to regress the compressed vector from the input pixels. Xie et al. [9] developed and compared two fully convolutional regression networks, using them to regress a cell spatial density map for detection. Xie et al. [10] proposed two types of units, i.e. voting units and weight units, and merged them into a layer to form a deep voting model to find nuclei in microscopy images. Chen et al. [11] performed mitosis detection by using a cascaded network, which combines an efficient coarse retrieval model and a knowledge transferred fine discrimination model. Xie et al. [12] proposed a CNN based structured regression model (SR-CNN), which fuses the predictions for each pixel with its neighbours to get a proximity patch and uses local maximum to find the cell location. Sirinukunwattana et al. [13] proposed a Spatially Constrained Convolutional Neural Network (SC-CNN), where spatial constraints are added to the last two layer of the network. With these two constraint layers, high probability values are spatially constrained to locate in the vicinity of the centers of nuclei. Based on SC-CNN, Kashif et al. [14] combined handcrafted features with raw data as the input of the network and improved the performance. Similarly, Wang et al. [15] used both handcrafted and CNN features for cell detection, and performed a two-stage cascaded ensemble to combine the results given by the two types of features. Zhou et al. [16] proposed a FCN based network which has two sibling branches that perform both cell detection and classification. By establishing the interaction between the detection branch and the classification branch, the network yields good performances on both tasks.

Class imbalance: Object detection methods, both classic and modern methods, are suffering from the problem that most datasets are highly imbalanced, only a few locations in a sample contain objects that we are interested in. This imbalance may cause the network to be more likely to

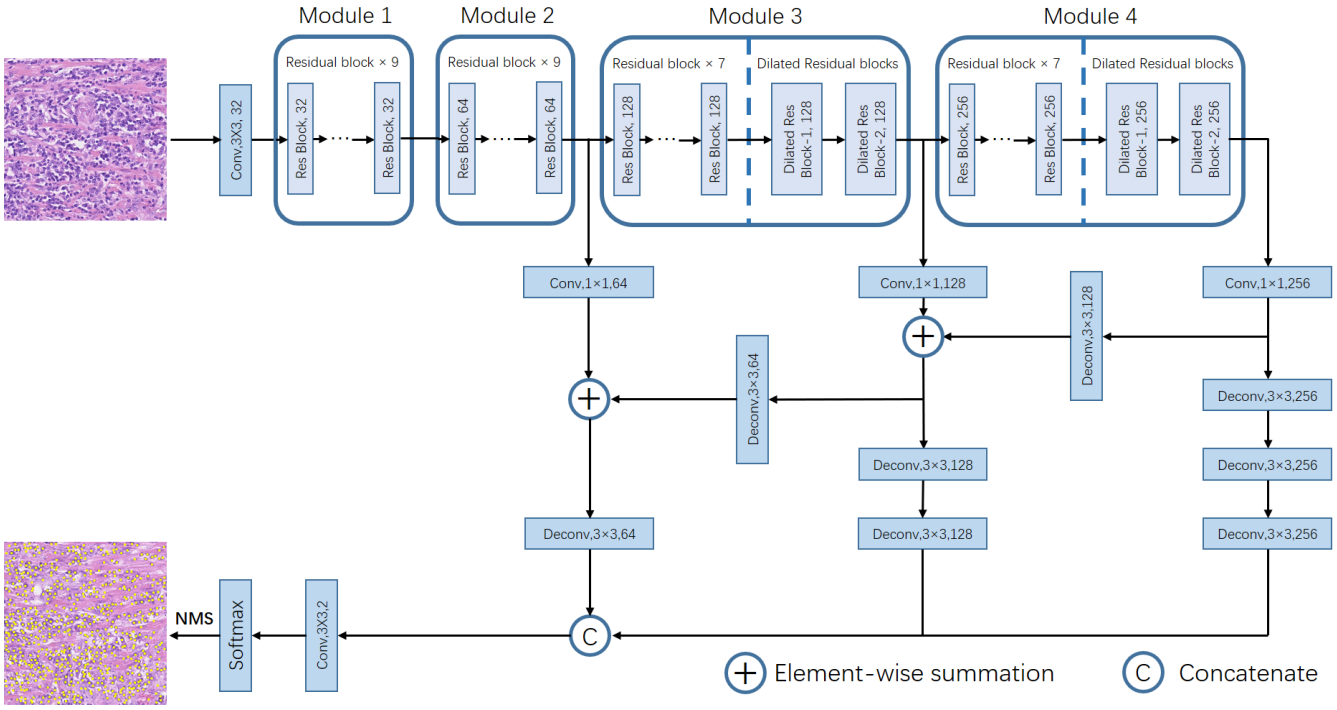


Fig. 2. An overview of our proposed network structure.

produce false negative results, decreasing the performance of the networks significantly. One approach [17]–[21] to solve the imbalance problem is to perform hard negative mining that selects hard negative samples during training phase based on loss [22] or based on modified region-based object detectors [17]. Another approach is to modify the loss to make it robust to the imbalance between classes. Focal loss, proposed in [22], is one of the modified losses. Instead of sampling hard negative samples, focal loss decreases the weights of easy samples, which often refer to the background pixels, to deal with the imbalance between classes.

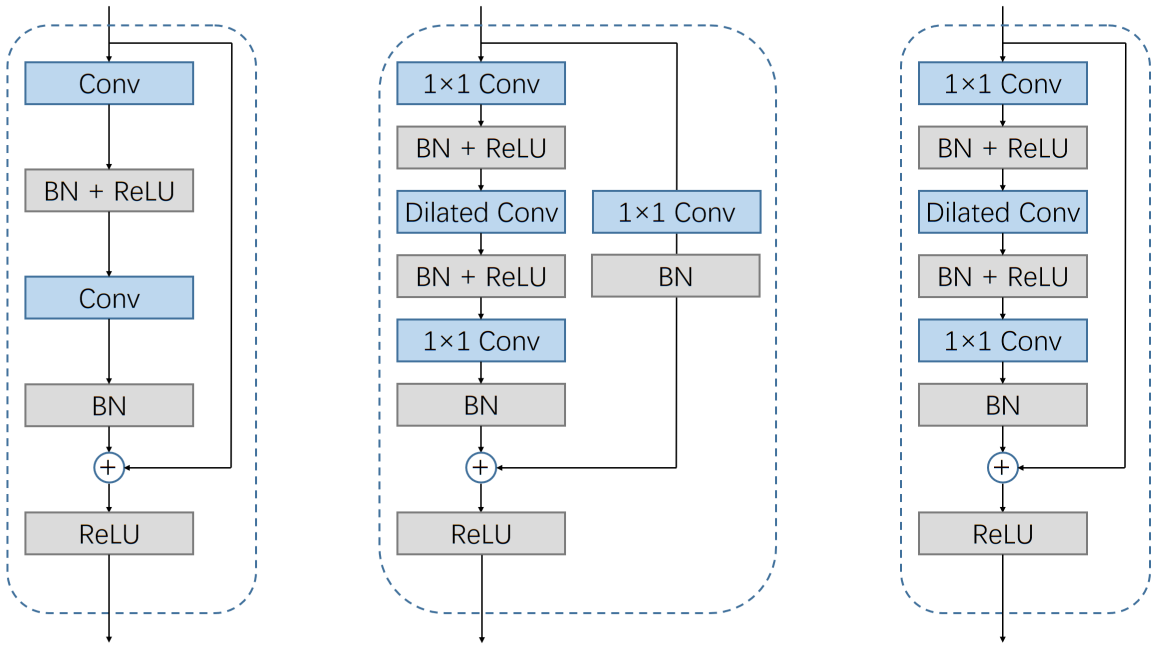
Encoder-decoder: The encoder-decoder structure has been proven to be successful in many computer vision tasks, such as object detection [23]–[25] and semantic segmentation [6], [26]–[33]. A typical encoder-decoder structure consists of two parts, (1) An encoder module that gradually reduces the size of feature maps to capture the high semantic information. (2) A decoder module that gradually recovers the spatial information. For example, Segnet [28] reserves the pooling indices and reuses them in decoder modules to upsample the feature maps. Unet [30] employs skip connection and concatenates feature maps from encoder module to the corresponding decoder module. In this paper, we take the advantage of FPN [23] and add a decoder structure on top of our encoder module to boost our model performance on cell detection task.

Dilated convolution: Dilated convolution has been shown in segmentation tasks with significant improvement on accuracy [26], [34], [35]. By enlarging receptive fields, dilated convolution can effectively incorporate surrounding contexts. Deeplab [34] exploited multi-scale features by employing multiple dilated convolutions with different dilation rates on spatial pyramid pooling. Wei et al. [35] employed dilated convolutions as Class Activation Mapping [36] under weakly supervised semantic segmentation. In this paper, we design a block that employ dilated convolution at encoder stage, which we call dilated residual blocks, to effectively capture the spatial contextual information of cells.

III. Methods

A. Encoder structure

As shown in Fig. 2, our encoder structure mainly consists of four residual modules, which extract the multi-scale features from the input image. To effectively train our deep encoder structure, we employ residual connection to avoid the issue of gradient vanish and gradient explosion, which commonly occur during training a deep neural network. A typical residual block contains a stack of three 3×3 convolution layers followed by batch normalization (BN) and ReLU non-linearity activation function. Nevertheless, to reduce computational cost, we remove one stack of convolution layer. As a result, there is a stack of two 3×3 convolution for every residual block, and each residual module contains nine residual blocks. We denote each residual block as $x_{i+1} = W_s x_i + F_i(x_i; W_s)$. W_s is used to downsample the feature map by a 1×1 convolution with stride of 2 at the beginning of the last



(a) Residual Block (b) Dilated Residual Block-1 (c) Dilated Residual Block-2

Fig. 3. Structure of residual block and dilated residual blocks.

three residual modules, and in other cases W_s means identity mapping.

Let H and W denote the height and width of the input image. We first use a 3×3 convolutional layer with 32 filters followed by BN and ReLU to extract shallow features. The feature maps are then passed through four residual modules hierarchically. Module 2, 3 and 4 downsample the feature map size by half, i.e. the feature map sizes in Module 2, 3 and 4 are (H/2, W/2), (H/4, W/4) and (H/8, W/8) respectively. These relatively large feature map sizes help to retain spatial information for the deconvoling operations in the decoder structure. The numbers of feature maps in four residual modules are 32, 64, 128, 256.

Due to the complexity of histology images, it is difficult to tell whether a pixel belongs to a cell or not if insufficient surrounding information is given. In order to extract features which contain surrounding information, the network should have a relatively large receptive field. Therefore, for the last two residual blocks in Module 3 and Module 4, we use dilated convolutional layers with a dilation rate of 2, namely dilated residual blocks, to enlarge reception fields while keeping the spatial information by having a relatively large feature map size. The structures of dilated residual blocks are shown in Fig. 3.

B. Decoder structure

As for decoder structure, we use a feature pyramid like structure to combine multi-scale feature maps to make use of both high semantic information and spatial information. Take Module 4 as an example, the outputs of Module 4 are first passed through a 1×1 convolutional layer with 256 filters followed by BN and ReLU, and then they are upsampled by a deconvolutional layer with a stride of 2. The upsampled feature maps are then concatenated with the outputs from Module 3, which have been passed through a 1×1 convolutional layer with 128 filters, and once again they are upsampled by a deconvolutional layer. This operation is repeated until the feature maps are recovered to the original size. We also employ deconvolutional layer to recover feature maps from Module 4, Module 3 and Module 2, to the origin size individually. Finally, we concatenate the three groups of feature maps and send them to a 3×3 convolutional layer followed by a softmax layer to generate the probability predictions for each pixel in the input image. The output is composed of two channels, representing the background and the nuclei respectively.

C. Loss Function

In histology images, the number of nuclei pixels is much smaller than the number of background pixels, resulting in the severe imbalance between positive samples and negative samples. This problem generally exists in different organs, e.g. colon and breast. Without a suitable solution, this imbalance will bring about the issue that the network is more likely to classify a pixel into a background pixel, decreasing the performance of the network significantly. In the meanwhile, easy negative samples, such as the white background space and benign gland in some histology images, contribute little to the training of the network. Nevertheless, they may have a large impact on the loss because of their huge amount, that reducing the network's ability of classifying nuclei.

Towards the targets of resolving both the class imbalance problem between foreground and background as well as degenerated model issues due to easy negative samples, we propose a variant focal loss [22] based on the binary cross-entropy loss function. Specifically, we introduce two weighting factors, α and β, to deal with the problem of easy samples. We denote b_{focal} to be the focal weighting factor for background and f_{focal} to be the focal weighting factor for foreground. They are defined by (1) and (2), where P_{bkg} is the probability of being a background pixel and P_{obj} is the probability of being a cell. Besides the focal weighting factors, we also use two weights W_{bkg} and W_{obj} to balance the number of background pixels and foreground pixels. So the loss function for background and foreground can be defined by (3) and (4), and the total loss of our network is the summation of these two losses.

$$b_{focal,i} = (1 - P_i^{bkg})^{\alpha} \tag{1}$$

$$f_{focal,i} = (1 - P_i^{obj})^{\beta} \tag{2}$$

$$L_{bkg,i} = W_{bkg}[y_i = 0] b_{focal,i} log(P_i^{bkg}) \tag{3}$$

$$L_{obj,i} = W_{obj}[y_i = 1] f_{focal,i} log(P_i^{obj}) \tag{4}$$

$$L = -\frac{1}{N}\sum_{i=1}^{N} L_{bkg,i} + L_{obj,i} \tag{5}$$

The focal parts b_{focal} and f_{focal} are designed to resolve the issue of easy samples. If a certain pixel is easy to be classified, i.e. the output probability of the network to this pixel is high, there is no need for the network to pay more attention to the pixel during the rest training phase. Reasonably, the weight of this pixel should be decreased so that the network can move its attention to pixels that are ambiguous. By applying the focal part, the weights of these easy samples decrease as their predicted probabilities rise.

The weights W_{bkg} and W_{obj} aim to alleviate the extreme imbalance between foreground and background. These two hyper-parameters should be set according to the numbers of background pixels and nuclei pixel.

By combining the two parts, the loss is able to strengthen the network's ability of detecting nuclei from complicated background regions.

D. Training Strategies

As our network is designed to perform cell detection as a pixel-wise classification task, we need a ground truth mask for every input image. While the dataset we use gives annotations in the form of points representing the center of every existing nucleus, directly generating a mask with a single pixel at the center of each nucleus is not sufficient for training our network, so we mark a circle with 3-pixel radius around each point as ground truth region to generate the mask labels.

During the testing phase, given a test image, the network gives out two probability maps representing background and nuclei regions. Then we conduct a non-maximum suppression to the nuclei probability map to get the exact locations for nuclei prediction.

TABLE I
EXPERIMENT RESULTS OF ABLATION STUDY.

Methods	Precision	Recall	F1
Full-w/decoder	0.770	0.868	0.816
Full-w/dilation	0.822	0.886	0.852
Full-crossentropy	0.831	0.915	0.871
Full-bilinear	0.815	0.905	0.857
Ours	**0.845**	**0.938**	**0.889**

IV. EXPERIMENTS

A. Dataset

To evaluate our proposed network, we conduct our experiments on a classic cell nuclei detection dataset [13]. This dataset contains 100 H&E stained histology images of colorectal adenocarcinomas and totally 29,756 nuclei were marked at the center of target nuclei for detection purposes. Many previous methods use this dataset to evaluate their work since it has been published in 2016.

B. Implementation Details

Data augmentation. We randomly divide 100 histology images into training, validation and test sets by a ratio of 7:1:2. We resize each image into size of 256×256 without cropping to keep all spatial information. We use images with sizes of 256×256 instead of original size mainly because of insufficient computational resources. Then for each image we do randomly data augmentation, to be specific, we randomly combine horizontal and vertical flipping, zooming, rotating, shearing and channel shifting. At testing phase, we resize each image into size of 256×256 just as we did at training phase, and then pass it into the network to generate output. As the output of our network is a score map with size of 256×256, we resize it back to original size of 500×500 to evaluate our result.

Training. The whole network is trained end to end from scratch using stochastic gradient descent with initial learning rate of 0.01 and 0.9 momentum. The learning rate is decayed to 0.001 at 100 epochs and 0.0001 for the next 50 epochs. The weight decay is set to 0.0001. All weights are initialized using Xavier uniform initializer. We trained our model on four GTX Titan X GPU with a total batch size of 28 for 250 epochs.

C. Evaluation Metrics

We harness the evaluation metrics which used in [16] to evaluate our network and further comparison with other methods. Specifically, we draw a circle with 6-pixel radius at each cell centroid. Prediction results that locate in the circles

are considered as true positives (TP) and the ones outside the circles are all considered as false positives (FP). The ground truth circles which contain no prediction results are considered as false negatives (FN). Then we use the following formula to calculate the precision, recall and F1 score respectively.

$$Precision = \frac{TP}{TP + FP}, \qquad Recall = \frac{TP}{TP + FN}$$

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

D. Experiment Results

Ablation study. We made an ablation study on certain parts of our network to demonstrate the effectiveness of our proposed network. As shown in Table I, Full-w/decoder represents that the model employs a FCN network without decoder structure. This structure upsamples the feature map from deep layer without utilizing the feature in shallow layer. Because feature map in low layer often contains various edge and contour information for both cell and gland, the result is relatively low without fully use of these information.

The Full-w/dilation represents model using our proposed encoder-decoder structure without replacing convolution layers with dilated convolution layers at Module 3 and Module 4. Comparing to our method, this model yield a lower score on precision, recall and f1. Because dilated convolution enables the network to have large receptive field such that the spatial information is able to be preserved.

The Full-crossentropy represents model using a weighted cross-entropy loss instead of our variant focal loss. The performance is lower by 0.18 in F1 score comparing to method using variant focal loss. Because of the severe imbalance between nuclei pixels and background pixels, the model is largely influenced by the background regions which are often easy samples and incline to predict a pixel as background.

The Full-bilinear represents using bilinear upsampling instead of deconvolution in the decoder structure. Upsampling layer is typically used along with feature pyramid methods [34] to recover the spatial information of feature maps at deep layers. However, since the smallest size of feature map is 1/8 of the input image, directly upsampling feature map 8 times may result in misalignment. Therefore we use multiple deconvolution layers, which can gradually recover the feature maps to the original input size as well as learning the parameters during the training phase. As can be seen from the table, the scores of using bilinear upsampling are lower than scores from our model.

Compared with the networks above, our proposed network achieves the highest F1 score of 0.889. By using dilated convolution, the feature maps at deep layers have a large receptive field, which enables them to have strong representative capabilities for the complicated nuclei regions and their surroundings. By applying multiple deconvolution layer, the feature maps are able to preserve spatial information. By using focal loss, the model are able to focus on detecting cells and hard negative cell samples, boosting the network performance.

TABLE II
COMPARISON WITH OTHER METHODS

Methods	Precision	Recall	F1
LIPSSyM	0.725	0.517	0.604
CRImage	0.657	0.461	0.542
SSAE	0.617	0.644	0.630
SR-CNN	0.783	0.804	0.793
SC-CNN	0.781	0.823	0.802
SFCN-OPI	0.819	0.874	0.834
Ours	**0.845**	**0.938**	**0.889**

Comparison with other methods. We compare our cell detection model against other six well-established or state-of-the-art methods on the same dataset under same evaluation metrics. We will briefly introduce the other five methods and focus more on SFCN-OPI which is the state-of-the-art method.

The first two methods are based on handcrafted features. Specifically, LIPSSyM [7] designs a local isotropic phase symmetry measure for cell and lymphocytes detection and CRImage [37] proposed a hierarchical multi-resolution model. We employ the results given by Sirinukunwattana et al. [13] that implemented these two methods on the dataset we use for comparison.

The rest four methods are CNN based. SSAE [38] uses stacked sparse autoencoder for unsupervised cell detection. The result is also implemented by Sirinukunwattana et al. [13]. SR-CNN [12] proposed a CNN based structured regression model, which fuses the predictions for each pixel with its neighbours to get a proximity and uses local maximum to find the cell location. SC-CNN [13] proposed a spatially constrained CNN by adding two extra layers that perform spatial constraints to the output of the CNN. SFCN-OPI [16] proposed a sibling FCN with object prior interaction that performs both cell detection and classification. The two sibling branches for detection and classification share the features extracted by shallow layers. As mentioned in [16], the two tasks are closely related so both branches can benefit from the shared features. And by adding an objectness prior interaction with the two sibling branches, which further integrates the features learned by the two branches, the network achieves the state-of-the-art performance on both detection and classification.

As can be seen from Table II, our network outperforms all the above methods on cell detection, setting a new state-of-the-art. We achieve highest scores in both precision and recall, and setting the F1 score to 0.889.

Qualitative results. We visualize the detection results of our proposed method on some typical histology images. As

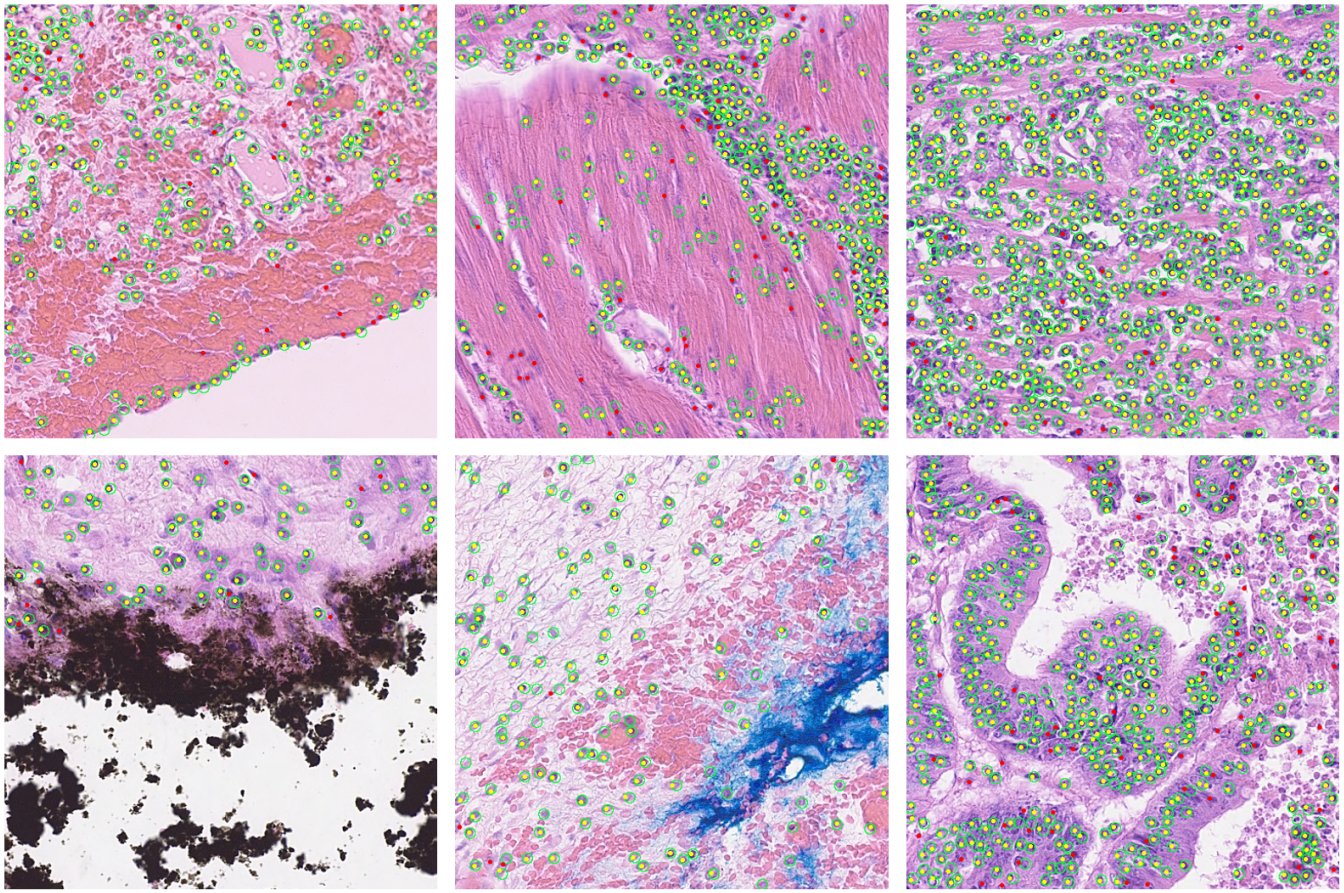


Fig. 4. Results of our proposed method. The ground truth regions are marked by green circles and the predicted locations are marked by dots. The yellow dots represent true positives and the red dots represent false positives.

shown in Fig. 4, the ground truth regions are marked by green circles with a radius of 6 pixels and the nuclei detected by our network are marked by dots. The true positives, which locate inside the ground truth regions, are marked by yellow dots, and the false positives, which locate outside the ground truth regions, are marked by red dots. As can be seen, our network succeeds in detecting nuclei of different types with large variance in size, shape and texture.

V. Conclusion

In this paper, we propose a novel network for cell detection. We use dilated convolution to integrate surrounding information to help classifying nuclei from complicated background. To leverage the balance between spatial information and semantic information, we use deconvolution to upsample the feature maps of different scales. Additionally, We propose a modified focal loss to deal with the severe imbalance between nuclei pixels and background pixels. We demonstrate the effectiveness of our proposed method by comparing with previous state-of-the-art methods on a typical colorectal histology image dataset and achieve a new state-of-the-art. Further researches will be made on this network structure to perform both nuclei detection and classification by adding a classification branch.

VI. Acknowledgements

We would like to thank Shanghai Intelligent Medical Tech Co.,Ltd and Changhai Hospital of Shanghai for their help in the research.

References

[1] M. M. Veta, "Breast cancer histopathology image analysis," Ph.D. dissertation, Utrecht University, 2014.

[2] Y. Al-Kofahi, W. Lassoued, W. Lee, and B. Roysam, "Improved automatic detection and segmentation of cell nuclei in histopathology images," *IEEE Transactions on Biomedical Engineering*, vol. 57, no. 4, pp. 841–852, 2010.

[3] M. Veta, P. J. van Diest, R. Kornegoor, A. Huisman, M. A. Viergever, and J. P. Pluim, "Automatic nuclei segmentation in he stained breast cancer histopathology images," *Plos One*, vol. 8, no. 7, p. e70221, 2013.

[4] F. Mualla, S. Scholl, B. Sommerfeldt, A. Maier, and J. Hornegger, "Automatic cell detection in bright-field microscope images using sift, random forests, and hierarchical clustering." *IEEE Transactions on Medical Imaging*, vol. 32, no. 12, pp. 2274–2286, 2013.

[5] E. Cosatto, M. Miller, H. P. Graf, and J. S. Meyer, "Grading nuclear pleomorphism on histological micrographs," in *International Conference on Pattern Recognition*, 2008, pp. 1–4.

[6] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," 2015.

[7] M. Kuse, Y. F. Wang, V. Kalasannavar, M. Khan, and N. Rajpoot, "Local isotropic phase symmetry measure for detection of beta cells and lymphocytes," *Journal of Pathology Informatics*, vol. 2, no. 2, p. S2, 2011.

[8] Y. Xue and N. Ray, "Cell detection in microscopy images with deep convolutional neural network and compressed sensing," 2018.

[9] W. Xie, J. A. Noble, and A. Zisserman, "Microscopy cell counting and detection with fully convolutional regression networks," pp. 1–10, 2016.

[10] Y. Xie, X. Kong, F. Xing, F. Liu, H. Su, and L. Yang, "Deep voting: A robust approach toward nucleus localization in microscopy images," 2015, p. 374.

[11] H. Chen, Q. Dou, X. Wang, J. Qin, and P. A. Heng, "Mitosis detection in breast cancer histology images via deep cascaded networks," in *Thirtieth AAAI Conference on Artificial Intelligence*, 2016, pp. 1160–1166.

[12] Y. Xie, F. Xing, X. Kong, H. Su, and L. Yang, "Beyond classification: Structured regression for robust cell detection using convolutional neural network," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2015, pp. 358–365.

[13] K. Sirinukunwattana, E. A. R. Shan, Y. W. Tsang, D. R. J. Snead, I. A. Cree, and N. M. Rajpoot, "Locality sensitive deep learning for detection and classification of nuclei in routine colon cancer histology images," *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1196–1206, 2016.

[14] M. N. Kashif, E. A. R. Shan, K. Sirinukunwattana, M. Arif, and N. Rajpoot, "Handcrafted features with convolutional neural networks for detection of tumor cells in histology images," in *IEEE International Symposium on Biomedical Imaging*, 2016, pp. 1029–1032.

[15] H. Wang, A. Cruz-Roa, A. Basavanhally, H. Gilmore, N. Shih, M. Feldman, J. Tomaszewski, F. Gonzalez, and A. Madabhushi, "Mitosis detection in breast cancer pathology images by combining handcrafted and convolutional neural network features," *J Med Imaging*, vol. 1, no. 3, p. 034003, 2014.

[16] Y. Zhou, Q. Dou, H. Chen, J. Qin, and P. A. Heng, "Sfcn-opi: Detection and fine-grained classification of nuclei using sibling fcn with objectness prior interaction," 2017.

[17] A. Shrivastava, A. Gupta, and R. Girshick, "Training region-based object detectors with online hard example mining," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 761–769.

[18] K. K. Sung, *Learning and example selection for object and pattern detection*. Massachusetts Institute of Technology, 1996.

[19] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *IEEE Computer Society Conference on Computer Vision Pattern Recognition*, 2001, p. 511.

[20] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," pp. 21–37, 2015.

[21] P. Felzenszwalb, R. Girshick, D. Mcallester, and D. Ramanan, "Visual object detection with deformable part models," *Communications of the Acm*, vol. 56, no. 9, pp. 97–105, 2013.

[22] T. Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," *IEEE Transactions on Pattern Analysis Machine Intelligence*, vol. PP, no. 99, pp. 2999–3007, 2017.

[23] T. Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 936–944.

[24] C. Y. Fu, W. Liu, A. Ranga, A. Tyagi, and A. C. Berg, "Dssd : Deconvolutional single shot detector," 2017.

[25] A. Shrivastava, R. Sukthankar, J. Malik, and A. Gupta, "Beyond skip connections: Top-down modulation for object detection," 2016.

[26] L. C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected crfs," *Computer Science*, no. 4, pp. 357–361, 2015.

[27] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE Transactions on Pattern Analysis Machine Intelligence*, vol. 39, no. 4, pp. 640–651, 2017.

[28] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for scene segmentation," *IEEE Transactions on Pattern Analysis Machine Intelligence*, vol. PP, no. 99, pp. 2481–2495, 2017.

[29] H. Noh, S. Hong, and B. Han, "Learning deconvolution network for semantic segmentation," in *IEEE International Conference on Computer Vision*, 2015, pp. 1520–1528.

[30] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2015, pp. 234–241.

[31] G. Lin, A. Milan, C. Shen, and I. Reid, "Refinenet: Multi-path refinement networks for high-resolution semantic segmentation," 2016.

[32] J. Fu, J. Liu, Y. Wang, and H. Lu, "Stacked deconvolutional network for semantic segmentation," 2017.

[33] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," pp. 6230–6239, 2017.

[34] L. C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," 2018.

[35] Y. Wei, H. Xiao, H. Shi, Z. Jie, J. Feng, and T. S. Huang, "Revisiting dilated convolution: A simple approach for weakly- and semi- supervised semantic segmentation," 2018.

[36] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2921–2929.

[37] Y. Yuan, H. Failmezger, O. M. Rueda, H. R. Ali, S. Grf, S. F. Chin, R. F. Schwarz, C. Curtis, M. J. Dunning, and H. Bardwell, "Quantitative image analysis of cellular heterogeneity in breast tumors complements genomic profiling," *Science Translational Medicine*, vol. 4, no. 157, p. 157ra143, 2012.

[38] J. Xu, L. Xiang, Q. Liu, H. Gilmore, J. Wu, J. Tang, and A. Madabhushi, "Stacked sparse autoencoder (ssae) for nuclei detection on breast cancer histopathology images," *IEEE Transactions on Medical Imaging*, vol. 35, no. 1, pp. 119–130, 2016.