2021 14th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI) | 978-1-6654-0004-6/21/$31.00 ©2021 IEEE | DOI: 10.1109/CISP-BMEI53629.2021.9624420

Cervical Cell Image Classification Based On Multiple Attention Fusion

Xin Su
School of Computer Science and Information Engineering
Hefei University of Technology
Hefei, China

Jun Shi
School of Software
Hefei University of Technology
Hefei, China

Yusheng Peng
School of Computer Science and Information Engineering
Hefei University of Technology
Hefei, China

Liping Zheng
School of Computer Science and Information Engineering
Hefei University of Technology
Hefei, China

Abstract—**Accurate classification of cervical cells is of great significance to the detection and treatment of cervical cancer. Over the years, convolutional neural network (CNN) has been successfully applied to cervical cell classification. Recently, the attention mechanism has been the research focus, which can learn local discriminant features. To further improve the performance of cervical cell image classification, we propose a novel cervical cell image classification method based on multiple attention fusion in this paper. Specifically, the Squeeze and Excitation (SE) and Spatial Attention Module (SAM) blocks are fused to learn the dependency between features from the channel and spatial directions respectively. In order to capture the long-range dependencies between features, the features embedded with SE and SAM are further fused with Disentangled Non-Local block (DNL). Experimental results on the publicly available cervical cell dataset SIPaKMeD show the effectiveness of our method.**

Keywords- Cervical cancer screening; Cervical cell classification; Attention mechanism; Feature fusion.

I. INTRODUCTION

Cervical cancer is one of the most common gynecological malignancies and the fourth most common cancer among women in the world [1]. Tumor pathology studies [2,3] have shown that if cervical cancer can be detected at an early stage it can be effectively cured. Therefore, regular screening for cervical cancer can significantly reduce the incidence and mortality of cervical cancer. Cervical cytology is one of the most common screening tests for prevention and early detection of cervical cancer. The pathologists are required to observe the cells on the smear under a microscope, and then identify positive cells, finally make a diagnosis empirically. Due to the large number of cervical cytology images, manual screening diagnosis is generally inefficient. Therefore, computer-aided diagnosis (CAD) is becoming an important tool in cervical cancer screening, which can improve the diagnosis efficiency of pathologists.

With the development of deep learning, convolutional neural networks (CNN) have been extensively used in CAD and especially in the field of cervical cytology analysis. Wieslander et al. [4] used VGG [5] and ResNet [6] to classify cervical cell images into benign cells and malignant cells. Plissiti et al. [7] proposed an annotated cervical cell image dataset SIPaKMeD, and applied VGG19 to classify five types of cervical cells. Zhang et al. [8] adapted a fine-tuned AlexNet for the cervical cell classification. CNN can only extract global features of images, but local features of cell images are also important for discriminant feature representation of cells. Therefore, the attention mechanism is introduced to explore the local features in the process of CNN feature learning. Hu et al. [9] proposed a channel-based attention mechanism squeeze and excitation (SE) block to capture dependencies between channels. Wang et al. [10] introduced a non-local operation for finding long-range dependencies. Compared with non-local, Yin et al. [11] presented disentangled non-local (DNL) block. By decoupling the standard non-local modules, the learning ability of pairwise term and unary term is improved, and the feature recognition ability of feature maps is enhanced.

In this paper, we apply multiple attention fusion for cervical cell image classification. In order to further enhance attention performance from different directions, we use three types of attention modules. The SE block is adapted to capture the dependence between channels and learn the importance of different feature channels adaptively. In addition to SE block, we also design SAM block to learn feature dependencies between locations in different spatial directions. The self-attention module disentangled non-local (DNL) block is also applied to enhance the ability of exploiting long-range dependencies between features. Finally, we connect SE and SAM blocks in series, and then fuse with DNL block. By combing three types of attention blocks, the network can learn more discriminant feature representation more usefully. The publicly available cervical cell dataset SIPaKMeD [7] is employed to evaluate our method, and experimental results show our method has better performance in cervical cell image classification.

The proposed method has the following two contributions:

1.The multiple attention fusion is applied for cervical cell image classification. The SE block is applied to adaptively learn the importance of different feature channels, and the SAM block is used to catch the spatial dependencies of features, finally DNL block is introduced to learn long-range dependencies between features.

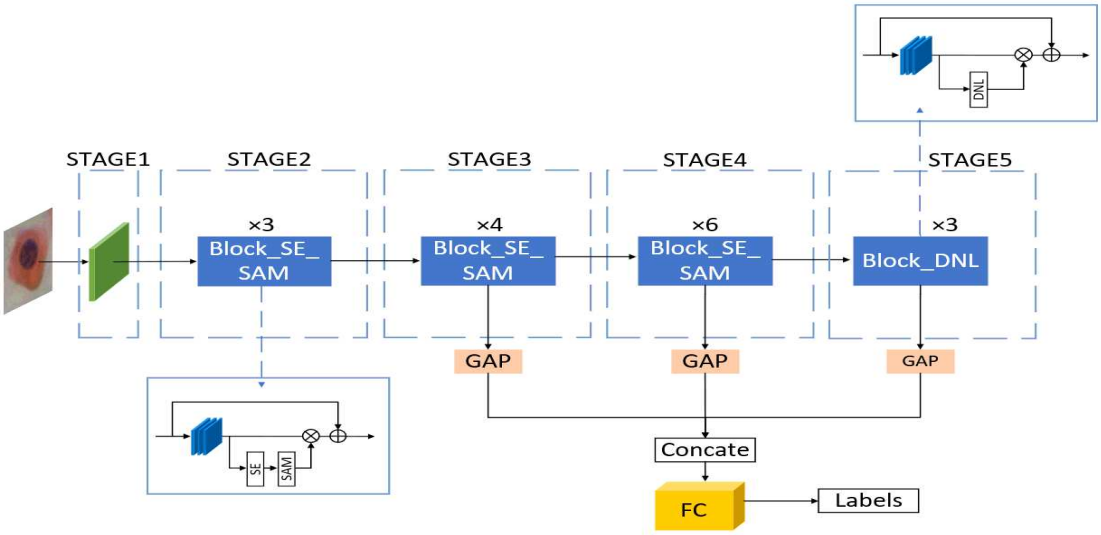

Figure 1.The pipeline of our method.

2. Our method is validated on the public cervical cell image SIPaKMeD dataset and the results show its effectiveness.

The reminder of this paper is organized as follows: Section 2 describes our method in detail. Section 3 presents the experimental results and analysis. Section 4 summarizes the conclusions.

II. METHODOLOGY

The pipeline of our method is shown in Figure 1. First, the images of cervical cells are fed into ResNet50[6], which is composed of 5 stages. The first stage is only composed of 7×7 convolution layers, and the remaining stages are composed of residual blocks. The SE and SAM blocks are fused in series and embedded into the first to third residual block of the network. In order to capture long-range dependencies between spatial pixels and reduce the computation complexity, DNL block [11] is embedded in the last residual block of ResNet. The feature maps generated by Stages 3-5 are stretched into feature vectors by global average pooling (GAP). All the generated feature vectors are concatenated to get the final feature vectors, and then inject into the fully-connected (FC) layer to get the prediction results.

A. Channel Attention generated by SENet

The core of SE [9] block is to learn the dependence relationship between channel-wise features. As shown in Figure 2, SE is composed of two-step operation: Squeeze and Excitation. The Squeeze operation is a global average pooling (GAP), which is used to aggregate cross-spatial features and obtain channel-wise feature responses. The weights for each channel of the feature map are learned through the excitation operation. A statistic z can be obtained by Squeeze operation which describes the global spatial information corresponding to the specific channel, and the formula is defined as

$$z = F_{sq}(u_c) = \frac{1}{H \times W}\sum_{i=1}^{H}\sum_{j=1}^{W} u_c(i,j). \quad (1)$$

where H and W represent the height and width of the feature maps in spatial dimensions respectively, and u_c *(i, j)* represents the pixel values of different channels in the position *(i, j)* of the image. The formula for the Excitation operation to calculate the activation S is represented as

$$S = F_{ex}(z, W) = \sigma\big(g(z,w)\big) = \sigma\big(W_2\delta(W_1 z)\big). \quad (2)$$

where σ means sigmoid activation function, and δ represents the ReLu function. W_1 and W_2 refer to the parameters of full connection layers. The Scale is the operation of multiplying the weights of the learned channels by the original features. The formula for the Scale operation is defined as

$$u'_c\,(i,j) = S * u_c(i,j). \quad (3)$$

where u_c refers to the original feature maps, and u'_c means the final feature maps.

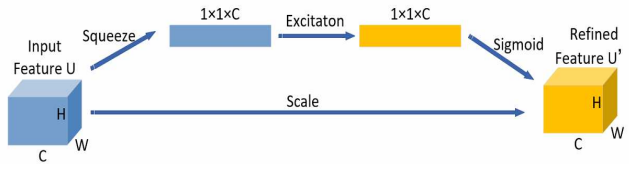

Figure 2. The structure of SE block.

B. Spatial Attention generated by SAM

The spatial attention module SAM learns the spatial dependencies of features by aggregating feature information of channel domains. The structure of SAM can be seen in Figure 3. Firstly, GAP is used to aggregate the channel features of the input feature maps, the formula for the GAP operation along channel direction is defined as

$$x(i,j) = F_{GAP}(u_c) = \frac{1}{C}\sum_{c=1}^{C} u_c(i,j). \quad (4)$$

where C refers to the channel of the feature maps, and x *(i, j)* refers to the feature at the position *(i, j)*. Then the 1×1 convolution is used to learn the weights G of the attention map, the formula of 1×1 Conv can be described as

$$G = F_{Conv} = Mx(i,j). \quad (5)$$

where M is a weight matrix learned by 1×1 convolution. The formula of Scale is defined as

$$u'_c\,(i,j) = G * u_c\,(i,j). \quad (6)$$

where u_c refers to the input feature maps, and u'_c means the final refined feature maps.

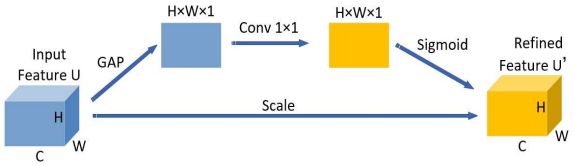

Figure 3. The structure of SAM block.

C. *Self-Attention generated by DNL*

The Disentangled Non-local (DNL) block [11] is an improvement on the non-local module. Compared with non-local, DNL decouples the pairwise term and unary term of the original non-local module by using independent SoftMax function and embedded matrix. The difficulty of the two terms joint learning is reduced, and the learning ability of each term for image features is significantly improved. The structure of DNL block is shown in Figure 4. In pairwise term, the feature map is fed into two parallel 1×1 convolution followed by a corresponding whiten operation which eliminate the impact of shared parameters on pairwise term. Then SoftMax is used to obtain the output features after the dot product. Meanwhile, the unary term uses an independent 1×1 convolution connected with SoftMax to obtain the normalized feature map. The output feature maps are transformed to features with the same size of the pairwise term by an expand operation. The outputs of pairwise term and unary term are added to get the attention response. Furthermore, 1×1 convolution is used to calculate the features at different positions of feature maps. The generated features and the attention information generated by paired and unary terms are multiplied to get the attention maps. Finally, the weighted feature maps are obtained by adding the original feature maps with attention maps.

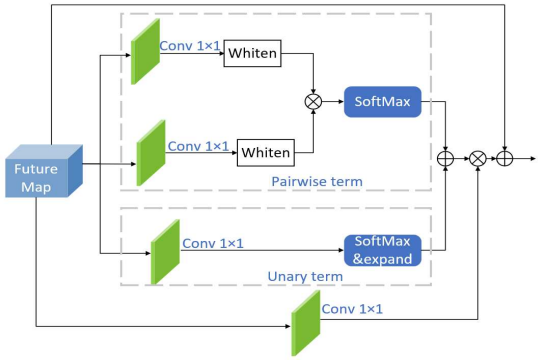

Figure 4. The structure of DNL block.

D. *Feature Fusion*

We embed the SE and SAM blocks in Stages 2-4 of the network. Both of them are fused together in series. For capturing long-range dependencies between spatial pixels, the DNL block is used in the last stage. In order to enhance the ability of feature representation, feature fusion is performed on the last three stages. Concretely, we use GAP to transform feature maps of Stages 3-5 into feature vectors. The generated feature vectors are concatenated to get the final feature vectors of cell images.

III. EXPERIMENT

In the experiment, the public cervical cell image dataset SIPaKMeD [7] is used to verify the classification accuracy of our method. It contains 4,049 annotated images of cells that are divided into five categories: Superficial Intermediate, Parabasal, Koilocytotic, Dyskeratotic and Metaplastic. Figure 5 shows example images of five types of cells.

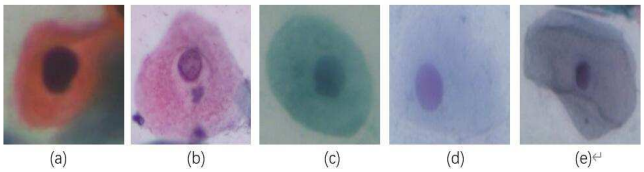

Figure 5. Cell images of five categories: (a)Dyskeratotic, (b)Koilocytotic, (c)Metaplastic, (d)Parabasal, (e)Superficial-Intermediate.

Our method is compared with six classification methods, namely ResNet [6], the method proposed by Plissiti et al. [7], ResNet50 embedded in SE block (denoted as SE), ResNet50 embedded in SAM block (denoted as SAM), ResNet50 embedded in DNL block (denoted as DNL), and SE and SAM blocks are connected in series and embedded in ResNet50 (denoted as SE + SAM). Follow the practice [7], we apply 5 training-test fold cross-validation to evaluate classification performance. Each cell image is adjusted to 80×80 pixels and augmented by flipping each cell image horizontally, vertically and in both directions. The batch size is set to 48 and the network is trained for 80 epochs by stochastic gradient descent (SGD). Cross-entropy loss is used to train the model. All the experiments are conducted on a computer with an Intel Core i7-9700X CPU of 3.60 GHz and a GPU of NVIDIA GTX 2080Ti.

Mean/standard deviation classification accuracies of the seven methods over 5 training-test folds are given in Table I. Obviously, ResNet50 [6] slightly outperforms the method proposed by Plissiti et al. [7]. Compared with ResNet50, the attention-based methods (i.e., SE, SAM and DNL) have better classification results. It can be explained that attention-based methods can extract local discriminant features, which is more conducive to cell classification. The classification results of individual-attention method are not as ideal as the methods based on attention fusion. The classification accuracy has been significantly improved by fusing the SE and SAM blocks, and feature dependencies in different directions can be learned effectively. Our method has more desirable classification accuracy than other attention-based methods. Based on SE and SAM fusion, our method uses DNL block to capture the long-range dependencies between features. The

fusion of three types of attention modules can effectively enhance the ability of feature representation.

TABLE I. COMPARISON OF SEVEN CLASSIFICATION METHODS ON SIPAKMED DATASET.

Method	Accuracy(%)
Plissiti et al. [7]	95.35±0.42
ResNet50	96.17±0.6
SE	98.23±0.3
SAM	97.17±0.2
DNL	97.90±0.3
SE + SAM	98.72±0.3
Ours	99.38±0.2

We also present the confusion matrices of these above methods as shown in Figure 6. As for the analysis of ResNet50 [6] without any attention module shown in Figure 6(a), the accuracies of the normal cells (i.e., Superficial/Intermediate and Parabasal) exceed 97%. The accuracies on the abnormal cells (i.e., Koilocytotic and Dyskeratotic) exceed 91%. The classification results of SE [9] on the normal cells exceed 99% shown in Figure 6(b). For the results of SAM block shown in Figure 6(c), the accuracies of the abnormal cells reach 93% and 96.64% respectively, which is higher than ResNet50. As shown in Figure 6(d), DNL has a better classification result for Parabasal cell than ResNet50. The classification results of SE + SAM are better than individual-attention methods. Particularly the classification results of all types of cells reach 98%, which can be seen in Figure 6(e). It indicates that SE and SAM blocks fusion can learn feature correlation of channel and spatial simultaneously, which leads to the better classification effect. However, the misclassification rate of Koilocytotic cell is 1.69%, and it is most easily misclassified into Metaplastic cell (1.08%). As displayed in Figure 6(f), our method has more desirable classification performance, the accuracies of all types of cells exceed 99%. This shows that multiple attention fusion enables the network to capture feature dependencies from channel, spatial and contextual directions.

IV. CONCLUSION

In this paper, we propose a new method for cervical cell image classification based on multiple attention fusion. It uses ResNet50 as the backbone network to extract deep cell features. The SE and SAM blocks are connected in series to capture feature correlation of feature maps from channel direction and spatial direction respectively. Finally, DNL block is fused to further learn the long-range dependencies between spatial pixels. Experiments on the SIPaKMeD dataset show that our method has better classification performance.

ACKNOWLEDGEMENT

This work was partly supported by the National Natural Science Foundation of China (grant no. 61906058, 61972128), partly supported by the Anhui Provincial Natural

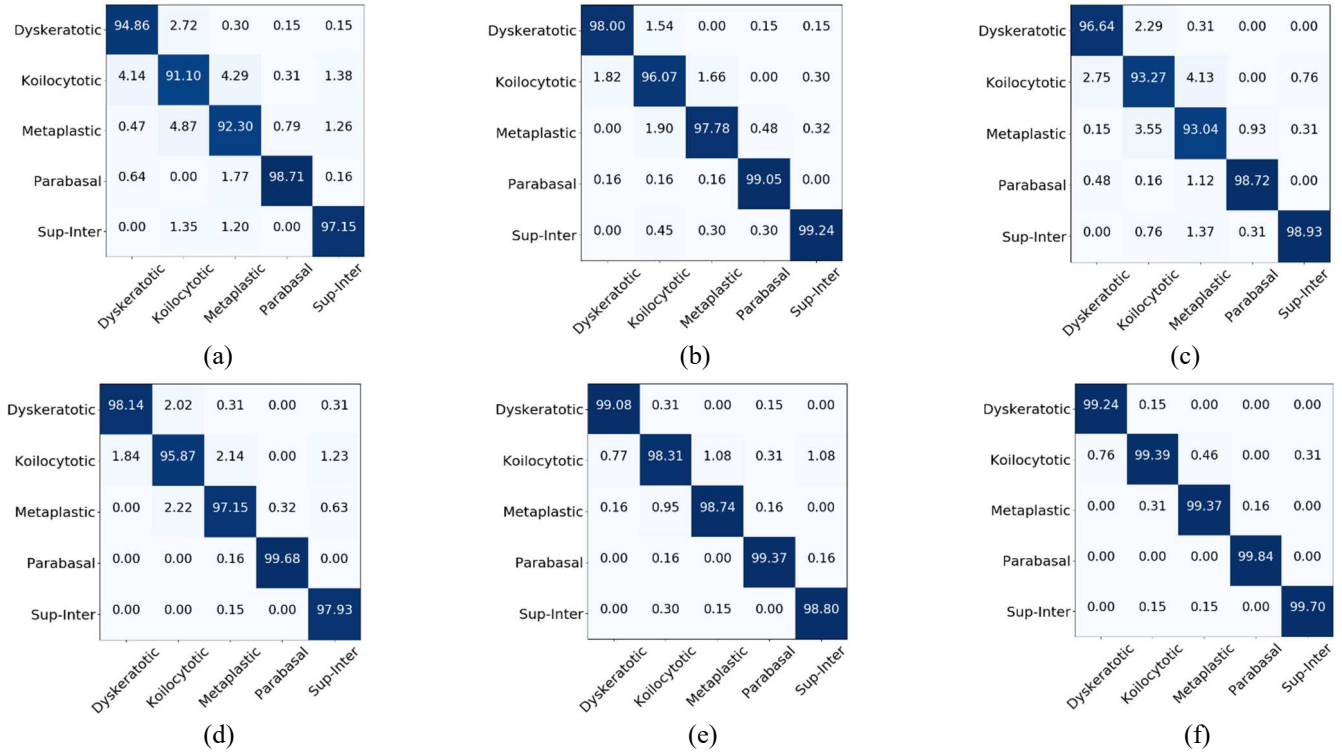

Figure 6. Confusion matrices using different attention for classification on SIPaKMeD dataset. (a) ResNet50. (b)SE. (c)SAM. (d) DNL. (e) SE + SAM. (f)Ours.

Science Foundation (grant no. 1908085MF210), and partly supported by the Fundamental Research Funds for the Central Universities of China (grant no. JZ2020YYPY0093).

REFERENCES

[1] A. Jemal, F. Bray, M. Center, J. Ferlay, E. Ward, and D. Forman, "Global cancer stat," CA Cancer J Clin, vol. 61, Jan 2011,pp. 69–90.

[2] T. Potapova, J. Zhu, and R. Li, "Aneuploidy and chromosomal instability: A vicious cycle driving cellular evolution and cancer genome chaos," Cancer metastasis reviews, vol. 32, May 2013, pp. 377–389.

[3] D. Garner, "Clinical application of dna ploidy to cervical cancer screening: A review," World journal of clinical oncology, vol. 5, Dec. 2014, pp. 931–965.

[4] I. S. Jacobs and C. P. Bean, "Fine particles, thin films and exchange anisotropy," in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.

[5] M. E. Plissiti, P. Dimitrakopoulos, G. Sfikas, C. Nikou, O. Krikoni, and A. Charchanti, "Sipakmed: A new dataset for feature and image based classification of normal and pathological cervical cells in pap smear images," in 2018 25th IEEE International Conference on Image Processing (ICIP), 2018, pp. 3144–3148.

[6] A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv, Sep 2014 ,pp. 1409-1556.

[7] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 37, no. 9, 2015, pp. 1904‑1916.

[8] A. Krizhevsky, I. Sutskever, and G. Hinton, "Imagenet classification with deep convolutional neural networks," Neural Information Processing Systems, vol. 25, Jan 2012.

[9] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," IEEE conference on computer vision and pattern recognition (CVPR), 2018, pp. 7132–7141.

[10] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," IEEE conference on computer vision and pattern recognition (CVPR), 2018, pp. 7794–7803.

[11] M. Yin, Z. Yao, Y. Cao, X. Li, Z. Zhang, S. Lin, and H. Hu,"Disentangled non-local neural networks," in European Conference on Computer Vision (ECCV), 2020, pp. 191–207.