IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received May 19, 2022, accepted June 7, 2022, date of publication June 13, 2022, date of current version June 23, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3182800

MSS-WISN: Multiscale Multistaining WBCs Instance Segmentation Network

MENG ZHAO[1], HONGXIA YANG[1], FAN SHI[1], (Member, IEEE), XINPENG ZHANG[1], YAO ZHANG[1], XUGUO SUN[2], AND HAO WANG[3], (Senior Member, IEEE)

[1]Key Laboratory of Computer Vision and System (Ministry of Education of China), Engineering Research Center of Learning-Based Intelligent System (Ministry of Education of China), School of Computer Science and Engineering, Tianjin University of Technology, Tianjin 300382, China
[2]School of Medical Laboratory, Tianjin Medical University, Tianjin 300070, China
[3]Department of Computer Science, Norwegian University of Science and Technology, 7491 Trondheim, Norway

Corresponding authors: Xinpeng Zhang (xpzhang1989@outlook.com) and Yao Zhang (zytju221@tju.edu.cn)

This work was supported by the National Natural Science Foundation of China under Grant 61903275, Grant 61902078, Grant 62020106004, Grant 92048301, and Grant 61906133.

ABSTRACT Accurate segmentation and detection (instance segmentation) of white blood cells (WBCs) from whole slide images remains a challenging task, as the WBCs vary widely in shapes, sizes, and colors caused by different cell subtypes and various staining techniques. In this paper, we propose a novel framework for end-to-end segmentation and detection of WBCs that are on multiple scales and stained by different techniques. We name the framework the multi-scale and multi-staining WBC instance segmentation network (MSS-WISN). The MSS-WISN consists of two parts: 1) a feature extraction network for strengthening the feature expression and minimizing the impact of different staining techniques, and 2) a feature fusion network for highlighting salient features and thereby eliminating the effect of scale variations. To verify the effectiveness of the MSS-WISN, we build a new dataset containing 302 Magenta stained images (collected by Tianjin Medical University) and 242 Wright stained images (from a public dataset). Experiments show that the proposed framework outperforms other state-of-the-art methods in terms of WBC detection and WBC segmentation, achieving the highest F1-Score (0.901) and Dice (0.902).

INDEX TERMS White blood cells, instance segmentation, strengthened feature expression, highlighted salient features.

I. INTRODUCTION

White blood cells (WBCs) plays a pivotal role in human immune system, and the WBCs total count or each subtpye count provides significant indicator for human health [1], [2]. Neutrophils, lymphocytes, monocytes, eosinophils, and basophils are the five subtypes of WBCs in the blood with decreasing amounts in normal human bodies [3]. Each subtype has its physiological function. Thus, accurate classification and segmentation of WBCs from whole slide images are essential foundations for clinical examination [3], [4]. To accomplish these tasks, the whole slide images are usually stained in preparation, which on one hand aids the cells in showing their characteristics, but on the other hand, can potentially hamper the effectiveness of image analysis due to color and intensity variations of staining [5].

The associate editor coordinating the review of this manuscript and approving it for publication was Kumaradevan Punithakumar.

Magenta and Wright staining are the two most commonly used methods for WBC whole slide images [6]. The WBCs stained by the two methods are presented in Fig. 1. The original cell sizes in Wright stained images are larger than those of Magenta stained. Note that there are only three subtypes in the Magenta stained images: neutrophil (neu), lymphocyte (lym), and monocyte (mon) since the other two subtypes have lost their activity in this staining.

Manual detection, segmentation, and classification of WBCs from a whole slide image are time-consuming, laborious, subjective, and fallible, especially when the WBCs have unclear texture or structure. Computer-aided approaches thus emerged as new automatic solutions [7]. Deep learning-based methods have achieved state-of-the-art performance in WBCs classification and segmentation [8]. But many challenges still exist. For example, segmentation and classification are always performed independently since ensembling the two procedures into a single framework

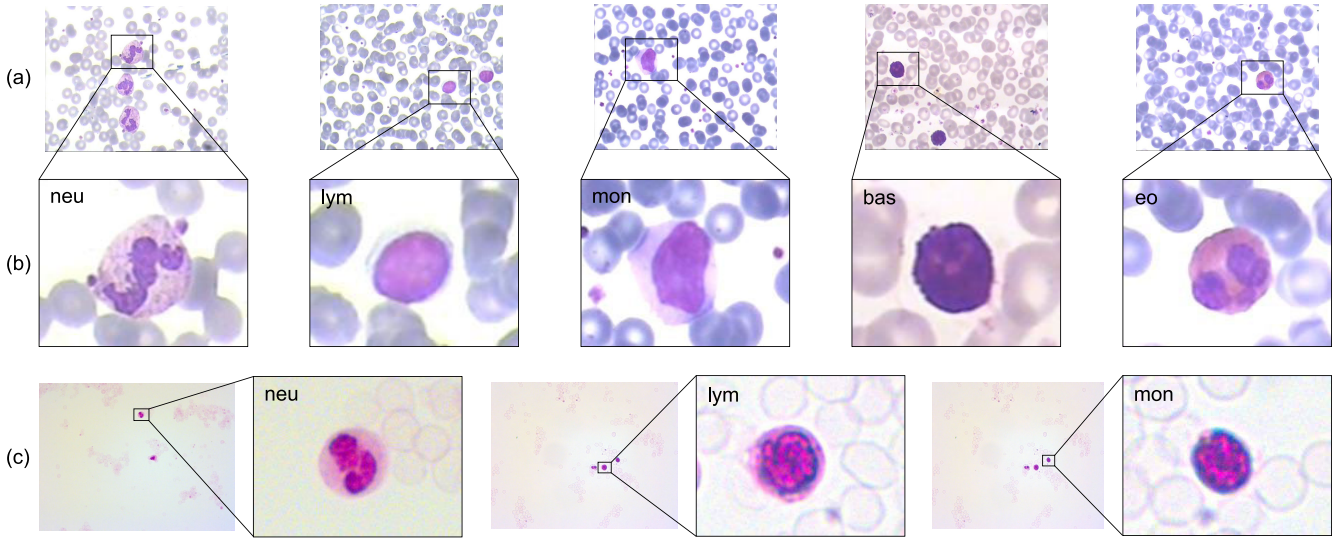

FIGURE 1. Five subtypes of WBCs under Wright (a and b) and Magenta (c) staining, respectively. In (a) and (b), there are five subtypes of WBCs: neu, lym, mon, bas and eo; while in (c), there are three subtypes of WBCs: neu, lym and mon.

requires an extra step to classify individual instances [9]. Moreover, the robustness of instance segmentation against various cell sizes and staining techniques is difficult to achieve [10].

To tackle the above-mentioned challenges, in this paper, we propose a framework, named multi-scale and multi-staining WBC instance segmentation network (MSS-WISN, MSS for short), to realize simultaneous detection, segmentation, and classification of WBCs from whole slide images. MSS contains a feature extraction module and a feature fusion module to address the problems caused by different staining techniques and multiple scales respectively. We also establish a dataset with 544 images and manually labeled them with segmentation masks and classification subtypes.

The main contributions of our method are as follows:

1. We build a new dataset composed of 544 WBC images with classification labels and masks: 302 Magenta stained images with three subtypes containing 322 neutrophils, 175 lymphocytes, and 93 monocytes; 242 Wright stained images with five subtypes containing 50 neutrophils, 52 lymphocytes, 48 monocytes, 53 basophils, and 39 eosinophils. The Magenta stained images were self-collected by Tianjin Medical University and the Wright stained images came from Leukocyte Images for Segmentation and Classification (LISC) [11] database.

2. We propose a new framework that realizes simultaneous detection, segmentation, and classification of WBCs on the above-mentioned dataset. The framework can strengthen feature expression and highlight salient features, which handles the difficulties caused by different staining methods and multi-scale sizes of WBCs.

The rest of this paper is organized as follows. In Section II, we review the methods used for the detection, segmentation, and classification for WBCs in recent years. Section III introduces the proposed network for cell instance segmentation. Section IV presents experimental results. Finally, Section V concludes the whole work and discusses emerging trends in WBC instance segmentation.

II. RELATED WORK

A. DETECTION AND SEGMENTATION OF CELLS AND NUCLEI

Fruitful results have been reported on detection and segmentation of cells and nuclei. Some traditional algorithms for detection include: distance transform (DT) [12], morphology operation [13], h-minima transform (HIT) [14], maximally stable extremal region (MSER) detection [15] and so on. For segmentation, clustering techniques were employed as powerful tools [16]–[18]. On the other hand, there are plenty of works investigating the image characteristic for segmentation, for example, edge boundary [19], colour space [6], [20], [21], and threshold [22]–[24] were separately studied. Zhang *et al.* [6] combined several traditional methods including color space decomposition and k-means clustering for segmentation, which achieved satisfying accuracy. Hannah *et al.* [25] developed the so-called soft covering rough k-means clustering (HSCRKM) method for leukemia and nucleus segmentation by combining the advantages of a soft covering rough set and the rough k-means clustering. Lu *et al.* [15] proposed a method based on accurate nucleus detection which, however, was easy to be affected by impurities or unrelated substances.

Deep learning-based methods have been used to solve specific challenges such as overlapping and blurred boundary segmentation. Self-supervised learning methods and Convolutional Neural Networks (CNNs) [26] are frequently involved. For example, in [27], a self-supervised learning method based on the topological structure of leukocytes and fuzzy boundary enhancement was used to segment WBCs. Two CNN-based object detection methods, the SSD (Single

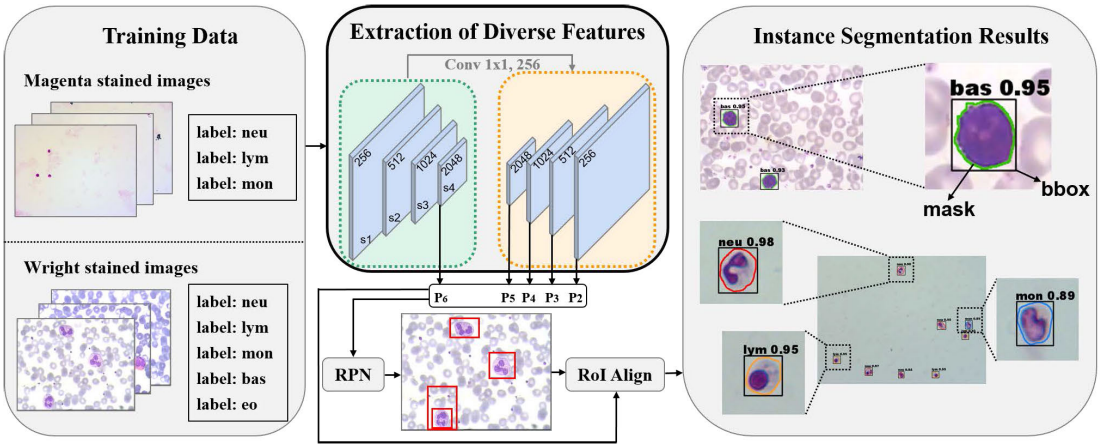

FIGURE 2. The overall architecture of the proposed framework-MSS. The original whole slide WBC images stained by Magenta or Wright are input into the network, followed by the extraction of the diverse features module, then the obtained feature maps P2-P6 are fed into the detection module to implement multiple tasks: bounding box prediction, subtype classification, and mask generation.

Shot Multibox Detector) and the YOLOv3 (You Only Look Once), were utilized in [28] for detecting WBCs. In [29], the authors successfully segmented the overlapped cervical cells in Pap smear images by resorting to the Mask R-CNN. However, none of the aforementioned methods considered the difficulties caused by multi-scale cells and different staining ways.

B. CLASSIFICATION OF CELLS AND NUCLEI

Some methods to realize both segmentation and classification have also come into being. In [30], a system was proposed to automatically segment, count, and classify the WBC of five types, but the segmentation accuracy on the cell boundary was limited. In [31], the authors put forward a machine-learning algorithm to detect and classify immature leukocytes and used the traditional random forest algorithm to classify immature WBCs based on morphological features. In [32], a simple CNN was utilized to identify and locate WBCs in Wright stained blood cell images. To resolve the problems of intra-nucleus variation and segregation of aggregated nucleus, in [33], a UNet-based architecture was integrated with residual blocks, densely connected blocks, and a fully convolutional layer in the encoder-decoder block, achieving better classification than the UNet or the Mask R-CNN. However, this method required two kinds of mask representation, which raised the complexity of data preprocessing.

C. MASK PREDICTION OF CELL AND NUCLEI

In addition to the segmentation and the classification, the general instance segmentation involves predicting a mask for each WBCs. There are two mainstream models: one-stage models and two-stage models. The one-stage model has the advantages of few parameters and economic training time, but its accuracy is lower than the level of the two-stage model [27], [34] since it, unlike the Region Proposal Network (RPN)-based methods [35], does not filter out the background and most negative samples. Mask R-CNN [27] is a popular two-stage instance segmentation model and sparks numerous variants, see examples in [29], [36], [37]. These Mask R-CNN variants are aimed at improving the general performance of instance segmentation by using specific characteristics of images.

Multi-scale cell instance segmentation has been considered in some of the above references. For example, in [38], a new box-based cell instance segmentation method was proposed using a keypoint graph to extract the bounding box for each cell. The highest average precision reached 0.88. Different staining techniques can result in distinctive feature representations for the same cell. The feature difference should be removed such that the original features of cells can be exhibited. ResNeXt [39] realized the split and reorganization of feature channels so that it could fuse feature maps from different subspaces, being an option for preserving the essential features regardless of staining methods.

III. METHOD

In this section, we present the proposed MSS framework in detail. In what follows, the overall architecture of MSS is given in Section III-A. Then, all the involved modules, i.e., the Se-ResNeXt, the CFPN, and the detection network are elaborated in Sections III-B, III-C, III-D respectively.

A. OVERALL ARCHITECTURE

The proposed framework consists of three parts: dataset construction, diverse features extraction, and WBCs detection and segmentation. The general flowchart is given in Fig. 2. Firstly, we construct a dataset of WBCs whole slide images. The details of the dataset can be seen in Section IV. Then, we introduce a novel features extraction network, say, the Se-ResNeXt, to extract diverse feature representations of the WBCs. The resulting feature maps which are denoted by

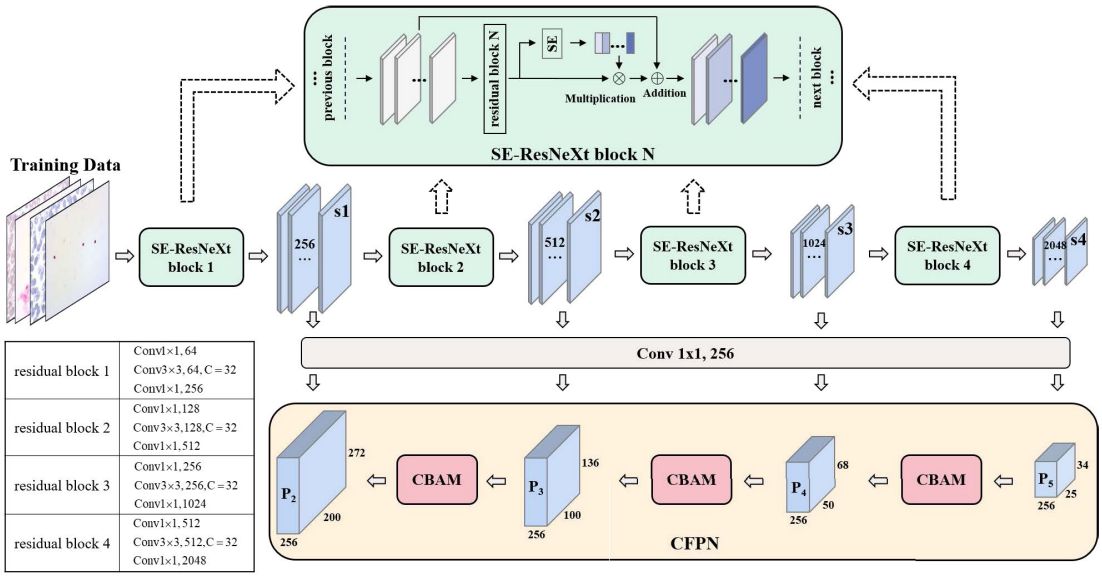

FIGURE 3. Detailed structure of diverse features extraction module.

s1, **s2**, **s3**, **s4** in Fig. 2 contain the deep semantic and the shallow information of the original images and are of multi-scales. To integrate the high-level and low level features maps **s1**, **s2**, **s3**, **s4**, a feature pyramid network (FPN) equipped with attention modules–Convolution block attention modules (CBAM)–is employed, which is called the CFPN. The details of the Se-ResNeXt and the CFPN are separately elaborated in Section III-B and Section III-C. At last, a WBCs detection network is developed taking the obtained feature maps as inputs, realizing three tasks simultaneously: bounding box prediction, subtype classification, and mask generation.

B. SE-ResNeXt FOR STRENGTHENING FEATURE REPRESENTATION OF GLOBAL FEATURES

To remedy the problem caused by different staining methods, we introduce a new feature extraction network structure, namely the SE-ResNeXt network, that combines the Squeeze and Excitation (SE) module [40] and the ResNeXt [39]. It is depicted in Fig. 3. SE-ResNeXt adopts a group-convolution structure containing 32 convolutions of 3×3 dimensions to capture diverse features of the original images. To strengthen the feature expression, a channel attention mechanism is employed. There are in total four SE-ResNeXt blocks, that are SE-ResNeXt block a, $a = 1, 2, 3, 4$. In the a-th block, the corresponding group-convolution outputs a set of feature maps of $64 \times 2^{a-1}$ channels. Then through the following $1 \times 1 \times (64 \times 2^{a+1})$ convolution layer, the feature maps are expanded by four times, yielding a collection of $64 \times 2^{a+1}$-channel feature maps.

The group-convolution structure in each SE-ResNeXt generates a feature subspace of the original images, a subspace containing different levels of feature representations from each other. Therefore, combining these diverse feature representations through the SE-ResNeXt can enhance the expression of the salient features of the WBCs which are only slightly affected by the staining methods. As a result, the accuracy of instance segmentation can be maintained regardless of the staining methods.

Intuitively, some feature maps are perhaps more significant and more correlated with the WBCs segmentation performance than others. However, the classical ResNeXt, despite the common merits of the group-convolution, uniformly allocates the weights on the feature maps. Moreover, the input feature maps of each ResNeXt block, after being processed by the group-convolution, lose the connection among different channels to some extent. To reallocate the channel weights and strengthen the connection, the SE module, an attention-based structure, trained to generate attention weights for the feature channels, is inserted at the end of the residual block of each SE-ResNeXt block. The re-weighted feature maps are then added to the clean feature maps, being the inputs to the next SE-ResNeXt block. In this way, more nonlinearities are included in the SE-ResNeXt to exhibit the complex correlations among the channels with marginal extra parameters and calculations.

In the a-th SE-ResNeXt block, the global average pooling is performed on the input feature maps of the SE module to obtain a set of $1 \times 1 \times (64 \times 2^{a+1})$ vectors which contain the global channel information. Then, the vectors are fed into a convolution and nonlinear network which outputs the weights of the input feature maps' channels. Therefore, more informative channels can be assigned larger weights with the aid of the SE module.

C. CFPN FOR HIGHLIGHTING SALIENT FEATURES IN MULTI-SCALE FEATURES

As can be seen in Fig. 1, the scales of WBCs vary a lot due to different image resolutions and different staining ways. The

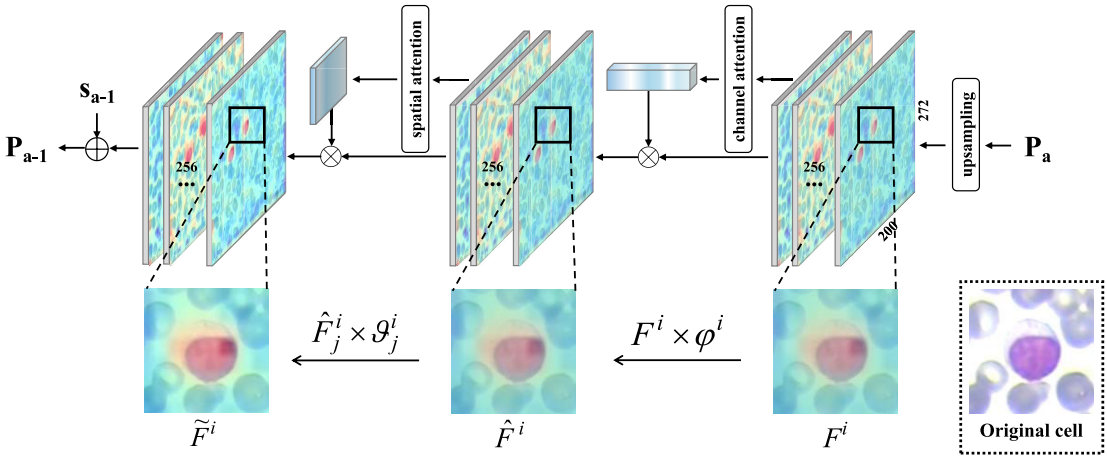

FIGURE 4. The CBAM–detailed process from P_{a+1} to P_a ($a = 2, 3, 4$).

cells in the Magenta stained whole slide images are relatively smaller than those in Wright stained images. To solve the problem of multi-scales, we use the FPN, a pyramid structure that can build and aggregate high-level semantic feature maps at all scales. But still, the FPN is incapable of putting adequate attention to the primary objects in each specific scale. The WBCs are smaller targets compared to the negative samples such as red blood cells, and the count of the WBCs is also less than that of the negative samples. As such, an imbalance problem of the WBCs dataset exists. To achieve good segmentation of the WBCs, we need to highlight the feature expression of positive samples and blur the negative samples on each given scale. Meanwhile, more attention should be paid to the positive samples to reduce the impact of data imbalance. The CBAM [41] which can integrate both the channel and the spatial attention is introduced to solve the aforementioned challenges.

The CFPN is depicted in Fig. 3. There are four layers from the up to the bottom, whose corresponding feature maps are denoted in descending order, as $\mathbf{P}_5, \mathbf{P}_4, \mathbf{P}_3, \mathbf{P}_2$. The dimensions of channels of the feature maps $\mathbf{s}_1, \mathbf{s}_2, \mathbf{s}_3, \mathbf{s}_4$ are unified as 256 through a $1 \times 1 \times 256$ convolution layer. The resulting feature maps thereby share the same channel dimension with $\mathbf{P}_2, \mathbf{P}_3, \mathbf{P}_4, \mathbf{P}_5$. The feature map $\mathbf{s}_4$, after convolution, flows directly into the CFPN structure as the up layer features $\mathbf{P}_5$. In the a-th layer, the feature maps $\mathbf{P}_{a+1}$ from the previous layer are upsampled to the same size as the laterally-connected feature maps $\mathbf{s}_{a-1}$, and then fed into the CBAM for attention-oriented operation, at last merged with $\mathbf{s}_{a-1}$, obtaining $\mathbf{P}_a$. The main process in each layer of CFPN is illustrated by the zoom-in figure in Fig. 4. After being upsampled, $\mathbf{P}_a$ is input into the CBAM module for sequentially undergoing the channel attention operation and the spatial attention operation. The channel attention aims at assigning great weights to the channels that are highly related to the salient features of the targets. On the other hand, the spatial attention tends to assign great weights to the regions of the targets. The two attention mechanisms are presented as follows.

1) CHANNEL ATTENTION

To aggregate the channel information, the input feature maps are compressed along the spatial dimensions separately using the average-pooling and the max-pooling, attaining two one-dimensional vectors. The vectors are then put into a shared network which is composed of a hidden layer and a multi-layer perceptron (MLP), to produce a set of weights for the corresponding channels. Let F^i represent the i-th channel feature map, and ϕ^i be the produced weight, then the new i-th channel feature map is derived as:

$$\hat{F}^i = F^i \times \phi^i. \tag{1}$$

2) SPATIAL ATTENTION

The spatial attention module focuses on the inter-spatial relationship of the features. The average-pooling and the max-pooling are also used for aggregating the spatial information and generating two $H_a \times W_a \times 1$ feature maps. By concatenating them along the channel dimension and putting the results into a 7×7 convolution and a nonlinear activation function, the spatial weights of the feature maps are obtained. Denote $\tilde{F}_j^i$ the i-th channel, j-th spatial position of the feature maps, and ϑ_j^i the weight assigned to it. The weighted feature map can be computed as follows:

$$\tilde{F}_j^i = \hat{F}_j^i \times \vartheta_j^i. \tag{2}$$

The evolution of a WBC through the operations of upsampling and channel-spatial attention is given in Fig. 4. The target is visually highlighted and the background is suppressed. Comparing to $\mathbf{s_{a-1}}$, the features $\tilde{F}$ refined from $\mathbf{P_a}$ contains more local information. Pixel-wisely adding $\mathbf{s_{a-1}}$ and $\tilde{F}$ yields $\mathbf{P_{a-1}}$, the input feature maps to the next layer. All the achieved feature maps $\mathbf{P}_5, \mathbf{P}_4, \mathbf{P}_3, \mathbf{P}_2$ are used for classification and segmentation of the WBCs.

D. DETECTION NETWORK

The detection network consists of a Region Proposal Network (RPN) [35] to generate a set of rectangular object proposals

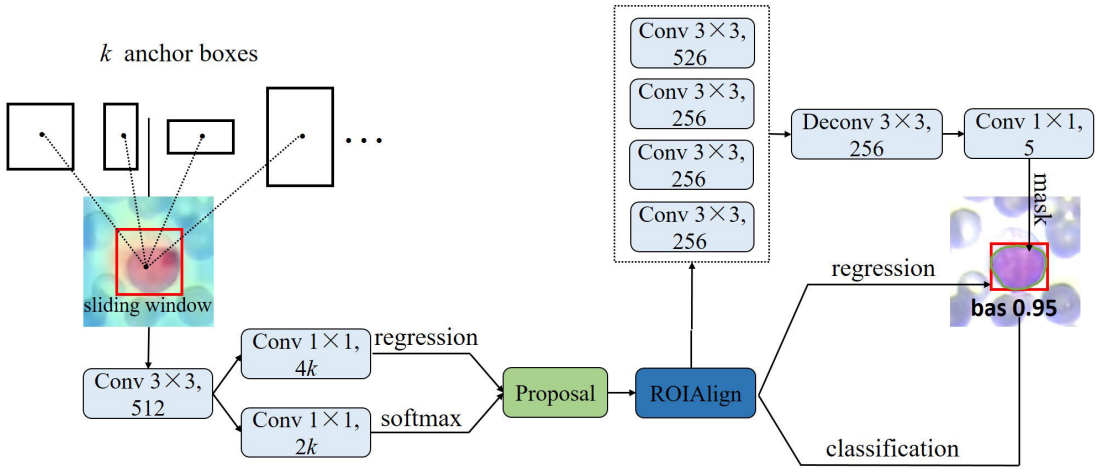

FIGURE 5. Detailed process of the detection network.

and three branches to refine the bounding box, classify the WBCs, and segment the WBCs. The flowchart is presented in Fig. 5. The feature maps $\mathbf{P}_2$, $\mathbf{P}_3$, $\mathbf{P}_4$, $\mathbf{P}_5$ and $\mathbf{s}_4$ enter the RPN through a $3 \times 3 \times 512$ convolution layer. We attach k region proposals to each spatial location that the convolutions slide by. Therefore, there are approximate $W \times H \times k$ proposals totally, where H and W are the height and width of the feature maps. Each region proposal also called an anchor, is associated with a scale and a ratio. In our setting, scales 4, 8, 16, 32, 64 and ratios 0.5, 1.0, 1.5, 2.0 are used, indicating a total of 20 ($k = 20$) anchors at each sliding location. Then the feature maps with anchors are fed into two parallel 1×1 convolutions, the one with dimension $2k$ used for classifying the proposals based on if they contain objects, and the other one of dimension $4k$ used for regressing a bounding box for each proposal. We assign a proposal positive if the proposal has an IoU bigger than 0.5 with any ground-truth box and negative otherwise. A bounding box is encoded with 4 parameters, which are the coordinate of the centroid, the height, and the width. With these definitions, the $1 \times 1 \times 2k$ convolutions combined with Softmax will output $2k$ scores for the proposals on each sliding location, predicting the proposals' classification. On the other hand, the regression layer outputs the parameters of k boxes. To train the RPN, we employ the following loss [35]:

$$L_{\text{RPN}} = \frac{1}{N_{\text{cls}}} \sum_i L_{\text{cls}}(p_i, p_i^*) + \lambda \frac{1}{N_{\text{reg}}} \sum_i p_i^* L_{\text{reg}}(t_i, t_i^*), \quad (3)$$

where

$$L_{\text{cls}}(p_i, p_i^*) = -(p_i^* \log p_i + (1 - p_i^*) \log(1 - p_i)) \quad (4)$$

is the proposal classification loss and

$$L_{\text{reg}}(t_i, t_i^*) = \text{smooth}_{L1}(t_i - t_i^*) \quad (5)$$

is the box regression loss; p_i, p_i^* represent the predicted label probability (0.0 to 1.0) and the ground-truth label (0 or 1), respectively; t_i and t_i^* denote the predicted box parameters and the ground-truth box parameters; N_{cls} represents mini-batch size; N_{reg} represents the number of anchor locations; λ is a balancing parameter.

Having acquired the positive proposals–the regions of interest (ROI)–the next stage is to in parallel predict the bounding box, the classification of the WBCs, and the segmentation for the WBCs within them. Before that, it is necessary to extract the features from the ROIs. To this we apply the RoIAlign method [27] which involves a series of operations including sampling, bilinear interpolation, and max-pooling, as a result, generating a set of ROI-feature maps with uniform size. The details of the RoIAlign can be seen in [27]. The ROI-feature maps are then mapped to two feature vectors by fully connected layers, one for classifying the WBCs and the other for predicting the bounding boxes. Besides this, a third branch containing several layers of convolutions and deconvolutions is introduced to generate masks for semantic segmentation of the WBCs. The overall loss for these three tasks is

$$L = L_{\text{cls}} + L_{\text{box}} + L_{\text{mask}}, \quad (6)$$

which is defined as identical to that given in [27].

IV. EXPERIMENTS

We conducted several experiments to illustrate the validity of the proposed framework MSS. We start with the dataset construction and the experimental settings and then present comparative studies and ablation studies.

A. EXPERIMENT SETUP

1) DATASET

We construct a WBCs dataset by integrating a set of 302 Magenta stained WBC images from Tianjin Medical University and 242 Wright stained images from the public dataset LISC. The images from the two categories have different resolutions–Magenta stained images are 2592×1944 and Wright stained images are 720×576. Most of the Magenta

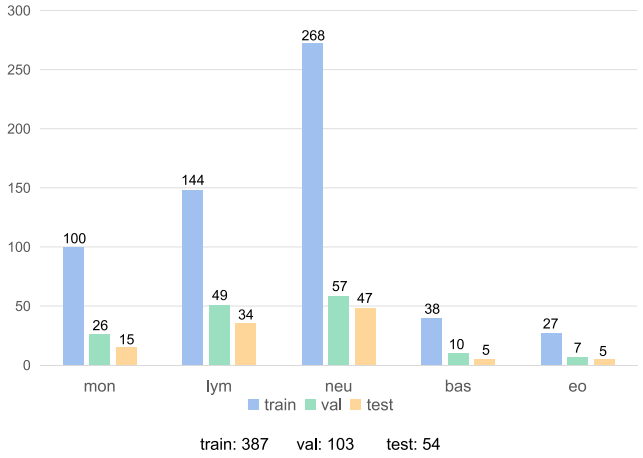

FIGURE 6. Dataset distribution. The portion of five subtypes of WBCs is about 3 : 4 : 7 : 1 : 1 (mon:lym:neu:bas:eo). The ratio of the training, validation, and testing is about 7 : 2 : 1.

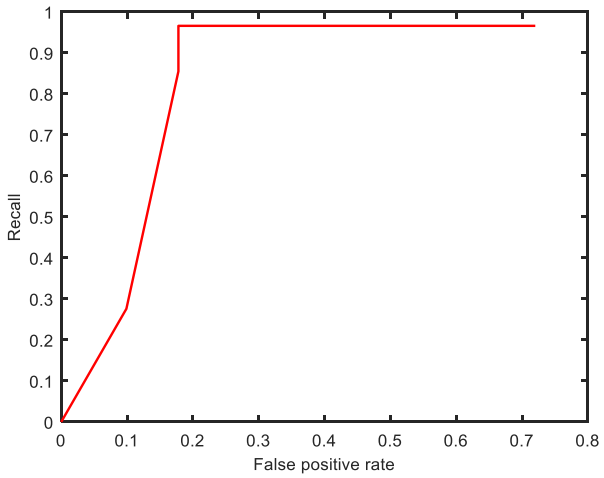

FIGURE 7. ROC curve of the proposed method.

TABLE 1. Distribution of single and multi-cellular images in each stained data.

Staining ways	Single	Multicellular
Magenta	103	109
Wright	167	8

stained images are multi-cellular while most of the Wright stained images are single-cellular. When training the model, 212 magenta stained images and 175 wright stained images are utilized, accounting for approximately 70% of the total amount of the dataset. The details of the training data are listed in Table 1 and Figure 6.

For the whole 544 images dataset, we make pixel-wise boundary annotation and give each cell a class label with the help of pathologists from Tianjin Medical University.

2) DATA PREPROCESSING

We firstly unify all the images with a fixed width of 1000, altering their lengths according to their original size ratios. Then, the images are padded into 1300×1000. Data augmentation is implemented through random image flipping and pixel value normalization.

3) METRICS

Instance segmentation includes a multi-class classification task that requires to be evaluated with convincing metrics. Therefore, we adopted two metrics, the average precision (AP) and the average recall (AR), which are originally proposed in [42] for accessing the classification of the COCO multi-class dataset. AP is a significant metric to access the overall precision of all classes. On the other hand, AR mainly depicts the overall missing detection ratio [27]. The target is assigned positive to a label if its IoU is overlapped more than a threshold with the ground-truth. In the experiments, we preset 10 thresholds (e.g., 0.50:0.05:0.95), under each of which we access the corresponding average precision and average recall over all classes. Then, we averaged the 10 precision and recalls separately to obtain an averaged AP and AR. High rates of such AP and AR imply an accurate classification and pixel segmentation of an algorithm, and strong robustness with a wide range of IoU thresholds. Besides, we withdraw two independent APs calculated on 0.5 and 0.75 IoUs as complementary to the averaged AP, which are broadly used in research.

However, the AP and AR are exclusively focused on the precision and recall performance, so the F_1-score which comprehensively depicts both the two aspects is necessary for evaluating a classification performance. The $F1$-score can be obtained as

$$\text{F1-score} = 2 \times \frac{\text{AP} \cdot \text{AR}}{\text{AP} + \text{AR}}. \tag{7}$$

To quantify the segmentation for the WBCs, we use the Dice Similarity Coefficient (Dice) to calculate the overlap of the predicted mask (*pred*) and ground-truth mask (*true*) as the following equation

$$\text{Dice} = 2 \times \frac{pred \cap true}{pred \cup true} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}}. \tag{8}$$

The Receiver Operating Characteristic (ROC) curve is used to show the diagnostic ability of the proposed method for WBCs classification. An ROC curve is constructed by plotting the true positive rate (TPR) against the false positive rate (FPR). The true positive rate (the Recall)is the proportion of observations that were correctly predicted to be positive out of all positive observations (TP/(TP+FN)), where TP and FN denote pixel-wise true positive and false negative, respectively. Similarly, the false positive rate is the proportion of observations that are incorrectly predicted to be positive out of all negative observations (FP/(TN+FP)). FP and TN are pixel-wise false positive and true negative. Figure 7 shows the ROC curve of the proposed method, where we can see the optimal recall value is attained at 0.2 FPR. The curve implies that most of WBCs are correctly recognized and only a few regions are wrongly classified as WBCs.

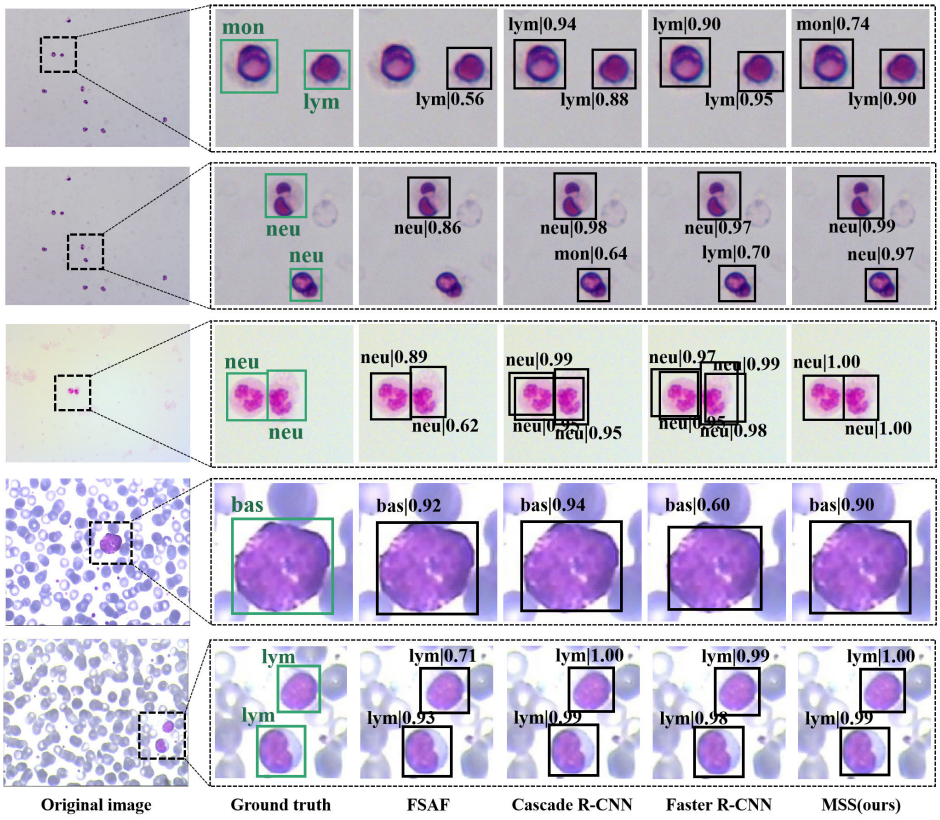

FIGURE 8. Comparison of different methods in detection results. Row (1)-(3) show the images with Magenta staining, and every magnified image contains two WBCs. Row (4)-(5) show the images with Wright staining. Column (1) shows the original images. Column (2) shows the ground-truth of the classification and location of WBCs. Columns (3)-(6) show the detection results of FSAF, Cascade R-CNN, Faster R-CNN, and MSS(ours), respectively.

TABLE 2. Comparison with different methods in instance segmentation performance.

	AP (IoU=0.50:0.95)	AP (IoU=0.50)	AP (IoU=0.75)	AR	F1-Score	Dice
One-stage						
FSAF [44]	0.695	0.853	0.794	0.836	0.844	-
SSD512 [45]	0.622	0.786	0.743	0.792	0.789	-
Two-stage						
Faster R-CNN [36]	0.625	0.894	0.791	0.739	0.809	-
Cascade R-CNN [46]	0.723	0.874	0.834	0.813	0.842	-
Mask R-CNN [28]	0.710	0.888	0.814	0.819	0.852	0.840
Mask R-CNN + PISA [47]	0.749	0.878	0.842	0.820	0.848	0.857
Mask R-CNN + GRoIE [48]	0.738	0.886	0.831	0.823	0.853	0.842
KG [39]	0.737	0.968	0.875	0.416	0.582	0.574
MSS(ours)	**0.800**	**0.925**	**0.887**	**0.878**	**0.901**	**0.902**

B. COMPARISON WITH OTHER METHODS

We compare MSS to some state-of-the-art methods on the same dataset. These methods include FSAF and SSD512 which are two one-stage detection methods, KG, and several variants of Fast R-CNN such as Faster R-CNN, Cascade R-CNN, Mask R-CNN, Mask R-CNN+PISA, Mask R-CNN+GRoIE. The experiment results are shown in Table 2 and Table 3, where we can see MSS achieves the best scores in terms of all metrics except that it performs lower than KG on AP(IoU=0.50). However, AP(IoU=0.50:0.95) and AP(IoU=0.75) are more reliable metrics than AP(IoU=0.50), thus it is still fair to say MSS outperforms KG. On individual category detection, MSS scores the highest in lym and mon detection while only trivially lower than the highest ones in neu, bas, eo detection.

We also visualize the outputs of FSAF, Cascade R-CNN, Faster R-CNN, KG, and MSS in Fig. 8 and Fig. 9. As can be seen, FSAF fails to detect some small mon and neu cells; R-CNN and Faster R-CNN both misclassify the mon cell into the type of lym, and wrongly identify the new cell as mon and lym respectively. As shown in Fig. 9, KG fails to detect small cells in Magenta staining images (in rows (1), (3), and (4)), especially when the small cells are adjacent to others.

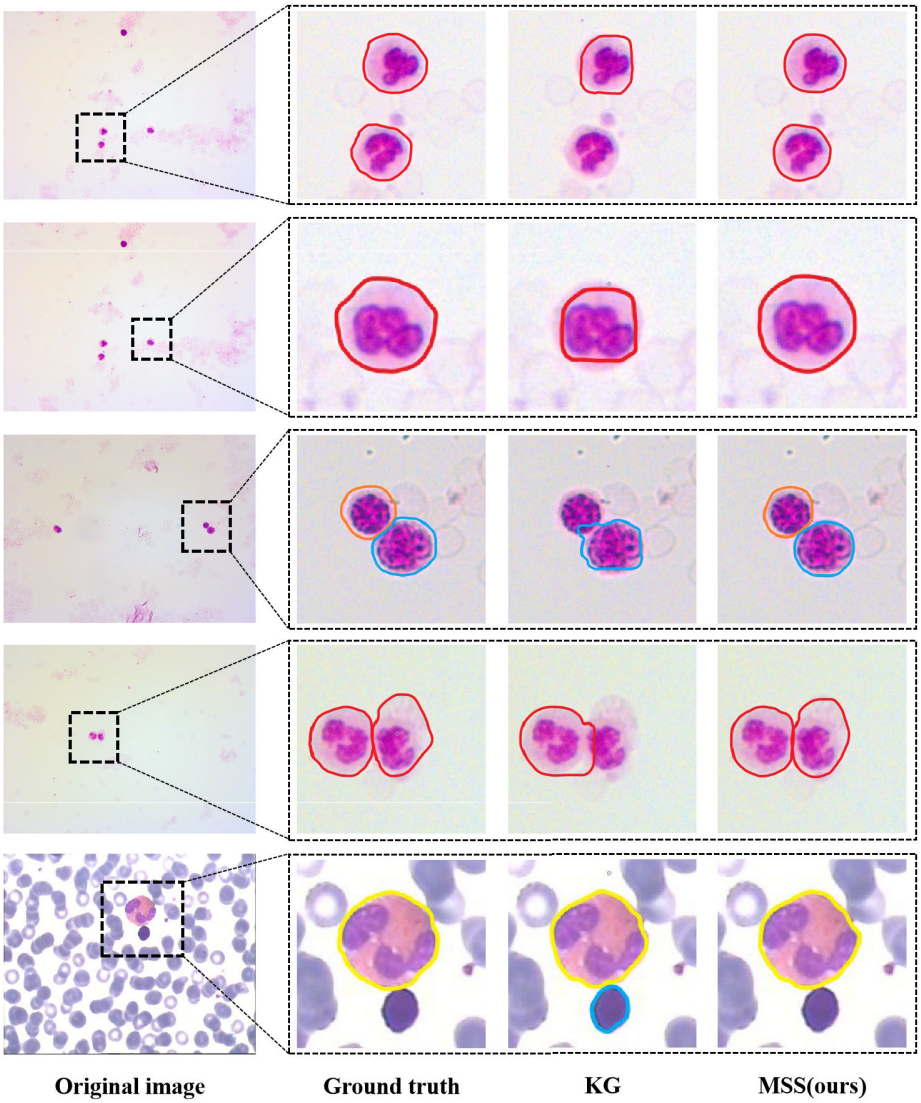

FIGURE 9. Comparison with KG in instance segmentation results. The contours in red, orange, yellow, and blue represent neu, lym, eo, and mon, respectively. Row (1)-(4) show the images with Magenta staining. Row (5) shows the images with Wright staining. Object adhesion exists in the last three rows. In the last row, the smaller one in the two adjacent objects is an impurity. Column (1) shows the original images. Column (2) shows the ground truth of instance segmentation of WBCs. Columns (3)-(4) show the instance segmentation results of KG and MSS(ours), respectively.

TABLE 3. Comparison of different detection methods on each individual category.

	neu	lym	mon	bas	eo
One-stage					
FSAF [44]	0.759	0.689	0.458	0.876	0.694
SSD512 [45]	0.700	0.569	0.331	0.864	0.646
Two-stage					
Faster R-CNN [36]	0.695	0.638	0.366	0.738	0.688
Cascade R-CNN [46]	0.745	0.662	0.414	0.907	0.889
Mask R-CNN [28]	0.723	0.666	0.493	0.854	0.883
Mask R-CNN + PISA [47]	0.781	0.680	0.420	0.945	0.925
Mask R-CNN + GRoIE [48]	0.746	0.662	0.443	0.907	0.923
MSS(ours)	**0.792**	**0.759**	**0.544**	**0.976**	**0.927**

We have also experimented 5-fold cross-validation, with the results listed in Table 4 and 5. In each fold of cross-validation, the ratios of training, validation, and testing data are set as 7: 2: 1 approximately. As can be seen, the 5 groups of experiments are even in terms of the metrics AP (IoU=0.50:0.95), AP (IoU=0.5), AP (IoU=0.75), AR, F1-score, and Dice, indicating that the dataset we construct is sufficient for describing the distribution of the WBCs features and thus competent for training a qualified model.

C. ABLATION STUDY

In this part, we make ablation experiments to testify the effectiveness of each module in our framework.

1) EXPLORATION IN CFPN STRUCTURE

In this part, we discuss the influence of different orders of upsampling, channel attention, and spatial attention in each CBAM. We allocate the operation of upsampling before, after, and in the middle of the two attention operations, and

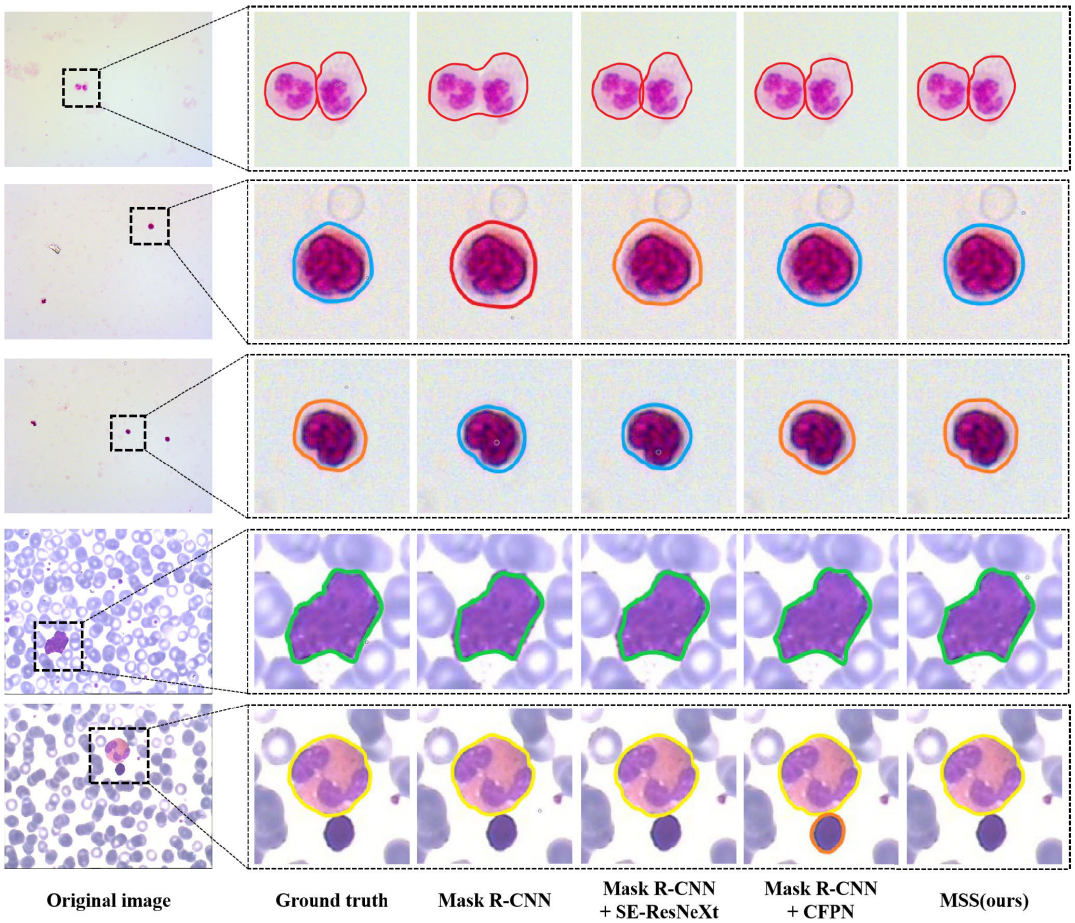

FIGURE 10. Individual effect comparison of each module. The contours in red, orange, yellow, blue, and green represent neu, lym, eo, mon, and bas, respectively. Rows (1)-(3) show the images with Magenta staining, and rows (4)-(5) show the images with Wright staining. Column (1) shows the original images. Column (2) shows the ground truth of instance segmentation of WBCs. Columns (3)-(6) show the instance segmentation results of Mask R-CNN, Mask R-CNN with SE-ResNeXt, Mask R-CNN with CFPN, and MSS(ours), respectively.

TABLE 4. Results of cross validation evaluating by the metrics.

Cross-validation	AP (IoU=0.50:0.95)	AP (IoU=0.50)	AP (IoU=0.75)	AR	F1-score	Dice
1-fold	0.800	0.925	0.887	0.878	0.901	0.902
2-fold	0.762	0.913	0.830	0.825	0.867	0.891
3-fold	0.784	0.918	0.884	0.845	0.880	0.896
4-fold	0.829	0.951	0.940	0.878	0.913	0.925
5-fold	0.806	0.933	0.921	0.856	0.893	0.908
Averaged results	0.796	0.928	0.892	0.856	0.891	0.904

TABLE 5. Averaged results of cross validation on each category.

Cross-validation	neu	lym	mon	bas	eo
1-fold	0.792	0.759	0.544	0.976	0.927
2-fold	0.618	0.814	0.587	0.910	0.884
3-fold	0.824	0.606	0.646	0.936	0.905
4-fold	0.781	0.761	0.785	0.919	0.900
5-fold	0.855	0.785	0.636	0.869	0.886
Average results	0.774	0.745	0.640	0.922	0.900

conduct three groups of experiments. The results are given in Table 6, from which we can observe that the best score is attained with the order of up-ch-sp. Putting upsampling before attention may be beneficial to restoring the details in the graph which the attention can use efficiently for producing weights. Then, we switch the channel attention and spatial attention, obtaining a new collection of results also given in Table 6. At last, we arrange the two attentions in parallel right after the upsampling, which leads to the last entry of Table 6. It turns out that the attention operations in tandem are better than in parallel, and the channel attention is better before than after the spatial attention. The reason is not clear, but a hypothesis is that the cascading effect through the two attentions is stronger than the parallel effect in detaching the feature maps of different WBCs in the feature space.

2) STRENGTHENED FEATURE EXPRESSION

To analyze the effect of SE-ResNeXt, we conduct three groups of experiments with results shown in Fig. 10. Mask

TABLE 6. Experiments on the concrete structure of CFPN.

		AP (IoU=0.50:0.95)	AP (IoU=0.50)	AP (IoU=0.75)	AR	F1-Score
Parallel	up-(ch & sp)	0.769	0.902	0.864	0.845	0.873
Tandem	ch-sp-up	0.773	0.923	0.861	0.842	0.880
	ch-up-sp	0.767	0.921	0.853	0.834	0.877
	up-sp-ch	0.770	0.913	0.865	0.833	0.871
	up-ch-sp	**0.800**	**0.925**	**0.887**	**0.878**	**0.901**

TABLE 7. Ablations of SE-ResNeXt and CFPN for MSS.

SE-ResNeXt	CFPN	AP (IoU=0.5)	AR
×	×	0.888	0.819
✓	×	0.909 (↑0.021)	0.828 (↑0.009)
×	✓	0.915 (↑0.027)	0.837 (↑0.018)
✓	✓	**0.925** (↑0.037)	**0.878** (↑0.059)

R-CNN is the baseline of the proposed network. In the Magenta staining condition (rows (1)-(3)), Mask R-CNN fails to separate adjacent cells and tends to misclassify the lym and the mon. Replacing ResNet in Mask R-CNN with the new SE-ResNeXt, the feature expression is strengthened such that the adjacent cells are accurately separated. But the classification accuracy for small cells is still limited (for example see rows (2) and (3)). The misclassification is resulted from the fact that the network is insensitive to multi-scale features. The second row of Table 7 also tells us that introducing SE-ResNeXt can make a better performance in terms of AP (IoU=0.50) and AR.

3) HIGHLIGHTED SALIENT FEATURES

Combining the Mask R-CNN and the CFPN, the small cells can be identified in the Magenta stained images (see row (1)-(3) of Fig. 10). It is worth noting that some impurities having similar size to WBCs are mistakenly detected, as shown in row (5). But after using the SE-ResNeXt module, this mistake is corrected. As illustrated in Table 7, the CFPN+Mask R-CNN also improves the segmentation results compared to the baseline Mask R-CNN. It indicates that the CBAM is functional in generating heavy weights for important channels and spatial regions. When both the SE-ResNeXt and the CFPN are used, the network obtains the highest AP and AR scores for instance segmentation. Overall, the MSS with each component playing positive roles achieves the highest scores compared to the state-of-the-art methods.

V. CONCLUSION

In this paper, we propose an instance segmentation network named MSS to realize simultaneous detection, segmentation, and classification of white blood cells from whole slide images. MSS not only achieves satisfying performance in both magenta and wright staining images but also solves the problem of multi-scale cell detection. We elaborate the explanations on the experimental results and ablation study, demonstrating the feasibility of the proposed framework. Our future work is to explore semi-supervised or weakly-supervised methods to save the cost of manual annotating.

REFERENCES

[1] P. Pandey, V. Kyatham, D. Mishra, and T. R. Dastidar, "Target-independent domain adaptation for WBC classification using generative latent search," *IEEE Trans. Med. Imag.*, vol. 39, no. 12, pp. 3979–3991, Dec. 2020.
[2] T. A. M. Elhassan, M. S. M. Rahim, T. T. Swee, S. Z. M. Hashim, and M. Aljurf, "Feature extraction of white blood cells using CMYK-moment localization and deep learning in acute myeloid leukemia blood smear microscopic images," *IEEE Access*, vol. 10, pp. 16577–16591, 2022.
[3] R. Roy and S. Sasi, "Classification of WBC using deep learning for diagnosing diseases," in *Proc. 2nd Int. Conf. Inventive Commun. Comput. Technol. (ICICCT)*, Apr. 2018, pp. 1634–1638.
[4] I. T. Young and I. L. Paskowitz, "Localization of cellular structures," *IEEE Trans. Biomed. Eng.*, vol. BME-22, no. 1, pp. 35–40, Jan. 1975.
[5] R. Macsween, "Theory and practice of histological techniques," *J. Clin. Pathol.*, vol. 30, no. 11, p. 1089, 1977.
[6] C. Zhang, X. Xiao, X. Li, Y.-J. Chen, W. Zhen, J. Chang, C. Zheng, and Z. Liu, "White blood cell segmentation by color-space-based K-means clustering," *Sensors*, vol. 14, no. 9, pp. 16128–16147, 2014.
[7] D. Karimi, H. Dou, and A. Gholipour, "Medical image segmentation using transformer networks," *IEEE Access*, vol. 10, pp. 29322–29332, 2022.
[8] L. Ma, R. Shuai, X. Ran, W. Liu, and C. Ye, "Combining DC-GAN with ResNet for blood cell image classification," *Med. Biol. Eng. Comput.*, vol. 58, no. 6, pp. 1251–1264, Jun. 2020.
[9] Y. Zhou, H. Chen, H. Lin, and P.-A. Heng, "Deep semi-supervised knowledge distillation for overlapping cervical cell instance segmentation," 2020, *arXiv:2007.10787*.
[10] F. Xing and L. Yang, "Robust nucleus/cell detection and segmentation in digital pathology and microscopy images: A comprehensive review," *IEEE Rev. Biomed. Eng.*, vol. 9, pp. 234–263, 2016.
[11] S. H. Rezatofighi and H. Soltanian-Zadeh, "Automatic recognition of five types of white blood cells in peripheral blood," *Computerized Med. Imag. Graph.*, vol. 35, no. 4, pp. 333–343, Jun. 2011.
[12] U. Adiga, R. Malladi, R. Fernandez-Gonzalez, and C. O. D. Solorzano, "High-throughput analysis of multispectral images of breast cancer tissue," *IEEE Trans. Image Process.*, vol. 15, no. 8, pp. 2259–2268, Aug. 2006.
[13] X. Yang, H. Li, and X. Zhou, "Nuclei segmentation using marker-controlled watershed, tracking using mean-shift, and Kalman filter in time-lapse microscopy," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 53, no. 11, pp. 2405–2414, Nov. 2006.
[14] M. E. Plissiti and C. Nikou, "Overlapping cell nuclei segmentation using a spatially adaptive active physical model," *IEEE Trans. Image Process.*, vol. 21, no. 11, pp. 4568–4580, Nov. 2012.
[15] Z. Lu, G. Carneiro, and A. Bradley, "An improved joint optimization of multiple level set functions for the segmentation of overlapping cervical cells," *IEEE Trans. Image Process.*, vol. 24, no. 4, pp. 1261–1272, Apr. 2015.
[16] P. K. Mondal, U. K. Prodhan, M. S. Al Mamun, M. A. Rahim, and K. K. Hossain, "Segmentation of white blood cells using fuzzy C means segmentation algorithm," *IOSR J. Comput. Eng.*, vol. 1, no. 16, pp. 1–5, 2014.
[17] N. Ghane, A. Vard, A. Talebi, and P. Nematollahy, "Segmentation of white blood cells from microscopic images using a novel combination of K-means clustering and modified watershed algorithm," *J. Med. Signals Sens.*, vol. 7, pp. 92–101, Apr./Jun. 2017.

[18] K. Al-Dulaimi, A. Al-Sabaawi, R. D. Resen, J. J. Stephan, and A. Zwayen, "Using adapted JSEG algorithm with fuzzy C mean for segmentation and counting of white blood cell and nucleus images," in *Proc. IEEE Asia–Pacific Conf. Comput. Sci. Data Eng. (CSDE)*, Dec. 2019, pp. 1–7.
[19] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 740–755.
[20] S. N. M. Safuan, R. Tomari, W. N. W. Zakaria, and N. Othman, "White blood cell counting analysis of blood smear images using various segmentation strategies," in *Proc. AIP Conf.*, 2017, vol. 1883, no. 1, pp. 1–8.
[21] J. Zhao, M. Zhang, Z. Zhou, J. Chu, and F. Cao, "Automatic detection and classification of leukocytes using convolutional neural networks," *Med. Biol. Eng. Comput.*, vol. 55, no. 8, pp. 1287–1301, Aug. 2016.
[22] I. Cseke, "A fast segmentation scheme for white blood cell images," in *Proc. 11th Int. Conf. Pattern Recognit. (IAPR)*, 1992, pp. 530–533.
[23] L. Yan, Z. Rui, M. Lei, C. Yihui, and Y. Di, "Segmentation of white blood cell from acute lymphoblastic leukemia images using dual-threshold method," *Comput. Math. Methods Med.*, vol. 2016, Apr. 2016, Art. no. 9514707.
[24] N. Salem, N. M. Sobhy, and M. E. Dosoky, "A comparative study of white blood cells segmentation using Otsu threshold and watershed transformation," *J. Biomed. Eng. Med. Imag.*, vol. 3, no. 3, Jun. 2016.
[25] H. Inbarani and A. T. Azar, "Leukemia image segmentation using a hybrid histogram-based soft covering rough K-means clustering algorithm," *Electronics*, vol. 9, no. 1, p. 188, Jan. 2020.
[26] K. AL-Dulaimi, J. Banks, K. Nugyen, A. Al-Sabaawi, I. Tomeo-Reyes, and V. Chandran, "Segmentation of white blood cell, nucleus and cytoplasm in digital haematology microscope images: A review–challenges, current and future potential techniques," *IEEE Rev. Biomed. Eng.*, vol. 14, pp. 290–306, 2021.
[27] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 2961–2969.
[28] Q. Wang, S. Bi, M. Sun, Y. Wang, D. Wang, and S. Yang, "Deep learning approach to peripheral leukocyte recognition," *PLoS ONE*, vol. 14, no. 6, Jun. 2019, Art. no. e0218808.
[29] Y. Zhou, H. Chen, J. Xu, Q. Dou, and P. A. Heng, "IRNet: Instance relation network for overlapping cervical cell segmentation," in *Proc. Med. Image Comput. Comput. Assist. Intervent. (MICCAI)*, 2019, pp. 640–648.
[30] S. Nazlibilek, D. Karacor, T. Ercan, M. H. Sazli, O. Kalender, and Y. Ege, "Automatic segmentation, counting, size determination and classification of white blood cells," *Measurement*, vol. 55, no. 3, pp. 58–65, 2014.
[31] S. Dasariraju, M. Huo, and S. McCalla, "Detection and classification of immature leukocytes for diagnosis of acute myeloid leukemia using random forest algorithm," *Bioengineering*, vol. 7, no. 4, p. 120, Oct. 2020.
[32] H. Kutlu, E. Avci, and F. Özyurt, "White blood cells detection and classification based on regional convolutional neural networks," *Med. Hypotheses*, vol. 135, Feb. 2020, Art. no. 109472.
[33] E. Hussain, L. B. Mahanta, C. R. Das, M. Choudhury, and M. Chowdhury, "A shape context fully convolutional neural network for segmentation and classification of cervical nuclei in pap smear images," *Artif. Intell. Med.*, vol. 107, Jul. 2020, Art. no. 101897.
[34] Y. Li, H. Qi, J. Dai, X. Ji, and Y. Wei, "Fully convolutional instance-aware semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2359–2367.
[35] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.
[36] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path aggregation network for instance segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8759–8768.
[37] Z. Huang, L. Huang, Y. Gong, C. Huang, and X. Wang, "Mask scoring R-CNN," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 6409–6418.
[38] J. Yi, P. Wu, Q. Huang, H. Qu, B. Liu, D. J. Hoeppner, and D. N. Metaxas, "Multi-scale cell instance segmentation with keypoint graph based bounding boxes," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2019, pp. 369–377.
[39] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1492–1500.
[40] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.
[41] S. Woo, J. Park, J. Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 3–19.
[42] T. Y. Lin, M. Maire, S. Belongie, J. Hays, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 740–755.
[43] C. Zhu, Y. He, and M. Savvides, "Feature selective anchor-free module for single-shot object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 840–849.
[44] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Y. Fu, and A. C. Berg, "SSD: Single shot MultiBox detector," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2016, pp. 21–37.
[45] Z. Cai and N. Vasconcelos, "Cascade R-CNN: Delving into high quality object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6154–6162.
[46] Y. Cao, K. Chen, C. C. Loy, and D. Lin, "Prime sample attention in object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 11583–11591.
[47] L. Rossi, A. Karimi, and A. Prati, "A novel region of interest extraction layer for instance segmentation," 2020, *arXiv:2004.13665*.

MENG ZHAO received the Ph.D. degree from Tianjin University, in 2016. She is currently an Associate Professor with the Tianjin University of Technology. She was granted the Alain Bensoussan Fellowship by European Research Consortium for Informatics and Mathematics, in 2020. Her research interests include medical image processing and computer vision.

HONGXIA YANG is currently pursuing the master's degree in computer science with the Tianjin University of Technology. Her research interests include medical image analysis and computer vision.

FAN SHI (Member, IEEE) received the Ph.D. degree in optics from Nankai University, Tianjin, China, in 2012. From June 2018 to August 2019, he was a Research Scholar with West Virginia University. He is currently an Associate Professor with the Tianjin University of Technology, Tianjin. His research interests include machine vision, pattern recognition, and optics.

XINPENG ZHANG received the Ph.D. degree from Tiangong University, Tianjin, China, in 2017. From 2018 to 2020, he was a Postdoctoral Researcher with the Guangdong University of Technology, Guangzhou, China. He is currently a Lecturer with the School of Computer Science and Engineering, Tianjin University of Technology, Tianjin. His current research interests include image processing and deep learning.

YAO ZHANG received the B.Eng. and Ph.D. degrees in control engineering from Tianjin University. He is currently an Assistant Professor with the Tianjin University of Technology. His research interests include nonlinear control theory and machine learning.

XUGUO SUN received the M.D. and Ph.D. degrees from the School of Medicine, Kumamoto University, Japan. He is currently a Professor in experimental diagnostics with Tianjin Medical University, China. His research interests include methodological study of clinical medical examination and methodological study on artificial intelligence detection of cytopathology. He is the Chairperson of Inspection and Testing Technology Industry Alliance, Tianjin, China.

HAO WANG (Senior Member, IEEE) received the B.Eng. and Ph.D. degrees in computer science and engineering from the South China University of Technology. He is currently an Associate Professor with the Department of Computer Science, Norwegian University of Science and Technology, Norway. He has published more than 120 papers in reputable international journals and conferences. His research interests include big data analytics, health informatics, and safety-critical systems. He is a member of IEEE IES Technical Committee on Industrial Informatics.