IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 28 March 2024, accepted 8 April 2024, date of publication 11 April 2024, date of current version 18 April 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3387832

RESEARCH ARTICLE

VB-SOLO: Single-Stage Instance Segmentation of Overlapping Epithelial Cells

LICHUAN LI[1], WEI CHEN[1,3], AND JIE QI[2]
[1]School of Communication and Information Engineering, Xi'an University of Science and Technology, Xi'an, Shaanxi 710054, China
[2]Orthopedic Department, Shaanxi Provincial People's Hospital, Xi'an 710064, China
[3]Xi'an Key Laboratory of Network Convergence Communication, Xi'an, Shaanxi 710054, China

Corresponding author: Jie Qi (qj_dentist@163.com)

ABSTRACT The instance segmentation of overlapping cells in smear images of epithelial cells is challenging due to the significant overlap and adhesion between the cells' translucent cytoplasm. In this paper, an improved single-stage instance segmentation network called VoVNet-BiFPN-SOLO (VB-SOLO) is proposed to address this problem. The model takes SOLOv2 model as its main frame. Firstly, the backbone network uses Efficient Channel Attention (ECA) to optimize the VoVNetv2 network to increase the information interaction across channels and enhance the extraction of cell instance features. Secondly, the bi-directional feature pyramid network (BiFPN) is introduced to connect with the new backbone. BiFPN can achieve the weighted fusion of features with different resolutions from bottom to top and keep more shallow semantic information in the network. Finally, the Convolutional Block Attention Module (CBAM) is added to the mask branch to improve cell segmentation results in feature maps. Experimental results on the publicly available datasets CISD and Cx22 demonstrate the effectiveness of the VB-SOLO model, achieving a DC_P of 0.966 and 0.940 and a FNR_O of 0.055 and 0.03. Compared to the original SOLOv2 algorithm, the proposed method achieved improvements in DC_P of 1.3% and 1.1% respectively. Additionally, comparative tests with multiple instance segmentation networks have shown that the proposed improved network can achieve a better balance between segmentation accuracy and efficiency. The experimental results demonstrate the effectiveness of the proposed network improvements and the potential of single-stage instance segmentation networks in overlapping cell image segmentation.

INDEX TERMS Biomedical imaging, cervical cancer, convolutional neural networks, deep learning, image segmentation, instance segmentation, SOLOv2.

I. INTRODUCTION

Automated cell image segmentation is a crucial domain of research in medical image processing [1]. Its primary goal is to segment cell regions in images to extract features such as morphology, color, contour, and nucleoplasm ratio of each cell. These extracted features serve as a foundation for quantitative analysis and qualitative evaluation of cytopathological images. Furthermore, cell segmentation plays a pivotal role in the investigation of cell counting and tracking, which constitutes a critical component of computer-aided medical diagnosis.

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

However, during the automatic segmentation process of cell images, there are typically multiple challenges: (1) Due to lighting, staining, or equipment factors in the production process, cell images often possess characteristics such as low resolution, poor contrast, complex background, and impurities, which are not beneficial for segmentation algorithms. (2) Cell images have diverse cell types with varying morphologies, irregular shapes of some cell contours, and there may be a significant number of cell overlaps or adhesions. (3) Acquiring and annotating cell images involves higher costs compared to natural images, and algorithm models often lack sufficient training samples.

An example includes a smear image of urothelial cells for bladder cancer screening [2] and a smear image of cervical

epithelial cells for cervical cancer screening [3]. After Pap staining of these two types of epithelial samples, the smear images exhibit a significant number of cell clumps that overlap or adhere to each other, and the cytoplasm appears irregularly contoured and translucent. This presents a substantial challenge in accurately identifying cell boundaries for cell segmentation and classification [4]. In current clinical practice, this smear screening process is performed by experienced physicians, which is a time-consuming and laborious task [5]. Repetitive and tedious manual smear reading often leads to errors and misjudgments, which cause delays in the timely diagnosis of disease and subsequent treatment. Therefore, it is crucial to develop and optimize automatic cell segmentation algorithm to alleviate the workload of medical professionals and enhance diagnostic efficiency and accuracy.

Traditional methods were widely used for cell instance segmentation before the application of deep learning models to segment overlapping cells. Classic machine learning algorithms utilized hand-crafted features, which were then inputted into classification algorithms such as Level Set [6], [7], [8], Shape-coding [9], Region-based [10], [11] and Watershed-based [12]. However, in recent years, deep learning methods, represented by Convolutional Neural Networks (CNN), have emerged as the preferred approach for cytoplasm segmentation of overlapping cells [13], [14].

Deep learning-based image segmentation techniques can be categorized into two major types: semantic segmentation and instance segmentation. Semantic segmentation is used to classify all pixel points in an image and assign category labels. Currently, UNET [15] and its modifications [16], [17], [18] are extensively utilized in semantic segmentation of microscopic medical cell images. However, a significant drawback of semantic segmentation is its inability to discern between different instances of the same category. Conversely, instance segmentation integrates the principles of object detection and semantic segmentation to not merely classify the pixel points of digital images, but also distinguish distinct instances belonging to the same category. Deep learning-based instance segmentation methods can be broadly classified into two main types: single-stage and two-stage methods. The two-stage instance segmentation algorithms can be further subdivided into top-down box-based detection algorithms and bottom-up segmentation-based algorithms. Mask R-CNN [19], which is based on the Region Proposal Network (RPN), is the most classical two-stage instance segmentation algorithm and is widely used in medical image instance segmentation [20], [21], [22], [23]. One-stage instance segmentation methods can be further classified into anchor-based [24], [25], [26], [27], [28] and anchor-free [29], [30], [31] methods depending on whether anchor frames are used. The single-stage algorithm aims to achieve better segmentation results by going beyond the limits of the RPN.

In the field of overlapping cell segmentation, CNN-based methods can be broadly classified into two categories. The first category involves combining the semantic segmentation results of the nucleus or cytoplasm with subsequent image processing algorithms to achieve instance segmentation of overlapping cells. For example, Mahyari and Dansereau [32] proposed a three-phase scheme focused on multi-layer image processing for overlapping cell image segmentation. The first two steps obtain the semantic segmentation results of the trained convolutional neural network and perform rough cell segmentation based on the multilayer randomised wandering map technique, respectively. In the third stage, a Hungarian algorithm is used to optimize the assignment of individual pixel positions for the final cell segmentation. Similarly, Zhang et al. [33] identified strong contour points based on semantic segmentation results of cell nuclei and a boundary tracking algorithm. They utilized a combined approach to obtain cell boundaries and extract overlapping cell boundaries based on semantic segmentation. However, these methods accomplish the segmentation task through a multi-stage process, where performance degradation at any one of these stages can affect the final segmentation result.

The second category involves employing a two-stage instance segmentation method based on Region Proposal Network (RPN) for overlapped cell segmentation. For example, Hao et al. [34] developed a Convolutional Network Model for Region Proposal Segmentation (CRP-PSN). The cell region detection and localization network CRPN is utilized for cell region detection and localization, identifying cervical cells and providing the target area for segmentation. The segmentation network PSN is used to complete the pixel-level segmentation of cervical cells in the target area. Similarly, Chen and Zhang [35] utilized Mask R-CNN to annotate and train the cell boundary, and then completed the instance segmentation of the cell boundary. Zhou et al. [36] proposed a novel instance relationship network (IRNET) to achieve the instance segmentation of overlapped cell units through the study of the interaction relationship between cell instances. Specifically, IRM was added to the original Mask R-CNN model to model the interaction between instances, and the whole process was completed through end-to-end training.

While the previously mentioned deep learning-based methods have demonstrated promising results, both types of approaches currently possess inherent limitations. The first category relies on the initial segmentation results obtained from the semantic segmentation network and the parameter configurations for subsequent processes. In contrast, the two-stage framework employed in the second category relies on the evaluation of a substantial number of region proposal boxes within the network. However, this approach consumes a significant amount of computational resources.

This paper introduces an improved single-stage instance segmentation network called VB-SOLO for achieving segmentation of overlapping cells in epithelial cell smear images

using the CISD dataset [37] and the Cx22 dataset [38]. The main contributions of our work are summarized as follows:

- Improved instance segmentation model SOLOv2 [31] network: The complexity and diversity of cell image features are addressed in this paper by replacing the original feature extraction backbone network with the VoVNetv2 [39] backbone network, which can fuse both shallow and deep network features. The FPN network is also improved to BiFPN [40], allowing for bi-directional weighted feature fusion of the network. Moreover, the CBAM module [41] is added behind the mask feature map to generate better segmentation results.
- Single-stage instance segmentation network applied to overlapping cell segmentation: An end-to-end single-stage instance segmentation network for overlapping epithelial cells is implemented.
- The segmentation effectiveness of the proposed method is evaluated on two public datasets, and the results show that our method achieves higher segmentation accuracy and stronger generalization capability than other parties. In addition, the effectiveness of the improvements to the network is tested and validated.

II. METHOD

The SOLOv2 network is an instance segmentation algorithm that employs a single-stage approach to differentiate objects by their central location and shape. It utilizes a fully convolutional, frameless, and group-free paradigm to dynamically segment each instance in the image [31]. To achieve this, the network divides the input image into an $S \times S$ grid. Initially, the image is processed by the ResNet and FPN, resulting in a fused feature map I with dimensions $H \times W \times E$ (where H, W, and E represent the height, width, and number of channels of the feature map, respectively). This fused feature map is further divided into an $S \times S \times E$ feature map.

For the classification branch, the feature map undergoes multiple convolutional layers and ultimately unfolds into a feature map of size $S \times S \times C$ (Here, C denotes the number of channels of the classification feature map), representing the number of categories of the target, where each channel corresponds to a particular category.

In the segmentation branch, the SOLOv2 network stands out for its ability to employ dynamic convolutional kernels. This branch is further split into a dynamic convolutional kernel branch and a feature branch. The feature map I obtained from the feature pyramid, serves as the input of the segmentation branch. It then passes through the convolutional kernel branch and feature branch to obtain the dynamic convolution kernel G (with dimensions $S \times S \times D$, where D represents the number of channels of the dynamic convolution kernel) and the feature map F (with dimensions $H \times W \times E$), respectively. The feature map is subsequently convolved using the dynamic convolution kernel G to calculate the mask.

This paper introduces three main modifications to the existing SOLOv2 network. The network structure is improved by using the enhanced VoVNetV2 as the backbone network of Solov2, which enhances the feature extraction of overlapping cell images. Additionally, the original FPN network is replaced with BiFPN, which introduces weights to achieve a better balance of feature information at different scales. An attention module is also added after the original mask feature branch to pay more attention to the location information of cell segmentation. The improved network structure is illustrated in FIGURE 1.

A. BACKBONE

This paper utilizes the VoVNetv2 [39] architecture as the backbone network. VoVNetv2 is an efficient network that utilizes a tandem approach to aggregate shallow features, deepen the relationship between feature maps, and improve the utilization of shallow features in deeper layers. It also incorporates a residual structure to prevent gradient disappearance and improve detection accuracy. Compared to ResNet [42], VoVNetv2 provides a more diverse and superior feature representation.

The main structure of VoVNetv2, used in this study, is illustrated in FIGURE 2. The VoVNetv2 takes a cell smear image of $512 \times 512 \times 3$-sized as input and consists of five convolutional stages. The first stage comprises three 3×3 convolutional layers, followed by four stages that incorporate One-Shot Aggregation (OSA) modules [39]. Each OSA module consists of five 3×3 convolutional layers and one 1×1 convolutional layer. In this module, each layer of the input is connected in two ways: one connection is to the 3×3 convolutional layer to produce a feature map with a larger perceptual field, and the other connection is to the final feature map output layer to aggregate sufficient features. The aggregated output layer undergoes a 1×1 convolution operation to obtain a diverse feature map $X_{div}^{C \times W \times H}$.

The effective Squeeze-Excitation (eSE) channel attention module then utilizes the $A_{eSE}^{c \times 1 \times 1}$ channel attention feature descriptors to the diverse feature map $X_{refine}^{C \times W \times H}$ to obtain richer information. Eventually, the initial input features of each layer are added to the refined feature map $X_{refine}^{C \times W \times H}$ through residual connections. Each OSA phase ends with a 3×3 maximum pooling layer spanning 2 for downsampling. This aggregation method allows for the aggregation of intermediate features at once, greatly improving the computational efficiency of media access and graphics processors while maintaining strong connections.

In this paper, we propose replacing the eSE module in the original structure with the Efficient Channel Attention (ECA) module [43]. The ECA module uses an adaptive convolution kernel size k that depends on the number of channels for convolution. This is different from the eSE module, which uses a fixed 1×1 convolution kernel. The calculation of k is given by Equation 1:

$$k = \psi(C) = \left| \frac{log_2(C)}{\gamma} + \frac{b}{\gamma} \right|_{odd}, \tag{1}$$

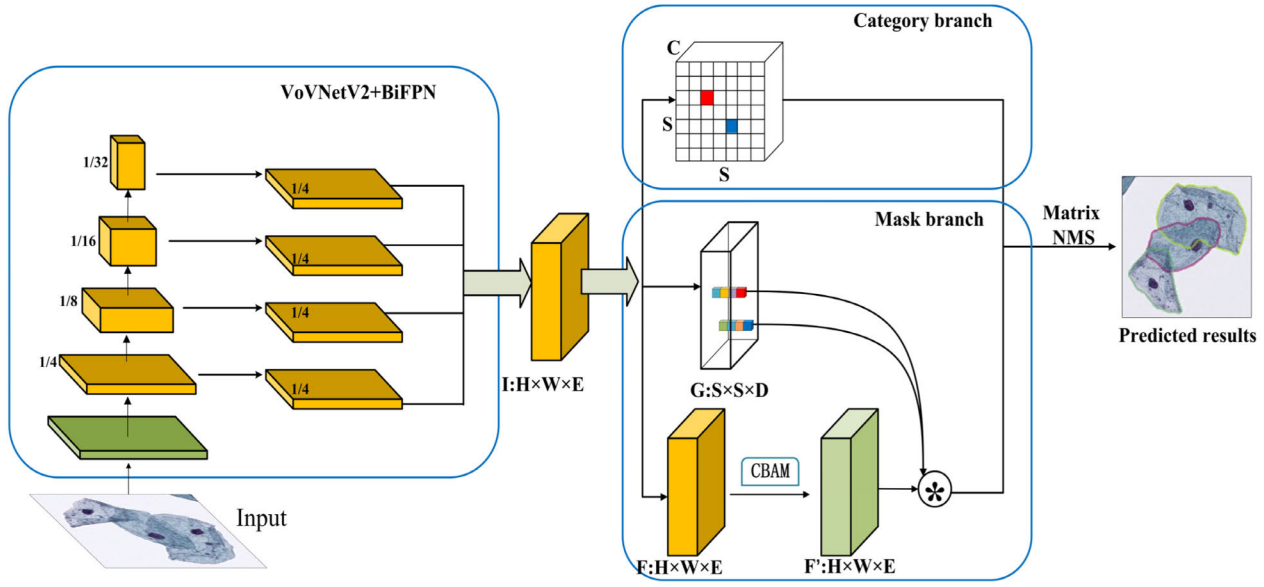

FIGURE 1. Overview of the proposed SOLOv2.

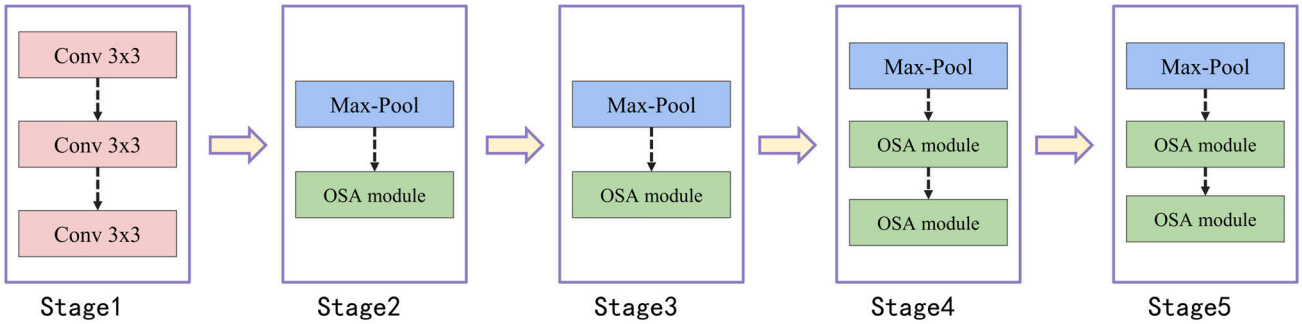

FIGURE 2. Specific operation process of VoVNet-v2-39.

where C denotes the number of feature channels and $\gamma = 2$ and $b = 1$ denote the two hyperparameters. Compared with the eSE channel attention module, the ECA module increases the information interaction across channels and can bring significant performance improvement. The OSA module used in this experiment is depicted in FIGURE 3.

B. FEATURE PYRAMID NETWORK

In the SOLOv2 network, the classical FPN structure is used to integrate feature layer information. Nevertheless, with increasing network depth, the shallow information in the original FPN is passed to the top layer with information loss. This is problematic for cells, as the shallow layer information contains important semantic information such as location and contour. To address this issue, this paper adopts the Bidirectional Feature Pyramid Network (BiFPN) structure, which effectively fuses the shallow layer information with the deep layer information.

BiFPN [39] was developed by the Google team as an improved network structure, built upon PANet [44], and FIGURE 4 depicts the schematic structure of the network. Compared to the feature pyramid network (FPN) structure in the original network, BiFPN bi-directionally incorporates features from layers 3 to 6 of the original network and considers that if a node has only one input edge, then it contributes the least to the network. Therefore, the feature fusion nodes in layers 3 and 6 are removed to reduce the computational effort. At the same time, a cross-scale connectivity approach is used to add an edge to fuse features in the feature extraction network directly with features relative to size in the bottom-up path to maintain more shallow semantic information in the network without losing too much relatively deep semantic information.

In contrast to conventional feature fusion, BiFPN utilizes a weighted fusion mechanism to distinguish and merge diverse input features, enabling it to learn the significance of each feature. BiFPN achieves similar accuracy to SoftMax-based fusion while exhibiting faster computational speed. Formula (2) describes the fast normalization method used by BiFPN.

$$Out = \sum_i \frac{w_i}{E + \sum_i w_i} \cdot In_i. \tag{2}$$

Equation (2) defines w_i as the weight, which is constrained to be non-negative under the ReLu activation function. The

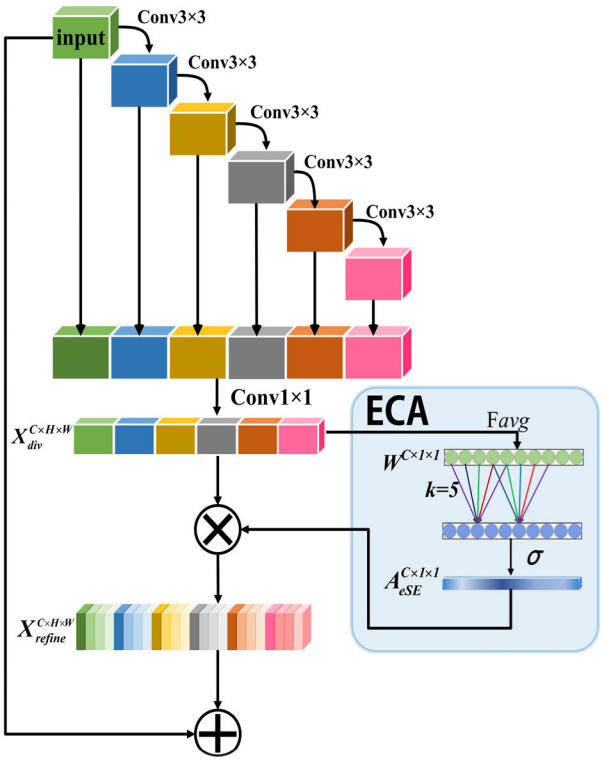

FIGURE 3. Structure of OSA module of VoVNetv2.

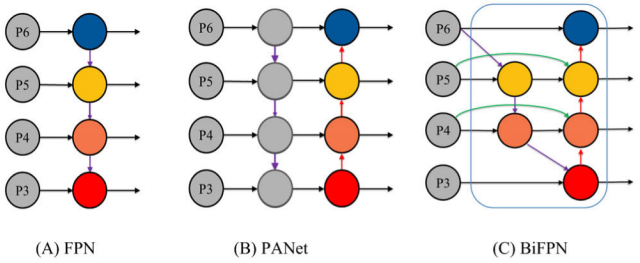

FIGURE 4. Feature pyramid network design (a) is the original FPN; (b) is the PANet network structure; (c) is the BiFPN module structure.

parameter ε is utilized to minimize the risk of numerical instability and is assigned a low value. In_i refers to the input features, whereas *Out* denotes the outcome of the weighted feature fusion.

C. CBAM MODULE

In the field of computer vision, attention mechanisms enable networks to selectively focus on the most relevant parts of image information, much like the human visual system. To mitigate the impact of background noise on feature maps, Convolutional Block Attention Module (CBAM) [41] has been incorporated into the mask branch of the original SOLOv2 network. This helps to enhance the focus of the fused feature maps on mask-generated regions of instances.

FIGURE 5 demonstrates that the CBAM is not only attentive to channel domain information but also to spatial domain information, which is particularly relevant for mask generation tasks that require attention to cell spatial distribution. The channel attention module generates the channel attention map by learning relationships between features in the channel domain, while the spatial attention module generates the spatial attention map by learning relationships between features in the spatial domain. These maps are then used to adjust the features, improving the cell segmentation results in the feature map.

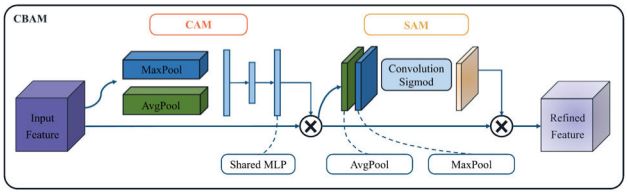

FIGURE 5. Convolutional block attention module.

TABLE 1. The details for CISD and Cx22 datasets.

Dataset	Image	Cells	Size	Preparation
CISD	3911	9378	512	Papanicolaou smear
Cx22	500	5833	512	Liquid-based cytology

III. DATASET AND EVALUATION METRICS

A. DATASETS

The Cell Instance Segmentation Dataset (CISD) is proposed to enhance the adoption of deep learning instance segmentation networks for the analysis of overlapping cells [37]. It contains 3911 EDF (extended depth of field) cell smear images, each containing at least two contact or overlapping totaling 9378 epithelial cells. All samples were taken from 30 digital cytology slides, and these smears were stained using different Pap stains. Cytological smears were routinely generated from healthy human urine samples using a Hologic ThinPrep 5000 processor and an Agilent Dako CoverStainer.

The Cx22 dataset is a public dataset for deep learning training of cervical cell instance segmentation [38]. The dataset images were produced by a platform consisting of a Nikon ELIPSE Ci slide scanner and a 3-megapixel digital camera. The dataset was derived from 686 cervical cytology images with a resolution of 2048 × 1536. The sub-images of the dataset were normalized to 512 × 512 by padding white pixels around their boundaries after cropping. With the guidance of biomedical professionals, a total of 500 labeled images and 5833 labeled cell instances were annotated.

The cell images in the CISD dataset are stored as JPG files and are accompanied by a JSON file containing instance masks encoded in the RLE format. The Cx22 dataset stored data using MATLAB.Mat files with hdf5 data format. For convenience, these files were converted into image and mask files to JPEG format using Python code. The details of the datasets are as shown in Table 1.

B. IMPLEMENTATION DETAILS

The hardware environment for training and testing was as follows: Intel Xeon (R) Gold 6148 CPU 2.40GHz x 80 and

NVIDIA GeForce RTX 3090 GPU. All experiments were conducted using the same hardware environment. The experimental operating system is based on Ubuntu 20.04, Python 3.7, and the PyTorch 1.8.0 framework.

During the training process, the epoch number was set to 150. The network was trained with the SGD optimizer with a momentum of 0.9, which is the default value for the MMDetection framework. A starting learning rate of 0.001 was used, with a decay factor of 0.1 applied at the 80th and 130th epochs to reduce the learning rate. The batch size was uniformly set to 16. In the instance segmentation experiments, the CISD and CX22 datasets were reprocessed to COCO format.

C. EVALUATION METRICS

Four common evaluation metrics were selected to assess the performance of the instance segmentation network. The first metric is the pixel based DC_P (the dice similarity coefficient), where a unit in the ground truth is considered well segmented if a segmented region in the detection is above a specific threshold with its DC_P. If the Intersection over Union (IoU) between the predicted mask and the associated ground-truth is greater than 0.7, the segmented cell is considered a true positive because this threshold was chosen in the challenge of ISBI'14 (The 2014 IEEE International Symposium on Biomedical Imaging) to define good segmentation [45]. Then there is the instance false-negative rate FNR_O (Object-wise false-negative rate): the proportion of unidentified cell instances (missed reports) among all labeled instances. Finally, the pixel-wise true positive rate (TPR_P) and the pixel-wise false positive rate (FPR_P) are shown at the pixel level, and higher values of TPR_P and lower values of FPR_P imply a better segmentation quality of cell segmentation results. The DC_P, TPR_P and FPR_P are calculated as follows equations. Among these metrics, TP is true positive, which indicates that positive samples are correctly determined as positive samples. Conversely, FP represents false positive, denoting negative samples inaccurately labeled as positive. FN represents false negatives, indicating positive samples erroneously classified as negative. Lastly, TN represents true negatives, signifying accurately identified negative samples.

$$DC_p = \frac{2\,|GT \cap PR|}{|GT \cup PR|}, \tag{3}$$

$$TPR_P = \frac{tp}{tp + fn}, \tag{4}$$

$$FPR_P = \frac{fp}{fp + tn}. \tag{5}$$

The segmentation speed of the algorithm was evaluated using Frames Per Second (FPS).

IV. RESULTS AND COMPARISON

A. EFFECTIVENESS OF THE PROPOSED METHOD

First, we conducted separate experiments on the CISD dataset and CX22 dataset and compared other different deep learning segmentation network models for overlapping cell segmentation. Two-stage example segmentation model package expansion: ANCIS [46], a method of a box-based instance segmentation method which was successfully applied to touching neural cell segmentation; Chen and Zhang [35], the method of segmentation of overlapping cervical cells with masked regional convolutional neural networks (Mask R-CNN). One-stage instance segmentation networks are YOLACT++ [27], an anchor-based single-stage instance segmentation network, and SOLOv2.

Table 2 shows the results on the CISD test set, from which we know that our approach has a DC_P 1.5% higher than Mask R-CNN, 2.2% higher than YOLACT, and 1.3% higher than SOLOv2. For FNRO and other metrics, VB-SOLO also has the best results compared to other segmentation models. In terms of inference speed, while YOLACT has the fastest inference speed, it tends to have lower accuracy. The method described in this paper achieves relatively good segmentation accuracy under conditions where the inference speed is similar to Mask R-CNN.

Table 3 shows the results on the CX22 test set. The cell images in the CX22 dataset have lower resolution, more complex backgrounds, and more tasks for small target cell identification segmentation compared to the CISD dataset. Table 3 shows that VB-SOLO outperforms Chen's method by 0.7% in terms of DC_P and YOLACT++ by 2.2% in terms of DC_P. In addition, VB-SOLO also achieves 1.1% higher DC_P compared to SOLOv2. However, in terms of FNR_O, VB-SOLO performs worse than Mask R-CNN, but better than other segmentation models. This may be due to the inherent advantage of Mask R-CNN in small target extraction segmentation, while still achieving a low miss detection rate compared to other networks. It can be seen that the method proposed in this paper is slightly higher than the segmentation accuracy of the two-stage instance segmentation algorithm, Mask RCNN, and the segmentation accuracies of the other single-stage instance segmentation algorithms are lower than that of the algorithm proposed in this article. In terms of segmentation speed, the FPS of the algorithm proposed in this paper is higher than that of Mask R-CNN 's algorithm.

B. ABLATION EXPERIMENTS

To validate the BiFPN module, we conducted comparative experiments with various feature pyramid modules on the CISD dataset. The experiments were performed using VoVNetV2-39-ECA as the backbone network. We explored different feature pyramid modules, such as PANet [44], which first connects from the bottom to the top and then loops from the top to the bottom. NAS-FPN [47] utilized neural architecture search to discover more effective cross-scale feature network topologies, including the BiFPN module introduced in this study.

Table 4 reveals that the traditional top-down original FPN network has the fastest inference speed at 41.0. However, it is constrained by one-way information flow, resulting in lower accuracy. While NAS-FPN performs slightly better

TABLE 2. Quantitative comparison against other methods on the CISD test set.

Method	DC_P	FNR_O	TPR_P	FPR_P	FPS
ANCIS [46]	0.909±0.030	0.094±0.229	0.884±0.002	0.0360±0.0308	——
Mask R-CNN [35]	0.951±0.037	0.064±0.136	0.945±0.047	0.0101±0.0169	32.5
YOLACT [26, 27]	0.944±0.039	0.088±0.163	0.933±0.043	0.0113±0.0175	**52.9**
SOLOv2 [31]	0.953±0.033	0.066±0.106	0.958±0.042	0.0082±0.0153	38.6
VB-SOLO	**0.966±0.031**	**0.055±0.118**	**0.964±0.035**	**0.0070±0.0140**	35.1

TABLE 3. Quantitative comparison against other methods on the Cx22 test set.

Method	DC_P	FNR_O	TPR_P	FPR_P	FPS
ANCIS [46]	0.894±0.023	0.091±0.127	0.877±0.011	0.0033±0.0029	——
Mask R-CNN [35]	0.933±0.044	**0.030±0.061**	0.916±0.065	0.0009±0.0019	30.8
YOLACT [27]	0.921±0.045	0.044±0.070	0.893±0.042	0.0009±0.0019	**54.3**
SOLOv2 [31]	0.929±0.015	0.038±0.057	0.903±0.022	0.0009±0.0017	38.2
VB-SOLO	**0.940±0.016**	0.035±0.069	**0.917±0.021**	**0.0008±0.0019**	35.3

TABLE 4. Comparison of performance of different feature pyramid modules.

NECK	DC_P	FNR_O	TPR_P	FPR_P	FPS
FPN	0.955	0.066	0.959	0.0088	**41.0**
PANet[44]	0.964	0.061	0.932	0.0093	39.6
NAS-FPN[47]	0.960	**0.055**	0.942	0.0086	22.2
BiFPN[39]	**0.965**	0.056	**0.964**	**0.0082**	35.3

than PANet, its inference speed noticeably decreases. The BiFPN mentioned in this paper, through additional weighted feature fusion, maintains a similar inference speed to PANet while achieving higher segmentation accuracy.

In order to further verify the effectiveness of the proposed structure, its detection effect is compared with that of the original module, and the results are shown in Table 5. As can be seen from Table 5 after replacing the backbone network with the unimproved VoVNetV2, the DC_P decreases by 1.1% compared with the original model, but the FPS improves by 1.9. After replacing the backbone network with ECA-VoVNetV2, the segmentation accuracy is similar to that of the original model, but the FPS improves by 2.9%. Obviously, the addition of ECA-VoVNetV2 speeds up the processing speed of the model. When FPN is replaced by BiFPN, the DC_P of the model increases by 0.8% and the FPS decreases by 4.9%. After adding the CBAM alone, the segmentation effect is slightly improved compared with the original model, which verifies the effectiveness of the algorithm proposed in this paper. When all three improvement points are added to the network, the segmentation performance of the network comes to the best.

C. QUALITATIVE COMPARISON

Figure 6 shows a representative sample of challenging cases in the CISD dataset, such as highly overlapping (first row) and mutually adherent (second row) cytoplasm. Figure 6(e) shows the results of the VB-SOLO in this paper and compares Figure 6(b): YOLACT, Figure 6(c): Mask R-CNN, and Figure 6(d): SOLOv2. In the first row of images, the anchor-based two-stage instance segmentation algorithm inhibits the screening of overlapping candidate frames due to the highly overlapping cytoplasm, and cellular misses occur. The same single-stage instance segmentation network of YOLACT++ and the original SOLOV2 algorithm showed confusing segmentation boundaries. In the next row of the adherent cell image, the YOLACT++ algorithm showed a missed detection of one cell under two cell overlays, and the segmentation boundaries of the other algorithms showed less satisfactory segmentation results compared with the labeled map. Compared with other algorithms, our modified SOLOV2 obtained a better segmentation result.

FIGURE 7 shows a representative sample of challenging cases in the CISD dataset, such as cytoplasm with high overlap (first row) and dense small targets (second and third rows). For the adherent overlapping cell images in the first row, YOLACT++ and SOLOv2 also show a poorly segmented boundary, while Mask R-CNN shows cytoplasmic over-segmentation below the red rectangle. For the cytoplasmic segmentation of dense small targets (second and third rows), both Mask R-CNN and SOLOv2 show cell misses (top left of the white circle in the second row and bottom left of the red box in the third row), compared with which our proposed method detects the overlapping cells under the small targets and gives relatively better segmentation results.

D. LIMITATIONS

The method proposed in this paper is based on two datasets, the CISD dataset and the Cx22 dataset. In both datasets, areas that are difficult for cell experts to delineate the actual contours are chosen to be ignored by the original dataset due to severe overlapping of fuzzy boundaries, weak cytoplasmic contrast, and the presence of mucus, blood, and inflammatory cells [38]. Therefore, if the cells are extremely overlapped, the

TABLE 5. Ablation study for the design of the proposed method.

VoVNetV2	ECA	BiFPN	CBAM	DC_P	FNR_O	TPR_P	FPR_P	FPS
				0.953±0.033	0.066±0.106	0.958±0.002	0.0082±0.0153	38.6
√				0.942±0.020	0.078±0.220	0.948±0.035	0.0088±0.0230	40.7
√	√			0.955±0.068	0.066±0.154	0.959±0.047	0.0088±0.0171	**41.0**
		√		0.961±0.022	0.057±0.216	0.965±0.035	0.0080±0.0140	33.7
√	√	√		0.965±0.023	0.056±0.118	0.964±0.033	0.0082±0.0153	35.3
			√	0.956±0.047	0.063±0.149	**0.966±0.005**	0.0073±0.0220	37.7
√	√	√	√	**0.966±0.031**	**0.055±0.118**	0.964±0.035	**0.0070±0.0140**	35.1

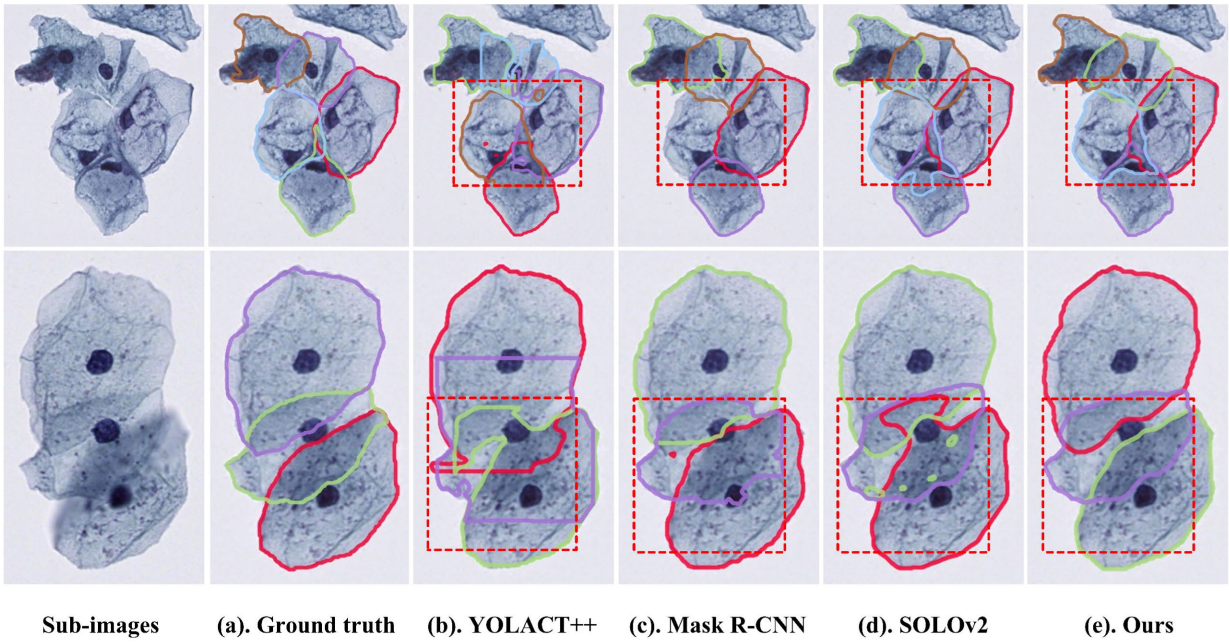

FIGURE 6. Qualitative results of overlapping cervical cell segmentation on the CISD test set (each closed curve denotes an individual instance). Rectangles show the main differences among different methods.

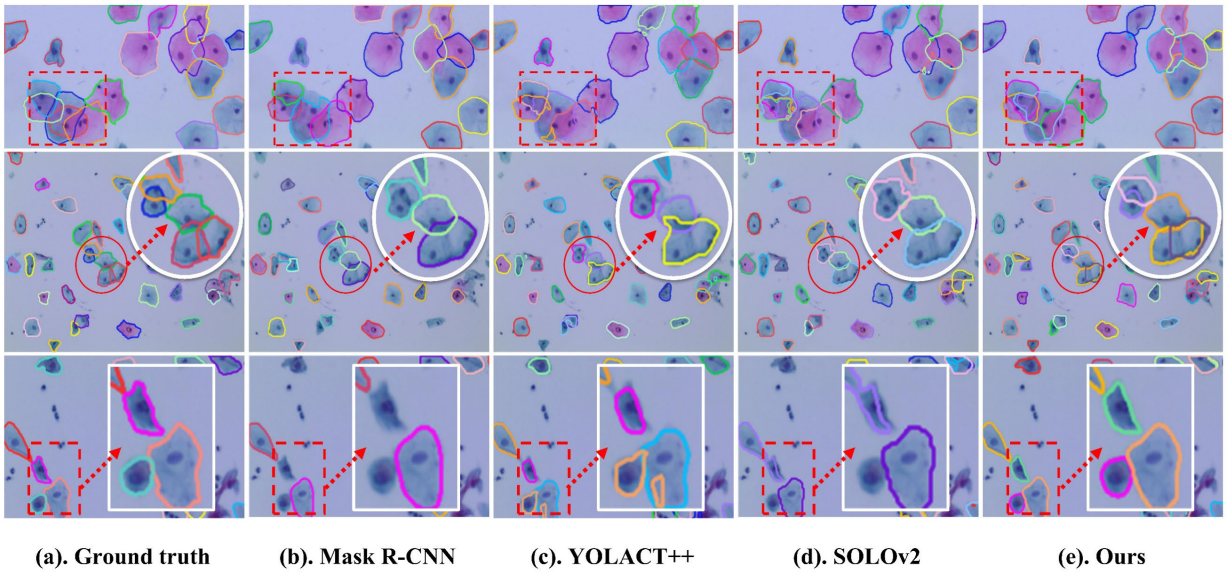

FIGURE 7. Qualitative results of overlapping cervical cell segmentation on the CX22 test set (each closed curve denotes an individual instance). The red blocks show the main differences among different methods. The white block shows a magnification of the missed area of the cell image.

performance of the proposed method decreases. In addition to this, as shown in the first row of **Figure 6**, when the cells are located at the edge of the image, the proposed method also automatically ignores the segmentation of the cells because

the cells take up too small a proportion of the picture and are incomplete.

In addition, some studies have begun to address the problem of screening cervical cell carcinoma on whole-slide images (WSIs) [48]. Although the proposed method can barely solve this problem, the processing time for segmenting large pixel images is long. Therefore, it is not practical to segment the overlapping cytoplasm of whole sections using the proposed method.

V. CONCLUSION

The proposed approach VB-SOLO in this paper is based on modifying the SOLOV2 network to perform instance segmentation of overlapping cells in cytology images. By using an end-to-end single-stage instance detection approach that does not rely on the Region Proposal Network, our method is relatively more generalizable. The effectiveness of our proposed method is demonstrated through quantitative and qualitative results, which show that our approach achieves accurate instance segmentation of overlapping cells in cytology images. The experiments performed on two publicly available datasets, namely CISD and Cx22, revealed that the VB-SOLO model attained a DC_P (Dice Coefficient for Precision) of 0.966 and 0.940, as well as a FNR_O (False Negative Rate for Overlapping) of 0.055 and 0.035 in the task of segmenting overlapping epithelial cells. Our segmentation accuracy outperforms the current mainstream single-stage instance segmentation algorithms and is equal to the two-stage instance segmentation algorithms, but the segmentation speed is faster than the two-stage instance segmentation algorithms.

These experimental results serve as compelling evidence for the efficacy of the network enhancements proposed in this study and underscore the potential of single-stage instance segmentation networks in the context of overlapping cell image segmentation. In our upcoming work, our primary research focus will be to further enhance edge contour segmentation. Simultaneously, we will utilize techniques like knowledge distillation and channel pruning to reduce the algorithm's size while maintaining its accuracy, ultimately aiming for seamless integration into embedded devices.

REFERENCES

[1] T. Wen, B. Tong, Y. Liu, T. Pan, Y. Du, Y. Chen, and S. Zhang, "Review of research on the instance segmentation of cell images," *Comput. Methods Programs Biomed.*, vol. 227, Dec. 2022, Art. no. 107211, doi: 10.1016/j.cmpb.2022.107211.

[2] C. Muralidaran, P. Dey, R. Nijhawan, and N. Kakkar, "Artificial neural network in diagnosis of urothelial cell carcinoma in urine cytology," *Diagnostic Cytopathol.*, vol. 43, no. 6, pp. 443–449, Jun. 2015, doi: 10.1002/dc.23244.

[3] L. Allahqoli, A. S. Laganà, A. Mazidimoradi, H. Salehiniya, V. Günther, V. Chiantera, S. Karimi Goghari, M. M. Ghiasvand, A. Rahmani, Z. Momenimovahed, and I. Alkatout, "Diagnosis of cervical cancer and pre-cancerous lesions by artificial intelligence: A systematic review," *Diagnostics*, vol. 12, no. 11, p. 2771, Nov. 2022. [Online]. Available: https://www.mdpi.com/2075-4418/12/11/2771

[4] N. Chantziantoniou, A. D. Donnelly, M. Mukherjee, M. E. Boon, and R. M. Austin, "Inception and development of the papanicolaou stain method," *Acta Cytologica*, vol. 61, nos. 4–5, pp. 266–280, 2017, doi: 10.1159/000457827.

[5] A. Manna, R. Kundu, D. Kaplun, A. Sinitca, and R. Sarkar, "A fuzzy rank-based ensemble of CNN models for classification of cervical cytology," *Sci. Rep.*, vol. 11, no. 1, p. 14538, Jul. 2021, doi: 10.1038/s41598-021-93783-8.

[6] A. M. Braga, R. C. P. Marques, F. N. S. Medeiros, J. F. S. R. Neto, A. G. C. Bianchi, C. M. Carneiro, and D. M. Ushizima, "Hierarchical median narrow band for level set segmentation of cervical cell nuclei," *Measurement*, vol. 176, May 2021, Art. no. 109232, doi: 10.1016/j.measurement.2021.109232.

[7] Z. Lu, G. Carneiro, and A. P. 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, doi: 10.1109/TIP.2015.2389619.

[8] A. Gharipour and A. W.-C. Liew, "Segmentation of cell nuclei in fluorescence microscopy images: An integrated framework using level set segmentation and touching-cell splitting," *Pattern Recognit.*, vol. 58, pp. 1–11, Oct. 2016, doi: 10.1016/j.patcog.2016.03.030.

[9] Z. Lu, G. Carneiro, and A. P. Bradley, "Automated nucleus and cytoplasm segmentation of overlapping cervical cells," in *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2013*. Berlin, Germany: Springer, 2013, pp. 452–460.

[10] C. Panagiotakis and A. Argyros, "Region-based fitting of overlapping ellipses and its application to cells segmentation," *Image Vis. Comput.*, vol. 93, Jan. 2020, Art. no. 103810, doi: 10.1016/j.imavis.2019.09.001.

[11] J. Zhang, Z. Hu, G. Han, and X. He, "Segmentation of overlapping cells in cervical smears based on spatial relationship and overlapping translucency light transmission model," *Pattern Recognit.*, vol. 60, pp. 286–295, Dec. 2016, doi: 10.1016/j.patcog.2016.04.021.

[12] A. Tareef, Y. Song, H. Huang, D. Feng, M. Chen, Y. Wang, and W. Cai, "Multi-pass fast watershed for accurate segmentation of overlapping cervical cells," *IEEE Trans. Med. Imag.*, vol. 37, no. 9, pp. 2044–2059, Sep. 2018, doi: 10.1109/TMI.2018.2815013.

[13] T. Wan, S. Xu, C. Sang, Y. Jin, and Z. Qin, "Accurate segmentation of overlapping cells in cervical cytology with deep convolutional neural networks," *Neurocomputing*, vol. 365, pp. 157–170, Nov. 2019, doi: 10.1016/j.neucom.2019.06.086.

[14] Y. Song, E.-L. Tan, X. Jiang, J.-Z. Cheng, D. Ni, S. Chen, B. Lei, and T. Wang, "Accurate cervical cell segmentation from overlapping clumps in pap smear images," *IEEE Trans. Med. Imag.*, vol. 36, no. 1, pp. 288–300, Jan. 2017.

[15] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. 18th Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, vol. 9351. Cham, Switzerland: Springer, 2015, pp. 234–241.

[16] W. Chen and M. Zhu, "Leukocyte segmentation method based on adaptive retinex correction and U-Net," *Comput. Math. Methods Med.*, vol. 2022, pp. 1–12, Jul. 2022, doi: 10.1155/2022/9951582.

[17] Y. Zhao, C. Fu, S. Xu, L. Cao, and H.-F. Ma, "LFANet: Lightweight feature attention network for abnormal cell segmentation in cervical cytology images," *Comput. Biol. Med.*, vol. 145, Jun. 2022, Art. no. 105500, doi: 10.1016/j.compbiomed.2022.105500.

[18] G. J. Chowdary, G. Suganya, M. Premalatha, and P. Yogarajah, "Nucleus segmentation and classification using residual SE-UNet and feature concatenation approach incervical cytopathology cell images," *Technol. Cancer Res. Treatment*, vol. 22, Jan. 2023, Art. no. 153303382211348, doi: 10.1177/15330338221134833.

[19] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2980–2988, doi: 10.1109/ICCV.2017.322.

[20] M. Vania and D. Lee, "Intervertebral disc instance segmentation using a multistage optimization mask-RCNN (MOM-RCNN)," *J. Comput. Design Eng.*, vol. 8, no. 4, pp. 1023–1036, Jun. 2021, doi: 10.1093/jcde/qwab030.

[21] P. Wang, W. Hu, J. Zhang, Y. Wen, C. Xu, and D. Qian, "Enhanced rotated mask R-CNN for chromosome segmentation," in *Proc. 43rd Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Nov. 2021, pp. 2769–2772, doi: 10.1109/EMBC46164.2021.9630695.

[22] A. Jha, H. Yang, R. Deng, M. E. Kapp, A. B. Fogo, and Y. Huo, "Instance segmentation for whole slide imaging: End-to-end or detect-then-segment," *J. Med. Imag.*, vol. 8, no. 1, Jan. 2021, Art. no. 014001.

[23] Y. Zhang, J. Chu, L. Leng, and J. Miao, "Mask-refined R-CNN: A network for refining object details in instance segmentation," *Sensors*, vol. 20, no. 4, p. 1010, Feb. 2020, doi: 10.3390/s20041010.
[24] K. Du, X. Wang, Y. Yan, Y. Lu, and H. Wang, "EGNet: A novel edge guided network for instance segmentation," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2022, pp. 3868–3872, doi: 10.1109/ICIP46576.2022.9897497.
[25] H. Liu, R. A. Rivera Soto, F. Xiao, and Y. Jae Lee, "YolactEdge: Real-time instance segmentation on the edge," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2021, pp. 9579–9585, doi: 10.1109/ICRA48506.2021.9561858.
[26] D. Bolya, C. Zhou, F. Xiao, and Y. J. Lee, "YOLACT: Real-time instance segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9156–9165, doi: 10.1109/ICCV.2019.00925.
[27] D. Bolya, C. Zhou, F. Xiao, and Y. J. Lee, "YOLACT++ better real-time instance segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 2, pp. 1108–1121, Feb. 2022, doi: 10.1109/TPAMI.2020.3014297.
[28] W. Lin, J. Chu, L. Leng, J. Miao, and L. Wang, "Feature disentanglement in one-stage object detection," *Pattern Recognit.*, vol. 145, Jan. 2024, Art. no. 109878, doi: 10.1016/j.patcog.2023.109878.
[29] T. Zhang, S. Wei, and S. Ji, "E2EC: An end-to-end contour-based method for high-quality high-speed instance segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 4433–4442, doi: 10.1109/CVPR52688.2022.00440.
[30] E. Xie, P. Sun, X. Song, W. Wang, X. Liu, D. Liang, C. Shen, and P. Luo, "PolarMask: Single shot instance segmentation with polar representation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 12190–12199, doi: 10.1109/CVPR42600.2020.01221.
[31] X. Wang, R. Zhang, T. Kong, L. Li, and C. Shen, "SOLOv2: Dynamic and fast instance segmentation," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, 2020, pp. 17721–17732.
[32] T. L. Mahyari and R. M. Dansereau, "Multi-layer random Walker image segmentation for overlapped cervical cells using probabilistic deep learning methods," *IET Image Process.*, vol. 16, no. 11, pp. 2959–2972, Sep. 2022, doi: 10.1049/ipr2.12531.
[33] H. Zhang, H. Zhu, and X. Ling, "Polar coordinate sampling-based segmentation of overlapping cervical cells using attention U-Net and random walk," *Neurocomputing*, vol. 383, pp. 212–223, Mar. 2020, doi: 10.1016/j.neucom.2019.12.036.
[34] X. Hao, L. Pei, W. Li, Y. Liu, and H. Shen, "An improved cervical cell segmentation method based on deep convolutional network," *Math. Problems Eng.*, vol. 2022, Mar. 2022, Art. no. 7383573, doi: 10.1155/2022/7383573.
[35] J. Chen and B. Zhang, "Segmentation of overlapping cervical cells with mask region convolutional neural network," *Comput. Math. Methods Med.*, vol. 2021, Oct. 2021, Art. no. 3890988, doi: 10.1155/2021/3890988.
[36] Y. Zhou, H. Chen, J. Xu, Q. Dou, and P.-A. Heng, "IRNet: Instance relation network for overlapping cervical cell segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput. Assist. Intervent.* Cham, Switzerland: Springer, 2019, pp. 640–648.
[37] A. Bouyssoux, R. Fezzani, and J.-C. Olivo-Marin, "Cell instance segmentation using Z-stacks in digital cytology," in *Proc. IEEE 19th Int. Symp. Biomed. Imag. (ISBI)*, Mar. 2022, pp. 1–4, doi: 10.1109/ISBI52829.2022.9761495.
[38] G. Liu, Q. Ding, H. Luo, M. Sha, X. Li, and M. Ju, "Cx22: A new publicly available dataset for deep learning-based segmentation of cervical cytology images," *Comput. Biol. Med.*, vol. 150, Nov. 2022, Art. no. 106194, doi: 10.1016/j.compbiomed.2022.106194.
[39] Y. Lee and J. Park, "CenterMask: Real-time anchor-free instance segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 13903–13912, doi: 10.1109/CVPR42600.2020.01392.
[40] M. Tan, R. Pang, and Q. V. Le, "EfficientDet: Scalable and efficient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10778–10787, doi: 10.1109/CVPR42600.2020.01079.
[41] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, Sep. 2018, pp. 3–19.
[42] Z. Wu, C. Shen, and A. van den Hengel, "Wider or deeper: Revisiting the ResNet model for visual recognition," *Pattern Recognit.*, vol. 90, pp. 119–133, Jun. 2019, doi: 10.1016/j.patcog.2019.01.006.
[43] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "ECA-Net: Efficient channel attention for deep convolutional neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 11531–11539, doi: 10.1109/CVPR42600.2020.01155.
[44] K. Wang, J. H. Liew, Y. Zou, D. Zhou, and J. Feng, "PANet: Few-shot image semantic segmentation with prototype alignment," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9196–9205, doi: 10.1109/ICCV.2019.00929.
[45] Z. Lu, G. Carneiro, A. P. Bradley, D. Ushizima, M. S. Nosrati, A. G. C. Bianchi, C. M. Carneiro, and G. Hamarneh, "Evaluation of three algorithms for the segmentation of overlapping cervical cells," *IEEE J. Biomed. Health Informat.*, vol. 21, no. 2, pp. 441–450, Mar. 2017, doi: 10.1109/JBHI.2016.2519686.
[46] J. Yi, P. Wu, M. Jiang, Q. Huang, D. J. Hoeppner, and D. N. Metaxas, "Attentive neural cell instance segmentation," *Med. Image Anal.*, vol. 55, pp. 228–240, Jul. 2019, doi: 10.1016/j.media.2019.05.004.
[47] G. Ghiasi, T.-Y. Lin, and Q. V. Le, "NAS-FPN: Learning scalable feature pyramid architecture for object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7029–7038.
[48] M. Hamdi, E. M. Senan, B. Awaji, F. Olayah, M. E. Jadhav, and K. M. Alalayah, "Analysis of WSI images by hybrid systems with fusion features for early diagnosis of cervical cancer," *Diagnostics*, vol. 13, no. 15, p. 2538, Jul. 2023. [Online]. Available: https://www.mdpi.com/2075-4418/13/15/2538

LICHUAN LI is currently pursuing the M.S. degree in electronic information with Xi'an University of Science and Technology, Xi'an, China. His research interests include computer vision, artificial intelligence, image processing, and optoelectronic detection.

WEI CHEN received the B.S. and M.S. degrees from Zhejiang University, and the Ph.D. degree from the University of Chinese Academy of Sciences. He is currently an Associate Professor with the School of Communication and Information Engineering, Xi'an University of Science and Technology. His current research interests include the areas of computer vision, artificial intelligence, image processing, and optoelectric detection.

JIE QI received the B.S., M.S., and Ph.D. degrees from West China Medical Center, Sichuan University. She is currently a Chief Physician with the Orthopedic Department, Shaanxi Provincial People's Hospital. Her current research interests include the areas of anatomical feature recognition, image segmentation, and image analysis.