Computational and Structural Biotechnology Journal 20 (2022) 4360–4368

COMPUTATIONAL ANDSTRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

Method article

Dual supervised sampling networks for real-time segmentation of cervical cell nucleus

Die Luo [a], Hongtao Kang [a], Junan Long [b], Jun Zhang [c], Li Chen [d], Tingwei Quan [a], Xiuli Liu [a,*]

[a] Britton Chance Center for Biomedical Photonics, Wuhan National Laboratory for Optoelectronics-Huazhong University of Science and Technology, China
[b] Department of Software Engineering, Collage of Computer Science, South-central Minzu University, China
[c] National Key Laboratory of Science and Technology on MultiSpectral Information Processing, School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, China
[d] Department of Pathology, Hubei Cancer Hospital, Tongji Medical College, Huazhong University of Science and Technology, China

ARTICLE INFO

Article history:
Received 24 April 2022
Received in revised form 9 August 2022
Accepted 9 August 2022
Available online 13 August 2022

Keywords:
Segmentation
Cervical cell
Dual supervised sampling

ABSTRACT

The morphology of the cervical cell nucleus is the most important consideration for pathological cell identification. And a precise segmentation of the cervical cell nucleus determines the performance of the final classification for most traditional algorithms and even some deep learning-based algorithms. Many deep learning-based methods can accurately segment cervical cell nuclei but will cost lots of time, especially when dealing with the whole-slide image (WSI) of tens of thousands of cells. To address this challenge, we propose a dual-supervised sampling network structure, in which a supervised-down sampling module uses compressed images instead of original images for cell nucleus segmentation, and a boundary detection network is introduced to supervise the up-sampling process of the decoding layer for accurate segmentation. This strategy dramatically reduces the convolution calculation in image feature extraction and ensures segmentation accuracy. Experimental results on various cervical cell datasets demonstrate that compared with UNet, the inference speed of the proposed network is increased by 5 times without losing segmentation accuracy. The codes and datasets are available at https://github.com/ldrunning/DSSNet.

1. Introductions

Cervical cancer is the fourth leading malignant tumor that threatens women's lives, and the deaths account for about 8% of all female cancer deaths [1]. It g-enerally takes ten years or more from precancerous lesions to cervical cancer. During this period, cervical cytology screening can detect cancerous or precancerous cells. Thus, cervical cytology screening is an effective way to reduce the incidence rate and mortality of cervical cancer [2]. redIn cervical cytology screening, cell classification is mainly based on features including the area and color of nucleus, the smoothness of nuclear membranes and nuclear-cytoplasmic ratio (N/C), and texture [3–5]. Among them, the cell nucleus is the important consideration for reporting cervical cytology. And a precise nucleus segmentation determines the performance of the final classification for some deep learning-based algorithms [4,6–9] and all classical approaches [10–13]. So, an accurate and fast segmentation of the cell nucleus is crucial in the computer-aided diagnosis systems, but it is still challenging at present. The challenge stems from the following facts. There are tens of thousands of cervical cells in one WSI, among which the number of lesion cells ranges from a few to dozens. The WSI has much poor contrast and uneven staining regions [14,15], and it is complicated to separate the lesion cells from the normal cells [6,16]. Moreover, segmenting all cell nuclei in a WSI for identifying lesion cells is highly time-consuming for some current state-of-the-art methods.

Numerous methods have been proposed for cell nucleus segmentation. Threshold [5,17] segmentation distinguishes the nucleus from its surrounding region by finding an appropriate grayscale threshold and requires image preprocessing and subsequent morphological operations to improve the segmentation accuracy. K-means [7,18] utilizes the grayscale information of cells and considers factors such as cell color, texture, and gradient

* Corresponding author.
E-mail addresses: luodie@hust.edu.cn (D. Luo), khtao@hust.edu.cn (H. Kang), 201921155123@mail.scuec.edu.cn (J. Long), junzhang@hust.edu.cn (J. Zhang), quantingwei@hust.edu.cn (T. Quan), xlliu@mail.hust.edu.cn (X. Liu).

https://doi.org/10.1016/j.csbj.2022.08.023
2001-0370/

distribution, which improves the algorithm's robustness. SVM [19,20] integrates more texture features such as spectrum, shape, and gradient features of the image in the process of nucleus segmentation, and can be applied to complex backgrounds. These algorithms are computationally fast and meet real-time requirements. However, these algorithms require stable grayscale, color, and contrast information of the dataset and are only applicable in environments where pathological slides' color and texture styles differ slightly. Pathological slides have great dispersion in color style and image quality in the actual diagnosis due to the slicing process, staining methods, and imaging instruments [21–23]. As a result, these algorithms still have some difficulties in their actual use.

In recent years, deep learning-based methods have been predominant in cell nucleus segmentation. UNet [24] is a widely used segmentation method, and its network structure has been furthermore developed. PGU-net+ [8] replaces the convolutional module of UNet with a residual structure, and the model is progressively trained using multiple scales. UNet++ [25] modifies long-range skip connection to short-range connection, enabling flexible feature fusion in the decoder, and its network is designed as a collection of UNet at different depths. nnUNet [26] automatically optimizes the network training process and can independently set the structure and hyperparameters of the network according to the characteristics of the training datasets. Mask-RCNN + CRF [9] generates robust but rough cell nucleus segmentation and then uses the conditional random field to refine the coarse segmentation. Cellpose [27] has been demonstrated to be a generalist segmentation method that originates from network structure design and train dataset construction. These methods significantly improve the segmentation accuracy from highly varied images due to their efficiency in capturing complex features of the cell nucleus. However, little attention has been paid to segmentation speed.

This paper proposes a fast segmentation network named Dual-Supervised Sampling Network (DSSNet). In the DSSNet, an autoencoder structure is adopted to perform supervised dimensionality reduction on the original image [28]. The original image is encoded into multiple compressed image blocks for feature extraction, which vastly reduces the subsequent calculation in the network. Second, residual blocks are cascaded to extract the image features without pooling operations. The feature map size remains unchanged to avoid the loss of spatial information. Finally, the ground truth boundary of the nucleus is used to supervise the up-sampling process for recovering the high-resolution information of the decoding layer. DSSNet is evaluated on a variety of cervical cell nucleus datasets. The results demonstrate that DSSNet can improve 5 times segmentation speed with the same level of segmentation accuracy as UNet.

2. Network structure

The dual supervised sampling network includes supervised down-sampling module, feature extraction module, and supervised up-sampling module, as shown in Fig. 1. In the supervised down-sampling module, the original image is supervised dimensionality reduction into 30 compressed images with 1/8 original image size. The down-sampling process is managed by the super-resolution network. These compressed images are fed into the feature extraction module for feature extraction. The feature extraction module consists of 10 cascade residual blocks, in which there is no down-sampling operation to avoid the loss of feature information. In the supervised up-sampling module, the edge detection network is designed to supervise the up-sampling process of the decoding layer features for better segmentation. Note that both super-resolution and edge detection networks are only used in training.

2.1. Supervised down-sampling module

The supervised down-sampling module consists of the down-sampling network and the super-resolution network [29], both regarded as autoencoder (Fig. 2). Different from the purpose of down-sampling in existing semantic segmentation networks, the role of this module is to obtain a set of compressed low-resolution images to replace the original images for segmentation. This operation is similar to the reverse process of video super-resolution [30,31]. The down-sampling network structure is shown in Fig. 2(a). First, the original image passes through a convolutional layer to obtain a feature map with a width of 30. Then we use two branches to compress the feature. One branch adopts maximum pooling to obtain nonlinear compression features; the other adopts two convolution operations to obtain linear compression features. We repeat the above operation three times to get 30 low-resolution images with 1/8 original image size, which is the final compressed image block. The super-resolution network (Fig. 2(b)) includes convolution layers, residual blocks, and pixel-shuffle layer [32,33]. Its function is to reconstruct the image from the compressed low-resolution image block. The reconstrued image is required to be similar to the original image as much as possible. Based on this consideration, the loss function to supervise the down-sampling process is given below.

$$SSIM(x,y) = \frac{(2u_x u_y + C_1)(2\sigma_{xy} + C_2)}{(u_x^2 + u_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \tag{1}$$

$$Loss_{SSIM} = 1 - SSIM(x,y) \tag{2}$$

where SSIM [34] is the structural similarity that measures the similarity between two images, x is the original image, and y is the image reconstructed by the super-resolution network, u_x and u_y are the mean value of the pixel values of the original image and the reconstructed image, σ_x and σ_y are the variance of the original image and the reconstructed image, and σ_{xy} is the covariance of the two images. C_1 and C_2 are constant coefficients to avoid zero in the calculation formula.

2.2. Feature extraction module

The basic structure of the feature extraction module is the residual network [35]. We modify the residual structure for better segmentation. Compared with the current structure [37,36], the modified residual block has two convolution layers side by side for increasing the receptive field of the feature layer, named wide-bottleneck residual block(W-NECK) (Fig. 3). The feature extraction module consists of a convolutional layer and ten cascade wide-bottleneck residual blocks and has no down-sampling layer, which avoids the loss of spatial information. The cascaded residual blocks will transfer spatial information block by block and increase the semantic information [38,39]. The function of the cascaded residual blocks is shown in Expression (3). r^k represents the semantic information obtained from the k-th residual block. $\phi(\cdot)$ represents the functional expression of the residual block. We can be seen from Expression (3) that the semantic information obtained by the last residual block is the superposition of the previous semantic information of different depths. Therefore, using cascade residual blocks for feature extraction can get rich semantic information.

$$r^k = \phi(r^{k-1}) + r^{k-1} = \phi(\phi(r^{k-2}) + r^{k-2}) + \phi(r^{k-2}) + r^{k-2} \tag{3}$$

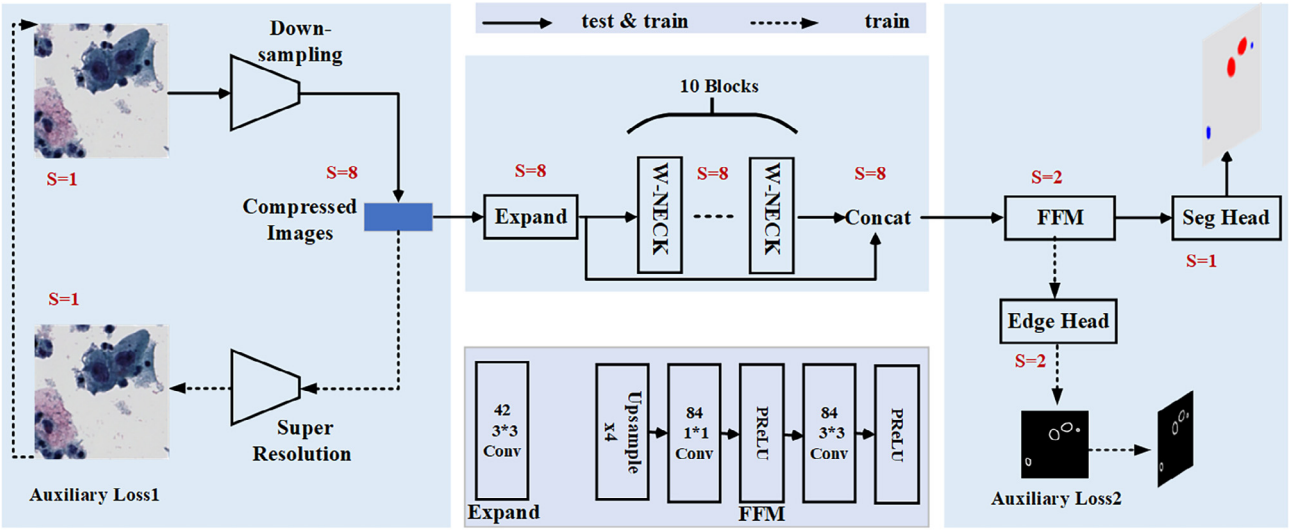

Fig. 1. Schematic diagram of network structure. S is a multiple of the down-sampling of the picture.

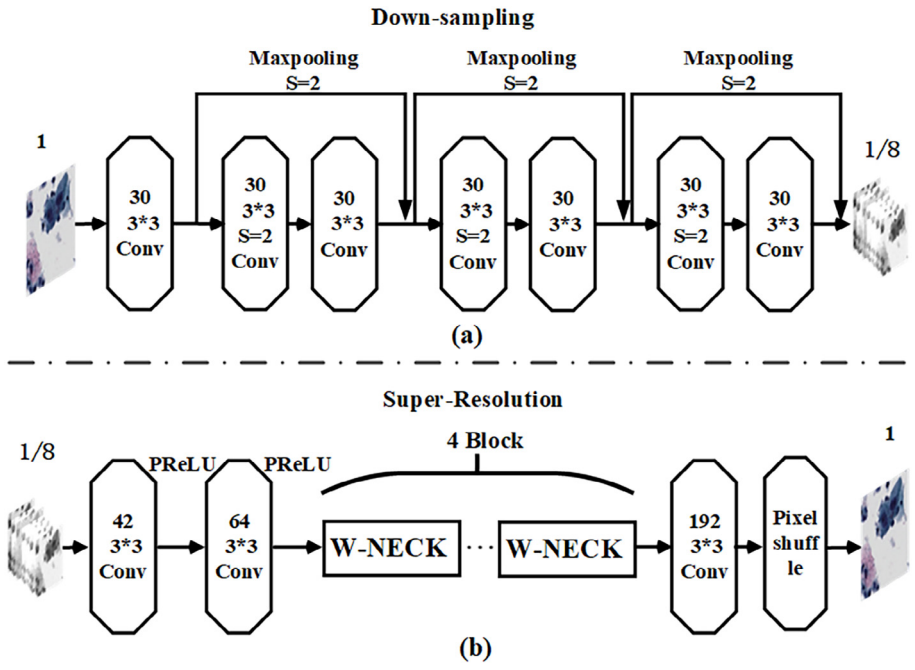

Fig. 2. Illustration of Supervised Down-sampling Module. (a) is the structure of down-sampling network. (b) is the structure of super-resolution network.

2.3. Supervised up-sampling module

The supervised up-sampling module is the decoding layer of the network. It includes feature fusion layers (FFM), segmentation network, and edge detection network [40,41]. Both edge detection and segmentation networks are composed of convolution layers (Fig. 4). We can see from the entire network structure that the input of the feature extraction module has delicate spatial information, and the output has rich semantic information. The feature fusion module (FFM) is adopted to fuse the two feature maps, as shown in Fig. 1FFM). We use edge detection to supervise the fusion process to ensure that the up-sampled decoding layer contains fine details and rich semantics. The $L1$ loss is used to measure edge detection results, and the expression is as follows (See Fig. 5)

$$Loss_{ED} = |\theta_d - g_d| \tag{4}$$

θ_d presents the manually labeled nucleus contour, and g_d presents the predicted contour. The cross-entropy is used to measure the loss of segmentation. i presents every pixel in the image, $p(i)$ presents the output expectation, and $q(i)$ presents the actual category distribution.

$$\begin{aligned} Loss_{SEG} &= H(p,q) \\ &= -\sum_i (p(i)\log q(i) + (1-p(i))log(1-q(i))) \end{aligned} \tag{5}$$

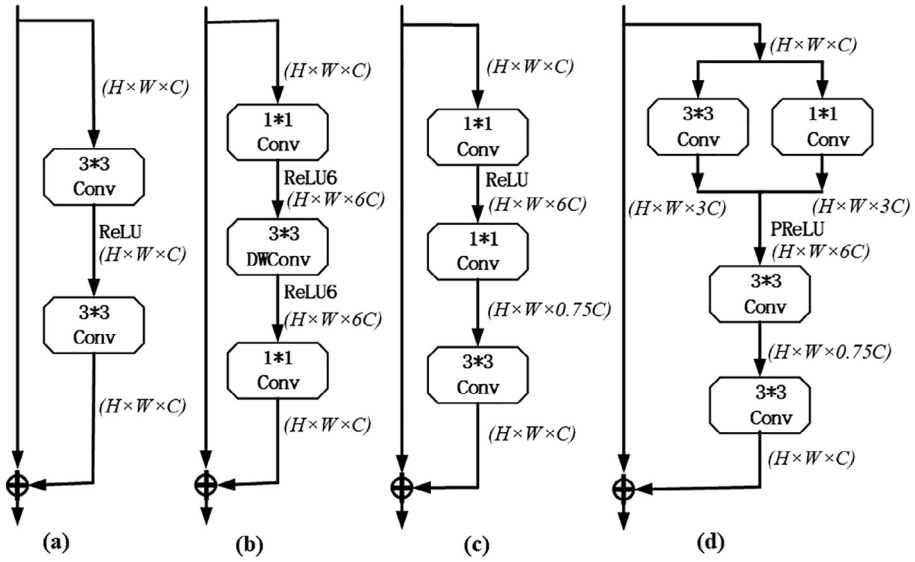

Fig. 3. Illustration of residual structure. (a) refers to the basic residual structure [35]. (b) refers to the mobile inverted bottleneck structure proposed in MobileNetV2 [36]. (c) refers to the bottleneck residual structure proposed in WDSR [37]. (d) refers to the W-NECK.

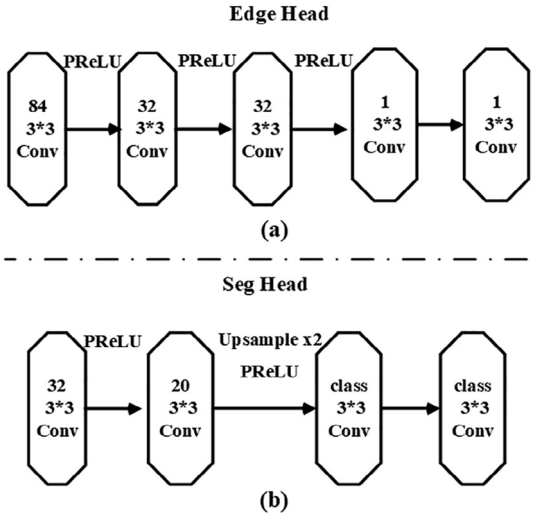

Fig. 4. Detailed design of the Edge Head and the Seg Head.

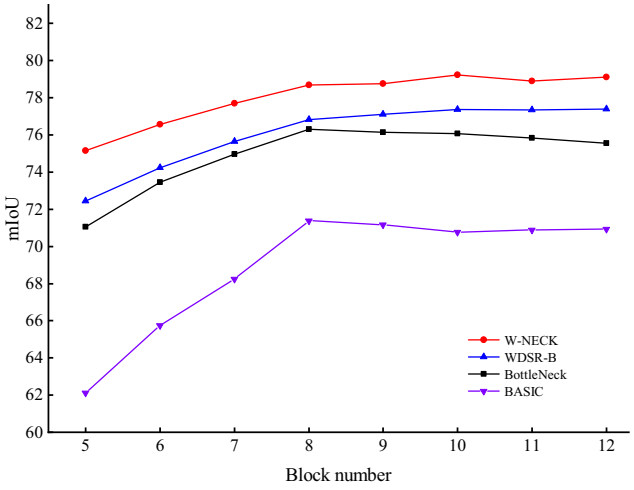

Fig. 5. Influence of different residual blocks on network accuracy.

2.4. Loss function

The loss function is divided into the structural similarity loss of super-resolution reconstruction, the $L1$ loss of edge detection, the cross-entropy loss of the segmentation result. The structural similarity loss ensures that the down-sampled feature blocks can recover the original image, which plays the same role in feature extraction for segmentation as the original image. The $L1$ loss ensures that the up-sampled feature layer has delicate boundary information and can restore high-resolution target contours. The cross-entropy loss is to supervise the feature extraction of the whole network. The hyperparameters λ_1 and λ_2 are set to be 0.25 and 0.4 in our analysis. The total losses are given below.

$$Loss = \lambda_1 * Loss_{SSIM} + \lambda_2 * Loss_{ED} + Loss_{SEG} \quad (6)$$

3. Experiments

3.1. Dataset introduction

The datasets used in the experiments are shown in Table 1. The private datasets TJ_sparse and TJ_dense are provided by Tongji Hospital and cropped from WSIs. In both TJ_sparse and TJ_dense datasets, the nucleus is divided into two categories. The former is the negative nucleus and positive nucleus whose shape is large and deformed, and the latter is the negative nucleus and positive glandular nucleus. This classification is to test the ability of DSSNet in detecting a specific type of positive nucleus. Mendeley-LBC [42] is collected from 460 participants and contains four classes of cervical cells. We select Low squamous intraepithelial lesion (LSIL) and Negative intraepithelial malignancy (NIL) from four classes for segmentation. ISBI2014 [43,44] is a part of the ISBI Cervical Cell segmentation challenge, and all those images are grayscale images with annotated nuclei. DIC–HeLa [45] is HeLa cell on a fjat glass recorded by differential interference contrast (DIC) microscopy, and it's a part of the ISBI cell tracking challenge.

3.2. Network implementation details

We conduct all experiments on PyTorch 1.6.0 with Nvidia GeForce GTX 1080Ti. We use mini-batch stochastic gradient des-

Table 1
The image segmentation datasets used in our experiments.

Dataset	Property	Image channels	Image size	Classes of nucleus	Train	Test
TJ_sparse	private	3	512 × 512	2	3200	400
TJ_dense	private	3	512 × 512	2	1200	200
Mendeley-LBC	public	3	2048 × 1536	1	600	126
ISBI2014	public	1	512 × 512	1	45	900
DIC–HeLa	public	1	512 × 512	1	84	84

cent (SGD) with momentum 0.9, weight decay 5e-4. The batch size is set to 4, and the training epoch is set to 400. The initial learning rate is 0.001, and it is decayed by cosine annealing with a period of 400.

3.3. Ablation experiment

Ablation experiments are performed on TJ_sparse. In the training and testing phases, the input size of the image is 512 × 512. The running speed of the network is calculated from the average FPS rate of 1000 iterations measured on one GPU card. The evaluation index of the network is mIoU. mIoU represents the average IoU for all classes. $IoU = \frac{P \cap G}{P \cup G}$. P represents the actual segmentation area of one class of object, and G represents the area of ground truth. The IoU coefficient represents the ratio of the intersection of the areas overlapped by the actual segmentation area and the ground truth area to their union.

3.3.1. Influence of down-sampling feature layer width on network accuracy

We quantify the effect of down-sampling feature layer width on network accuracy. Down-sampling feature layer width refers to the number of feature channels that are the output of the down-sampling network (Fig. 2). It can be regarded as the number of low-resolution images that replace the original image for segmentation. In the experiments, the down-sampling feature layer ranges from 9 to 36, while other network structures are unchanged. SSIM measures the similarity between the original image and its corresponding image reconstructed from the low-resolution images. The high SSIM suggests that the group of low-resolution images recover their corresponding original image well and thus can replace the original image for segmentation. This viewpoint is verified in Table 2. Both SSIM and mIoU increase then keep stable as the number of low-resolution images increases. Namely, the group of low-resolution images has a more substantial capacity to represent the original image. A small number of low-resolution images (<15) cannot well reconstruct the original image, which behaves as the low SSIM (<90%) and thus results in low mIoU (<76%). The prediction speed of the network changes relatively slowly because other layers are fixed and independent of the down-sampling feature layer. The down-sampling feature layer width is set to 30 for better prediction accuracy in all experiments, considering speed and accuracy.

3.3.2. Influence of the depth of the DSSNet on network accuracy

In the Dual Supervised Sampling Network, the cascade residual blocks are designed to extract the features of the group of down-sampled images. Here, the number of these residual blocks is regarded as the depth of DSSNet. In Table 3, as DSSNet depth increases, the segmentation accuracy of DSSNet strictly increases and then keeps stable. At the same time, the network's speed is slowing down rapidly. It indicates that the depth of the network needs to be set in an appropriate range. Based on the tradeoff between inference speed and accuracy, the number of cascade residual blocks is set to 10 in all experiments.

3.3.3. Segmentation performances of different residual structures

We use W-NECK, WDSR-B, BottleNeck, and basic residual block as the residual structure of the feature extraction module, respectively, and compare the segmentation performances of four modules on TJ_sparse dataset. W-NECK can be regarded as the modified WDSR-B. Both WDSR-B and W-NECK adopt expansion and squeeze block, and can realize the regularization of features, contributing to higher segmentation accuracy than the other two residual blocks. In W-NECK, the semantic information of the original feature is extracted using 1 × 1 and 3 × 3 convolution operations, while only 1 × 1 convolution operation in WDSR-B. W-NECK is easier to obtain sparse and non-sparse features and has a larger receptive field. W-NECK can provide more accurate segmentation results in comparing WDSR-B and thus employed in the DSSNet.

3.4. Effectiveness of supervising modules in networks

We verify the advantages of supervised down- and up-sampling processes. The experiments are performed on four cervical datasets, including two private and two public datasets. The backbone network and the backbone network with supervised down-sampling and the backbone network with both supervised down- and up-sampling are denoted by SOLO, SOLO + SR, SOLO + SR + E D, respectively. SOLO + SR + ED provides the best segmentation results on these four testing datasets (Table 4). To verify the effectiveness of the supervising module in the network structure, we

Table 2
Influence of different feature layer widths on network accuracy.

Width of feature layer	SSIM(%)	FPS	mIoU(%)
36	92.92	145	78.97
33	92.90	151	79.14
30	92.82	156	79.23
27	92.61	159	79.05
24	92.48	161	78.65
21	92.12	164	78.53
18	92.11	166	78.16
15	90.02	169	77.58
12	88.75	171	75.64
9	86.25	173	74.73

Table 3
Influence of the number of residual blocks on network accuracy.

Number of blocks	Parameters (M)	FPS	mIoU(%)
13	1.93	141	79.17
12	1.79	145	79.11
11	1.66	150	78.89
10	1.53	156	79.23
9	1.39	160	78.75
8	1.26	166	78.68
7	1.12	172	77.69
6	0.99	177	76.56
5	0.85	184	75.15
4	0.72	192	73.25

Table 4
Accuracy of the network under different structures.

Dataset	SOLO(mIoU)	SOLO + SR(mIoU)	SOLO + SR + ED(mIoU)
TJ_sparse	75.97 ± 1.01	78.05 ± 0.67	79.23 ± 0.54
TJ_dense	66.28 ± 1.96	68.32 ± 0.63	69.27 ± 0.53
Mendeley-LBC	72.62 ± 2.09	74.02 ± 0.82	74.58 ± 0.63
ISBI2014	82.07 ± 1.34	83.42 ± 0.83	84.45 ± 0.75

use four cervical datasets to compare the accuracy of the network segmentation under different structures. We can see from Table 4 that supervised sampling provides more accurate and stable segmentation. The super-resolution network improves the accuracy of the backbone network by 2%, and the edge detection network can further enhance the accuracy by 1%. These results suggest that SR can effectively supervise the down-sampling process, ensuring that the down-sampling operation can compress the original image and retain enough spatial information of the original image. It is also proven that ED can guide the pixel filling in the up-sampling process and ensure that the feature layer has delicate boundary information. These results indicate that the dual supervised sampling is valuable for cell nucleus segmentation.

3.5. Comparative experiments between DSSNet network with other networks

We compare the segmentation performance of DSSNet and some segmentation networks on four cervical datasets and DIC–Hale dataset (Table 5). The input size of the network on the TJ_sparse, TJ_dense, and DIC–HeLa is 512 × 512. On Mendeley-LBC and ISBI2014, the network crops and scales the original image; the final input size are 1024 × 768 and 768 × 768, respectively. The accuracy refers to the accuracy of the network on the test set. The network speeds in Table 5 are obtained by calculating the inference time with the input size of 512 × 512. DSSNet has the same level of inference speed as DFANet [48] and is much faster than other segmentation networks. Like the real-time segmentation networks, DSSNet has a small size of network parameters, which is far lower than other generic semantic networks. In the accuracy comparison, DSSNet behaves best among these real-time segmentation networks, i.e., ERFNet [47], DFANet. The accuracy of DSSNet has the same as UNet and its modified versions, but the speed is improved by 5 times. These results conclude that DSSNet is a competitive method for segmenting the nucleus.

We also show some segmentation results derived from DSSNet and UNet. In TJ_sparse dataset, we select two images in which nucleus regions are polluted (Fig. 6a&6b). DSSNet identifies the polluted nucleus and segments them while UNet fails. Red and blue represent the positive and negative nucleus, respectively. In the ISBI2014 dataset, it is demonstrated that DSSNet is superior to UNet in some low-contrast nucleus segmentation (Fig. 6c&6d). These results suggest that DSSNet can access plenty of semantic information for nucleus segmentation.

3.6. Networks test on WSIs

We evaluate the segmentation of DSSNet and UNet on WSIs. 100 WSIs collected from Tongji Hospital are used for testing. The inference time of DSSNet on one WSI is about 190s versus 900s for UNet. DSSNet segments all nuclei in the whole slice (Fig. 7a), and identifies the positive nucleus with large size and deformed shape (red label in Fig. 7b). We enlarge the subregions of the image in Fig. 7b and their corresponding nucleus segmentation. DSSNet accurately segments positive nucleus and negative cervical nuclei under the interference of mucus and inflammatory nucleus. In addition, we randomly select 8000 regions with a size of 700 × 700 from these 100 WSIs as the test set, and resize them into 512 × 512 images. We manually diagnosed and labeled all nuclei in the test set. There are 1197 positive nuclei, and the rest are negative. We calculate the segmentation accuracy of the two networks. DSSNet achieves 76.26 mIoU versus 75.15 mIoU for UNet. Meanwhile, we calculate the classification accuracy of the two networks on the negative and positive nuclei as independent instances. The results are shown in Table 6. The precision of DSSNet is much higher than that of UNet when the recall of DSSNet and UNet is close.The test results show that DSSNet can accurately segment nuclei even in complex environment, and has higher classification accuracy. At the same time, it is better than UNET in segmentation speed.

3.7. Robustness experiments

We evaluate the robustness of DSSNet on some different smear slice images. We select four datasets from different medical institutes or laboratory, and the datasets are TJ_sparse, WN, UN, and HB. TJ_sparse comes from Tongji Hospital, WN comes from Wuhan National Laboratory for Optoelectronics-Huazhong University of Science and Technology, UN comes from Wuhan Union Hospital of China, and HB comes from Hubei Provincial Women and Children's Hospital. The WN slice was prepared and stained by ourselves, following the standard protocols, and the cell suspension we used came from the remaining samples after the completion of their diagnosis and treatment process in Tongji Hospital. The four datasets are produced in with three common sedimentation methods. Among them, TJ and HB are prepared with the membrane sediment method, UN is prepared with the natural sediment method, and WN is prepared with the centrifugal sediment method. The size of each image in the four datasets is 512 × 512. And the nuclei in the datasets are manually labeled into two categories: the negative nucleus and the positive nucleus whose shape is large and deformed. TJ_sparse contains 3200 training sets and 400 test sets. The other three kinds of datasets only include 400

Table 5
Speed and accuracy comparison of DSSNet and others segmentation networks on five datasets.

Network	**Parameters (M)**	**FPS***	mIoU (%)				
			TJ_sparse	TJ_dense	Mendeley-LBC	ISBI2014	DIC–HeLa
PSPNet[[46]]	67.45	1	81.86 ± 0.28	72.43 ± 0.25	76.33 ± 0.47	–	–
UNet[[24]]	31.03	31	77.93 ± 0.48	69.14 ± 0.51	74.21 ± 0.55	84.24 ± 0.66	88.68 ± 0.58
UNet++[[25]]	36.63	14	79.57 ± 0.39	69.36 ± 0.47	74.68 ± 0.45	84.32 ± 0.62	88.63 ± 0.54
PGU-net +[[8]]	14.84	37	78.76 ± 0.59	68.89 ± 0.74	74.37 ± 0.67	83.83 ± 0.86	88.07 ± 0.87
ERFNet[[47]]	2.06	83	78.03 ± 0.86	69.09 ± 1.23	74.51 ± 1.36	83.45 ± 0.78	88.12 ± 1.45
DFANet[[48]]	2.02	120	77.37 ± 0.91	68.27 ± 0.85	73.55 ± 0.93	82.78 ± 0.75	86.86 ± 0.97
DSSNet	1.53	156	79.23 ± 0.54	69.27 ± 0.63	74.58 ± 0.69	84.45 ± 0.75	87.55 ± 0.98

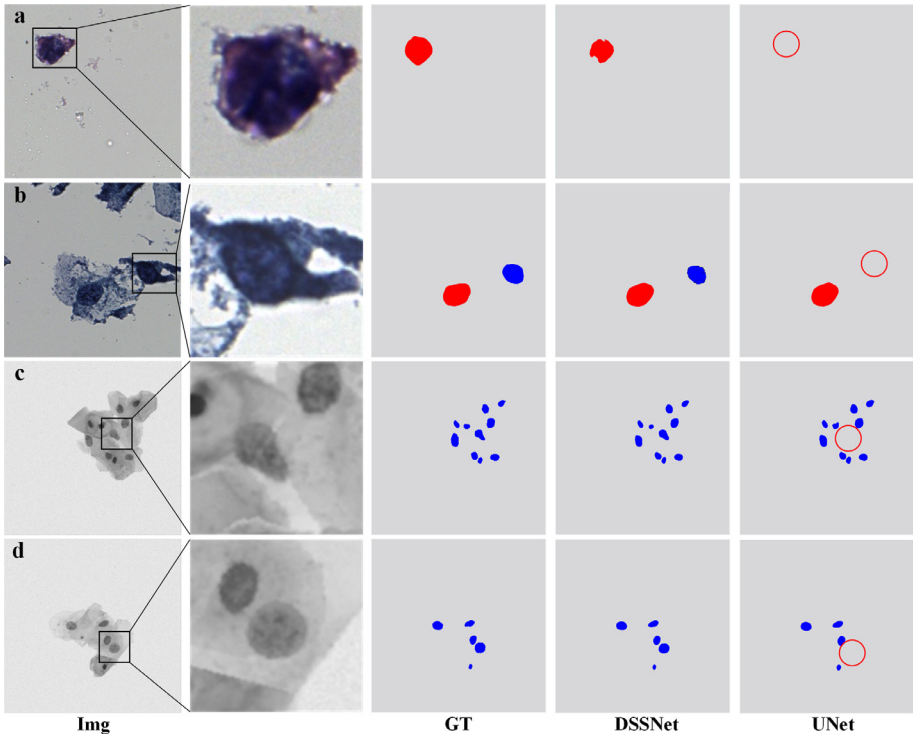

Fig. 6. Qualitative examples of the segmentation produced by DSSNet compared to the ground truth labels and UNet. From left to right: Input image, local magnification, ground-truth label, prediction of DSSNet, and prediction of UNet. Images of a and b come from the TJ_sparse test set. Images of c and d come from the ISBI2014 test set.

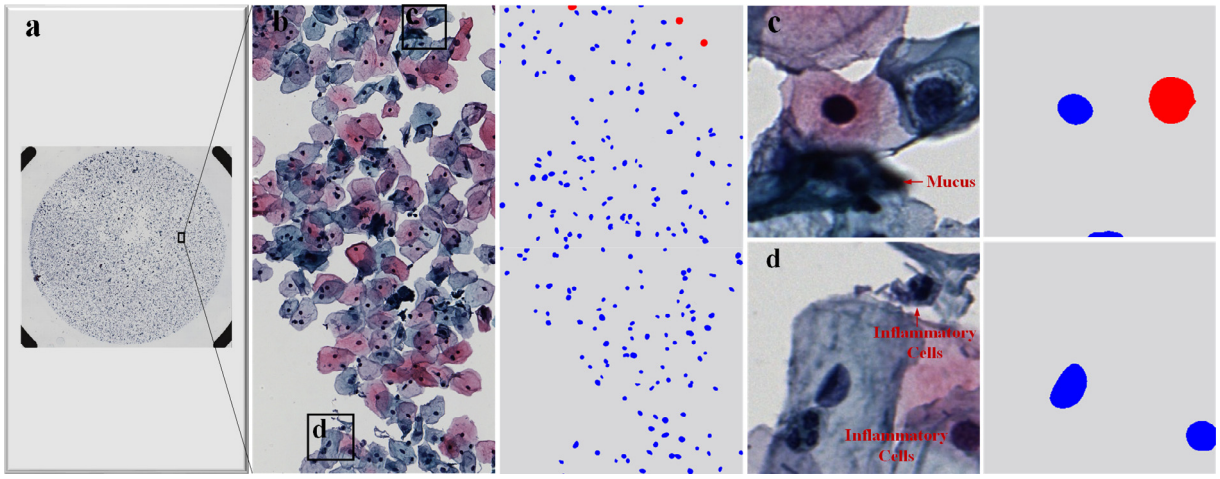

Fig. 7. A schematic diagram of network segmentation on WSI. a is one of the cervical cytology slides. b is a large field selected in a. c and d are two small regions in b.

Table 6
Classification accuracy comparison of DSSNet and UNet on WSI.

Network	Total positive	Return positive	True positive	Precision (instance)	Recall (instance)
DSSNet	1197	1216	1175	96.63%	98.16%
UNet	1197	1307	1159	89.29%	96.83%

Total positive refers to the total number of positive nuclei in the test data set. Return positive refers to the number of positive nuclei judged by the network. True positive refers to the number of true positive nuclei in the positive nuclei judged by the network.

test sets. The two models are trained on the TJ_sparse training set and tested on four test sets, respectively. The images from the four test sets are shown in Fig. 8. In the images, red and blue represent the positive with large size and deformed shape and negative nucleus, respectively. Four datasets have color style and image quality dispersion due to the different preparation methods. There are obvious differences in contrast between nucleus and cytoplasm and the background color of the whole image. The test results are shown in Table 7. The speed is the actual inference time of the network on the test set. We can see from the table that the segmenta-

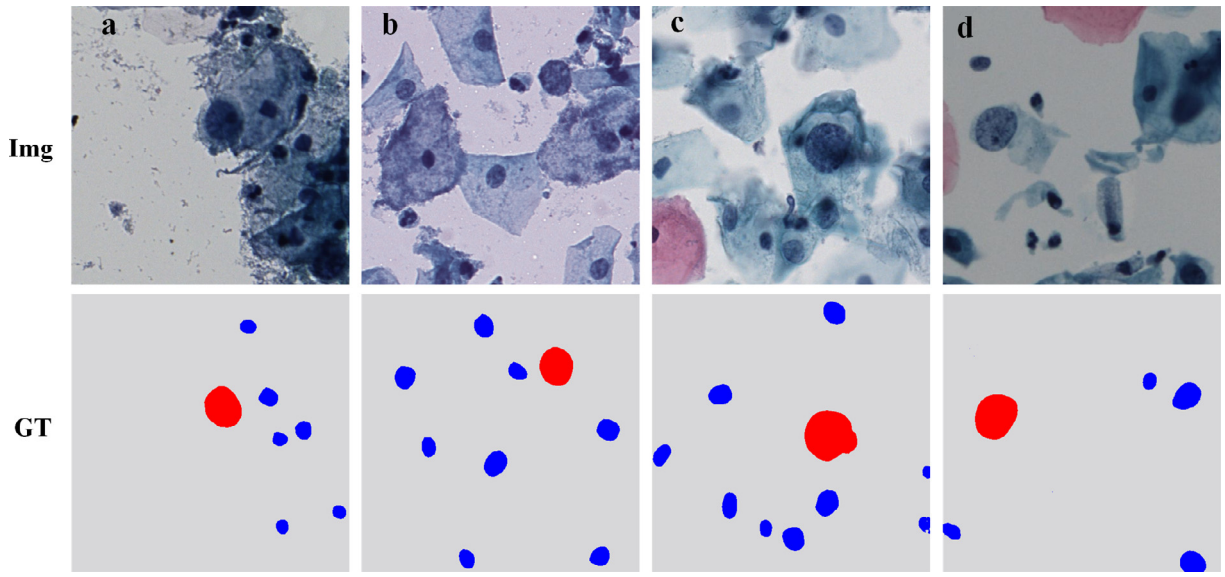

Fig. 8. Four different styles of cervical cell datasets. Images of a-d are from TJ_sparse, WN, HB, and UN datasets, respectively. The images of the first line represent the original images (Img), and the images of the second line represent their corresponding ground-truth label (GT).

Table 7
Segmentation accuracy comparison of DSSNet and UNet on different datasets.

Network	FPS	mIoU (%)			
		TJ_sparse	WN	HB	UN
UNet	31	77.93	77.15	76.24	72.97
DSSNet	156	79.23	78.43	78.28	75.16

tion accuracy of the models on the other three untrained data is close to that on TJ_sparse. The loss of segmentation accuracy is less than 5%. The results show that the trained segmentation model is robust to some untrained cervical data.

4. Conclusions

In this paper, we propose a dual-supervised network for the segmentation of cervical nuclei. The supervised down-sampling can significantly improve the model segmentation speed, and the supervised up-sampling can improve the final segmentation accuracy. The dual-supervised approach is a generic architecture, and it can be extended to detection and instance segmentation tasks. The experimental results show that compared with mainstream networks, our method considers both speed and accuracy, and has an excellent practical application prospect in the field of medical image.

CRediT authorship contribution statement

Die Luo: Conceptualization, Methodology, Software, Writing-Original draft, Formal analysis. **Hongtao Kang:** Validation, Writing- reviewing & editing. **Junan Long:** Writing- reviewing & editing. **Jun Zhang:** Investigation, Supervision. **LI Chen:** Data curation. **Tingwei Quan:** Supervision, Funding acquisition. **Xiuli Liu:** Investigation, Data curation, Supervision, Writing- reviewing & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This authors would like to thank Tongji Hospital, Huazhong University of Science and Technology, Wuhan Union Hospital of China, and Hubei Provincial Women and Children's Hospital for their support in data acquisition and lesion cells labeling. We gratefully acknowledge the support of Wuhan National Laboratory for Optoelectronics in slide scanning.This work was supported in part by the National Natural Science Foundation of China under Grant 81771913, the Science Fund for Creative Research Group of China under Grant 61721092.

References

[1] Bray F, Ferlay J, Soerjomataram I, Siegel RL, et al. Global cancer statistics 2018: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 2018;68(6):394–424.
[2] Sarwar A, Sheikh AA, Manhas J, Sharma V. Segmentation of cervical cells for automated screening of cervical cancer: a review. Artif Intell Rev 2020;53 (2):2341–79.
[3] Liu Y, Ma J, Li X, Liu X, et al. Discrimination of cervical cancer cells via cognition-based features. J Innov Opt Health Sci 2020;13(01):1–10.
[4] Song Y, Zhang L, Chen S, Ni D, et al. Accurate segmentation of cervical cytoplasm and nuclei based on multiscale convolutional network and graph partitioning. IEEE Trans Biomed Eng 2015;62(10):2421–33.
[5] Zhang L, Kong H, Chin CT, et al. Segmentation of cytoplasm and nuclei of abnormal cells in cervical cytology using global and local graph cuts. Comput Med Imaging Graphics 2014;38(5):369–80.
[6] Zhu X, Li X, Ong K, Zhang W, et al. Hybrid ai-assistive diagnostic model permits rapid tbs classification of cervical liquid-based thin-layer cell smears. Nat Commun 2021;12(1):1–12.
[7] Riana D, Tohir H, Hidayanto AN. Segmentation of overlapping areas on pap smear images with color features using k-means and otsu methods. In: 2018 Third International Conference on Informatics and Computing (ICIC), IEEE. p. 1–5.
[8] Zhao J, Dai L, Zhang M, Yu F, et al. PGU-Net+: Progressive growing of U-Net+ for automated cervical nuclei segmentation. International Workshop on Multiscale Multimodal Medical Imaging, Springer 2019:51–8.
[9] Liu Y, Zhang P, Song Q, Li A, et al. Automatic segmentation of cervical nuclei based on deep learning and a conditional random field. IEEE Access 2018;6:53709–21.
[10] Plissiti Marina E, Nikou Christophoros, Charchanti Antonia. Combining shape, texture and intensity features for cell nuclei extraction in Pap smear images. Pattern Recogn. Lett. 2011;32(6):838–53.
[11] Li Kuan, Lu Zhi, Liu Wenyin, et al. Cytoplasm and nucleus segmentation in cervical smear images using Radiating GVF Snake. Pattern Recogn 2012;45 (4):1255–64.
[12] Plissiti Marina E, Nikou Christophoros, Charchanti Antonia. Automated detection of cell nuclei in pap smear images using morphological

reconstruction and clustering. IEEE Trans Inf Technol Biomed 2010;15 (2):233–41.
[13] Cheng Jierong, Rajapakse Jagath C, et al. Segmentation of clustered nuclei with shape markers and marking function. IEEE Trans Biomed Eng 2008;56 (3):741–8.
[14] Kale A, Aksoy S.Segmentation of cervical cell images. In 2010 20th International Conference on Pattern Recognition; 2010. pp. 2399–2402.
[15] Song Y, Tan E-L, Jiang X, Cheng J-Z, et al. Accurate cervical cell segmentation from overlapping clumps in pap smear images. IEEE Trans Med Imaging 2016;36(1):288–300.
[16] Cheng S, Liu S, Yu J, Rao G, et al. Robust whole slide image analysis for cervical cancer screening using deep learning. Nat Commun 2021;12(1):1–10.
[17] Tareef A, Song Y, Cai W, Huang H, et al. Automatic segmentation of overlapping cervical smear cells based on local distinctive features and guided shape deformation. Neurocomputing 2017;221:94–107.
[18] Guan T, Zhou D, Liu Y. Accurate segmentation of partially overlapping cervical cells based on dynamic sparse contour searching and gvf snake model. IEEE J Biomed Health Inf 2014;19(4):1494–504.
[19] Zhang J, Liu Y. Cervical cancer detection using svm based feature screening. In: International Conference on Medical Image Computing and Computer-Assisted Intervention Springer. p. 873–80.
[20] Zhao M, Wu A, Song J, Sun X, Dong N. Automatic screening of cervical cells using block image processing. Biomed Eng Online 2016;15(1):1–20.
[21] Chen X, Yu J, Cheng S, Geng X, et al. An unsupervised style normalization method for cytopathology images, Computational and Structural. Biotechnol J 2021;19:3852–63.
[22] Kang H, Luo D, Feng W, Zeng S, Quan T, et al. Stain style transfer using transitive adversarial networks. Front Med 2021;8(1):1–12.
[23] Cai S, Xue Y, Gao Q, Du M, et al. Stain style transfer using transitive adversarial networks. In: International Workshop on Machine Learning for Medical Image Reconstruction. Springer; 2019. p. 163–72.
[24] Ronneberger O, Fischer P, Brox T. U-Net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-assisted Intervention Springer. p. 234–41.
[25] Zhou Z, Siddiquee MMR, Tajbakhsh N, Liang J. UNet++: Redesigning skip connections to exploit multiscale features in image segmentation. IEEE Trans Med Imaging 2019;39(6):1856–67.
[26] Isensee F, Jaeger PF, Kohl SA, Petersen J, Maier-Hein KH. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nat Methods 2021;18(2):203–11.
[27] Stringer C, Wang T, Michaelos M, Pachitariu M. Cellpose: A generalist algorithm for cellular segmentation. Nat Methods 2021;18(1):100–6.
[28] Hinton GE, Salakhutdinov RR. Reducing the dimensionality of data with neural networks. Science 2006;313(5786):504–7.
[29] Caballero J, Ledig C, Aitken A, Acosta A, et al. Real-time video super-resolution with spatio-temporal networks and motion compensation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. p. 4778–87.
[30] Kappeler A, Yoo S, Dai Q, Katsaggelos AK. Video super-resolution with convolutional neural networks. IEEE Trans Comput Imaging 2016;2 (2):109–22.
[31] Jo Y, Oh SW, Kang J, Kim SJ. Deep video super-resolution network using dynamic upsampling filters without explicit motion compensation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. p. 3224–32.
[32] Shi W, Caballero J, Huszár F, Totz J, et al. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. p. 1874–83.
[33] Lim B, Son S, Kim H, Nah S, et al. Enhanced deep residual networks for single image super-resolution. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. p. 136–44.
[34] Wang Z, Bovik AC, Sheikh HR, Simoncelli EP. Image quality assessment: from error visibility to structural similarity. IEEE Trans Image Process 2004;13 (4):600–12.
[35] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. p. 770–8.
[36] Sandler M, Howard A, Zhu M, Zhmoginov A, Chen L-C. MobileNetV2: Inverted residuals and linear bottlenecks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. p. 4510–20.
[37] Yu J, Fan Y, Yang J, Xu N, et al. Wide activation for efficient and accurate image super-resolution. arXiv preprint arXiv:1808.08718.
[38] Huang G, Liu Z, Van Der Maaten L, Weinberger KQ. Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. p. 4700–8.
[39] Chen Y, Li J, Xiao H, Jin X, Yan S, Feng J. Dual path networks. Advances in Neural Information Processing Systems 2017;30:4470–8.
[40] Li Y, Kamata S-I, Liu H. Edge-guided hierarchically nested network for real-time semantic segmentation. In: 2019 IEEE International Conference on Signal and Image Processing Applications (ICSIPA), IEEE. p. 296–301.
[41] Fan Z, Liu H, He J, Zhang M, Du X. MPDNet: A 3d missing part detection network based on point cloud segmentation. In: ICASSP 2021–2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE. p. 1810–4.
[42] Hussain E, Mahanta LB, Borah H, Das CR. Liquid based-cytology pap smear aataset for automated multi-class diagnosis of pre-cancerous and cervical cancer lesions. Data Brief 2020;30:105589.
[43] Nosrati M, Hamarneh G. A variational approach for overlapping cell segmentation. ISBI Overlapping Cervical Cytology Image Segmentation Challenge 2014:1–2.
[44] Lu Z, Carneiro G, Bradley AP. An improved joint optimization of multiple level set functions for the segmentation of overlapping cervical cells. IEEE Trans Image Process 2015;24(4):1261–72.
[45] van G. Dic-hela, the netherlands cappellen erasmus medical center, rotterdam. http://data.celltrackingchallenge.net/training-datasets/DIC-C2DH-HeLa.zip.
[46] Zhao H, Shi J, Qi X, Wang X, Jia J. Pyramid scene parsing network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. p. 2881–90.
[47] Romera E, Alvarez JM, Bergasa LM, Arroyo R. ERFNet: Efficient residual factorized convnet for real-time semantic segmentation. IEEE Trans Intell Transp Syst 2017;19(1):263–72.
[48] Li H, Xiong P, Fan H, Sun J. DFANet: Deep feature aggregation for real-time semantic segmentation, in. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. p. 9522–31.