Hierarchical Spatial Pyramid Network For Cervical Precancerous Segmentation By Reconstructing Deep Segmentation Networks

Zhu Meng [a] Zhicheng Zhao [a,b] Fei Su [a,b] Limei Guo [c]
[a] Beijing University of Posts and Telecommunications, China
[b] Beijing Key Laboratory of Network System and Network Culture, China
[c] Third Hospital, Peking University Health Science Center, China
{bamboo, zhaozc, sufei}@bupt.edu.cn, guolimei@bjmu.edu.cn

Abstract

Cervical cancer is one of the leading causes of cancer death in women aged 20 to 39 years, which emphasizes the importance of cervical precancerous diagnosis and treatment. Although there are many attempts on medical image processing, the research on the automatic diagnosis of cervical precancerous pathology is still scarce. In this paper, a challenging end-to-end automatic segmentation task for cervical precancerous diagnosis is focused. Specifically, considering that the diagnosis of cervical lesions relies heavily on spatial information, a hierarchical spatial pyramid network (HSP-Net) is proposed to enhance the representation ability of cervical structural features. First, a vertical hierarchical spatial pyramid (V-HSP) network is devised to aggregate the multiscale information during the feature extraction of the encoder. Second, a horizontal hierarchical spatial pyramid (H-HSP) network is designed to fuse information of multiscale receptive fields before and after cascading features from different branches. Experiments on the public dataset MTCHI demonstrate that HSP-Net achieves the state-of-the-art performance, reflecting the potential to assist doctors and patients clinically.

1. Introduction

Since the mid-1970s, the survival rates of all the most common cancers have improved except uterine cervix and uterine corpus; cervical cancer is one of the leading causes of cancer death in women aged 20 to 39 years with 10 premature deaths per week [31]. Fortunately, mild cervical precancerous lesions is curable when detected early, and visual inspection on hematoxylin and eosin (H&E) stained cervical pathological slides is a popular screening method. However, the giga resolution of pathological images (e.g., up to $168,960px \times 99,072px$ for one image from MTCHI dataset[1]) places high demands on the professionalism and concentration of pathologists. Therefore, automated processing with deep learning can assist pathologists in diagnosis in terms of efficiency and accuracy. Although deep learning has achieved excellent performance in natural image processing (e.g., ResNet [12] and GoogLeNet [33] in classification; Faster R-CNN [28] and Mask R-CNN [13] in detection; U-Net [29], FCN [19], and DeepLab v3+ [6] in segmentation), it still faces many challenges in the processing of pathological images: (1) Different H&E stained slides vary in hue, saturation and contrast, due to the factors such as laboratory protocols, source manufacturers, scanners, concentration, and even staining time. (2) Considering the progression of cervical lesions, different lesion grades cannot be completely accurately distinguished, and thus adjacent categories are similar to each other. (3) The diagnosis of pathologists is affected by subjectivity and experience, and pixel-level annotations exist inevitable errors, which introduces noise to the data fitting. (4) The resolution of a cervical whole slide image (WSI) far exceeds that of natural images, which greatly increases the burden of the graphics processing unit (GPU).

Deep networks have achieved outstanding performance in the classification tasks of pathological images [17]. Conventional methods usually crop a giga-WSI into multiple small patches for processing. The output diagnostic map is composed of the prediction results of the patches. Since each patch in the classification task corresponds to only one pixel in the output map, the size of the final diagnostic map is much smaller than the original image. However, the area of the cervical lesion is usually small. The diagnostic map generated by classification will miss many small lesions. Variants of U-Net are widely utilized in medical image segmentation tasks, such as brain tumor segmentation [9] and blood vessel segmentation [36]. Considering that the original

[1] https://mcprl.com/html/dataset/MTCHI.html

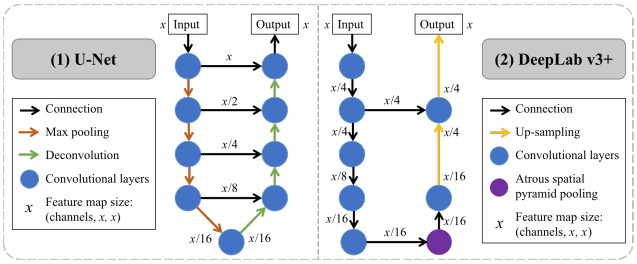

Figure 1. (1) U-Net gradually aggregates features of different scales through skip-connections to integrate contextual space information. (2) DeepLab v3+ expands the receptive fields through the atrous spatial pyramid pooling (ASPP) to obtain spatial features.

U-Net output size is smaller than the input size, the variants usually slightly modify the convolutional layers through zero padding to obtain equal-sized segmentation results. As shown in Figure 1 (1), U-Net gradually merges the information of the encoder and decoder through skip connections. In previous work, TriUpSegNet [24] implemented segmentation instead of classification based on DeepLab v3+, to ensure pixel-by-pixel diagnosis for cervical WSIs. As shown in Figure 1 (2), DeepLab v3+ applies ResNet as the encoder to extract semantic features, and ASPP to expand the receptive fields. However, typical segmentation networks have several shortcomings in the precancerous segmentation task of cervical pathological images. For example, the ASPP module of DeepLab v3+ merely expands the receptive fields of one feature map, and the integration of multiscale information is insufficient. In addition, the encoder of U-Net-like networks cannot be flexibly replaced with a common pretrained network, and the deconvolution operation increases the amount of parameters. To address above problems, we propose a HSP-Net to improve the segmentation performance of cervical pathological images. First, different from the decoder of the U-Net which cascades the features from the encoder after deconvolution gradually, a vertical hierarchical spatial pyramid (V-HSP) structure is devised to cascade the bilinear up-sampled features from different nodes of the encoder directly. Without deconvolution, V-HSP structure contains fewer parameters than the U-Net. Second, inspired by the ASPP structure of DeepLab v3+, the down-sampling blocks (DS-Blocks) are applied instead of conventional convolutions to fuse information of multiscale receptive fields before and after cascading the features output from the encoder, which forms a horizontal hierarchical spatial pyramid (H-HSP) network. Combining the V-HSP and H-HSP structures, HSP-Net achieves outstanding results in the task of cervical precancerous pathological segmentation which is very concerned about the spatial information.

The contributions can be summarized as follows:

- To obtain rich information from multiple perspective scales, the V-HSP structure is proposed to fuse multiscale features during the feature extraction of the encoder. Four output feature maps from different nodes of the encoder are first bilinear up-sampled without deconvolution and then aggregated directly instead of cascading step by step to save parameters.

- To avoid the loss of structural information caused by deep convolutions, the H-HSP structure is built. DS-Blocks with parallel convolutions of different dilations are assigned instead of conventional convolutions to aggregate information from multiscale receptive fields multifoldly.

- Experiments on the public histopathological dataset MTCHI demonstrate the effectiveness of our HSP-Net (the combination of H-HSP and V-HSP structures) through cross validation. In addition, the HSP-Net also outperforms the recently published state-of-the-art (SOTA) segmentation algorithms for cervical precancerous lesions on the test set of MTCHI dataset.

The remainder of the paper proceeds as follows: Section 2 introduces recent works about the automatic diagnosis of the cervix and pathology. Section 3 is concerned with the specific construction of HSP-Net, including the V-HSP and the H-HSP structures. Section 4 analyses the ablation experiments, cross validation, and comparison with other methods. Section 5 provides the conclusion.

2. Related Work

2.1. Cervical automatic diagnosis

In recent years, most of the explorations of cervical computer-aided diagnosis have delved in Pap smear images. The issue is often addressed through the following aspects: (1) semantic segmentation of nuclei, cytoplasm, and background; (2) accurate extraction of overlapping cell edges; and (3) abnormal identification of each cell. For example, Song et al. [32] explored the topic of overlapping cell edge extraction on the dataset of challenge ISBI 2015 [20]. Zhang et al. [39] achieved high accuracy for single-cell classification on the public dataset Herlev [15]. Although previous studies have reported the effectiveness of computer-aided algorithms in Pap smear screening, it is still insufficient on cervical pathology automatic diagnosis. Considering that the diagnosis of cervical pathology is largely dependent on structural features, the algorithms for cervical Pap smear images cannot be directly transferred to pathological precancerous segmentation. Some previous algorithms [8] [10] [1] cut some relatively simple cervical tissues into three layers along the parallel direction of the basement membrane, and extracted features to jointly

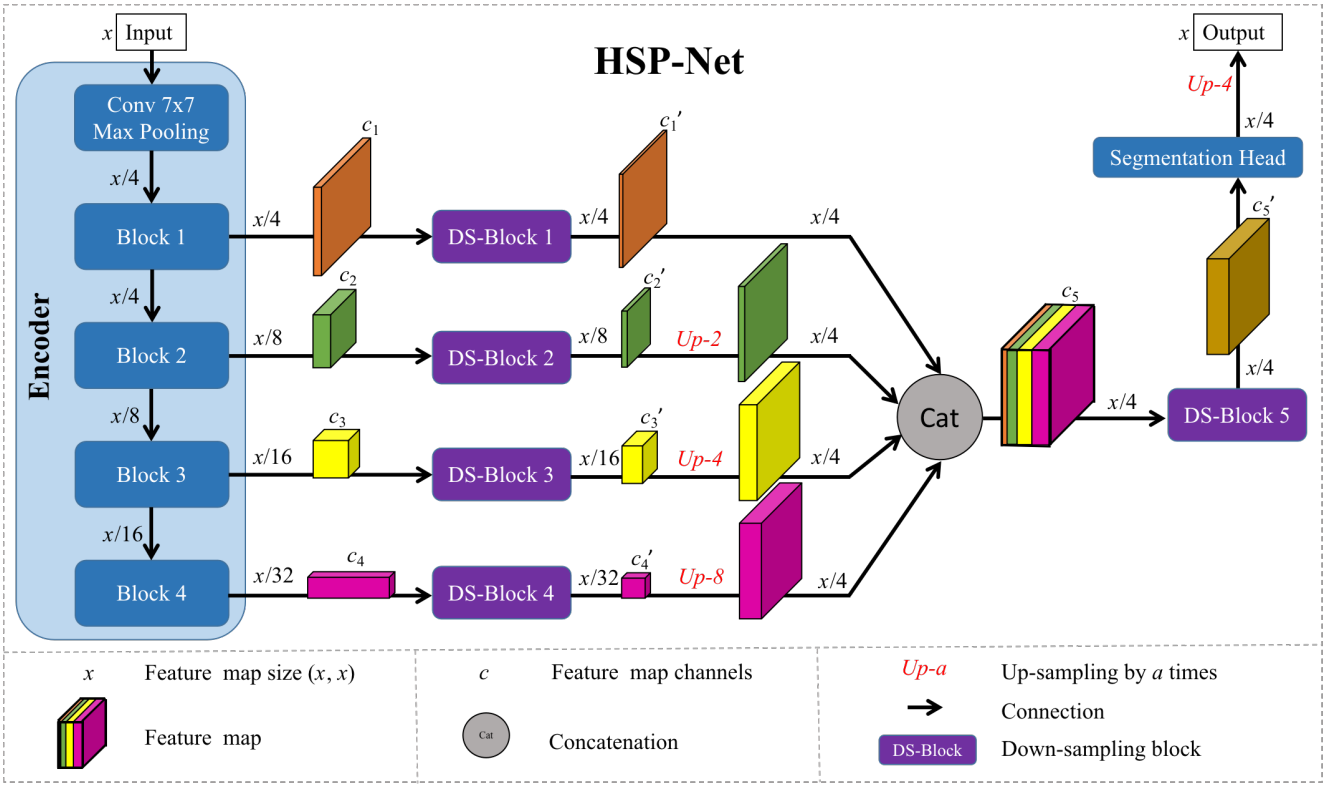

Figure 2. The architecture of HSP-Net. The encoder consists of the blocks of ResNet. The multiscale outputs from the encoder are concatenated together to form the V-HSP structure. DS-Blocks are used to reduce dimensions and aggregate features of multiscale receptive fields. Multiple DS-Blocks form the H-HSP structure.

predict the tissue progression. Wang et al. [37] determined the position of the basement membrane by adversarial neural networks. However, they manually selected samples containing basement membranes for exploration, which is unsuitable for practical applications when the tissue is incomplete. In this paper, we address the task of cervical precancerous diagnosis through the exploration of CNN structure based on a complicated pathology dataset.

2.2. Histological automatic diagnosis

Since the resolution of WSI is much higher than the input of the CNN, most of the previous algorithms usually cut WSIs into small patches containing regions of interest (RoIs) for processing. The 2018 grand challenge on breast cancer histology images (BACH) [2] attracted many attempts on breast pathology classification. They took advantage of image preprocessing, transfer learning, weakly supervised learning, and attention mechanism to classify the breast image patches with size of $2,048px \times 1,536px$ into four categories (i.e., normal, benign, in situ carcinoma, and invasive carcinoma), and achieved high accuracy [22] [16] [7]. The WSI diagnostic inference map was usually obtained by stitching the classification results of cropped patches [18] [17]. Some algorithms sent the WSI feature maps obtained from the classification network to the segmentation network, to further improve the diagnostic accuracy [34] [11]. The output diagnostic maps of these strategies with the classification network as backbones were often much smaller than the original WSI size. In addition, many outstanding instance segmentation algorithms were widely investigated for pathological patches [5] [27] [38]. In this paper, we explore the pixel-by-pixel diagnosis of pathology, i.e., semantic segmentation, to focus on small lesions in cervical tissue.

3. Methodology

3.1. HSP-Net

The overall architecture of HSP-Net is shown in Figure 2. HSP-Net gradually extracts semantic features through an encoder. Considering the scarcity of annotated cervical images, the encoder is extracted from classification networks pretrained on the dataset ImageNet [30]. Here, HSP-Net adopts the backbone of ResNet as the encoder. The encoder can flexibly choose ResNet-34 or ResNet-101 according to the trade-off between time and accuracy. For the input image X with side length x, during the feature extraction by the encoder, the side length of the feature map

gradually decreases, and the number of channels gradually increases accordingly, so as to ensure sufficient information. HSP-Net sets four intermediate results of the encoder as output nodes. The output feature maps of the encoder are $X_{En} =< X_{En1}, X_{En2}, X_{En3}, X_{En4} >$, where X_{Eni} is the output of Block i. Note that the minimum size of the output feature map of the encoder is $x/32$, which is different from $x/16$ of DeepLab v3+. The feature maps of the four branches differ in size and channel, thus converging rich multiscale information in subsequent convolutions. The numbers of X_{En} feature map channels $< c_1, c_2, c_3, c_4 >$ vary according to the complexity of the encoder, and $c_1 < c_2 < c_3 < c_4$ is always satisfied.

3.2. V-HSP

V-HSP extracts and aggregates features of different scales from the encoder. Unlike using multiple deconvolutions in the U-Net, V-HSP fuses the multiscale features from the encoder by using bilinear up-sampling which is parameter-free. The features of the four branches are cascaded together directly instead of stepwise convolutions. To further reduce the parameters and balance the multiscale information, the features of the four branches are dimensionally reduced before concatenation. The features with different scales are dimension-reduced by DS-Blocks. DS-Blocks reduce the number of channels from $< c_1, c_2, c_3, c_4 >$ to $< c'_1, c'_2, c'_3, c'_4 >$. Then, the feature maps from DS-Block 2, 3, 4 are up-sampled to the size of the feature map from DS-Block 1. Finally, the aggregation of feature maps with different scales forms the V-HSP structure. Specifically, the output is

$$X_{Cat} = \begin{array}{l} \mathcal{C}\{\mathcal{D}(X_{En1}), \mathcal{U}[\mathcal{D}(X_{En2})]^2, \\ \mathcal{U}[\mathcal{D}(X_{En3})]^4, \mathcal{U}[\mathcal{D}(X_{En4})]^8\}, \end{array} \quad (1)$$

where $\mathcal{C}(\cdot)$ denotes the concatenation of feature maps, $\mathcal{D}(\cdot)$ is the dimension reduction operation, and $\mathcal{U}(\cdot)^i$ denotes up-sampling by i times. The size of X_{Cat} is the same as that of X_{En1}. The channel number of X_{Cat} is the sum of those after dimension reduction, namely, $c_5 = c'_1 + c'_2 + c'_3 + c'_4$.

3.3. H-HSP

DS-Blocks are adopted to reduce the dimension and aggregate features with multiscale receptive fields as shown in Figure 3. To avoid the loss of spatial information during deep convolutions, five DS-Blocks are adopted to form a H-HSP structure. The input X_{In} of a DS-Block with size (c, x', x') is fed in parallel to five sub-blocks, namely, a global pooling, a convolutional layer with kernel size 1, and three atrous convolutional layers with dilations of 6, 12, and 18. The output of the global pooling is then up-sampled to the same size as other sub-block outputs. Note that the numbers of the five sub-block channels are all $c'(c' < c)$. And thus the feature map size after concatenation is $(5c', x', x')$. Then, a convolutional layer with kernel size 1 is assigned to reduce the dimension to c'. The numbers of input channels c and output channels c' of the five DS-Blocks in HSP-Net are shown in Table 1. The segmentation head contains three convolutional layers with kernel size of 3×3. Finally, the HSP-Net output is

$$X_{Out} = \mathcal{U}\{\mathcal{V}[\mathcal{D}(X_{Cat}), 3]^3\}^4, \quad (2)$$

where $\mathcal{V}(\cdot, i)^j$ denotes j convolutional layers with kernel size i. The size of X_{Out} is x, and the channel number of X_{Out} is set according to the task.

All of the convolutional layers in HSP-Net except the last one for segmentation, are activated by ReLU function and followed with a batch normalization layer.

Table 1. The input and output feature map sizes of the five DS-Blocks. The input image size of HSP-Net is $(3, x, x)$. The input and output feature map sizes of a DS-Block are (c, x', x') and (c', x', x').

		ResNet-34		ResNet-101	
DS-Block	x'	c	c'	c	c'
DS-Block 1	$x/4$	64	32	256	32
DS-Block 2	$x/8$	128	64	512	64
DS-Block 3	$x/16$	256	128	1024	128
DS-Block 4	$x/32$	512	256	2048	256
DS-Block 5	$x/4$	256	256	256	256

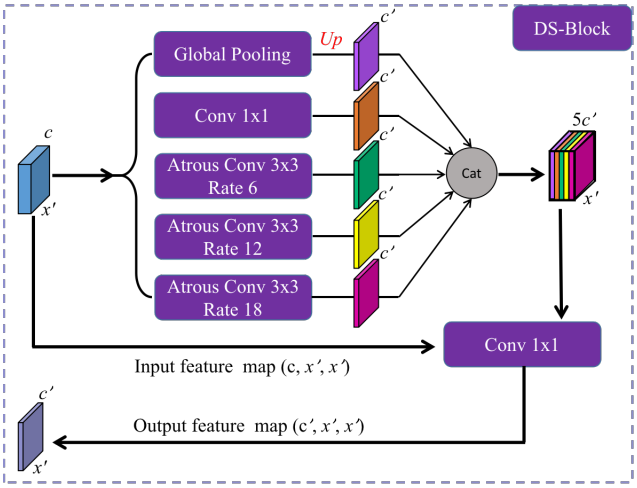

Figure 3. The DS-Block structure. The input of DS-Block is dimension-reduced by five parallel operations, including a global average pooling and four convolutions with different rates. Then the features from the five branches are cascaded to aggregate features with different receptive fields. Finally, a 1×1 convolution is applied to reduce the dimension again and obtain the output.

Table 2. Cross validation on the MTCHI dataset. HSP-Net is superior to the DeepLab v3+ without post-processing. When the Gauss-like post-processing is combined, performance of HSP-Net trained with AE-Loss is better than that of TriUpSegNet trained with DC-loss.

Network	Loss	Post	Dice	mIoU	AP
DeepLab v3+	CE-Loss	x	(0.5033±0.1218)	(0.3678±0.1098)	(0.5611±0.0984)
HSP-Net	CE-Loss	x	(0.5125±0.1551)	(0.3856±0.1457)	(0.5693±0.1306)
TriUpSegNet-A	DC-Loss	√	(0.5321±0.1473)	(0.4065±0.1399)	(0.5913±0.1363)
TriUpSegNet-B	DC-Loss	√	(0.5343±0.1458)	(0.4048±0.1389)	(0.5967±0.1261)
HSP-Net	AE-Loss	√	**(0.5416±0.1681)**	**(0.4186±0.1624)**	**(0.6086±0.1353)**

4. Experiments

4.1. Data and implementation

The experiments are conducted on the public dataset MTCHI. There are 101 cervical precancerous regions selected by pathologists for training and evaluation. The images are annotated pixel-by-pixel into normal or cervical intraepithelial neoplasia (CIN). According to the severity of precancerous lesions, CIN is further divided into CIN1, CIN2, and CIN3. The test set consists of 39 regions for evaluation. We cropped the images in the training set at $\times 10$ magnification with a size of $400px \times 400px$ and a stride of $100px$. After discarding the patches with foreground proportions of less than 20%, 7,724 image patches are used for training. The network is optimized with stochastic gradient descent, and the batch size is set to 16. The learning rate of the encoder of HSP-Net is initialized with 0.001, and decreased by using the cosine annealing strategy. The learning rates of the other blocks are ten times of that of the encoder. The experimental results of the 30th epoch are stored for comparison. The experiments are implemented via Pytorch [26], and a single NVIDIA Tesla-T4 GPU with 16 GB RAM.

4.2. Evaluation metrics

The effectiveness of the algorithms on cervical precancerous segmentation is measured by three evaluation metrics, namely, Dice coefficient, mean intersection over union (mIoU), and average precision (AP). Specifically, the three metrics are defined as follows:

$$\text{Dice} = \frac{1}{4}\sum_{i=1}^{4}\frac{2 \mid P_i \cap T_i \mid}{\mid P_i \mid + \mid T_i \mid}, \tag{3}$$

$$\text{mIoU} = \frac{1}{4}\sum_{i=1}^{4}\frac{P_i \cap T_i}{P_i \cup T_i}, \tag{4}$$

$$\text{AP} = \frac{1}{N}\sum_{j=1}^{N} x_j \begin{cases} x_j = 1 & (y_j = t_j) \\ x_j = 0 & (y_j \neq t_j) \end{cases}, \tag{5}$$

where P_i denotes the regions predicted to be category i ($i = 1, 2, 3, 4$ denote normal, CIN1, CIN2, and CIN3, respectively) and T_i denotes the truth regions; N is the number of pixels, y_j denotes the predicted category for a pixel, and t_j denotes the ground truth.

Table 3. Ablation experiments of DS-Blocks. DS-Blocks before and after the cascade are both effective. The performance is significantly improved when all five DS-Blocks are adopted.

DS 1-4	DS 5	Dice	mIoU	AP
x	x	0.5646	0.4065	0.5802
√	x	0.7029	0.5599	0.7152
x	√	0.7134	0.5673	0.7370
√	√	0.7390	0.5998	0.7533

Table 4. Experiments of the encoder and the loss function.

Encoder	Loss	Dice	mIoU	AP
ResNet-34	CE-Loss	0.6845	0.5421	0.7004
ResNet-101	CE-Loss	0.7390	0.5998	0.7533
ResNet-101	AE-Loss	0.7517	0.6150	0.7646

4.3. Ablation experiments

The ablation experiments of DS-Blocks are conducted with the same encoder (ResNet-101) and cross-entropy loss (CE-Loss). The results are shown in Table 3, where"√" denotes the aforementioned DS-Block is adopted, and "x" denotes only a convolutional layer with kernel size 1 is adopted. When DS-Block 1-4 are assigned without DS-Block 5, the results are already better than the baselines (results of U-Net and DeepLab v3+ in Table 5). The results increase when all DS-Blocks are utilized, which implies the effectiveness of the HSP-Net. Since the utility of the adaptive elastic loss (AE-Loss) has been demonstrated in cervical segmentation in [23], the AE-Loss is used to further improve the performance. As shown in Table 4, HSP-Net obtains high accuracy when the encoder is ResNet-101 because of more parameters. All of the subsequent experiments are on the basis of the ResNet-101 encoder.

4.4. Cross validation

Due to the limited number of cervical images in the test set, cross validation is conducted to demonstrate the validity of HSP-Net on the whole dataset. Similar to TriUpSegNet,

Table 5. Comparison with previous methods. HSP-Net without post-processing is obviously better than other methods without post-processing, and even better than some methods with post-processing. With Gauss-like post-processing, HSP-Net achieves better performance than the SOTA results with fewer parameters.

Network	Parameters	Loss	Post	Dice	mIoU	AP
FCN32s [19]	18.64M	CE-Loss	x	0.3882	0.2749	0.4231
U-Net [29]	31.03M	CE-Loss	x	0.4015	0.2820	0.4739
ENS-UNet [21]	34.18M	CE-Loss	x	0.4195	0.3039	0.4811
FCN16s [19]	18.64M	CE-Loss	x	0.4308	0.3214	0.4508
HookNet [35]	48.97M	CE-Loss	x	0.4652	0.3218	0.4382
Res-UNet [4]	65.45M	CE-Loss	x	0.5059	0.3770	0.5690
SegNet [3]	28.44M	CE-Loss	x	0.5260	0.4016	0.6118
UNET 3+ [14]	26.98M	CE-Loss	x	0.5600	0.4122	0.5721
DeepLab v3+ [6]	59.34M	CE-Loss	x	0.6445	0.5091	0.6921
DeepLab v3+ [6]	59.34M	AE-Loss [23]	x	0.7002	0.5569	0.7233
Ensemble-A [25]	59.35M	CE-Loss	x	0.7060	0.5626	0.7362
Ensemble-B [25]	84.13M	CE-Loss	x	0.7261	0.5829	0.7492
Ensemble-C [25]	84.13M	CE-Loss	x	0.7321	0.5910	0.7477
HSP-Net (Ours)	69.70M	AE-Loss [23]	x	0.7517	0.6150	0.7646
TriUpSegNet-A [24]	60.06M	DC-Loss [24]	√	0.7395	0.6030	0.7628
TriUpSegNet-A [24]	59.47M	DC-Loss [24]	√	0.7413	0.6043	0.7620
Ensemble-A [25]	59.35M	CE-Loss	√	0.7559	0.6255	0.7930
Ensemble-B [25]	84.13M	CE-Loss	√	0.7699	0.6404	**0.8004**
Ensemble-C [25]	84.13M	CE-Loss	√	0.7700	0.6403	0.7930
HSP-Net (Ours)	69.70M	AE-Loss [23]	√	**0.7822**	**0.6549**	0.7976

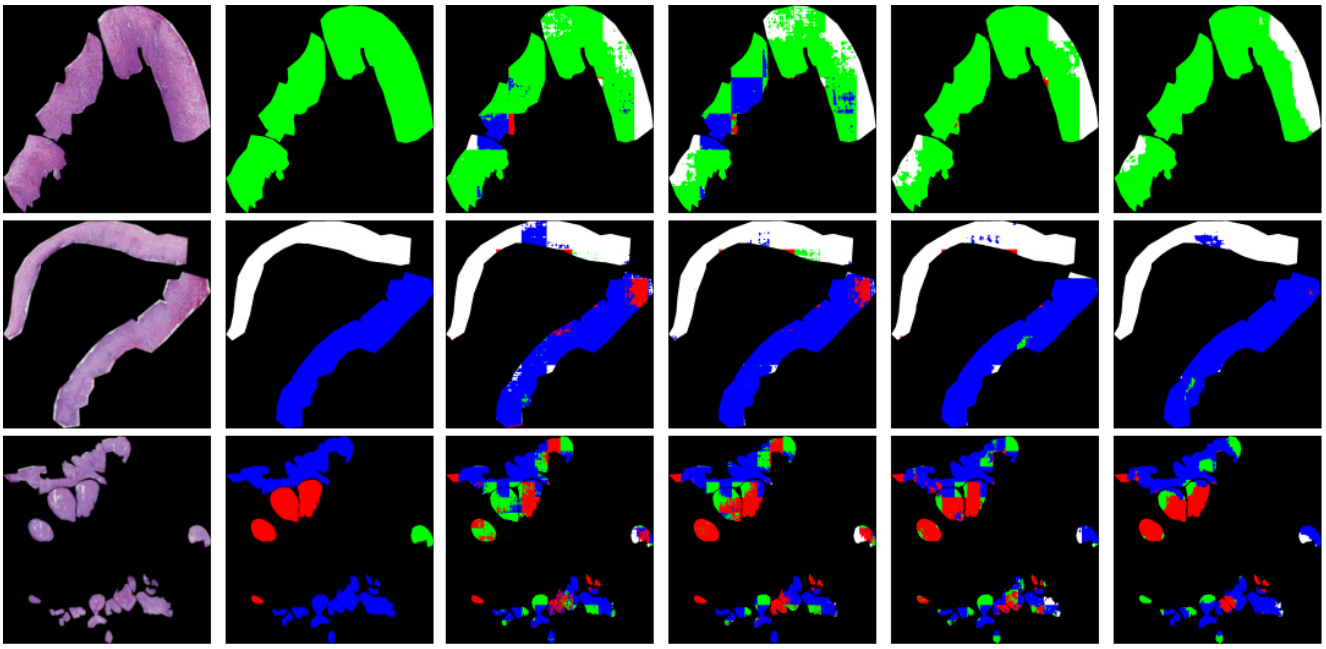

Figure 4. Three WSIs from the test set and their corresponding segmentation results. The black, white, green, blue, and red regions denote background, normal, CIN1, CIN2, and CIN3, respectively. HSP-Net performs better than other methods in detail.

a four-fold cross validation is conducted and the test set is regarded as one of the folds. TriUpSegNet is a recently published segmentation network derived from DeepLab v3+, and it achieves good performance on the MTCHI dataset when the distribution consistency loss function (DC-Loss) and Gauss-like post-processing are combined. Table 2 compares the means and standard deviations of DeepLab v3+, TriUpSegNet, and our HSP-Net. Since the original training set contains more complex samples than the test set, the average results are lower than the test set results, but the cross validation can still compare the stability of the algorithms on the whole dataset. Without post-processing and special loss functions, the HSP-Net performs better than the baseline DeepLab v3+. When the Gauss-like post-processing from [24] is applied to our HSP-Net trained with AE-Loss, the results are further improved and better than the two specific structures of the TriUpSegNet, which demonstrates the stability and superiority of the HSP-Net.

4.5. Comparison with SOTA networks

Table 5 compares HSP-Net with previous SOTA methods. The Dice coefficient of HSP-Net trained with AE-Loss without post-processing is 0.7517, which is significantly higher than the published best result 0.7321 with fewer parameters. The results of HSP-Net without post-processing are even superior to those of TriUpSegNet with Gauss-like post-processing. Ensemble-A, -B, and -C are the networks assembling classification and segmentation together. For fair comparison, only results without additional training data are included in this paper. Ensemble-A, -B, and -C with post-processing are conducted by averaging the predictions of overlapping pixels and the stride is set to 96 which is 25% of the side length. TriUpSegNet and HSP-Net adopt Gauss-like post-processing which averages Gauss-weighted predictions and the stride is set to 100. After post-processing, the Dice coefficient of HSP-Net reaches 0.7822 which is obviously higher than 0.7559 of Ensemble-A. Although Ensemble-C achieves good results, it contains more parameters than HSP-Net. To compare the performance visually, three examples from the test set are shown in Figure 4. Note that for fair comparison, all the cropped patches are stitched without overlapping except the last column. HSP-Net performs more accurate at segmentation details than others. Images in the last column are patches stitched with Gauss-like post-processing. The segmentation map exists noises at the edge of the cervical tissue due to insufficient information which can be avoided by post-processing.

5. Conclusion

In this paper, a HSP-Net is proposed to improve the accuracy of cervical pathology precancerous segmentation. Considering that the diagnosis of cervical lesions strongly depends on the spatial structures, HSP-Net focuses on the aggregation of spatial information. Specifically, the V-HSP structure is adopted to fuse multiscale features from the encoder, and H-HSP structure is combined to interweave structural information of multiscale receptive fields. The HSP-Net is simple but efficient. The experiments on the public dataset MTCHI demonstrate the effectiveness of HSP-Net. Although the accuracy of automatic diagnosis of cervical pathological images has room for improvement, it shows great potential for clinical applications with the advancement of algorithms.

Acknowledgement

This work is supported by The National Key Research and Development Program of China (2020YFB2104604), Chinese National Natural Science Foundation (U1931202, 62076033), and BUPT Innovation and Entrepreneurship Support Program (2021-YC-T031, 2021-YC-A230).

References

[1] Haidar A AlMubarak et al. A hybrid deep learning and handcrafted feature approach for cervical cancer digital histology image classification. *International journal of healthcare information systems and informatics*, 14(2):66–87, 2019.

[2] Guilherme Aresta et al. Bach: Grand challenge on breast cancer histology images. *Medical image analysis*, 56:122–139, 2019.

[3] Vijay Badrinarayanan et al. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(12):2481–2495, 2017.

[4] Kaili Cao and Xiaoli Zhang. An improved res-unet model for tree species classification using airborne high-resolution images. *Remote Sens.*, 12(7):1128, 2020.

[5] Hao Chen et al. Dcan: deep contour-aware networks for accurate gland segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2487–2496, 2016.

[6] Liang-Chieh Chen et al. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the european conference on computer vision*, pages 801–818, 2018.

[7] Sai Saketh Chennamsetty et al. Classification of breast cancer histology image using ensemble of pre-trained neural networks. In *International conference image analysis and recognition*, pages 804–811. Springer, 2018.

[8] Soumya De et al. A fusion-based approach for uterine cervical cancer histology image classification. *Computerized medical imaging and graphics*, 37(7-8):475–487, 2013.

[9] Xue Feng et al. Brain tumor segmentation using an ensemble of 3d u-nets and overall survival prediction using radiomic features. *Frontiers in computational neuroscience*, 14:25, 2020.

[10] Peng Guo et al. Nuclei-based features for uterine cervical cancer histology image analysis with fusion-based

classification. *IEEE journal of biomedical and health informatics*, 20(6):1595–1607, 2015.

[11] Zichao Guo et al. A fast and refined cancer regions segmentation framework in whole-slide breast pathological images. *Scientific reports*, 9(1):1–10, 2019.

[12] Kaiming He et al. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[13] Kaiming He et al. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[14] Huimin Huang et al. Unet 3+: A full-scale connected unet for medical image segmentation. In *IEEE international conference on acoustics, speech and signal processing*, pages 1055–1059. IEEE, 2020.

[15] Jan Jantzen and George Dounias. Special session proceedings of the nisis-2006 symposium: The pap smear benchmark. 2006.

[16] Scotty Kwok. Multiclass classification of breast cancer in whole-slide images. In *International conference image analysis and recognition*, pages 931–940. Springer, 2018.

[17] Byungjae Lee and Kyunghyun Paeng. A robust and effective approach towards accurate metastasis detection and pn-stage classification in breast cancer. In *International conference on medical image computing and computer-assisted intervention*, pages 841–850. Springer, 2018.

[18] Yun Liu et al. Detecting cancer metastases on gigapixel pathology images. *arXiv preprint arXiv:1703.02442*, 2017.

[19] Jonathan Long et al. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

[20] Zhi Lu et al. An improved joint optimization of multiple level set functions for the segmentation of overlapping cervical cells. *IEEE transactions on image processing*, 24(4):1261–1272, 2015.

[21] Zhu Meng et al. Ens-unet: End-to-end noise suppression u-net for brain tumor segmentation. In *40th annual international conference of the IEEE engineering in medicine and biology society*, pages 5886–5889. IEEE, 2018.

[22] Zhu Meng et al. Multi-classification of breast cancer histology images by using gravitation loss. In *IEEE international conference on acoustics, speech and signal processing*, pages 1030–1034. IEEE, 2019.

[23] Zhu Meng et al. Adaptive elastic loss based on progressive inter-class association for cervical histology image segmentation. In *IEEE international conference on acoustics, speech and signal processing*, pages 976–980, 2020.

[24] Zhu Meng et al. Triple up-sampling segmentation network with distribution consistency loss for pathological diagnosis of cervical precancerous lesions. *IEEE Journal of Biomedical and Health Informatics*, Early Access, 2020.

[25] Zhu Meng et al. A cervical histopathology dataset for computer aided diagnosis of precancerous lesions. *IEEE Transactions on Medical Imaging*, Early Access, 2021.

[26] Adam Paszke et al. Pytorch: An imperative style, high-performance deep learning library. In *Proc. Adv. Neural Inf. Process. Syst.*, pages 8026–8037, 2019.

[27] Hui Qu et al. Improving nuclei/gland instance segmentation in histopathology images by full resolution neural network and spatial constrained loss. In *International conference on medical image computing and computer-assisted intervention*, pages 378–386. Springer, 2019.

[28] Shaoqing Ren et al. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

[29] Olaf Ronneberger et al. U-net: Convolutional networks for biomedical image segmentation. In *International conference on medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[30] Olga Russakovsky et al. Imagenet large scale visual recognition challenge. *Int. J. Comput. Vis.*, 115(3):211–252, 2015.

[31] Rebecca L Siegel et al. Cancer statistics, 2020. *CA-A cancer journal for clinicians*, 70(1):7–30, 2020.

[32] Youyi Song et al. Segmentation of overlapping cytoplasm in cervical smear images via adaptive shape priors extracted from contour fragments. *IEEE transactions on medical imaging*, 38(12):2849–2862, 2019.

[33] Christian Szegedy et al. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

[34] Shusuke Takahama and otheers. Multi-stage pathological image classification using semantic segmentation. In *Proceedings of the IEEE international conference on computer vision*, pages 10702–10711, 2019.

[35] Mart van Rijthoven et al. Hooknet: multi-resolution convolutional neural networks for semantic segmentation in histopathology whole-slide images. *Med. Image Anal.*, 68, 2020.

[36] Bo Wang et al. Dual encoding u-net for retinal vessel segmentation. In *International conference on medical image computing and computer-assisted intervention*, pages 84–92. Springer, 2019.

[37] Du Wang et al. Adversarial neural networks for basal membrane segmentation of microinvasive cervix carcinoma in histopathology images. In *International conference on machine learning and cybernetics*, volume 2, pages 385–389. IEEE, 2017.

[38] Jingru Yi et al. Multi-scale cell instance segmentation with keypoint graph based bounding boxes. In *International conference on medical image computing and computer-assisted intervention*, pages 369–377. Springer, 2019.

[39] Ling Zhang et al. Deeppap: deep convolutional networks for cervical cell classification. *IEEE journal of biomedical and health informatics*, 21(6):1633–1643, 2017.