A Cervical Histopathology Dataset for Computer Aided Diagnosis of Precancerous Lesions

Zhu Meng, Zhicheng Zhao, *Member, IEEE*, Bingyang Li, Fei Su, *Member, IEEE*, and Limei Guo

***Abstract*— Cervical cancer, as one of the most frequently diagnosed cancers worldwide, is curable when detected early. Histopathology images play an important role in precision medicine of the cervical lesions. However, few computer aided algorithms have been explored on cervical histopathology images due to the lack of public datasets. In this article, we release a new cervical histopathology image dataset for automated precancerous diagnosis. Specifically, 100 slides from 71 patients are annotated by three independent pathologists. To show the difficulty of the task, benchmarks are obtained through both fully and weakly supervised learning. Extensive experiments based on typical classification and semantic segmentation networks are carried out to provide strong baselines. In particular, a strategy of assembling classification, segmentation, and pseudo-labeling is proposed to further improve the performance. The Dice coefficient reaches 0.7833, indicating the feasibility of computer aided diagnosis and the effectiveness of our weakly supervised ensemble algorithm. The dataset and evaluation codes are publicly available. To the best of our knowledge, it is the first public cervical histopathology dataset for automated precancerous segmentation. We believe that this work will attract researchers to explore novel algorithms on cervical automated diagnosis, thereby assisting doctors and patients clinically.**

***Index Terms*— Cervical histopathology, classification, dataset, segmentation, weakly supervised learning.**

I. Introduction

CERVICAL cancer is one of the leading causes of cancer death in women aged 20 to 39 years with 10 premature deaths per week [1]. It is observed that cervical lesion is a continuous disease from mild dysplasia to cervical cancer [2]. Fortunately, cervical precancerous lesions can be identified and treated clinically to reduce the risk of developing invasive cancer. Routinely, once a cervical lesion is detected through the examinations of Pap smear, human papillomavirus (HPV), and colposcopy, the histopathological screening considered as the gold standard, is adopted to finalize subsequent treatments. Thus, the accuracy and efficiency of cervical histopathological screening are of vital importance. However, the giga-pixels of whole slide images (WSIs) place high demands on the professionalism and concentration of pathologists. Therefore, computer aided diagnosis (CAD) is on urgent demand.

Deep learning has shown great potential in CAD since the emergence of some histopathology datasets, such as CAMELYON16 [3] and BACH [4] for breast cancer, DigestPath [5] for colon cancer, and PAIP [6] for liver cancer etc. However, a recent study of CAD for pan-cancer shows that the correlation between the algorithm of cervical cancer and pathologist-estimated tumor purity is the lowest among 42 tissue types [7], which highlights the variety of the cervix from other tissues and the necessity for special study on the cervix. Nevertheless, to the best of our knowledge, there is no specially designed histopathology public dataset for CAD of cervical precancerous lesions. The scarcity of public data further hinders the development of related algorithms. Therefore, we release a new public dataset called MTCHI to help researchers without medical background to delve and compare the automated algorithms. The MTCHI dataset contains 100 cervical WSIs at 10× magnification. Specifically, 20 WSIs containing 101 regions of interest (RoIs) are provided with pixel-level annotations, and additional 80 ones have image-level annotations. Considering diagnostic subjectivity and experience, the data are annotated into four categories (i.e., normal, CIN 1, CIN 2, and CIN 3) by three independent pathologists according to the severity of cervical lesions as described in [8].

Automated precancerous diagnosis of cervical histopathology may encounter multiple challenges. First, the acquisition location and incision direction of the biopsied tissues determine the appearance of the cervical basement membrane in the histopathology images, leading to uncertainty of spatial morphology and high demands on the ability of algorithms to identify diversity data. Second, cervical carcinogenesis is developed from mild lesion to cancer gradually, and the lesion grading is subjective without precise quantification criteria, which causes lots of annotation noises. In addition, compared

Manuscript received January 11, 2021; accepted February 10, 2021. Date of publication February 18, 2021; date of current version June 1, 2021. This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant U1931202 and Grant 62076033, in part by the Beijing Municipal Science and Technology Commission under Grant Z201100007520001 and Grant Z131100004013036, and in part by the BUPT Excellent Ph.D. Students Foundation under Grant CX2019217. *(Zhu Meng and Zhicheng Zhao contributed equally to this work.) (Corresponding author: Zhicheng Zhao.)*

Zhu Meng and Bingyang Li are with the School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing 100876, China (e-mail: bamboo@bupt.edu.cn; 2015210044@bupt.edu.cn).

Zhicheng Zhao and Fei Su are with the School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing 100876, China, and also with the Beijing Key Laboratory of Network System and Network Culture, Beijing 100876, China (e-mail: zhaozc@bupt.edu.cn; sufei@bupt.edu.cn).

Limei Guo is with the Department of Pathology, School of Basic Medical Sciences, Third Hospital, Peking University Health Science Center, Beijing 100044, China (e-mail: guolimei@bjmu.edu.cn).

Digital Object Identifier 10.1109/TMI.2021.3059699

with tissues such as breast and colon, cervical tissues usually shape into strip with small areas, resulting in the fitting difficulty of deep models. Furthermore, pixel-level annotated data are scarce, which hinders the generalization performance of the fully supervised algorithms. In this article, extensive experiments are conducted based on the analysis of image patches cropped from the WSIs, and strong baselines are provided for future algorithm comparison on the MTCHI dataset. Specifically, both fully and weakly supervised learning are considered and the ensemble of classification, segmentation, and pseudo-labeling achieves the best accuracy.

The remainder of this article is organized as follows: Section II introduces related datasets and deep learning algorithms. Section III describes our dataset construction and evaluation metrics. Section IV discusses the evaluated methods, including fully supervised and weakly supervised approaches. Section V presents the experiments and discussion. Section VI provides the conclusions.

II. Related Work

A. Previous Datasets

1) Previous Cervical Datasets: CAD on cervical Pap smear images has attracted much attention because of public datasets. For example, Herlev dataset [9] focuses on the segmentation and classification of the nucleus and cytoplasm of a single cell on Pap smear images. In addition, the ISBI 2014 [10] and ISBI 2015 [11] datasets are designed to extract the boundaries of individual cytoplasm and nucleus from real and synthetic overlapping cervical cytology images. These datasets pay much attention on the features of cytology, including the segmentation of nucleus and cytoplasm, the overlap and separation between cells, and the lesion grade of each cell. Differently, cervical histopathology images contain rich information of tissue structure, concerning about both histology and cytology. Therefore, a public cervical histopathology image dataset is urgently needed.

2) Previous Histopathology Datasets: CAMELYON16 is a large histopathology dataset about the detection of cancer metastasis on sentinel lymph nodes of breast cancer patients. It contains 400 WSIs with giga-pixels, attracting generous researchers to delve in CAD of histopathology images. PatchCamelyon [12] extracts small patches with size of $96px \times 96px$ from CAMELYON16 to assign a simple and direct binary metastasis classification task. BreaKHis [13] contains 7,909 breast cancer histopathology images ($700px \times 460px$) acquired from 82 patients for benign and malignant classification of tumors. BreastPathQ [14] scores cancer cellularity for tumor burden assessment in 2,579 breast pathology patches ($512px \times 512px$) at $20\times$ magnification. BACH attracts many algorithms via promoting a detailed microscopy breast image classification (normal, benign, in situ carcinoma, and invasive carcinoma) with patches and WSIs at $20\times$ magnification. DigestPath evaluates algorithms through signet ring cell detection from 155 patients and 872 colonoscopy tissue screening slices ($3,000px \times 3,000px$ on average) of gastric mucosa and intestine. PAIP contains 100 liver WSIs with multiple magnifications to detect and segment areas of carcinogenic cells, and calculate the area of the tumor burden. In the aforementioned datasets, cells of different grades in the breast, intestine, and liver possess distinct morphological differences. The exploration on pan-cancer has shown that algorithms that perform well in other cancers encounter obstacles in cervical lesions [7]. Therefore, a public cervical histopathology image dataset is required to specifically explore the CAD of cervical precancerous lesions.

B. Related Methods

1) CAD for Cervical Histopathology Images: The CAD of cervical precancerous histopathology images strongly depends on the extraction of structural features, which is extremely challenging. Thus, some algorithms attempted to start with simple samples. For example, algorithms from [15]–[17] ingeniously selected and divided simple samples into upper, middle, and lower layers parallel to the basement membrane. Each layer was further cut into several small patches to extract features such as color, texture, cell distribution, and deep learning semantic information. These features were finally fused to determine the lesion grade of the whole tissue. However, accurately cutting a sample into three layers is complicated in practical applications. Wang *et al.* [18] achieved basement membrane segmentation via a generative adversarial network [19], but the basement membrane is occasionally invisible because of incomplete tissue. All of the abovementioned experiments were conducted on the basis of private datasets.

2) Fully Supervised Learning for Histopathology Images: Convolutional neural networks (CNNs) have achieved remarkable results in natural image processing; accordingly, the transfer learning for automatic processing of histopathology images has become increasingly popular. The WSIs were often first cropped into patches with a small size by sliding a window. The patches were then classified or segmented through CNNs, thereby stitching into a diagnostic result map. For example, Fu *et al.* [7] performed a pan-cancer computational histopathology analysis with Inception-v4 [20]. The lightweight ShuffleNet [21] was used in [22] to identify microsatellite instability and mismatch-repair deficiency in colorectal tumors. ResNet-34 [23], VGG-16 [24] and Inception-v4 were adopted in [25] to detect the invasive breast cancer. Wang *et al.* [26] utilized GoogLeNet [27] with hard example guided training to locate tumors in breast and colon images. HookNet [28], a semantic segmentation network derived from U-Net [29], was designed to aggregate patch features of multiple resolutions for breast and lung cancer. In particular, many outstanding algorithms have been proposed since the establishment of CAMELYON16 dataset. Liu *et al.* [30] achieved high accuracy with Inception-v3 [31]. Lin *et al.* [32] proposed the ScanNet based on a modified VGG-16 network by replacing the last three fully connected layers with fully convolutional layers to avoid the boundary effect of network predictions. Takahama *et al.* [33] extracted patch features from a classification model (GoogLeNet) and then input them into a segmentation model to obtain probability heatmaps. Guo *et al.* [34] proposed a similar method, but only applied the segmentation stage to the tumor regions detected by the

classification model. Khened *et al.* [35] assembled U-Net and DeepLab v3+ [36] together to improve the diagnosis performance.

3) Weakly Supervised Learning for Histopathology Images: Recently, pseudo-labeling [37] has been recognized as an effective way to utilize unlabeled histopathology data. Routinely, the limited labeled data are trained first, and then the unlabeled patches are predicted by the model. If the confidence probability of a patch exceeds the threshold, it is added to the training set, and the predicted category is regarded as the positive pseudo-label. Tokunaga [38] implemented weakly supervised learning by generating both positive and negative pseudo-labels according to the proportion of the adenocarcinoma subtypes. Li *et al.* [39] generated self-loop uncertainty as a pseudo-label to augment the training set and boost the organ segmentation accuracy. Cheng *et al.* [40] assembled predictions of similar patches as the pseudo-label of a given patch to counteract its noisy label. Chaw *et al.* [41] trained a teacher model with labeled data first and generated pseudo-labels for the unlabeled data to train a student model, then fine-tuned the student model on original labeled dataset. The teacher-student loop was applied iteratively for reducing the annotation burden of colorectal tissue samples.

III. The MTCHI Dataset

A. Dataset Construction

The data in the MTCHI dataset are provided by Singularity.AI Technology. 400 cervical histopathology slides were collected in China and scanned with pixel size $0.226\mu m$ at $40\times$ magnification in 2018. Then 100 characteristic slides from 71 patients who underwent cervical biopsy were carefully selected for research purpose only. The MTCHI dataset contains only digital images without any patient-related personal information. All the experiments were performed under the approval of ethnic guidance of the hospitals. The selected data are all hematoxylin and eosin stained slides containing normal or precancerous cervical tissues. Considering that pathologists usually diagnose cervical lesions at $10\times$ magnification or lower resolution, the scanned images are down-sampled for 4 times and stored in PNG (Portable Network Graphics) format in 3-channel RGB (Red-Green-Blue) for exploration. The data are divided into two subgroups. One subgroup contains 20 slides for fully supervised learning. The other contains 80 slides for weakly supervised learning. Specifically, considering the complexity of tissue structure and cervical lesions, 101 RoIs without background from 20 slides are pixel-level annotated, which is suitable to explore fully supervised classification and segmentation algorithms. These 20 slides are subdivided into a training set and a test set in the ratio of 15:5. To be fair, the patients of the training set and the test set are independent from each other. Additional 80 slides are obtained from 53 different patients. These slides are annotated with image-level multi-labels. Actually, image-level annotations are easier to obtain than pixel-level annotations; thus coarse-grained information is valuable for clinical applications to improve the performance of the algorithms. Because the resolution of histopathology slides exceeds the load capacity of the graphics processing unit (GPU), slides are often cropped to small-sized patches for processing. However, image-level annotations are only responsible for the partial regions in the slide; thus, the cropped patches cannot match the label accurately and contain a lot of noise, which requires weakly supervised algorithms. The data and annotations of MTCHI dataset are publicly available in this website.[1]

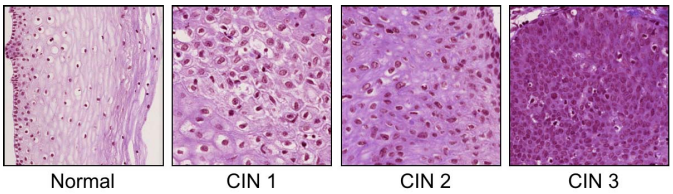

Fig. 1. Characteristic image patches of normal cervical tissue, and lesions of CIN 1, CIN 2, and CIN 3 in the MTCHI dataset.

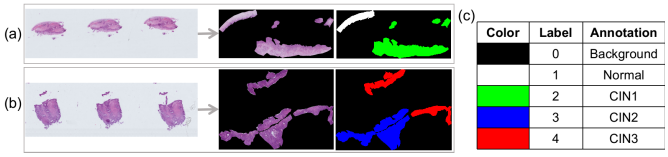

Fig. 2. Data with pixel-level annotations are obtained by cropping the RoIs from the slides. The pixels outside the regions are regarded as background. (a) and (b) are examples of data acquisition. The data are annotated pixel by pixel according to the description (c).

B. Reference Annotation Protocol

Cervical tissues in MTCHI dataset are grading into normal or cervical intraepithelial neoplasia (CIN) as described in [8]. According to the dysplastic degrees of cervical precancerous lesions, CIN is further divided into three grades, i.e., CIN 1, CIN 2, and CIN 3. As shown in Fig. 1, representative features of normal (a), mild dysplasia (CIN 1) (b), moderate dysplasia (CIN 2) (c), and severe dysplasia (CIN 3) (d) are shown. Due to the subjectivity of CIN diagnosis, the data in the MTCHI dataset are annotated by three independent experienced pathologists. Annotator A has 5 years of diagnostic experience, while Annotator B and Annotator C are experts with more than 10 years of diagnostic experience.

1) Pixel-Level Annotation: The 101 key regions cropped from 20 slides are annotated pixel by pixel. Since the pixel-level annotation plays an important role in the performance of both fully and weakly supervised algorithms, these data are directly annotated by Annotator B and checked by Annotator C. As shown in Fig. 2, the RoIs are first cropped from the slides, then each pixel is annotated with a label $i (i \in \{0, 1, 2, 3, 4\})$, where 0, 1, 2, 3, 4 represent background, normal, CIN 1, CIN 2, and CIN 3, respectively. The masks containing pixel-level annotations are stored with the same length and width as the images, i.e., gray scale masks in PNG format.

2) Image-Level Annotation: The 80 slides from 53 patients with image-level annotations are independently annotated by Annotator A and Annotator C. Each slide is provided with

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

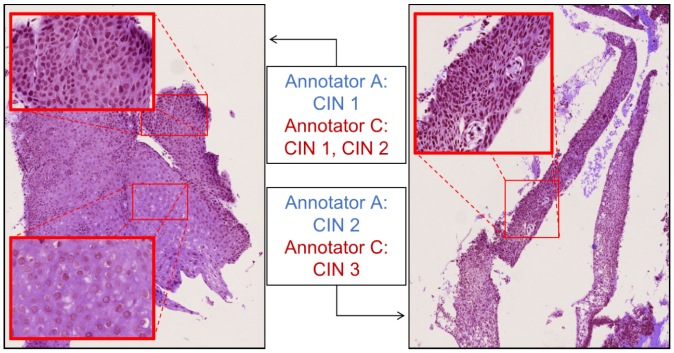

Fig. 3. Contrast of image-level annotations from Annotator A and Annotator C.

multiple labels. For example, a slide is annotated as "CIN 1 and CIN 2", if both lesions can be found in the slide with obvious corresponding regions. Therefore, a slide has at least one label and up to four labels (i.e., normal, CIN 1, CIN 2, and CIN 3). However, cervical tissue is diverse in morphology, and the diagnoses are almost experiential. Annotator A and Annotator C work independently. The annotation differences of slides by the two annotators are contrasted in Fig. 3. Only 51.25% of the slides are identified consistently by the two annotators, indicating the difficulty of cervical precancerous diagnosis. In two diagnostic results, the information from the professional Annotator C is recommended. The multi-labels for these data indicate that the slides contain the relevant regions, but do not specify the corresponding locations. When high resolution slides are cut into small patches for processing, the label of each patch is unknown. Therefore, the 80 slides are valuable materials for unsupervised and weakly supervised algorithms.

C. Evaluation Metrics

The performance of the algorithms are evaluated by pixel-level ground truth. Four evaluation metrics are applied to compare the automated diagnostic results with the ground truth to fairly measure the performance of the models. First, the Dice coefficient, a commonly used evaluation metric in medical image segmentation tasks, is used to measure the degree of coincidence between the prediction and the truth. The Dice coefficient is defined as

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

where P_i denotes the regions predicted to be category i ($i = 1, 2, 3, 4$ denotes normal, CIN 1, CIN 2, and CIN 3, respectively) and T_i denotes the truth. Second, mean Intersection over Union (mIoU), a commonly used evaluation metric in natural image segmentation, is also applied in this task. The definition of mIoU is slightly different from the Dice coefficient. mIoU is defined as

$$\text{mIoU} = \frac{1}{4}\sum_{i=1}^{4}\frac{P_i \cap T_i}{P_i \cup T_i}. \tag{2}$$

Third, average precision (AP) which is used to calculate the average accuracy of pixel classification, can be expressed as

$$\text{AP} = \frac{1}{N}\sum_{j=1}^{N} x_j \begin{cases} x_j = 1, & (y_j = t_j) \\ x_j = 0, & (y_j \neq t_j), \end{cases} \tag{3}$$

where N is the number of pixels, y_j denotes the predicted category, and t_j denotes the ground truth. Finally, different pathologists may annotate the difficult samples as adjacent different categories. Thus, window precision (WP) which is proposed to ignore the deviation of adjacent false prediction, can be defined as

$$\text{WP} = \frac{1}{N}\sum_{j=1}^{N} x_j \begin{cases} x_j = 1, & (||y_j - t_j|| <= 1) \\ x_j = 0, & (||y_j - t_j|| > 1). \end{cases} \tag{4}$$

Note that forcing the category of the entire region by counting the categories of most pixels is not recommended because many complex regions have been removed in advance for the test images. Although the tricky approach yields better results, such method is not optimal means for future exploration. In addition, the background is ignored in the evaluation.

IV. Evaluated Methods

Considering that the high resolution of WSIs exceeds the load capacity of GPUs, experiments on the MTCHI dataset are carried out by image patch analysis. The slides from the pixel-annotated training set are first cropped into small patches ($400px \times 400px$) with an overlapping stride of $100px$. Then the patches with foreground proportions of less than 20% are discarded before training. The remaining 7,724 patches are used for training the fully supervised classification and segmentation, as well as extracting pseudo-labeled patches from the image-annotated slides for weakly supervised learning. The networks used for classification and segmentation are described in Section IV-A and Section IV-B. The strategy of assembling classification and segmentation is introduced in Section IV-C. The weakly supervised learning strategy for MTCHI is discussed in Section IV-D. Post-processing is presented in Section IV-E.

A. Fully Supervised Classification

Popular classification networks are adopted in the fully supervised experiments, including MobileNet-v2 [42], VGG, GoogLeNet, Inception-v3, DenseNet [43], and ResNet. All classification networks are initialized with parameters pre-trained on the ImageNet [44]. For a pixel-annotated patch, the classification truth is the category with the largest area except background, i.e., one of normal, CIN 1, CIN 2, and CIN3. The learning rate is initialized with 0.001, and decreased by using the cosine annealing strategy. The experimental results of the 30th epoch are stored for comparison when cross-entropy loss (CE-loss) is used to constrain the model fitting, while the 50th ones are stored when adaptive elastic loss (AE-loss) [45] is used. The outputs of the classification networks are the confidence probabilities of the four categories, and the category with the largest confidence probability is regarded as the prediction result. Each patch

corresponds to a single diagnostic result. Therefore, it is necessary to fill the whole patch to obtain the diagnosis heatmap of the same size as the input patch. The structures of the classification networks are described below.

1) MobileNet-v2: MobileNet-v2 contains very few network parameters to complete approximately the same function as traditional convolutions, thereby accelerating the feedforward process. Specifically, depthwise separable convolutions are embedded instead of regular convolutions. An expansion layer is assigned before the depthwise convolution to expand the feature channels, and a projection layer is assigned after it to reduce the dimensions. The expansion, depthwise, and projection layers form a bottleneck residual block. Multiple blocks are stacked to extract patch features.

2) VGG: Convolutional layers with the same kernel size of 3×3 are stacked in the VGG network. The number of feature-map channels is increased step by step through the convolutional layers. The feature-map size is down-sampled five times through five max pooling layers with a stride of 2. Three fully connected layers are assigned at the end of the VGG network to obtain the classification results. In addition to the 5 max pooling layers and the 3 fully connected layers, VGG-16 contains 8 convolutional layers and VGG-19 contains 11 convolutional layers. Due to the large number of parameters, VGG has a good ability to extract features but is time-consuming.

3) GoogLeNet and Inception-v3: GoogLeNet, also known as Inception-v1, contains fewer parameters than VGG. It is composed of multiple Inception modules. An Inception module aggregates convolutional layers ($1 \times 1, 3 \times 3, 5 \times 5$), pooling operations ($3 \times 3$) to deal with different scales. Dimension reductions and projections are judiciously applied before the convolutions with large kernel sizes. As to Inception-v3, the Inception modules are improved through the factorization into small convolutions.

4) DenseNet: DenseNet-121 and DenseNet-169 are stacked by dense blocks. In a dense block, each layer takes all preceding feature-maps as input. Each layer can access the gradient directly from the loss and the original input image, enabling implicit deep supervision.

5) ResNet: Different from VGG-19, ResNet replaces the max pooling layers except the first one using convolutions with a stride of 2. For ResNet-18 and ResNet-34, two 3×3 convolutional layers constitute a residual block. The input of the residual block is connected to the output of the convolutional layers to avoid the gradient disappearance during training. When the networks are deeper (e.g., ResNet-50 and ResNet-101), the residual block is composed of three convolutional layers with kernel sizes of 1×1, 3×3, and 1×1. Multiple residual blocks are stacked to extract features.

B. Fully Supervised Segmentation

When there are multiple categories in one patch, it is an inaccurate practice to take the category with the most pixels in the patch as the label, so the pixel-wise classification (i.e., the semantic segmentation) is required. The semantic networks used for experiments in this article include FCN [46], SegNet [47], DeepLab v3+, U-Net and its variants such as ENS-UNet [48], Res-UNet [49], UNET3+ [50], and HookNet. The learning rate is initially set to 0.02 and decreased by a factor of 0.5 after every 10 epochs. The networks are all trained by CE-loss. The output of the semantic segmentation network has the same length and width as the input image. Semantic segmentation can be regarded as a pixel-level classification, and thus, each pixel has a corresponding prediction result. The segmentation network structures are introduced below.

1) FCN: The fully connected layers of VGG are replaced by convolutional layers, constituting a fully convolutional network (FCN). During feature extraction process by VGG, feature-map size is reduced from x to $1/32x$. In FCN32s, the feature-map with size $1/32x$ is return to the original size through a deconvolution layer to obtain segmentation results. FCN16s first up-samples the feature-map at size $1/32x$, then sums it to the feature-map at size $1/16x$, and finally recovers it to the input size by a deconvolution layer.

2) SegNet: SegNet employs the structure of encoder-decoder to implement semantic segmentation. Here, VGG-16 is regarded as an encoder to extract features. The decoder consists of convolutional layers and up-sampling operations. During the max pooling in the encoder, pooling locations are stored as indexes for the up-sampling of the decoder.

3) U-Net: U-Net was originally designed for cell segmentation, and its output size is smaller than the input size. To obtain the diagnostic heatmap with the same size as the input patch, all convolutional layers in U-Net are modified with zero padding. The encoder extracts high-dimensional semantic features by max pooling and convolutional layers, while the decoder gradually restores the feature size through convolutional and deconvolutional layers. Four sets of different scale features in the encoder and decoder are cascaded by skip connections to improve the position accuracy of the segmentation results.

4) U-Net Variants (ENS-UNet, Res-UNet, UNET3+): ENS-UNet inserts a noise suppression block in every skip connection path of U-Net. Res-UNet replaces all convolutional layers and skip connections in U-Net with residual blocks. In the decoder of UNET3+, features are densely concatenated with all shallow features from the encoder.

5) HookNet: The U-Net-like structure without skip connections is treated as one branch of the HookNet. The structures of the two branches in the HookNet are exactly the same, but the input patches have different fields of view. Specifically, a $400px \times 400px$ patch is resized to $284px \times 284px$ as input to the first context branch. The same patch is center cropped by $284px \times 284px$ as input to the target branch. The second layer of the decoder in the context branch is center cropped and cascaded to the first layer of the decoder in the target branch. Since the channel number in the HookNet-x is elevated gradually, the output channel number x of the first convolutional layer determines the parameter quantity. Experiments on the basis of HookNet-16 and HookNet-64 are conducted in this article.

6) DeepLab v3+: DeepLab v3+ adopts ResNet as the encoder to extract features. The high-dimensional features

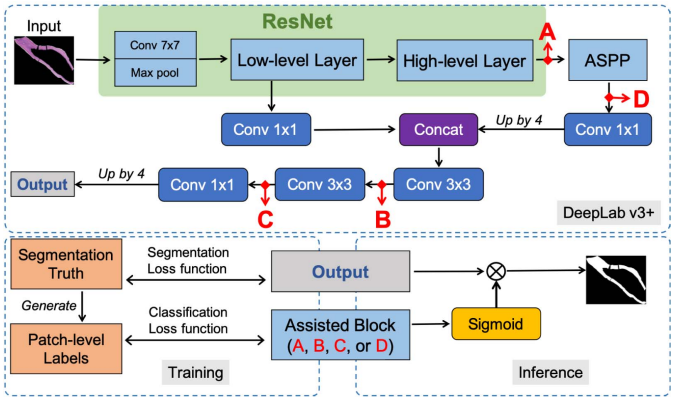

Fig. 4. The ensemble architecture of classification and segmentation is constructed on the basis of DeepLab v3+. The assisted block can be assigned to point A, B, C, or D. The parameters of the assisted block are trained with classification loss according to patch-level labels. In the inference process, the outputs of the segmentation branch and the assisted block branch are combined by multiplication.

from the end of the encoder are fed into an Atrous Spatial Pyramid Pooling (ASPP) module. The ASPP module consists of a global pooling, a 1×1 convolution, and three 3×3 atrous convolutions with rates of 6, 12, and 18, thereby aggregating features from multiple fields of view. The outputs from the ASPP are first up-sampled through bilinear interpolation, then cascaded with the features from the first layer of the encoder. The feature-map which combines the location and the semantic information is up-sampled by four times to obtain the final semantic segmentation results.

C. Fully Supervised Ensemble

On the one hand, the block effect exists in the results of the classification because all pixel in a patch are predicted as the same class. On the other hand, pepper noise in the outputs of segmentation inevitably undermines the consistency of the results. Therefore, we combine classification and segmentation together to make them complementary. To make the parameters as few as possible, classification and segmentation are not implemented by two models separately. Instead, they share the same encoder to extract features. For convenience, the segmentation network DeepLab v3+ is combined with classification assisted blocks to realize the ensemble of classification and segmentation. As shown in Fig. 4, the assisted block can be connected after the classification backbone with pretrained parameters, that is, points A, B, C, or D after ResNet. At point A and D, the assisted block is composed of a global average pooling and a fully connected layer for calculating the classification loss. At point B and C, the assisted block is composed of three 3×3 convolutional layers, a global average pooling, and a fully connected layer. The parameters of the assisted block are constrained by the CE-loss, while the previous shared parameters are jointly optimized by classification and segmentation loss functions, thereby helping to improve the fitting performance of the segmentation model. To make more use of the information learned by the assisted block, during the inference process, the output of the assisted block is normalized to [0,1] by the sigmoid function, and then used to adjust the segmentation output distribution by multiplication. The final output will balance the advantages of classification and segmentation.

Algorithm 1: Construction of the Weakly Supervised Training Set With Pseudo-Labels

Input: Images with pixel-level annotations I_{pixel}, images with image-level annotations I_{img} and their labels T_{img}.
Output: Training set S for weakly supervised learning.
Parameters: Classification network Net, maximum limits on the number of pseudo-labeled patches N_p, proportion of pseudo-labeled patches involved in training $r\%$.
Step 1: Construct a pure training set S_0. I_{pixel} are cropped into patches with overlapping and patches with foreground less than 20% are discarded.
Step 2: Net is trained with S_0 to obtain a feature extraction model M.
Step 3: Calculate the number of times c_i that class $i (i \in [1, 4])$ appears in T_{img}.
Step 4: Select appropriate pseudo-labeled patches on every image $I_x \in I_{img}$ with multi-labels $T_x \in T_{img}$. **for** I_x *in* I_{img} **do**
    1. Crop overlapping patches on I_x;
    2. Predict the confidence probabilities for each patch;
    for i *in* T_x **do**
        1. Confidence probability ranking of all patches for category i;
        2. The first $N_p/(4 \times c_i)$ patches with high probability are added to the candidate pseudo-label set S_p.
    end
end
Step 5: $N_p \times r\%$ patches are selected.
foreach *[Normal, CIN 1, CIN 2, CIN 3]* **do**
    1. The corresponding confidence probabilities of the patches in S_p are reranked.
    2. The first $N_p \times r\%/4$ patches with high probability are retained and the rest removed.
end
Step 6: $S = S_0 + S_p$.

D. Weakly Supervised Learning

Pixel-level annotations are laborious, while image-level annotations are relatively abundant. Therefore, algorithms based on image-level annotations are essential to continuously improve model performance. However, cervical histopathology images are large and the labels only represent the category without the specific location. Therefore, the labels of the cropped patches are unknown. Pseudo-labeling strategy is adopted in our experiments to take advantage of the weak data. Specifically, the new training set with pseudo-labels is constructed by the process described in Algorithm 1. For weakly supervised classification, the pseudo-label is the category with the largest confidence probability from the feature extractor M. For weakly supervised segmentation, the ground truth is

TABLE I
RESULTS OF FULLY SUPERVISED CLASSIFICATION. OVERLAPPING IS BETTER THAN NON-OVERLAPPING FOR MOST NETWORKS. MOBILENET-V2 WORKS THE FASTEST, RESNET-101 WORKS THE BEST FOR NON-OVERLAPPING, AND VGG-19 WORKS THE BEST FOR OVERLAPPING

Loss	Network	Para	FLOPs	Non-overlapping				50% Overlapping				75% Overlapping			
				Dice	mIoU	AP	WP	Dice	mIoU	AP	WP	Dice	mIoU	AP	WP
CE-loss	MobileNet-v2	2.23M	0.94G	0.6456	0.4955	0.6508	0.9432	0.6953	0.5527	0.7100	0.9372	0.7183	0.5771	0.7307	0.9514
	VGG-16	134.28M	45.32G	0.6323	0.4902	0.6854	0.9530	0.6448	0.5160	0.7042	0.9599	0.6629	0.5372	0.7237	0.9593
	VGG-19	139.59M	57.56G	0.6916	0.5587	0.7249	**0.9636**	0.6716	0.5473	0.7183	0.9739	0.6807	0.5571	0.7236	**0.9793**
	GoogLeNet	5.60M	4.43G	0.5683	0.4265	0.6325	0.9412	0.5841	0.4636	0.6847	0.9411	0.5924	0.4750	0.6934	0.9545
	Inception-v3	21.79M	9.47G	0.6419	0.4810	0.6643	0.9318	0.7015	0.5499	0.7281	0.9285	0.7106	0.5640	0.7414	0.9409
	DenseNet-121	6.96M	8.47G	0.5778	0.4539	0.6512	0.9513	0.5913	0.4771	0.6852	0.9483	0.6041	0.4931	0.6989	0.9579
	DenseNet-169	12.49M	10.04G	0.6362	0.5153	0.7147	0.9368	0.6518	0.5381	0.7406	0.9445	0.6714	0.5597	0.7587	0.9493
	ResNet-34	21.29M	10.80G	0.6044	0.4622	0.6517	0.9023	0.6476	0.5186	0.7011	0.9200	0.6639	0.5398	0.7226	0.9271
	ResNet-50	23.52M	12.10G	0.5937	0.4581	0.6593	0.9216	0.6309	0.5071	0.7108	0.9330	0.6515	0.5301	0.7317	0.9384
	ResNet-101	42.51M	23.05G	**0.7218**	**0.5850**	**0.7472**	0.9417	0.6883	0.5475	0.7177	0.9337	0.7054	0.5658	0.7320	0.9391
AE-loss	MobileNet-v2	2.23M	0.94G	0.6591	0.5103	0.6663	0.9386	0.7099	0.5708	0.7244	0.9499	0.7172	0.5817	0.7321	0.9563
	VGG-16	134.28M	45.32G	0.6367	0.5100	0.7083	0.9457	0.6572	0.5380	0.7235	0.9411	0.6752	0.5621	0.7477	0.9456
	VGG-19	139.59M	57.56G	0.7055	0.5701	0.7321	0.9609	**0.7239**	**0.5887**	**0.7519**	**0.9760**	**0.7476**	**0.6187**	**0.7765**	0.9789
	GoogLeNet	5.60M	4.43G	0.5874	0.4477	0.6420	0.9319	0.6175	0.4987	0.7051	0.9388	0.6268	0.5116	0.7102	0.9487
	Inception-v3	21.79M	9.47G	0.6520	0.4919	0.6698	0.9318	0.7199	0.5714	0.7381	0.9424	0.7321	0.5870	0.7494	0.9513
	DenseNet-121	6.96M	8.47G	0.5959	0.4787	0.6869	0.9428	0.6061	0.5079	0.7116	0.9433	0.6124	0.5119	0.7167	0.9547
	DenseNet-169	12.49M	10.04G	0.6015	0.4857	0.6875	0.9082	0.6257	0.5192	0.7313	0.9000	0.6297	0.5297	0.7420	0.9010
	ResNet-34	21.29M	10.80G	0.6519	0.5165	0.7082	0.9170	0.6311	0.5099	0.7078	0.9285	0.6773	0.5608	0.7502	0.9349
	ResNet-50	23.52M	12.10G	0.6147	0.4848	0.6936	0.9201	0.6627	0.5391	0.7436	0.9293	0.6656	0.5442	0.7499	0.9303
	ResNet-101	42.51M	23.05G	0.7003	0.5498	0.7154	0.9438	0.6981	0.5574	0.7219	0.9348	0.7228	0.5838	0.7446	0.9470

the mask filled with the pseudo-label. In addition, assigning the same pseudo-label to an entire patch is coarse with false pixel-level labels. Thus, for the ensemble of classification and segmentation, pseudo learning is only applied to the classification branch to balance the information and noise. Considering that the assisted blocks at point B and C need more parameters, only the block at point A is used for weakly supervised ensemble. First, the mixed training set S is used to train the ResNet backbone and the assisted block for 30 epochs. Second, the weights of the classification branch are fixed and the rest segmentation layers are trained by using the original training set S_0 for 10 epochs with the learning rate of 0.001. In inference process, the output of the assisted block is normalized by the sigmoid function and multiplied with the segmentation output. Note that the backbone contains atrous convolution to aggregate the information of multiple receptive fields, namely, the classification branch in the ensemble is not exactly the same as ResNet. Thus, it is retrained instead of directly loading the weights of weakly supervised ResNet.

E. Post-Processing

The diagnosis results of patches are stitched into the entire diagnostic map. A simple method is to stitch non-overlapping patches (0% overlapping). However, there are noises in stitching due to insufficient information at the edge of the patch. Therefore, the patches are cropped with overlapping. Specifically, the network output of a pixel $\vec{P} =< p_1, p_2, p_3, p_4 >$ denotes the confidence probability of < normal, CIN 1, CIN 2, CIN 3>. When N pixels overlap, the fusion confidence probability is $\vec{P} = \frac{1}{N}\sum_{k=1}^{N}\vec{P}_k$. The final diagnosis result of the pixel is the category with the highest confidence probability in $\vec{P}$. The larger the overlap area is, the more context information is gathered in each pixel, but with more computing resources.

V. RESULTS AND DISCUSSION

A. Fully Supervised Learning

1) Fully Supervised Classification: The classification networks described in Section IV-A are all adopted for the experiments on the MTCHI dataset. As shown in Table I, although the networks play their advantages in different aspects, there are obvious differences in speed and accuracy when applied to cervical precancerous CAD. MobileNet-v2 runs fastest because its floating point operations (FLOPs) is only 0.94G. Although Inception-v3 and ResNet-101 are very deep, they are still faster than VGG-19 because of their special connection modes, namely, the Inception block and the residual block. The training performance of AE-loss is better than that of CE-loss in most experiments, because AE-loss fits the relationship between categories of cervical precancerous lesions better. Generally overlapping post-processing achieves good results because it makes up for the lack of foreground information of one patch. However, there are also some special cases of overlapping post-processing which leads to the decrease of accuracy. For example, in Fig. 5 (b), the wrong prediction of a patch misleads the diagnosis of adjacent patches. In Fig. 5 (c), although the whole region is annotated as CIN 1 by pathologists, some regions are actually normal tissues which can be clearly displayed by 75% overlapping. Although the overlapping post-processing in (c) is closer to the real situation, it inevitably causes the loss of accuracy when compared with the ground truth.

2) Fully Supervised Segmentation: The results of fully supervised segmentation networks are fairly compared without overlapping in Table II. Compared with U-Net, the

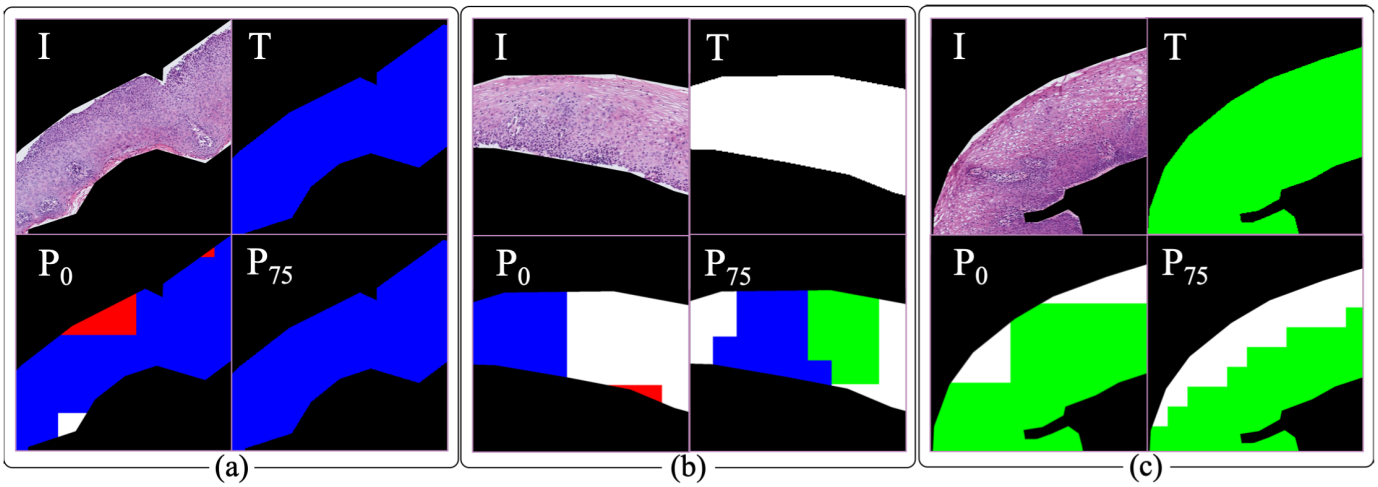

Fig. 5. Three examples from ResNet-101 trained with CE-loss. *I*, *T*, P_0 and P_{75} represent the input image, the ground truth, non-overlapping result and 75% overlapping result, respectively. The white, green, blue and red parts represent normal tissue, lesions of CIN 1, CIN 2 and CIN 3 respectively. (a) The 75% overlapping makes up for the lack of edge information in non-overlapping. (b) The wrong prediction results mislead the judgement of adjacent patches in overlapping post-processing. (c) The upper left corner of the sample is closer to the normal tissue rather than CIN 1, and the result of overlapping is closer to the real situation than the annotation.

TABLE II
RESULTS OF FULLY SUPERVISED SEGMENTATION AND ENSEMBLE WITH ASSISTED BLOCK. ENSEMBLE-A, B, AND C REPRESENT THE RESULTS OF ASSISTED BLOCK AT POINT A, B, AND C

Overlap	Network	Para	FLOPs	Dice	mIoU	AP	WP
0%	HookNet-16	3.07M	2.76G	0.3804	0.2609	0.3700	0.7140
	FCN32s	18.64M	45.25G	0.3882	0.2749	0.4231	0.8841
	U-Net	31.03M	54.20G	0.4015	0.2820	0.4739	0.8523
	ENS-UNet	34.18M	125.74G	0.4195	0.3039	0.4811	0.8784
	FCN16s	18.64M	45.25G	0.4308	0.3214	0.4508	0.8716
	HookNet-64	48.97M	43.37G	0.4652	0.3218	0.4382	0.7649
	Res-UNet	65.45M	744.56G	0.5059	0.3770	0.5690	0.9203
	SegNet	28.44M	109.94G	0.5260	0.4016	0.6118	0.9140
	UNET3+	26.98M	447.05G	0.5600	0.4122	0.5721	0.9148
	DeepLab v3+	59.34M	49.97G	0.6500	0.5180	0.7035	0.9167
	Ensemble-A	59.35M	49.97G	0.7060	0.5626	0.7362	0.9561
	Ensemble-B	84.13M	66.28G	0.7261	0.5829	0.7492	0.9423
	Ensemble-C	84.13M	66.28G	0.7321	0.5910	0.7477	0.9471
50%	Ensemble-A	59.35M	49.97G	0.7450	0.6109	0.7807	0.9601
	Ensemble-B	84.13M	66.28G	0.7671	0.6362	0.7954	0.9543
	Ensemble-C	84.13M	66.28G	0.7621	0.6301	0.7832	0.9547
75%	Ensemble-A	59.35M	49.97G	0.7559	0.6255	0.7930	**0.9626**
	Ensemble-B	84.13M	66.28G	0.7699	**0.6404**	**0.8004**	0.9549
	Ensemble-C	84.13M	66.28G	**0.7700**	0.6403	0.7930	0.9569

U-Net-variants (ENS-UNet, Res-UNet, and UNET3+) increase the computational burden while improving the accuracy. Compared with HookNet-16, HookNet-64 contains more parameters and improves the learning ability of the model. Although DeepLab v3+ is a deep network with more than 100 layers, some parameters of ResNet-101 pre-trained on ImageNet can be used in the encoder. Thus, it can quickly fit and achieve good segmentation results in cervical segmentation task. Since the up-sampling operation based on bilinear interpolation is weight-free, the FLOPs of DeepLab v3+ are similar to those of other shallow networks. Therefore, DeepLab v3+ is superior in both speed and accuracy.

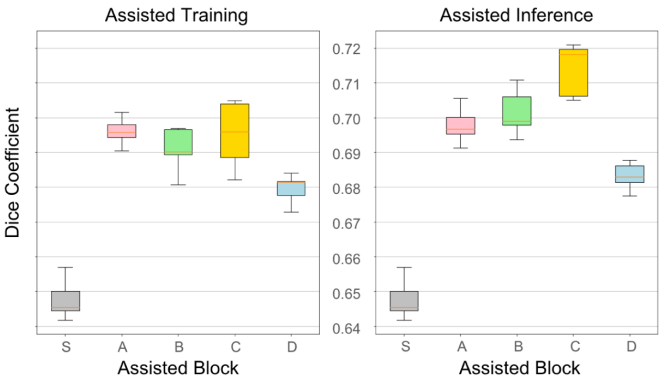

Fig. 6. The box plots of the segmentation results, where each set of experiments was repeated five times. The S in the abscissa represents segmentation networks without assisted blocks. The left results are the segmentation outputs of the models trained with assisted block A, B, C, or D, but without the inference multiplication. The right results are the segmentation results with assisted inference.

3) Fully Supervised Ensemble: The classification loss of the assisted block and the segmentation loss jointly restrain the parameter iteration of the ensemble equally. Since the skeleton network parameters are shared, the assisted block will help the fitting of the shared parameters during the training process. Due to the large connected area of cervical lesions, the patch cut out from the central area of a RoI belongs to a single category. Thus, in the inference process, the outputs of the assisted block can be assigned as weights to the different channels of the segmentation feature-maps to suppress the pepper noise. Fig. 6 shows the box plots of the segmentation results. Each set of experiments in the box plots was repeated five times. The left box plot reflects the segmentation Dice coefficients with assisted blocks in the training process. The right one reflects the Dice coefficients with different assisted blocks in the inference process. S indicates the comparison results without assisted blocks. A, B, C, and D indicate

TABLE III

WEAKLY SUPERVISED CLASSIFICATION RESULTS WITH 75% OVERLAPPING. ACC REPRESENTS THE ACCURACY COMPARED WITH PATCH-LEVEL GROUND TRUTH. THE RESULTS USING RESNET-101 AS *M* FOR EXTRACTING PSEUDO-LABELED PATCHES IS BETTER THAN THOSE USING MOBILENET-V2. WHEN RESNET-101 IS ADOPTED AS *M*, INCEPTION-V3 SHOWS THE BEST EFFECT WHEN ALL PSEUDO DATA ARE INTRODUCED, RESNET-101 SHOWS THE BEST EFFECT WHEN 50% ONES ARE INTRODUCED, AND VGG-19 SHOWS THE BEST EFFECT WHEN 25% ONES ARE INTRODUCED

Training Set	Network	M: MobileNet-v2					M: ResNet-101				
		Dice	mIoU	AP	WP	Acc	Dice	mIoU	AP	WP	Acc
S_p	VGG-19	0.4195	0.3054	0.4150	0.8707	0.3790	0.3939	0.2810	0.3847	0.8222	0.3644
	Inception-v3	0.6437	0.5126	0.6531	0.9417	0.5221	0.4931	0.3868	0.4921	0.8854	0.3935
	ResNet-101	0.4115	0.3086	0.4106	0.8159	0.3319	0.5101	0.4192	0.5370	0.9302	0.4339
	MobileNet-v2	0.2941	0.1997	0.2823	0.6900	0.3208	0.3809	0.2817	0.3707	0.7790	0.3366
$S_p + S_0$	VGG-19	0.6904	0.5722	0.7394	0.9760	0.5610	0.6821	0.5589	0.7396	0.9424	0.5673
	Inception-v3	0.7200	0.5896	**0.7769**	0.9542	0.5670	**0.7785**	**0.6503**	**0.8022**	0.9646	0.5850
	ResNet-101	0.7349	0.5976	0.7596	0.9493	0.5875	0.7118	0.5822	0.7633	0.9492	0.5898
	MobileNet-v2	0.6790	0.5594	0.7360	0.9418	0.5414	0.6264	0.4996	0.6732	0.9075	0.5174
$50\%S_p + S_0$	VGG-19	0.6309	0.4985	0.6856	**0.9819**	0.5496	0.6755	0.5474	0.7230	**0.9791**	0.5727
	Inception-v3	0.7442	0.6065	0.7717	0.9734	0.5673	0.7588	0.6196	0.7751	0.9701	0.5841
	ResNet-101	0.7288	0.5863	0.7479	0.9535	**0.5936**	0.7730	0.6406	0.7908	0.9625	**0.6169**
	MobileNet-v2	0.6414	0.5087	0.6926	0.9440	0.5446	0.6749	0.5401	0.7194	0.9423	0.5683
$25\%S_p + S_0$	VGG-19	**0.7477**	**0.6165**	0.7670	**0.9819**	0.5847	0.7567	0.6290	0.7803	0.9746	0.5831
	Inception-v3	0.7237	0.5794	0.7403	0.9647	0.5610	0.7570	0.6175	0.7709	0.9600	0.5819
	ResNet-101	0.7364	0.5939	0.7513	0.9512	0.5917	0.7586	0.6241	0.7795	0.9587	0.6037
	MobileNet-v2	0.6355	0.4955	0.6599	0.9436	0.5363	0.6642	0.5308	0.6992	0.9402	0.5544

the connection positions of the assisted blocks in Fig. 4. According to the Dice coefficients, the assisted block can significantly improve the segmentation performance, and the assisted inference can further enhance the overall effect. Due to the sharp decrease in the channels of the feature-map at point D, the model cannot fit the classification labels well, which leads to a weaker improvement effect than the assisted blocks at the other three points. The results of repeated experiments show that although the assisted blocks of B and C sometimes bring great effects, the performance is unstable due to a large number of parameters. Table II compares the results of the ensemble of classification and segmentation, where Ensemble-A, B, C represent the assisted block at point A, B, C, respectively. The Dice coefficient of Ensemble-A is 0.706, while that of DeepLab v3+ is 0.65, indicating that the ensemble algorithms show great advantages over the single segmentation networks. The accuracy of Ensemble-A, B, and C are relatively similar, but Ensemble-A is faster and more stable.

B. Weakly Supervised Learning

Since the pseudo-labels in weakly supervised learning are not absolutely accurate, the pseudo dataset S_p should not be too large. In our experiments, the pseudo-label number N_p of Algorithm 1 is set to 4000. Considering that the feature extractor M is trained on pixel-annotated data, MobileNet-v2 which is the fastest and ResNet-101 which is the most accurate in non-overlapping fully supervised classification, are selected.

1) Weakly Supervised Classification: According to Section V-A.1, MobileNet-v2 is superior in speed, while VGG-19, Inception-v3, and ResNet-101 are superior in accuracy. Thus, the four networks are adopted for the experiments of weakly supervised classification. Table III shows the results with different amount of pseudo training data. The results of ResNet-101 as the feature extractor M are significantly better than those of MobileNet-v2, indicating that the accuracy of pseudo-labels is crucial for weakly supervised learning. The test results are very poor when only the pseudo dataset S_P is used without the pixel-annotated training set S_0, suggesting that the difference between the test set and the training set is too large, and many of the extracted pseudo-labels are wrong. Thus, it is necessary to control the number of pseudo-labels. In addition, different networks have different ability to resist noise. MobileNet-v2 is greatly affected by the noisy data, and the performance becomes worse after the introduction of pseudo-labels. VGG-19 is capable to fit data because of many parameters, and is also easy to fluctuate because of noise. Thus, when the pseudo data are relatively pure, the Dice coefficient reaches relatively high accuracy. The Dice coefficients of ResNet-101 and Inception-v3 reach approximate 0.77 when 50% and 100% pseudo-label data are introduced. Therefore, the key of improving the effect of weakly supervised learning is increasing the purity of pseudo-label data.

2) Weakly Supervised Segmentation: DeepLab v3+ is adopted in the experiments of weakly supervised segmentation because it yields good performance in fully supervised segmentation (Section V-A.2). Since the pseudo-label only represents the existence of corresponding category in the patch, so it is coarse and improper to fill the segmentation truth with the pseudo-label. As shown in Table IV, the Dice coefficient reaches 0.674 (75% overlapping) when 50% S_p is used for training, which is only slightly improved compared with fully supervised segmentation 0.65 (non-overlapping). Therefore, pseudo-labeling is unsuitable for the weakly supervised segmentation because of too much noise.

TABLE IV
RESULTS OF WEAKLY SUPERVISED SEGMENTATION WITH DIFFERENT PSEUDO EXTRACTOR *M*. ALL EXPERIMENTS ARE CONDUCTED BASED ON DEEPLAB V3+ WITH 75% OVERLAPPING

M	Training Set	Dice	mIoU	AP	WP
MobileNet-v2	S_p	0.1913	0.1410	0.3510	0.8874
	$S_p + S_0$	0.6516	0.5094	0.6937	**0.9322**
	$50\%S_p + S_0$	0.6665	0.5263	0.6993	0.9193
	$25\%S_p + S_0$	0.6379	0.5055	0.6941	0.9052
ResNet-101	S_p	0.1929	0.1425	0.3549	0.8876
	$S_p + S_0$	0.6344	0.4962	0.6867	0.9276
	$50\%S_p + S_0$	**0.6740**	**0.5315**	**0.7033**	0.9243
	$25\%S_p + S_0$	0.6410	0.5057	0.6910	0.9109

TABLE V
RESULTS OF WEAKLY SUPERVISED ENSEMBLE WITH ASSISTED BLOCK AT POINT A. THE RESULTS OBVIOUSLY SURPASS THOSE OF FULLY SUPERVISED ENSEMBLE-A AND EVEN ENSEMBLE-C

Loss	Non-overlapping				75% Overlapping			
	Dice	mIoU	AP	WP	Dice	mIoU	AP	WP
CE-loss	0.7487	0.6087	0.7674	0.9567	0.7783	0.6472	0.7991	0.9637
AE-loss	**0.7529**	**0.6146**	**0.7707**	**0.9571**	**0.7833**	**0.6531**	**0.8017**	**0.9640**

3) Weakly Supervised Ensemble: As mentioned above, pseudo-labeling is valuable in classification instead of segmentation. Thus, pseudo dataset S_p is only used for training the classification branch in the ensemble, while the segmentation branch is still trained with pixel-level annotated S_0. Considering that the training of classification and segmentation is separate, Ensemble-B and C are difficult to converge due to more parameters and the lack of segmentation loss; thus, only Ensemble-A is adopted for weakly supervised ensemble learning. The Dice coefficient of Ensemble-A reaches 0.7833 in Table V, which is significantly higher than 0.7559 in Table II. Therefore, image-level annotated data are valuable for cervical precancerous CAD.

C. Discussion

The classification network transferred from the natural image processing tasks can quickly fit the cervical histopathology data. Classification networks are prone to misjudgement due to insufficient foreground information, which is offset by overlapping post-processing through aggregating the diagnosis information of adjacent patches. The segmentation networks accurately locate the boundaries of the lesion regions, but the segmentation performance is decreased because of the holes inside the prediction heatmaps. By assigning an assisted block to the segmentation network, the advantages of both classification and segmentation are combined and good performance is achieved.

In fact, pixel-level annotation is time-consuming, while image-level annotation is relatively abundant. Despite the image-level annotation, the label of the cropped patch is unknown due to the large size of the cervical histopathology image. Therefore, weakly supervised learning is a latent. Pseudo-labeling is adopted in this article for weakly supervised learning. Experiments show that the accuracy of the classification model for extracting pseudo-label data is critical to the purity of the training set which strongly affects the final performance. The anti-noise performance of networks is different, so the scale of pseudo-label dataset should be balanced according to the learning ability of networks. Considering the limitations of pseudo-label learning for segmentation, the performance of the ensemble algorithm can be improved by effectively using pseudo-label dataset on the classification branch.

The public dataset provided in this article aims to attract researchers to pay attention to the automated diagnosis assistance of the cervix. In addition to the benchmarks described in this article, different directions can be explored to meet the requirements of pathologists:

- The original WSI size is far beyond the GPU memory limit, and the cropped patch has a limited field of view. Therefore, it is expected to explore methods based on multi-scale image fusion to gather rich structural information to simulate the pathologists' process of determining the lesion location and grade at different resolutions.
- Due to the complexity of cervical tissue structure, pathologists are subjective in the diagnosis of lesions. In addition, absolute accuracy cannot be achieved when labeling lesion regions with polygonal lines on low- resolution images. Therefore, weakly supervised learning that only uses annotations as references is encouraged to resist labeling errors introduced by various factors.
- It is much easier to obtain unlabeled data. Hence, unsupervised learning algorithms are strongly recommended.

VI. CONCLUSION

In this article, a new cervical histopathology image dataset called MTCHI is introduced, and a precancerous task is designed to evaluate the performance of automated diagnosis. Four evaluation metrics (Dice coefficient, mIoU, AP, and WP) are provided particularly for this task. Both fully and weakly supervised algorithms are discussed. Extensive experiments based on classification and segmentation networks are carried out to demonstrate the feasibility of CAD on cervical precancerous lesions. The high accuracy of the ensemble of fully and weakly supervised strategies demonstrates the potential of unlabeled data in improving the performance. The dataset is publicly available for researchers to reproduce and explore novel algorithms, and finally is helpful for diagnosis.

REFERENCES

[1] R. L. Siegel, K. D. Miller, and A. Jemal, "Cancer statistics, 2020," *CA, A Cancer J. Clinicians*, vol. 70, no. 4, pp. 7–30, 2020.
[2] R. M. Richart and B. A. Barron, "A follow-up study of patients with cervical dysplasia," *Amer. J. Obstetrics Gynecol.*, vol. 105, no. 3, pp. 386–393, Oct. 1969.
[3] B. E. Bejnordi *et al.*, "Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer," *J. Amer. Med. Assoc.*, vol. 318, no. 22, pp. 2199–2210, Dec. 2017.
[4] G. Aresta *et al.*, "Bach: Grand challenge on breast cancer histology images," *Med. Image Anal.*, vol. 56, pp. 122–139, Aug. 2019.
[5] J. Li *et al.*, "Signet ring cell detection with a semi-supervised learning framework," in *Proc. IPMI*, 2019, pp. 842–854.
[6] Y. J. Kim *et al.*, "PAIP 2019: Liver cancer segmentation challenge," *Med. Image Anal.*, vol. 67, Jan. 2020, Art. no. 101854.

[7] Y. Fu *et al.*, "Pan-cancer computational histopathology reveals mutations, tumor composition and prognosis," *Nature Cancer*, vol. 1, no. 8, pp. 800–810, Aug. 2020.

[8] T. M. Darragh *et al.*, "The lower anogenital squamous terminology standardization project for HPV-associated lesions: Background and consensus recommendations from the college of American pathologists and the American society for colposcopy and cervical pathology," *Arch. Pathol. Lab. Med.*, vol. 136, no. 10, pp. 1266–1297, 2012.

[9] J. Jantzen, J. Norup, G. Dounias, and B. Bjerregaard, "Pap-smear benchmark data for pattern classification," in *Proc. Nature Inspired Smart Inf. Syst. (NiSIS)*, 2005, pp. 1–9.

[10] Z. Lu *et al.*, "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.

[11] 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.

[12] B. S. Veeling, J. Linmans, J. Winkens, T. Cohen, and M. Welling, "Rotation equivariant CNNs for digital pathology," in *Proc. MICCAI*. Cham, Switzerland: Springer, 2018, pp. 210–218.

[13] F. A. Spanhol, L. S. Oliveira, C. Petitjean, and L. Heutte, "A dataset for breast cancer histopathological image classification," *IEEE Trans. Biomed. Eng.*, vol. 63, no. 7, pp. 1455–1462, Jul. 2016.

[14] M. Peikari, S. Salama, S. Nofech-Mozes, and A. L. Martel, "Automatic cellularity assessment from post-treated breast surgical specimens," *Cytometry A*, vol. 91, no. 11, pp. 1078–1087, Nov. 2017.

[15] P. Guo *et al.*, "Nuclei-based features for uterine cervical cancer histology image analysis with fusion-based classification," *IEEE J. Biomed. Health Informat.*, vol. 20, no. 6, pp. 1595–1607, Nov. 2016.

[16] S. De *et al.*, "A fusion-based approach for uterine cervical cancer histology image classification," *Comput. Med. Imag. Graph.*, vol. 37, nos. 7–8, pp. 475–487, Oct. 2013.

[17] H. A. AlMubarak *et al.*, "A hybrid deep learning and handcrafted feature approach for cervical cancer digital histology image classification," *Int. J. Healthcare Inf. Syst. Informat.*, vol. 14, no. 2, pp. 66–87, Apr. 2019.

[18] D. Wang, C. Gu, K. Wu, and X. Guan, "Adversarial neural networks for basal membrane segmentation of microinvasive cervix carcinoma in histopathology images," in *Proc. Int. Conf. Mach. Learn. Cybern. (ICMLC)*, Jul. 2017, pp. 385–389.

[19] I. Goodfellow *et al.*, "Generative adversarial networks," *Commun. ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[20] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, inception-ResNet and the impact of residual connections on learning," 2016, *arXiv:1602.07261*. [Online]. Available: https://arxiv.org/abs/1602.07261

[21] X. Zhang, X. Zhou, M. Lin, and J. Sun, "ShuffleNet: An extremely efficient convolutional neural network for mobile devices," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6848–6856.

[22] A. Echle *et al.*, "Clinical-grade detection of microsatellite instability in colorectal tumors by deep learning," *Gastroenterology*, vol. 159, no. 4, pp. 1406–1416, 2020.

[23] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[24] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: https://arxiv.org/abs/1409.1556

[25] H. Le *et al.*, "Utilizing automated breast cancer detection to identify spatial distributions of tumor-infiltrating lymphocytes in invasive breast cancer," *Amer. J. Pathol.*, vol. 190, no. 7, pp. 1491–1504, Jul. 2020.

[26] Y. Wang *et al.*, "Pathological image classification based on hard example guided CNN," *IEEE Access*, vol. 8, pp. 114249–114258, 2020.

[27] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9.

[28] M. van Rijthoven, M. Balkenhol, K. Siliņa, J. van der Laak, and F. Ciompi, "HookNet: Multi-resolution convolutional neural networks for semantic segmentation in histopathology whole-slide images," *Med. Image Anal.*, vol. 68, Feb. 2021, Art. no. 101890.

[29] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. MICCAI*. Cham, Switzerland: Springer, 2015, pp. 234–241.

[30] Y. Liu *et al.*, "Detecting cancer metastases on gigapixel pathology images," 2017, *arXiv:1703.02442*. [Online]. Available: https://arxiv.org/abs/1703.02442

[31] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2818–2826.

[32] H. Lin, H. Chen, Q. Dou, L. Wang, J. Qin, and P.-A. Heng, "ScanNet: A fast and dense scanning framework for metastastic breast cancer detection from whole-slide image," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2018, pp. 539–546.

[33] S. Takahama *et al.*, "Multi-stage pathological image classification using semantic segmentation," in *Proc. ICCV*, Oct. 2019, pp. 10702–10711.

[34] Z. Guo *et al.*, "A fast and refined cancer regions segmentation framework in whole-slide breast pathological images," *Sci. Rep.*, vol. 9, no. 1, pp. 1–10, Dec. 2019.

[35] M. Khened, A. Kori, H. Rajkumar, B. Srinivasan, and G. Krishnamurthi, "A generalized deep learning framework for whole-slide image segmentation and analysis," 2020, *arXiv:2001.00258*. [Online]. Available: https://arxiv.org/abs/2001.00258

[36] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. ECCV*, Sep. 2018, pp. 801–818.

[37] D.-H. Lee, "Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks," in *Proc. ICML Workshop*, Jun. 2013, vol. 3, no. 2, pp. 1–6.

[38] H. A. Tokunaga, "Negative pseudo labeling using class proportion for semantic segmentation in pathology," in *Proc. ECCV*, 2020, pp. 1–17.

[39] Y. Li *et al.*, "Self-loop uncertainty: A novel pseudo-label for semi-supervised medical image segmentation," in *Proc. MICCAI*. Cham, Switzerland: Springer, 2020, pp. 614–623.

[40] H.-T. Cheng *et al.*, "Self-similarity student for partial label histopathology image segmentation," in *Proc. ECCV*. Cham, Switzerland: Springer, 2020, pp. 117–132.

[41] S. Shaw *et al.*, "Teacher-student chain for efficient semi-supervised histology image classification," in *Proc. ICLR Workshop*, 2020, pp. 1–6.

[42] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4510–4520.

[43] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4700–4708.

[44] O. Russakovsky *et al.*, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.

[45] Z. Meng *et al.*, "Adaptive elastic loss based on progressive inter-class association for cervical histology image segmentation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 976–980.

[46] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.

[47] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017.

[48] Z. Meng, Z. Fan, Z. Zhao, and F. Su, "ENS-UNet: End-to-end noise suppression U-Net for brain tumor segmentation," in *Proc. EMBC*, Jul. 2018, pp. 5886–5889.

[49] K. Cao and X. Zhang, "An improved Res-UNet model for tree species classification using airborne high-resolution images," *Remote Sens.*, vol. 12, no. 7, p. 1128, Apr. 2020.

[50] H. Huang *et al.*, "UNet 3+4: A full-scale connected unet for medical image segmentation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*. May 2020, pp. 1055–1059.