Deep Learning Methods for Lung Cancer Segmentation in Whole-Slide Histopathology Images—The ACDC@LungHP Challenge 2019

Zhang Li, Jiehua Zhang, Tao Tan, Xichao Teng, Xiaoliang Sun, Hong Zhao, Lihong Liu, Yang Xiao, Byungjae Lee, Yilong Li, Qianni Zhang, Shujiao Sun, Yushan Zheng, Junyu Yan, Ni Li, Yiyu Hong, Junsu Ko, Hyun Jung, Yanling Liu, Yu-cheng Chen, Ching-wei Wang, Vladimir Yurovskiy, Pavel Maevskikh, Vahid Khanagha, Yi Jiang, Li Yu, Zhihong Liu, Daiqiang Li, Peter J. Schüffler, Qifeng Yu, Hui Chen, Yuling Tang, and Geert Litjens

***Abstract*—Accurate segmentation of lung cancer in pathology slides is a critical step in improving patient care. We proposed the ACDC@LungHP (Automatic Cancer Detection and Classification in Whole-slide Lung Histopathology) challenge for evaluating different computer-aided diagnosis (CADs) methods on the automatic diagnosis of lung cancer. The ACDC@LungHP 2019 focused on segmentation (pixel-wise detection) of cancer tissue in whole slide imaging (WSI), using an annotated dataset of 150 training images and 50 test images from 200 patients. This paper reviews this challenge and summarizes the top 10 submitted methods for lung cancer segmentation. All methods were evaluated using metrics using the precision, accuracy, sensitivity, specificity, and DICE coefficient (DC). The DC ranged from 0.7354±0.1149 to 0.8372±0.0858. The DC of the best method was close to the inter-observer agreement (0.8398±0.0890). All methods were based on deep learning and categorized into two groups: multi-model method and single model method. In general, multi-model methods were significantly better ($p<0.01$) than single model methods, with mean DC of 0.7966 and 0.7544, respectively. Deep learning based methods could potentially help pathologists find suspicious regions for further analysis of lung cancer in WSI.**

***Index Terms*—Artificial intelligence, convolutional neural networks, deep learning, lung cancer.**

I. Introduction

LUNG cancer is the top cause of cancer-related death in the world. According to the 2009-2013 SEER (Surveillance, Epidemiology, and End Results) database, the 5-year survival rate of lung cancer patients is approximately 18% [1].

Manuscript received November 30, 2019; revised June 16, 2020 and October 12, 2020; accepted November 11, 2020. Date of publication November 20, 2020; date of current version February 4, 2021. This work was supported in part by the National Natural Science Funding of China (No.61801491), in part by the Natural Science Funding of Hunan Province (No. 2019JJ50728), and in part by the NIH/NCI Cancer Center Support under Grant (P30 CA008748). *(Zhang Li, Jiehua Zhang and Tao Tan equally contributed to this work.)* *(Corresponding authors: Zhang Li; Hui Chen; Yuling Tang.)*

Zhang Li, Jiehua Zhang, Xichao Teng, Xiaoliang Sun, Hong Zhao, and Qifeng Yu are with the College of Aerospace Science and Engineering, National University of Defense Technology, Changsha 410073, China, and also with the Hunan Provincial Key Laboratory of Image Measurement and Vision Navigation (e-mail: zhangli_nudt@163.com; zhangjiehua_nudt@outlook.com; tengari@buaa.edu.cn; alexander_sxl@nudt.edu.cn; shamrock-zhao@hotmail.com; yuqifeng@vip.sina.com).

Tao Tan is with the Department of Mathematics and Computer Science, Eindhoven University of Technology, Eindhoven, MB 5600, The Netherlands, and also with ScreenPoint Medical, Nijmegen, EC 6525, The Netherlands (e-mail: t.tanl@tue.nl).

Lihong Liu and Yang Xiao are with Pingan Technology, Shenzhen 518000, China.

Byungjae Lee is with Lunit, Inc., Seoul, Korea.

Yilong Li and Qianni Zhang are with the School of Electrical Engineering and Computer Science, Queen Mary University of London, London, U.K.

Shujiao Sun is with the Image Processing Center, School of Astronautics, Beihang University, Beijing 102206, China, and also with the Beijing Advanced Innovation Center for Biomedical Engineering, Beijing 100191, China.

Digital Object Identifier 10.1109/JBHI.2020.3039741

Yushan Zheng is with the Beijing Advanced Innovation Center for Biomedical Engineering, Beihang University and Image Processing Center, Beijing 102206, China, and also with the School of Astronautics, Beihang University, Beijing 102206, China.

Junyu Yan and Ni Li are with AstLab, School of Automation, Beihang University, Beijing, China.

Yiyu Hong and Junsu Ko are with the R&D Center, Arontier Company Ltd., Seoul, Korea.

Hyun Jung and Yanling Liu are with Frederick National Laboratory, Frederick, MD, USA.

Yu-cheng Chen and Ching-wei Wang are with the Center of Computer Vision and Medical Imaging, the Graduate Institute of Biomedical Engineering, and the Graduate Institute of Applied Science and Technology, National Taiwan University of Science and Technology and AI Explore, Taipei, Taiwan.

Vladimir Yurovskiy and Pavel Maevskikh are with the Research Department Skychain Global, Yekaterinburg, Russia.

Vahid Khanagha is with the Audio Solutions Team, Motorola Solutions, Inc. Plantation, FL, USA.

Yi Jiang and Daiqiang Li are with the Second Xiangya Hospital, Central South University, Changsha, China (e-mail: jiangyi76; lidqxf@163.com).

Li Yu is with the Lensee Biotechnology Company Ltd., Ningbo, China (e-mail: 22379006@qq.com).

Zhihong Liu is with the Hunan Cancer Hospital, Central South University, Changsha, China (e-mail: liuzhihong214@163.com).

Peter J. Schüffler is with the Memorial Sloan Kettering Cancer Center, USA (e-mail: schueffp@mskcc.org).

Hui Chen and Yuling Tang are with the First Hospital of Changsha City, Changsha, China (e-mail: tyl71523@sina.com).

Geert Litjens is with the Radboud University Medical Center in Nijmegen, The Netherlands (e-mail: geert.litjens@radboudumc.nl).

For patients with the early stage, resectable cancer, the 5-year survival rate is about 34%, but for unresectable cancer, the 5-year survival rate is less than 10%. Therefore, early detection and diagnosis of lung cancer are the key important steps in improving patient treatment outcomes. According to the National Comprehensive Cancer Network (NCCN) guidelines, for image-suspected tumors, histopathological assessment of biopsies obtained via fiberoptic bronchoscopy should be performed for the diagnosis [2], [3].

Assessment of biopsy tissue by a pathologist is the golden standard for lung cancer diagnosis. However, the diagnostic accuracy is less than 80% [4]. The major histological subtypes of malignant lung disease are squamous carcinoma, adenocarcinoma, small cell carcinoma, and undifferentiated carcinoma. Correctly assessing these subtypes on biopsy is paramount for correct treatment decisions. However, the number of qualified pathologists is too small to meet the substantial clinical demands, especially in countries such as China, with a significant population of lung cancer patients. Recently, the results from the largest randomized control lung screening trial, the National Lung Screening Trial (NLST), led to the implementation of lung cancer screening with low-dose Computed Tomography in the United States in 2015. Moreover, the results from the second-largest randomized control trial, the Dutch-Belgian lung cancer screening trial (NELSON), also show the benefits of implementing lung cancer screening. The implementation in the U.S. and the possible implementation of lung cancer screening in Europe will likely lead to a substantial amount of whole-slide histopathology images biopsies and resected tumors. At the same time, the workload and the shortage of pathologists are severe. An artificial intelligence (AI) system might efficiently solve the problems mentioned above by an automatic assessment of lung biopsies.

Digital pathology has been gradually introduced in pathological clinical practice. Digital pathology scanners could generate high-resolution WSIs (up to 160 nm per pixel). It facilitates the development of automatic analysis algorithms for reducing the burden and improving the performance of pathologists. Most recently, a large number of deep learning (DL) methods have been proposed for automatic image analysis of WSIs from the cell level to the image level [5]–[20].

At the cell level, DL methods were used in mitosis detection [21]–[23], nucleus detection [24]–[26] and cell classification [27], [28]. These proposed methods were all based on convolutional neural networks (CNNs). At the tissue level, CNNs were proposed for segmentation (e.g., segmenting glands for grading adenocarcinomas [29]). Moreover, contour information [30], handcrafted features [31], [32], multi-loss [33]–[35] were incorporated into CNNs to obtain more reliable tissue segmentation results.

At the image level, a three-layer CNN was first introduced to detect invasive ductal breast carcinoma and showed a comparable result (65.40% accuracy) with classifiers relying on specific handcrafted features [36]. CNNs were also used in the detection of prostate cancer [37], pancreas cancer [38], renal diseases [12], stomach disease [16], kidney cancer [39], and colon cancer [40].

Deeper CNN, such as GoogLeNet [41], AlexNet [42], VGG [43] and ResNet [44], was transferred to breast cancer classification [45] and prostate cancer prediction [46]. In the CAMELYON16 challenge [47], the 1st rank team ensembled two GoogLeNets to elevate the AUC of classification of lymph node metastases to 99.4%. Several challenges in medical imaging also significantly advanced the pathology image analysis community, such as mitosis detection challenges in ICPR 2012,[1] CAMELYON16[2] and CAMELYON17[3] for identifying breast cancer metastases. In particular, the CAMELYON16 was the first challenge to offer WSIs a large number of annotations, which is essential for training deeper CNNs.

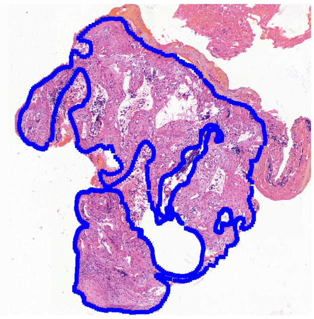


Fig. 1. Pathological WSI with annotations for cancer regions.

With the breakthrough of DL methods in medical image analysis and increasing of available public WSIs for developing a specific CNN, we believe that the CNN could be leveraged to give pathologists more reliable objective results or even help pathologists to improve the cancer diagnostic level. However, after assessing recent review papers [10], [11], we found very few articles discussing the applications of CNNs to histopathological images of lung cancer. Furthermore, no public datasets of WSI were available to evaluate such algorithms. A recent paper that used CNNs on lung cancer detection was only on cytological image [48]. The size of each image was limited (only around 1k*1 k pixels), and the appearance of this image was quite different from the hematoxylin&eosin (H&E) stained image that we used in this paper. The recent research [49] suggested that image features automatically extracted from WSIs can predict the prognosis of lung cancer patients and thereby contribute to precision oncology by machine learning classifiers.

To further explore the potential application of DL on WSI for lung cancer diagnosis, we proposed the ACDC@LungHP challenge which is the first challenge at addressing lung cancer detection and classification using WSI, to our best knowledge [45]. This manuscript is a summary of the first stage of ACDC@LungHP (in conjunction with ISBI2019) that focused on the segmentation of cancer tissue in WSI. The sample of pathological WSI with annotations is shown in Fig. 1.

II. Materials

A. Patient Recruitment

For the ACDC@LUNGHP challenge, 200 lung patients were recruited in this study at the Department of Pulmonary Oncology in the First Hospital of Changsha, from January 2016 to

[1] http://people.idsia.ch/~juergen/deeplearningwinsMICCAIgrandchallenge.html

[2] https://camelyon16.grand-challenge.org/

[3] https://camelyon17.grand-challenge.org/

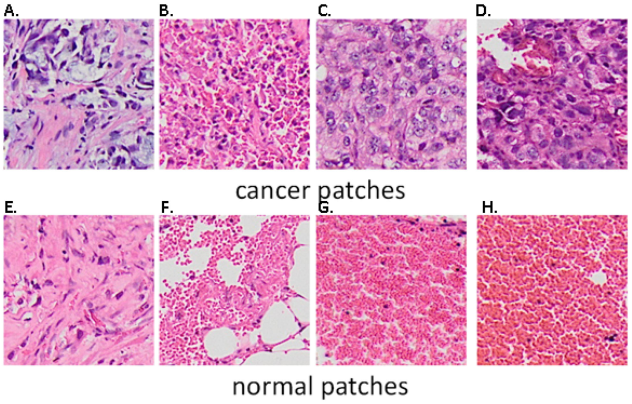


Fig. 2. Example of tumor patches and normal patches.

November 2017. According to the American Joint Committee on Cancer (AJCC) staging system, patients firstly diagnosed with lung/bronchus cancer (site: C34.1-C34.9; histology type: adenocarcinoma, squamous cell carcinoma, and small cell carcinoma) were recruited. Other inclusion criteria included: 1) pathologically confirmed patients with surgery biopsy maintained; 2) no radiotherapy before surgery; 3) aged between 30 and 90 yr. The exclusion criteria were: 1) multiple primary cancers; 2) metastatic lung cancer; 3) patients with immune-deficiency or organ-transplantation history; 4) patients who did not provide informed consent. This study was approved by the Ethics Committee of the First Hospital of Changsha. Informed consent was obtained from each patient before the examination. Necessary demographic and clinical information for each patient, such as age, gender, stage, pathology, etc. were collected.

B. Data Preparation

Histological slides were stained with H&E scanned by a digital slide scanner (3DHISTECH Pannoramic 250) with objective magnifications of 20x. The close look of different tissues in the slides can be seen in Fig. 2. One can see that the patch colors were quite different even among the patches from normal tissue due to the staining variability. The appearance of the cancer regions was also quite different because of the different cancer types. For instance, Fig. 2.(A) and (B) represent small-cell lung cancer, and Fig. 2.(C) and (D) represent squamous cell lung cancer and adenoid cell lung cancer. Fig. 2.(E)-(H) are normal patches.

In total, 200 H&E stained slides were scanned and digitized. We randomly split those 200 slides into training and test sets. 150 slides with annotation were released as the training set. 50 slides were held as the test set. The main types of cancer were included in our data: squamous cell carcinoma, small cell carcinoma, and adenocarcinoma. The ratio of them was approximately 6:3:1. One pathologist with 30 years of experience (the director of the pathology department) annotated the cancer regions for all 200 slides (See Fig. 1). We also asked the second pathologist (with 20 years experience) to annotate the test set only. The annotation of the second pathologist was only used for accessing the inter-observer variability. Participants were allowed to use their own training data for pre-training. All data were uploaded to Microsoft OneDrive, Google Drive, and Baidu Pan for participants from different regions. Whole-Slide images were released in the TIFF format. Manual annotations were in XML format.

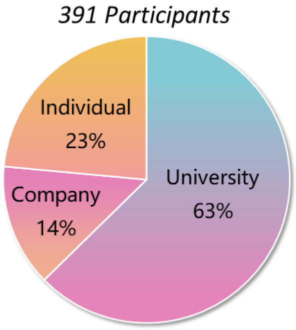


Fig. 3. The distribution of participants from ACDC@LungHP 2019.

In the clinical practice, more than one sample from the same biopsy were scanned. If samples had a similar shape, the pathologist only annotated one sample in the WSI. Participants were suggested to use ASAP[4] to make a bounding box themselves to exclude the unused samples.

III. ACDC@LUNGHP Challenge Summary

A. Challenge Overview

The first stage of the ACDC@LUNGHP challenge focused on detecting and segmenting lung carcinoma in WSI. The segmentation as a potential aid could quickly help pathologists to identify suspicious regions. At this stage, 495 participants submitted the challenge applications, and 391 of them were confirmed as valid participants (with required registration information). Each team was allowed to submit their result three times per day. 25 participants successfully submitted their results before the closing time. The distribution of the participants is shown in Fig. 3.

The Dice coefficient (DC) was computed to evaluate the agreement between the automatic segmentation and the manual annotation by the pathologist. The DC was defined as:

$$Dice = \frac{2|GT \cap RES|}{|GT| + |RES|} \tag{1}$$

where the GT and RES are ground truth from the pathologist and result of automatic segmentation, respectively. The top 10 teams were selected from the final participants. The overall comparison could be seen in Table I. The DC ranges from 0.7354 to 0.8372. Based on the model ensembling strategy (See the following sections), the methods from top 10 teams could be categorized into two groups: multi-model method and single model method. Other criteria, such as label refine, pre-processing and pre-training strategy are also summarized in Table I.

[4] https://github.com/GeertLitjens/ASAP

TABLE I
OVERALL COMPARISON OF TOP 10 TEAMS FOR THE ACDC@LUNGHP CHALLENGE 2019

	Team	**Task: Lung cancer detection**		**Deep Learning Algorithm**			
		Mean DC	**Rank**	**Label refine**	**Architecture**	**Preprocessing**	**Comments**
Multi Model	**PATECH**	0.8372	1	√	DenseNets& dilation block with U-Net	Color normalization; Ostu to refine label	Capture more context information and multi-scale feature; Ensemble of models by changing loss function
	Byungjae Lee	0.8297	2	√	ResNet50& DeepLab V3+	Multi data augmentations; Ostu to refine label	Initialize encoder with ImageNet pre-trained weights
	Turbolag	0.7968	3		U-Net& ConvCRF	Multi-resolution training data	Multiple networks; enhance the boundary accuracy
	ArontierHYY	0.7638	6	√	Mdrn80+DenseNet& ResNet	Tile labeling strategy	Ensemble of 16 models
	Newhyun00	0.7552	7	√	DenseNet103	Select clean labels	”Co-teaching” method made training deep neural networks robustly
Single Model	**CMIAS**	0.7700	4	√	DenseNet121& FCN	Locate the tissue regions by a bounding box	Combination of two networks
	Jorey	0.7659	5	√	IncRes+ACF& CRF	Ostu to refine label; Divided into 3 classes (tumor; normal;mix) and mix Mix file into other classes	Feature fusing by using multi-atrous convolution
	AIExplore	0.7510	8		FCN	None	Training in the AI Explore platform; Using a large momentum in SGD; Pre-trained network
	Skyuser	0.7456	9		ResNet18	Multi data augmentations;	Classifier-based approach; Fast, small and robust network
	Vahid	0.7354	10		Small-FCN-512	None	Designed a custom FCN; Pre-trained network

B. Methods Based on Single Model

The single model methods only used the individual model as their architectures. The mean DC for single model methods was 0.7544. An overall comparisons of single model methods could be seen in Table I).

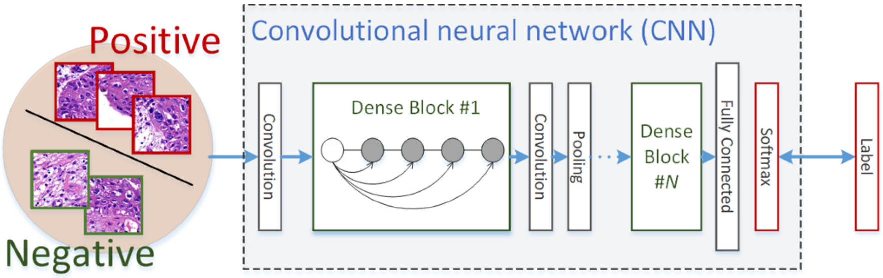
The rank #4 team combined advantages of a CNN and a fully convolutional network (FCN) [50] to improve the accuracy of segmentation. At first, a bounding box was manually annotated to locate the tissue regions. CNN was based on DenseNet-121 structure [51] with two output neurons, and the FCN was based on the DenseNet structure consisting of three dense blocks. The first dense block was with five convolutional layers, and the other two were with eight convolutional layers. The architecture of their model is shown as Fig. 4. Intel Core i7-7700 k CPU and a GPU of Nvidia GTX 1080Ti were used for training. They used cross-entropy with softmax output as the loss function, and Adam as the optimizer for CNN structure. The dice loss and focal loss were set as loss function, and the SGD with momentum was set as the optimizer for FCN structure.

The rank #5 team integrated the Atrous fusing module and CNN feature extractor to build their networks (See Fig. 5). They combined ResNet and Inception V2 (IncRes), which replaced eight middle blocks of ResNet18 with Inception's module. The WSI was split into big patches (with size of $768 * 768 * 3$) in the data pre-processing step, and nine small patches were extracted uniformly from each big patch. After feature extraction using IncRes, the multi-atrous convolution was used for feature fusion [52]. The big patches were assigned to TUMOR, NORMAL,and MIX according to the annotation. They mixed MIX patches into TUMOR and NORMAL to keep the balance of the training data. In their experiments, four parallel atrous convolution modules were used to fuse all features with different dilation ratios. The Convolutional Conditional Random Field (CRF) [53] after the concatenate layer was connected. The CRF did not involve in the training stage, but used to modify the output results. They used four NVIDIA GTX 1080Ti 12 GB GPU and set the learning rate to 1e-3 for beginning 40 epochs, 7e-4 for the last 20 epochs. The loss function was set as BCEWithLogitsLoss. They illustrated that the model combining IncRes, atrous convolution module, and CRF gave the best segmentation performance.

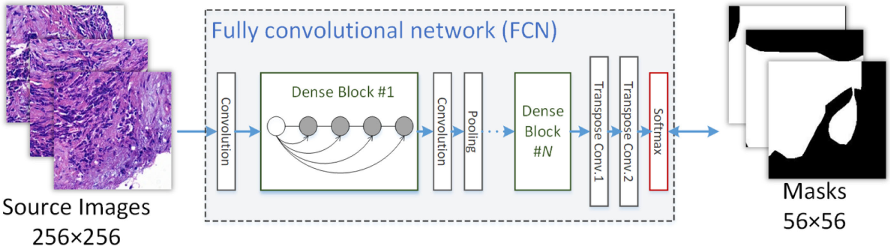
The rank #8 team used a fast deep learning-based model. They put all training sets into the FCN [50] in the AI Explore platform [54]. After training, the test set was tested by the AI Explore platform for real-time lung whole slide segmentation. They used NVIDIA GeForce GTX 1080 Ti to train the model and the SGD with large momentum to avoid the multiple local gradient minimums. The learning rate set to 1e-10.

The rank #9 team used a classification method by labeling large regions instead of distinct pixels. They trained a ResNet18 model with multiple data augmentation methods. An adaptation of threshold was used for cell detection. The training was on a single NVIDIA GeForce GTX 1060, Adam was used as the optimizer with a learning rate set to 1e-4.

The rank #10 team processed WSIs in large patches with no overlap to capture more context. They evaluated three alternative networks: Small-FCN-16, Small-FCN-32 and Small-FCN-512.



(a) patch classification based on CNN.



(b) pixel-wise segmentation based on FCN.

Fig. 4. The network architecture proposed by rank #4 team.

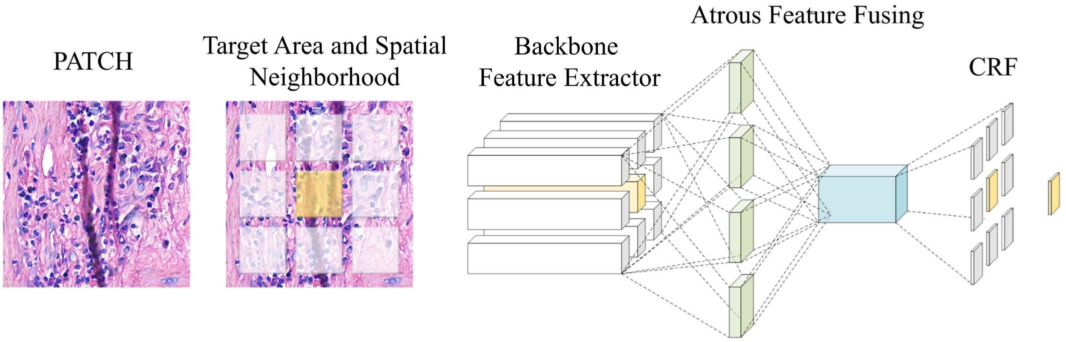

PATCH
Target Area and Spatial Neighborhood
Backbone Feature Extractor
Atrous Feature Fusing
CRF

Fig. 5. Model architecture used by rank #5 team.

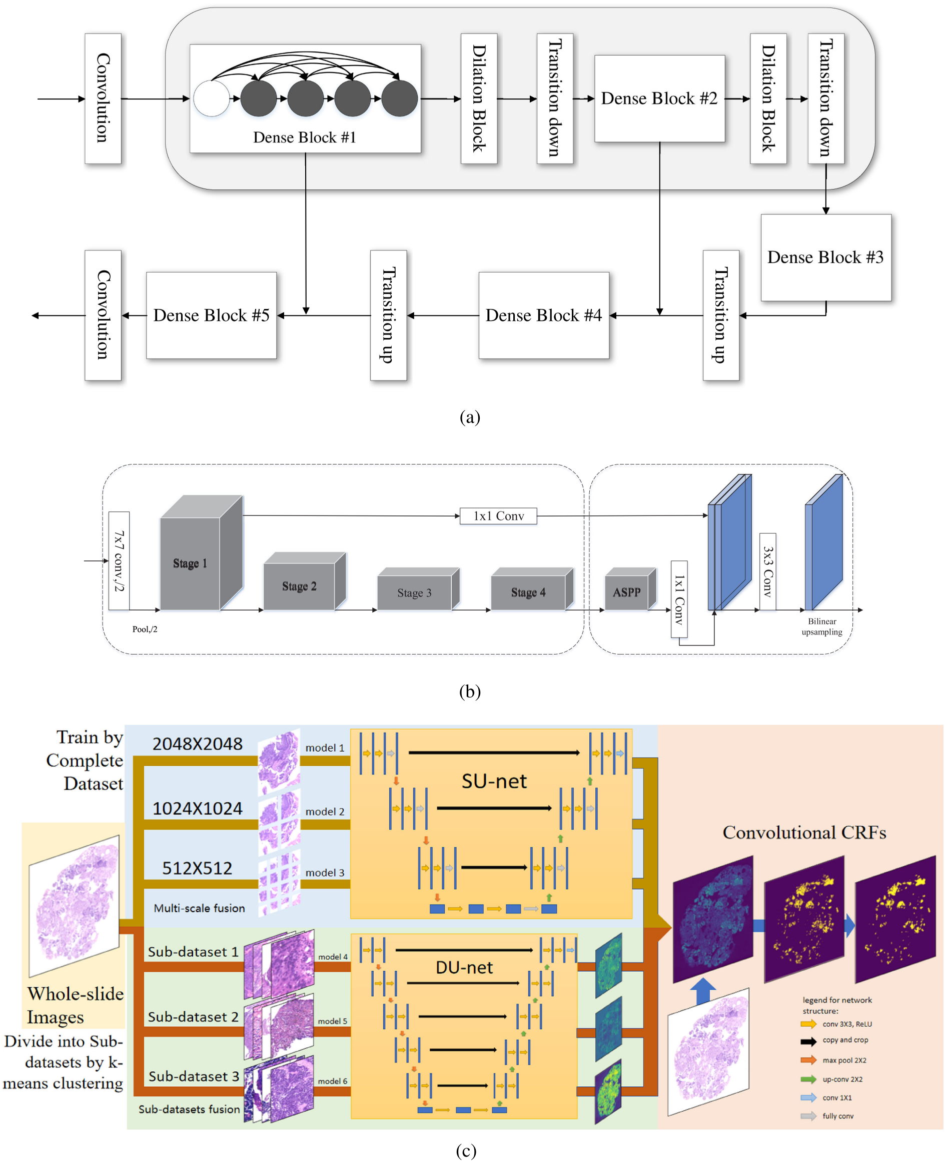
For locating the cancer region rather than exact boundaries, they used 4×4 convolutional filters to increase the receptive field at a different level. They also used Imagenet-FCN to train their model. The training was on the NVIDIA Pascal GPU. Adam optimizer, with a decaying learning rate that started with 1e-4, was used to optimize the weights of these networks. The cross-entropy was set as a loss function. They compared different small networks and selected small-FCN-32 with Imagenet-FCN as their final model.

C. Methods Based on Multi-Model

In general, the single model is not flexible enough to solve complex problems [55], such as the segmentation of lung cancer regions. Furthermore, training multiple models could significantly improve the generalized performance than only using single model [55].

The rank #1 team combined the DenseNets and dilation block to work with U-Net. DenseNet [51] connected each layer to every other layer in a feed-forward fashion (See Fig. 6(a)). The U-Net has an encoder-decoder structure with skip connections that enables efficient information flow [56]. In the dilation block, with the same convolution kernel size, different dilation rates could be utilized to obtain multi-scale features and more context information. The dilation rate (1, 3, 5) with 3x3 kernel were concatenated as the input of the convolution. The dense block was constructed by four layers. They trained different models by changing the loss function through weights and choose the best-performing model to ensemble. This model was more sensitive to tiny lesions and able to capture more context information and multi-scale feature. They used four GPU on Tesla M60 and Adam optimization with default parameters $\beta_1 = 0.9$, $\beta_2 = 0.999$ for training, set the initial learning rate to 2∗10-4, and then divided learning rate by 20 in every 20 epochs. The loss function was a combination of dice function and cross-entropy.

The rank #2 team refined labels by removing the background within the tumor area and performed data augmentation at the training step. The ResNet50 was used as an encoder network



(a)

(b)

(c)

Fig. 6. Network architectures proposed by top three teams.(From top to bottom: rank #1, rank #2 and rank #3).

to extract semantic information. The DeepLab V3+ was used for upsampling. They also modified the ResNet architecture to adapt to the task as described below: 1) Down-sampling step in stage4 was eliminated by changing first convolution layer stride 2 to 1; 2) All convolution layers in stage4 had been altered to use atrous rate 1 to 2; 3) Global average pooling layer was removed and attached DeepLab V3+ decoder;4) All convolution layers in DeepLab V3+ decoder used separable convolution. The model is shown as Fig. 6(b). In the experiments, they used ImageNet pre-trained weights for encoder and Adam as the optimizer, set the initial learning rate to 1e-4. The loss function was a combination of the cross-entropy loss and soft dice loss. They trained CNN models with five-fold cross-validation and ensembled five models from cross-validation training.

The rank #3 team proposed a multi-scale U-net fusion model with the CRF [53] (See Fig. 6(c)). The framework fused networks in two ways: multi-scale fusion and sub-datasets fusion. In multi-scale fusion, three models were trained on the whole training set of three resolutions (576, 1152, and 2048 pixels). The network structure was a modified U-net called SU-net (shallow U-net), which focused on the local details of tumor cells. They removed one downsampling and one upsampling steps in the original U-net and added a fully connected layer before every remaining downsampling and upsampling steps. The SU-net included three times of downsampling and upsampling, consisting of 24 layers in total. In sub-datasets fusion: the dataset was divided into three sub-datasets by k-means algorithm [57]. Each sub-dataset was in the same image resolution of 512 pixels and trained on a DU-net (deep U-net model). The DU-net added one additional downsampling and upsampling stages, consisting of 28 layers. In the experiments, they used the soft-max combining with the cross entropy loss as the loss function.

The rank #6 team used existing classical models, including ResNet101, ResNet152, DenseNet201, DenseNet264 and

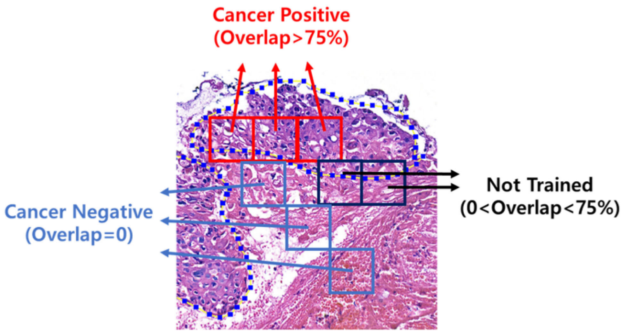


Fig. 7. Tile labeling strategy of rank #6 team.

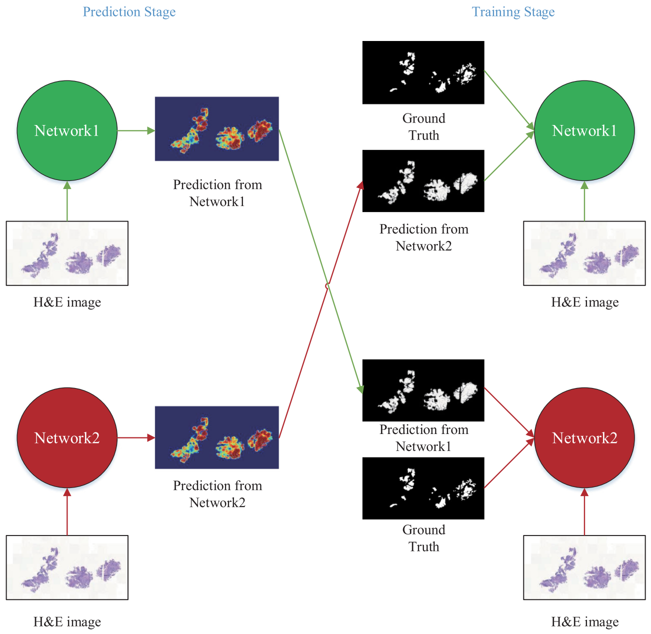


Fig. 8. Schematic diagram of pixel-level Co-teaching algorithm.

Mdrn80 (a short version of the network that DeepMind [58]). They used the tile labeling strategy to label the cancer regions (See Fig. 7). Tile overlapped more than 75% with annotated cancer region was defined as a positive cancer tile, and the tile without overlapping the cancer region was a negative tile. Other tiles were not used for training. They trained and ensemble 16 models to conduct the experiments. They used three NVIDIA RTX Titan GPUs and Adam optimization with a learning rate of 1e-4 for training. Cross-entropy was set as the loss function.

The rank #7 team used “Co-teaching” to train networks (See Fig. 8). Co-teaching [59] aims to clean the noisy label. The proposed method trained two networks simultaneously. In each mini-batch of data, each network viewed its small-loss instances as useful knowledge and taught such instances to its peer network for updating the parameters. Comparing with the original Co-teaching algorithm, the main difference was the dynamic drop rate $\Re(T)$, which controlled the number of clean-instances selected for training. It was used to avoid the training error from a network to be directly transferred back on itself (See Algorithm.1). They used the fully convolutional DenseNet (FC-DenseNet) 103 network as a backbone. The two FC-DenseNet 103 networks were trained from scratch simultaneously using the same data. In their experiment, the networks were trained using four NVIDIA GeForce GTX 1080 Ti GPUs, and an Adam optimizer was used with an initial learning rate of 1.5e-4.

Algorithm 1: Pixel-level Co-teaching Algorithm.

```
Input: $w_f$ and $w_g$, learning rate $\eta$, epoch $T_{max}$, iteration $N_{max}$;
for T=1,2,...,$T_{max}$ do
    Shuffle the training set D;
    for N=1,2,...,$N_{max}$ do
        Fetch J and L from D;
        Obtain $\overline{L}_f = \sigma[f_{w_f}(J)] > 0.5$;
        Obtain $\overline{L}_g = \sigma[f_{w_g}(J)] > 0.5$;
        Update $w_f = w_f - \eta \nabla l_f(f, \overline{L}_g, L)$
        Update $w_g = w_g - \eta \nabla l_g(f, \overline{L}_f, L)$
    end for
end for
Output: $w_f$ and $w_g$
```

A detailed description of the top 10 methods will be uploaded to our challenge website.[5]

IV. Result and Discussion

A. Comparisons of Top 10 Methods

The box-plot of the DC for test set of the top 10 teams is shown in Fig. 9. The inter-observer variability between the two pathologists was also assessed using the mean DC, which was 0.8398 (See Fig. 9). And mean DC of multi-model methods and single model methods for each test image is shown in Fig. 10. All teams got a relative high DC on the NO.27 test image. The DC ranged from 0.8653 to 0.9435. This sample was well prepared during H&E staining like most of the training datasets, and the cancer tissue was clearly shown in this image. One could see typical results from two different teams in Fig. 11(c) and (d). The tissue is shown in Fig. 11(a) and (b).

In contrast, most of the teams got relative low DC on the NO.41 test image (between 0.2-0.5, see Fig. 9). Only rank #1 team got a high DC (0.9458), among others. A visual comparison could be found in Fig. 12(c) and (d). NO.41 was an example of highly differentiated squamous cell carcinoma. The abnormal cells were similar to normal cells in this case. And the color appearance of this slide was not consistent with other slides due to the off-standard H&E staining process. Models from most of the teams might not be generalized enough to deal with this problem.

In order to better evaluate the performance of top 10 methods on the test set, we listed the mean DC on the images of the squamous cell carcinoma (SCC), small cell carcinoma (SCLC), and adenocarcinoma (ADC) (See Table II). The result illustrated that the accuracy of the segmentation depends on how the cancer cells grow. There were no other components in the squamous cell carcinoma nest, so the segmentation accuracy was higher than the other two types. However, small cell carcinoma spread along

[5] http://acdc-lunghp.grand-challenge.org

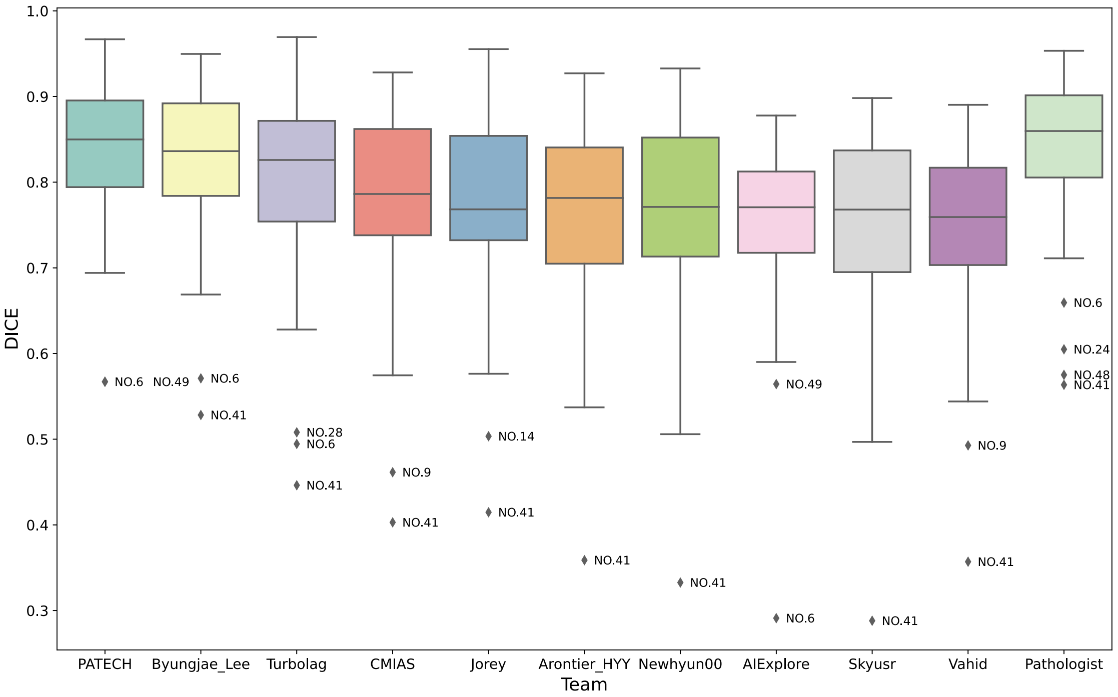
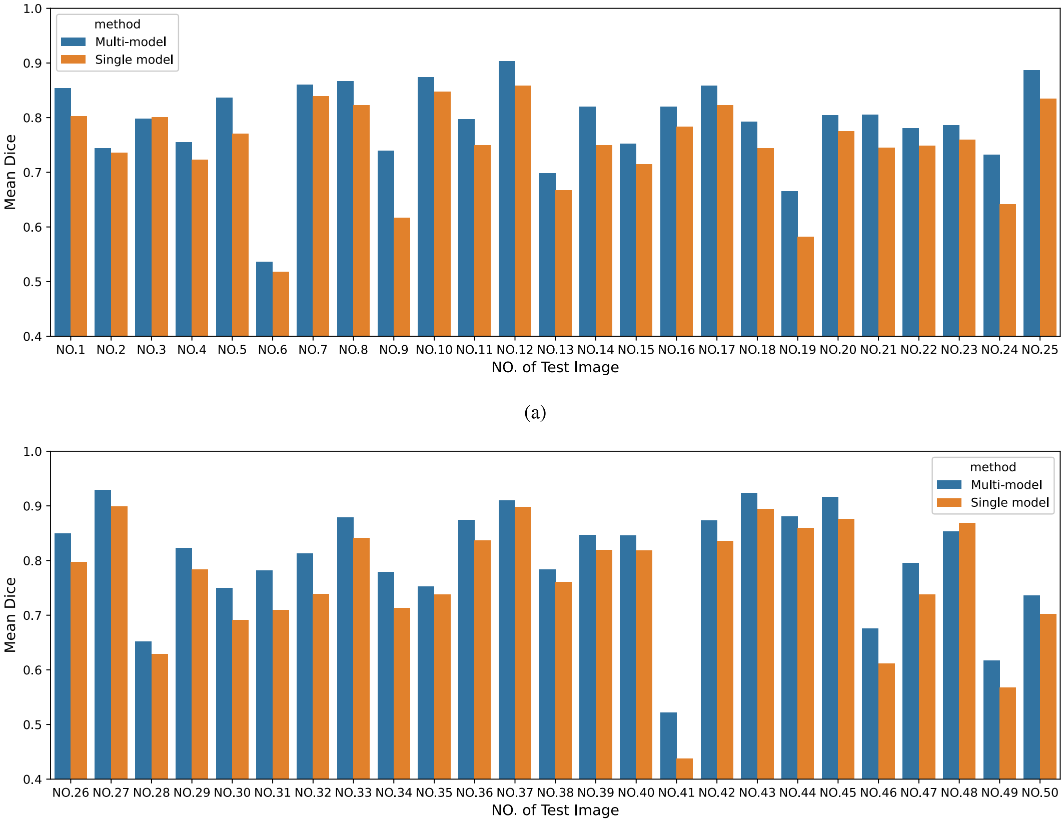


Fig. 9. Comparisons of top 10 teams on the test set.



(a)

(b)

Fig. 10. Mean DC of multi-model methods and single model methods for all 50 test image. (a) test image NO.1-NO.25.(b) test image NO.26-NO.50.

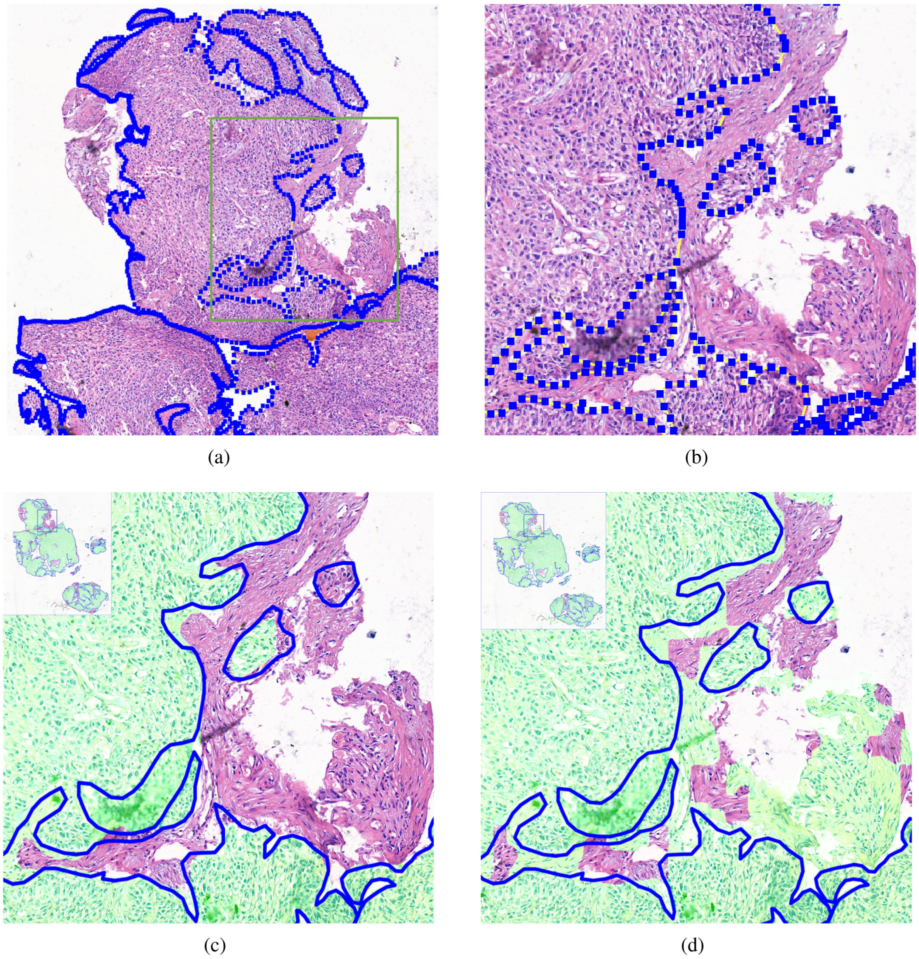


Fig. 11. Pathological WSI for test image NO.27. (a) image with annotation (blue line). (b) selected patch of (a). (c) and (d) results of rank #2 team (DC = 0.9435) and rank #8 team (DC = 0.8653).

TABLE II
COMPARISONS OF MULTI-MODEL AND SINGLE MODEL METHODS ON THREE TYPES OF LUNG CANCER

	SCC	SCLC	ADC
Multi Model	0.8205	0.7521	0.7888
Single Model	0.7797	0.7186	0.7468
All	0.8001	0.7353	0.7678

the sparse fibrous interstitium and gaps, and its cytoplasm was minimal. The adhesion between the cells was inferior, and it was easy to loosen, plant, transfer. Also, the cells were squeezed and deformed during the biopsy, resulting in unclear boundaries. So high performance was hard to be achieved for SCLC. ADC grew along the alveolar wall, and there were too many vascular interstitial components that may affect the segmentation accuracy.

B. Multi-Model v.s. Single Model

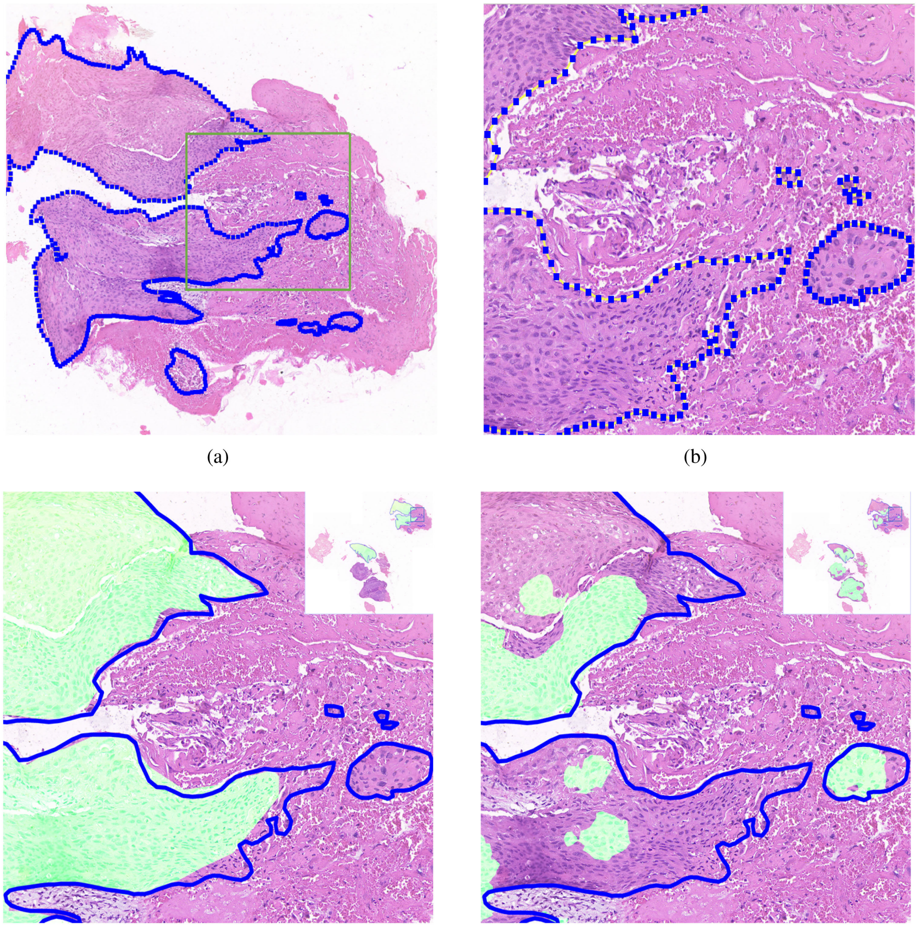
The sign rank test was used to evaluate differences of DC between multi-model and single model methods (based on Fig. 10). The multi-model methods gave significantly better results (p=1.0872e-09) than single model methods. Besides comparing the DC, we also calculated accuracy, precision, sensitivity, and specificity of detection for the top 10 methods (See Table III).

We can see from Table III that sensitivity and specificity of multi-model methods were generally higher than single model methods. We can see that different types of cancer tissue were with different appearances. The current challenge may difficult to provide enough data for all types of cancer. Using a single model might not be sufficient in identifying specific types of cancer. Through model fusion, we could combine multiple models' performance and reduce the probability of missed inspections.

C. Pre-Trained Model v.s. No Pre-Trained Model

Transfer learning is a commonly used method in the AI community. Using the pre-trained model for fine-tuning can reduce training time and achieve better results in several applications. Three teams used ImageNet pre-trained weights to initialize their models. In the challenge, the methods using pre-trained models did not outperform the method that learning from scratch. This might be because the digital pathology domain is inherently different from the ImageNet domain.

The CAMELYON16, TUPAC, and CAMELYON17 challenges aimed at detecting the micro- and macro- metastases in the lymph node in H&E stained WSIs (CAMELYON16/17) and assessing tumor proliferation in breast cancer (TUPAC). Using



(a)

(b)

(c)

(d)

Fig. 12. Pathological WSI for test image NO.41. (a) image with annotation (blue line). (b) selected patch of (a). (c) and (d) results of rank #1 team (DC = 0.9458) and rank #7 team (DC = 0.3327).

TABLE III
QUANTITATIVE COMPARISONS OF MULTI-MODEL AND SINGLE MODEL METHODS ON TEST SET

	Rank	Mean.DC	Accuracy	Precision	Sensitivity	Specificity
Multi Model	1	**0.8372±0.0858**	**0.9505**	**0.7929**	**0.9052**	0.9531
	2	**0.8297±0.0867**	**0.9508**	**0.7996**	**0.8628**	**0.9609**
	3	**0.7968±0.1081**	**0.9462**	**0.7646**	0.8469	**0.9585**
	6	0.7638±0.1107	0.9289	0.7163	0.8558	0.9404
	7	0.7552±0.1237	0.9307	0.7312	0.8199	0.9489
Single Model	4	0.7700±0.1177	0.9375	0.7701	0.8003	**0.9567**
	5	0.7659±0.1130	0.9369	0.7376	0.8151	0.9514
	8	0.7510±0.0973	0.9231	0.6829	**0.8596**	0.9279
	9	0.7465±0.1188	0.9319	0.7428	0.7672	0.9563
	10	0.7354±0.1149	0.9212	0.6830	0.8462	0.9316

The top 3 methods were shown in bold format.

these data for pre-training might get good results. However, none of the teams used a pre-training model from those data.

D. Label Refine

Experienced pathologist annotated cancer regions using ASAP software. We intended to make relative rough labels for the training set (e.g., label contains background region shown as Fig. 13) to evaluate the robustness of the methods in dealing with label noise. All backgrounds and normal tissues were kept in the training set as well. It makes tumor tissue and the normal area extremely unbalanced in the training set. Therefore, label refine is one of the significant issues that should be taken into consideration in this challenge to keep data balanced.

Several teams used different methods to refine the label. Three teams (rank #1, #2, #5) used the Otsu algorithm to remove the background area in tumor tissue labeled and obtain a tighter boundary of the cancer region. The team (rank #4) located the tissue region by a bounding box and filtered the blank areas using

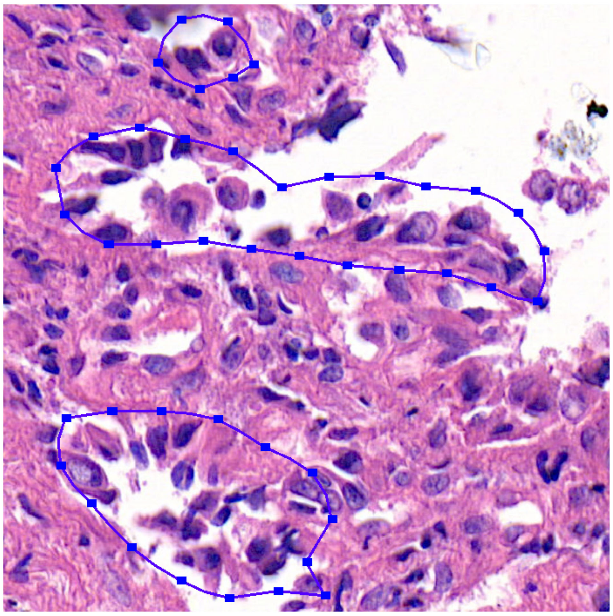


Fig. 13. The sample with label noise (annotation contains background region).

a threshold. The team (rank #6) used a tile labeling strategy in their method and removed background by the percentage of pixel values above 200 in grayscale space. The team (rank #7) used the "Co-teaching" algorithm to refine noisy annotation. The team (rank #10) increased the receptive field at a different level, and they tried to label regions of WSIs rather than finding the exact boundaries.

We found that teams using the Otsu algorithm that removing background gave relatively higher DC. The preprocess for removing the label noise (such as the background in the label area) is essential for model training for the challenge despite the network design.

V. Conclusion

In this paper, the ACDC@LungHP challenge was summarized. The current stage of the challenge focused on lung cancer segmentation. 200 slides were used for this challenge, and methods from the top 10 teams were selected for comparison. In general, multi-model method was relatively better than single model-based methods. The results showed the potentiality of using deep learning for accurate lung cancer diagnosis on WSI.

All submitted methods were based on deep learning, but the networks were quite different. Methods based on multi-model outperformed single model method (mean DC of a single model is 0.7544±0.0991 and multi-model is 0.7966±0.0898). Unlike fine-tuning for other computer vision tasks, the submitted methods did not benefit too much from the ImageNet pre-trained models. The pre-processing for the label noise during the training stage is crucial since our training data was not accurately labeled for test set.

In the coming second stage of this challenge, we will focus on classifying the primary lung cancer subtypes (e.g., squamous carcinoma, adenocarcinoma) using WSI biopsy. At least 1000 slides collected from multiple medical centers will be released in 2020. We believe that the experiences of the first stage will greatly help digital pathology communities to achieve better performance for the second stage.

Acknowledgment

The authors would like to thank Ping Liu, Jun Tang, Tao Xu, Jun XU, Shanshan Wan, Ke Lou, Hui Li, Keyu Li and Yusheng Yan for collecting all the images.

References

[1] A. Jemal *et al.*, "Annual report to the nation on the status of cancer, 1975–2014," *Featuring Survival, J. Nat. Cancer Inst.*, vol. 109, no. 9, 2017, Art. no. djx030.

[2] H. J. Kim *et al.*, "Outcome of incidentally detected airway nodules," *Eur. Respir. J.*, vol. 47, no. 5, pp. 1510–1517, 2016.

[3] M. Andolfi *et al.*, "The role of bronchoscopy in the diagnosis of early lung cancer: A review," *J. Thorac. Dis.*, vol. 8, no. 11, pp. 3329–3337, 2016.

[4] J. S. Thomas *et al.*, "How reliable is the diagnosis of lung cancer using small biopsy specimens? Report of a UKCCCR Lung Cancer Working Party," *Thorax*, vol. 48, no. 11, pp. 1135–9, 1993.

[5] C. L. Srinidhi, O. Ciga, and A. L. Martel, "Deep neural network models for computational histopathology: A survey," *Med. Image Anal.*, vol. 67, 2021, Art. no. 101813.

[6] M. Veta *et al.*, "Predicting breast tumor proliferation from whole-slide images: The TUPAC16 challenge," *Med. Image Anal.*, vol. 54, pp. 111–121, 2019.

[7] Z. Zhang *et al.*, "Pathologist-level interpretable whole-slide cancer diagnosis with deep learning," *Nature Mach. Intell.*, vol. 1, pp. 236–245, 2019.

[8] G. Xu *et al.*, "CAMEL: A weakly supervised learning framework for histopathology image segmentation," in *IEEE/CVF Int. Conf. Comput. Vis.*,, 2019, pp. 10681–10690.

[9] K. Bera *et al.*, A. "Artificial intelligence in digital pathology new tools for diagnosis and precision oncology," *Nature Rev. Clin. Oncol.*, vol. 16, pp. 703–715, 2019.

[10] G. Litjens *et al.*, "A survey on deep learning in medical image analysis," *Med. Image. Anal.*, vol. 42, pp. 60–88, 2017.

[11] D. Shen, G. Wu, and H. I. Suk, "Deep learning in medical image analysis," *Annu. Rev. Biomed. Eng.*," vol. 19, pp. 221–248, 2017.

[12] G. Bueno *et al.*, "Glomerulosclerosis identification in whole slide images using semantic segmentation," *Comput. Methods Programs Biomed.*, vol. 184, 2019, Art. no. 105273.

[13] K. Sirinukunwattana *et al.*, "Gland segmentation in colon histology images: The GlaS challenge contest," *Med. Image Anal.*, vol. 35, pp. 489–502, 2016.

[14] Z. Swiderska-Chadaj *et al.*, "Learning to detect lymphocytes in immunohistochemistry with deep learning," *Med. Image Anal.*, vol. 58, 2019, Art. no. 101547.

[15] W. Bulten *et al.*, "Epithelium segmentation using deep learning in H & E-stained prostate specimens with immunohistochemistry as reference standard," *Sci. Rep.*, vol. 9, no. 1, 2019, Art. no. 864.

[16] H. Sharma *et al.*, "Deep convolutional neural networks for automatic classification of gastric carcinoma using whole slide images in digital histopathology," *Comput. Med. Imag. Graph.*, vol. 61, pp. 2–13, 2017.

[17] M. Shaban *et al.*, "A novel digital score for abundance of tumour infiltrating lymphocytes predicts disease free survival in oral squamous cell carcinoma," *Sci. Rep.*, vol. 9, 2019, Art. no. 13341.

[18] M. Shahedi *et al.*, "Head and neck cancer detection in digitized whole-slide histology using convolutional neural networks," *Sci. Rep.*, vol. 9, 2019, Art. no. 14043.

[19] P. J. Schüffler *et al.*, "Mitochondria-based renal cell carcinoma subtyping: Learning from deep vs. flat feature representations," in *Proc. 1st Mach. Learn. Healthcare Conf., Mach. Learn. Healthcare*, 2016, vol. 56, pp. 191–208, *Proc. Mach. Learn. Res.*.

[20] D. J. Ho *et al.*, “Deep interactive learning: An efficient labeling approach for deep learning-based osteosarcoma treatment response assessment,” *Med. Image Comput. Comput. Assist. Interv.*, 2020, vol. 12265, pp. 540–549.

[21] H. Wang *et al.*, “Mitosis detection in breast cancer pathology images by combining handcrafted and convolutional neural network features,” *J. Med. Imag. (Bellingham)*, vol. 1, no. 3, 2014, Art. no. 034003.

[22] A. Shkolyar *et al.*, “Automatic detection of cell divisions (mitosis) in live-imaging microscopy images using convolutional neural networks,” in *Proc. IEEE 37th Annu. Int. Conf. Eng. Med. Biol. Soc.*, 2015, pp. 743–746.

[23] C. D. Malon and E. Cosatto, “Classification of mitotic figures with convolutional neural networks and seeded blob features,” *J. Pathol. Inform.*, vol. 4, p. 9, 2013.

[24] Y. Xie *et al.*, “Beyond classification: Structured regression for robust cell detection using convolutional neural network,” *Med. Image Comput. Comput. Assist. Interv.*, vol. 9351, pp. 358–365, 2015.

[25] Y. Xie *et al.*, “Deep voting: A robust approach toward nucleus localization in microscopy images,” *Med. Image Comput. Comput. Assist. Interv.*, vol. 9351, pp. 374–382, 2015.

[26] F. Xing, T. C. Cornish, T. Bennett, D. Ghosh, and L. Yang, “Pixel-to-pixel learning with weak supervision for single-stage nucleus recognition in Ki67 images,” *IEEE Trans. Biomed. Eng.*, vol. 66, no. 11, pp. 3088–3097, Nov. 2019.

[27] Z. Gao, L. Wang, L. Zhou, and J. Zhang, “HEp-2 cell image classification with deep convolutional neural networks,” *IEEE J. Biomed. Health Inform.*, vol. 21, no. 2, pp. 416–428, Mar. 2017.

[28] S. Bauer *et al.*, (2016) Multi-Organ cancer classification and survival analysis,” 2016, *arXiv e-prints 1606*.

[29] H. Chen, X. Qi, L. Yu and P. Heng, “DCAN: Deep contour-aware networks for accurate gland segmentation,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2487–2496.

[30] H. Chen *et al.*, “DCAN: Deep contour-aware networks for object instance segmentation from histology images,” *Med. Image. Anal.*, vol. 36, pp. 135–146, 2017.

[31] Y. Xu *et al.*, “Gland instance segmentation using deep multichannel neural networks,” *IEEE Trans. Biomed. Eng.*, vol. 64, no. 12, Dec. 2017.

[32] M. Gadermayr *et al.*, CNN cascades for segmenting whole slide images of the kidney, 2017, *arXiv:1708.00251*.

[33] A. BenTaieb, J. Kawahara, and G. Hamarneh, “Multi-loss convolutional networks for gland analysis in microscopy,” in *IEEE Int. Symp. Biomed. Imag.*, 2016, pp. 642–645.

[34] T. Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, “Focal loss for dense object detection,” in *IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 2999–3007.

[35] F. Milletari, N. Navab, and S.-A. Ahmadi, “V-net: Fully convolutional neural networks for volumetric medical image segmentation,” in *Proc. 4th Int. Conf. 3D Vis.*, 2016, pp. 565–571.

[36] N. Wahab, A. Khan, and Y.S. Lee, “Two-phase deep convolutional neural network for reducing class skewness in histopathological images based breast cancer detection,” *Comput. Biol. Med.*, vol. 85, pp. 86–97, 2017.

[37] W. Bulten *et al.*, “Automated gleason grading of prostate biopsies using deep learning”, 2019, *arXiv:1907.07980*.

[38] F. Xing, T. C. Cornish, T. Bennett, D. Ghosh, and L. Yang, “Pixel-to-pixel learning with weak supervision for single-stage nucleus recognition in Ki-67 images,” *IEEE Trans. Biomed. Eng.*, vol. 66, no. 11, pp. 3088–3097, Nov. 2019.

[39] T. de Bel *et al.*, “Automatic segmentation of histopathological slides of renal tissue using deep learning,” *Med. Imag. Digit. Pathol.*, vol. 10581, 2018, Art. no. 1058112.

[40] M. N. Gurcan *et al.*, “Deep learning for tissue microarray image-based outcome prediction in patients with colorectal cancer,” *SPIE Med. Imag.*, vol. 9791, 2016.

[41] C. Szegedy *et al.*, “Going deeper with convolutions,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recogni. (CVPR)*, Boston, MA, USA, 2015, pp. 1–9, doi: 10.1109/CVPR.2015.7298594.

[42] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” *Commun. ACM*, vol. 60, no. 6, pp. 84–90, 2017.

[43] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” in *Proc. Int. Conf. Learn. Represent.*, 2015.

[44] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition”, in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, 2016, pp. 770–778, doi: 10.1109/CVPR.2016.90.

[45] D. Wang *et al.*, “Deep learning for identifying metastatic breast cancer”, 2016, *arXiv e-prints 1606*.

[46] A. J. Schaumberg, M. A. Rubin, and T. J. Fuchs, “H&E-stained whole slide image deep learning predicts SPOP mutation state in prostate cancer”, 2018, *arXiv-064279*.

[47] B. E. Bejnordi *et al.*, “Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer,” *JAMA*, vol. 318, no. 22, pp. 2199–2210, 2017.

[48] A. Teramoto, T. Tsukamoto, Y. Kiriyama, and H. Fujita, “Automated classification of lung cancer types from cytological images using deep convolutional neural networks,” *Bio. Med. Res. Int.*, vol. 2017, 2017, Art. no. 4067832.

[49] K. H. Yu *et al.*, “Predicting non-small cell lung cancer prognosis by fully automated microscopic pathology image features,” *Nature Commun.*, vol. 7, 2016, Art. no. 12474.

[50] J. Long, E. Shelhamer, T. Darrell, “Fully convolutional networks for semantic segmentation,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.* , 2015, pp. 3431–3440.

[51] G. Huang *et al.*, “Densely connected convolutional networks,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2261–2269.

[52] L. C. Chen *et al.*, “DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol, 40, no. 4, pp. 834–848, Apr. 2018.

[53] M. T. T. Teichmann, R. Cipolla, “Convolutional CRFs for semantic segmentation,” 2018, *arXiv:1805.04777*.

[54] AI explore platform for real time whole slide segmentation, http://aiexploredb.ntust.edu.tw/

[55] H. Li, X. Wang, and S. Ding, “Research and development of neural network ensembles: A survey,” *Artif. Intell. Rev.*, vol. 49, pp. 455–479, 2018.

[56] R. Evans *et al.*, “De novo structure prediction with deeplearning based scoring,” in *13th Crit. Assessment Tech. Protein Struct. Prediction (Abstracts)*, 2018.

[57] T. Kanungo, D. M. Mount, N. S. Netanyahu, C. D. Piatko, R. Silverman, and A. Y. Wu, “An efficient k-means clustering algorithm: Analysis and implementation,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 881–892, Jul. 2002.

[58] O. Ronneberger, P. Fischer, and T. Brox, “U-Net: convolutional networks for biomedical image segmentation,” in *Proc. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F.(eds.) MICCAI LNCS*, 2015, vol. 9351, pp. 234–241.

[59] B. Han *et al.*, “Co-teaching: Robust training of deep neural networks with extremely noisy labels,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, vol. 31, pp. 8527–8537.