Weakly Supervised Deep Learning for Whole Slide Lung Cancer Image Analysis

Xi Wang, Hao Chen, *Member, IEEE*, Caixia Gan, Huangjing Lin, Qi Dou, *Member, IEEE*, Efstratios Tsougenis, Qitao Huang, Muyan Cai, and Pheng-Ann Heng, *Senior Member, IEEE*

***Abstract*—Histopathology image analysis serves as the gold standard for cancer diagnosis. Efficient and precise diagnosis is quite critical for the subsequent therapeutic treatment of patients. So far, computer-aided diagnosis has not been widely applied in pathological field yet as currently well-addressed tasks are only the tip of the iceberg. Whole slide image (WSI) classification is a quite challenging problem. First, the scarcity of annotations heavily impedes the pace of developing effective approaches. Pixelwise delineated annotations on WSIs are time consuming and tedious, which poses difficulties in building a large-scale training dataset. In addition, a variety of heterogeneous patterns of tumor existing in high magnification field are actually the major obstacle. Furthermore, a gigapixel scale WSI cannot be directly analyzed due to the immeasurable computational cost. How to design the weakly supervised learning methods to maximize the use of available WSI-level labels that can be readily obtained in clinical practice is quite appealing. To overcome these challenges, we present a weakly supervised approach in this article for fast and effective classification on the whole slide lung cancer images. Our method first takes advantage of a patch-based fully convolutional network (FCN) to retrieve discriminative blocks and provides representative deep features with high efficiency. Then, different context-aware block selection and feature aggregation strategies are explored to generate globally holistic WSI descriptor which is ultimately fed into a random forest (RF) classifier for the image-level prediction. To the best of our knowledge, this is the first study to exploit the potential of image-level labels along with some coarse annotations for weakly supervised learning. A large-scale lung cancer WSI dataset is constructed in this article for evaluation, which validates the effectiveness and feasibility of the proposed method. Extensive experiments demonstrate the superior performance of our method that surpasses the state-of-the-art approaches by a significant margin with an accuracy of 97.3%. In addition, our method also achieves the best performance on the public lung cancer WSIs dataset from The Cancer Genome Atlas (TCGA). We highlight that a small number of coarse annotations can contribute to further accuracy improvement. We believe that weakly supervised learning methods have great potential to assist pathologists in histology image diagnosis in the near future.**

***Index Terms*—Deep learning, histology image analysis, weakly supervised learning, whole slide images (WSIs).**

I. Introduction

LUNG cancer is the leading cause of cancer death in both men and women in the U.S. [1]. Appropriate treatment for lung cancer patients primarily depends on the type of lung carcinoma, such as small cell lung cancer (15%) or nonsmall cell lung cancer (85%) [2]. The most common nonsmall cell lung cancer can be divided into several subtypes that are named based upon the tumor cells, such as adenocarcinomas (ADCs) and squamous cell carcinomas (SCs), as shown in Fig. 1. A range of diagnostic tests can be used to diagnose lung cancer, including chest X-ray, computerized tomography (CT), magnetic resonance imaging (MRI), and needle biopsy. Among these approaches, histopathological image analysis serves as the gold standard for lung cancer diagnosis.

Classification of carcinoma types and assessment of aggressiveness are essential for the following targeted treatment. In the clinical practice, carcinoma is routinely identified by experienced pathologists through checking of tissue slide stained with hematoxylin and eosin (H&E) under a high-power microscopy, which is a labor-intensive and time-consuming task [e.g., it takes an experienced histopathologist about 15 min to half an hour to check one whole slide image (WSI)], as it usually requires pathologists to look through large swathes of normal tissue regions to eventually recognize the malignant areas. In addition, lots of mimics share a similar appearance to cancer regions, which should be distinguished carefully. Therefore, automated analysis technique is highly demanded in the pathological field, which could considerably

Manuscript received May 6, 2019; accepted August 1, 2019. This work was supported in part by the Hong Kong Innovation and Technology Commission through ITSP Tier 3 under Project ITS/041/16, and through ITSP Tier 2 Platform under Project ITS/426/17FP, in part by the National Natural Science Foundation of China under Project U1813204, and in part by the Shenzhen Science and Technology Program under Grant JCYJ20180507182410327. This article was recommended by Associate Editor D. Goldgof. *(Corresponding authors: Hao Chen; Muyan Cai.)*

X. Wang and H. Lin are with the Department of Computer Science and Engineering, The Chinese University of Hong Kong, Hong Kong, China (e-mail: xiwang@cse.cuhk.edu.hk).

H. Chen and E. Tsougenis are with Department of AI Image Research, Imsight Medical Technology Company Ltd., Shenzhen 518057, China (e-mail: hchen@cse.cuhk.edu.hk).

C. Gan is with the Hexian Memorial Hospital, Southern Medical University, Guangzhou 511400, China.

Q. Dou is with the Department of Computing, Imperial College London, London SW7 2AZ, U.K.

Q. Huang and M. Cai are with the State Key Laboratory of Oncology in South China, Collaborative Innovation Center for Cancer Medicine, Sun Yat-sen University Cancer Center, Guangzhou 510060, China (e-mail: caimy@sysucc.org.cn).

P.-A. Heng is with the Department of Computer Science and Engineering, The Chinese University of Hong Kong, Hong Kong, China and also with the Guangdong Provincial Key Laboratory of Computer Vision and Virtual Reality Technology, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China.

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCYB.2019.2935141

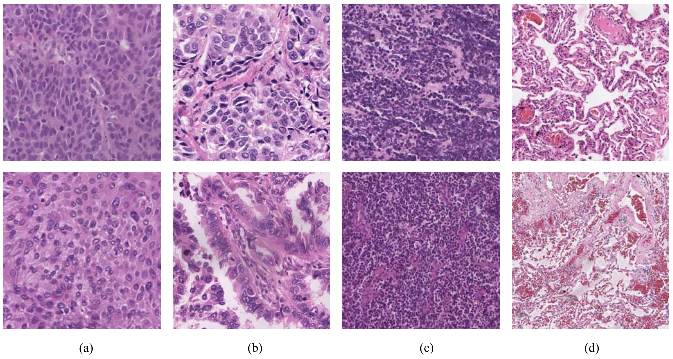


Fig. 1. Examples of different types of lung carcinoma in each column, including (a) SC, (b) ADC, (c) SCLC, and (d) NORM cases.

ease the workload, speed up the diagnosis, and facilitate the in-time treatment.

During the last decade, the convolutional neural networks (CNNs) have achieved astounding achievements in many aspects of the medical imaging field, such as ultrasound [3], [4]; MRI [5], [6]; and CT [7], [8]. Also, lots of histopathology image recognition tasks have been well studied by researchers, for example, mitosis detection in breast cancer histopathology images [9], [10]; gland instance segmentation in colon images [11]; and nuclear atypia scoring for breast cancer assessment [12]. However, these histopathology images employed for research studies fall into the type of region of interests (ROIs) [9]–[11], [13], which are deliberately selected by experienced pathologists with a much smaller size (e.g., 1000×1000) from WSIs. Patch-level labels or even pixelwise annotation masks can be feasibly provided by pathologists for designing the effective algorithms at the training phase. Therefore, most of the approaches can be categorized into fully supervised learning methods inherently.

With the advent of whole slide scanning techniques, dramatically increasing interest has been shown on the WSI analysis which is much more challenging than the ROI-level analysis. For example, a gigapixel WSI contains more than billions of pixels (e.g., $74\,000 \times 76\,000$) on the highest resolution level, which poses great challenges to the image-level classification. Downsampling WSIs into thumbnails is not feasible as lots of intrinsic information and fine details would be lost. Alternatively, it is more reasonable to perform analysis on small patches with fine details cropped from high-resolution WSIs, which is similar to the ROI-level analysis, but in a much larger scale as millions of patches should be taken into consideration for the WSI analysis.

By and large, high-level tasks can be mainly categorized into three branches: 1) tumor detection or segmentation; 2) cancer prognosis; and 3) carcinoma classification. As for the tumor segmentation task, researchers tend to address this problem through two steps, starting with candidate selection and following tumor confirmation. Initially, CNNs served as the patch-level classifier [14]–[20]. It aimed to select suspicion exemplars or candidates from WSIs. In this regard, various deep networks (e.g., GoogLeNet [21], AlexNet [22], and VGG-16 [23]) were evaluated for comparison [17]. In contrast, Lin *et al.* [18] built a novel framework by leveraging fully convolutional network (FCN) for efficient inference while reconstructing dense predictions to ensure the detection accuracy. Rich spatial information plays an important role in tumor detection, and it has been explored in [19] and [24]. For instance, a 2D Long Short-Term Memory was utilized to aggregate the context from a grid of neighboring patches [24], while Bejnordi *et al.* [19] proposed a context-aware stacked CNN to take advantage of the spatial information within WSIs. Furthermore, a range of magnification levels of WSIs were also considered to improve the performance [17], [20]. In general, tumor confirmation is an indispensable part for accurate detection among candidates. The normalized cumulative histogram with percentile analysis [15] and connected component features of WSI probability maps [15], [17], [18] were widely utilized in the second-stage decision fusion model. Handcrafted features, including the local binary patterns, HSD color histogram, topological features (e.g., Voronoi diagram and Delaunay triangulation), were also popular in [14], [19], [24], and [25]. Random forest (RF), support vector machine (SVM), and multilayer perceptron (MLP) [17], [18], [24]–[26] were commonly used for making the final decision. These fully supervised methods largely rely on careful annotations of cancer regions.

Cancer prognosis based on WSIs is a hot research topic which aims to assist clinicians in making early decision on treatments in the context of access to patient's histopathology images. Currently, many approaches have been proposed to provide more accurate survival predictions [27]–[30]. However, these methods depend on handcrafted features extracted from manually selected ROIs whereas in lack of the ability to learn discriminative patterns automatically from WSIs. For instance, Yu *et al.* [28] and Luo *et al.* [29] utilized the software *CellProfiler* to extract a huge number of handcrafted features from annotated tumor regions and then used a regularized machine-learning method for the top feature selection and cancer prognosis classification. In contrast, Zhu *et al.* [31] obtained the survival prediction by utilizing the WSI-level labels only. Clustering was applied on phenotypes of patches by K-means algorithm, followed by the feature selection via CNN and feature aggregation for final prediction. Although Wang *et al.* [32] successfully made lung cancer survival prediction based on shape and boundary features of tumor regions automatically detected by CNN, it indeed requires large number of annotations to train the detector.

Regarding the high-level WSI classification task [33]–[39], due to the lack of detailed annotations of cancer regions, previous studies applied domain-specific handcrafted features to depict morphological character, texture, and statistic property of malignant tumor [33], [34], [40] along with the unsupervised clustering methods (e.g., K-means) and feature embedding for classification [41]. In addition, various weakly supervised methods, such as multiple instance learning (MIL) [42], [43], have been adopted to address this problem by automatically extracting the refined valuable information from coarse-labeled patches [44]–[50]. Handcrafted features (e.g., color histogram, local binary pattern, and SIFT) are also extensively studied in these MIL methods [44]–[46], which actually require considerable efforts to design and validate. Recently, an MIL framework [48] utilized sparse label assignment in supervised training to classify the entire mammogram of breast cancers with size of about 3000 × 2000. However, this idea cannot easily generalize to WSIs since the latter is along with a much larger resolution. In the Camelyon Grand Challenge 2016 (CAMELYON16) [17], [18], hundreds of carefully annotated WSIs are provided to train a fully supervised deep network for automatically detecting the metastatic breast cancer in WSIs. However, the acquisition of careful annotation of WSI in a large scale is fairly prohibitive, if not impossible, in practice as it usually takes several hours to well annotate a WSI for a specialized pathologist. Therefore, it would be appealing to train the cancer region detector with a minimum annotation (e.g., image-level labels) [35], [36], [38] which can be more easily acquired in clinical practice. Xu *et al.* [35] used CNN activations trained by ImageNet to extract off-the-shelf features in a patch-wise fashion, followed by feature embedding to represent the WSI. Since these features are quite generic with limited specific presentation, this method could not succeed in complicated tasks. Lately, an EM-based method proposed by Hou *et al.* [36] was the first to combine the patch-based CNN with supervised decision fusion. Initially, an EM-based method with CNN was used to identify the discriminative patches in WSIs; then, a count-based feature fusion model performed the image-level prediction. Although this method has been verified effective on two WSI datasets, it marginally exceeded other fundamental methods at the cost of huge computation on the iteration of training and inference [36]. Moreover, a study [39] recently achieved fairly inspiring results on nonsmall cell lung cancer WSIs classification by using Inception-V3 [51] network; however, this success was under the condition of good-quality WSIs.

Overall, the existing approaches solving these WSI-level tasks either highly depend on elaborated delineation of tumor regions to provide the annotation masks or require careful design of handcrafted features, which actually pose a heavy burden to histopathologists and researchers, respectively. In order to overcome the aforementioned challenges, in this article, we propose a weakly supervised learning method for fast and effective classification of whole slide lung cancer images with a minimum annotation effort from pathologists. First, our method merely requires image-level labels and a small number of coarse annotations that are readily obtained in practice, which significantly saves a lot of labor cost on annotations. Second, a powerful FCN is adopted to automatically learn the representative features which are much superior to the handcrafted ones. The major contributions of this article are as follows.

1) We develop a novel approach addressing the lung cancer WSI classification problem. To the best of our knowledge, this is the first study that explores the weakly supervised learning on WSI classification with image-level labels as well as a small number of coarse annotations. We take advantage of FCN for efficient prediction and extracting useful deep features and propose several context-aware block selection and feature aggregation strategies.
2) We build the largest fine-grained whole slide lung cancer histopathology image dataset, composed of 939 WSIs. This dataset contains comprehensive types of lung carcinoma (i.e., two subtypes of nonsmall cell lung cancers and small cell lung cancer) and the normal type.
3) The proposed method achieves the state-of-the-art performance on two independent datasets. Extensive experiments on our dataset demonstrate that the context-aware block selection and WSI feature aggregation from multiple instances can provide high-quality holistic feature representation for WSIs. Our method achieves an accuracy of 97.3% on our dataset and an AUC of 85.6% on the public dataset from TCGA. Both results outperform the previous methods by a large margin.

II. Methods

Due to intrinsic properties of whole slide histopathology images (e.g., high resolution and heterogeneity of tumors), it is hardly possible to tackle the WSI classification task by one step, even a large number of careful annotations are available [17], [18]. Fig. 2 shows the architecture of the proposed method. It consists of three parts. The first part is a patch-based CNN that aims to predict the cancer likelihood of WSIs, referred as *discriminative patch prediction*. In the second part of the *context-aware block selection*, the spatially contextual information is taken into consideration when selecting the features from retrieved blocks. Finally, we aggregate features from multiple representative instances in the *context-aware feature selection and aggregation* part; hence, each WSI can be represented by a global feature descriptor that summarizes the most indicative information. The global feature descriptor is eventually fed into a standard RF classifier for WSI-level prediction. This procedure shares the similar idea of "vocabulary-based paradigm" [41] in embedded-space MIL, referred as *WSI feature aggregation and classification*.

A. Discriminative Patch Prediction

1) Preprocessing: In general, a WSI might contain a large proportion (e.g., ranging from 40% to 80%) of white background which is actually irrelevant for cancer analysis. Removal of such noninformative regions could greatly reduce the computational cost while ensuring the validity of training samples. Hence, we apply OTSU algorithm [52], a traditional method for image thresholding to eliminate the majority of irrelevant background while maintaining the tissue regions for training, as shown in Fig. 3. This process can be significantly accelerated by using a multilevel mapping strategy that is proposed in [18].

2) Fast Fully Convolutional Network and Efficient Training Strategy: Efficiency is a key issue concerned in clinical practice. How to quickly process a gigapixel histopathology image is one of the biggest challenges for researchers. Unlike [17], [35], and [36], where CNN scans the entire image in a

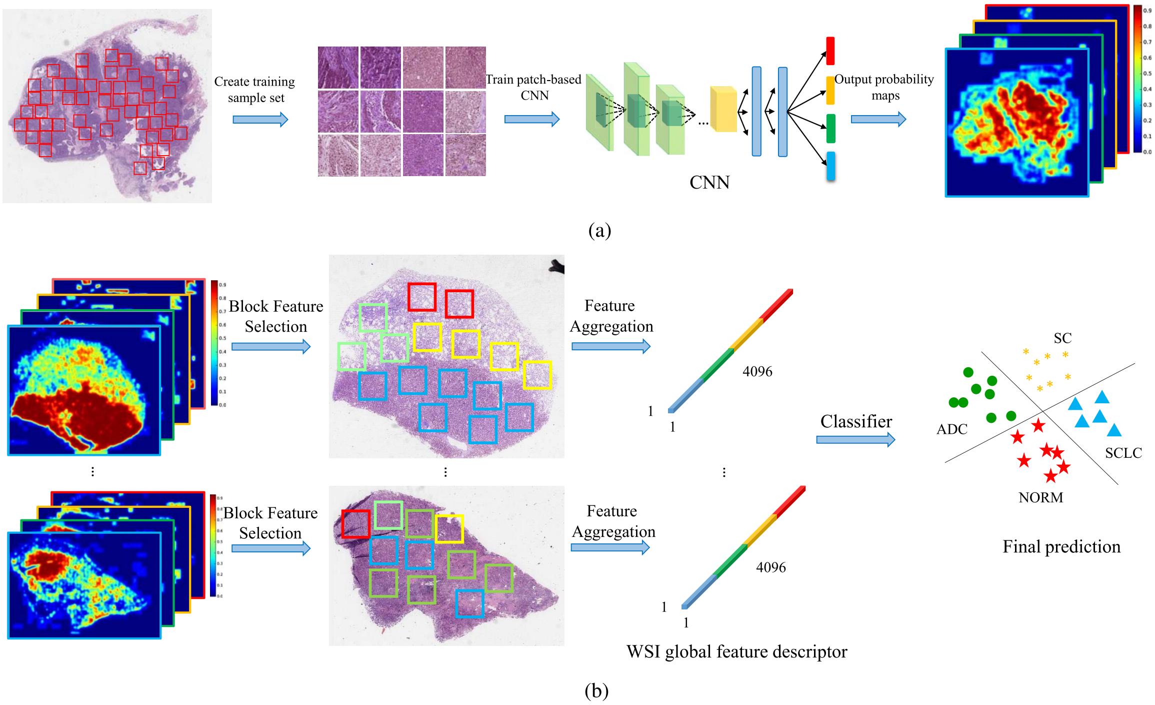


Fig. 2. Overview of the proposed method. (a) Discriminative patch prediction. A patch-based CNN is used to find discriminative regions. (b) Context-aware feature selection and aggregation. By imposing spatial constraint, features from discriminative blocks are selected and aggregated for the WSI classification.

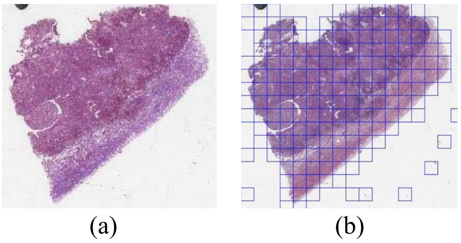


Fig. 3. (a) Raw WSI and (b) segmented tissue regions denoted by blue rectangles.

patchwise fashion, requiring lots of time at the inference phase, we adopt a modified FCN, ScanNet [18], as the patch-prediction model. This network can be flexibly trained with extensively augmented samples in a patchwise way while it can leverage the efficiency of an FCN architecture at prediction phase, as illustrated in Fig. 4. The architecture of ScanNet is based on a modified VGG-16 [23] network by replacing the last three fully connected layers with fully convolutional layers. It can enjoy the transferred features learned from a large set of nature images [53]. In addition, all the padding operations of convolutional layers in the standard VGG-16 are removed to avoid the boundary effect of the FCN predictions. Based on this modification, ScanNet can process blocks in an arbitrary size fast by leveraging the advantage of FCN. Finally, all probability tiles generated from blocks are stitched together to form the probability map of WSIs.

During the training stage, training patches are generated on-the-fly in the data preparation process, which could not only save the memory space but also achieve flexible data augmentation in the meanwhile. However, GPU is frequently idle because a heavy I/O bottleneck occurs due to waiting for the preparation of training samples. To alleviate this problem, we adopt the asynchronous sample prefetching mechanism [18]. The producer/consumer scheme is implemented by multiple processes. Specifically, several producer processes run on CPU to generate training patches while only one consumer process runs on GPU to consume the training samples. This asynchronism breaks the dependency between the producer and the consumer, which would bring a lot of performance improvement in efficiency.

3) Weighted Loss Function for Weakly Supervised Learning: There are at least two challenges for fully supervised learning of WSI analysis. First, it is quite difficult and tedious to obtain accurately pixelwise annotations. Second, there exist ambiguous regions that can not be well distinguished, even for histology experts. However, making use of a large number of available image-level labels and a small number of coarsely annotated WSIs can be feasible in practice. In this article, we are the first to explore weakly supervised learning on WSI classification with image-level labels as well as a small number of coarse annotations of tumor regions. We require the pathologists to annotate the abnormal regions in a scratch way by drawing polygons as shown in Fig. 5. As the annotation is quite coarse (i.e., not all annotated areas are precisely occupied by tumor and vice-versa), it is not safe to take all annotated regions as positive patches, and nonannotated counterparts as negative ones at the training stage. A more reasonable way is to impose a larger weight on these annotated regions as they carry more confidence for manifestation of being carcinoma. In other words, more severe penalty is given when the CNN misclassifies annotated regions, which guides the CNN to

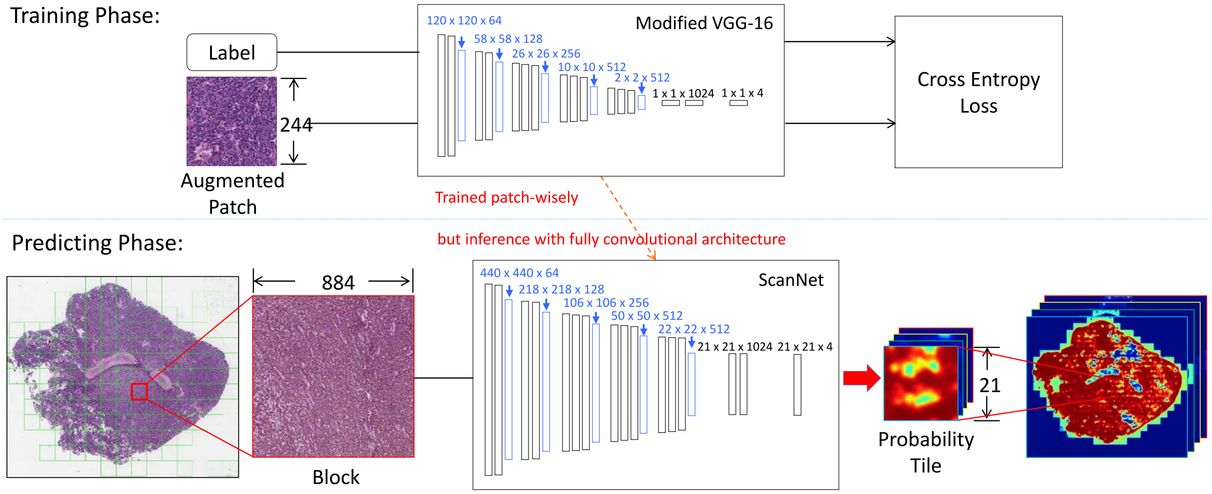


Fig. 4. Illustration of fast patch prediction with ScanNet.

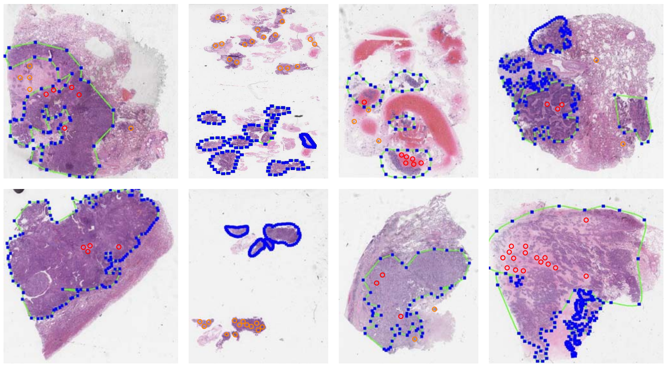


Fig. 5. Illustration of coarse annotations by experienced histopathologists. Green lines with blue dots are the coarse annotations from histopathologists and majority of them are correctly annotated. Note that the orange circles denote cancer regions that are not annotated while the red circles indicate noncancer regions that are enclosed in annotations.

learn more useful and discriminative patterns. Specifically, we train a patch-based CNN by minimizing the following weighted cross-entropy loss function:

$$\mathcal{L} = \sum_{x \in \mathcal{M}} \sum_{\ell=1}^{C} -\alpha y_\ell \log P(q = \ell | x; \mathcal{W}, b) + \sum_{x \notin \mathcal{M}} \sum_{\ell=1}^{C} -y_\ell \log P(q = \ell | x; \mathcal{W}, b) + \lambda \|\mathcal{W}\|_2^2 \quad (1)$$

where $\theta = \{\mathcal{W}, b\}$ denotes the parameters of our CNN model, $P(q = \ell | x)$ is the output probability for the ℓth class given the input subwindow x in which $q \in \{1, 2, \ldots, C\}$, and y_ℓ corresponds to the WSI-level label. C is the total number of classes. $\mathcal{M}$ denotes the coarse annotation mask set. α is the balance weight between annotated region classifier and nonannotated region classifier. In this article, a range of values of $\alpha \in [1, 5]$ have been considered, while α is eventually set as 2 by *grid search*. Moreover, λ controls the tradeoff between the data loss term and regularization term.

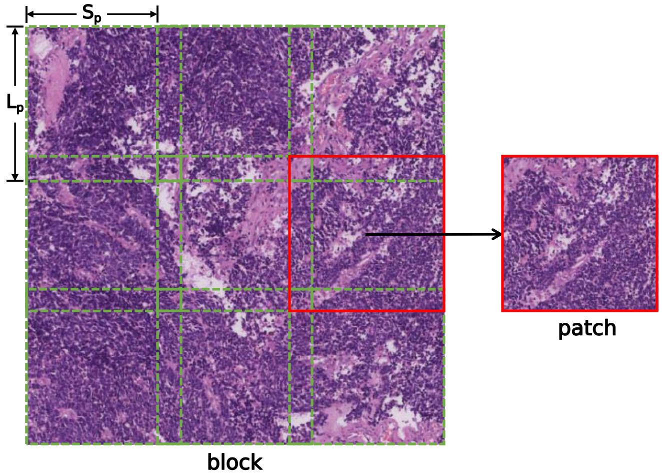


Fig. 6. Simplified example of a block with 3 × 3 overlapped patches.

B. Context-Aware Block Selection

We hypothesize that patches with higher predicted probability for a specific class are more likely to be true. Thus, the features extracted from such regions would be more reliable than those from regions with lower probability. Previous study [36] utilized all the discriminative patches and the corresponding features. However, this method always leads to feature redundancy during inference as CNN densely slides over the WSI and patches share the overlapped regions with their neighbors. On the other hand, due to the heterogeneity of histopathological characteristics, there exist outliers or mimics having high probabilities in WSIs. They usually exert a negative effect on the quality of subsequent WSI holistic feature representation, which eventually degrade the performance of image-level classification.

In order to tackle the aforementioned issues, we take the rich contextual information into account for better feature selection. Here, a *block* refers to a large grid that consists of a number of overlapped patches as shown in Fig. 6, then a WSI can be regarded as a composition of many blocks. In general, a tumor area has a larger size than a patch does, resulting in high probability scores appearing in a concentrated region. In other words, the average probability of tumor region would certainly be high. On the contrary, even if an outlier carrying a high probability value falls in a normal tissue block, it is easy to be filtered out due to the low average probability of such a block. Explicitly, we denote a *block* as

$$A = \begin{bmatrix} I_{1,1} & I_{1,2} & \dots & I_{1,n} \\ I_{2,1} & I_{2,2} & \dots & I_{2,n} \\ \vdots & \vdots & \vdots & \vdots \\ I_{n,1} & I_{n,2} & \dots & I_{n,n} \end{bmatrix}$$

where n is the number of overlapped patches in each row (column), and $I_{i,j}$ is the patch located at the ith row and the jth column in block A, where $i, j \in \{1, 2, \dots, n\}$. Specifically, the size of the block is: $L_b = L_p + S_p \times (n-1)$, where L_p represents the patch size and S_p is the patch stride [the distance between two adjacent patches in the same row (column)], which are 244 and 32, respectively, according to the architecture of the network shown in Fig. 4. In fact, the S_p is determined by the overall downsampling rate of FCN. During inference, for each patch $I_{i,j}$, the FCN outputs a feature embedding $f_{i,j} \in \mathbb{R}^{1024}$ from the penultimate convolutional layer and a probability vector $p_{i,j} = (p^1_{i,j}, p^2_{i,j}, \dots, p^C_{i,j})$ from the last layer, where $p^\ell_{i,j}$ means the probability score of patch $I_{i,j}$ for the ℓth class. For each class ℓ, the average probability within a block is calculated by $\bar{p}^\ell = (1/n^2) \sum_{i=1}^{n} \sum_{j=1}^{n} p^\ell_{i,j}$, which is used to identify the discriminative block by judging whether it exceeds a certain threshold τ. It is worth noting that in practice, we take blocks as inputs that go through the trained FCN directly for better efficiency and convenience; then, the resulting probability maps are averaged to obtain the average probability values. The hyperparameters are determined by cross-validation, that is, $\tau = 0.3$, $L_b = 884$, and $n = 21$.

C. WSI Feature Aggregation and Classification

A good holistic feature descriptor is essentially required to classify a WSI. Intuitively, it should integrate the global information from all cancer types and noncancer type. We call them positive evidence and negative evidence, respectively. Specifically, the positive evidence can well support the existence of cancer class that is consistent with the ground-truth label. In contrast, the negative evidence can manifest the absence of any other classes.

The general procedure to obtain the holistic representation of WSI consists of three-stage feature aggregations. First of all, we perform feature aggregation within each discriminative block. This can be regarded as patch-level feature fusion. The outcome of this phase is called *block descriptor* which is supposed to represent one block. Afterward, we fuse the information of all discriminative blocks to obtain the specific class feature, which is called *class descriptor*. It can support the existence or absence of a certain class. Eventually, all class descriptors are concatenated together to interpret the WSI, which is referred as *global descriptor*.

1) Block Descriptor: In the first stage, there are three different strategies to aggregate features within a discriminative block. The first approach is called *MaxFeat*, which takes the feature f of the patch with the highest probability as the block descriptor

$$\mathcal{B}^\ell = f_{i^*,j^*} \text{ s.t. } \left(i^*, j^*\right) = \arg\max_{i,j}\ p^\ell_{i,j} \tag{2}$$

where $\mathcal{B}^\ell \in \mathbb{R}^{1024}$ denotes the block descriptor for the class ℓ. The second strategy is the fusion of all patch-level features with equal contributions, called *AvgFeat*

$$\mathcal{B}^\ell = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} f_{i,j}. \tag{3}$$

Similarly, the last strategy considers all features within the same block, but the contribution of each individual patch-level feature to the block descriptor is directly proportional to its probability score. This strategy is referred as *WeightFeat* as follows:

$$\mathcal{B}^\ell = \sum_{i=1}^{n} \sum_{j=1}^{n} p^\ell_{i,j} f_{i,j}. \tag{4}$$

2) Class Descriptor: In the second stage, we try two approaches to aggregate block descriptors to obtain the class descriptor. One is to simply take average of all the discriminative block descriptors as the class descriptor, called *Mean-pool*

$$\mathcal{C}^\ell = \frac{1}{N} \sum_{k=1}^{N} \mathcal{B}^\ell_k \tag{5}$$

where $\mathcal{C}^\ell \in \mathbb{R}^{1024}$ represents the class descriptor for the class ℓ, and N denotes the number of discriminative blocks of the class ℓ. The other method is 3-norm pool [54] as depicted in (6), denoted by *Norm3*, which is also utilized in [35] to aggregate features

$$\mathcal{C}^\ell = \left(\sum_{k=1}^{N} \left(\mathcal{B}^\ell_k \right)^3 \right)^{\frac{1}{3}}. \tag{6}$$

In such a way, all class descriptors could have the same dimension.

3) Global Descriptor: All the class descriptors are concatenated together to generate the global descriptor $\mathcal{G} = \{\mathcal{C}^1, \mathcal{C}^2, \dots, \mathcal{C}^C\}$. The detailed processes of feature selection and aggregation are illustrated in Fig. 7 and Fig. 2(b) accordingly. Finally, the global descriptors with good holistic representation are fed into a standard RF classifier for WSI-level predictions.

III. Experiments and Results

In this article, two datasets are utilized to evaluate the performance of the proposed method. The relevant information and the corresponding results are as follows.

A. Experiments on SUCC Dataset

1) Dataset and Preprocessing: First, we constructed a large-scale dataset in collaboration with Sun Yat-sen University Cancer Center (SUCC), State Key Laboratory of Oncology in South China. This dataset contains comprehensive lung cancer classes and it is composed of 939 digitalized histology WSIs collected from 871 lung cancer patients and 68

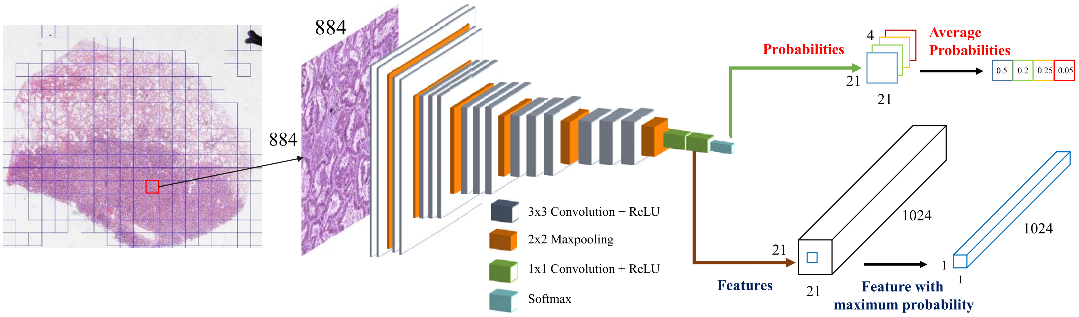


Fig. 7. Context-aware feature selection. A block, larger than a patch, with size $L_b \times L_b$ is regarded as discriminative only if its average probability exceeds a certain threshold τ. Then features are extracted from this block according to different strategies, for example, *MaxFeat*.

TABLE I
DATA DISTRIBUTION OF *SUCC* DATASET

		Carcinoma	SC	ADC	SCLC	NORM
Training	D1	59	21	20	18	-
	D2	642	267	293	82	-
	D3	-	-	-	-	53
Testing	D4	170	73	77	20	15

healthy subjects. Eight hundred and seventy one lung cancer WSIs are further diagnosed into three fine-grained categories of lung carcinoma, that is, SC, ADC, and small-cell lung carcinoma (SCLC) with 361, 390, and 120 WSIs accordingly, as shown in Fig. 1. All WSIs are obtained by using a Leica Aperio AT2 scanner at a 4× magnification with 0.25 μm/pixel resolution, and are stored in multiple zoom levels (3 or 4) with a pyramid-like structure. At the finest magnification, the WSIs come with an average size of 74 000 × 76 000 pixels. Within this dataset, merely 59 images were nonexhaustively annotated by a panel of experienced pathologists as illustrated in Fig. 5.

This dataset is divided into different groups as shown in Table I. We consider 59 annotated images as D1. Besides, the remaining 812 cancer images carrying the WSI-level label only are further split into 642 (D2) and 170 (D4) images for training and testing, respectively. Analogously, noncancer [normal (NORM)] WSIs are also divided into two parts containing 53 (D3) and 15 (D4) images for training and testing accordingly.

Considering that processing the images at the finest magnification level would be intractable due to the huge computation, we downsampled each WSI by a factor of 4 to the resolution of 1 μm/pixel during preprocessing. To enrich the training set, we applied several data augmentation techniques, including rotation, translation, flipping, and color jittering. Specifically, we first cropped a larger size tile at a random scaling ratio within the tissue region in the WSI, followed by a horizontal or vertical flipping with a fixed probability (i.e., 0.25). Afterward, random rotation with a degree ranging from 0 to 360 was applied. Then, we cropped the tile to have the desirable dimension (244 × 244 × 3). Finally, color jittering was employed in R, G, and B channel, respectively.

2) *Experimental Settings:*

a) *Configuration of training datasets:*

1) *M1:* In this experiment, D1 (59 cancer WSIs) and D3 (53 noncancer WSIs) are used for patch-based CNN training. All patches extracted from D1 and D3 only convey the WSI-level labels. Note that the coarse annotation masks are not utilized during the CNN training.
2) *M2:* It is quite similar to M1 except that the weighted loss function is employed during training, which gives higher penalty to patches extracted from the annotated regions.
3) *M3:* The training set consists of D1, D2, and D3 (i.e., 701 cancer images and 53 noncancer images). Note that the coarse annotation masks are not utilized during the CNN training.
4) *M4:* Analogous to M3, M4 utilizes the same training data: D1, D2, and D3. The only difference is that weighted loss function is applied to use the coarse annotations.

b) *Configuration of feature aggregation methods:* We implement different feature aggregation methods that are commonly used in the previous works to obtain the image-level prediction of each WSI.

1) *MajorityVoting:* We employ the CNN on the testing WSI and obtain a score map. The prediction of each location votes to the four classes and we take the category with majority vote as the prediction for the image.
2) *AveragePooling:* We calculate the average probability of the locations on the test WSI score map for each class channel. The category with the highest average probability is taken as the image-level prediction.
3) *MaxPooling:* We select the maximum probability of the locations on the test WSI score map for each class channel. The category with the highest max-pooling probability is taken as the image-level prediction.
4) *Count-Based RF:* We count the numbers of all cancer types and noncancer type prediction in test WSI score map to form a prediction histogram of classes. The four-bit histogram is fed into an RF classifier for the image-level prediction.
5) *Component-Based RF:* For each test WSI score map, the connected component with the largest area for each class is chosen as the ROI. Then, we obtain different features of this ROI, including maximum probability, average probability, area, eccentricity, convex area, orientation, extent, equivalent diameter, solidity, major axis length, minor axis length, and perimeter. Finally,

TABLE II
IMPACT OF α IN WEIGHTED LOSS FUNCTION ON CLASSIFICATION PERFORMANCE

	$\alpha = 1$	$\alpha = 2$	$\alpha = 3$	$\alpha = 4$	$\alpha = 5$
avgACC	0.779	**0.831**	0.795	0.804	0.790

TABLE III
RESULTS ON THE *SUCC* DATASET WITH DIFFERENT EXPERIMENTAL SETTINGS

Methods	M1	M2	M3	M4
MajorityVoting	0.708	0.719	0.665	0.697
AveragePooling	0.730	0.735	0.676	0.703
MaxPooling	0.530	0.681	0.616	0.627
Count-based RF	0.770	0.783	0.875	0.930
Component-based RF	0.748	0.759	0.909	0.935
CNN-AvgFeat-MeanPool-based RF	0.786	0.812	0.928	0.955
CNN-MaxFeat-MeanPool-based RF	0.732	0.824	0.953	0.971
CNN-WeightedFeat-MeanPool-based RF	0.767	0.858	0.932	0.960
CNN-AvgFea-Norm3-based RF	0.816	0.843	0.943	0.962
CNN-MaxFeat-Norm3-based RF	0.778	0.827	0.931	0.965
CNN-WeightedFeat-Norm3-based RF	0.789	0.811	0.941	**0.973**

TABLE IV
CONFUSION MATRIX OF WSI CLASSIFICATION RESULT

	Predictions			
Ground Truth	SC	ADC	SCLC	NORM
SC	72	0	1	0
ADC	2	75	0	0
SCLC	1	0	19	0
NORM	0	1	0	14

an RF classifier takes the feature vector as input to get the final prediction.

6) *CNN-AvgFeat-MeanPool-Based RF, CNN-WeightedFeat-MeanPool-Based RF, CNN-MaxFeat-MeanPool-Based RF:* Our proposed feature aggregation methods, *AvgFeat*, *WeightedFeat*, and *MaxFeat*, are used separately to obtain block descriptors; then, the Mean-pool is utilized to fuse block descriptors to obtain the class descriptor. After feature aggregation, an RF gives the WSI-level prediction based on the global descriptor.

7) *CNN-AvgFeat-Norm3-Based RF, CNN-WeightedFeat-Norm3-Based RF, CNN-MaxFeat-Norm3-Based RF:* Similarly, *AvgFeat*, *WeightedFeat*, and *MaxFeat* are used to generate block descriptors, respectively; then, 3-norm pool is adopted to aggregate block descriptors to obtain the class descriptor. After feature aggregation, an RF gives the WSI-level prediction based on the global descriptor.

3) Quantitative Evaluation: We employ *accuracy* as the evaluation criterion for this multiclass WSI classification task. We apply different training strategies with different block selection and feature aggregation methods as ablation studies to evaluate the contribution of each crucial component within our framework.

At first, to investigate the impact of the value of the hyperparameter α in the weighted loss function on WSI analysis, we utilize the weighted loss function with different values of α to train the network. Note that when $\alpha = 1$, no weight is imposed on annotations and each training patch contributes equally to the loss. With respect to each value of α, we apply four-fold cross-validation on 400 WSIs from our training set (D1, D2, and D3). In other words, the 400 WSIs are randomly split into four equal sized groups. Of the four groups, a single group is retrained as the validation set for testing the model, and the remaining three groups are used to train the model. This process is repeated four times. For each validation set, we calculated the mean accuracy (meanACC) of our proposed feature aggregation strategies (the last six methods in Table III). Then, we take the average of meanACCs from all validation sets as the final result (avgACC), as shown in Table II. Apparently, results of the weighted loss function ($\alpha \geq 2$) are better than that of the classical cross-entropy loss function ($\alpha = 1$). It indicates that emphasizing weight on annotations can boost the performance. As the value of α increases, it achieves the peak performance when $\alpha = 2$, but a drop is observed when α becomes larger. Therefore, we fix the value of α as 2 for the following experiments.

Once the optimum value of α is determined, we carry a thorough analysis on the *SUCC* dataset. The experimental results are listed in Table III. As for the first three simple prediction strategies, *MajorityVoting*, *AveragePooling*, and *MaxPooling*, the performance is not very competitive. The reason behind could be that these methods only rely on instances so that they are in lack of effective holistic information of WSIs. Besides, there is no gain of improvement by adding more training samples as these methods are quite sensitive to outliers, which can degrade the performance. With respect to *Count-based RF* and *Component-based RF*, there is a considerable boost on accuracy. These two methods integrate information from instances and create a global feature vector for the second-stage classifier, so they are more robust to output the WSI-level prediction. Inspiringly, context-aware CNN feature selection and aggregation methods outperform the preceding methods by a large margin, because not only the feature of multiple instances is used but also rich spatial information is taken advantage of. It is validated that the features learned by CNN are more representative than count-based histogram and component-based features. We can observe that the *CNN-MaxFeat-Norm3-based RF* classifier achieves the best result among all the WSI-level prediction strategies with the training setting M4. In addition, we can also notice that a small number of coarse annotations (M4) can contribute to the accuracy improvement compared to that without coarse annotations (M3).

With a closer observation of the predictions of each class as shown in the confusion matrix in Table IV (obtained by *M4-CNN-WeightedFeat-Norm3-based RF*), we can see that there are very few misclassifications within each class (i.e., only 1, 2, 1, and 1 misclassified WSI(s) for SC, ADC, SCLC, and NORM, respectively). Notably, no cancer WSIs are misclassified as normal, and it is very critical in clinical practice. Inspiringly, we achieve the Cohen's Kappa $k = 0.958$, reaching fairly good agreement with the histopathologists.

TABLE V
COMPARISON WITH THE STATE-OF-THE-ART METHODS ON THE *SUCC* DATASET

Methods	Accuracy
Pretrained-Feature-Norm3 [35]	0.874
EM-CNN-SVM [36]	0.914
M3-CNN-MaxFeat-MeanPool-based RF (Ours)	0.953
M4-CNN-WeightedFeat-Norm3-based RF (Ours)	**0.973**

We further implemented two state-of-the-art weakly supervised learning methods on WSI analysis for comparison. One is the EM-based CNN with a supervised decision fusion model proposed by Hou *et al.* [36]. In this approach, an EM-based CNN is used to iteratively eliminate the nondiscriminative patches. Then, the class histogram of the patch-level predictions is taken as the input to an SVM with *radial basis function* kernel to predict the image-level label. The other is a CNN activation feature-based method [35]. Each WSI is divided into a set of patches (336 × 336) that have 25% overlap with adjacent patches. Then, all patches are resized into 224 × 224 as the input of VGG-16 that is pretrained on ImageNet for feature extraction. Then, the 3-norm pooling is used to aggregate all features and feature dimension reduction is applied to remove irrelevant features. Finally, a linear SVM classifier outputs the WSI-level prediction. From Table V, we can see that our methods overwhelm these two approaches significantly.

It is known that the off-the-shelf features extracted by CNN pretrained on other domain are quite generic, so Pretrained-Feature-Norm3 [35] fails to achieve high accuracy. It implies that fine-tuning with histopathology images is necessary for CNN to learn more useful discriminative patterns. On the other hand, although EM-CNN-SVM [36] trains a patch-based CNN iteratively and employs the count-based histogram as the global feature at the second-stage classifier, the class histogram is less representative than the deep features learned by CNN. Obviously, our success primarily owes to the representative features from CNN in conjunction with context-aware feature selection and aggregation strategies since the quality of the global descriptor representing a WSI is crucial for WSI-level classification.

4) Qualitative Evaluation:

a) Discriminative region detection: Albeit our ultimate task is not tumor detection, our method can achieve such a goal simultaneously by retrieving the most discriminative regions. We invite a specialized pathologist to delineate the discriminative regions elaborately on a few testing WSIs. The results are depicted in Fig. 8. Here, Fig. 8(a) and (b) shows the WSI and ground truth (red regions in Fig. 8(b) denote carcinoma regions in cancer WSIs or normal tissue in the normal WSI) respectively, followed by heatmaps generated from CNN with the training setting M3 and M4 in Fig. 8(c) and (d) accordingly. Note that the first three rows are cancer cases, and the last row is a normal case. Clearly, despite that the patch-based CNN could learn discriminative patterns from WSIs without the aid of annotation information (M3), it sometimes fails to find the most discriminative regions, which might lead to feature deficiency for the positive evidence. On the contrary, robustness is obviously improved by using only a handful of annotations. The heated regions found by M4 is more consistent with the annotation from the pathologist. These visualization results validate that our system makes the diagnosis decisions based on really discriminative regions.

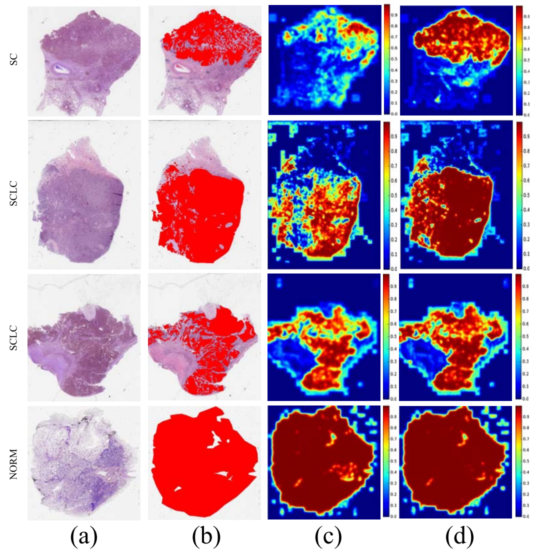


Fig. 8. Visualization of discriminative region detection. (a) WSI. (b) Ground Truth. (c) M3: Heatmap. (d) M4: Heatmap.

b) Feature embedding visualization: In order to evaluate the effect of different training data sources on global features of WSIs, we project those global features onto a 2-D space for visual comparison with t-distributed stochastic neighborhood embedding (t-SNE). In Fig. 9(a), results are obtained via VGG-16 network pretrained on ImageNet [35] while Fig. 9(b) shows our results from ScanNet training on lung cancer WSIs. It can be clearly seen from Fig. 9(b) that the proposed method distinctively maximizes the distance among different interclasses. Besides, although there are some hard samples near the decision boundary which are difficult to be identified correctly in Fig. 9(b), most of WSIs belonging to the same category form an individual cluster, indicating that the holistic feature representation obtained by our method is very helpful for image-level classification even using the standard RF classifier.

c) Failure cases analysis: There are a few failure cases misclassified by our method as illustrated in Fig. 10. Apparently, the patch-based CNN does not manage to retrieve the discriminative regions, which eventually leads to inappropriate global representation of the WSI. Failure cases may be due to the following reasons. First, both misclassified ADC and SC images belong to poorly differentiated cancer types. That means the morphological patterns of carcinoma within these WSIs might be ambiguous, so that the patch-based CNN is highly likely to misidentify such cancer areas. Second, within misclassified lung cancer WSIs, valid cancerous regions might occupy only a very small part, whereas normal areas dominate the entire image. The misclassified normal WSI has distinct structural features compared to other normal ones, leading to the misclassification.

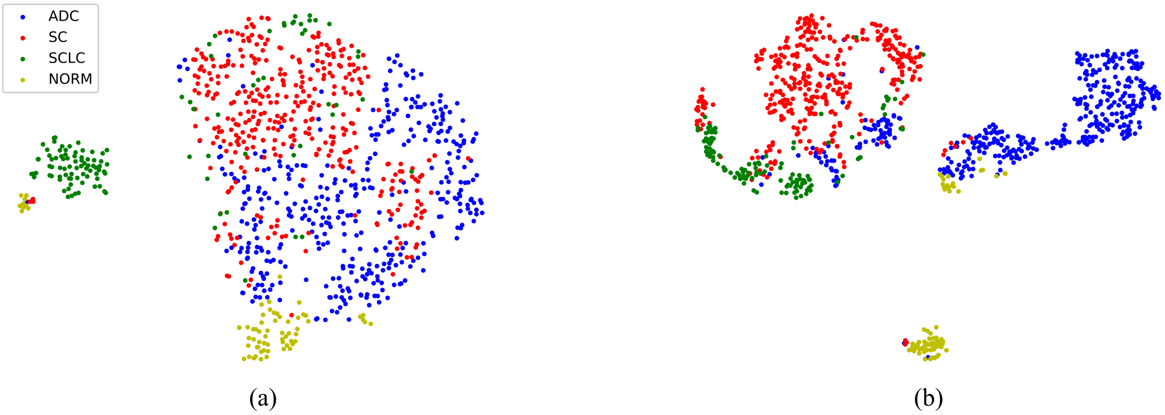


Fig. 9. 2-D projection of the holistic features used for second-stage classifier obtained from (a) pretrained VGG-16 network and (b) ScanNet trained on lung cancer WSIs with easier feature separation among interclasses. Best view in color.

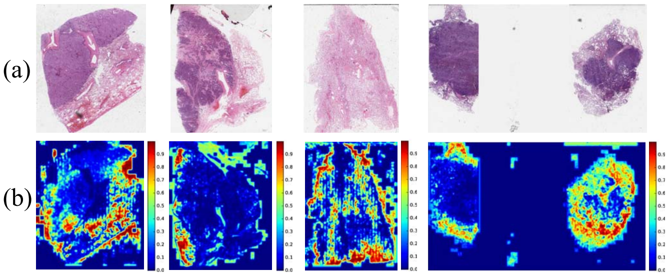


Fig. 10. Failure cases. (a) Upper panel denotes the entire slide images of different categories, ADC, SCLC, NORM, and SC, respectively. (b) Bottom panel shows the corresponding discriminative regions output by ScanNet with M4.

B. Experiments on TCGA Dataset

1) Dataset and Configurations: To verify the efficacy and generalization of the proposed method, we also conducted experiments on a publicly available lung cancer WSI dataset from the The Cancer Genome Atlas (TCGA).[1] We used 500 WSIs in good quality from the Genomic Data Commons database, composed of 250 ADC images and 250 SC images. We random split the dataset into training set and testing set, with 400 WSIs and 100 WSIs, respectively. There are equal numbers of two classes in training or testing set. Due to the deficiency of the annotations of the tumor regions in this dataset, the typical binary cross-entropy loss function is adopted to train the ScanNet. In addition, we employ *accuracy* and area under the receiver operating characteristic curve (AUC) as the evaluation criteria for this binary classification problem.

The experimental results are reported in Table VI, along with the comparison to the state-of-the-arts. Encouragingly, our proposed approach outperforms these methods considerably achieving the best performance on the two metrics with 5% and 4% gain on accuracy and AUC compared to EM-CNN-Fea-SVM [36] and with 5% and 2.4% gain on accuracy and AUC compared to Pretrained-Feature-Norm3 [35].

TABLE VI
COMPARISON WITH THE STATE-OF-THE-ART METHODS ON THE *TCGA* DATASET

Methods	Accuracy	AUC
Pretrained-Feature-Norm3 [35]	0.770	0.832
EM-CNN-Fea-SVM [36]	0.770	0.816
CNN-AvgFea-Norm3-based RF(Ours)	**0.820**	**0.856**

C. Implementation Details and Computation Cost

Our method was implemented with Python based on the open-source deep learning library Tensorflow on a workstation equipped with 3.5-GHz Intel Core i7-5930K CPU and two GPUs of Nvidia GeForce GTX Titan X. At the training phase, the network randomly cropped tiles from the WSI with the dimension of at least $\sqrt{2}$ larger than the training patch size, which could guarantee that the following arbitrary-degree rotation would not result in invalid region appearing in the training patches. While in the inference stage, the FCN densely scanned the WSI, where the *block* size was set as 884×884 and the outcomes of a block were a $21 \times 21 \times C$ probability tile and a 1024-bit feature vector. For the *SUCC* and *TCGA* datasets, C was 4 and 2, respectively. In order to avoid the boundary effect, the FCN scanned the WSI with a stride of 32×21, where 32 was the total downsampling rate by ScanNet and 21 was the edge length of the probability tile. Ultimately, all probability tiles generated from blocks were stitched together to obtain the probability map of WSIs.

The network was trained with stochastic gradient descent method [55]. We set the batch size as 100 and the learning rate as 0.0001 initially and then decreased gradually by a factor of 10 every 10 000 iterations. The parameters were randomly initialized from a Gaussian distribution ($\mu = 0, \sigma = 0.01$) and updated by a standard backpropagation.

Efficiency is a key factor for WSI analysis, hence, we investigate the speed of our method. In general, it took about 100 s to process one WSI with the dimension of $20\,500 \times 19\,500$ ($1\ \mu$m/pixel), which achieved more than hundreds of time faster than the traditional patchwise classification framework with the same stride. Considering the increasing number of

[1] https://portal.gdc.cancer.gov/

large-scale WSIs, the promising efficiency implies a great potential of our method in clinical practice.

IV. Discussion

Histopathology plays an essential role in cancer diagnosis and serves as the gold standard in many medical protocols. However, the examination of whole slide histopathology image is always the bottleneck for in-time treatments. Especially, in some developing countries (e.g., China), the supply of histopathologists severely falls short of demand, therefore, most of the histopathologists have to bear the heavy burden and cancer patients cannot be treated timely. Hence, automated analysis of histopathology is highly welcome to ease the workload and address the issue of medical manpower shortage on underserved populations and areas. Unlike tumor detection and segmentation tasks where a large number of pixelwise annotations are provided for researchers to design fully supervised learning algorithms, the WSI classification problem is much more challenging since only image-level labels are available in most cases. In addition, tumor regions are generally sparsely distributed in the entire image, mingling with a large proportion of noncancer regions, which results in more difficulties. Consequently, the ground-truth label of each individual patch extracted from the WSI is somehow ambiguous. Another notoriously tough issue in histopathology image analysis is ascribed to the large variation of tissue appearance. Even for the same type of carcinoma/noncancer, the morphological structures and textures are of great diversity, which is actually one of the major stumbling blocks to design a robust automated analysis tool. Although significant achievements have been realized in the past few years, many problems still remain unsolved. Most of the previous traditional weakly supervised learning methods (e.g., MIL-based techniques) are developed upon a great number of handcrafted features, which largely restrain their generality and transferability. In addition, current accessible public datasets of lung cancer WSIs (e.g., The *TCGA* cohort) merely contain ADC and SC classes, which definitely impede the development of a comprehensive multiclass classifier that is highly demanded in clinical practice.

Most of the aforementioned issues have been well tackled in this article. At first, a comprehensive dataset *SUCC* is built to satisfy the underlying requirement of identification for different lung cancer subtypes. Based upon this dataset, we propose a novel weakly supervised learning approach to address the WSI classification problem by exploring the deep learning feature selection and aggregation. The weighted loss function, to the best of our knowledge, is the first trial that exploits weakly supervised learning on WSI classification using image-level labels as well as a small number of coarse annotations. To incorporate rich spatial information, we propose different strategies, for example, *AvgFeat*, *WeightedFeat*, and *MaxFeat* for block descriptor, and *MeanPool* and *Norm3* for class descriptor, to obtain the holistic feature representation of WSI. Extensive experiments performed on two datasets (*SUCC* and *TCGA*) verify the effectiveness of our method. It surpasses other two weakly supervised learning methods by a significant margin.

Inherently, our method falls into the cohort of weakly supervised learning method, due to scarcity of elaborated delineation of cancer regions. Although our proposed weighted loss function demonstrates significant improvements, the annotated regions can be noisy along with majority of annotations being correct as illustrated in Fig. 5. It would be quite tedious for pathologists to indicate the noisy regions in a manual way. Furthermore, it might be extremely challenging for pathologists to discriminate ambiguous regions. Therefore, how to automatically select training samples with real ground-truth labels and eliminate noisy regions in the training process would be a very promising direction to improve the accuracy. We will explore this direction in the future work.

Although our method achieved fairly good results in the experiments, there are still some space for further improvement. First, the feature embedding of patches for the holistic descriptor is little complicated as a number of hyperparameters need to be determined. Automated feature selection and aggregation by adaptive learning would be more straightforward and attractive in the future. In addition, the framework in our method is not an end-to-end pipeline, where the feature assembling and classification are two individual modules. In the future work, we would replace the RF classifier with MLP classifier to make it end-to-end. Besides, the proposed method still could not properly deal with ambiguous regions in WSIs due to complex technique variations (e.g., variations of color/texture) and biological heterogeneities (e.g., cell type and cell state) that are always present in a large cohort. Furthermore, we plan to use more lung cancer datasets to validate the generalization capability of the proposed weakly supervised learning method.

V. Conclusion

In this article, we proposed a weakly supervised learning method to address the whole slide lung cancer image classification problem with minimum annotation effort. The exploited FCN can efficiently generate cancer likelihood that helps to retrieve discriminative regions. Besides, deep learning features extracted by CNN are the ideal substitute of handcrafted features. We proposed different context-aware block selection and feature aggregation strategies to obtain an effective holistic feature representation of the WSI. In order to validate the efficacy of the proposed method, we first constructed the largest fine-grained lung cancer WSI dataset *SUCC* for comprehensive analysis, and then evaluated our method on a public lung cancer WSIs dataset from TCGA. Extensive experiments corroborated the superiority of the proposed method which outperformed the state-of-the-art methods significantly on two datasets. We believe that the proposed method can alleviate the bottleneck of expert annotation cost and advance the progress of computer-aided histology image analysis.

References

[1] *Cancer Facts & Figures*, Amer. Chem. Soc., Washington, DC, USA, 2018.

[2] *Types and Staging of Lung Cancer*. Accessed: May 4, 2018. [Online]. Available: https://www.lungcancer.org/find_information/publications/163-lung_cancer_101/268-types_and_staging

[3] H. Chen *et al.*, "Ultrasound standard plane detection using a composite neural network framework," *IEEE Trans. Cybern.*, vol. 47, no. 6, pp. 1576–1586, Jun. 2017.

[4] X. Yang *et al.*, "Fine-grained recurrent neural networks for automatic prostate segmentation in ultrasound images," in *Proc. AAAI*, 2017, pp. 1633–1639.

[5] Q. Dou *et al.*, "Automatic detection of cerebral microbleeds from mr images via 3D convolutional neural networks," *IEEE Trans. Med. Imag.*, vol. 35, no. 5, pp. 1182–1195, May 2016.

[6] D. Nie, L. Wang, E. Adeli, C. Lao, W. Lin, and D. Shen, "3-D fully convolutional networks for multimodal isointense infant brain image segmentation," *IEEE Trans. Cybern.*, vol. 49, no. 3, pp. 1123–1136, Mar. 2019.

[7] H. Chen *et al.*, "Automatic localization and identification of vertebrae in spine CT via a joint learning model with deep neural networks," in *Proc. Int. Conf. Med. Image Comput. Comput. Assist. Intervent.*, 2015, pp. 515–522.

[8] X. Li, H. Chen, X. Qi, Q. Dou, C.-W. Fu, and P.-A. Heng, "H-Denseunet: Hybrid densely connected UNet for liver and tumor segmentation from CT volumes," *IEEE Trans. Med. Imag.*, vol. 37, no. 12, pp. 2663–2674, Dec. 2018.

[9] D. C. Cireşan, A. Giusti, L. M. Gambardella, and J. Schmidhuber, "Mitosis detection in breast cancer histology images with deep neural networks," in *Proc. Int. Conf. Med. Image Comput. Assist. Intervent.*, 2013, pp. 411–418.

[10] H. Chen, Q. Dou, X. Wang, J. Qin, and P.-A. Heng, "Mitosis detection in breast cancer histology images via deep cascaded networks," in *Proc. AAAI*, 2016, pp. 1160–1166.

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

[12] A. M. Khan, K. Sirinukunwattana, and N. Rajpoot, "A global covariance descriptor for nuclear atypia scoring in breast histopathology images," *IEEE J. Biomed. Health Informat.*, vol. 19, no. 5, pp. 1637–1647, Sep. 2015.

[13] H. Chang, Y. Zhou, A. Borowsky, K. Barner, P. Spellman, and B. Parvin, "Stacked predictive sparse decomposition for classification of histology sections," *Int. J. Comput. Vis.*, vol. 113, no. 1, pp. 3–18, 2015.

[14] T. Qaiser, K. Sirinukunwattana, K. Nakane, Y.-W. Tsang, D. Epstein, and N. Rajpoot, "Persistent homology for fast tumor segmentation in whole slide histology images," *Procedia Comput. Sci.*, vol. 90, pp. 119–124, Jul. 2016.

[15] G. Litjens *et al.*, "Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis," *Sci. Rep.*, vol. 6, May 2016, Art. no. 26286.

[16] A. Agarwalla, M. Shaban, and N. M. Rajpoot, "Representation-aggregation networks for segmentation of multi-gigapixel histology images," *arXiv preprint arXiv:1707.08814*, 2017.

[17] D. Wang, A. Khosla, R. Gargeya, H. Irshad, and A. H. Beck, "Deep learning for identifying metastatic breast cancer," *arXiv preprint arXiv:1606.05718*, 2016.

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

[19] B. E. Bejnordi *et al.*, "Context-aware stacked convolutional neural networks for classification of breast carcinomas in whole-slide histopathology images," *J. Med. Imag.*, vol. 4, no. 4, 2017, Art. no. 044504.

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

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

[22] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[23] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[24] B. E. Bejnordi, G. Litjens, M. Hermsen, N. Karssemeijer, and J. A. van der Laak, "A multi-scale superpixel classification approach to the detection of regions of interest in whole slide histopathology images," in *Proc. Med. Imag. Digit. Pathol.*, vol. 9420, 2015, Art. no. 94200H.

[25] D. Ganti, "Lung cancer subtype classification from whole slide histopathological images," Ph.D. dissertation, Comput. Sci. Eng., Univ. Texas at Arlington, Arlington, TX, USA, 2015.

[26] P. Courtiol, E. W. Tramel, M. Sanselme, and G. Wainrib, "Classification and disease localization in histopathology using only global labels: A weakly-supervised approach," *arXiv preprint arXiv:1802.02212*, 2018.

[27] H. Wang, F. Xing, H. Su, A. Stromberg, and L. Yang, "Novel image markers for non-small cell lung cancer classification and survival prediction," *BMC Bioinformat.*, vol. 15, no. 1, p. 310, 2014.

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

[29] X. Luo *et al.*, "Comprehensive computational pathological image analysis predicts lung cancer prognosis," *J. Thoracic Oncol.*, vol. 12, no. 3, pp. 501–509, 2017.

[30] P. Mobadersany *et al.*, "Predicting cancer outcomes from histology and genomics using convolutional networks," *Proc. Nat. Acad. Sci. USA*, vol. 115, no. 13, pp. E2970–E2979, 2018.

[31] X. Zhu, J. Yao, F. Zhu, and J. Huang, "WSISA: Making survival prediction from whole slide histopathological images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 7234–7242.

[32] S. Wang *et al.*, "Comprehensive analysis of lung cancer pathology images to discover tumor shape and boundary features that predict survival outcome," *Sci. Rep.*, vol. 8, no. 1, 2018, Art. no. 10393.

[33] A. M. Khan, K. Sirinukunwattana, and N. Rajpoot, "Geodesic geometric mean of regional covariance descriptors as an image-level descriptor for nuclear atypia grading in breast histology images," in *Proc. Int. Workshop Mach. Learn. Med. Imag.*, 2014, pp. 101–108.

[34] H. Chang, A. Borowsky, P. Spellman, and B. Parvin, "Classification of tumor histology via morphometric context," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2013, pp. 2203–2210.

[35] Y. Xu *et al.*, "Deep convolutional activation features for large scale brain tumor histopathology image classification and segmentation," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2015, pp. 947–951.

[36] L. Hou, D. Samaras, T. M. Kurc, Y. Gao, J. E. Davis, and J. H. Saltz, "Patch-based convolutional neural network for whole slide tissue image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2424–2433.

[37] B. Korbar *et al.*, "Deep learning for classification of colorectal polyps on whole-slide images," *J. Pathol. Informat.*, vol. 8, p. 30, Jul. 2017.

[38] S. Graham, M. Shaban, T. Qaiser, S. A. Khurram, and N. Rajpoot, "Classification of lung cancer histology images using patch-level summary statistics," in *Proc. Med. Imag. Digit. Pathol.*, vol. 10581, 2018, Art. no. 1058119.

[39] N. Coudray *et al.*, "Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning," *Nat. Med.*, vol. 24, no. 10, pp. 1559–1567, 2018.

[40] J. Barker, A. Hoogi, A. Depeursinge, and D. L. Rubin, "Automated classification of brain tumor type in whole-slide digital pathology images using local representative tiles," *Med. Image Anal.*, vol. 30, pp. 60–71, May 2016.

[41] J. Amores, "Multiple instance classification: Review, taxonomy and comparative study," *Artif. Intell.*, vol. 201, pp. 81–105, Aug. 2013.

[42] Y. Xiao, B. Liu, Z. Hao, and L. Cao, "A similarity-based classification framework for multiple-instance learning," *IEEE Trans. Cybern.*, vol. 44, no. 4, pp. 500–515, Apr. 2014.

[43] Y. Xia, L. Nie, L. Zhang, Y. Yang, R. Hong, and X. Li, "Weakly supervised multilabel clustering and its applications in computer vision," *IEEE Trans. Cybern.*, vol. 46, no. 12, pp. 3220–3232, Dec. 2016.

[44] Y. Xu, J. Zhang, E. I.-C. Chang, M. Lai, and Z. Tu, "Contexts-constrained multiple instance learning for histopathology image analysis," in *Proc. MICCAI*, vol. 7512, 2012, pp. 623–30.

[45] Y. Xu, J.-Y. Zhu, E. I.-C. Chang, M. Lai, and Z. Tu, "Weakly supervised histopathology cancer image segmentation and classification," *Med. Image Anal.*, vol. 18, no. 3, pp. 591–604, 2014.

[46] Y. Xu *et al.*, "Deep learning of feature representation with multiple instance learning for medical image analysis," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2014, pp. 1626–1630.

[47] C. Mercan, E. Mercan, S. Aksoy, L. G. Shapiro, D. L. Weaver, and J. G. Elmore, "Multi-instance multi-label learning for whole slide breast histopathology," in *Proc. Med. Imag. Digit. Pathol.*, vol. 9791, 2016, Art. no. 979108.

[48] W. Zhu, Q. Lou, Y. S. Vang, and X. Xie, "Deep multi-instance networks with sparse label assignment for whole mammogram classification," in *Proc. Int. Conf. Med. Image Comput. Comput. Assist. Intervent.*, 2017, pp. 603–611.

[49] H. Yoshida *et al.*, "Automated histological classification of whole-slide images of gastric biopsy specimens," *Gastric Cancer*, vol. 21, no. 2, pp. 249–257, 2018.

[50] C. Mercan, S. Aksoy, E. Mercan, L. G. Shapiro, D. L. Weaver, and J. G. Elmore, "Multi-instance multi-label learning for multi-class classification of whole slide breast histopathology images," *IEEE Trans. Med. Imag.*, vol. 37, no. 1, pp. 316–325, Jan. 2018.

[51] 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.*, 2016, pp. 2818–2826.

[52] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. SMC-9, no. 1, pp. 62–66, Jan. 1979.

[53] H. Chen, X. Qi, L. Yu, Q. Dou, J. Qin, and P.-A. Heng, "DCAN: Deep contour-aware networks for object instance segmentation from histology images," *Med. Image Anal.*, vol. 36, pp. 135–146, Feb. 2017.

[54] Y.-L. Boureau, J. Ponce, and Y. LeCun, "A theoretical analysis of feature pooling in visual recognition," in *Proc. 27th Int. Conf. Mach. Learn. (ICML)*, 2010, pp. 111–118.

[55] L. Bottou, "Large-scale machine learning with stochastic gradient descent," in *Proc. COMPSTAT*, 2010, pp. 177–186.

Xi Wang received the bachelor's degree in software engineering from Southwest University, Chongqing, China, in 2013, and the master's degree in computer science from Sichuan University, Chengdu, China, in 2016. She is currently pursuing the Ph.D. degree in computer science and engineering with The Chinese University of Hong Kong, Hong Kong, China.

Her current research interests include medical image analysis, deep learning, and weakly supervised learning.

Hao Chen (S'13–M'17) received the bachelor's degree in information engineering from Beihang University, Beijing, China, in 2013, and the Ph.D. degree in computer science and engineering from The Chinese University of Hong Kong, Hong Kong, China, in 2017.

His current research interests include medical image analysis, deep learning, object detection, and segmentation.

Dr. Chen was a recipient of the best paper awards for three of his research works.

Caixia Gan received the bachelor's degree in clinical pathology from Southern Medical University, Guangzhou, China, in 2015.

She was certified as a Resident in standardized training with the Pathology Department, Sun Yat-sen University Cancer Center, Guangzhou, in 2018. Since 2018, she has been with the Pathology Department, Affiliated Hexian Memorial Hospital, Southern Medical University. She researches on clinicopathology and intends to specialize in obstetrics and gynaecology.

Huangjing Lin received the bachelor's degree in computer science from Inner Mongolia University, Huhhot, China, in 2011, and the master's degree in information science and engineering from Xiamen University, Xiamen, China, in 2015. He is currently pursuing the Ph.D. degree in computer science and engineering with The Chinese University of Hong Kong, Hong Kong, China.

His current research interests include medical image analysis, deep learning, object detection, and segmentation.

Qi Dou (S'14–M'18) received the bachelor's degree in biomedical engineering from Beihang University, Beijing, China, in 2014, and the Ph.D. degree from the Department of Computer Science and Engineering, The Chinese University of Hong Kong, Hong Kong, China, in 2018.

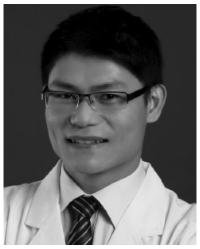
She is currently a Post-Doctoral Research Associate with the Biomedical Image Analysis Group, Department of Computing, Imperial College London, London, U.K. Her current research interests include medical image analysis, deep learning, and machine learning.

Efstratios Tsougenis received the bachelor's and Ph.D. degrees in production and management engineering from the Democritus University of Thrace, Komotini, Greece, in 2008 and 2013, respectively. During his Ph.D. he focused on computer vision and machine learning field.

In 2013, he was a Post-Doctoral Fellow with the Department of Computer Science and Engineering, Hong Kong University of Science and Technology, Hong Kong, China, under the supervision of Prof. C. K. Tang. He currently serves as the Technical Director of Imsight Medical Technology Company Ltd., Shenzhen, China, focusing on medical imaging analysis and deep learning. He also serves as a Referee in a number of established journals and conferences and a Guest Lecturer with the City University of Hong Kong, Hong Kong University Medicine School and Hong Kong Hospital Authority.

Qitao Huang received the bachelor's degree in medical laboratory from Guangzhou Medical University, Guangzhou, China, in 2016 and the Medical Information Management from Guangdong Pharmaceutical University, Guangzhou, in 2008.

His current research interests include pathological technology and AI in pathological diagnosis.



Muyan Cai received the M.D. degree and the Ph.D. degree in oncology from Sun Yat-sen University, Guangzhou, China, in 2005 and 2014, respectively.

He was a Post-Doctoral Fellow with the Dana-Farber Cancer Institute, Harvard Medical School, Boston, MA, USA, in 2019. His current research interests include tumor DNA repair, tumor metastatic mechanisms, and AI in pathological diagnosis. He has published lots of highly impact papers in well-known journals, including *Gut*, the *Journal of Clinical Investigation*, *Oncogene*, *PLoS Genetics*, and *EBioMedicine*.

Pheng-Ann Heng (M'89–SM'06) received the B.Sc. degree in computer science from the National University of Singapore, Singapore, in 1985, and the M.Sc. degree in computer science, the M.Art degree in applied math, and the Ph.D. degree in computer science from the Indiana University, Bloomington, IN, USA, in 1987, 1988, and 1992, respectively.

He is a Professor with the Department of Computer Science and Engineering, The Chinese University of Hong Kong, Hong Kong, China, where he has served as the Director of Virtual Reality, Visualization and Imaging Research Center since 1999 and has also served as the Department Chairman and the Head of Graduate Division. He has served as the Director of Center for Human–Computer Interaction, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China, since 2006 and has been appointed by the China Ministry of Education as a Cheung Kong Scholar Chair Professor in 2007. His current research interests include AI and VR for medical applications, surgical simulation, visualization, and graphics and human–computer interaction.