Constrained Deep Weak Supervision
for Histopathology Image Segmentation

Zhipeng Jia, Xingyi Huang, Eric I-Chao Chang and Yan Xu*

arXiv:1701.00794v1 [cs.CV] 3 Jan 2017

Abstract—In this paper, we develop a new weakly-supervised learning algorithm to learn to segment cancerous regions in histopathology images. Our work is under a multiple instance learning framework (MIL) with a new formulation, deep weak supervision (DWS); we also propose an effective way to introduce constraints to our neural networks to assist the learning process. The contributions of our algorithm are threefold: (1) We build an end-to-end learning system that segments cancerous regions with fully convolutional networks (FCN) in which image-to-image weakly-supervised learning is performed. (2) We develop a deep week supervision formulation to exploit multi-scale learning under weak supervision within fully convolutional networks. (3) Constraints about positive instances are introduced in our approach to effectively explore additional weakly-supervised information that is easy to obtain and enjoys a significant boost to the learning process. The proposed algorithm, abbreviated as DWS-MIL, is easy to implement and can be trained efficiently. Our system demonstrates state-of-the-art results on large-scale histopathology image datasets and can be applied to various applications in medical imaging beyond histopathology images such as MRI, CT, and ultrasound images.

Index Terms—Convolutional neural networks, histopathology image segmentation, weakly supervised learning, fully convolutional networks, multiple instance learning.

I. Introduction

HIGH resolution histopathology images play a critical role in cancer diagnosis, providing essential information to separate non-cancerous tissues from cancerous ones. A variety of classification and segmentation algorithms have been developed in the past [1], [2], [3], [4], [5], [6], [7], [8], focusing primarily on the design of local pathological patterns, such as morphological [2], geometric [1], and texture [9] features based on various clinical characteristics.

In medical imaging, supervised learning approaches [10], [11], [12] have shown their particular effectiveness in performing image classification and segmentation for modalities such

Xingyi Huang, and Yan Xu are with State Key Laboratory of Software Development Environment and Key Laboratory of Biomechanics and Mechanobiology of Ministry of Education and Research Institute of Beihang University in Shenzhen, Beihang University, Beijing 100191, China (email: huangxingyi102@126.com; xuyan04@gmail.com).

Zhipeng Jia, Eric I-Chao Chang, and Yan Xu are with Microsoft Research, Beijing 100080, China (email: zhipeng.jia@outlook.com; echang@microsoft.com; xuyan04@gmail.com).

Zhipeng Jia is with Institute for Interdisciplinary Information Sciences, Tsinghua University, Beijing 100084, China (email: zhipeng.jia@outlook.com).

as MRI, CT, and Ultrasound. However, the success of these supervised learning algorithms depends on the availability of a large amount of high-quality manual annotations/labeling that are often time-consuming and costly to obtain. In addition, well-experienced medical experts themselves may have a disagreement on ambiguous and challenging cases. Unsupervised learning strategies where no expert annotations are needed point to a promising but thus far not clinically practical direction.

In-between supervised and unsupervised learning, weakly-supervised learning in which only coarse-grained (image-level) labeling is required makes a good balance of having a moderate level of annotations by experts while being able to automatically explore fine-grained (pixel-level) classification [13], [14], [15], [16], [17], [18], [19], [20]. In pathology, a pathologist annotates whether a given histopathology image has a cancer or not; a weakly-supervised learning algorithm would hope to automatically detect and segment cancerous tissues based on a collection of histopathology (training) images annotated by expert pathologists; this process that substantially reduces the amount of work for annotating cancerous tissues/regions falls into the category of weakly-supervised learning, or more specifically multiple instance learning [13], which is the main topic of this paper.

Multiple instance learning (MIL) was first introduced by Dieterich et al. [13] to predict drug activities; a wealthy body of MIL based algorithms was developed thereafter [21], [22], [14]. In multiple instance learning, instances arrive together in groups during training, known as *bags*, each of which is assigned either a positive or a negative label (can be multi-class), but absent instance-level labels (as shown in Figure 1). In the original MIL setting [13], each bag consists of a number of organic molecules as instances; their task was to predict instance-level label for the training/test data, in addition to being able to perform bag-level classification. In our case here, each histopathology image with cancer or non-cancer label forms a bag and each pixel in the image is referred to as an instance (note that the instance features are computed based on each pixel's surroundings beyond the single pixel itself).

Despite the great success of MIL approaches [13], [14], [15] that explicitly deal with the latent (instance-level) labels, one big problem with many existing MIL algorithms is the use of pre-specified features [21], [14], [16]. Although algorithms like MILBoost [14] have embedded feature selection procedures, their input feature types are nevertheless fixed and pre-specified. To this point, it is natural to develop an integrated framework by combining the MIL concept with convolutional neural networks (CNN), which automatically learns rich hi-
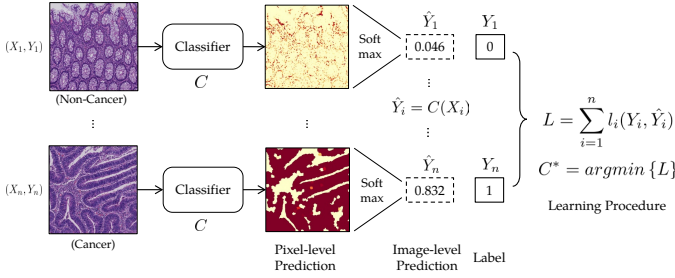
Fig. 1: Illustration of the learning procedure of a MIL algorithm. Our training dataset is denoted by $S = \{(X_i, Y_i), i = 1, 2, 3, \ldots, n\}$, where X_i indicates the ith input image, and $Y_i \in \{0, 1\}$ represents its corresponding manual label ($Y_i = 0$ refers to a non-cancer image and $Y_i = 1$ refers to a cancer image). Given an input image, a classifier C generates pixel-level predictions. Then, the image-level prediction $\widehat{Y}_i$ is computed from pixel-level predictions via a softmax function. Next, a loss between the ground truth Y_i and the image-level prediction $\widehat{Y}_i$ is computed for the ith input image, denoted by $l_i(Y_i, \widehat{Y}_i)$. Finally, an objective loss function L takes the sum of loss functions of all input images. The classifier C is learned by minimizing the objective loss function.

erarchical features for pattern recognition with state-of-the-art classification/recognition results. A previous approach that adopts CNN in a MIL formulation was recently proposed [17], but its greatest limitation is the use of image patches instead of full images, making the learning process slow and ineffective. For patch-based approaches: (1) image patch size has to be specified in advance; (2) every pixel as the center of a patch is potentially an instance, resulting in millions of patches to be extracted even for a single image; (3) feature extraction for image patches is not efficient. Beyond the patch-centric CNN framework is the image-centric paradigm where image-to-image prediction can be performed by fully convolutional networks (FCN) [23] in which features for all pixels are computed altogether. The efficiency and effectiveness of both training and testing by FCN family models have shown great success in various computer vision applications such as image labeling [23], [24] and edge detection [25]. An early version of FCN applied in MIL was proposed in [26] which was extended into a more advanced model [18].

In this paper, we first build an FCN based multiple instance learning framework to serve as our baseline algorithm for weakly-supervised learning of histopathology image segmentation. The main focus of this paper is the introduction of deep weak supervision and constraints to our multiple instance learning framework. We abbreviate our deep weak supervision for multiple instance learning as DWS-MIL and our constrained deep weak supervision for multiple instance learning as CDWS-MIL. The concept of deep supervision in the supervised learning was introduced in [27], which is combined with FCN for edge detection [25]. We propose a deep weak supervision strategy in which the intermediate FCN layers are expected to be further guided through weakly-supervised information within their own layers.

We also introduce area constraints that only require a small amount of additional labeling effort but are shown to be immensely effective. That is, in addition to the annotation of being a cancerous or non-cancerous image, we ask pathologists to give a rough estimation of the relative size (e.g 30%) of cancerous regions within each image; this rough estimation is then turned into an area constraint in our MIL formulation.

Our motivation to introduce area constrains is three-fold. First, having informative but easy to obtain expert annotation can always help the learning process and we are encouraged to seek information beyond being just positive or negative. There exists a study in cognitive science [28] indicating the natural surfacing of the concept of relative size when making a discrete yes-or-no decision. Second, our DWS-MIL formulation under an image-to-image paradigm allows the additional term of the area constraints to be conveniently carried out through back-propagation, which is nearly impossible to do if a patch-based approach is adopted [16], [17]. Third, having area constraints conceptually and mathematically greatly enhances learning capability; this is evident in our experiments where a significant performance boost is observed using the area constraints.

To summarize, in this paper we develop a new multiple instance learning algorithm for histopathology image segmentation under a deep weak supervision formulation, abbreviated as DWS-MIL. The contributions of our algorithm include: (1) DWS-MIL is an end-to-end learning system that performs image-to-image learning and prediction under weak supervision. (2) Deep weak supervision is adopted in each intermediate layer to exploit nested multi-scale feature learning. (3) Area constraints are also introduced as weak supervision, which is shown to be particularly effective in the learning process, significantly enhancing segmentation accuracy with very little extra work during the annotation process. In addition, we experiment with the adoption of super-pixels [29] as an alternative way to pixels and show their effectiveness in maintaining intrinsic tissue boundaries in histopathology images.

II. RELATED WORK

Related work can be divided into three broad categories: (1) directly related work, (2) weakly supervised learning in computer vision, and (3) weakly supervised learning in medical images.

A. Directly related work

Three existing approaches that are closely related to our work are discussed below.

Xu et al. [16] propose a histopathology image segmentation algorithm in which the concept of multiple clustered instance learning (MCIL) is introduced. The MCIL algorithm [16] can simultaneously perform image-level classification, patch-level segmentation and patch-level clustering. However, as mentioned previously, their approach is a patch-based system that is extremely space-demanding (requiring large disk space to store the features) and time-consuming to train. In addition, a boosting algorithm is adopted in [16] with all feature types pre-specified, but features in our approaches are automatically learned.

Pathak et al. present an early version of fully convolutional networks applied in a multiple instance learning setting [26] and they later generalize the algorithm by introducing a new loss function to optimize for any set of linear constraints on the output space [18]. Some typical linear constraints include

suppression, foreground, background, and size constraints. Compared with the generalized constrained optimization in their model, the area constraints proposed in this paper are simpler to carry out through back-propagation within MIL. Moreover, our formulation of deep weak supervision combined with area constraints demonstrates its particular advantage in histopathology image segmentation where only two-class (positive and negative) classification is studied.

Holistically-nested edge detector (HED) is developed in [30] by combining deep supervision with fully convolutional networks to effectively learn edges and object boundaries. Our deep weak supervision formulation is inspired by HED but we instead focus on a weakly-supervised learning setting as opposed to being fully supervised in HED. Our deep weak supervision demonstrates its power under an end-to-end MIL framework.

B. Weakly supervised learning in computer vision

A rich body of weakly-supervised learning algorithms exists in computer vision and we discuss them in two groupings: segmentation based and detection based.

Segmentation. In computer vision, MIL has been applied to segmentation in many previous systems [31], [32], [33], [34]. A patch-based approach would extract pre-specified image features from selected image patches [31], [32] and try to learn the hidden instance labeling under MIL. The limitations of these approaches are apparent, as stated before, requiring significant space and computation. More recently, convolutional neural networks have become increasingly popular. Pinheiro et al. [33] propose a convolutional neural network-based model which weights important pixels during training. Papandreous et al. [34] propose an expectation-maximization (EM) method using image-level and bounding box annotation in a weakly-supervised setting.

Object detection. MIL has also been applied to objection detection where the instances are now image patches of varying sizes, which are also referred to as sliding windows. The space for storing all instances are enormous and proposals are often used to limit the number of possible instances [35]. A lot of algorithms exist in this domain and we name a couple here. Cinbis et al. [36] propose a multi-fold multiple instance learning procedure, which prevents training from prematurely looking at all object locations; this method iteratively trains a detector and infers object locations. Diba et al. [37] propose a cascaded network structure which is composed of two or three stages and is trained in an end-to-end pipeline.

C. Weakly supervised learning in medical imaging

Weakly-supervised learning has been applied to medical images as well. Yan et al. [38] propose a multi-instance deep learning method by automatically discovering discriminative local anatomies for anatomical structure recognition; positive instances are defined as contiguous bounding boxes and negative instances (non-informative anatomy) are randomly selected from the background. A weakly-supervised learning approach is also adopted in Hou et al. [39] to train convolutional neural networks to identify gigapixel resolution histopathology images.

Though being promising, existing methods in medical imaging lack an end-to-end learning strategy for image-to-image learning and prediction under MIL.

III. METHOD

In this section, we present in detail the concept and formulation of our algorithms. First, we introduce our baseline algorithm, a method in spirit similar to the FCN-MIL method [26] but our method focuses on two-class classification whereas FCN-MIL is a multi-class approach with some preliminary results shown for natural image segmentation. We then discuss the main part of this work, deep weak supervision for MIL (DWS-MIL) and constrained deep weak supervision for MIL (CDWS-MIL).

A. Our Baseline

Here, we build an end-to-end MIL method as our baseline to perform image-to-image learning and prediction, in which the MIL formulation enables automatic learning of pixel-level segmentation from image-level labels.

We denote our training dataset by $S = \{(X_i, Y_i), i = 1, 2, 3, \ldots, n\}$, where X_i denotes the ith input image and $Y_i \in \{0, 1\}$ refers to the manual annotation (ground truth label) assigned to the ith input image. Here $Y_i = 0$ refers to a non-cancer image and $Y_i = 1$ refers to a cancerous image. Figure 1 demonstrates the basic concept. As mentioned previously, our task is to be able to perform pixel-level prediction learned from image-level labels and each pixel is referred to as an instance in this case. We denote $\widehat{Y}_{ik}$ to be the probability of the kth pixel being positive in the ith image, where $k = \{1, 2, \ldots, |X_i|\}$ and $|X_i|$ represents the total number of pixels of image X_i. If an image-level predictions $\widehat{Y}_i$ can be computed from all $\widehat{Y}_{ik}$s, then it can be used against the true image-level labels Y_i to calculate a loss $\mathcal{L}_{mil}$. The loss function we opt to use is the cross-entropy cost function:

$$\mathcal{L}_{mil} = \sum_i \left(\boldsymbol{I}(Y_i = 1) \log \widehat{Y}_i + \boldsymbol{I}(Y_i = 0) \log(1 - \widehat{Y}_i) \right),$$
(1)

where $\boldsymbol{I}(\cdot)$ is an indicator function.

Since one image is identified to be negative if and only if there does not exist any positive instances, $\widehat{Y}_i$ is typically obtained by $\widehat{Y}_i = \max_k \widehat{Y}_{ik}$, resulting in a *hard maximum* approach. However, there are two problems with the hard maximum approach: (1) It makes the derivative $\partial \widehat{Y}_i / \partial \widehat{Y}_{ik}$ discontinuous, leading to numerical instability; (2) $\partial \widehat{Y}_i / \partial \widehat{Y}_{ik}$ would be 0 for all but the maximum $\widehat{Y}_{ik}$, rendering the learner unable to consider all instances simultaneously. Therefore, a softmax function is often used to replace the hard maximum approach. We use *Generalized Mean (GM)* as our softmax function [14], which is defined as

$$\widehat{Y}_i = \left(\frac{1}{|X_i|} \sum_{k=1}^{|X_i|} \widehat{Y}_{ik}^r \right)^{1/r}.$$
(2)

The parameter r controls the sharpness and proximity to the hard function: $\widehat{Y}_i \to \max_k \widehat{Y}_{ik}$ as $r \to \infty$.

We replace classifier C in Figure 1 with a fully convolutional network (FCN) [23] using a trimmed VGGNet [40] under the MIL setting. To minimize the loss function via back propagation, we calculate $\partial\mathcal{L}_{mil}/\partial\widehat{Y}_{ik}$ from $\partial\mathcal{L}_{mil}/\partial\widehat{Y}_i$. By the chain rule of differentiation,

$$\frac{\partial\mathcal{L}_{mil}}{\partial\widehat{Y}_{ik}} = \frac{\partial\mathcal{L}_{mil}}{\partial\widehat{Y}_i}\frac{\partial\widehat{Y}_i}{\partial\widehat{Y}_{ik}}. \tag{3}$$

It suffices to know $\partial\widehat{Y}_i/\partial\widehat{Y}_{ik}$, whose analytical expression can be derived from the softmax function itself. Once $\partial\mathcal{L}_{mil}/\partial\widehat{Y}_{ik}$ is known, back propagation can be performed.

In Figure 2, a training image and its learned instance-level predictions are illustrated. Instance-level predictions are shown as a heatmap, which shows the probability of each pixel being cancerous. We use a color coding bar to illustrate the probabilities ranging between 0 and 1. Note that in the following figures, the instance-level predictions (segmentation) are all displayed as heatmaps and we no longer show the color coding bar for simplicity.

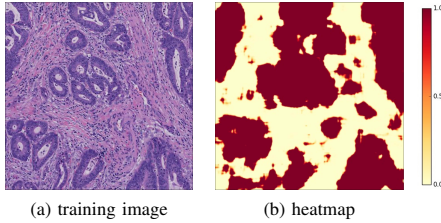(a) training image   (b) heatmap

Fig. 2: Probability map of an image for all instances. (a) Training image. (b) Instance-level probabilities (segmentation) of being positive (cancerous) by our baseline algorithm. The color coding bar indicates a probability ranging between 0 and 1.

B. Constrained Deep Weak Supervision.

After the introduction of our baseline algorithm that is an FCN-like model under MIL, we are ready to introduce the main part of our algorithm, constrained deep weak supervision for histopathology image segmentation.

We denote our training set as $S = \{(X_i, Y_i, a_i), i = 1, 2, 3, \ldots, n\}$, where X_i refers to the ith input image, $Y_i \in \{0, 1\}$ indicates the corresponding ground truth label for the ith input image, and a_i specifies a rough estimation of the relative area size of the cancerous region within image X_i. The kth pixel in the ith image is given a prediction of the probability being positive, denoted as $\widehat{Y}_{ik}$, where $k = \{1, 2, \ldots, |X_i|\}$ and there are $|X_i|$ pixels in the ith image. We denote parameters of the network as θ and the model is trained to minimize a total loss.

Deep weak supervision. Aiming to control and guide the learning process across multiple scales, we introduce deep weak supervision by producing side-outputs, forming the multiple instance learning framework with deep weak supervision, named DWS-MIL. The concept of side-output is similar to that defined in [30].

Suppose there are T side-output layers, then each side-output layer is connected with an accompanying classifier with the weights $w = (w^{(1)}, \ldots, w^{(T)})$, where $t = \{1, 2, \ldots, T\}$. Our goal is to train the model by minimizing a loss between output predictions and ground truth, which is described in the form of the cross-entropy loss function $l_{mil}^{(t)}$, indicating the loss produced by the tth side-output layer relative to image-level ground truth. The cross-entropy loss function in each side-output layer is defined as

$$l_{mil}^{(t)} = \sum_i \left(\boldsymbol{I}(Y_i = 1)\log\widehat{Y}_i^{(t)} + \boldsymbol{I}(Y_i = 0)\log(1 - \widehat{Y}_i^{(t)}) \right). \tag{4}$$

The loss function brought by the tth side-output layer is defined as :

$$l_{side}^{(t)}(\theta, w) = l_{mil}^{(t)}(\theta, w). \tag{5}$$

The objective function is defined as:

$$\mathcal{L}_{side}(\theta, w) = \sum_{t=1}^{T} l_{side}^{(t)}(\theta, w). \tag{6}$$

Deep weak supervision with constraints. Our baseline MIL formulation produces a decent result as shown in the experiments but still with room to improve. One problem is that the positive instances predicted by the algorithm tends to progressively outgrow the true cancerous regions. Here we propose to use an area constraint term to constrain the expansion of the positive instances during training and we name our new algorithm as constrained deep weak supervision, abbreviated as CDWS-MIL.

A rough estimation of the relative size of the cancerous region, a_i, is given by the experts during the annotation process. A measure of the overall "positiveness" of all the instances in each image is calculated as

$$v_i = \frac{1}{|X_i|}\sum_{k=1}^{|X_i|}\widehat{Y}_{ik}. \tag{7}$$

We then define an area constraint as an $L2$ loss:

$$l_{ac} = \boldsymbol{I}(Y_i = 1)\sum_i (v_i - a_i)^2. \tag{8}$$

Naturally the loss function for the tth side-output layer can be replaced by:

$$l_{side}^{(t)}(\theta, w) \leftarrow l_{mil}^{(t)}(\theta, w) + \eta_t \cdot l_{ac}^{(t)}(\theta, w), \tag{9}$$

where $l_{mil}^{(t)}(\theta, w)$ denotes the loss function generated in equation 4, $l_{ac}^{(t)}(\theta, w)$ is the area constraints loss, and η_t is a hyper-parameter specified manually to balance the two terms. Then, the objective loss function is still defined as the accumulation of the loss generated from each side-output layer, which is described in equation (6).

Fusion model. In order to adequately leverage the multi-scale predictions across all the layers, we merge the side-output layers with each other to generate a fusion layer. $\widehat{Y}_{side}^{(t)}$ is the predicted probability map at the tth side output layer. The output of the fusion layer is defined as

$$\widehat{Y}_{fuse} = \sum_{t=1}^{T} \alpha_t \widehat{Y}_{side}^{(t)}, \tag{10}$$

where α_t refers to the weight deduced for the probability map generated by the tth side-output layer.
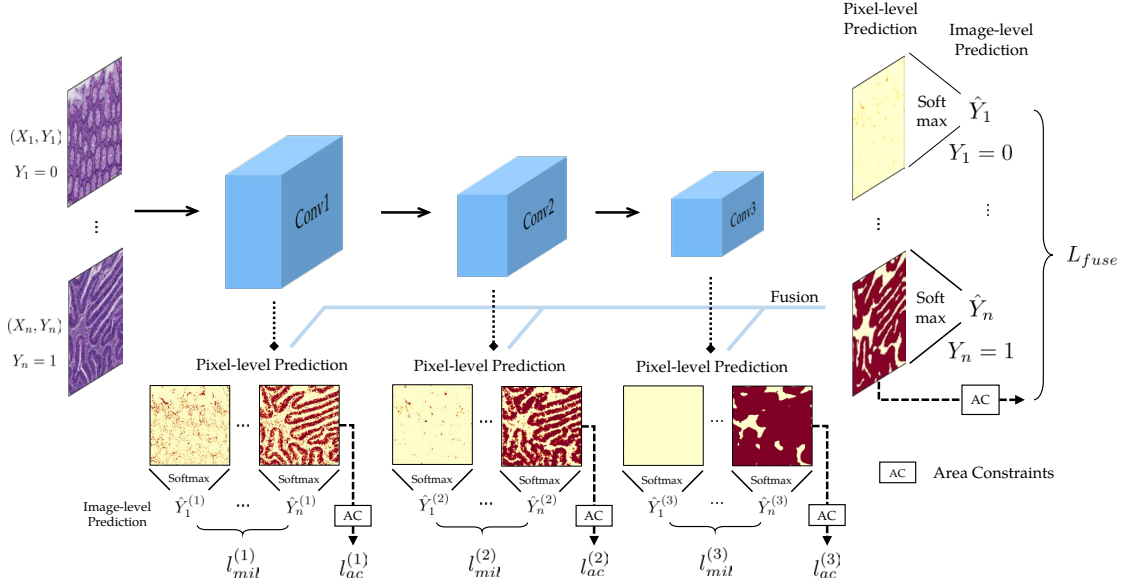
Fig. 3: Overview of our framework. Under the MIL setting, we adopt first three stages of the VGGNet and connect side-output layers with deep weak supervision under MIL. We also propose area constraints to regularize the size of predicted positive instances. To utilize the multi-scale predictions of the individual layers, we merge side-outputs via a weighted fusion layer. The overall model of equation (13) is trained via back-propagation using the stochastic gradient descent algorithm.

Then, the fusion loss function is given as:

$$\mathcal{L}_{fuse}(\theta, w) = l_{mil}^{(fuse)}(\theta, w) + \eta_{fuse} \cdot l_{ac}^{(fuse)}(\theta, w), \quad (11)$$

where $l_{mil}^{(fuse)}(\theta, w)$ is the MIL loss of $\widehat{Y}_{fuse}$ computed as equation (4), $l_{ac}^{(fuse)}(\theta, w)$ is the area constraints loss of $\widehat{Y}_{fuse}$ computed as equation (8), and η_{fuse} is a hyper-parameter specified manually to balance the two terms. The final objective loss function is defined as below:

$$\mathcal{L}(\theta, w) = \mathcal{L}_{side}(\theta, w) + \mathcal{L}_{fuse}(\theta, w). \quad (12)$$

In the end, we minimize the overall loss function by stochastic gradient descent algorithm during network training:

$$(\theta, w)^* = argmin_{\theta, w} \mathcal{L}(\theta, w). \quad (13)$$

To summarize, equation (13) gives the overall function to learn, which is under the general multiple instance learning with an end-to-end learning process. Our algorithm is built on top of fully convolutional networks with deep weak supervision and additional area constraints. The pipeline of our algorithm is illustrated in Figure 3. In our framework, we adopt the first three stages of the VGGNet and then the last convolutional layer of each stage is connected to side-output. Pixel-level prediction maps can be produced by each side-output layer and the fusion layer. The fusion layer takes a weighted average of all side-outputs. The MIL formulation guides the learning of the entire network to make pixel-level prediction for a better prediction of the image-level labels via softmax functions. In each side-output layer, the loss function l_{mil} is computed in the form of deep weak supervision. Furthermore, area constraints loss l_{ac} makes it possible to constrain the size of predicted cancerous tissues. Finally, the parameters of our network is learned by minimizing the objective function defined in equation (13) via back-propagation using the stochastic gradient descent algorithm.

C. Super-pixels

Treating each pixel as an instance may sometimes produce jagged tissue boundaries. We therefore alternatively explore another option of defining instances, super-pixels. Using super-pixels gives rise to a smaller number of instances and consistent elements that can be readily pre-computed using an over-segmentation algorithm [29]. Our effort starts with the SLIC method mentioned in [29] to generate super-pixels by grouping the input image pixels into a number of small regions. These super-pixels act as our instances but our main formulation stays the same as to minimize the overall objective function defined in equation (13).

IV. NETWORK ARCHITECTURE

We choose the 16-layer VGGNet [40] as the CNN architecture of our framework, which was pre-trained on the ImageNet 1K class dataset and achieved state-of-the-art performance in the ImageNet challenge [41]. Although ImageNet consists of natural images, which are different from histopathology images, several previous works [42] have shown that networks pre-trained on ImageNet are also very effective in dealing with histopathology images. The VGGNet has 5 sequential stages before the fully-connected layer. Within each stage, two or three convolutional layers are followed by a 2-stride pooling layer. In our framework, we trim off the 4th and 5th stages and only adopt the first three stages. Side-output layers are connected to last convolutional layer in each stage (see Table I). The side-output layer is a 1×1 convolutional layer of one-channel output with the sigmoid activation. This style of network architecture makes different side-output layers have different strides and receptive field sizes, resulting in side-outputs of different scales. Having three side-output layers, we add a fusion layer that takes a weighted average of side-outputs to yield the final output. Note due to the different

strides in different side-output layers, the sizes of different side-outputs are not same. Hence, before the fusion, all side-outputs are upsampled to the size of the input image by bilinear interpolation.

TABLE I: The receptive field size and stride in the VGGNet [40]. In our framework, the first three stages are used, and the bolded parts indicate convolutional layers linked to additional side-output layers.

layer	c1_2	c2_2	c3_3	c4_3	c5_3
rf size	**5**	**14**	**40**	92	196
stride	**1**	**2**	**4**	8	16

The reason for trimming the VGGNet. In histopathology images, tissues appear as local texture patterns. In the 4th and 5th stages of the VGGNet, the receptive field sizes (see Table I) become too large for local textures. Figure 4 shows side-outputs if all 5 stages of the VGGNet is adopted. From the figure, as the network going deeper, the receptive field size increases and the side output grows to be larger and coarser. In the 4th and 5th stages, the side-outputs almost fill the entire images, which becomes meaningless. Thus we ignore the 4th and 5th stages of the VGGNet in our framework, due to their overlarge receptive field size.

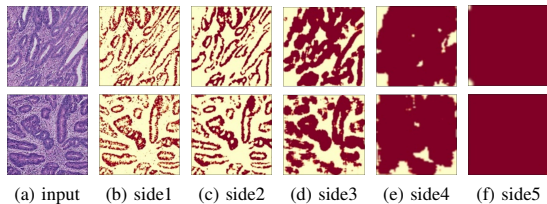(a) input   (b) side1   (c) side2   (d) side3   (e) side4   (f) side5

Fig. 4: Side-outputs from 5 stages of the VGGNet. As the network going deeper, the receptive field size increases and the side-output grows to be larger and coarser. In the 4th and 5th stages, almost all the pixels are recognized as positive, and then positive areas almost cover the entire images. Therefore, we trim off the 4th and 5th stages in our framework.

V. EXPERIMENTS

In this section, we first describe the implementation details of our framework. Two histopathology image datasets are used to evaluate our proposed methods.

A. Implementation

We implement our framework on top of the publicly available Caffe toolbox [43]. Based on the official version of Caffe, we add a layer to compute the softmax of the generalized mean for pixel-level predictions and a layer to compute the area constraints loss from pixel-level predictions.

Model parameters. The MIL loss is known to be hard to train, and special care is required for choosing training hyper-parameters. In order to reduce fluctuations in optimizing the MIL loss, all training data are used in each iteration (the mini-batch size is equal to the size of the training set). The network is trained with Adam optimizer [44], using a momentum of 0.9, a weight decay of 0.0005, and a fixed learning rate of 0.001. The learning rates of side-output layers are set to $1/100$ of the global learning rate. For the parameter of the generalized mean, we set $r = 4$.

Fusion layer. The fusion layer adopts the weighted average of side-output layers. At the first attempt, we initialize all the fusion weights to $1/3$, and let the model learn appropriate weights in the training phase. When the network converges, we observe that the outputs of the fusion layer are very close to the 3rd side-output layer, making the fusion results useless. The reason for this outcome is that for the deeper side-output layer, it has a lower MIL loss as a result of more discriminative features. To resolve the problem, we use fixed fusion weights instead of learning them. Based on cross-validation on training data, the fusion weights are finally chosen as $0.2, 0.35, 0.45$ for the three side-output layers, and a threshold of 0.5 is used to produce segmentation results.

Weight of area constraints loss. The weight of the area constraints loss is crucial for CDWS-MIL, since it directly decides the strength of constraints. Strong constraints may make the network unable to converge, while weak constraints have a little help with learning better segments. To decide the appropriate loss weight, we select a validation set from training data to evaluate different options. The loss weights of area constraints for the different side-output layers are decided separately. To achieve this, when deciding the loss weight for a side-output layer, only this layer has area constraints. Finally, loss weights of $2.5, 5, 10, 10$ are selected for the three side-output layers and the fusion layer.

B. Experiment A

Dataset A is a histopathology image dataset of colon cancer, which consists of 330 cancer (CA) and 580 non-cancer (NC) images. In this dataset, 250 cancer and 500 non-cancer images are used for training; 80 cancer and 80 non-cancer images are used for testing. These images are obtained from the NanoZoomer 2.0HT digital slide scanner produced by Hamamatsu Photonics with a magnification factor of 40, i.e. 226 nm/pixel. Each image has a resolution of $3,000 \times 3,000$. Two pathologists were asked to label each image to be cancerous or non-cancerous. When the two pathologists disagree on a particular image, they would discuss it with another senior pathologist to reach an agreement. For the evaluation purpose, we also ask the pathologists to annotate all the cancerous tissues for each image, which are only used in testing to evaluate our algorithms. For simplicity, in our table we use CA to refer to cancer images and use NC to refer to non-cancer images.

F-measure is used as the evaluation metric for experiments on *Dataset A*. Given the ground truth map G and the prediction map H, we define F-measure $= (2 \cdot \text{precision} \times \text{recall})/(\text{precision} + \text{recall})$ in which precision $= |H \cap G|/|H|$ and recall $= |H \cap G|/|G|$. For images with label $Y = 1$, the prediction map consists of pixels with 1 as the pixel-level prediction, and the ground truth map is the annotated cancerous regions. For images with label $Y = 0$, the prediction map consists of pixels with 0 as the pixel-level prediction, and the ground truth map is the entire image.

Comparisons. Table II summarizes the results of our proposed algorithms and other methods on *Dataset A*. In all experiments, images are resized to 500×500 pixels for time-efficiency. In MIL-Boosting, a patch size of 64×64 pixels and

a stride of 4 pixels are used for both training and testing, and other settings follow [16]. To show the effectiveness of area constraints, we also integrate area constraints into our baseline, denoted as "our baseline w/ AC" in the table. From the table, DWS-MIL and CDWS-MIL surpass other methods by large margins, and constrained deep weak supervision contributes an improvement of 7.3% than our baseline method (0.835 vs 0.778). Figure 9 shows some examples of segmentation results by these methods.

TABLE II: Performance of various methods on *Dataset A*.

Method	F-measure of CA	F-measure of NC
MIL-Boosting	0.684	0.997
our baseline	0.778	0.998
our baseline w/ AC	0.815	0.998
DWS-MIL	0.817	0.999
CDWS-MIL	**0.835**	0.997

Less training data. To observe how the amounts of training data influence our baseline method, we train our baseline with less training data. Table III summarizes the results, and Figure 5 shows some samples of segmentation results that use different amounts of training data. Given more training data, the performance of segmentation is better. In the case of less training data, the segmentation results tend to be larger than the ground truth. This observation can be explained by analyzing the MIL formulation. From the expression of the MIL loss, identifying more pixels as positive in a positive image always results in a lower MIL loss. With a smaller amount of negative training images, it is easier to achieve this objective.

TABLE III: Performance of our baseline trained with less training data.

Training data	F-measure of CA		F-measure of NC	
(Pos,Neg)	w/o AC	w/ AC	w/o AC	w/ AC
20% (50,100)	0.758	0.801	0.997	0.997
40% (100,200)	0.762	0.809	0.997	0.999
60% (150,300)	0.778	0.805	0.999	0.999
80% (200,400)	0.779	0.813	0.998	0.999
100% (250,500)	0.778	0.815	0.998	0.998

(a) input   (b) gt   (c) 20%   (d) 40%   (e) 60%   (f) 80%   (g) 100%
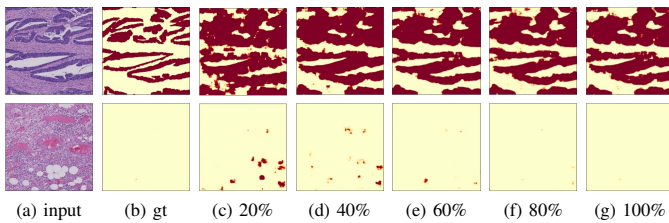
Fig. 5: Differences in results with different amounts of training data: (a) The input images. (b) Ground truth labels. (c) Results that use 20% of training data. (d) Results that use 40% of training data. (e) Results that use 60% of training data. (f) Results that use 80% of training data. (g) Results that use all the training data.

Area constraints. From Table III, the area constraints enable our baseline method to achieve a competitive accuracy with a small training set. Equipped with area constraints, our baseline method using 20% of training data achieves better accuracy than using all training data without area constraints. Figure 6 shows some samples of segmentation results by using

and not using area constraints. It is clear that area constraints achieve the goal of constraining the model to learn smaller segmentations, which significantly improves segmentation accuracy for both cancer images and non-cancer images. When not using area constraints, the segmentation results are much larger than the ground truth, and also have the tendency to cover entire images. In contrast, when the area constraints loss is integrated with the MIL loss, the fact that too many pixels are identified as positive will yield a large area constraint loss to compete with the MIL loss. To achieve a balance between the MIL loss and the area constraints loss, it only learns the most confident pixels as positive, as proven in Figure 6. Table IV summarizes results of the baseline methods, DWS-MIL, CDWS-MIL and MIL-Boosting using 20% of training data. Comparing CDWS-MIL in Table IV with other methods in Table II, CDWS-MIL outperforms other methods using only 20% of training data. In addition, constrained deep weak supervision contributes an improvement of 8.2% over our baseline method (0.820 vs 0.758), which is larger than the condition that all training data are used.

TABLE IV: Performance of various methods with 20% training data.

Method	F-measure of CA	F-measure of NC
MIL-Boosting	0.635	0.995
our baseline	0.758	0.997
our baseline w/ AC	0.800	0.997
DWS-MIL	0.808	0.998
CDWS-MIL	**0.820**	0.998

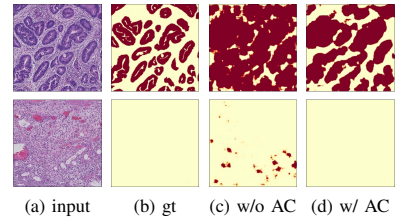(a) input   (b) gt   (c) w/o AC   (d) w/ AC

Fig. 6: Comparison of using and not using area constraints: (a) The input images. (b) Ground truth labels. (c) Results of our baseline. (d) Results of our baseline w/ AC. The area constraints loss constrains the model to learn better segmentations.

Deep weak supervision. To illustrate the effectiveness of deep weak supervision, Table V summarizes segmentation accuracies of the different side-outputs and Figure 7 shows some examples of the different side-outputs. From Table V, we observe that segmentation accuracy improves from lower layers to higher ones. Figure 7 shows pixel-level predictions (segmentation) of side-output layer 1, side-output layer 2, and side-output layer 3. This is understandable since the receptive fields of CNN become increasingly bigger from lower layers to higher ones. Histopathology images typically observe local texture patterns. The final fusion layer that combines all the intermediate layers achieves the best result.

Super-pixels. We conduct experiments to compare DSW-MIL and DSW-MIL w/ super-pixel. We adopt the SLIC method [29] to generate super-pixels. The average F-measures of DSW-MIL w/ super-pixel on cancer images and non-cancer images are 0.818 and 0.999, respectively. Figure 8 shows

TABLE V: Performance of different side-output layers. The first line: DWS-MIL; The second line: CDWS-MIL.

F-measure of CA				F-measure of NC			
side1	side2	side3	fusion	side1	side2	side3	fusion
0.666	0.747	0.783	0.817	0.984	0.994	0.997	0.999
0.660	0.783	0.819	0.835	0.984	0.994	0.997	0.997

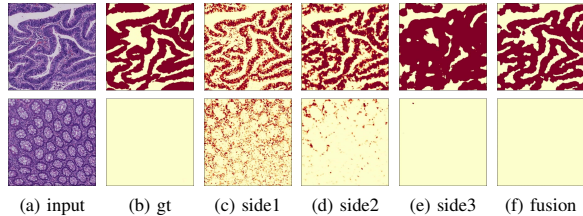(a) input   (b) gt   (c) side1   (d) side2   (e) side3   (f) fusion

Fig. 7: Results of side-output layers: (a) The input images. (b) Ground truth labels. (c) Results of side-output 1. (d) Results of side-output 2. (e) Results of side-output 3. (f) Results by final fusion. The figure shows a nesting characteristic of segmentation outputs from the lower side-output layer to the higher side-output layer. The final fusion balances pros and cons of different side outputs, and achieves better segmentation results than all of them.

some samples of the segmentation results of the two methods. In histopathology images, super-pixels adhere well to tissue edges, resulting in more accurate segmentations. The adoption of super-pixels can help to predict more detailed boundaries.
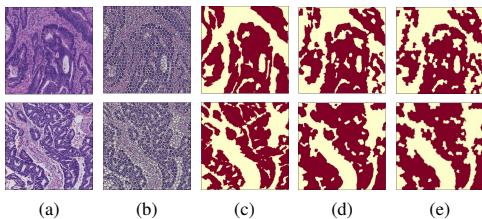

(a)   (b)   (c)   (d)   (e)

Fig. 8: Comparisons of DWS-MIL and DWS-MIL w/ super-pixel: (a) The input images. (b) Results generated by SLIC method [29]. (c) Ground truth labels. (d) Results of DWS-MIL. (e) Results of DWS-MIL w/ super-pixel. Some detailed edges can be recognized with the help of super-pixels.

Advantages of CDWS-MIL. MIL-Boosting in comparison is a patch-based MIL approach. The bags in their MIL formulation are composed of patches sampled from input images. Figure 9 shows some samples of segmentation results of CDWS-MIL and MIL-Boosting, demonstrating that in some cases (like the 2nd row in the figure), MIL-Boosting completely fails to learn the correct segmentations, while in other cases (like the 5th row in the figure), CDWS-MIL and MIL-Boosting both learn roughly correct segmentations, but CDWS-MIL learns much more elaborate ones. There are three advantages of our framework CDWS-MIL over MIL-Boosting: (1) CDWS-MIL is an end-to-end segmentation framework, which can learn more detailed segmentations than the patch-based MIL-Boosting; (2) Deep weak supervision enables CDWS-MIL to learn from multiple scales, and the fusion output balances outputs of different scales to achieve the best accuracy; (3) Area constraints in CDWS-MIL are straightforward, while being hard to be integrated into patch-based methods like MIL-Boosting.

C. Experiment B

Dataset B is a histopathology image dataset of 30 colon cancer images and 30 non-cancer images which are referred as tissue microarrays (TMAs). The dataset is randomly selected from the dataset in [16]. All images have a resolution of 1024×1024 pixels, and the rough estimations of the portion of cancerous regions have 8 levels $0.05, 0.1, 0.15, \ldots, 0.4$. They are annotated in the same way as *Dataset A*.

We conduct experiments to compare MIL-Boosting with our proposed method CDWS-MIL on *Dataset B*. All experiments are conducted with 5-fold cross-validation, and the evaluation metric is the same on *Dataset A*. The average F-measures of CDWS-MIL on cancer images and non-cancer images are 0.622 and 0.997, respectively. The average F-measures of MIL-Boosting on cancer images and non-cancer images are 0.449 and 0.993, respectively. Figure 10 shows some samples of the segmentation results of these two methods.

VI. CONCLUSION

In this paper, we have developed an end-to-end framework under deep weak supervision to perform image-to-image segmentation for histopathology images. To preferably learn multi-scale information, deep weak supervision is developed in our formulation. Area constraints are also introduced in a natural way to seek for additional weakly-supervised information. Experiments demonstrates that our methods attain the state-of-the-art results on large-scale challenging histopathology images. The scope of our proposed methods are quite broad and they can be widely applied to a range of medical imaging and computer vision applications.

REFERENCES

[1] A. Esgiar, R. Naguib, B. Sharif, M. Bennett, and A. Murray, "Fractal analysis in the detection of colonic cancer images," *IEEE Transactions on Information Technology in Biomedicine*, vol. 6, no. 1, pp. 54–58, 2002.

[2] P.-W. Huang and C.-H. Lee, "Automatic classification for pathological prostate images based on fractal analysis," *TMI*, vol. 28, no. 7, pp. 1037–1050, 2009.

[3] A. Madabhushi, "Digital pathology image analysis: opportunities and challenges," *Imaging in Medicine*, vol. 1, no. 1, pp. 7–10, 2009.

[4] S. Park, D. Sargent, R. Lieberman, and U. Gustafsson, "Domain-specific image analysis for cervical neoplasia detection based on conditional random fields," *TMI*, vol. 30, no. 3, pp. 867 –78, 2011.

[5] A. Tabesh, M. Teverovskiy, H.-Y. Pang, V. Kumar, D. Verbel, A. Kotsianti, and O. Saidi, "Multifeature prostate cancer diagnosis and gleason grading of histological images," *TMI*, vol. 26, no. 10, pp. 1366–78, 2007.

[6] A. Madabhushi and G. Lee, "Image analysis and machine learning in digital pathology: Challenges and opportunities," *Medical Image Analysis*, vol. 33, no. 6, pp. 170–175, 2016.

[7] M. N. Gurcan, L. E. Boucheron, A. Can, A. Madabhushi, N. M. Rajpoot, and B. Yener, "Histopathological image analysis: A review," *IEEE Reviews in Biomedical Engineering*, vol. 2, no. 25, pp. 147–171, 2009.

[8] J. Tang, R. M. Rangayyan, J. Xu, I. El Naqa, and Y. Yang, "Computer-aided detection and diagnosis of breast cancer with mammography: recent advances," *IEEE Transactions on Information Technology in Biomedicine*, vol. 13, no. 2, pp. 236–251, 2009.
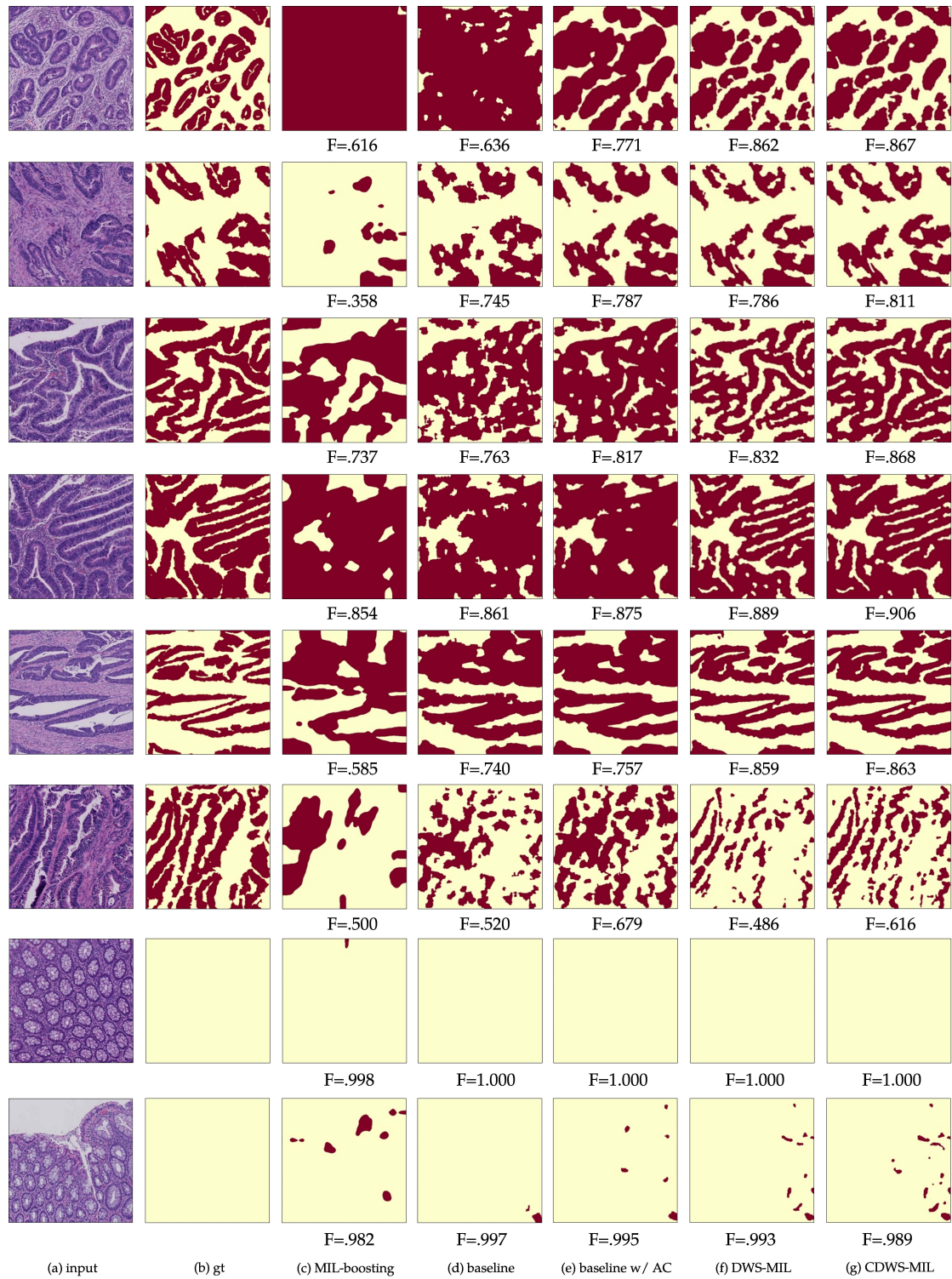
Fig. 9: Segmentation results on *dataset A*: (a) Input images. (b) Ground truth labels. (c) Results by MIL-Boosting. (d) Results by our baseline. (e) Results by our baseline w/ AC. (f) Results by DWS-MIL. (g) Results by CDWS-MIL. Compared with MIL-Boosting (patch-based), our proposed DWS-MIL and CDWS-MIL produce significantly improved results due to the characteristics we introduced in this paper.

F=.507 F=.774

F=.097 F=.702

F=.641 F=.860

F=.435 F=.752

F=.341 F=.499

F=.000 F=.310

F=1.000 F=1.000

F=.981 F=.995

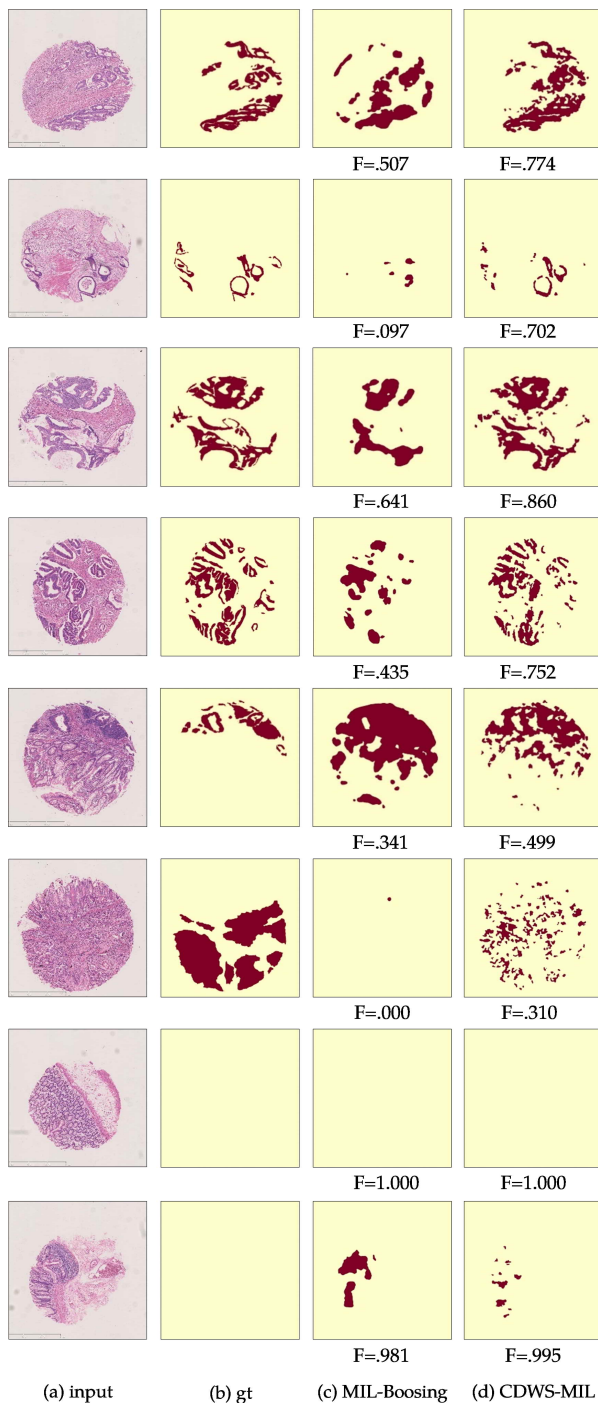(a) input   (b) gt   (c) MIL-Boosting   (d) CDWS-MIL

Fig. 10: Segmentation results on *dataset B*: (a) Input images. (b) Ground truth labels. (c) Results by MIL-Boosting. (d) Results by CDWS-MIL. Compared with MIL-Boosting (patch-based), CDWS-MIL produces significantly improved results due to the characteristics we introduced in this paper.

  [9] J. Kong, O. Sertel, H. Shimada, K. L. Boyer, J. H. Saltz, and M. N. Gurcan, "Computer-aided evaluation of neuroblastoma on whole-slide histology images: Classifying grade of neuroblastic differentiation," *Pattern Recognition*, vol. 42, no. 6, pp. 1080–1092, 2009.

[10] Y. Zheng, A. Barbu, B. Georgescu, M. Scheuering, and D. Comaniciu, "Four-chamber heart modeling and automatic segmentation for 3-d cardiac ct volumes using marginal space learning and steerable features," *TMI*, vol. 27, no. 11, pp. 1668–1681, 2008.

[11] Z. Tu, "Auto-context and its application to high-level vision tasks," in *CVPR*, 2008, pp. 1–8.

[12] A. Criminisi and J. Shotton, *Decision forests for computer vision and medical image analysis*. Springer Science & Business Media, 2013.

[13] T. G. Dietterich, R. H. Lathrop, and T. Lozano-Pérez, "Solving the multiple instance problem with axis-parallel rectangles," *Artificial Intelligence*, vol. 89, no. 1, pp. 31–71, 1997.

[14] C. Zhang, J. C. Platt, and P. A. Viola, "Multiple instance boosting for object detection," in *NIPS*, 2005, pp. 1417–1424.

[15] O. Maron and A. L. Ratan, "Multiple-instance learning for natural scene classification," in *ICML*, 1998, pp. 341–349.

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

[17] Y. Xu, T. Mo, Q. Feng, P. Zhong, M. Lai, and E. I.-C. Chang, "Deep learning of feature representation with multiple instance learning for medical image analysis," in *ICASSP*, 2014, pp. 1626–1630.

[18] D. Pathak, P. Krahenbuhl, and T. Darrell, "Constrained convolutional neural networks for weakly supervised segmentation," in *ICCV*, 2015, pp. 1796–1804.

[19] Y. Xu, J. Zhu, E. Chang, and Z. Tu, "Multiple clustered instance learning for histopathology cancer image classification, segmentation and clustering," in *CVPR*, 2012, pp. 964–971.

[20] Y. Xu, J. Zhang, E. Chang, M. Lai, and Z. Tu, "Contexts-constrained multiple instance learning for histopathology image analysis," in *MICCAI*, 2012, pp. 623–630.

[21] S. Andrews, I. Tsochantaridis, and T. Hofmann, "Support vector machines for multiple-instance learning," in *NIPS*, 2002, pp. 561–568.

[22] O. Maron and T. Lozano-Pérez, "A framework for multiple-instance learning," in *NIPS*, 1998, pp. 570–576.

[23] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *CVPR*, 2015, pp. 3431–3440.

[24] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected crfs," in *ICLR*, 2015.

[25] Y. Xie, X. Kong, F. Xing, F. Liu, H. Su, and L. Yang, "Deep voting: A robust approach toward nucleus localization in microscopy images," in *MICCAI*, 2015.

[26] D. Pathak, E. Shelhamer, J. Long, and T. Darrell, "Fully convolutional multi-class multiple instance learning," in *ICLR*, 2014.

[27] C.-Y. Lee, S. Xie, P. Gallagher, Z. Zhang, and Z. Tu, "Deeply-supervised nets," in *AISTATS*, 2015.

[28] N. S. F. USA, "Converging technologies for improving human performance," *Annals of the New York Academy of Sciences*, vol. 1013, pp. 223–226, 2004.

[29] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "Slic superpixels compared to state-of-the-art superpixel methods," *TPAMI*, vol. 34, no. 11, pp. 2274–2282, 2012.

[30] S. Xie and Z. Tu, "Holistically-nested edge detection," in *ICCV*, 2015, pp. 1395–1403.

[31] Q. Li, J. Wu, and Z. Tu, "Harvesting mid-level visual concepts from large-scale internet images," in *CVPR*, 2013, pp. 851–858.

[32] X. Wang, B. Wang, X. Bai, W. Liu, and Z. Tu, "Max-margin multiple instance dictionary learning," in *ICML*, 2013, pp. 846–854.

[33] P. O. Pinheiro and R. Collobert, "From image-level to pixel-level labeling with convolutional networks," in *CVPR*, 2015, pp. 1713–1721.

[34] G. Papandreou, L.-C. Chen, K. P. Murphy, and A. L. Yuille, "Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation," in *ICCV*, 2015, pp. 1742–1750.

[35] J.-Y. Zhu, J. Wu, Y. Xu, E. Chang, and Z. Tu, "Unsupervised object class discovery via saliency-guided multiple class learning," *TPAMI*, vol. 37, no. 4, pp. 862–875, 2015.

[36] R. G. Cinbis, J. Verbeek, and C. Schmid, "Weakly supervised object localization with multi-fold multiple instance learning," *TPAMI*, vol. 39, no. 1, pp. 189–203, 2016.

[37] A. Diba, V. Sharma, A. Pazandeh, H. Pirsiavash, and L. Van Gool, "Weakly supervised cascaded convolutional networks," *arXiv preprint arXiv:1611.08258*, 2016.

[38] Z. Yan, Y. Zhan, Z. Peng, S. Liao, Y. Shinagawa, S. Zhang, D. N. Metaxas, and X. S. Zhou, "Multi-instance deep learning: Discover discriminative local anatomies for bodypart recognition," *TMI*, vol. 35, no. 5, pp. 1332–1343, 2016.

[39] L. Hou, D. Samaras, T. M. Kurc, Y. Gao, J. E. Davis, and J. H. Saltz, "Efficient multiple instance convolutional neural networks for gigapixel resolution image classification," *arXiv preprint arXiv:1504.07947*, 2015.

[40] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *ICLR*, 2015.

[41] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "Imagenet large scale visual recognition challenge," 2014.

[42] Y. Xu, Z. Jia, Y. Ai, F. Zhang, M. Lai, and E. I.-C. Chang, "Deep convolutional activation features for large scale brain tumor histopathology image classification and segmentation," in *ICASSP*, 2015, pp. 947–951.

[43] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," *arXiv preprint arXiv:1408.5093*, 2014.

[44] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *ICLR*, 2015.