Constrained Deep Weak Supervision for Histopathology Image Segmentation

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

***Abstract*— In this paper, we develop a new weakly supervised learning algorithm to learn to segment cancerous regions in histopathology images. This paper is under a multiple instance learning (MIL) framework 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 (FCNs) in which image-to-image weakly-supervised learning is performed; 2) we develop a DWS formulation to exploit multi-scale learning under weak supervision within FCNs; and 3) constraints about positive instances are introduced in our approach to effectively explore additional weakly supervised information that is easy to obtain and enjoy 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 the state-of-the-art results on large-scale histopathology image data sets 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]–[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]–[15] have shown their particular effectiveness in performing image classification and segmentation for modalities such 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 [16]–[23]. In pathology, a pathologist annotates whether a given histopathology image has 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 [16], which is the main topic of this paper.

Multiple instance learning (MIL) was first introduced by Dietterich *et al.* [16] to predict drug activities; a wealthy body of MIL based algorithms was developed thereafter [17], [24], [25]. 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 instance-level labels are absent (as shown in Figure 1). In the original MIL setting [16], 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

Manuscript received May 17, 2017; revised July 4, 2017; accepted July 4, 2017. Date of publication July 7, 2017; date of current version October 25, 2017. This work was supported in part by Microsoft Research through the eHealth Program, in part by the Beijing National Science Foundation in China, under Grant 4152033, and in part by the Technology and Innovation Commission of Shenzhen in China under Grant shenfagai2016-627. *(Corresponding author: Yan Xu.)*

Z. Jia is with the Institute for Interdisciplinary Information Sciences, Tsinghua University, Beijing 100084, China, and also with Microsoft Research, Beijing 100080, China (e-mail: zhipeng.jia@outlook.com).

X. Huang is with the State Key Laboratory of Software Development Environment, Key Laboratory of Biomechanics and Mechanobiology, Ministry of Education, Research Institute of Beihang University in Shenzhen, Beihang University, Beijing 100191, China (e-mail: huangxingyi102@126.com).

E. I.-C. Chang is with Microsoft Research, Beijing 100080, China (e-mail: echang@microsoft.com).

Y. Xu is with the State Key Laboratory of Software Development Environment, Key Laboratory of Biomechanics and Mechanobiology, Ministry of Education, Research Institute of Beihang University in Shenzhen, Beihang University, Beijing 100191, China, and also with Microsoft Research, Beijing 100080, China (e-mail: xuyan04@gmail.com).

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

Digital Object Identifier 10.1109/TMI.2017.2724070

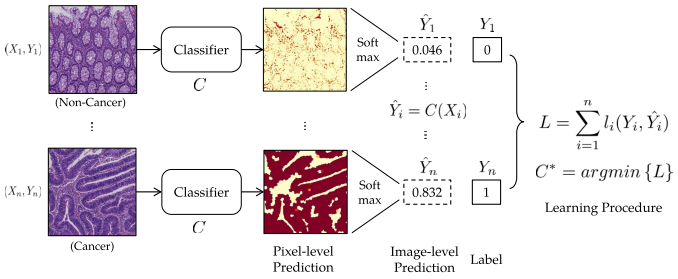


Fig. 1. Illustration of the learning procedure of a MIL algorithm. Our training dataset is denoted by $S = \{(X_i, Y_i), i = 1, 2, 3, \ldots, n\}$, where X_i indicates the ith input image, and $Y_i \in \{0, 1\}$ represents its corresponding manual label ($Y_i = 0$ refers to a non-cancer image and $Y_i = 1$ 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.

based on each pixel's surroundings beyond the single pixel itself).

Despite the great success of MIL approaches [16]–[18] that explicitly deal with the latent (instance-level) labels, one big problem with many existing MIL algorithms is the use of pre-specified features [17], [19], [24]. Although algorithms like MILBoost [17] 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 hierarchical 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 [20], 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) [26] 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 [26], [27] and edge detection [28]. An early version of FCN applied in MIL was proposed in [29] which was extended into a more advanced model [21].

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 [30], which is combined with FCN for edge detection [28]. 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 [31] 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 [19], [20]. 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 [32] 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.* [19] propose a histopathology image segmentation algorithm in which the concept of multiple clustered instance learning (MCIL) is introduced. The MCIL algorithm [19] 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 [19] 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 [29] and they later generalize the algorithm by introducing a new loss function to optimize for any set of linear constraints on the output space [21]. 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 [33] 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.

1) Segmentation: In computer vision, MIL has been applied to segmentation in many previous systems [34]–[37]. A patch-based approach would extract pre-specified image features from selected image patches [34], [35] 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 and Collobert [36] propose a convolutional neural network-based model which weights important pixels during training. Papandreou *et al.* [37] propose an expectation-maximization (EM) method using image-level and bounding box annotation in a weakly-supervised setting.

2) 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 [38]. A lot of algorithms exist in this domain and we name a couple here. Cinbis *et al.* [39] 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.* [40] 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.* [41] 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.* [42] to train convolutional neural networks to identify gigapixel resolution histopathology images.

Though 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 [29] 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). The flowchart of our algorithm is illustrated in Figure 3.

A. Our Baseline

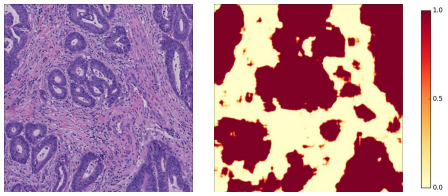
Here, we build an end-to-end MIL method as our baseline to perform image-to-image learning and prediction, in which the MIL formulation enables automatic learning of pixel-level segmentation from image-level labels.

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 $\widehat{Y}_{ik}$ as 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 prediction $\widehat{Y}_i$ 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:

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

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

Since one image is identified as negative if and only if there do 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 this 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



(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 between 0 and 1.

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 [17], which is defined as

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

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) [26] using a trimmed VGGNet [43] 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.

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.

1) 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, called DWS-MIL. The concept of side-output is similar to that which is defined in [33].

Supposing there are T side-output layers, each side-output layer is connected with an accompanying classifier with weights $w = (w^{(1)}, \ldots, w^{(T)})$, where $t = \{1, 2, \ldots, T\}$. The output probability map of the t-th side-output layer is denoted as $\widehat{Y}_i^{(t)}$. 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 t-th side-output layer relative to image-level ground truth. The cross-entropy loss function in each side-output layer is defined as

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

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

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

The objective function is defined as:

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

2) Deep Weak Supervision With Constraints: Our baseline MIL formulation produces a decent result as shown in the experiments but still with room for improvement. One problem is that positive instances predicted by the algorithm tend to progressively outgrow true cancerous regions. Here we propose using 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 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

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

which is a soft measure with the merit of being continuous and differentiable. We then define an area constraint as an $L2$ loss:

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

The loss on the area term will be activated only when $Y_i = 1$ *and* $v_i > a_i$. Therefore, for cancerous images where $v_i < a_i$ or non-cancerous images, no penalty will be introduced.

Naturally the loss function for the t-th side-output layer can be updated from Equation (5) to:

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

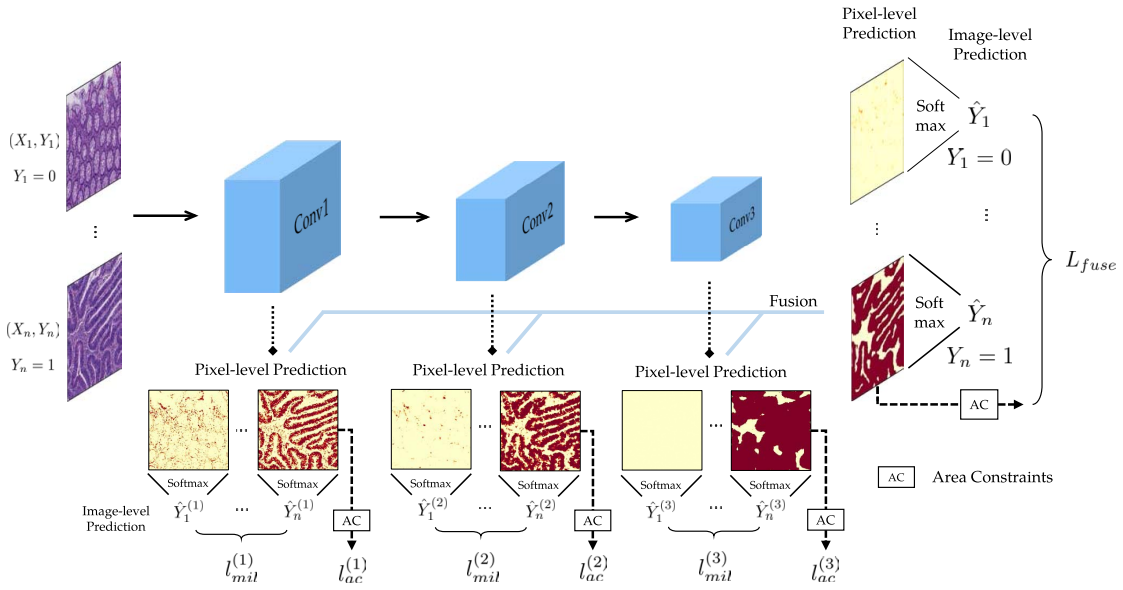


Fig. 3. Overview of our framework. Under the MIL setting, we adopt first three stages of the VGGNet and connect side-output layers with deep weak supervision under MIL. We also propose area constraints to regularize the size of predicted positive instances. To utilize the multi-scale predictions of individual 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.

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).

3) Fusion Model: In order to adequately leverage the multi-scale predictions across all layers, we merge the side-output layers with each other to generate a fusion layer. The output of the fusion layer is defined as

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

where α_t refers to the weight learned for the probability map generated by the t-th side-output layer. The fusion layer adopts the weighted average of side-output layers. In the training phase, considering the three-stage network architecture, we initialize all fusion weights to the average value $1/3$, and let the model learn appropriate weights. 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. Thus, in the test phase, we adopt a strategy, using fixed weights, to preserve multi-scale information. The appropriate weights for the test phase are decided based on the cross-validation of training data. For the three side-output layers, the chosen fusion weights are 0.2, 0.35 and 0.45.

The fusion loss function is then given as:

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

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:

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

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

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

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 constraint loss l_{ac} makes it possible to constrain the size of predicted cancerous tissues. Finally, the parameters of our network are 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 for defining instances, super-pixels.

TABLE I
THE RECEPTIVE FIELD SIZE AND STRIDE IN THE VGGNET [43]. IN OUR FRAMEWORK, THE FIRST THREE STAGES ARE USED. THE BOLD 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

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 [32]. A number of super-pixel based approaches for medical image segmentation have been previously proposed. Yu *et al.* [44] develop an automatic tumor segmentation framework, in which a simple linear iterative cluster (SLIC) algorithm is utilized to aggregate nearby pixels into super-pixels. Soltaninejad *et al.* [45] present a fully automated approach for brain tumor detection and segmentation, where super-pixels using the SLIC algorithm are obtained for tissue segmentation.

In this case, each super-pixel, instead of the pixel, is an instance, which has shown advantages in computer vision to reduce the computational complexity and retain sharp segmentation boundaries. In our paper, we use SLIC method to generate a number of super-pixels (atomic regions). At the last sigmoid layer of the neural network, pixel-level probabilities are produced. The probability for each super-pixel is the average of all the pixels in the super-pixel. 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 [43] as the CNN architecture of our framework, which is pre-trained on the ImageNet 1K class dataset and has achieved state-of-the-art performance in the ImageNet challenge [46]. Although ImageNet consists of natural images, which are different from histopathology images, several previous works [47]–[49] 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. Due to the different strides in different side-output layers, the sizes of different side-outputs are not the same. Hence, before the fusion, all side-outputs are upsampled to the size of the input image by bilinear interpolation.

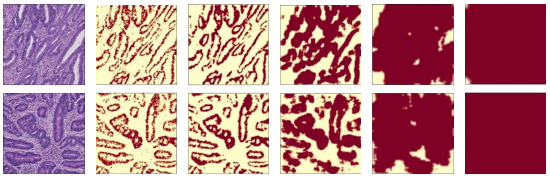


Fig. 4. Side-outputs from 5 stages of the VGGNet. As the network goes deeper, the receptive field size increases and the side-output grows larger and coarser. In the 4th and 5th stages, almost all pixels are recognized as positive, and then positive areas almost cover the entire images. Therefore, we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

1) 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 are adopted. As is shown in the figure, as the network goes 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.

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 [50]. 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. All experiments are conducted on Tesla K40 with 12G Memory.

1) 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 [51], 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 $r = 4$.

2) 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, a five-fold cross-validation is conducted in the experiments. The loss weights of area constraints for the different side-output layers are decided separately. Weights η_t of $2.5, 5, 10, 10$ are therefore selected for the three side-output layers and the fusion layer.

B. Experiment A

1) Dataset: *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 $3,000 \times 3,000$. Due to memory limits, the original $3,000 \times 3,000$ pixels can not be loaded directly. Thus, in all experiments, images are resized to 500×500 pixels. For simplicity, we use CA to refer to cancer images and use NC to refer to non-cancer images.

2) Annotations: During training, two image-level annotations for each image are given by the human pathologists, indicating each image as cancerous or not and a rough estimation ({0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, or 1.0}) about the cancerous region proportion to the entire image. For the evaluation purpose, the general annotation procedure can be summarized as follows: pixel-level annotations for each pixel being a cancerous or non-cancerous are provided by two pathologists. (1) if the overlap is larger than 80% for the two cancerous regions labeled by the two pathologists, we take the intersection of the two regions as the ground truth; (2) if the overlap is smaller than 80%, or if a cancerous region is annotated by one pathologist but neglected by another one, a third senior pathologist will step in to help decide whether to consider this region as a cancerous or not.

3) Evaluations: Both the best F-measure for boundaries using a fixed scale (ODS) [52] and the F-measure for regions, are used as evaluation metrics. We follow the definition in [52] for the ODS evaluation. All of the NC images have no boundaries. Thus, we just list the ODS results for all of the CA images in the following tables. F-measure for regions is defined as: given the ground truth map G and the prediction map H, 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.

4) Comparisons With Weakly Supervised Algorithms: Experiments have been conducted to compare the performance of our methods with some other weakly supervised algorithms. 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 [19]. 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 Table II, both F-measure and ODS evaluation results lead us to the conclusion that DWS-MIL and CDWS-MIL surpass other weak supervised methods by margins. Constrained deep weak supervision contributes an improvement of 7.3% in F-measure than our baseline method (0.835 vs 0.778) and 21.4% in ODS (0.559 vs 0.345). Figure 5 shows some examples of segmentation results by these methods.

TABLE II

PERFORMANCE OF VARIOUS METHODS ON *Dataset A*. ALL EXPERIMENTS ARE CONDUCTED ON GPU EXCEPT THAT MIL-BOOSTING AND SILC FOR GENERATING SUPER-PIXELS ARE PERFORMED ON CPU. SP IS THE ABBREVIATION FOR SUPER-PIXELS GENERATED BY SILC. THE RUNNING TIME OF THESE ALGORITHMS ARE OBTAINED BY AVERAGING MULTIPLE RUNS TO REMOVE OTHER FACTORS THAT AFFECT THE MEASURE OF RUNNING TIME SPEED. VARIATIONS OF VARIOUS WEAKLY SUPERVISED METHODS PROPOSED IN THE PAPER DIFFERENT MOSTLY IN THE TRAINING BUT THEY MOSTLY SHARE THE SAME ARCHITECTURE IN TESTING; THEIR RUN TIME SPEEDS ARE THEREFORE APPROXIMATELY THE SAME

Method	Running time (s)	ODS (boundaries)	F-measure (regions)	
		CA	CA	NC
weakly supervised:				
MIL-Boosting	2.02	0.285	0.684	0.997
baseline	0.12	0.345	0.778	0.998
baseline w/ AC	0.12	0.383	0.815	0.998
DWS-MIL	0.12	0.541	0.817	0.999
CDWS-MIL	0.12	**0.559**	**0.835**	0.997
baseline w/ SP	0.16	0.371	0.782	0.999
baseline w/ AC+SP	0.16	0.407	0.817	0.998
DWS-MIL w/ SP	0.16	0.541	0.818	0.999
CDWS-MIL w/ SP	0.16	**0.566**	**0.836**	0.999
fully supervised:				
FCN-32s	0.22	0.424	0.834	0.995
FCN-16s	0.23	0.638	0.875	0.997
FCN-8s	0.25	0.642	0.876	0.997
U-Net [14]	0.37	0.523	0.808	0.962
DCAN [13]	0.25	0.644	0.879	0.997

5) Comparisons With Fully Supervised Algorithms: Experiments have been also conducted to compare the performance of our methods against the best performers, the full supervised algorithms. Here, we adopt FCN [26] including standard FCN-32s, FCN-16s and FCN-8s, U-Net [14] and DCAN [13] for comparison. FCN-8s learns to fuse coarse, high layer information with fine, low information, combining predictions from both the final layer and the last two pooling layers. Similarly, FCN-16s fuses predictions from the last pooling layer and the final layer. For FCN, we finetune the publicly available VGG-16 model using SGD. The learning rate is set to 0.0001. For the U-Net algorithm, we follow the publicly available network and code. The learning rate is set to 0.0001. For the DCAN algorithm, we merely leverage the multi-level contextual features with auxiliary supervision for our segmentation task. The learning rate is set to 0.001.

From table II, there is a lingering gap between the performance achievable with full supervised algorithms and that achieved by existing weakly supervised algorithms. By proposing the constrained deep weak supervision for multiple instance learning, our approaches surpass other weakly supervised methods by a significant margin, reaching near

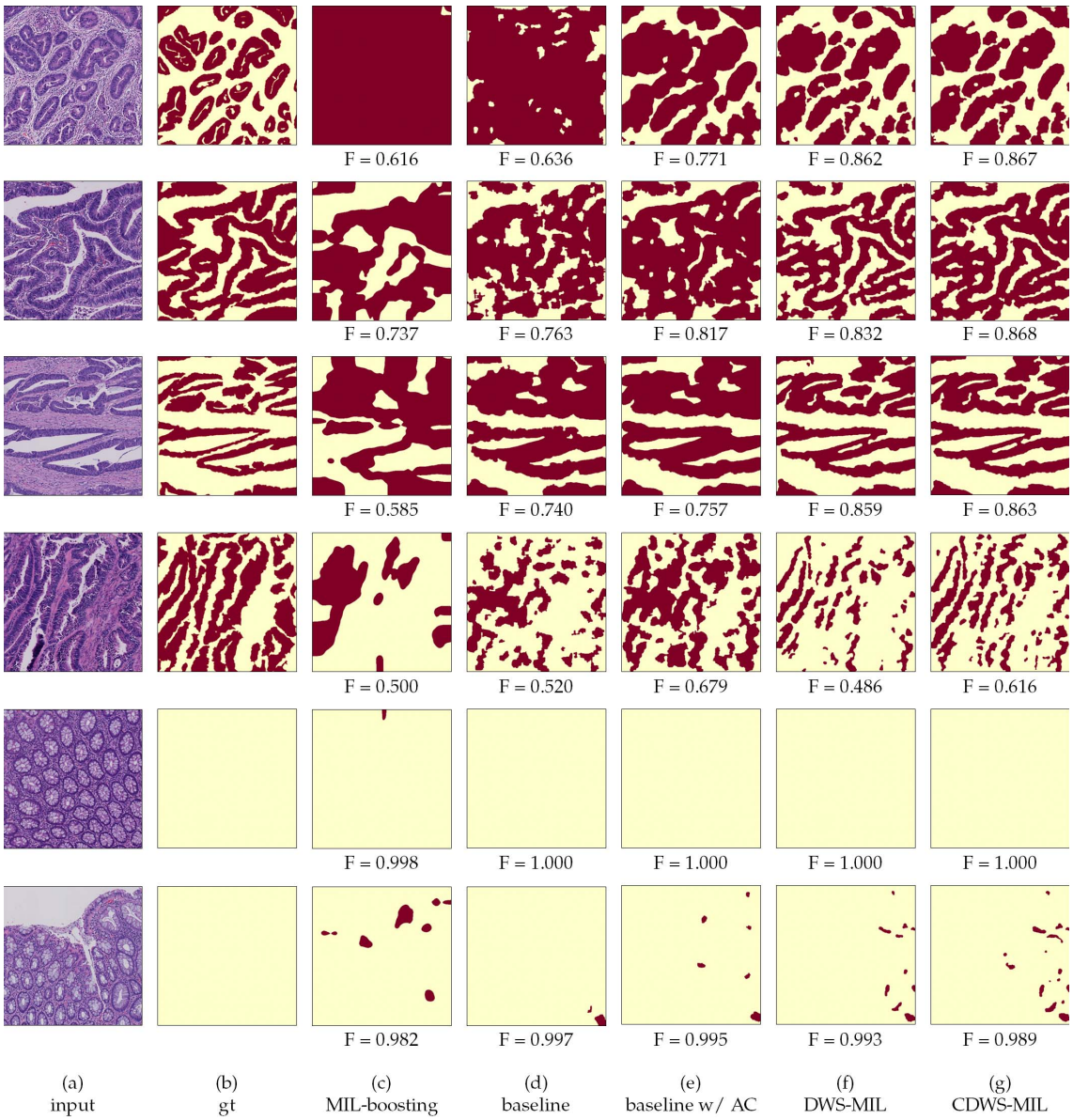


Fig. 5. Segmentation results on *Dataset A*: (a) Input images. (b) Ground truth labels. (c) Results by MIL-Boosting. (d) Results by our baseline. (e) Results by our baseline w/ AC. (f) Results by DWS-MIL. (g) Results by CDWS-MIL. Compared with MIL-Boosting (patch-based), our proposed DWS-MIL and CDWS-MIL produce significantly improved results due to the characteristics we introduced in this paper.

state-of-the-art performance. Figure 6 shows some examples of segmentation results by these full supervised methods.

Notwithstanding the state-of-the-art performance of these supervised learning algorithms, it depends on a large amount of high-quality manual annotations/labeling that are often time-consuming and costly to obtain. By contrast, weakly-supervised learning substantially reduces the amount of work for annotating cancerous tissues/regions. Take our dataset as an example. It takes at least ten minutes for a 3000×3000 colon histopathology image to be annotated in detail. Labeling our 160 test images would be a one-week workload for an expert. In contrast to fully supervised algorithms, no more than a minute is needed for a 3000×3000 image to be annotated in the weakly supervised setting, including the cancerous proportion. Moreover, a large quantities of researchers have applied weakly supervised algorithms to segmentation, such as [19], [34]–[37], [41], and [42]. Our system demonstrates state-of-the-art results on histopathology image datasets and can be applied to various applications in medical imaging beyond histopathology images.

6) Less Training Data: To observe how different amounts of training data influence our baseline method, we train our baseline with less training data. Table III summarizes the results, and Figure 7 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

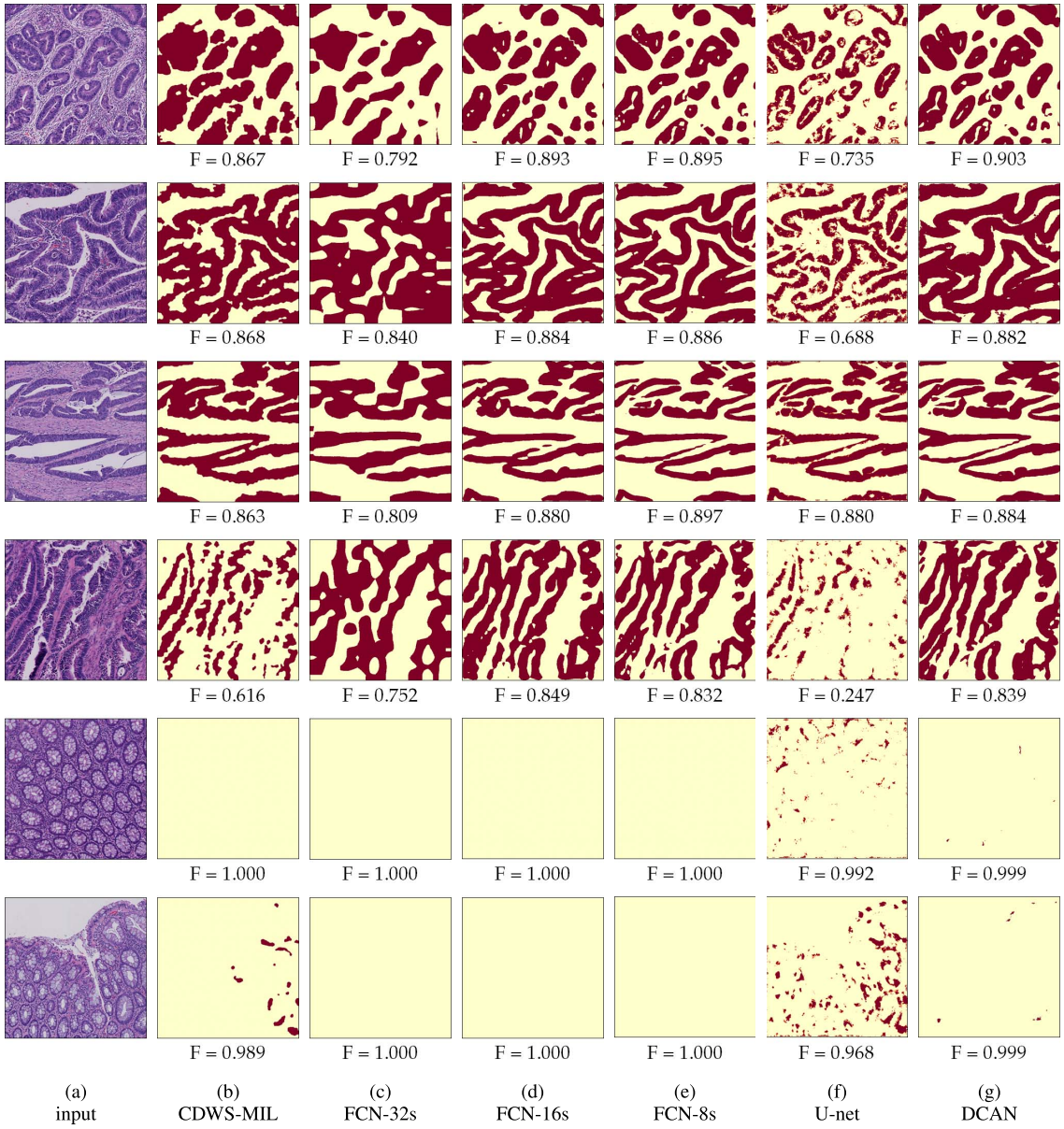


Fig. 6. Comparisons of segmentation results with fully supervised algorithms on *Dataset A*: (a) Input images. (b) Results by CDWS-MIL. (c) Results by FCN-32s. (d) Results by FCN-16s. (e) Results by FCN-8s. (f) Results by U-Net. (g) Results by DCAN. Compared with various fully supervised algorithms, our proposed CDWS-MIL produce near-state-of-the-art results due to the characteristics we introduced in this paper.

TABLE III
PERFORMANCE OF OUR BASELINE TRAINED WITH LESS TRAINING DATA

Training data (Pos,Neg)	ODS (boundaries)		F-measure (regions)			
	CA		CA		NC	
	w/o AC	w/ AC	w/o AC	w/ AC	w/o AC	w/ AC
20%(50,100)	0.337	0.376	0.758	0.801	0.997	0.997
40%(100,200)	0.334	0.378	0.762	0.809	0.997	0.999
60%(150,300)	0.344	0.380	0.778	0.805	0.999	0.999
80%(200,400)	0.345	0.382	0.779	0.813	0.998	0.999
100%(250,500)	0.345	0.383	0.778	0.815	0.998	0.998

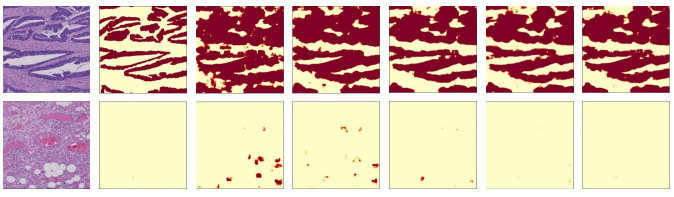


Fig. 7. Differences in results with different amounts of training data: (a) The input images. (b) Ground truth labels. (c) Results that use 20% of training data. (d) Results that use 40% of training data. (e) Results that use 60% of training data. (f) Results that use 80% of training data. (g) Results that use all the training data.

image always results in a lower MIL loss. With a smaller amount of negative training images, it is easier to achieve this objective.

7) Area Constraints: From Table III, the area constraints enable our baseline method to achieve a competitive rate of accuracy with a small training set. Equipped with area

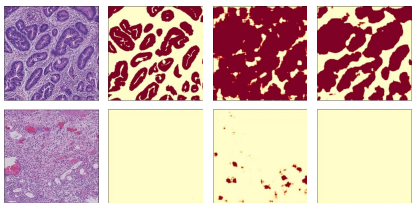


Fig. 8. Comparison of using and not using area constraints: (a) The input images. (b) Ground truth labels. (c) Results of our baseline. (d) Results of our baseline w/ AC. The area constraints loss constrains the model to learn better segmentations.

TABLE IV
PERFORMANCE OF VARIOUS METHODS WITH 20% TRAINING DATA

Method	ODS (boundaries)	F-measure (regions)	
	CA	CA	NC
weakly supervised:			
MIL-Boosting	0.269	0.635	0.995
baseline	0.337	0.758	0.997
baseline w/ AC	0.376	0.801	0.997
DWS-MIL	0.527	0.808	0.998
CDWS-MIL	**0.536**	**0.820**	0.998
baseline w/ SP	0.361	0.762	0.997
baseline w/ AC+SP	0.407	0.805	0.997
DWS-MIL w/ SP	0.534	0.808	0.998
CDWS-MIL w/ SP	**0.547**	**0.823**	0.998
fully supervised:			
FCN-32s	0.408	0.799	0.995
FCN-16s	0.612	0.836	0.994
FCN-8s	0.632	0.839	0.993
U-Net [14]	0.465	0.796	0.956
DCAN [13]	0.638	0.843	0.993

constraints, our baseline method using 20% of training data achieves better accuracy than using all training data without area constraints.

Figure 8 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 illustrated in Figure 8. Table IV summarizes results of different methods 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 in F-measure (0.820 vs 0.758) and 19.9% in ODS (0.536 vs 0.337).

TABLE V
PERFORMANCE OF DIFFERENT SIDE-OUTPUT LAYERS EVALUATED BY F-MEASURE. THE FIRST LINE: DWS-MIL; THE SECOND LINE: CDWS-MIL

F-measure (regions) of CA				F-measure (regions) 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

TABLE VI
PERFORMANCE OF DIFFERENT SIDE-OUTPUT LAYERS EVALUATED BY ODS (FOR BOUNDARIES) OF CA IMAGES. THE FIRST LINE: DWS-MIL; THE SECOND LINE: CDWS-MIL

side1	side2	side3	fusion
0.461	0.515	0.335	0.541
0.479	0.551	0.361	0.559

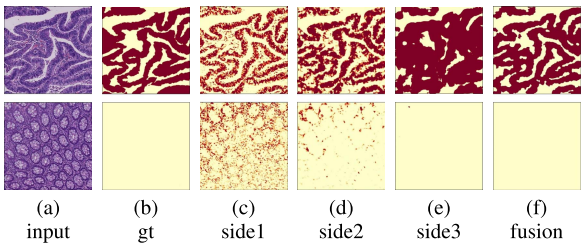


Fig. 9. Results of side-output layers: (a) The input images. (b) Ground truth labels. (c) Results of side-output 1. (d) Results of side-output 2. (e) Results of side-output 3. (f) Results by final fusion. The figure shows a nesting characteristic of segmentation outputs from the lower side-output layer 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.

8) Deep Weak Supervision: To illustrate the effectiveness of deep weak supervision, Tables V and VI summarize segmentation accuracies of the different side-outputs and Figure 9 shows some examples of the different side-outputs. From Tables V and VI, we observe that segmentation accuracy improves from lower layers to higher ones. Figure 9 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.

9) Super-Pixels: We conduct experiments to compare various weakly supervised algorithms and that with super-pixels. Performance of various methods with super-pixels are summarized in Table II, IV, VII. For simplification, wtih super-pixels is denoted as "w/ SP" in the table. We adopt the SLIC method [32] to generate super-pixels and each image produces roughly 900 super-pixels. Figure 10 shows 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. One of the main advantages of super-pixels is its scalability to dierent histology stains and dierent types of glandular tissues, as it is based mainly on

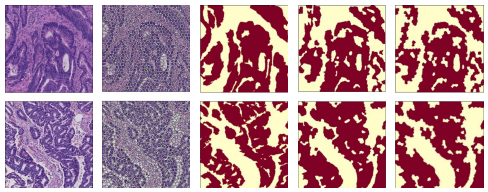


Fig. 10. Comparisons of DWS-MIL and DWS-MIL w/ super-pixel: (a) The input images. (b) Results generated by SLIC method [32]. (c) Ground truth labels. (d) Results of DWS-MIL. (e) Results of DWS-MIL w/ super-pixel. Some detailed edges can be recognized with the help of super-pixels.

TABLE VII
PERFORMANCE OF VARIOUS METHODS ON *Dataset B*. THE EXPERIMENT SETTING IS IDENTICAL TO TABLE II

Method	Running time (s)	ODS (boundaries)	F-measure (regions)	
		CA	CA	NC
weakly supervised:				
MIL-Boosting	8.10	0.205	0.449	0.993
baseline	0.48	0.371	0.599	0.996
baseline w/ AC	0.48	0.384	0.607	0.996
DWS-MIL	0.48	0.409	0.616	0.996
CDWS-MIL	0.48	**0.436**	**0.622**	0.997
baseline w/ SP	0.52	0.380	0.602	0.996
baseline w/ AC+SP	0.52	0.393	0.610	0.996
DWS-MIL w/ SP	0.52	0.410	0.615	0.996
CDWS-MIL w/ SP	0.52	**0.447**	**0.622**	0.998
fully supervised:				
FCN-32s	0.89	0.769	0.628	0.995
FCN-16s	0.94	0.820	0.678	0.995
FCN-8s	1.14	0.833	0.679	0.995
U-Net [14]	1.48	0.670	0.602	0.958
DCAN [13]	1.01	0.835	0.683	0.995

the spatial arrangements of cells rather than texture and color information. The adoption of super-pixels help to predict more detailed boundaries.

10) 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 5 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 CDWS-MIL framework over MIL-Boosting: (1) CDWS-MIL is an end-to-end segmentation framework, which can learn more detailed segmentations than 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 difficult to be integrated into patch-based methods like MIL-Boosting.

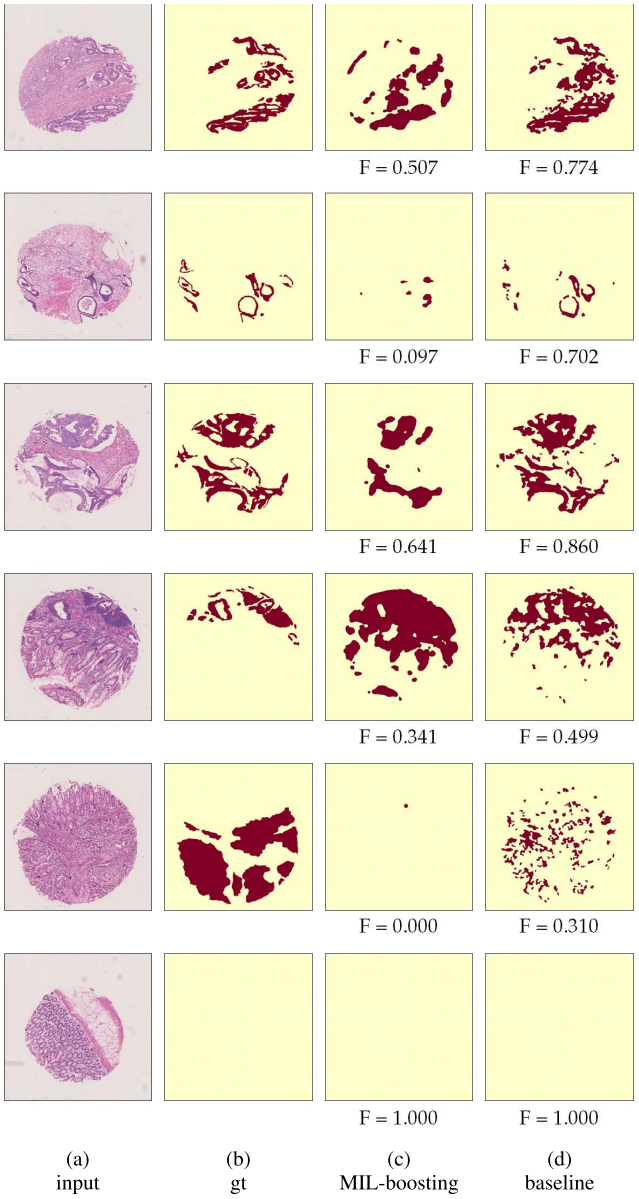


Fig. 11. Segmentation results for *dataset B*: (a) Input images. (b) Ground truth labels. (c) Results by MIL-Boosting. (d) Results by CDWS-MIL. Compared with MIL-Boosting (patch-based), CDWS-MIL produces significantly improved results due to the characteristics we outline in this paper.

C. Experiment B

Dataset B is a histopathology image dataset of 30 colon cancer images and 30 non-cancer images which are referred to as tissue microarrays (TMAs). The dataset is randomly selected from the dataset in [19]. All images have a resolution of 1024 × 1024 pixels. Considering that there is a great deal of blank background for each image, we select an interval of 0.05 for the proportion of cancerous regions. They are annotated in the same way as *Dataset A*. All experiments are conducted with 5-fold cross-validation, and the evaluation metric is the same for *Dataset A*.

We conduct experiments to compare other algorithms with our proposed method CDWS-MIL on *Dataset B*. The comparison experiments all follow *Experiment A*.

Table VII summarizes the results of our proposed algorithms and other methods on *Dataset B*. Figure 11 shows some samples of the segmentation results of these two methods.

D. Running Time

Comparisons of running time using different algorithms have been summarized in Table II, IV, VII. All experiments are conducted on one computer using Tesla K40 GPU with 12G memory except that MIL-Boosting and super-pixels are performed on Intel(R) Xeon(R) CPU e5-2650 v2@2.60GHz.

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 demonstrate that our methods achieve state-of-the-art results on large-scale challenging histopathology images. The scope of our proposed methods are quite broad and they can be applied to a wide range of medical imaging and computer vision applications.

Our method using weak supervision attains performance close to state-of-the-art methods with supervision. However, the results still can be further improved. Some typical failure cases are shown in Figures 5 and 11. In some cancerous images, narrow glandular cavities can not be correctly labeled and segmented. In images of no cancers, sometimes small areas are still being classified as cancerous regions. Our experiments show that constraints are beneficial for improving segmentation accuracy. In the future, we will extend constraints to include other semantic information such as shape and contextual information.

Acknowledgment

The authors would like to thank the Lab of Pathology and Pathophysiology of Zhejiang University in China for providing data and support.

References

[1] A. N. Esgiar, R. N. G. Naguib, B. S. Sharif, M. K. Bennett, and A. Murray, "Fractal analysis in the detection of colonic cancer images," *IEEE Trans. Inf. Technol. Biomed.*, vol. 6, no. 1, pp. 54–58, Mar. 2002.

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

[3] A. Madabhushi, "Digital pathology image analysis: Opportunities and challenges," *Imag. Med.*, vol. 1, no. 1, pp. 7–10, 2009.

[4] S. Y. Park, D. Sargent, R. Lieberman, and U. Gustafsson, "Domain-specific image analysis for cervical neoplasia detection based on conditional random fields," *IEEE Trans. Med. Imag.*, vol. 30, no. 3, pp. 867–878, Mar. 2011.

[5] A. Tabesh *et al.*, "Multifeature prostate cancer diagnosis and Gleason grading of histological images," *IEEE Trans. Med. Imag.*, vol. 26, no. 10, pp. 1366–1378, Oct. 2007.

[6] A. Madabhushi and G. Lee, "Image analysis and machine learning in digital pathology: Challenges and opportunities," *Med. Image Anal.*, 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 Rev. Biomed. Eng.*, vol. 2, no. 2, pp. 147–171, Feb. 2009.

[8] J. Tang, R. M. Rangayyan, J. Xu, I. E. Naqa, and Y. Yang, "Computer-aided detection and diagnosis of breast cancer with mammography: Recent advances," *IEEE Trans. Inf. Technol. Biomed.*, vol. 13, no. 2, pp. 236–251, Mar. 2009.

[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 Recognit.*, 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," *IEEE Trans. Med. Imag.*, vol. 27, no. 11, pp. 1668–1681, Nov. 2008.

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

[12] A. Criminisi and J. Shotton, *Decision Forests for Computer Vision and Medical Image Analysis*. Heidelberg, Germany: Springer, 2013.

[13] H. Chen, X. Qi, L. Yu, and P.-A. Heng, "DCAN: Deep contour-aware networks for accurate gland segmentation," in *Proc. CVPR*, 2016, pp. 2487–2496.

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

[15] F. Xing, X. Shi, Z. Zhang, J. Cai, Y. Xie, and L. Yang, "Transfer shape modeling towards high-throughput microscopy image segmentation," in *Proc. MICCAI*, 2016, pp. 183–190.

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

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

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

[19] 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.

[20] 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 *Proc. ICASSP*, 2014, pp. 1626–1630.

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

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

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

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

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

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

[27] 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 *Proc. ICLR*, 2015, pp. 1–2.

[28] 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 *Proc. MICCAI*, 2015, pp. 374–382.

[29] D. Pathak, E. Shelhamer, J. Long, and T. Darrell, "Fully convolutional multi-class multiple instance learning," in *Proc. ICLR*, 2014, pp. 1–4.

[30] C.-Y. Lee, S. Xie, P. Gallagher, Z. Zhang, and Z. Tu, "Deeply-supervised nets," in *Proc. AISTATS*, 2015, pp. 562–570.

[31] M. C. Roco and W. S. Bainbridge, *Converging Technologies for Improving Human Performance*. Amsterdam, The Netherlands: Springer, 2003.

[32] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "SLIC superpixels compared to state-of-the-art superpixel methods," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 11, pp. 2274–2282, Nov. 2012.

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

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

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

[36] P. O. Pinheiro and R. Collobert, "From image-level to pixel-level labeling with convolutional networks," in *Proc. CVPR*, Jun. 2015, pp. 1713–1721.
[37] 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 *Proc. ICCV*, 2015, pp. 1742–1750.
[38] J.-Y. Zhu, J. Wu, Y. Xu, E. Chang, and Z. Tu, "Unsupervised object class discovery via saliency-guided multiple class learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 4, pp. 862–875, Apr. 2015.
[39] R. G. Cinbis, J. Verbeek, and C. Schmid, "Weakly supervised object localization with multi-fold multiple instance learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 1, pp. 189–203, 2016.
[40] A. Diba, V. Sharma, A. Pazandeh, H. Pirsiavash, and L. van Gool. (2016). "Weakly supervised cascaded convolutional networks." [Online]. Available: https://arxiv.org/abs/1611.08258
[41] Z. Yan *et al.*, "Multi-instance deep learning: Discover discriminative local anatomies for bodypart recognition," *IEEE Trans. Med. Imag.*, vol. 35, no. 5, pp. 1332–1343, May 2016.
[42] L. Hou, D. Samaras, T. M. Kurc, Y. Gao, J. E. Davis, and J. H. Saltz. (2015). "Efficient multiple instance convolutional neural networks for gigapixel resolution image classification." [Online]. Available: https://arxiv.org/abs/1504.07947
[43] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. ICLR*, 2015, pp. 1–14.
[44] N. Yu *et al.*, "A superpixel-based framework for automatic tumor segmentation on breast DCE-MRI," *Proc. SPIE*, vol. 9414, p. 94140O, Mar. 2015.
[45] M. Soltaninejad *et al.*, "Automated brain tumour detection and segmentation using superpixel-based extremely randomized trees in flair mri," *Int. J. Comput. Assist. Radiol. Surgery*, vol. 12, no. 2, pp. 183–203, 2017.
[46] O. Russakovsky *et al.*, "Imagenet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Mar. 2014.
[47] 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 *Proc. ICASSP*, 2015, pp. 947–951.
[48] H. Chen, X. Wang, and P. A. Heng, "Automated mitosis detection with deep regression networks," in *Proc. ISBI*, 2016, pp. 1204–1207.
[49] V. Murthy, L. Hou, D. Samaras, T. M. Kurc, and J. H. Saltz. (2016). "Center-focusing multi-task CNN with injected features for classification of glioma nuclear images." [Online]. Available: https://arxiv.org/abs/1612.06825
[50] Y. Jia *et al.* (2014). "Caffe: Convolutional architecture for fast feature embedding." [Online]. Available: https://arxiv.org/abs/1408.5093
[51] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. ICLR*, 2015, pp. 1–15.
[52] P. Arbeláez, M. Maire, C. Fowlkes, and J. Malik, "Contour detection and hierarchical image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 5, pp. 898–916, May 2011.