# DSNet: A Dual-Stream Framework for Weakly-Supervised Gigapixel Pathology Image Analysis

Tiange Xiang, Yang Song, *Member, IEEE*, Chaoyi Zhang, Dongnan Liu, *Member, IEEE*, Mei Chen, *Senior Member, IEEE*, Fan Zhang, Heng Huang, *Member, IEEE*, Lauren O'Donnell, and Weidong Cai, *Member, IEEE*

***Abstract*— We present a novel weakly-supervised framework for classifying whole slide images (WSIs). WSIs, due to their gigapixel resolution, are commonly processed by patch-wise classification with patch-level labels. However, patch-level labels require precise annotations, which is expensive and usually unavailable on clinical data. With image-level labels only, patch-wise classification would be sub-optimal due to inconsistency between the patch appearance and image-level label. To address this issue, we posit that WSI analysis can be effectively conducted by integrating information at both high magnification (local) and low magnification (regional) levels. We auto-encode the visual signals in each patch into a latent embedding vector representing local information, and down-sample the raw WSI to hardware-acceptable thumbnails representing regional information. The WSI label is then predicted with a Dual-Stream Network (DSNet), which takes the transformed local patch embeddings and multi-scale thumbnail images as inputs and can be trained by the image-level label only. Experiments conducted on three large-scale public datasets demonstrate that our method outperforms all recent state-of-the-art weakly-supervised WSI classification methods.**

***Index Terms*— Weakly-supervised training, image classification, whole slide images.**

## I. Introduction

CONVOLUTIONAL Neural Networks (CNNs) have shown their extraordinary capabilities in feature extraction in image analysis tasks. With the help of CNNs, great progress has been made in pathology image analysis [1]–[6]. However, whole slide images (WSIs) cannot be analyzed the same way as typical digital images due to their gigapixel resolution. To achieve a feasible and effective WSI analysis (e.g. classification), most of the existing methods adopt a patch-based approach that first divides the WSI into small patches and then predicts the WSI-level label based on the predicted labels of individual patches [7]–[13]. Unfortunately, such fully-supervised approach relies on patch-level ground truth annotations, which require time consuming efforts from pathology experts. Meanwhile, image-level labels of WSIs are more readily attainable, therefore we seek to answer the question of whether a weakly supervised approach employing image level WSI labels can achieve comparable performance.

We are mindful to ensure that our weakly-supervised approach is able to fit into real clinical workflow, which pathologists typically analyze WSIs by first skimming through a thumbnail of the WSI to look for possible regions of interest that may contain cancerous tissues. Then, they zoom in on the regions of interest to investigate the local details. While the above process resembles the attention mechanism [14], [15], such a mechanism cannot be applied directly to WSIs due to their gigabyte sizes. We propose a whole image-based approach where gigapixel WSIs are first transformed into two matrices in compact latent spaces representing local and regional descriptors of the raw WSI. Then, a Dual-Stream Network (DSNet) integrates the compressed representations through a stack of Concurrent Bottleneck blocks. The two matrices are adaptively aggregated by assigning importance scores through both stream-wise and channel-wise attentions. Intuitively, our DSNet uses the coarse visual clues implied in thumbnails to guide detailed analysis at zoomed-in regions of interest. Our weakly-supervised framework is outlined in Figure 1.

The overall technical novelty of our work is three-fold: **(i)** We proposed a global-local framework that achieves weakly supervised WSI analysis based on multi-scale thumbnails and neural encoded embeddings, which differs from [16] that classifies WSI embedding directly and [17] that uses raw RGB images (and patches) only. **(ii)** We improved the unsupervised WSI encoding schema [16] by adapting S-CAE [18] with

Manuscript received 3 February 2022; revised 3 March 2022; accepted 6 March 2022. Date of publication 9 March 2022; date of current version 1 August 2022. *(Corresponding author: Weidong Cai.)*

Tiange Xiang, Chaoyi Zhang, Dongnan Liu, and Weidong Cai are with the School of Computer Science, The University of Sydney, Sydney, NSW 2008, Australia (e-mail: chrix0928@gmail.com; chaoyi.zhang@sydney.edu.au; dongnan.liu@sydney.edu.au; tom.cai@sydney.edu.au).

Yang Song is with the School of Computer Science and Engineering, University of New South Wales, Kensington, NSW 2052, Australia (e-mail: yang.song1@unsw.edu.au).

Mei Chen is with Microsoft Corporation, Redmond, WA 98052 USA (e-mail: may4mc@gmail.com).

Fan Zhang and Lauren O'Donnell are with the Brigham and Women's Hospital, Harvard Medical School, Boston, MA 02115 USA (e-mail: fzhang@bwh.harvard.edu; odonnell@bwh.harvard.edu).

Heng Huang is with the Department of Electrical and Computer Engineering, University of Pittsburgh, Pittsburgh, PA 15261 USA (e-mail: henghuanghh@gmail.com).

Digital Object Identifier 10.1109/TMI.2022.3157983

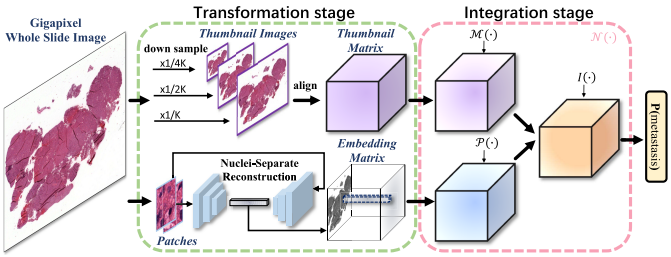


Fig. 1. Our WSI classification framework consists of a **Transformation** stage followed by an **Integration** stage. Our **Integration** stage is achieved by the proposed DSNet $\mathcal{N}(\cdot)$, which is comprised of different building blocks $\mathcal{M}(\cdot)$ and $\mathcal{P}(\cdot)$ for each stream, and ends up with an adaptive aggregation $\mathcal{I}(\cdot)$ of the two streams.

non-trivial upgrades to achieve nuclei-tissue separate encoding to sift out non-informative visual redundancies. **(iii)** We optimized the network architecture considering the sparsity issues in WSI analysis, and proposed multiple novel operators and modules for better effectiveness and efficiency. We validated our method on three large-scale benchmark datasets Camelyon16 [9], TCGA-LUSC [19], and BCNB [20], and our method outperforms recent state-of-the-art weakly-supervised methods with significantly fewer network parameters.

Our framework can be easily extended to other WSI-related tasks, such as thumbnail-wise tumor region segmentation and the metastasis localization. With abundant local and regional information provided by our framework, our DSNet can be easily modified to an encoder-decoder structure to fit different supervision signals.

## II. Related Work

### A. Weakly-Supervised WSI Classification

Classifying images that are too large to fit into the computer memory is a challenging topic [1], [3]. State-of-the-art classification methods on extremely high resolution images, such as satellite images [21], [22] and WSIs [7]–[10], [23]–[25] require fully annotated patch-level annotations as a guarantee, which are expensive to acquire and unavailable for most datasets.

Without patch-wise annotations, WSI classification can be approached in a weakly-supervised manner with image-wise [16] or patch-wise [26], [27] designs. As a representative image-wise approach, Tellez *et al.* [16] used a Generative Adversarial Network [28] to compress WSI into the latent space with a smaller size, and then employed a standard deep learning model to make predictions.

Differing from image-wise methods, patch-wise methods followed a Multi Instance Learning (MIL) paradigm, where groups of instances (bags) are analyzed based on the group label. When determining the presence of tumor or metastasis in WSIs, positive bags indicate WSIs with at least one patch containing diseases and negative bags when all patches in the WSIs are classified negative. Chen *et al.* [29] designed a rectified cross-entropy loss to model cancer metastasis patterns. Hou *et al.* [30] proposed the usage of a fusion model that sifts out discriminative patches by utilizing a patch level CNN. Instead of selecting discriminative patches only, Chikontwe *et al.* [26] designed a center loss that maps patch embeddings to a single image embedding centroid and reduced the intra-class variations. [31] proposed a similar multi-scale patch-wise approach for analysing WSIs with the help of domain adversarial learning. With the help of GANs, [32] proposed a graph convolutional network for learning bag-level representations. However, the proposed VAE-GAN relies on extra domain knowledge and annotations from the WSI experts. In a recent work, [33] designed a multi-task classification framework for predicting the source of metastasis in WSI slides with low resource costs. However, most of the patch-wise weakly-supervised methods have some evident drawbacks: localizing discriminative patches can be difficult and inaccurate due to their restricted receptive field. Also, the patch-wise processing requires inference of all patches, which is time consuming. Although the workflows of DSNet and GLNet [17] share a similar design, our framework differs from GLNet in four aspects. *(i)* Both branches in GLNet take the raw RGB images (in different resolutions) as inputs, while DSNet only accepts down-sampled thumbnail images in the global branch. In the local branch, DSNet is fed with the abstract embedding matrix that sifts out the visual redundancies and contains only meaningful semantics, which fosters the optimization of DSNet further. *(ii)* We considered the sparsity issue in WSIs and designed multiple sparsity-aware operators to be incorporated in the DSNet. *(iii)* Multiple novel operators (multi-scale convolution block, concurrent bottleneck and adaptive aggregation) were designed that prove to be effective. *(iv)* Our framework was designed in a resource-aware style with both minimum network parameters and inference latency, while GLNet focuses on segmentation accuracy only.

### B. Attention Mechanisms

The attention mechanism can be interpreted as enhancing informative signals while suppressing useless ones [34]–[37]. It has been applied to a wide variety of visual applications recently [38]–[42]. For example, Wang *et al.* [14] used extra encoder-decoder style branches to model the spatial attention at multiple levels in a CNN. The proposed network significantly boosts performance in image classification by denoising the intermediate feature maps. Hu *et al.* [43] investigated attention from channel-wise perspective. The proposed SENet self-recalibrates different feature maps regardless of their spatial signals. Furthermore, Roy *et al.* [44] and Woo *et al.* [45] introduced novel attention strategies by combining the spatial-wise and channel-wise attentions to further enhance the ability of the CNNs. Following the above methods, GSoP [46] designed a second-order pooling operator, which utilizes the covariance matrix for learning the channel-wise attentions. SKNet [47] proposed to adjust receptive field size of convolution kernels dynamically based on multi-scale features. In a recent work, ECA-Net [48] proposed a more efficient channel attention framework, that attention logits can be learned with a constant number of parameters. Unlike the above methods, we apply concurrent stream-wise and channel-wise attentions for an adaptive aggregation of features encoded at different magnifications.

## III. Method

### A. Framework

As outlined in Figure 1, we define our image-level weakly-supervised WSI classification framework as a two-stage sequential process: an efficient **Transformation** from gigapixel WSIs into a compact latent space, followed by an effective **Integration** that captures and fuses the underlying patterns encoded in the transformed representations.

During the **Transformation** stage, our framework takes the gigapixel WSI $\mathbf{X}$ as input and converts it into two correlated compact representations, which are the *thumbnail matrix* $\mathbf{T}$ and the *embedding matrix* $\mathbf{V}$, such that:

$$||\mathbf{T}|| \ + \ ||\mathbf{V}|| \ \ll ||\mathbf{X}||, \tag{1}$$

where $||\cdot||$ denotes the number of pixels. Given the constraint of Eq. 1, we adopt an unsupervised transformation strategy to ensure the latent representations keep as much discriminative information in $\mathbf{X}$ as possible. Subsequent to the **Transformation** stage, our **Integration** process (Sec. III-C) can be formulated as:

$$\underset{\theta}{\text{argmax}}\, \mathbf{Pr}\big(\mathcal{N}\left(\mathbf{T}, \mathbf{V}; \theta\right), y\big), \tag{2}$$

where $y$ is the WSI label, $\mathcal{N}(\cdot)$ and $\theta$ are our DSNet and its model parameters, respectively.

The *thumbnail matrix* $\mathbf{T}$ of $\mathbf{X}$ is comprised of thumbnail images at different magnification levels. Practically, thumbnails of different levels can be obtained by downsampling $\mathbf{X}$ using a set of pre-defined factors {K} to hardware-acceptable sizes. We then construct $\mathbf{T}$ by nearest neighbor interpolation of the smaller thumbnails to align with the finest-scale image. The resultant matrix is regarded as a global representation of $\mathbf{X}$ that encodes coarse-scale multi-level regional information.

### B. Patch Encoding

Sharing the same spatial correspondences as the *thumbnail matrix* $\mathbf{T}$, the *embedding matrix* $\mathbf{V} = \{\mathbf{v}_{i,j}\}$ of WSI $\mathbf{X} = \{\mathbf{x}_{i,j}\}$ encodes local information at the patch level. We assume that not all visual information in a patch $\mathbf{x}_{i,j} \in \mathbb{R}^{\mathrm{S}\times\mathrm{S}\times 3}$ contributes to cancer diagnosis, thus $\mathbf{V}$ is constructed by removing visual redundancies in every patch of $\mathbf{X}$. We denote the representative vector $\mathbf{v}_{i,j} \in \mathbb{R}^{1\times 1\times \mathrm{C}}$ as the compressed patch $\mathbf{x}_{i,j}$. An autoencoder network is adopted to formulate the mapping between $\mathbf{v}_{i,j}$ and $\mathbf{x}_{i,j}$ in an unsupervised style by reconstructing $\mathbf{x}_{i,j}$. The autoencoder is then slid throughout all spatial locations $i, j$ with a stride of S in both directions to generate a complete set of non-overlapping representation vectors $\mathbf{v}_{i,j}$ that eventually construct the *embedding matrix* $\mathbf{V}$. The above process maps a patch in the input WSI to a pixel in $\mathbf{V}$ with visual redundancies removed.

Our patch encoding process is outlined in Figure 2. It follows an autoencoder structure based on the **S**parse **C**onvolutional **A**uto**E**ncoder (S-CAE) [18] with two branches to encode the foreground and background embeddings separately. The encodings are then separately decoded and summed up together to reconstruct the input patch. Differing from the mixed encoding in conventional autoencoders, we distinguish

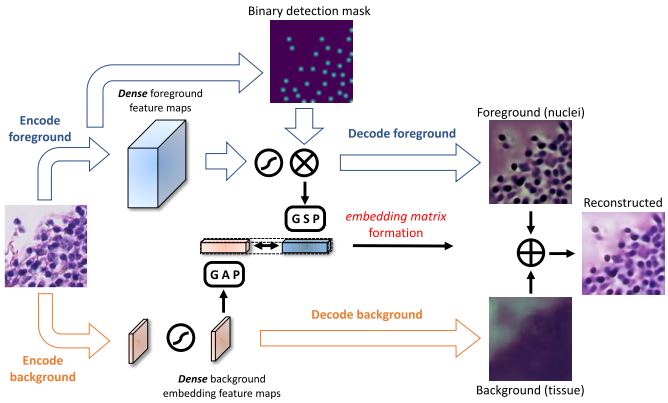


Fig. 2. **Our embedding matrix encoding process.** Representative vector of input patch is constructed as the concatenation of the pooling results of sparse foreground embeddings and dense background embeddings.

the nuclei from tissues to achieve a foreground-background separate encoding.

We encode the background features as small dense embedding feature maps $\mathbf{B}$. The majority of the patch background is tissue with relatively homogeneous color and texture, which does not carry sufficient diagnostic information. Therefore, a set of compact feature maps is adequate to describe the background tissue patterns.

Unlike background tissues, the foreground nuclei contain more complex patterns and require more descriptive feature representations. The dispersed nuclei can be encoded by activating the tensor units near the nuclei only. The small amount of nuclei existed in each patch would lead to *sparse* foreground embedding feature maps $\mathbf{F}$. To ensure the cross-wise sparsity [18] of the nuclei in any foreground feature map $\mathbf{F}_l \in \mathbf{F}$, a binary detection mask $\mathbf{M}$ can be generated. $\mathbf{M}^{i,\ j}$ indicates whether a nucleus appears at location $(i, j)$ and the corresponding $\mathbf{F}_l^{i,\ j}$ is activated, such that:

$$\mathbf{M}^{i,\ j} = \mathbb{1}\big(\sum_{l} \mathbb{1}(\mathbf{F}_l^{i,\ j} \neq 0) > 0\big), \tag{3}$$

where $\mathbb{1}(\cdot)$ is an indication function that returns 1 if the condition is true and 0 otherwise. Note that our binary detection mask $\mathbf{M}$ can be directly computed based on $\mathbf{F}$, and no gradient is required. The gradient flow from decoders to encoders is then ensured by using an extra branch.

Eq. 3 yields that, if any feature channel of a pixel in $\mathbf{F}$ carries useful information, we presume a nucleus can be observed at such location, and set the mask value to 1. However, such condition can be hardly satisfied in practice, and an alternative strategy is to pre-set a fixed number of neurons with the smallest values to be activated in $\mathbf{M}$. In reality, the number of nuclei in different patches would vary significantly, and units in $\mathbf{M}$ are better activated adaptively. We therefore set an adaptive sparsity rate $\rho$ to control the activation rates of the nuclei across different patches. With $\rho$, $\mathbf{M}$ is now computed by thresholding $\mathbf{F}$ at $\rho_{th}$ percentile for all spatial locations. The sparsity rate is then updated following the running average approach [49] with a pre-defined momentum. A high $\rho$ encodes fewer nuclei and a low $\rho$ confuses foreground with

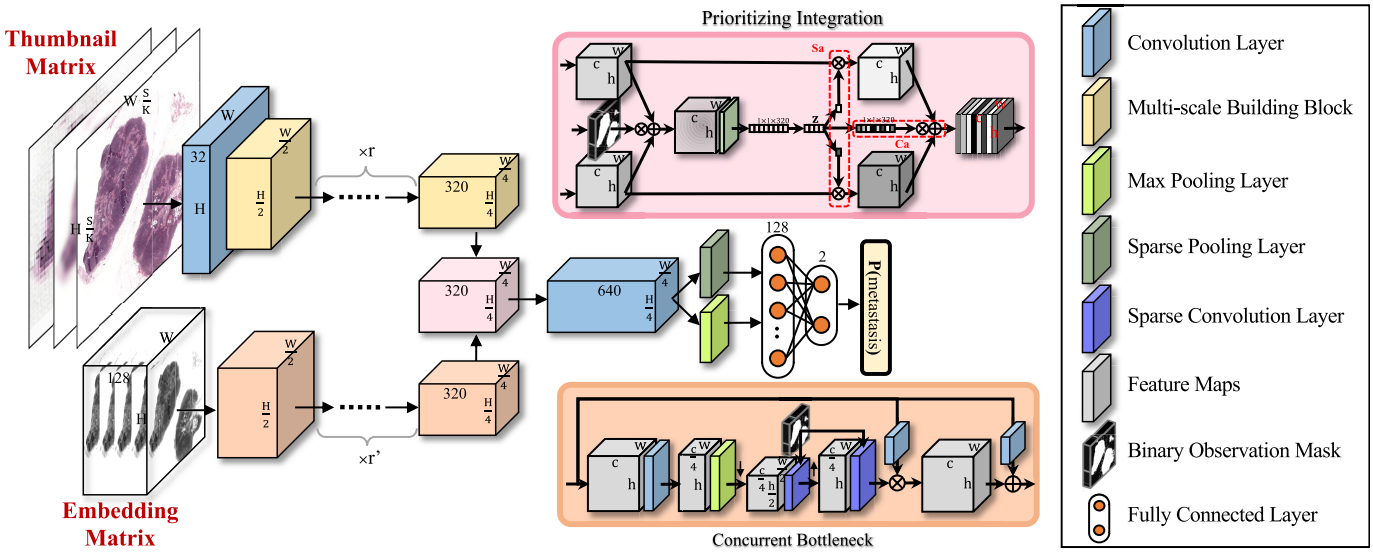


Fig. 3. **Overview of our DSNet architecture.** Relative sizes of feature maps are shown in this figure. '×r' denotes repeating the same block by r times. **Sa** denotes the stream-wise attention and **Ca** the channel-wise attention. More basic operators are shown in Figure 4.

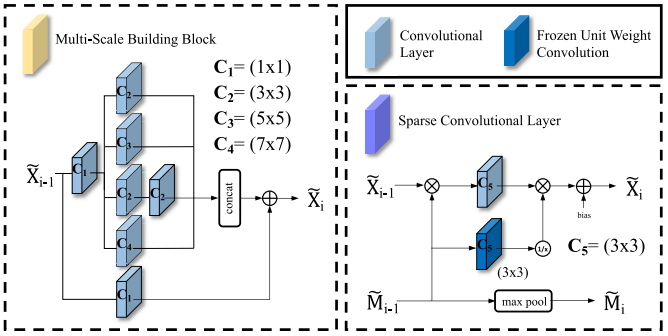


Fig. 4. **Details of the basic operators used in DSNet**. $\tilde{\mathrm{X}}$ and $\tilde{\mathrm{M}}$ denote feature maps and binary observation mask, respectively. "concat" denotes the concatenation operation, "max pool" denotes max pooling operation and $1/x$ denotes element-wise inversion. Kernel sizes are illustrated as well.

more background tissues, hence the optimal $\rho$ can be decided from a binary search performed on each dataset. Finally, we get the sparse foreground feature maps **F** as the element-wise multiplication of **M** and the densely encoded foreground features.

Sigmoid gates are employed to normalize the embedding feature maps by restricting the features within (0, 1). By global averaging on dense background embedding **B** and sparse **F** (discarding the 0 elements), the representative vector $\mathbf{v}_{i,\ j}$ for the patch $\mathbf{x}_{i,\ j}$ can be obtained as the concatenation of the background and foreground embedding features. Our DSNet $\mathcal{N}$ then takes and fuses *thumbnail matrix* **T** and *embedding matrix* **V** to produce the final prediction.

Compared to the original S-CAE [18], instead of adopting as a simple nuclei detection tool, the encoded sparse foreground and dense background features are further processed to be better adopted into our DSNet. To this end, we upgrade the original S-CAE architecture along the following dimensions: *(i)* We used more advanced building blocks [50], [51] to extract low-level features in the image patches. *(ii)* We alleviated the artifacts shown in the decoding process by optimizing the upsampling technique with bilinear resizing. *(iii)* We incorporated a feature bottleneck (Fig. 2) by using global average pooling and the proposed global sparse pooling to transfer the task from nuclei detection (the original objective of S-CAE) to our embedding matrix encoding.

Converging patch reconstruction loss proves that there exists at least one function (the learned decoder) that is able to understand and capture the encoded information in **V**. Therefore, with supervision, our DSNet may be also capable of decoding the underlying patterns implied in **V**. At the unsupervised **Transformation** stage, no diagnostic information is expected to be encoded in **V**, as we delegate the semantic reasoning job to the DSNet during the **Integration** stage with the supervision of WSI labels.

### C. Dual-Stream Network

Our **Integration** stage is to take **T** and **V** of a WSI and infer the WSI-level label through the proposed DSNet. The design challenges of this stage include: **(1)** how to efficiently capture the encoded patterns in **V** and **T**, **(2)** how to adaptively aggregate **V** and **T** towards the final prediction, and **(3)** how to alleviate the sparsity issue encountered in the representation matrices. To accomplish **(1)**, we design basic building blocks to extract discriminative features from each representation matrix. To achieve **(2)**, we employ an attention-based weighting process to obtain the aggregation adaptively. To tackle **(3)**, we replace the dense operators with the sparsity-aware ones to function on such sparse data.

*1) Building Blocks:* The *thumbnail matrix* **T** is comprised of raw RGB images at different magnifications. We adopt the commonly used Multi-Scale blocks [2] to extract varying range regional information in **T**. For better computational efficiency, we replace the standard convolutions with their separable variants [52]. Unlike the highly squeezed thumbnail images, the neural encoded **embedding matrix V** implies

great spatial-wise redundancies. The intuition is that in WSIs, adjacent patches usually have similar appearances and semantics, hence it would be sufficient to apply convolutions on the local representative of a neighborhood (e.g. the maximum values) instead of all nearby units. To better discriminate local features in $\mathbf{V}$, we employ a max pooling layer to create spatial bottlenecks. A paired upsampling layer is then utilized to map the max pooled tensor back to its original size, and we insert the convolutions between and after the spatial bottleneck. We embed the spatial bottleneck into a channel bottleneck block [50], constructing the Concurrent Bottleneck building blocks for $\mathbf{V}$.

*2) Adaptive Aggregation:* At the end of our DSNet, the two stream features are eventually aggregated toward the final label prediction. We achieve this in a self-attention manner. First, a sigmoid gated stream-wise attention [47] score is generated to scale all features in one stream, whose residual of 1 is multiplied to the other stream broadcastly. We then learn a subsequent set of channel-wise attention [43] to the element-wise summation of the two scaled streams.

*3) Sparsity in the Embedding Matrix:* Usually, more than 50% of a WSI is covered by connective tissue, which does not provide useful diagnostic information, and can be excluded during pre-processing (Sec. IV-C). Sifting out the large number of uninformative patches and replacing them with zero introduces significant sparsity in $\mathbf{V}$. Standard dense operators (e.g. pooling, normalization) without ignoring the zero values perform poorly on such sparse WSI representations. To tackle the problem, we apply sparse convolution [53] on the sparse $\mathbf{V}$ for basic feature aggregation. Compared to the conventional convolution operator, an extra binary observation mask is constructed indicating whether a pixel in $\mathbf{V}$ (a patch in WSI) is valid or not.

To better handle the sparsity in $\mathbf{V}$, we modify the dense pooling and normalization operators by masking the sparse $\mathbf{V}$ first, and then calculate on the non-zero values only. Conventional average calculation operator averages the values across all spatial locations. However, applying such operator directly on a sparse tensor might suffer from a significant information under-weighting, as a large amount of zeros are also taken into the calculation. With the help of the observation mask, we derive the sparse mean by neglecting all zero values and averaging over the masked entries only. We modify the variance calculation operator on sparse data by using the observation mask as well. We compute the sparsity-aware mean and variance as:

$$\mathtt{mean}_s(\mathbf{F}) = \frac{\sum_{i,j}(\mathbf{F}_{i,j} * \mathbf{O}_{i,j})}{\sum_{i,j}\mathbf{O}_{i,j}}, \tag{4}$$

$$\mathtt{var}_s(\mathbf{F}) = \frac{\sum_{i,j}(\mathbf{F}_{i,j} - \mathtt{mean}_s(\mathbf{F}))^2 - \beta}{\sum_{i,j}\mathbf{O}_{i,j}}, \tag{5}$$

where $\mathbf{F}$ is a sparse feature map, $\mathbf{O}$ is the binary observation mask, and overweight $\beta$ is computed as $\sum_{i,j}(1-\mathbf{O}_{i,j}) * \mathtt{mean}_s(\mathbf{F})^2$ indicating the over-weighted mean values brought by the zero entries.

*4) Network Structure:* As the name suggests, our DSNet processes the two matrices $\mathbf{T}$ and $\mathbf{V}$ in two separate streams with stacks of the aforementioned multi-scale blocks and concurrent bottleneck blocks. The local and regional feature maps are eventually fed into the adaptive aggregation block to assign importance scores to each of the streams and their integration respectively. Similar to [54], the classification head starts with a point-wise convolution that maps the aggregated features to a higher dimension. A sparse pooling layer together with a max pooling layer are then employed to generate global descriptors by summarizing over spatial dimensions. Our DSNet ends with two fully-connected layers to regress the final classification scores. Figure 3 illustrates an overview of our DSNet architecture, and the building block details are presented in Table I.

TABLE I
BUILDING BLOCK PARAMETER DETAILS IN OUR DSNET. MS DENOTES THE MULTI-SCALE BLOCK, CB THE CONCURRENT BOTTLENECK BLOCK

| Stream | Block | #in | #out | kernel size | stride |
|---|---|---|---|---|---|
| Thumbnail | Conv | 3 * L | 32 | 7 | 2 |
| Thumbnail | MS | 32 | 64 | 3,5,7 | 1 |
| Thumbnail | MS | 64 | 144 | 3,5,7 | 2 |
| Thumbnail | MS | 144 | 256 | 3,5,7 | 2 |
| Thumbnail | MS | 256 | 320 | 3,5,7 | 1 |
| Embedding | CB | 128 | 144 | 5 | 2 |
| Embedding | CB | 144 | 224 | 5 | 1 |
| Embedding | CB | 224 | 256 | 5 | 2 |
| Embedding | CB | 256 | 320 | 5 | 1 |

## IV. EXPERIMENTS

In this section, empirical experiments are performed to validate the effectiveness of the proposed weakly-supervised image-level method in WSI classification task.

### A. Dataset Description and Preparation

In this work, our method was evaluated on three large-scale public available datasets: Camelyon16 [9], The Cancer Genome Atlas Lung Squamous Cell Carcinoma project (TCGA-LUSC) [19], and Early Breast Cancer Core-Needle Biopsy WSI (BCNB) [20].

*1) Camelyon16:* The Camelyon16 dataset contains 400 sentinel lymph node Hematoxylin and Eosin (H&E) stained WSIs from breast cancer patients. We followed the official dataset split with 270 training WSIs and 130 validation WSIs. 110 of the 270 training WSIs are positive cases (metastasis) and the rest 160 are negative cases (normal). The 130 validation WSIs comprised of 50 positive cases and 80 negative cases.

To construct the embedding matrix $\mathbf{V}$, raw WSIs are segmented into $256\times256$ patches to be fed into our S-CAE. There is an average number of 140296 patches in the Camelyon16 WSIs. However, training S-CAE with all of the data is time consuming and unnecessary, we randomly drawn 600 patches from each of the 20 randomly selected WSIs in the positive training set, negative training set, and validation set respectively to conduct network training. In order to better encode

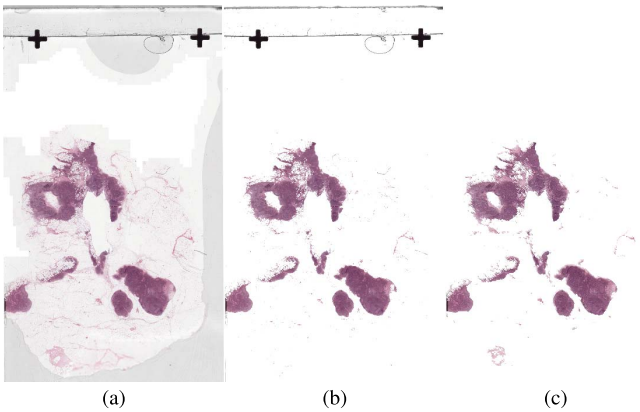

Fig. 5. **Pre-processing approaches.** (a) Raw WSI input. (b) Pre-processed by Otsu algorithm [55]. (c) Pre-processed by Run-length Encoding (RLE) [56] sorting and threshold (ours). (Background colours are reverted for better visualization.)

the patches with a rich distribution of nuclei, we ensure at least 90% of the training patches are drawn with all three channels' intensities being less than 70% of the maximum intensity, which represents samples from rich nucleus regions.

*2) TCGA-LUSC:* The Cancer Genome Atlas (TCGA) is a public database that collects H&E stained WSIs across various cancers and patients. We evaluated our method on 500 WSIs randomly sampled from the Lung Squamous Cell Carcinoma (LUSC) project for binary classification with the WSI-level label available only. The 500 WSIs contain equal number of primary solid tumor images and solid tissue normal images. We randomly chose 400 WSIs (with 200 WSIs in each class) as the training set, and the rest as the validation set. For fair comparisons, we reproduced all competing methods using the same dataset split.

The embedding matrix construction process is similar to the one used for Camelyon16. The WSI slide size is considerably smaller than the ones in Camelyon16, with an average number of 4872 patches. Therefore, we drawn only 100 patches from each of the 50 randomly selected WSIs in the positive training set, negative training set, positive validation set, and negative validation set respectively.

*3) BCNB:* Early Breast Cancer Core-Needle Biopsy WSI (BCNB) dataset includes core-needle biopsy whole slide images (WSIs) of early breast cancer patients and the corresponding clinical data. There are WSIs collected from a total of 1058 patients with only one WSI sampled from each patient. The entire dataset is officially [20] split into a training set, validation set, and testing set consisting of 630, 210, and 218 WSIs respectively. The WSIs are annotated from multiple perspectives including age, tumor size, tumor type, ER, PR, HER2, HER2 expression, histological grading, surgical, Ki67, molecular subtype, number of lymph node metastases, and the metastatic status of auxiliary lymph node (ALN). Following the official guideline [20], we benchmarked our DSNet on this dataset for weakly supervised ALN classification with N0 as the negative class and N(+) as the positive class. Unlike metastasis detection as in Camelyon16 and TCGA-LUSC, classifying the metastatic status of auxiliary lymph node in such a large dataset is much more challenging. The embedding matrix and thumbnail matrix construction processes are identical to the ones used for Camelyon16.

### B. Evaluation Metrics

Various metrics are used to evaluate the performance in the experiments: classification accuracy, precision, recall, and $F_1$-score. In addition, Receiver Operating Characteristic (ROC) curve and Area Under ROC (AUC) are also evaluated for comparing the performance of different methods.

### C. Image Pre-Processing

Before preparing all of the datasets, we applied extra WSI pre-processing procedures to exclude tissue fluid and blood to enhance the accuracy of our method and to reduce computation costs. We then crop the tightest bounding box on the foreground pixels to focus on the slide regions only. Note that we did not stain normalize the WSIs in our experiments, as we found that such pre-processing empirically led to poorer performances.

WSI pre-processed results with different approaches are visualized in Figure 5. The commonly used Otsu [55] algorithm cannot effectively remove the ink marks on the WSI slides. By applying the Run-Length Encoding (RLE) [56] on the average pixel intensity of patches, a threshold can be easily decided between the large encoding patches (rich pixel variations, i.e. patches with cells) and the small encoding patches (i.e. patches with tissues/inks).

### D. Implementations

We implemented all experiments on a NVIDIA RTX 2080 GPU using the Keras framework [58].

*1) Hyper-Parameter Settings:* There are three main hyper-parameters considered in our framework: embedding matrix channel number C, patch size S, and the greatest thumbnail downsampling rate K. $C = 128$ and $S = 256$ are directly borrowed from Tellez *etal*. [16]. For the thumbnail image downsampling rate K, we followed a simple intuition that sets $K = 128$ to be half of the patch size. In this way, the thumbnail image will contain at least four times more pixels than the embedding matrix and provide much more abundant regional information. All of these three hyper-parameters are adopted as default without any explicit tuning. Moreover, in our early trials, we found that tuning these hyperparameters contributed little to the overall framework performance. For better consistency and generalization ability, we used the most intuitive hyperparameters and kept them unchanged across all our experiments.

*2) S-CAE Implementation Details:* In all experiments related to S-CAE, we use SGD optimizer with Nesterov momentum to minimize the mean squared error (MSE) between inputs and outputs. The initial learning rate is set to 0.03, weight decay to 0.00001, momentum to 0.8, and batch size to 8. We train the S-CAE for 6 epochs for each of the selected WSIs on both datasets. All patches are extracted at the highest magnification level. Foreground and background pooled embeddings are concatenated with a split of 3:1.

Training patches are extracted in size 256 × 256 and then resized to 112 × 112 to fit the S-CAE. 600 patches are drawn from each of the 20 randomly selected WSIs in the positive training set, negative training set, and validation set respectively. In order to better encode the patches with a rich distribution of nuclei, we ensure at least 90% of the training patches are drawn with all three channels' intensities being less than 70% of the maximum intensity, which represents samples from rich nucleus regions.

*3) DSNet Implementation Details:* Our DSNet is constructed as described in Sec. III-C, which is supervised by a single cross-entropy loss. Leaky ReLU is used as the activation function for convolutions.

For training configurations, the loss of DSNet is minimized through the gradient descent algorithm with AdamW optimizer [59]. The initial learning rate is set to $1e^{-4}$ with warm up from $1e^{-6}$ and following a step decay by a factor of 5 at every 30 epochs. We set weight decay to $1e^{-5}$. All models are trained for 200 epochs from scratch with early stopping at the epoch with the highest AUC score. We augment all training data by standard image flipping, rotating, and random cropping with discarding at most 7 pixels along the height or width.

## E. Benchmark on Camelyon16

Our method is first evaluated on the dataset from Camelyon16 [9] challenge for detection of cancer metastasis at WSI-level.

We compared with the fully-supervised method (Camelyon16 challenge winner) [8], state-of-the-art weakly-supervised methods through patch-wise approaches [29], [30] and image-wise approach [16]. We also conducted experiments on a naive image-wise approach, by using ResNet-50 to classify the down-sampled WSI thumbnail image directly. The fully-supervised method requires ground truth annotations to indicate patch-level labels during their training phase, while only the WSI-level label is used in our method during the whole training process.

We reproduced all weakly-supervised methods [16], [29], [30], [32] for fair comparison[1] except for the fully-supervised method [8] on Camelyon16, which we report the result from their paper directly. Note that all weakly-supervised methods claimed using standard networks for classification, and we keep the training configurations and classifier networks (ResNet-50 [50]) consistent across all weakly-supervised approaches for fair comparisons.

Table II reports the performances of the methods in terms of AUC scores and total number of network parameters.[2] It can be seen that the naive image-level weakly-supervised approach performs the worst, as large amount of visual information is lost when downsizing the WSIs. On the contrary, with the help of a previous encoding of raw WSIs, image-wise approaches are able to maintain useful diagnostic clues to a great extent and achieve superior performances than the patch-wise approaches. Particularly, our DSNet, with an advanced encoder and a dedicated classifier, eventually achieves an AUC of 0.760 that outperforms recent state-of-the-art weakly-supervised counterparts by a safe margin. Noteworthy, our proposed DSNet requires only 1.1 M trainable parameters, which is around 4% of the standard ResNet-50 classifier. The ROC curves of our method along with the state-of-the-art weakly supervised methods are plotted in Figure 6.

[1] [32] originally relies on extra annotations from the WSI experts, while its unavailable in our weakly-supervised settings.

[2] For image-wise approaches, patch encoders require extra parameters, which are 1.8M for [16] and 2.4M for our S-CAE.

TABLE II
AUC SCORES OF FULLY-SUPERVISED, PATCH-WISE, IMAGE-WISE WEAKLY-SUPERVISED METHODS OVER THE VALIDATION SETS OF CAMELYON16 (CAM16) AND TCGA-LUSC (TCGA). '-' DENOTES UNAVAILABLE RESULTS. CLASSIFIER PARAMETERS ARE REPORTED AS WELL

| Supervision | Methods | Cam16 | TCGA | #Params |
|---|---|---|---|---|
| Fully | Wang *et al.*. [8] | 0.925 | - | - |
| Patch-wise Weakly | Hou *et al.*. [57] | 0.666 | 0.973 | 25.6 M |
| | Chen *et al.*. [29] | 0.643 | 0.975 | 25.6 M |
| | VAE-GAN [32] | 0.695 | 0.976 | 32.5 M |
| Image-wise Weakly | Naive | 0.524 | 0.956 | 25.6 M |
| | Tellez *et al.*. [16] | 0.717 | 0.976 | 25.6 M |
| | **DSNet (ours)** | **0.760** | **0.986** | **1.1** M |

TABLE III
AUC SCORES OF THE BASELINE IMAGE-LEVEL METHOD, AND OUR DSNET OVER THE VALIDATION SET AND TESTING SET OF BCNB

| Supervision | Methods | Val. | Test. | #Params |
|---|---|---|---|---|
| Patch-wise Weakly | Xu *et al.* [20] | 0.808 | 0.816 | 4.1 M |
| Image-wise Weakly | Naive | 0.573 | 0.554 | 25.6 M |
| | DSNet (ours) | 0.797 | 0.803 | **1.1** M |

## F. Benchmark on TCGA-LUSC

Evaluation results are presented in Table II, the component study results are presented in Table IV, and the comparison ROC curves are plotted in Figure 7. Compared to other weakly-supervised counterparts, our method achieves the best performance in terms of the AUC scores.

## G. Benchmark on BCNB

We further evaluate our DSNet and the naive image-wise baseline on a larger dataset for classifying the metastatic status of auxiliary lymph node. The comparison results of the method proposed in [20] and the naive image-wise baseline are reported in Table III. With the minimum network size, our DSNet significantly surpasses the naive image-wise counterpart and achieves on par results compared to the state-of-the-art method on this dataset [20].

Although the patch-wise state-of-the-art method proposed in [20] yields slightly better results than ours, their method *(i)* used extra data (ImageNet) for model pre-training, while all of our models were trained using in-distribution data only; *(ii)* was built with nearly 4 times more trainable parameters.

TABLE IV
ABLATION STUDY ON THE CAMELYON16 VALIDATION SET AND TCGA-LUSC VALIDATION SET. WE DENOTE **SA** AS THE STREAM-WISE ATTENTION, **CA** THE CHANNEL-WISE ATTENTION

| Methods | Camelyon16 | | | | | TCGA-LUSC | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Accuracy | AUC | Precision | Recall | $F_1$-score | Accuracy | AUC | Precision | Recall | $F_1$-score |
| w/o sparsity | 67.8% | 0.635 | 0.209 | **0.750** | 0.327 | 89.0% | 0.977 | 0.860 | 0.915 | 0.887 |
| w/o embedding stream | 57.7% | 0.536 | 0.439 | 0.360 | 0.396 | 83.0% | 0.959 | 0.867 | 0.780 | 0.821 |
| w/o thumbnail stream | 60.9% | 0.686 | 0.209 | 0.450 | 0.286 | 89.0% | 0.965 | 0.950 | 0.780 | 0.857 |
| w/o multi-scale block | 59.1% | 0.670 | 0.674 | 0.467 | 0.552 | 85.0% | 0.974 | 0.780 | 0.907 | 0.839 |
| w/o concurrent bottlenecks | 66.0% | 0.656 | 0.395 | 0.515 | 0.447 | 89.0% | 0.968 | 0.840 | 0.933 | 0.884 |
| w/o spatial bottleneck | 60.0% | 0.696 | **0.721** | 0.477 | 0.574 | 82.0% | 0.964 | 0.680 | 0.944 | 0.791 |
| w/o **Sa** w/o **Ca** | 67.0% | 0.714 | 0.395 | 0.586 | 0.472 | 90.0% | 0.975 | 0.880 | 0.916 | 0.898 |
| w/o **Sa** w. **Ca** | 63.5% | 0.695 | 0.349 | 0.517 | 0.417 | 85.0% | 0.975 | 0.780 | 0.907 | 0.839 |
| w. **Sa** w/o **Ca** | 65.3% | 0.709 | 0.349 | 0.556 | 0.429 | 84.0% | 0.975 | 0.780 | **0.886** | 0.830 |
| **Full DSNet** | **69.6%** | **0.760** | 0.558 | 0.600 | **0.578** | **93.0%** | **0.986** | **1.000** | 0.877 | **0.935** |

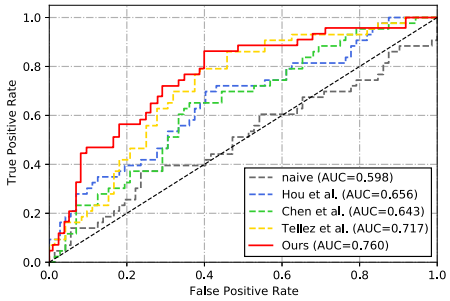


Fig. 6. ROC curves of comparisons to the state-of-the-art weakly-supervised methods on the Camelyon16 validation set.

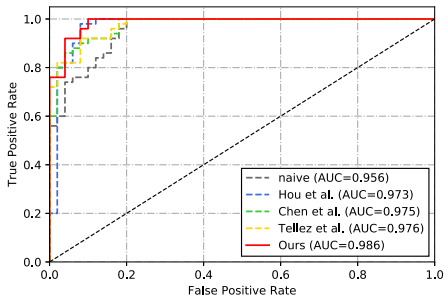


Fig. 7. ROC curves of comparisons to the state-of-the-art weakly-supervised methods on the TCGA-LUSC validation set.

## V. ANALYSIS

In this section, we provide in-depth analysis and additional ablation studies for a comprehensive understanding of the proposed method. Unless explicitly specified, the extensive experiments were conducted on the Camelyon16 dataset, with the identical training protocols introduced in Sec. IV-D.

### A. Complexity

*1) Space Complexity:* The hardware-unaffordable image size is the primary challenge to WSI analysis. Our method transforms the gigabyte original WSI to a latent space in a unsupervised style for reducing space complexity. Given the thumbnail downsampling factor K and the number of level L, the *thumbnail matrix* achieves a compression ratio of $\frac{K^2}{L}$. With the patch size S, the *embedding matrix* is able to achieve a compression ratio of $\frac{3S^2}{C}$. We then represent the original WSI by the two matrices, with the final compression ratio (w.r.t. # pixels) of $3S^2K^2/(CK^2 + (\sum_{l=0}^{L-1} 2^{2l}3)S^2)$. With our adopted setting (S=256, C=128, K=128, L=3), a compression ratio of 517 can be achieved.

*2) Time Complexity:* Time complexity is another essential consideration in our method development. The depth-wise design [52] was adopted as our basic convolutional operator to speed up network inference and lower the computation burden. When the thumbnail/embedding matrices are available, DSNet inference on a WSI with $10^9$ pixels takes about *44.1 ms* on a NVIDIA RTX 2080 GPU and *273.2 ms* on an Intel Core i7 CPU. Moreover, our entire pipeline, which includes: *(i)* WSI file loading from a hard disk; *(ii)* image preprocessing (patch dividing + patch sifting + matrices formation); and *(iii)* DSNet inference, processes a WSI with $10^9$ pixels in *126.1 s* on a CPU. The overall runtime can be reduced to *63.6 s* if a GPU is available. All running times were measured over an average of 10 WSIs at 40× magnification level.

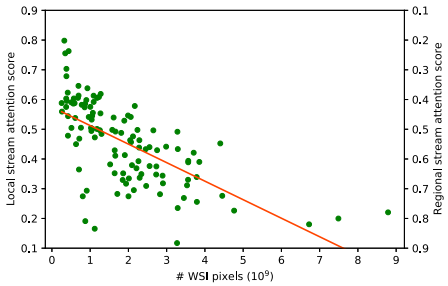


Fig. 8. Stream-wise attention score *v.s.* # WSI pixels on the Camelyon16 validation set. A linear fit on the attributes (red line) is shown.

### B. Ablation Studies

We examine the impact of individual components by removing each of them from the network. To assess the necessity of considering sparsity issues in WSIs, experiments are carried out by using dense operators only. Then, the entire embedding stream or thumbnail stream is removed from DSNet to verify whether the local or regional information is useful to the metastasis diagnosis. Subsequently, we examined the effective-

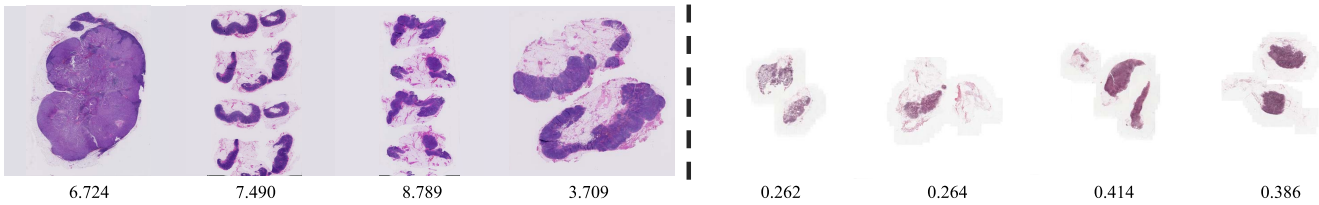


Fig. 9. **Visualization of local and regional attentive WSI examples**. Slide areas ($\times 10^9$ pixel) are shown in this figure. The WSIs on the left are paid more attention on regional information (<30% attention scores are assigned to local stream) and the right ones are paid more attention on local information (>80% attention scores are assigned to local stream).

TABLE V
COMPARISON OF USING DIFFERENT LEVELS OF THUMBNAIL IMAGES ON THE BASIS OF DOWNSAMPLING RATIO K=128

| $\frac{K}{2^\alpha}$ | Accuracy | AUC | Precision | Recall | Comp ratio |
|---|---|---|---|---|---|
| $\alpha$=[0] | 67.8% | 0.736 | 0.558 | 0.571 | **1404** |
| $\alpha$=[0,1] | 67.0% | 0.759 | **0.604** | 0.553 | 1046 |
| $\alpha$=[0,1,2] | **69.6%** | **0.760** | 0.558 | 0.600 | 517 |
| $\alpha$=[0,1,2,3] | 68.7% | 0.755 | 0.535 | 0.590 | 171 |
| $\alpha$=[0,1,2,3,4] | 67.0% | 0.737 | 0.233 | **0.714** | 56 |

ness of the multi-scale block by simply replacing it with $3 \times 3$ convolution only, and aligned the total network parameters. We further assess the impact of the concurrent bottleneck block and adaptive aggregation block separately. The ablation results are shown in Table IV. Without considering the sparsity issue, the network can hardly detect metastasis and tends to make predictions toward the negative class. The designed concurrent bottleneck block also yields a significant impact on our network by raising the AUC score for more than 10%. Surprisingly, we found that even without adaptive aggregation, our method can still achieve a competitive result at 0.714 AUC. However, the performance of DSNet could be harmed when only one attention operator is used. Our hypothesis is that the features suppressed in one perspective (e.g. channel-wise significance) could be more indicative in another perspective (e.g. stream-wise significance) [45]. Therefore, a single measurement of the feature significance in our adaptive aggregation module could degenerate feature representation. When integrating all the components into our network, the full DSNet reaches the state-of-the-art performance at 69.6% classification accuracy and 0.760 AUC.

### C. Level of Thumbnail Images

The thumbnail matrix of a WSI is constructed as a stack of thumbnail images captured at different magnified levels. We study the impact of using thumbnails at different magnifications as input to our DSNet. In Table V, we achieve 0.736 AUC by only using the image at the highest magnification on a downsampling basis of K=128, which already surpasses the image-wise state-of-the-art counterpart [16]. Compromising on the compression ratio, our method reaches the highest AUC score by using thumbnail images from three different levels.

### D. Patch Encoding Strategy

Here, we evaluate different patch encoding strategies by using foreground encoding only, background encoding only, and the mix encoding without distinguishing foreground or background. Table VI shows that using both separated foreground and background encodings yields the best performance at nearly all metrics. Similar to [16], by using a mixed encoding, we can also obtain a good result.

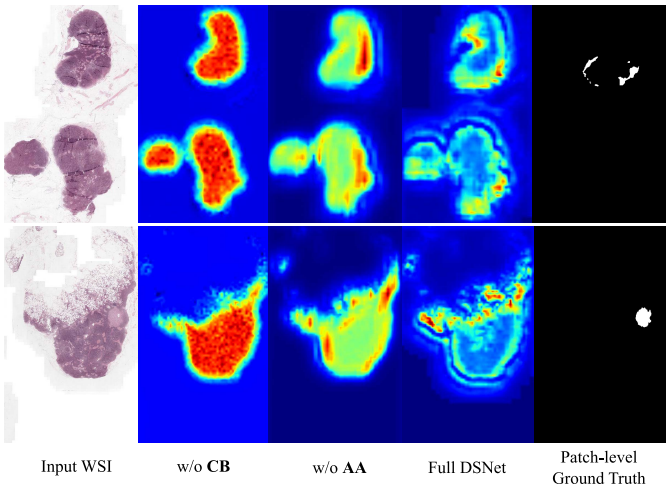


Fig. 10. **Grad-CAM visualizations of our *weakly-supervised* DSNet**. **AA** denotes our Adaptive Aggregation operator and **CB** denotes our Concurrent Bottleneck design.

TABLE VI
COMPARISON OF DIFFERENT PATCH ENCODING STRATEGIES

| Encoding | Accuracy | AUC | Precision | Recall | $F_1$-score |
|---|---|---|---|---|---|
| Foreground | 57.4% | 0.650 | 0.302 | 0.406 | 0.347 |
| Background | 58.3% | 0.585 | 0.233 | 0.400 | 0.294 |
| Mixed | 67.0% | 0.696 | 0.233 | **0.667** | 0.345 |
| Separated | **69.6%** | **0.760** | **0.558** | 0.600 | **0.578** |

### E. Stream Importance

We then analyze the importance of the two streams regrading of the learned stream-wise attention scores. The stream-wise attention score distribution on the Camelyon16 validation set with respect to the WSI size is shown in Figure 8. As the figure illustrates, our DSNet is more favorable to assign higher attention score to the local stream (embedding matrix stream) when the size of input WSI is small and to the regional stream (thumbnail matrix stream) when the size is large. To understand whether our DSNet can demonstrate certain correlations between WSI size and the attention score, we first calculate an absolute Pearsons correlation coefficient of *0.656* and fit

a linear regression model on the two attributes (Figure 8). We further visualize some instances with most local attention scores and most regional attention scores in Figure 9. From the Pearsons coefficient, the fitted linear model, and individual instances, we found that our DSNet indeed has the ability to assign different stream-wise attention scores regarding of WSI slide areas. Interestingly, pathologists are supposed to pay more attention to regional clues when diagnosing larger WSIs and focus more on local details for smaller WSIs, which is consistent with our findings.

### F. Grad-CAM Visualization

Grad-CAM [60] is a commonly used algorithm to visualize the region of network interest toward the prediction of a target class. We use Grad-CAM to highlight the area that our DSNet determines as evidences to diagnose a WSI as containing metastasis. The gradients are collected at the last $1 \times 1$ convolutional layer towards the positive class. In Figure 10, we present the comparison results by removing different components from the network to inspect the changes of network focus. When removing the concurrent bottleneck, the base network can hardly discriminate local semantics, results in paying attention to all valid regions. Without patch-level labels, our weakly-supervised method is also able to indicate reasonable regions that may show potential cancer metastasis.

## VI. Conclusion

We propose a novel framework for classifying whole slide images in a weakly-supervised manner to alleviate the burden of expert annotation that is required in fully supervised approaches. Our method consists of two sequential stages, which utilize an efficient neural network to first roughly skim through a WSI thumbnail and then zoom into regions of interest for more descriptive details. Specifically, our framework transforms the raw WSI in an unsupervised manner into a local embedding matrix and a regional thumbnail matrix, and then employs the proposed DSNet to adaptively integrate the transformed matrices to infer the WSI-level label through two separated streams consisting of stacks of efficient yet effective building blocks. We demonstrate that our image-level weakly-supervised approach surpasses recent state-of-the-art methods on three large-scale public benchmark datasets with the least network parameters.

Our proposed weakly-supervised approach processes gigapixel WSIs at their compact representations with minimum computational costs and hardware burdens. With such highly efficient two stream design, regional guidances implied in thumbnail images could be useful in all kinds of WSI analysis applications. Therefore, we believe the same framework can be easily extended to many other tasks such as metastasis segmentation and anomaly detection in future works.

## References


[1] H. Tokunaga, Y. Teramoto, A. Yoshizawa, and R. Bise, "Adaptive weighting multi-field-of-view CNN for semantic segmentation in pathology," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 12597–12606.

[2] L. Li, M. Xu, X. Wang, L. Jiang, and H. Liu, "Attention based glaucoma detection: A large-scale database and CNN model," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 10571–10580.

[3] S. Maksoud, K. Zhao, P. Hobson, A. Jennings, and B. C. Lovell, "SOS: Selective objective switch for rapid immunofluorescence whole slide image classification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3862–3871.

[4] T. Xiang, C. Zhang, D. Liu, Y. Song, H. Huang, and W. Cai, "BiO-Net: Learning recurrent bi-directional connections for encoder-decoder architecture," in *Proc. Int. Conf. Med. Image Comput.-Assist. Intervent.* Lima, Peru: Springer, 2020, pp. 74–84.

[5] Y. Song, Q. Li, H. Huang, D. Feng, M. Chen, and W. Cai, "Low dimensional representation of Fisher vectors for microscopy image classification," *IEEE Trans. Med. Imag.*, vol. 36, no. 8, pp. 1636–1649, Aug. 2017.

[6] X. Wang *et al.*, "BiX-NAS: Searching efficient bi-directional architecture for medical image segmentation," in *Proc. Int. Conf. Med. Image Comput.-Assist. Intervent.* Strasbourg, France: Springer, 2021.

[7] S. Wang *et al.*, "RMDL: Recalibrated multi-instance deep learning for whole slide gastric image classification," *Med. Image Anal.*, vol. 58, Dec. 2019, Art. no. 101549.

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

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

[10] T. Araújo *et al.*, "Classification of breast cancer histology images using convolutional neural networks," *PLoS ONE*, vol. 12, no. 6, Jun. 2017, Art. no. e0177544.

[11] D. Liu *et al.*, "PDAM: A panoptic-level feature alignment framework for unsupervised domain adaptive instance segmentation in microscopy images," *IEEE Trans. Med. Imag.*, vol. 40, no. 1, pp. 154–165, Jan. 2021.

[12] Y. Zheng *et al.*, "Histopathological whole slide image analysis using context-based CBIR," *IEEE Trans. Med. Imag.*, vol. 37, no. 7, pp. 1641–1652, Jul. 2018.

[13] C. Mercan, S. Aksoy, E. Mercan, L. G. Shapiro, D. L. Weaver, and J. G. Elmore, "Multi-instance multi-label learning for multi-class classification of whole slide breast histopathology images," *IEEE Trans. Med. Imag.*, vol. 37, no. 1, pp. 316–325, Jan. 2017.

[14] F. Wang *et al.*, "Residual attention network for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3156–3164.

[15] A. Vaswani *et al.*, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[16] D. Tellez, G. Litjens, J. van der Laak, and F. Ciompi, "Neural image compression for gigapixel histopathology image analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 2, pp. 567–578, Feb. 2021.

[17] W. Chen, Z. Jiang, Z. Wang, K. Cui, and X. Qian, "Collaborative global-local networks for memory-efficient segmentation of ultra-high resolution images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 8924–8933.

[18] L. Hou *et al.*, "Sparse autoencoder for unsupervised nucleus detection and representation in histopathology images," *Pattern Recognit.*, vol. 86, pp. 188–200, Feb. 2019.

[19] *T. Cancer Genome*. Accessed: Oct. 2020. [Online]. Available: https://atlashttps://tcga-data.nci.nih.gov/tcga/

[20] F. Xu *et al.*, "Predicting axillary lymph node metastasis in early breast cancer using deep learning on primary tumor biopsy slides," *Frontiers Oncol.*, vol. 11, p. 4133, Oct. 2021.

[21] S. Sankaran *et al.*, "Low-altitude, high-resolution aerial imaging systems for row and field crop phenotyping: A review," *Eur. J. Agronomy*, vol. 70, pp. 112–123, Oct. 2015.

[22] C. Yuan, Y. Zhang, and Z. Liu, "A survey on technologies for automatic forest fire monitoring, detection, and fighting using unmanned aerial vehicles and remote sensing techniques," *Can. J. Forest Res.*, vol. 45, no. 7, pp. 783–792, 2015.

[23] X. Wang *et al.*, "Weakly supervised learning for whole slide lung cancer image classification," in *Proc. Med. Imag. Deep Learn.*, 2018, pp. 1–10.

[24] H. Lin, H. Chen, S. Graham, Q. Dou, N. Rajpoot, and P.-A. Heng, "Fast ScanNet: Fast and dense analysis of multi-gigapixel whole-slide images for cancer metastasis detection," *IEEE Trans. Med. Imag.*, vol. 38, no. 8, pp. 1948–1958, Aug. 2019.

[25] M. S. Hosseini, J. A. Z. Brawley-Hayes, Y. Zhang, L. Chan, K. Plataniotis, and S. Damaskinos, "Focus quality assessment of high-throughput whole slide imaging in digital pathology," *IEEE Trans. Med. Imag.*, vol. 39, no. 1, pp. 62–74, Jan. 2020.

[26] P. Chikontwe, M. Kim, S. J. Nam, H. Go, and S. H. Park, "Multiple instance learning with center embeddings for histopathology classification," in *Proc. Int. Conf. Med. Image Comput.-Assist. Intervent.* Lima, Peru: Springer, 2020, pp. 519–528.

[27] T. Vu, P. Lai, R. Raich, A. Pham, X. Z. Fern, and U. A. Rao, "A novel attribute-based symmetric multiple instance learning for histopathological image analysis," *IEEE Trans. Med. Imag.*, vol. 39, no. 10, pp. 3125–3136, Oct. 2020.

[28] I. Goodfellow *et al.*, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2014, pp. 2672–2680.

[29] H. Chen *et al.*, "Rectified cross-entropy and upper transition loss for weakly supervised whole slide image classifier," in *Proc. Int. Conf. Med. Image Comput.-Assist. Intervent.* Shenzhen, China: Springer, 2019, pp. 351–359.

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

[31] N. Hashimoto *et al.*, "Multi-scale domain-adversarial multiple-instance CNN for cancer subtype classification with unannotated histopathological images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3852–3861.

[32] Y. Zhao *et al.*, "Predicting lymph node metastasis using histopathological images based on multiple instance learning with deep graph convolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 4837–4846.

[33] M. Y. Lu *et al.*, "AI-based pathology predicts origins for cancers of unknown primary," *Nature*, vol. 594, no. 7861, pp. 106–110, Jun. 2021.

[34] L. Itti and C. Koch, "Computational modelling of visual attention," *Nature Rev. Neurosci.*, vol. 2, no. 3, p. 194, 2001.

[35] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 11, pp. 1254–1259, Nov. 1998.

[36] B. A. Olshausen, C. H. Anderson, and D. C. Van Essen, "A neurobiological model of visual attention and invariant pattern recognition based on dynamic routing of information," *J. Neurosci.*, vol. 13, no. 11, pp. 4700–4719, Nov. 1993.

[37] M. Chen, *Computer Vision for Microscopy Image Analysis*. New York, NY, USA: Academic, 2020.

[38] Y. Chen, X. Dai, M. Liu, D. Chen, L. Yuan, and Z. Liu, "Dynamic convolution: Attention over convolution kernels," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 11030–11039.

[39] D. Zoran, M. Chrzanowski, P.-S. Huang, S. Gowal, A. Mott, and P. Kohli, "Towards robust image classification using sequential attention models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9483–9492.

[40] X. Hu, Z. Zhang, Z. Jiang, S. Chaudhuri, Z. Yang, and R. Nevatia, "Span: Spatial pyramid attention network for image manipulation localization," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2020, pp. 312–328.

[41] Y. A. Mejjati, C. F. Gomez, K. I. Kim, E. Shechtman, and Z. Bylinskii, "Look here! A parametric learning based approach to redirect visual attention," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2020, pp. 343–361.

[42] H. Zhang *et al.*, "ResNeSt: Split-attention networks," 2020, *arXiv:2004.08955*.

[43] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 7132–7141.

[44] A. G. Roy, N. Navab, and C. Wachinger, "Concurrent spatial and channel 'squeeze & excitation' in fully convolutional networks," in *Proc. Int. Conf. Med. Image Comput.-Assist. Intervent.* Granada, Spain: Springer, 2018, pp. 421–429.

[45] S. Woo, J. Park, J.-Y. Lee, and I. So Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 3–19.

[46] Z. Gao, J. Xie, Q. Wang, and P. Li, "Global second-order pooling convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 3024–3033.

[47] X. Li, W. Wang, X. Hu, and J. Yang, "Selective kernel networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 510–519.

[48] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "ECA-Net: Efficient channel attention for deep convolutional neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 11534–11542.

[49] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn.*, vol. 37, Jul. 2015, pp. 448–456. [Online]. Available: http://dl.acm.org/citation.cfm?id=3045118.3045167

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

[51] K. Han, Y. Wang, Q. Tian, J. Guo, C. Xu, and C. Xu, "GhostNet: More features from cheap operations," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 1580–1589.

[52] A. G. Howard *et al.*, "MobileNets: Efficient convolutional neural networks for mobile vision applications," 2017, *arXiv:1704.04861*.

[53] J. Uhrig, N. Schneider, L. Schneider, U. Franke, T. Brox, and A. Geiger, "Sparsity invariant CNNs," in *Proc. Int. Conf. 3D Vis. (3DV)*, Oct. 2017, pp. 11–20.

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

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

[56] A. H. Robinson and C. Cherry, "Results of a prototype television bandwidth compression scheme," *Proc. IEEE*, vol. 55, no. 3, pp. 356–364, Mar. 1967.

[57] L. Hou, D. Samaras, T. M. Kurc, Y. Gao, J. E. Davis, and J. H. Saltz, "Patch-based convolutional neural network for whole slide tissue image classification," 2015, *arXiv:1504.07947*.

[58] F. Chollet *et al.* (2015). *Keras*. [Online]. Available: https://keras.io

[59] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proc. Int. Conf. Learn. Represent.*, 2018, pp. 1–19.

[60] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 618–626.