# Learning to Segment Breast Biopsy Whole Slide Images

Sachin Mehta*[1], Ezgi Mercan*[1], Jamen Bartlett[2], Donald Weaver[2], Joann Elmore[1], and Linda Shapiro[1]

[1]University of Washington, Seattle, WA, USA
Email: {sacmehta, jelmore}@uw.edu, {ezgi, shapiro}@cs.washington.edu
[2]University of Vermont, Burlington, VT, USA
Email: {jamen.bartlett, donald.weaver}@uvmhealth.org

## Abstract

*We trained and applied an encoder-decoder model to semantically segment breast biopsy images into biologically meaningful tissue labels. Since conventional encoder-decoder networks cannot be applied directly on large biopsy images and the different sized structures in biopsies present novel challenges, we propose four modifications: (1) an input-aware encoding block to compensate for information loss, (2) a new dense connection pattern between encoder and decoder, (3) dense and sparse decoders to combine multi-level features, (4) a multi-resolution network that fuses the results of encoder-decoders run on different resolutions. Our model outperforms a feature-based approach and conventional encoder-decoders from the literature. We use semantic segmentations produced with our model in an automated diagnosis task and obtain higher accuracies than a baseline approach that employs an SVM for feature-based segmentation, both using the same segmentation-based diagnostic features.*

## 1. Introduction

Breast cancer is traditionally diagnosed with histopathological interpretation of the biopsy samples on glass slides by pathologists. Whole slide imaging (WSI) is a technology that captures the contents of glass slides in a multi-resolution image. With the developments in whole slide imaging, it is now possible to develop computer-aided diagnostic tools that support the decision-making process of medical experts. Until recently, the use of WSIs was limited to non-clinical purposes such as research, education, obtaining second opinions, and archiving, but they have been approved for diagnostic use in the US starting April 2017 [2].

Automated cancer detection from digital slides is a well-studied task in the computer vision community [10] and several image datasets have been developed for malignant tumors [3, 39, 4]; however, little work exists in differentiating the full spectrum of breast lesions from benign to pre-invasive lesions, and to invasive cancer [11]. Pre-invasive lesions presents a more difficult classification scenario than the binary classification task of invasive cancer detection. It requires careful analysis of epithelial structures in the breast biopsy images. In this paper, we propose a state-of-the-art semantic segmentation system to produce a tissue label image (Figure 1) for the WSIs of breast biopsies that can lead to an automated diagnosis system.

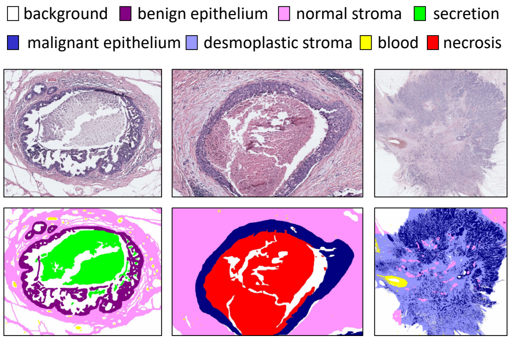


**Figure 1:** The set of tissue labels used in semantic segmentation: (top row) three example cases from the dataset and (bottom row) the pixel labels provided by a pathologist. Best viewed in color.

Our system builds on the encoder-decoder networks that are the state-of-the-art approaches for semantic segmentation. However, conventional architectures are not directly applicable to whole slide breast biopsy images with dimensions in gigapixels. A sliding window approach to crop fixed-sized images from WSIs is promising [22], but dividing the large structures limits the context available to convolutional neural networks (CNNs) and affect the segmentation performance. Unlike general image datasets (e.g. [26, 14, 37]), breast biopsy images have objects of interest in varied sizes. For some WSIs, the diagnosis is made while

*Contributed equally.

looking at the whole image, while others require the detection of a small structure at high resolutions. Simply using a sliding window with a constant size causes loss of information available at different resolutions.

This paper proposes a new multi-resolution encoder-decoder architecture that was specifically designed to handle the challenges of the breast biopsy semantic segmentation problem. The architecture is described in detail, and a rigorous set of experiments is applied to compare its segmentation performance to multiple different other models. Finally, the network is used in a set of diagnostic classification experiments that further show its benefits.

## 2. Related Work

Following the success of CNNs in image classification tasks [38, 36, 21], they have been extended for dense prediction tasks such as semantic segmentation [35, 31, 6]. Unlike object proposal-based methods [17, 18], fully convolutional networks (FCN) have enabled end-to-end training and have shown efficient feature learning. These methods are widely used for segmenting both natural [35, 31, 6, 8] and medical images [33, 15, 30, 42].

FCN-based networks generate coarse segmentation masks and several techniques have been proposed to address this limitation such as skip-connections [35, 33, 15], atrous/dilated convolutions [8, 41], deconvolutional networks [31, 6, 15, 33, 7], and multiple input networks (e.g. different scales [9, 43, 27] or streams [16]). These methods process the input sources either independently [8, 9, 16, 25] or recursively [32, 12]; thus exploit the features from multiple levels to refine the segmentation masks. Additionally, conditional random fields (CRFs) have been used to further refine the segmentation results [44, 8, 41].

Several CNN-based methods have been applied for segmenting medical images (e.g. EM [33], brain [15], gland [7], and 3D MR [42] images). Yet, segmenting breast biopsy images, with a full range of diagnosis from benign to invasive, still remains a challenge. Our approach applies previous work on encoder-decoders (e.g. [6, 15]) and improves upon them with carefully designed components that address their limitations on WSI applications.

## 3. Breast Biopsy Dataset

Our dataset contains 240 breast biopsies selected from the Breast Cancer Surveillance Consortium [1] affiliated archives in New Hampshire and Vermont. The cases span a wide range of diagnoses that mapped to four diagnostic categories: benign, atypia, ductal carcinoma *in-situ* (DCIS), and invasive cancer. The original H&E (heamatoxylin and eosin) stained glass slides were scanned using an iScan CoreoAu® in 40× magnification. A technician and an experienced breast pathologist reviewed each digital image, rescanning as needed to obtain the highest quality. The average image size for the 240 WSIs was $90,000 \times 70,000$ pixels.

| **Diagnostic Category** | **#ROI** (training) | **#ROI** (test) | **#ROI** (total) | **Avg. size** (pixels) |
|---|---|---|---|---|
| Benign | 4 | 5 | 9 | $9K \times 9K$ |
| Atypia | 11 | 11 | 22 | $6K \times 7K$ |
| DCIS | 12 | 10 | 22 | $8K \times 10K$ |
| Invasive | 3 | 2 | 5 | $38K \times 44K$ |
| **Total** | **30** | **28** | **58** | $10K \times 12K$ |

**Table 1:** Distribution of diagnostic categories and average image sizes from the segmentation subset.

All 240 digital slides were interpreted by an expert panel of three pathologists to produce an expert consensus diagnosis for each case. Experts also provided one or more regions of interest (ROIs) supporting the expert consensus diagnosis on each WSI. Since some cases had more than one ROI per WSI, the final set includes 102 benign, 128 atypia, 162 DCIS and 36 invasive ROIs.

To describe the structural changes that lead to cancer in the breast tissue, we produced a set of eight tissue labels in collaboration with an expert pathologist: (1) *benign epithelium*: the epithelial cells in the benign and atypia categories, (2) *malignant epithelium*: the bigger and more irregular epithelial cells from the DCIS and invasive cancer categories, (3) *normal stroma*: the connective tissue between the regular ductal structures in the breast, (4) *desmoplastic stroma*: proliferated stromal cells associated with tumor, (5) *secretion*: benign substance secreted from the ducts, (6) *necrosis* the dead cells at the center of the ducts in the DCIS and invasive cases, (7) *blood*: the blood cells, which are rare but have a very distinct appearance, and (8) *background*: the pixels that do not contain any tissue.

Although some labels are not critical for diagnosis, our tissue label set was intended to cover all the pixels in the images. Due to the expertise needed for labeling and the size of the biopsy images, we randomly selected a subset of 40 cases (58 ROIs) to be annotated by a pathologist. Table 1 summarizes the distribution of four diagnostic categories in training and test sets as well as average image sizes. Figure 1 shows three example images along with their pixel-wise labels provided by the pathologist.

## 4. Background

Encoder-decoder networks are state-of-the-art networks for segmenting 2D (e.g. [33, 15]) as well as 3D (e.g. [30, 42]) medical images. In a conventional encoder, the transition between two subsequent encoding blocks, $l^{th}$ and $(l+1)^{th}$, can be formulated as [26, 36]: $\mathbf{x}_e^{l+1} = \mathcal{F}_e(\mathbf{x}_e^l)$. In a class of encoder networks, called residual networks, the input and output of the $l^{th}$ block are combined to improve the gradient flow [21]:

$$\mathbf{x}_e^{l+1} = \mathcal{F}_e(\mathbf{x}_e^l) + \mathbf{x}_e^l \tag{1}$$

where $\mathcal{F}_e(\mathbf{x}_e^l)$ is a function comprising two $3 \times 3$ convolution operations. This block is referred as a Residual Convolutional Unit (RCU) (Figure 2a).

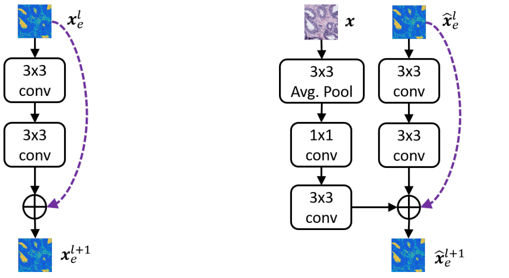
In a conventional decoder (Figure 3a), the transition between two subsequent decoding blocks, $l^{th}$ and $(l+1)^{th}$, can be formulated as [35, 8, 31, 6]: $\mathbf{x}_d^l = \mathcal{F}_d(\mathbf{x}_d^{l+1})$. To improve the gradient flow between the encoder and the decoder, the output of the $l^{th}$ encoding block and the corresponding decoding block can be combined as [15, 33]:

$$\check{\mathbf{x}}_d^l = \mathbf{x}_e^l + \mathcal{F}_d(\mathbf{x}_d^{l+1}) \tag{2}$$

where $\mathcal{F}_d(\mathbf{x}_d^l)$ is a decoding function that performs a $3 \times 3$ deconvolution operation. Such an encoder-decoder network with skip-connections between encoding and decoding blocks is called a *residual encoder-decoder* (Figure 3b). The deconvolution operation: (1) up-samples the feature maps, and (2) reduces the dimensionality of the feature maps. Note that the deconvolutional filters are capable of learning the non-linear up-sampling operations [35].

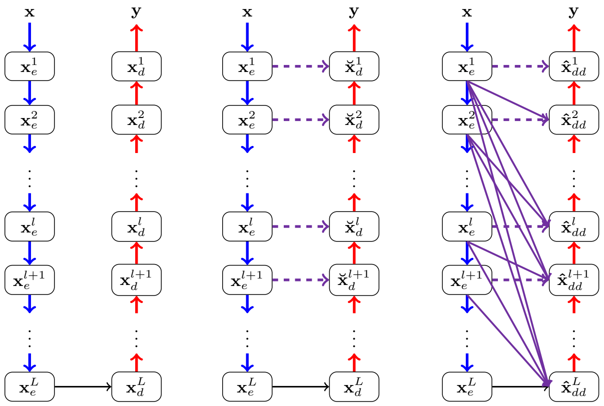
## 5. Proposed Encoder-Decoder Network

We propose a new encoder-decoder architecture to address the challenges that semantic segmentation of breast biopsies presents. Our network incorporates four new features: (1) input-aware encoding blocks (IA-RCU) that reinforces the input inside the encoder to compensate the loss of information due to down-sampling operations, (2) a densely connected decoding network and (3) an additional sparsely connected decoding network to efficiently combine the multi-level features aggregated by the encoder, and (4) a multi-resolution network for context-aware learning, which combines the output of different resolutions using a densely connected fusion block. Our network makes use of long-range skip-connections with identity and projection mappings in the encoder, the decoder, and fusion block to efficiently back-propagate the information to the input source and prevent the vanishing gradients; thereby helps train our network efficiently end-to-end. An overview of our network is illustrated in Figure 4 with details below.



**(a)** RCU [21] **(b)** Our Input-Aware RCU

**Figure 2:** Different type of encoding blocks: (a) residual convolutional unit (RCU) and (b) the proposed input-aware residual convolutional unit (IA-RCU).



**(a)** Plain [6, 31, 40] **(b)** Residual [33, 15] **(c)** Ours densely connected

**Figure 3:** (a, b) Conventional and (c) ours densely connected encoder-decoder networks with $L$ encoding and decoding blocks. These networks take an input $\mathbf{x}$ and generates an output $\mathbf{y}$. Here, $\dashrightarrow$ and $\longrightarrow$ represents residual and dense links between the encoder and the decoder.

### 5.1. Input-aware encoding blocks (IA-RCU)

The down-sampling operations in the encoder result in a loss of spatial information. To compensate the loss of spatial information, we introduce an input-aware encoding block (IA-RCU) that reinforces the input image at different levels of the encoder for better encoding of the spatial relationships and learned features. The IA-RCU, sketched in Figure 2b, introduces an additional path which can be viewed as a different connectivity pattern that establishes a direct link between an input image and any encoding stage, making each encoding block *aware of the input image*; thereby allowing gradients to flow back directly to the input paths. Additionally, the IA-RCU allows the encoding blocks to learn the features relevant to the input. The IA-RCU can be mathematically defined as:

$$\hat{\mathbf{x}}_e^{l+1} = \mathbf{x}_e^{l+1} + \mathcal{F}_{IA}(\mathbf{x}) \tag{3}$$

where $\mathcal{F}_{IA}(\mathbf{x})$ represents an input-aware mapping to be learned. $\mathcal{F}_{IA}$ is a composite function comprising a $3 \times 3$ average pooling operation that sub-samples the input image $\mathbf{x}$ to the same size as the encoding block $\mathbf{x}_e^{l+1}$, followed by $1 \times 1$ and $3 \times 3$ convolution operations that first projects the sub-sampled image to the same vector space as the encoding block $\mathbf{x}_e^{l+1}$ (Eq. 1) and then computes the dense features.

### 5.2. Densely Connected Decoding Blocks

Unlike a plain encoder-decoder network (Figure 3a), the skip-connections in the residual encoder-decoder network (Figure 3b) establishes a direct link between the encoding block and corresponding decoding block, which helps to improve the information flow. To further improve the information flow, we introduce direct connections between a decoding block and all encoding blocks that are at the same or lower level (Figure 3c). The $l^{th}$ decoding block receives the output feature maps from encoding blocks 1 to $l$. Dense connections can be defined as a modification to Eq. 2:

$$\hat{\mathbf{x}}_{dd}^l = \mathcal{F}_d(\mathbf{x}_d^{l+1}) + \sum_{i=1}^{l} \mathcal{F}_D(\hat{\mathbf{x}}_e^i) \tag{4}$$

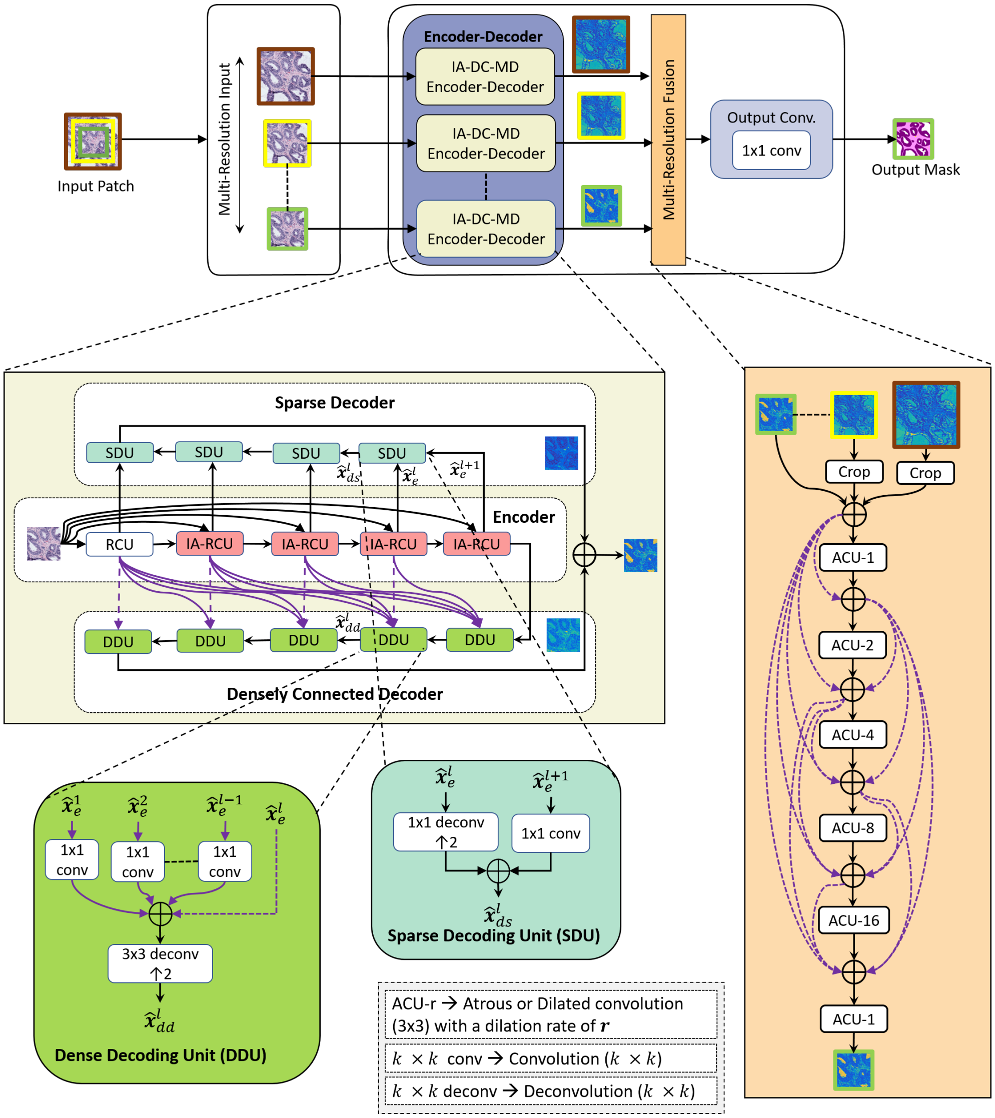


**Figure 4:** Our multi-resolution encoder-decoder network that incorporates input-aware encoding blocks, sparse and densely connected decoding networks, and densely connected fusion block. Different components in our architecture makes use of identity and projection mappings; thereby helping in back-propagating the information directly to the input paths efficiently. ⟶ and - - → links denotes the identity and projection links. The number of channels at different levels of encoder, densely connected decoder, and sparse decoder follow the following sequences: $64 \rightarrow 64 \rightarrow 128 \rightarrow 256 \rightarrow 512$, $256 \rightarrow 128 \rightarrow 64 \rightarrow 64 \rightarrow C$, and $C \rightarrow C \rightarrow C \rightarrow C$. Best viewed in color.

$\mathcal{F}_D(\hat{\mathbf{x}}^i_e)$ is the dense connection mapping to be learned. $\mathcal{F}_D$ consists of a $1 \times 1$ convolution operation, which projects the feature maps of the $i^{th}$ encoding block $\hat{\mathbf{x}}^i_e$ to the same vector space as $\mathbf{x}^l_d$.

## 5.3. Multiple Decoding Paths

For a given input image $\mathbf{x}$, we aim to efficiently combine the low- and mid-level features of the encoding network with high-level features to generate a pixel-level semantic segmentation mask. To do so, we must invert the loss of resolution from down-sampling. Using previous work [31, 6, 33, 15], we augment the encoder network with the bottom-up refinement approach. We introduce two decoding networks, densely connected and sparse, that decode the encoded input into a $C$-dimensional output, where $C$ represents the number of classes in the dataset. Figure 4 shows our network with multiple decoding paths.

The densely connected decoder stacks the densely connected decoding blocks, defined in Eq. 4, to decode the encoded feature maps into $C$-dimensional space. Because of the dense connections between the encoder and the decoder, we call this decoder a *densely connected decoder*. The sparse decoder projects the high-dimensional feature maps of each encoding block into $C$-dimensional vector spaces, which are then combined using a bottom-up approach. A sparse decoding function $\mathcal{F}_S$ can be formulated as:

$$\hat{\mathbf{x}}_{ds}^l = \mathcal{F}_S(\{\hat{\mathbf{x}}_e^l, \hat{\mathbf{x}}_e^{l+1}\}) \tag{5}$$

$\mathcal{F}_S(\{\hat{\mathbf{x}}_e^l, \hat{\mathbf{x}}_e^{l+1}\})$ is a function consisting of $1 \times 1$ deconvolutional and convolutional operations that projects high-dimensional encoder feature maps to $C$-dimensional vector space. Additionally, deconvolution operation up-samples the feature maps of $\hat{\mathbf{x}}_e^{l+1}$ to the same size as $\hat{\mathbf{x}}_e^l$. Because of the $1 \times 1$ convolution/deconvolutional operations involved, we call this decoder a *sparse decoder*.

### 5.4. Multiple Resolution Input

A sliding-window approach is promising for segmenting large biopsy images, however, the size of the patch determines the context available to the CNN model. Such an approach divides the bigger structures into smaller patches and may hurt the performance of the CNN method, especially at the border of the patch. To make the CNN model aware of the surrounding information, we introduce a multi-resolution network, which consists of the composition of $P$ instances of the encoder-decoder network (Figure 4). The $p^{th}$ instance takes the input patch $\mathbf{x}_p$ and generates the $C$-dimensional output $\mathbf{y}_p$. The spatial dimensions of each instance are different. A cropping function $\mathcal{F}_{Cr}(\mathbf{y}_p)$ takes the output of the $p^{th}$ instance and centrally crops it to produce the output $\hat{\mathbf{y}}_p$, which has the same dimensions as $\mathbf{y}_P$. After cropping, a multi-resolution fusion function $\mathcal{F}_{Mr}(\{\hat{\mathbf{y}}_1, \cdots, \hat{\mathbf{y}}_{P-1}, \mathbf{y}_P\})$ is applied to fuse the output of these $P$ network instances to produce the output $\mathbf{y}^{pred}$.

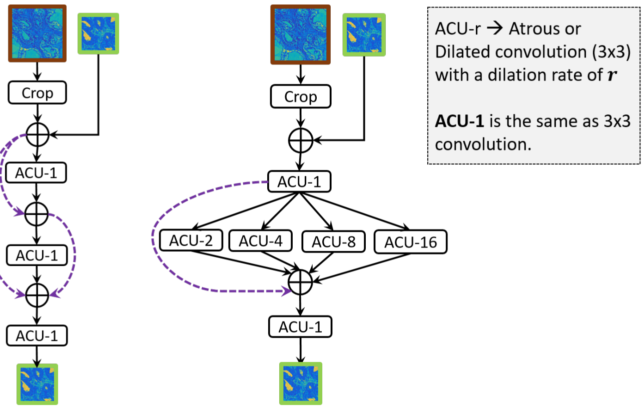
The multi-resolution fusion function $\mathcal{F}_{Mr}$, visualized in Figure 4, first combines the $P$ instances using an element-wise sum operation and then extracts the dense features using a stack of $3 \times 3$ dilated or atrous convolution operations with different dilation rates $r$. A traditional context module [41] may suffer from degradation problem and impede the information flow. Following Huang *et al.*[23], we introduce direct identity mappings from any layer to its subsequent layers to improve the information flow in the fusion block. We combine the output of any layer with the preceding layers using an element-wise sum operation.

## 6. Experiments and Results

To evaluate each proposed mechanism, we trained and tested eight encoder-decoder networks as summarized in Table 2. We compared our model to two conventional models: a plain encoder-decoder network [6] (Figure 3a) and a residual encoder-decoder [15] (Figure 3b). Then, we ran ablation studies by removing IA-RCU blocks (A1), multiple decoders (A2), and both IA-RCU blocks and multiple decoders (A3). We ran all models with a single encoder-decoder network using single resolution input and with multiple encoder decoders using multiple resolution inputs. Finally, to compare with our fusion approach for multi-resolution inputs, we implemented two alternative fusion methods (Figure 5): Fusion-A, with a standard stack of convolutional blocks, and Fusion-B, with a spatial pyramid pooling method using atrous or dilated convolutions [8]. We used two resolutions in multi-resolution models but our network can be easily extended to many resolutions.

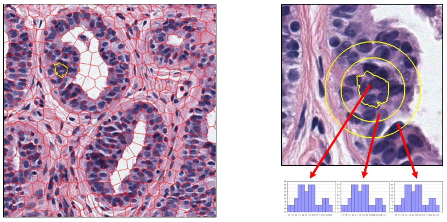
**Superpixel and SVM-based Baseline:** For purpose of comparison, we also implemented a traditional feature-based segmentation method as a baseline. We refer to this method as SP-SVM. We used the SLIC algorithm [5] to segment H&E images into superpixels of size 3,000 pixels. From each superpixel, we extracted color histograms on L*a*b* channels and LBP texture histograms [19] on the H&E channels. We used the color deconvolution algorithm [34] to separate the H&E channels. A superpixel size of 3,000 pixels was selected to have approximately one or two epithelial cells in one superpixel in order to capture detailed duct structures. To improve the classification, we included two circular neighborhoods around each superpixel in feature extraction. The color and texture histograms calculated from the superpixels and circular neighborhoods were concatenated to produce one feature vector for each superpixel. Figure 6 illustrates the two circular neighborhoods from which the same features were extracted and appended to the superpixel feature vector.

**Training Details:** We split 58 images (regions of interest marked and annotated by the experts) into training (N=30)


Crop
ACU-1
ACU-1
ACU-1
Crop
ACU-1
ACU-2
ACU-4
ACU-8
ACU-16
ACU-1
ACU-r → Atrous or Dilated convolution (3x3) with a dilation rate of r
ACU-1 is the same as 3x3 convolution.

**(a)** Fusion-A **(b)** Fusion-B

**Figure 5:** Different fusion strategies for multi-resolution network.



(a) superpixel segmentation (b) neighborhoods

**Figure 6:** Initial superpixel segmentation and the circular neighborhoods used to increase the superpixel classification accuracy for supervised segmentation. Best viewed in color.

and test (N=28) sets. For the single-resolution networks, we cropped patches of size $256 \times 256$ with an overlap of 56 pixels at different WSI resolutions ($5\times$ and $10\times$). For the multi-resolution networks, for each $256\times256$ patch, we created another patch by including a 64-pixel border area (see Figure 4). When necessary, we used symmetric padding to complete the patches. We obtained $5,312$ patches from the training set (N=30). To augment the data, we used standard augmentation strategies, such as random rotations, horizontal flips, and cropping, resulting in a total of 25,992 patches. We used a 90:10 ratio for splitting these patches into training and validation sets.

We trained all of our models end-to-end using stochastic gradient descent with a fixed learning rate of 0.0005, momentum of 0.9, weight decay of 0.0005, and a batch size of 10 on a single NVIDIA GTX-1080 GPU. We initialized encoder weight with ResNet-18 [21] trained on the ImageNet dataset [26]. We choose ResNet-18, because it: (1) is fast at inference, (2) requires less memory per image, and (3) learns less parameters while delivering accuracy similar to VGG [36] on the ImageNet. We initialized decoder weights as suggested in [20]. We did not use dropout, following the practice of [24, 21]. We used an inverse class probability weighting scheme to deal with the class imbalance. Motivated by He *et al.*[21], we applied batch normalization [24] and ReLU [20] operations after every convolution or deconvolution or atrous/dilated convolution operation, with the exception of RCU and IA-RCU blocks where second ReLU is performed after the element-wise sum operation.

For the superpixel and SVM-based baseline, we concatenated color and texture histograms to train an SVM that classifies super-pixels into eight tissue labels. To address the non-uniform distribution of the tissue labels and ROI size variation, we sampled 2,000 superpixels for each of the eight labels (if possible) from each image. We used the same training and test sets to evaluate the SP-SVM method.

## 6.1. Segmentation Results

We evaluated our results using three metrics commonly used for semantic segmentation [7, 35, 6]: (1) F1-score (F1), (2) mean region Intersection over Union (mIOU), and (3) global pixel accuracy (PA). Table 2 summarizes the performance of different encoder-decoder models and feature-based baseline. The impact of each of our modifications along with a comparison with the feature-based segmentation method are discussed below.

**Residual vs Dense Connections:** The residual encoder-decoder has a $0.5\%$ higher pixel accuracy (PA) than the plain encoder-decoder, and our model with dense connections (A3) has a $2\%$ higher PA than plain encoder-decoder under both single and multiple resolution settings. On an average, dense connections improve the accuracy (across different metrics) by at least $1\%$ without significantly increasing the number of parameters of the network.

**RCU vs IA-RCU:** Replacing the IA-RCU with conventional RCUs (A1) in our model reduces accuracy (both F1 and PA) by about $4\%$ under single resolution and $7\%$ under multiple resolutions. Furthermore, A2 with IA-RCU has $2\%$ higher accuracy than A3 with RCUs under multiple resolution setting. Figure 7 visualizes the activation maps of different encoding blocks at different spatial resolutions in which RCUs lose information about small structures in lower spatial dimensions, while the IA-RCUs help in retaining this information.

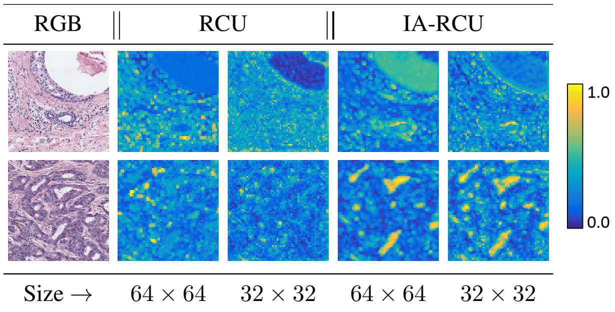


**Figure 7:** Visualization of activation maps of different encoding blocks at different spatial resolutions. IA-RCU compensates the loss of spatial information due to down-sampling operations and helps in learning features that are relevant with respect to input. For visualization, we have scaled the activation maps to the same spatial dimensions. Best viewed in color.

| | Dense Conn. | Multi-Dec. | IA-RCU | Single resolution | | | | Multiple resolution | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | **# Params** | **F1** | **mIOU** | **PA** | **# Params** | **F1** | **mIOU** | **PA** |
| Plain Enc-Dec [6] | | | | 12.80 M | 0.507 | 0.376 | 0.575 | 25.61 M | 0.513 | 0.381 | 0.593 |
| Residual Enc-Dec [15] | | | | 12.80 M | 0.510 | 0.381 | 0.586 | 25.61 M | 0.517 | 0.386 | 0.597 |
| **Our Model** | ✓ | ✓ | ✓ | 13.00 M | 0.554 | 0.418 | 0.642 | 26.03 M | **0.588** | **0.442** | **0.700** |
| A1 | ✓ | ✓ | | 12.93 M | 0.517 | 0.385 | 0.608 | 25.85 M | 0.529 | 0.390 | 0.631 |
| A2 | ✓ | | ✓ | 12.99 M | 0.517 | 0.387 | 0.601 | 25.98 M | 0.540 | 0.407 | 0.633 |
| A3 | ✓ | | | 12.92 M | 0.519 | 0.390 | 0.607 | 25.84 M | 0.524 | 0.392 | 0.611 |
| Ours + Fusion-A | ✓ | ✓ | ✓ | NA | NA | NA | NA | 26.03 M | 0.535 | 0.402 | 0.631 |
| Ours + Fusion-B | ✓ | ✓ | ✓ | NA | NA | NA | NA | 26.00 M | 0.554 | 0.419 | 0.658 |
| SP-SVM | | NA | | NA | 0.365 | 0.258 | 0.485 | NA | NA | NA | NA |

**Table 2:** Quantitative comparison of different methods on the Breast Biopsy dataset.

**Single vs Multiple Decoders:** Replacing multiple decoders with a single decoder in A2 reduces the pixel accuracy of our full model by 4% with single resolution and 7% with multiple resolutions. Furthermore, A1 has 2% higher pixel accuracy than A3 under multiple resolution setting. The pixel accuracy does not change from A3 to A1 under the single resolution setting.

**Single vs Multiple Resolutions:** For all models, multi-resolution inputs improve the performance up to 6% in pixel accuracy. All metrics increase from single resolution to multi-resolution for all models. Although the improvement in accuracy is small, multi-resolution input leads to better segmentation results (see Figure 8 and Figure 9).

**Different Fusion Methods:** The overall F1-score of our model with our fusion scheme (Figure 4) is about 6% and 4% higher than Fusion-A and Fusion-B (Figure 5), respectively.

**Inference Time and Number of Parameters:** The impact on inference time and number of parameters learned by both single and multi-resolution networks is reported in Figure 10. Multi-resolution network utilize the hardware resources efficiently by executing multiple encoder-decoder networks simultaneously and therefore, the impact on inference time is not drastic. The multi-resolution networks are merely 0.2× slower than the single resolution network while learning almost 2× more parameters.

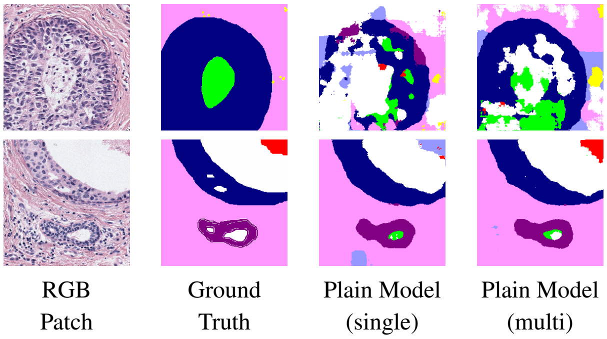


**Figure 8:** Patch-wise predictions of Plain Encoder-Decoder network with single and multiple resolution input. Multi-resolution input helps in improving the predictions, especially at the patch borders. Best viewed in color.

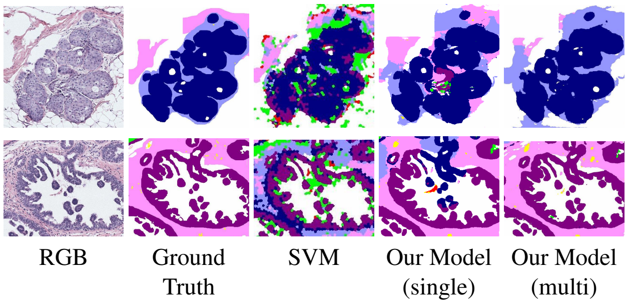


**Figure 9:** ROI-wise predictions: first row depicts an invasive case while the second row depicts a benign case. Best viewed in color.

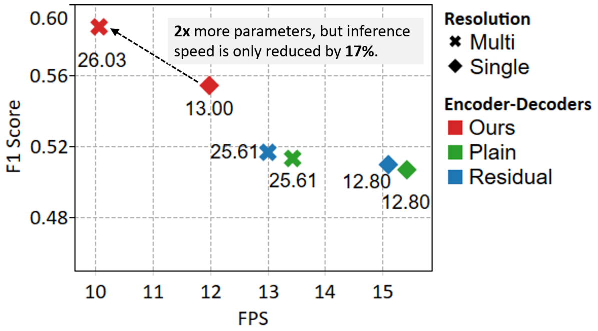


**Figure 10:** Impact on inference time and number of parameters learned at different resolutions. Number of parameters are in million and are listed next to the corresponding data point. Inference time is measured on NVIDIA GTX-1080 GPU and is an average across 3 trials for 20 samples of size 384 × 384. Here, FPS refers to frames (or patches) processed per second. Best viewed in color.

**Comparison with Feature-Based Baseline:** Since the SP-SVM method used only single resolution images, we compared it to our model's performance with single resolution input. Our model outperformed the SP-SVM method across all metrics.

## 6.2. Diagnostic Classification

Semantic segmentation provides a powerful abstraction for diagnostic classification. We designed a set of experiments to show the descriptive power of the tissue label segmentation in automated diagnosis. To this end, we used the full set of ROIs (N=428) to predict the consensus diagnosis assigned by the expert panel. We trained and tested two types of classifiers, an SVM and a multi-layer perceptron (MLP), for four classification tasks: (1) 4-class (benign vs. atypia vs. DCIS vs. invasive); (2) invasive vs. non-invasive (benign, atypia and DCIS); (3) benign vs. non-benign (atypia and DCIS); and (4) atypia vs. DCIS. The last three tasks were designed to imitate the diagnostic decision making process of pathologists while the first one is the naive approach.

We applied our model with single and multiple-resolutions and SP-SVM-based baseline to all the images in our dataset (N=428) to get tissue label segmentations. For diagnostic features, we calculated the frequency and co-occurrence histograms of superpixel tissue labels, using the majority pixel label for the CNN approach that labels pixels. We trained SVMs and MLPs for the four classification tasks in a 10-fold cross-validation setting and repeated the experiments 10 times. During training, we sub-sampled the data to have a uniform distribution of diagnostic classes.

**Results:** The accuracies for four diagnostic classification tasks are given in Table 3. The features calculated from segmentation masks produced by our model outperforms the

| | Diagnostic Classifier: **SVM** | | | | | | Diagnostic Classifier: **MLP** | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **SP-SVM** | | **Our Model (single)** | | **Our Model (multi)** | | **SP-SVM** | | **Our Model (single)** | | **Our Model (multi)** | |
| | all labels | no stroma | all labels | no stroma | all labels | no stroma | all labels | no stroma | all labels | no stroma | all labels | no stroma |
| **4-class** | 35.5% | 32.1% | 44.5% | 36.3% | **45.9%** | 36.3% | 45.0% | 38.6% | **54.5%** | 46.4% | 54.2% | 45.2% |
| **invasive** | 64.7% | 44.6% | 78.4% | 58.4% | **90.7%** | 63.4% | 69.0% | 57.8 % | 69.0% | 64.1% | **76.0%** | 68.7% |
| **benign** | 55.0% | **67.7%** | 44.7% | 65.3% | 40.0% | 61.0% | 61.1% | 60.3% | **66.5%** | 66.2% | 65.8% | 64.2% |
| **atypia-DCIS** | 66.34% | 59.2% | 84.69% | **85.1%** | 84.07% | 82.8% | 74.28% | 68.5% | 85.03% | **87.7%** | 82.07% | 81.3% |

**Table 3:** Diagnostic classification accuracies for different classification methods

SP-SVM method with both classifiers, with the exception of classification of benign cases with SVM. In particular, multi-resolution input improves the segmentation of desmoplastic stroma label significantly (Figure 11), which is easily identifiable in lower-resolutions and an important tissue type for diagnosing breast cancer [28]. Incorporating input from larger surrounding tissue helps the model identify tumor-associated desmoplastic stroma, in turn, it improves the classification of invasive cases (90.7% with the multi-resolution model and SVM classifier).

## 7. Discussion

Diagnostic classification with the full range of breast diagnoses is a difficult problem. In a previous study, a group of pathologists interpreted the same digital slides of breast biopsies [13] and achieved accuracies of 70%, 98%, 81% and 80% for the tasks of 4-class, invasive vs. (benign-atypia-DCIS), (atypia-DCIS) vs. benign, and DCIS vs. atypia respectively. Semantic segmentation provides a powerful abstraction so that simple features with diagnostic classifiers, like SVM and multi-layer perceptron, perform well in comparison to pathologists.

Multi-resolution input increases the context of the model and improve the segmentation of the labels pathologists identify in lower resolutions; such as desmoplastic stroma. Furthermore, our fusion block outperforms the alternative fusion blocks, most likely due to its high effective receptive field. The effective receptive field of our block (Figure 4) is 65 × 65 while the effective receptive fields of the fusion blocks in Figure 5a and 5b are 7 × 7 and 37 × 37. In addition to quantitative evaluation, our model results in smoother borders for the segmented regions while the SP-SVM method is limited to color similarity for initial segmentation and has much smaller context than networks.

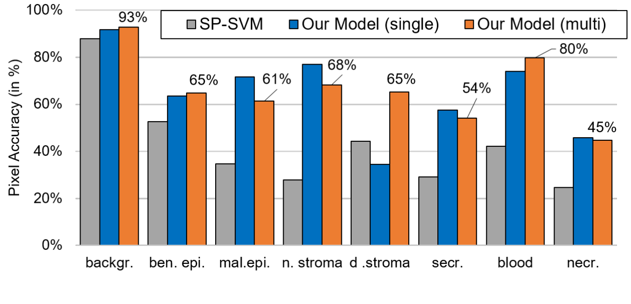


**Figure 11:** Segmentation accuracy for different labels. Best viewed in color.

An automated diagnosis system should operate on whole slide images. Since the whole-slide-level annotations were not available on our data, we validated our model on regions of interest that were identified, diagnosed, and annotated by the experts. Our method can easily be applied to WSIs for segmentation or can be used in combination with a region of interest identifier for classification [29]. Our future work involves developing a system for simultaneous ROI localization, segmentation, and diagnostic classification on WSIs.

## 8. Conclusions

Our model outperforms traditional encoder-decoders and the SP-SVM-baseline both qualitatively and quantitatively (see Figure 9). It also improves the F1-score and mIOU of conventional networks by at least 7% and the global pixel accuracy by 11% for multiple resolution settings. This improvement is mainly due to the long-range direct connections that are established between input and output either using identity or projection mappings. These long-range connections helps in back-propagating the information directly to the input paths efficiently and therefore, improves the flow of information inside the network and eases the optimization.

We showed that our semantic segmentation provides powerful features for diagnosis. With hand-crafted or learned features for diagnosis, our model is promising for a computer-aided system for breast cancer diagnosis. Though we study breast biopsy images in this paper, our system can be easily extended to other types of cancer.

## Acknowledgements

Research reported in this publication was supported by the National Cancer Institute awards R01 CA172343, R01 CA140560, and KO5 CA104699. The content is solely the responsibility of the authors and does not necessarily represent the views of the National Cancer Institute or the National Institutes of Health. We thank Ventana Medical Systems, Inc. (Tucson, AZ, USA) for the use of iScan Coreo Au WSI system, and HD View SL for the source code used to build our digital viewer.

The authors thank Hannaneh Hajishirzi and anonymous reviewers for their thorough and helpful comments.

## References

[1] Breast Cancer Surveillance Consortium. http://www.bcsc-research.org/. [Online accessed: 7 September, 2017].

[2] FDA allows marketing of first whole slide imaging system for digital pathology. https://www.fda.gov/newsevents/newsroom/pressannouncements/ucm552742.htm. [Online accessed: 7 September, 2017].

[3] The Cancer Genome Atlas. http://cancergenome.nih.gov. [Online accessed: 7 September, 2017].

[4] Tumor Proliferation Assessment Challenge 2016, MICCAI. http://tupac.tue-image.nl/. [Online accessed: 7 September, 2017].

[5] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk. SLIC superpixels compared to state-of-the-art superpixel methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(11):2274–2281, 2012.

[6] V. Badrinarayanan, A. Kendall, and R. Cipolla. Segnet: A deep convolutional encoder-decoder architecture for scene segmentation. *TPAMI*, 2017.

[7] H. Chen, X. Qi, L. Yu, and P.-A. Heng. DCAN: Deep Contour-Aware Networks for Accurate Gland Segmentation. In *CVPR*, 2016.

[8] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *arXiv preprint arXiv:1606.00915*, 2016.

[9] L.-C. Chen, Y. Yang, J. Wang, W. Xu, and A. L. Yuille. Attention to scale: Scale-aware semantic image segmentation. In *CVPR*, 2016.

[10] A. Cruz-Roa, H. Gilmore, A. Basavanhally, M. Feldman, S. Ganesan, N. N. Shih, J. Tomaszewski, F. A. González, and A. Madabhushi. Accurate and reproducible invasive breast cancer detection in whole-slide images: A deep learning approach for quantifying tumor extent. In *Scientific reports*, 2017.

[11] F. Dong, H. Irshad, E. Y. Oh, M. F. Lerwill, E. F. Brachtel, N. C. Jones, N. W. Knoblauch, L. Montaser-Kouhsari, N. B. Johnson, L. K. F. Rao, B. Faulkner-Jones, D. C. Wilbur, S. J. Schnitt, and A. H. Beck. Computational pathology to discriminate benign from malignant intraductal proliferations of the breast. *PLoS ONE*, 9(12):e114885, 2014.

[12] D. Eigen and R. Fergus. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In *ICCV*, 2015.

[13] J. G. Elmore, G. M. Longton, P. A. Carney, B. M. Geller, T. Onega, A. N. A. Tosteson, H. D. Nelson, M. S. Pepe, K. H. Allison, S. J. Schnitt, F. P. O'Malley, and D. L. Weaver. Diagnostic Concordance Among Pathologists Interpreting Breast Biopsy Specimens. *JAMA*, 313(11):1122, mar 2015.

[14] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes challenge: A retrospective. *International Journal of Computer Vision*, 2015.

[15] A. Fakhry, T. Zeng, and S. Ji. Residual deconvolutional networks for brain electron microscopy image segmentation. *IEEE Transactions on Medical Imaging*, 2017.

[16] C. Farabet, C. Couprie, L. Najman, and Y. LeCun. Learning hierarchical features for scene labeling. *TPAMI*, 2013.

[17] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014.

[18] B. Hariharan, P. Arbeláez, R. Girshick, and J. Malik. Hypercolumns for object segmentation and fine-grained localization. In *CVPR*, 2015.

[19] D. C. He and L. Wang. Texture unit, texture spectrum, and texture analysis. *IEEE Transactions on Geoscience and Remote Sensing*, 28(4):509–512, 1990.

[20] K. He, X. Zhang, S. Ren, and J. Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *ICCV*, 2015.

[21] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[22] 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 *CVPR*, 2016.

[23] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[24] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, 2015.

[25] A. Karpathy et al. Large-scale video classification with convolutional neural networks. In *CVPR*, 2014.

[26] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[27] G. Lin, A. Milan, C. Shen, and I. Reid. Refinenet: Multipath refinement networks with identity mappings for high-resolution semantic segmentation. In *CVPR*, 2017.

[28] Y. Mao, E. T. Keller, D. H. Garfield, K. Shen, and J. Wang. Stromal cells in tumor microenvironment and breast cancer. *Cancer and Metastasis Reviews*, 32(1-2):303–315, 2013.

[29] E. Mercan et al. Localization of diagnostically relevant regions of interest in whole slide images. In *ICPR*, 2014.

[30] F. Milletari, N. Navab, and S.-A. Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *3D Vision (3DV), 2016 Fourth International Conference on*, pages 565–571. IEEE, 2016.

[31] H. Noh, S. Hong, and B. Han. Learning deconvolution network for semantic segmentation. In *ICCV*, 2015.

[32] P. Pinheiro and R. Collobert. Recurrent convolutional neural networks for scene labeling. In *ICML*, 2014.

[33] O. Ronneberger, P. Fischer, and T. Brox. *U-Net: Convolutional Networks for Biomedical Image Segmentation*. 2015.

[34] A. C. Ruifrok and D. A. Johnston. Quantification of histochemical staining by color deconvolution. *Analytical and Quantitative Cytology and Histology*, 23(4):291–299, 2001.

[35] E. Shelhamer, J. Long, and T. Darrell. Fully convolutional networks for semantic segmentation. *TPAMI*, 2017.

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

[37] K. Sirinukunwattana et al. Gland segmentation in colon histology images: The glas challenge contest. *CoRR*, abs/1603.00275, 2016.

[38] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *CVPR*, 2015.

[39] M. Veta et al. Assessment of algorithms for mitosis detection in breast cancer histopathology images. *Medical Image Analysis*, 20(1):237 – 248, 2015.

[40] J. Yang, B. Price, S. Cohen, H. Lee, and M.-H. Yang. Object contour detection with a fully convolutional encoder-decoder network. In *CVPR*, 2016.

[41] F. Yu and V. Koltun. Multi-scale context aggregation by dilated convolutions. In *ICLR*, 2016.

[42] L. Yu, X. Yang, H. Chen, J. Qin, and P.-A. Heng. Volumetric convnets with mixed residual connections for automated prostate segmentation from 3d mr images. In *AAAI*, pages 66–72, 2017.

[43] S. Zagoruyko et al. A multipath network for object detection. In *BMVC*, 2016.

[44] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, and P. H. Torr. Conditional random fields as recurrent neural networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1529–1537, 2015.