Learning to Segment Breast Biopsy Whole Slide Images

Sachin Mehta, Ezgi Mercan, Jamen Bartlett, Donald Weaver, Joann Elmore, and Linda Shapiro

Outline

- Introduction
- Our encoder-decoder architecture
  - Input-aware residual convolutional units
  - Densely connected decoding units
  - Multi-resolution input
- Results

Introduction

- CNNs are the state-of-the-art architectures for segmenting natural and medical images.
- However, CNNs can't be directly applied to breast biopsy WSI images due to their size.

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**	$10K \times 12K$

Introduction

- To segment these images, a simple strategy is to use a sliding window-based approach
- Diving these large tissue structures limits the context available to CNNs and may affect the segmentation performance.
- We introduced a new multi-resolution encoder-decoder architecture that was specifically designed to handle the challenges of the breast biopsy semantic segmentation problem.

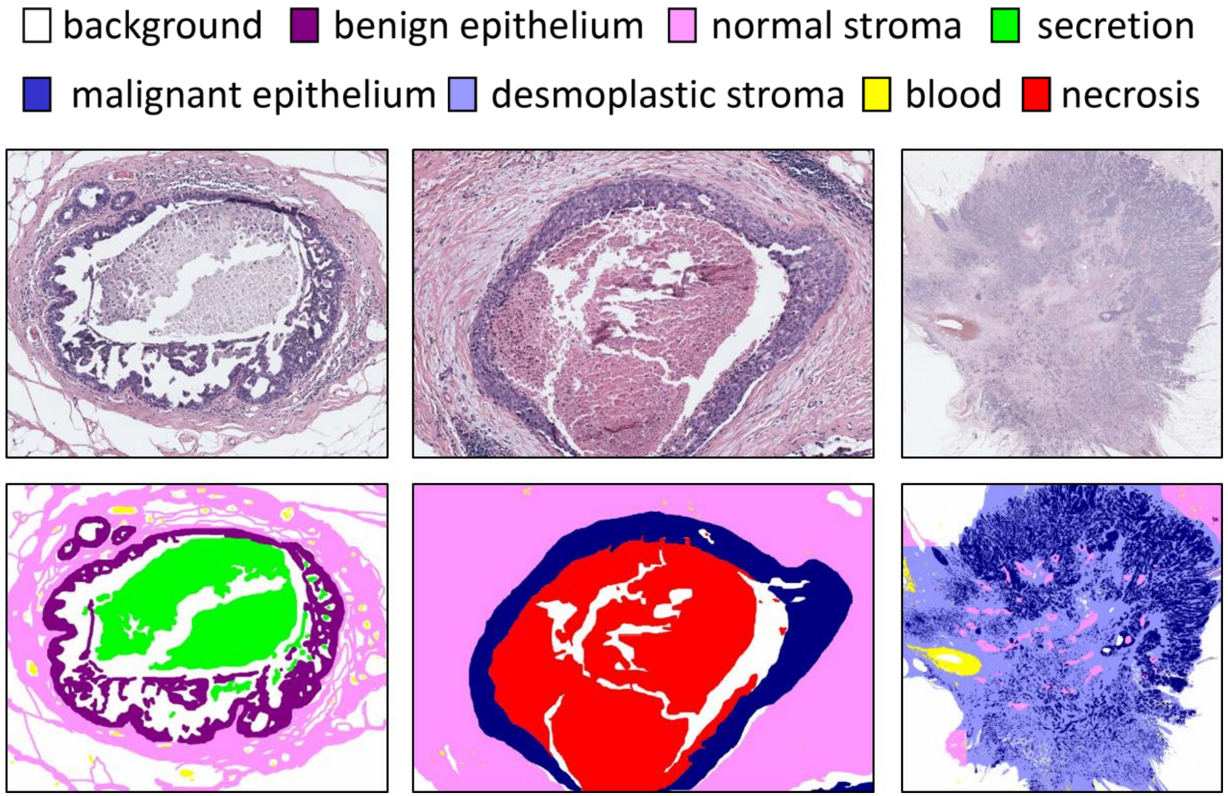


Figure: The set of tissue labels used in semantic segmentation. Note that the objects of interest (or tissues) are variable in size.

Encoder-decoder Network for Segmenting WSIs

Overview of encoder-decoder network

- Encoder-decoder network comprises of two networks:
  - Encoder
  - Decoder
- Encoder aggregate features at multiple spatial resolutions by performing different operations such as convolution and down-sampling operations.
- Decoder tries to invert the loss of spatial resolution due to down-sampling operations in the encoder.

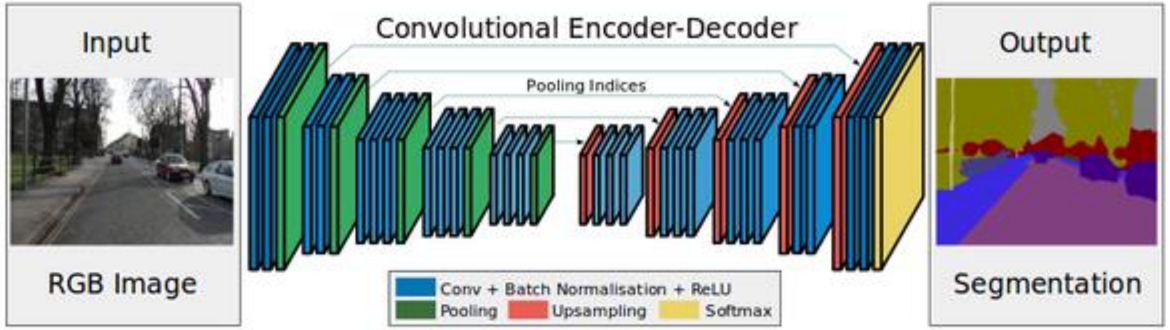
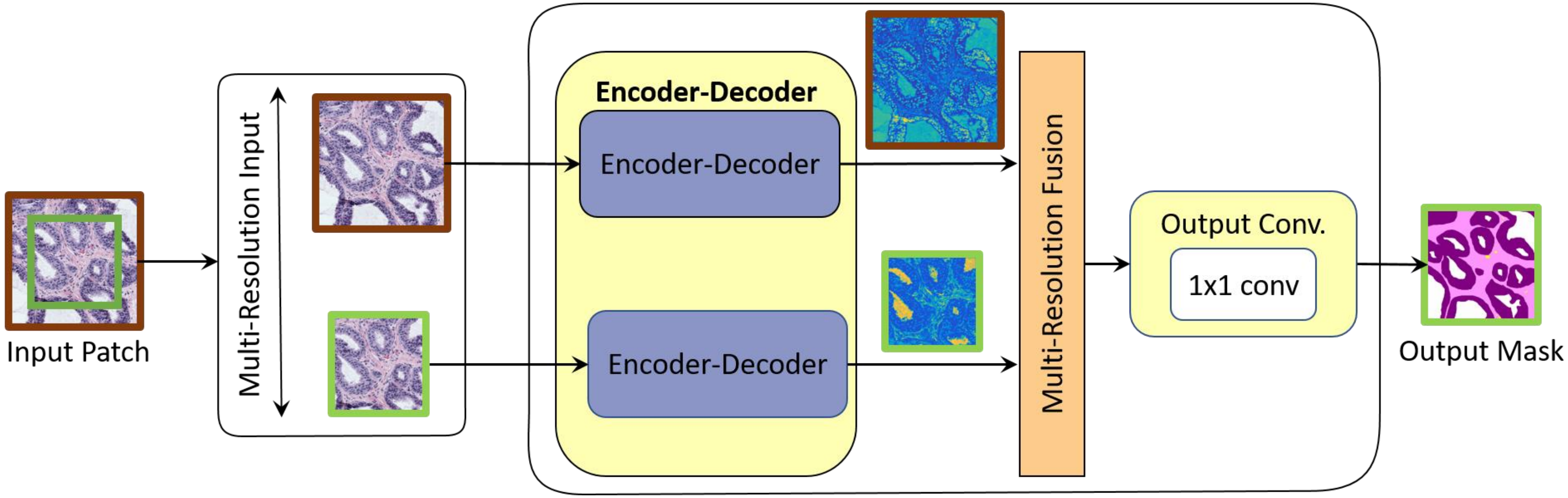


Figure: Convolutional Encoder-Decoder Network*

* **Image Source:** Badrinarayanan, V., Kendall, A. and Cipolla, R., 2017. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *TPAMI*

Single vs Multi-resolution Network

- Patch-based approach divides large tissue structures into smaller structures and limits the context (surrounding tissue information) available to CNNs.
- Patch-based approach may affect the segmentation performance.
- To make the CNN model aware of the surrounding information, we introduce a multi-resolution network



Patch-wise Predictions: Single vs Multi-resolution Network

background | benign epithelium | normal stroma | secretion
malignant epithelium | desmoplastic stroma | blood | necrosis

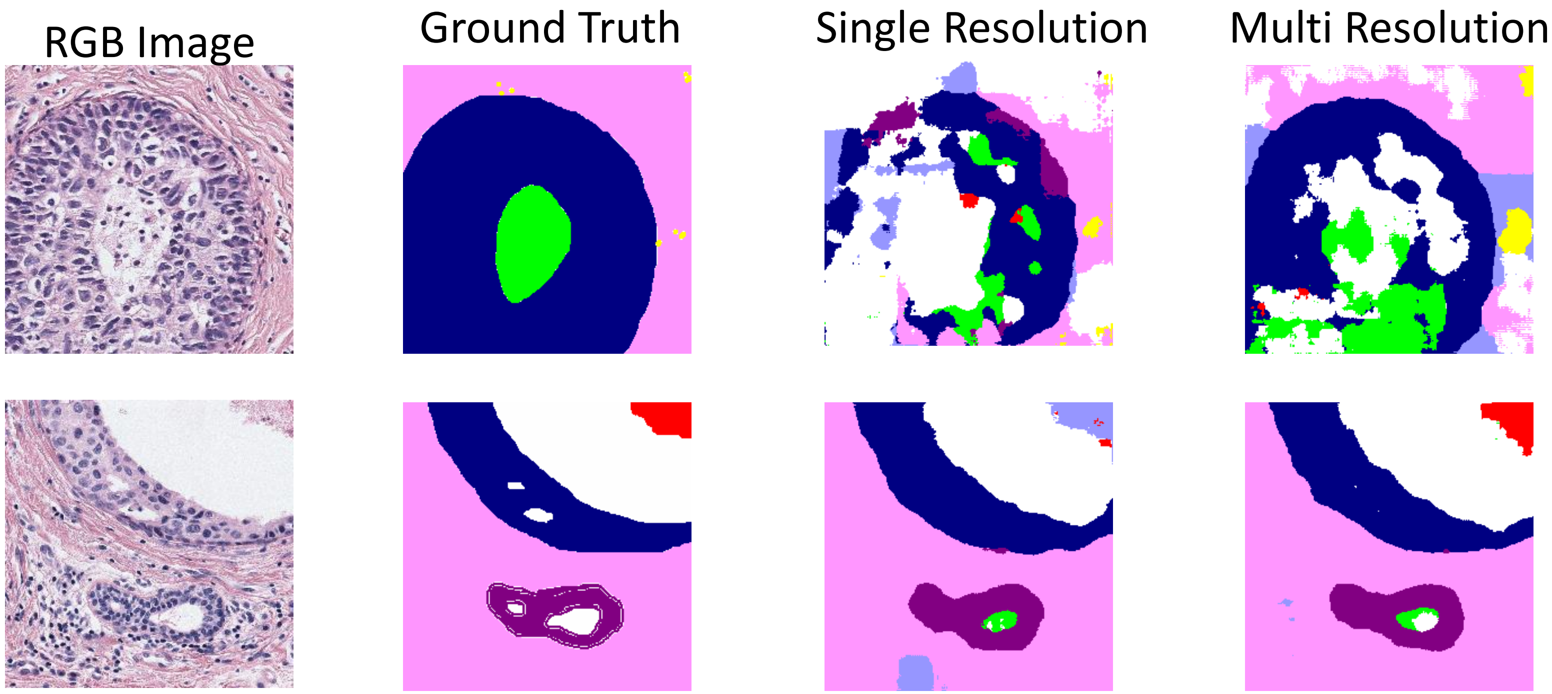
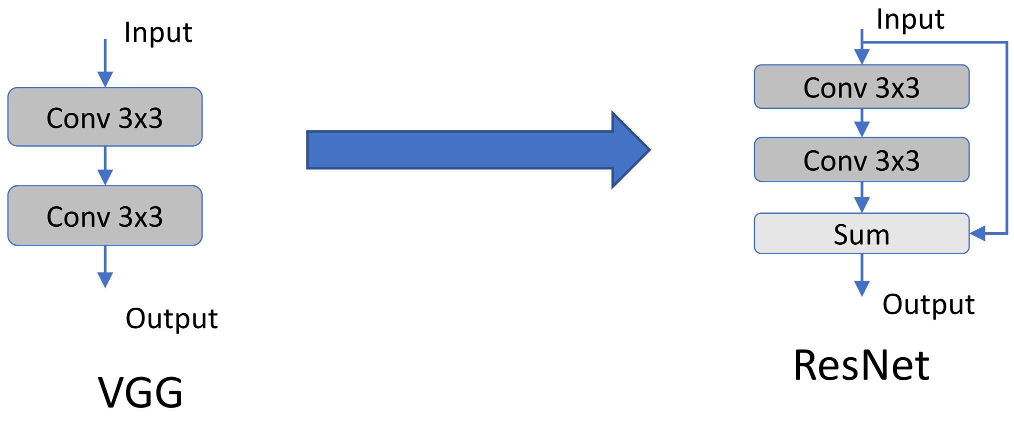


Figure: Patch-wise predictions of Plain Encoder-Decoder network with single and multiple resolution input. Multi-resolution input helps in improving the predictions, especially at the patch borders.

Overview of Convolutional Units

- Convolutional unit is a composite function comprising of convolutional layers, non-linearity operations (such as ReLU) and batch normalization.
- Two popular convolutional units are:



ResNet adds a bypass connection between input and output of the convolutional block to improve the information flow inside the network and avoid the vanishing gradient problem.

Source:
VGG: Simonyan, Karen, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." *ICLR, 2015*
ResNet: He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Deep residual learning for image recognition." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770-778. 2016.

Input-aware Residual Convolution Units

- As we increase the depth of the network, we learn coarse features about the objects. These coarse features are useful for object classification, but not for segmentation.

Input Image　　Ground Truth　　Output of FCN-32s

Figure: Output of FCN-32s

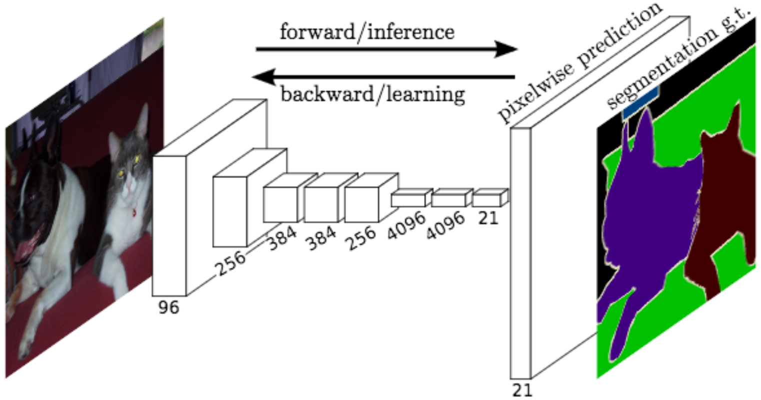
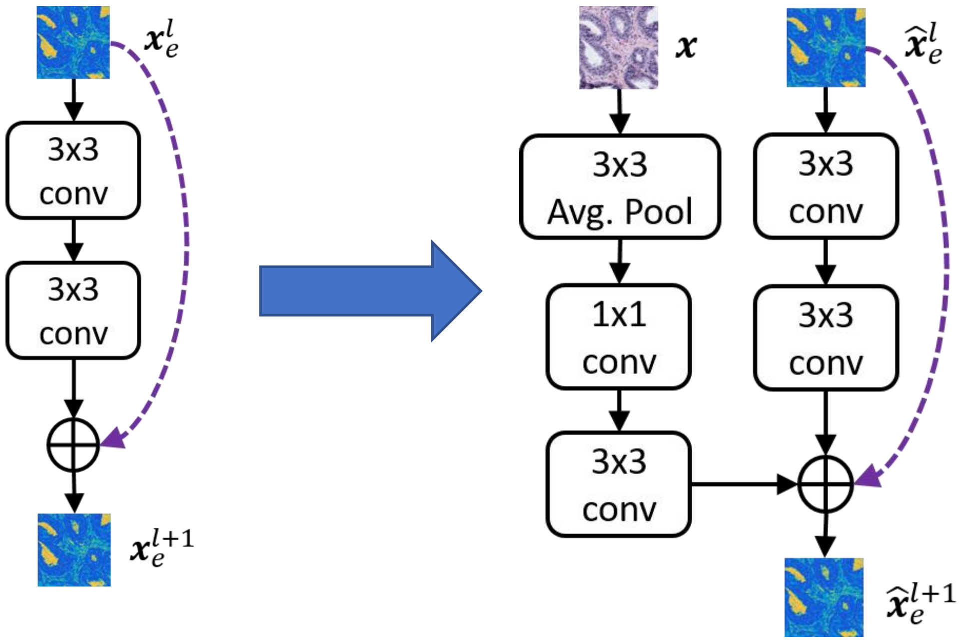


Figure: FCN-32s architecture that upsamples the last CNN layer (VGG) output by 32x, so that input image and segmentation output are of the same resolution

Source: Long, Jonathan, Evan Shelhamer, and Trevor Darrell. "Fully convolutional networks for semantic segmentation." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3431-3440. 2015.

Input-aware Residual Convolution Units

- We introduce an input-aware residual convolutional unit that reinforces the input at different spatial levels of CNNs to learn input-specific features



RCU IA-RCU

Activation Map Visualization

Residual Convolutional Unit (RCU) vs Input Aware Residual Convolutional Unit (IA-RCU)

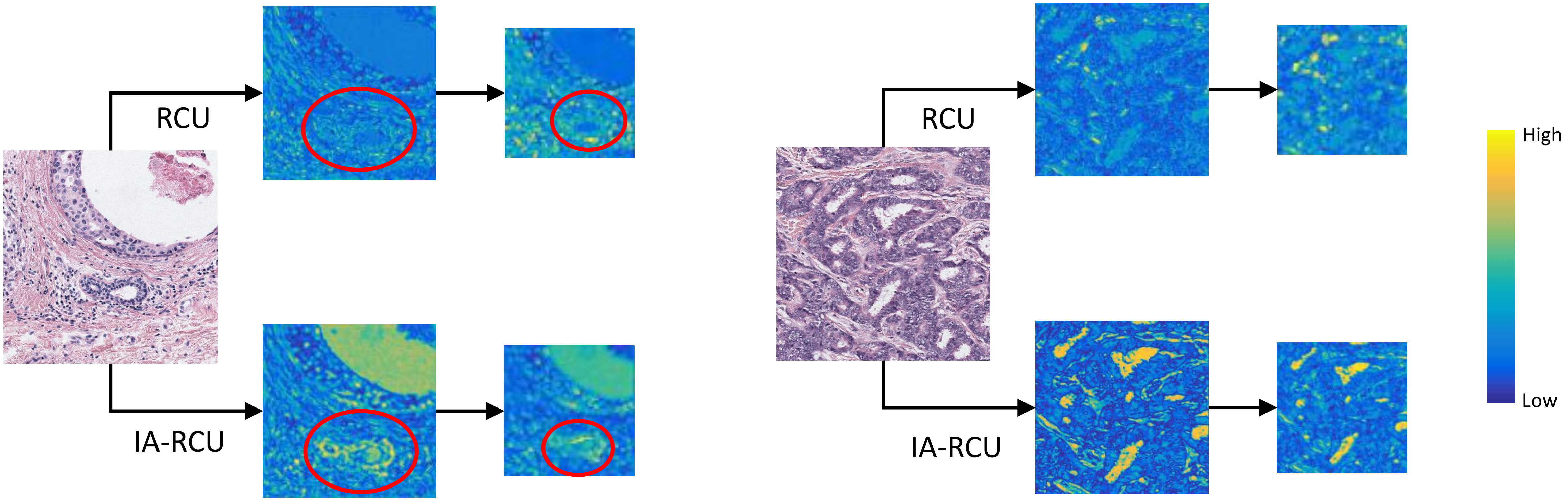


Figure: Two examples visualizing the activation maps at different spatial resolutions. IA-RCU compensates the loss of spatial information due to down-sampling operations and helps in learning features that are relevant with respect to input.

Densely Connected Decoding Paths

- Similar to convolutional units, we can have skip connections between encoding and its corresponding decoding block.
- These skip-connections establishes a direct connection between encoder and decoder and improves 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-level or lower-level.
- These connections establishes **long-range connections** and promote feature reuse.

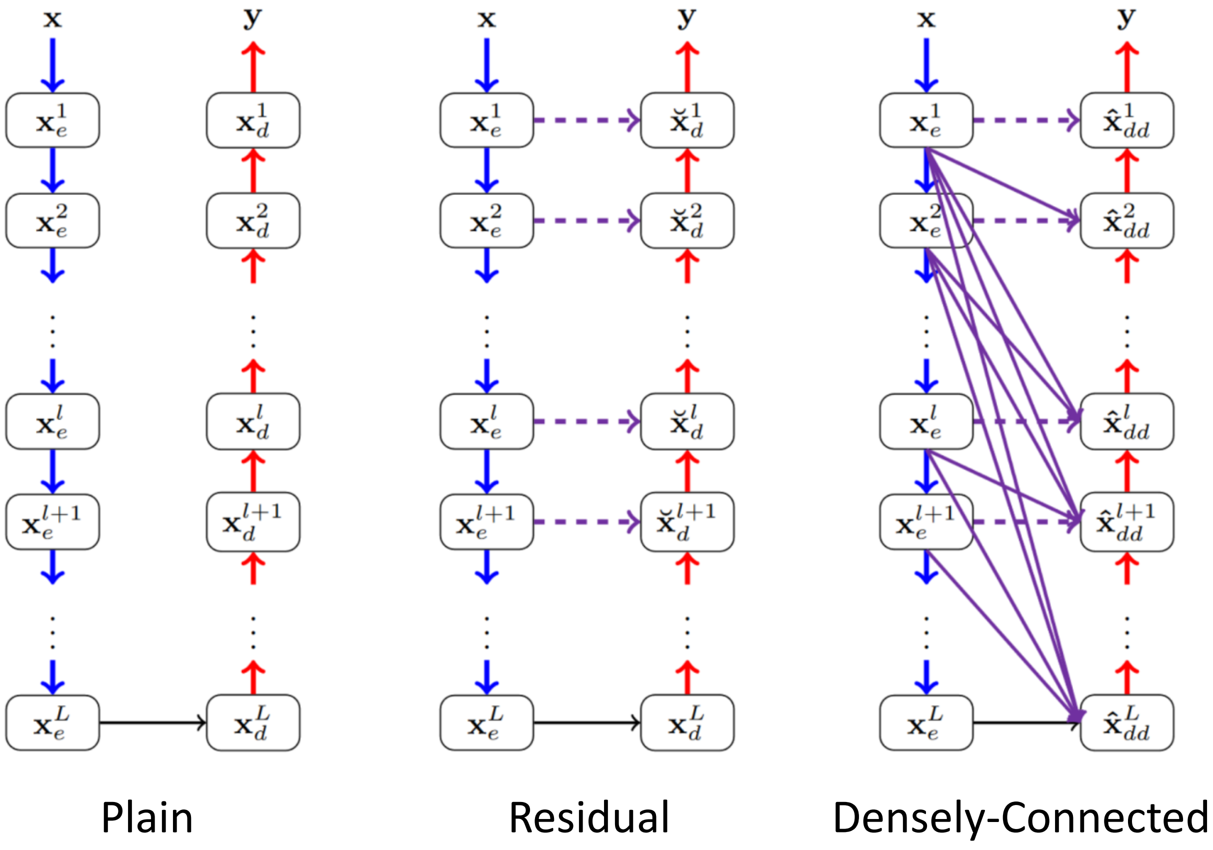
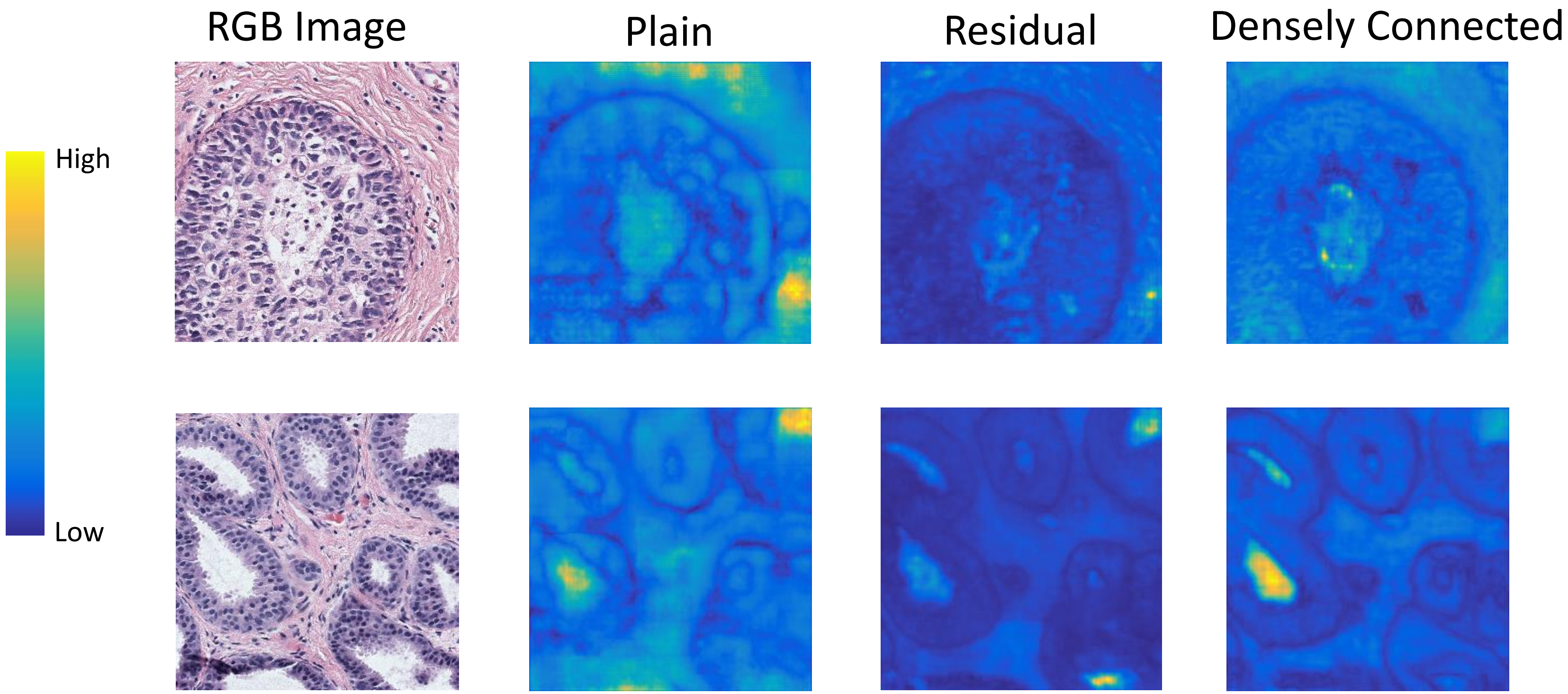


Figure: Different encoder-decoder architectures

Visualization of Activation Maps of different networks



- Features learned by the plain network are noisy.
- Residual network helps in refining the feature maps by combining the low-level and high-level information.
- Dense connections promote the feature reuse and helps in efficiently combining the low-level and high-level information.

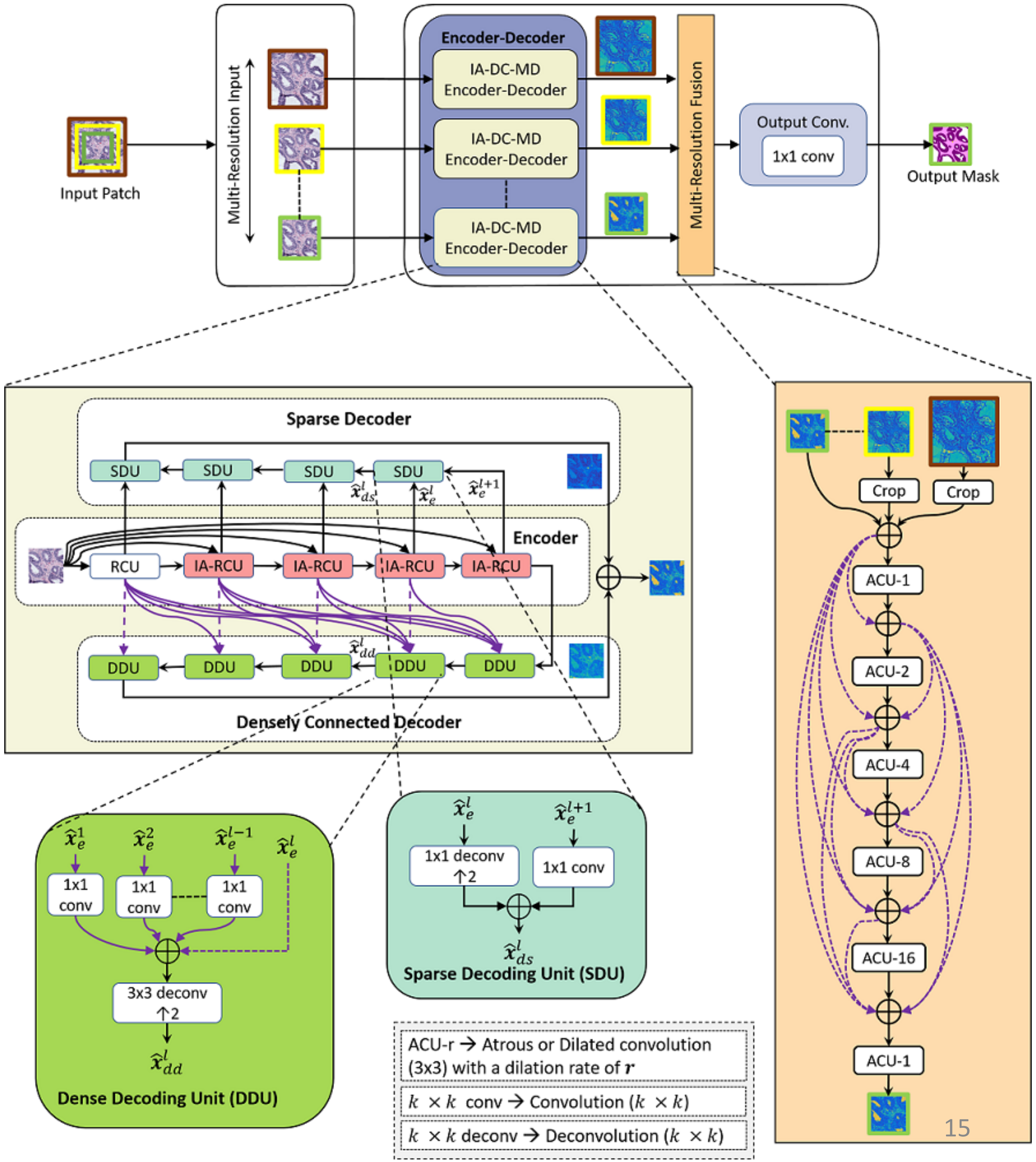
Our Encoder-Decoder Architecture for WSI Segmentation

Our encoder-decoder network for segmenting WSIs that incorporates:

- Multi-resolution input
- Input-aware residual convolutional units
- Densely connected decoding paths
- Sparse decoder

More details about the network architecture, see our paper: **Learning to Segment Breast Biopsy Whole Slide Images**, to appear *in IEEE Winter Conference in Computer Vision* (WACV-18)

Web Link: https://arxiv.org/abs/1709.02554



Results

Training details

- Training Set: 30 ROIs
  - 25,992 patches of size 256x256 with augmentation
  - Split into training and validation set using 90:10 ratio
- Test Set: 28 ROIs
- Evaluation metric:
  - Pixel accuracy
  - Mean Region Intersection over Union
  - F1-score
- Stochastic Gradient Descent for optimization
- Implemented in Torch
  - http://torch.ch/

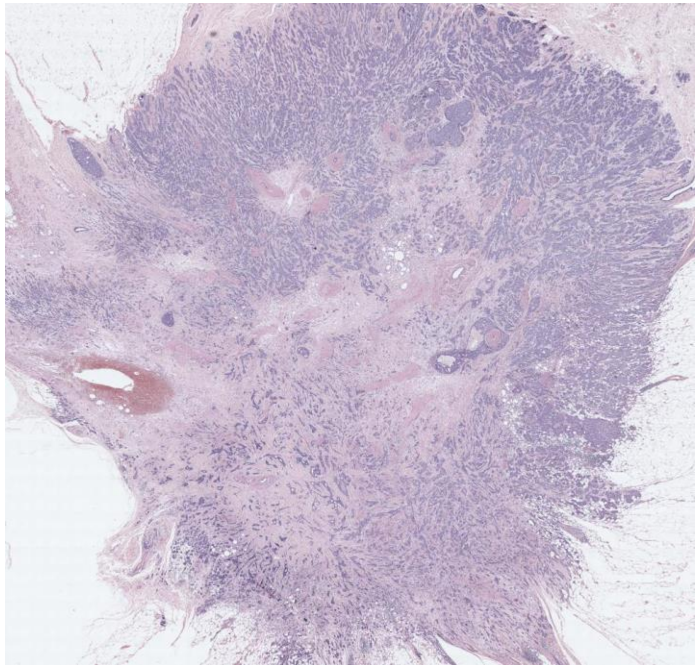
Results

	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

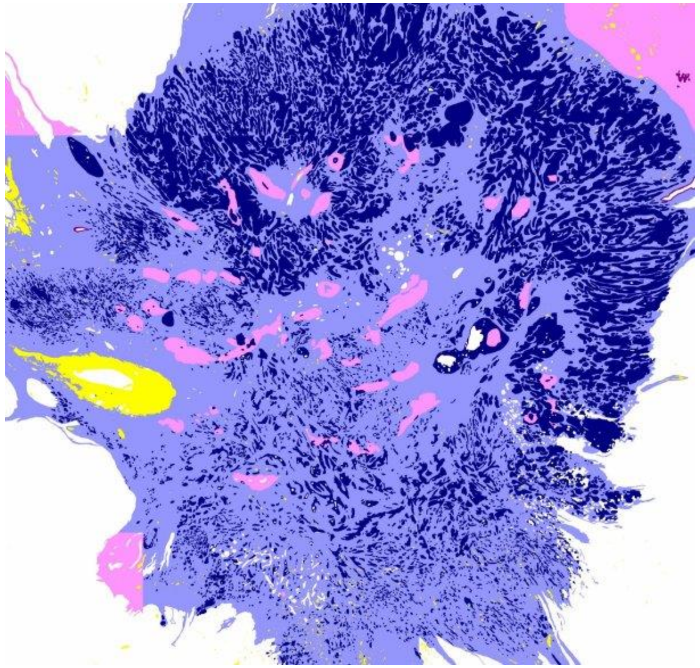
Key findings:

- **Singe-vs-multi-resolution:** For all models, multi-resolution improves the pixel accuracy by about 6%.
- **RCU-vs-IARCU:** IARCU improves the pixel accuracy by about 4% and 7% over RCUs (A1).
- **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.

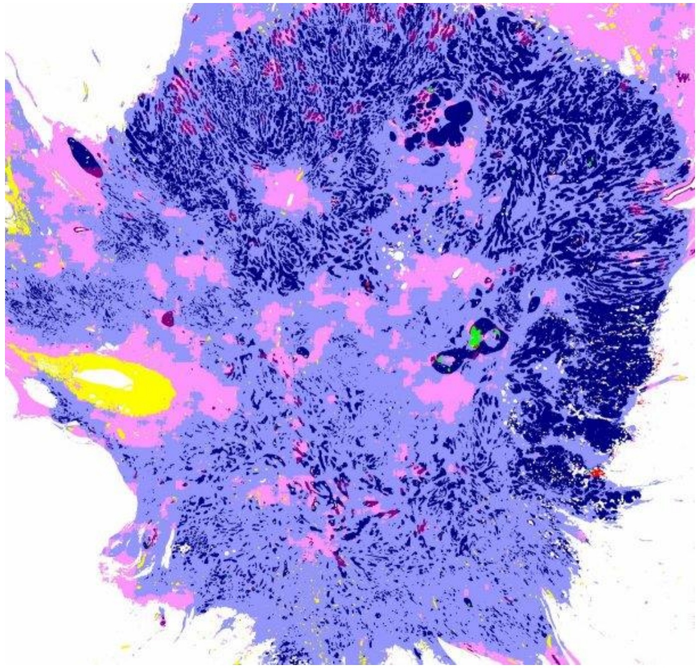
WSI Segmentation Results



RGB Image



Ground Truth



Predicted Semantic Mask

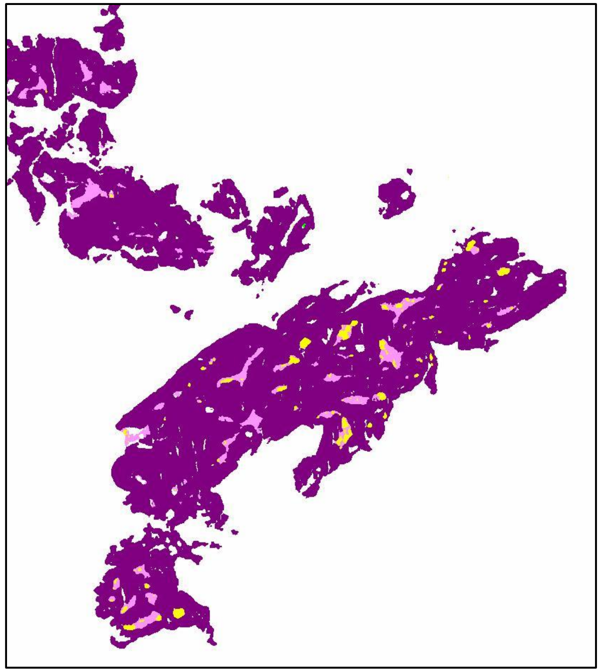
background benign epithelium normal stroma secretion

malignant epithelium desmoplastic stroma blood necrosis

WSI Segmentation Results

RGB Image

Ground Truth



Predicted Semantic Mask

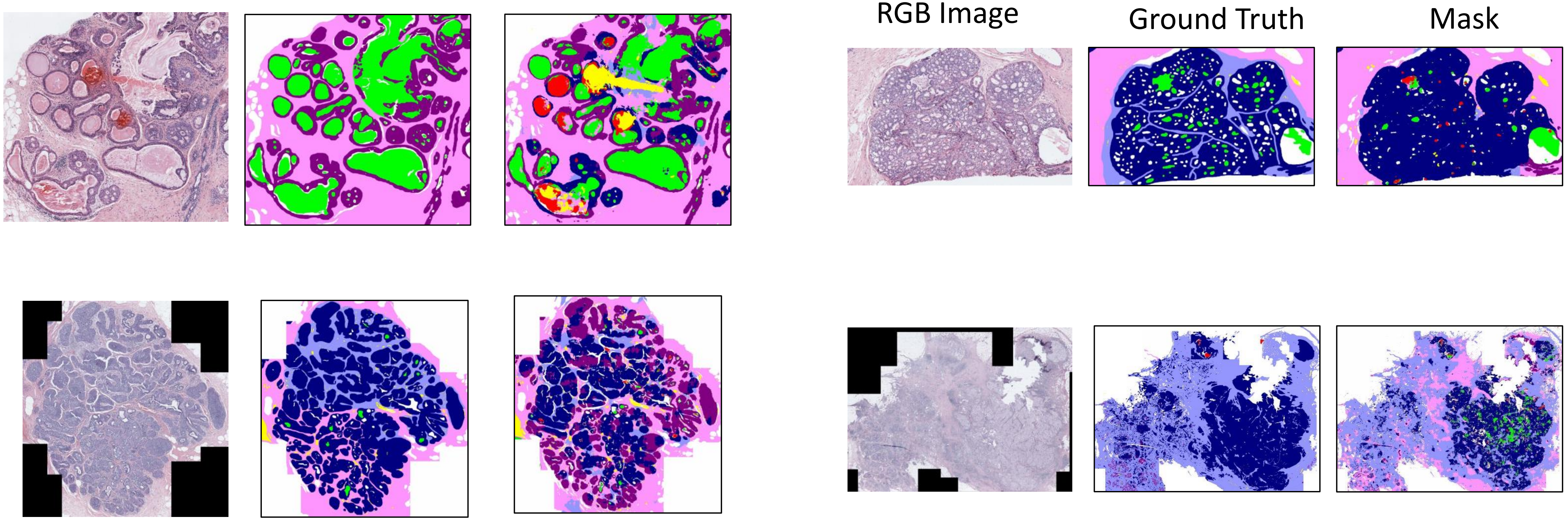
□ background ■ benign epithelium ■ normal stroma ■ secretion

■ malignant epithelium ■ desmoplastic stroma ■ blood ■ necrosis

WSI Segmentation Results

RGB Image | Ground Truth | Predicted Semantic Mask

RGB Image | Ground Truth | Predicted Semantic Mask



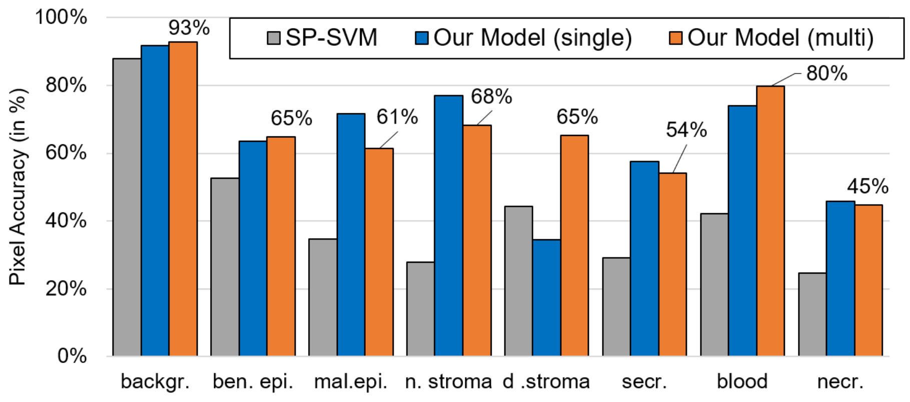
□ background ■ benign epithelium □ normal stroma ■ secretion

■ malignant epithelium □ desmoplastic stroma □ blood ■ necrosis

Application to Cancer Diagnosis

	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%

- Segmentation labels have high descriptive power and therefore, leads to good classification accuracy even with simple classifiers such as SVM and Multi-layer perceptron (MLP)
- Multi-resolution network improves the pixel-wise classification accuracy of stroma tissue; which is an important tissue type for identifying invasive cancer.



Thank You!!

For more details about this work, please check DIGITAL PATHOLOGY project here:
https://homes.cs.washington.edu/~shapiro/digipath.html