Gland Instance Segmentation Through Overlapping Contour Regions and Random Transformation Sampling

Nikita Zozoulenko
ContextVision AB
nikita.zozoulenko@gmail.com

Abstract

Manually separating glands is a hard, time-consuming and expensive task where the final result between multiple expert pathologists can vary massively. Developing an automated method can therefore save time, valuable resources, and lives. In this paper we propose a novel deep convolutional neural network architecture for the task of gland instance segmentation, achieving state of the art results on the MICCAI GlaS dataset. We employed a data augmentation strategy centered around random transformation sampling as a way to model the gland uncertainty at test time, increasing the final detection and segmentation accuracy. Furthermore, we utilized overlapping glandular contour regions to predict areas where two distinct objects are in close proximity to each other. We additionally implemented the previous state of the art in gland instance segmentation, MILD-Net, and failed to reproduce their reported results.

1 Introduction

Segmenting individual glands is an important part in histopathology in relation to adenocarcinomal cancer, since if part of a gland is cancerous, the whole gland should be classified as cancerous. This problem is especially hard because a gland's appearance differs depending on its shape, size, how the organic material was cut, what organ it comes from, as well as the histological grade of the gland. As a result, the labeling of two different expert pathologists can differ heavily. This has created an incentive to develop an automated method to classify individual glands.

Recent advancements in deep learning have lead to the creation of models capable of semantic segmentation of objects in images, classifying every pixel in an image given a set of class labels (see FCN (Long, Shelhamer, and Darrell 2014), U-Net (Ronneberger, Fischer, and Brox 2015), and DeepLabV3 (Chen et al. 2017)). Instance segmentation models, as opposed to a semantic segmentation models, were developed to semantically segment an image while also being able to detect and distinguish between two or more objects of the same semantic class. (see figure 1). This adds further complexity to the already challenging detection problem.

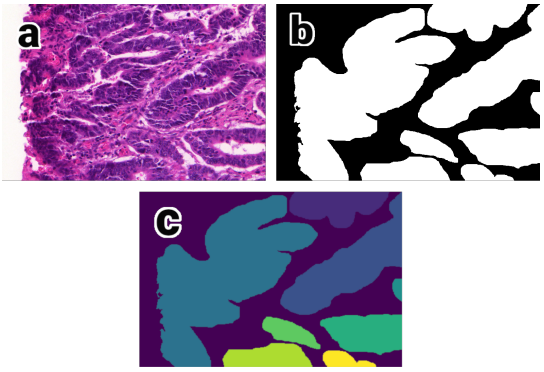

Figure 1: An example illustrating gland segmentation of an (a) H&E image using (b) semantic segmentation and (c) instance segmentation.

The objective of this paper was to improve the results of current gland instance segmentation models, aswell as to reproducing the previous state of the art which is not publically available. We propose a novel model which utilizes heavy data augmentation, smart ground truth targets and random transformation sampling. We implemented our model, as well as the previous state of the art, using PyTorch (Paszke et al. 2017).

2 Related Work

2.1 DCAN

DCAN (Chen et al. 2016) was the model which reached first place in the MICCAI 2015 Gland Segmentation Challenge (Sirinukunwattana et al. 2016). It used an FCN-style (Long, Shelhamer, and Darrell 2014) model architecture featuring two separate modules to predict the binary gland segmentation map and contour map simultaneously. The network was trained end to end through backpropagation by minimizing the sum of the softmax cross entropy loss, an auxiliary loss and L2 regularization. At test time the predicted gland contours were removed from the segmentation mask to separate inidividual gland instances, thus increasing detection results while decreasing segmentation accuracy. The model also employed transfer learning by pretraining on the Pascal VOC (Everingham et al. 2015) dataset.

2.2 Xu et al. 2016a

Xu et al. 2016a built on the work of DCAN by improving the contour detection with multi-scale deep supervision. Due to the information loss caused by the downsampling layers in the FCN structure, their proposed network uses a hierarchical structure using features from each layer in the network at multiple scales. This helped improve the predicted glandular boundaries.

2.3 Xu et al. 2016b

Instead of only predicting a segmentation and contour mask, Xu et al. 2016b added a third network head consisting of a Faster RCNN object detector (Ren et al. 2015) predicting bounding boxes. The final model combines its three network heads into a single network, predicting a binary gland segmentation mask, a binary gland contour mask, and a set of bounding boxes specifying the four coordinates of every predicted object. All three heads are used to create one final instance segmentation map.

2.4 MILD-Net

MILD-Net (Graham et al. 2018) presented a minimal information loss dilated network which utilized downsampling the input image and incorporation of its features in deeper layers of the network as a method of minimizing the information loss caused by maxpooling. Furthermore, atrous spatial pyramid pooling (ASPP) (Chen et al. 2017) was used together with dilated residual convolutions in an attempt to increase segmentation accuracy for glands of different shapes and sizes.

The authors proposed random transformation sampling as a method to generate an instance-wise uncertainty score for each gland. Random transformations of the input image are sampled at test time to create a distribution of predictions. The mean and variance of the predictions are then processed and are used to model the uncertainty of each gland. As a post-processing step, glands with high variance are removed and classified as a negative model prediction. MILD-Net currently reports state of the art results on gland instance segmentation.

3 Method

3.1 Dataset

For the task of gland instance segmentation, the MICCAI 2015 Gland Segmentation (GlaS) Challenge dataset was used. It consists of 164 small image pathes, majority of which are of sizes 522×775 pixels, from Haematoxylin and Eosin (H&E) stained slides. The dataset is split into three sets, one training set consisting of 84 images, and two test sets: testA and testB containing 60 and 20 images respectively. The H&E images are labeled by expert pathologists and contain an approximate 1:1 split of benign and malignant glands. A model is evaluated across three criteria judging detection performance, segmentation accuracy, and shape similarity through F1-score, object Dice and Hausdorff distance respectively.

Formally, a detected object instance is classified as a true positive (TP) if the predicted glandular object intersects with at least 50% of a ground truth object. The gland is considered a false positive (FN) if the intersection over union is below 50%. Any glands not detected are classified as false negatives (FN). Formally, F1-score ($F1$) is defined as:

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{1}$$

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

Let G and P be sets of instance-specific ground truth gland and predicted gland segmentation maps respectively. Additionally, let $\hat{G}$ be the set of ground truth glands where element j maximally overlaps P_j, and $\hat{P}$ be the set of predicted glands where element i maximally overlaps G_i. When there are no overlapping objects, the gland with the smallest Hausdorff distance is picked. The object wide Hausdorff distance (H_{obj}) and object-wide Dice score (D_{obj}) is defined as:

$$H_{obj}(G, P) = \frac{1}{2}\left(\sum_{i=1}^{|P|} \alpha_i H(S_i, \hat{G}_i) + \sum_{j=1}^{|G|} \beta_j H(G_i, \hat{S}_i)\right) \tag{4}$$

$$D_{obj}(G, P) = \frac{1}{2}\left(\sum_{i=1}^{|P|} \alpha_i D(S_i, \hat{G}_i) + \sum_{j=1}^{|G|} \beta_j D(G_i, \hat{S}_i)\right) \tag{5}$$

where

$$H(G, P) = \max\left(\sup_{x \in G} \inf_{y \in P} ||x - y||, \sup_{y \in P} \inf_{x \in G} ||x - y||\right) \tag{6}$$

$$D(G, P) = \frac{2|G \cap P|}{|G| + |P|} \tag{7}$$

$$\alpha_i = \frac{|P_i|}{\sum_{j=1}^{|P|} |P_j|} \tag{8}$$

$$\beta_j = \frac{|G_j|}{\sum_{k=1}^{|G|} |G_k|} \tag{9}$$

3.2 Models

We propose a novel deep convolutional neural network architecture which outputs three independent feature maps. The model uses uses a deep U-Net (Ronneberger, Fischer, and Brox 2015) style architecture with a ResNet-50 (He et al. 2015) encoder and decoder with skip connections to make use of lower level features and to reduce vanishing gradients. The *conv1* to *conv5_x* blocks are used from the ResNet-50 and are pretrained on ImageNet. Following the work on DeepLabV3 (Chen et al. 2017), the networks information loss is minimized by not downsampling more than a factor of 8 in the whole network. This is done by changing the stride of the *conv4_x* and *conv5_x* blocks in the encoder to be equal to one. Additionally, we employ Atrous Spatial Pyramid Pooling (ASPP) to increase segmentation accuracy with dilated convolutions (see figure 2).

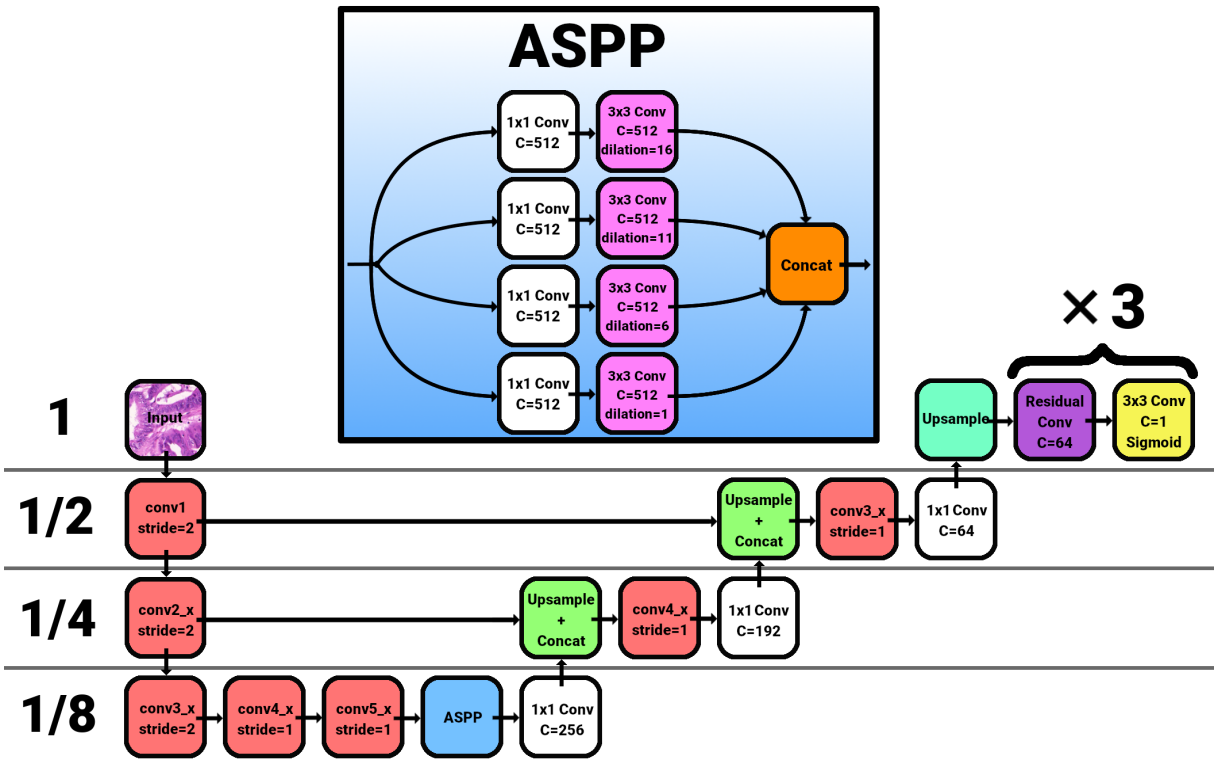

Figure 2: The model features an encoder-decoder architecture, downsampling and upsampling the input image by a factor of 8 over the course of the forward pass. The red blocks symbolize pretrained layers from ResNet-50 (He et al. 2015), where the stride of the *conv4_x* and the *conv5_x* blocks were changed to be equal to 1. ASPP denotes Atrous Spatial Pyramid Pooling. Every convolution is preceeded by batch normalization (Ioffe and Szegedy 2015). Bilinear interpolation is used as upsampling.

The network utilizes three independent heads at the end of the network to generate three separate feature maps predicting the binary gland segmentation mask, the binary contour mask of the edges of the glands, and a third mask of overlapping contour regions where there's a high probability that two glands are close in proximity.

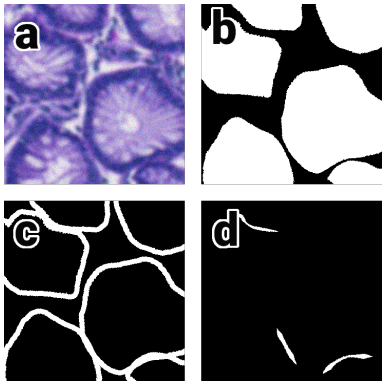

Figure 3: A figure illustrating a training example and ground truth including the (a) H&E image, (b) segmentation map, (c) contour map, and (d) overlapping contour regions.

The contour ground truth targets are generated by morphologically dilating each separate gland instance and subtracting the areas overlapping with the original segmentation mask. The final ground truth contour mask is formed by merging all gland contours (see figure 3). This way the predicted gland contours will minimally overlap with the gland segmentation map at test time. The model's final head predicts a feature map where two or more gland instance's contour regions overlap. The overlapping contour regions are removed from the segmentation mask during post-processing to increase gland instancing accuracy without decreasing segmentation performance.

For each network head the binary cross entropy (BCE) loss is computed with their respective ground truth labels, denoted by L_s for the gland segmentation mask, L_c for the contour mask, and L_o for the overlapping contour region mask. The network is trained by minimizing the total loss L_{total}, which is the sum of the binary cross entropy losses plus weighted L2-regularization. The constant λ was set to 1e−3:

$$L_{total} = L_s + L_c + L_o + \gamma||W||_2^2 \tag{10}$$

$$BCE = -y \log \hat{y} - (1 - y) \log(1 - \hat{y}) \tag{11}$$

The second model used was our implementation of the previous state of the art: MILD-Net. Many technical details

needed to fully reproduce MILD-Net were omitted from their paper, including but not limited to: no post-processing steps, no instancing algorithms, and an undefined auxiliary loss. As a results, we tested multiple methods and used hyperparameter tuning on the MILD-Net model for it to be able to be a fair fair comparison. We trained the MILD-Net model with the exact same architecture and loss function L_{MILD} as reported in their paper, with a total loss consisting of the segmentation loss L_s, the contour loss L_c and the auxiliary loss L_a. Softmax cross entropy was employed for L_s and L_c. The undefined auxiliary loss was interpreted as upsampling four feature maps from the middle of the network by a factor of 8 and computing the log loss with the contour and segmentation labels. The auxiliary loss was weighted by the variable w which was reduced by a factor of 10 after every 8th epoch, as the original paper states.

$$L_{MILD} = L_s + L_c + wL_a + \gamma||W||_2^2 \qquad (12)$$

Momentum and Adam optimizers were used to train the networks. A batch size of 3 was used for all experiments. The training time and learning rate varied depending on the amount of data augmentation and network architecture. We trained our best performing proposed model for 5000 iterations using a momentum of 0.9 and a learning rate of $1e{-}2$, decreasing it by a factor of 10 after 3000 and 4500 iterations.

3.3 Data Augmentation

Because of the small size of the dataset, machine learning models overfit the training data after just 3 epochs. This makes data augmentation the most important component when constructing a machine learning algorithm, considerably more impactful than the model architecture, and is often an overlooked detail in technical papers. To stress the importance of data augmentation, we made experiments showing that extensive data augmentation is required to obtain competitive results on the dataset.

Data augmentation techniques used in this paper are visualized in figure 4 and include random image crops, elastic warping, arbitrary $\theta \in [0, 360)$ degree rotations, gaussian, mean, and median blurs, unsharp masking, horizontal and vertical flipping, brightness shifting, contrast shifting, saturation shifting, hue shifting, uniform random noise, and random line of symmetry mirroring.

For each training example in a mini-batch, a 352×352 image patch is sampled, rotated by $\theta \in [0, 360)$ degrees, and warped elastically. Random line of symmetry augmentation entails drawing a random straight vertical, horizontal or vertical line across the image, and mirroring the image across the marked line of symmetry. Color jitter in the form of brightness, contrast, saturation, and hue shifting is later applied. One of the four different blurs is then randomly picked and applied to the image. Finally, random pixel noise is added to the training example, chosen uniformly in the range $[-x, x], x \in [15, 30]$.

3.4 Random Transformation Sampling

Random Transformation Sampling (RTS) (Graham et al. 2018) was employed at test time to increase the accuracy

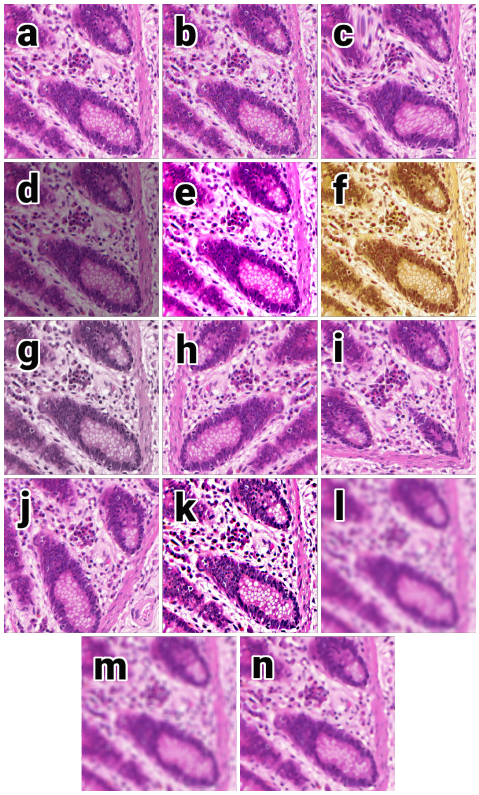

Figure 4: A visualization of the different types of data augmentation used (zoom in to view). They include (a) random cropping, (b) uniform noise, (c) elastic warping, (d) brightness shifting, (e) contrast shifting, (f) hue shifting, (g) saturation shifting, (h) mirroring, (i) line of symmetry, (j) rotation, (k) unsharp filtering, (l) gaussian blurring, (m) mean blurring, and (n) median blurring

Data Augmentation	Combined Test AB Set		
	F1	Dice	Hausdorff
Crop (Baseline)	0.786	0.804	115.55
Mirroring	0.804	0.847	112.11
Brightness Shifting	0.802	0.855	112.32
Contrast Shifting	0.793	0.833	114.84
Saturation Shifting	0.788	0.836	114.28
Hue Shifting	0.794	0.845	113.54
Uniform Noise	0.802	0.853	112.29
Gaussian Blur	0.815	0.851	111.90
Mean Blur	0.814	0.849	111.00
Median Blur	0.807	0.840	115.17
Unsharp Mask	0.802	0.840	114.35
Arbitrary Rotation	0.834	0.861	105.92
Elastic Warping	0.820	0.860	109.75
Line of Symmetry	0.800	0.831	111.10
All Augmentations	**0.892**	**0.876**	**77.40**

Table 1: **Data augmentation**. A table showing the performance of a single type of data augmentation used when training the proposed model.

of the models. RTS entails sampling n random transformations $\{f_1, f_2, ..., f_{n-1}, f_n\}$, and applying them to the input image at test time to generate a distribution of predictions. In our experiments random 90 degree rotations, horizontal mirroring, vertical mirroring, gaussian, mean, and median blurring, unsharp masking, brightness, contrast, hue, and saturation shifting, and uniform noise was used. The mean μ and variance σ^2 of the predictions can then be computed:

$$\mu = \frac{1}{n}\sum_{i=1}^{n} f_i(X, W) \tag{13}$$

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(f_i(X, W) - \mu)^2 \tag{14}$$

Here X is the input image and W are the network weights. This allows the model to compute where it is uncertain in its prediction. Let P be a set of predicted gland instances. The gland uncertainty score τ_k for instance k is defined as the expected variance of the instance P_k. Here N is the size of the predicted gland:

$$\tau_k = \frac{1}{N}\sum_{x,y \in P_k} \sigma^2_{x,y} \tag{15}$$

Glands with an uncertainty score below a threshold t are removed in a fully automated production environment, or can be passed on to a pathologist for further evaluation when used in conjunction with medical experts.

3.5 Post-processing

Heavy post-processing, similar to data augmentation, was found to be necessary for good results. An H&E image was first processed by the convolutional neural network with 50 RTS samples to predict the segmentation, contour and overlap binary maps. Morphological erosion and dilation, also known as morphological opening, was used to remove noise. The mean contour overlap regions were then subtracted from the newly modified segmentation map, and was finally thresholded with constant $h = 0.55$ to create the final binary segmentation map.

Connected component analysis was applied to make gland instances. As a final post-processing step, instances of sizes 35^2 pixels and instances with uncertainty greater than $t = 0.08$ were removed.

4 Experiments and Results

The first experiments which were conducted studied the effects of data augmentation on the model. For these experiments we used our proposed network architecture. A baseline was established using only random image crops. Every other type of data augmentation augments the cropped image. RTS was not employed and results show the mean of three independently trained models for each data augmentation technique. Performance was measured on the combined testAB set. Results are shown in table 1.

An ablation study was carried out with our proposed model as a baseline. We removed one variable at a time and observed its effect on the final result. These variables include

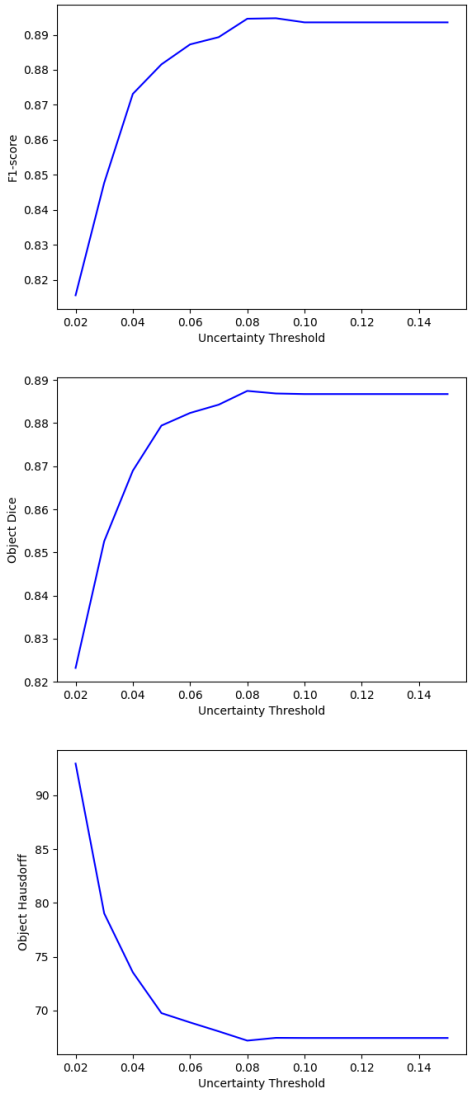

Figure 5: Graphs of the F1-score, object Dice, and object Hausdorff as a function of uncertainty threshold).

ASPP, increasing the downsampling (strides) in the encoder and upsampling decoder, using no ImageNet pretraining, using transposed convolutions instead of bilinear interpolation, and not using RTS. We also include our results from the reproduced MILD-Net. Table 2 presents the ablation study results.

The ablation study confirms that RTS, ASPP, limited input downsampling, transfer learning and bilinear interpolation are all critical components which increase the performance of the model. While ImageNet pretraining only slightly increases accuracy, the training time decreased from 20 000 to 5 000 iterations. Varying the gland uncertainty threshold t used in RTS also affects accuracy. In figure 5 we illustrate the F1-score, object Dice, and object Hausdorff as a function of the uncertainty threshold t. Best results were observed at $t = 0.08$.

Model	Test A			Test B		
	F1	Dice	Hausdorff	F1	Dice	Hausdorff
Baseline (RTS)	**0.916**	**0.900**	**58.32**	**0.821**	**0.861**	**87.27**
No ASPP	0.890	0.888	66.28	0.745	0.811	123.89
Stride=2 for conv4_x and conv5_x	0.852	0.847	75.15	0.702	0.0.731	130.58
No ImageNet pretraining	0.907	0.895	60.37	0.816	0.841	98.83
Transposed Conv	0.875	0.867	70.89	0.728	0.725	127.56
No RTS	0.910	0.900	58.52	0.788	0.819	109.85
Reproduced MILD-Net	0.808	0.853	59.87	0.628	0.760	136.97

Table 2: **Ablation study results**. A table presenting the results of the ablation study. All variables are kept constant except a single architectural design decision. Variables tested include ASPP, RTS, ImageNet transfer learning, network downsampling depth, and convolutional upsampling. Our reproduced MILD-Net is also shown. The best results are in bold.

Model	Test A			Test B		
	F1	Dice	Hausdorff	F1	Dice	Hausdorff
Proposed	**0.916**	0.900	58.32	0.821	**0.861**	**87.27**
Reproduced MILD-Net	0.808	0.853	59.87	0.628	0.760	136.97
MILD-Net	0.914	**0.913**	**41.54**	**0.844**	0.836	105.89
Xu et al. 2016b	0.893	0.908	44.13	0.843	0.833	116.82
MIMO-Net	0.913	0.906	49.15	0.724	0.785	133.98
Xu et al. 2016a	0.858	0.888	54.20	0.771	0.815	129.93
DCAN	0.912	0.897	45.42	0.716	0.781	160.35
ExB1	0.891	0.882	57.41	0.703	0.786	145.58
Freidburg2	0.870	0.876	57.09	0.695	0.786	148.47
CVML	0.652	0.644	155.43	0.541	0.654	176.24
LIB	0.777	0.781	112.71	0.306	0.617	190.45
vision4GlaS	0.635	0.737	107.49	0.527	0.610	210.10

Table 3: **GlaS comparison**. A comparison of the F1-score, object Dice, object Hausdorff distance for previous state of the art models (Sirinukunwattana et al. 2016) on the testA and testB splits of the GlaS dataset. The best result in a single category is marked in bold.

We were unable to reproduce the reported results of MILD-Net, suggesting that the details which were omitted from their paper were critical for cutting edge model performance. Our experiments showed that when keeping all variables constant except the model architecture, our network architecture outperformed MILD-Net. The results from the gland uncertainty modeling with our proposed model widely differs from what MILD-Net presented. While RTS uncertainty sampling did increase the accuracy of our model, it did not drastically improve our results, compared to the major increases in accuracy which MILD-Net presented by using very low gland uncertainty thresholds.

Additionally, while both our and MILD-Net's formulation of RTS and uncertainty thresholding is equivalent, our reported gland uncertainty scores differ by two orders of magnitude. Since each pixel segmentation is bounded by the range [0, 1], the maximal pixel variance, and thus gland uncertainty τ_k, is upper-bounded by 0.25. MILD-Net reports gland uncertainties in the approximate range [0, 9], suggesting that their implementation of RTS is different from the formal formulation of RTS presented in this paper. This difference might be the cause of the performance differences between our proposed model and the previous state of the art.

Our best performing model beat the current state of the art in gland instance segmentation in F1-score on test set A, and object Dice and object Hausdorff on test set B. Qualitative results of our proposed model can be viewed in figure 6, including the statistical measures computed through RTS.

5 Conclusion

We proposed a novel automated state of the art gland instance segmentation model to reduce the amount of time and resources needed for expert pathologists to annotate histological images, crucial for the detection of cancer. Our model utilized overlapping contour regions for precise detection of glandular boundaries. Additionally, we employed Random Transformation Sampling as a method to model instance specific gland uncertainty at test time, increasing the final detection and segmentation accuracy. This allows glands with an uncertainty score below a threshold to be removed in a fully automated production environment, or to be passed on to a pathologist for further evaluation when used in conjunction with human experts. We additionally implemented the previous state of the art in gland instance segmentation, MILD-Net, and failed to reproduce their reported results.

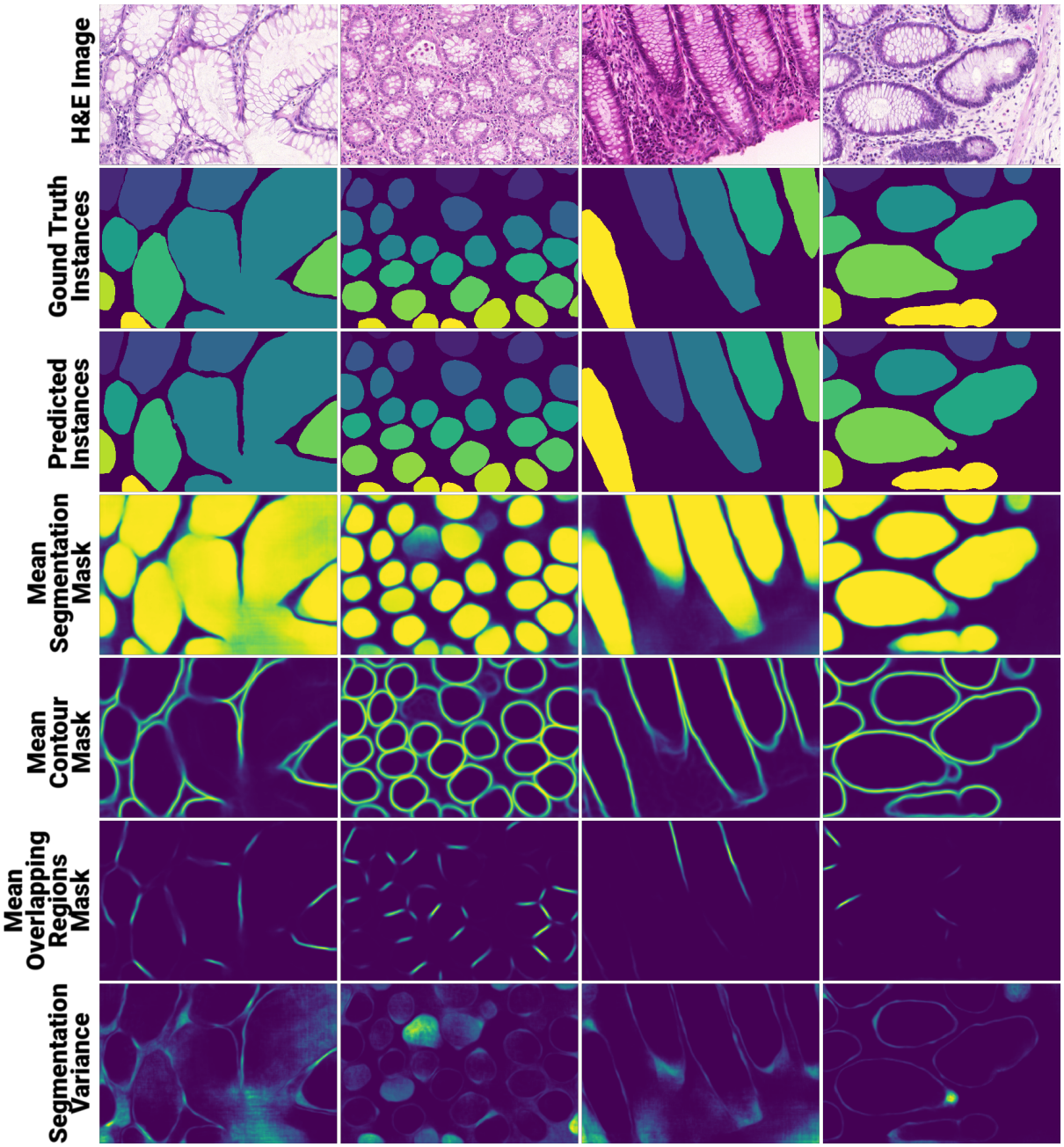

Figure 6: A figure showing predictions of our proposed model used on four different test examples. The image includes the H&E slide crop, the ground truth instances, the predicted instances, the mean segmentation mask, the mean contour mask, the mean overlapping contour regions, and the variance of the segmentation mask.

References

Chen, H.; Qi, X.; Yu, L.; and Heng, P.-A. 2016. Dcan: Deep contour-aware networks for accurate gland segmentation.

Chen, L.; Papandreou, G.; Schroff, F.; and Adam, H. 2017. Rethinking atrous convolution for semantic image segmentation. *CoRR* abs/1706.05587.

Everingham, M.; Eslami, S. M. A.; Van Gool, L.; Williams, C. K. I.; Winn, J.; and Zisserman, A. 2015. The pascal visual object classes challenge: A retrospective. *International Journal of Computer Vision* 111(1):98–136.

Graham, S.; Chen, H.; Dou, Q.; Heng, P.-A.; and Rajpoot, N. 2018. Mild-net: Minimal information loss dilated network for gland instance segmentation in colon histology images.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2015. Deep residual learning for image recognition. *CoRR* abs/1512.03385.

Ioffe, S., and Szegedy, C. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of the 32Nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, 448–456. JMLR.org.

Long, J.; Shelhamer, E.; and Darrell, T. 2014. Fully convolutional networks for semantic segmentation. *CoRR* abs/1411.4038.

Paszke, A.; Gross, S.; Chintala, S.; Chanan, G.; Yang, E.; DeVito, Z.; Lin, Z.; Desmaison, A.; Antiga, L.; and Lerer, A. 2017. Automatic differentiation in pytorch.

Ren, S.; He, K.; Girshick, R. B.; and Sun, J. 2015. Faster R-CNN: towards real-time object detection with region proposal networks. *CoRR* abs/1506.01497.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. *CoRR* abs/1505.04597.

Sirinukunwattana, K.; Pluim, J. P. W.; Chen, H.; Qi, X.; Heng, P.-A.; Guo, Y. B.; Wang, L. Y.; Matuszewski, B. J.; Bruni, E.; Sanchez, U.; Böhm, A.; Ronneberger, O.; Cheikh, B. B.; Racoceanu, D.; Kainz, P.; Pfeiffer, M.; Urschler, M.; Snead, D. R. J.; and Rajpoot, N. M. 2016. Gland segmentation in colon histology images: The glas challenge contest.