Instance segmentation for whole slide imaging: end-to-end or detect-then-segment

Aadarsh Jha,[a] Haichun Yang,[b] Ruining Deng,[a] Meghan E. Kapp,[b] Agnes B. Fogo,[b] and Yuankai Huo[a,*]

[a]Vanderbilt University, Department of Electrical Engineering and Computer Science, Nashville, United States

[b]Vanderbilt University Medical Center, Department of Pathology, Microbiology and Immunology, Nashville, United States

Abstract

Purpose: Automatic instance segmentation of glomeruli within kidney whole slide imaging (WSI) is essential for clinical research in renal pathology. In computer vision, the end-to-end instance segmentation methods (e.g., Mask-RCNN) have shown their advantages relative to detect-then-segment approaches by performing complementary detection and segmentation tasks simultaneously. As a result, the end-to-end Mask-RCNN approach has been the *de facto* standard method in recent glomerular segmentation studies, where downsampling and patch-based techniques are used to properly evaluate the high-resolution images from WSI (e.g., $>10,000 \times 10,000$ pixels on 40×). However, in high-resolution WSI, a single glomerulus itself can be more than 1000×1000 pixels in original resolution which yields significant information loss when the corresponding features maps are downsampled to the 28×28 resolution via the end-to-end Mask-RCNN pipeline.

Approach: We assess if the end-to-end instance segmentation framework is optimal for high-resolution WSI objects by comparing Mask-RCNN with our proposed detect-then-segment framework. Beyond such a comparison, we also comprehensively evaluate the performance of our detect-then-segment pipeline through: (1) two of the most prevalent segmentation backbones (U-Net and DeepLab_v3); (2) six different image resolutions (512×512, 256×256, 128×128, 64×64, 32×32, and 28×28); and (3) two different color spaces (RGB and LAB).

Results: Our detect-then-segment pipeline, with the DeepLab_v3 segmentation framework operating on previously detected glomeruli of 512×512 resolution, achieved a 0.953 Dice similarity coefficient (DSC), compared with a 0.902 DSC from the end-to-end Mask-RCNN pipeline. Further, we found that neither RGB nor LAB color spaces yield better performance when compared against each other in the context of a detect-then-segment framework.

Conclusions: The detect-then-segment pipeline achieved better segmentation performance compared with the end-to-end method. Our study provides an extensive quantitative reference for other researchers to select the optimized and most accurate segmentation approach for glomeruli, or other biological objects of similar character, on high-resolution WSI.

© *2021 Society of Photo-Optical Instrumentation Engineers (SPIE)* [DOI: 10.1117/1.JMI.8.1.014001]

1 Introduction

Understanding the underlying details of glomerular morphology through renal biopsy evaluation provides insights into various renal disorders.[1–3] Glomerular count and size are critical measurements in renal physiology. Abnormally large glomeruli, called glomerular hypertrophy,

---

is a hallmark of kidney injury in obesity-associated glomerulopathies and diabetic nephropathy.[4] The golden standard of characterizing glomerular size is to manually trace the contour of each glomerulus to achieve segmentation masks.[3] However, manual quantification of glomerular number and size requires exhaustive resources and is not scalable. In recent years, there has been a paradigm shift toward automatic glomerular instance segmentation, which aims to provide instance-level, pixel-wise annotation for each glomerulus driven by convolutional neural networks (CNNs).[5,6] The *de facto* standard method of instance segmentation of glomeruli, and more broadly the kidney, is Mask-RCNN,[7,8] an end-to-end pipeline which performs detection and instance segmentation simultaneously.[9] Since the end-to-end architecture of Mask-RCNN is designed for natural images (e.g., $\approx 1000 \times 1000$ pixels), both downsampling and tiling are utilized in order to leverage processing speeds and fit modern GPU memory when Mask-RCNN is applied to high-resolution whole slide imaging (WSI) (e.g., $>10,000 \times 10,000$ pixels on 40$\times$). However, a loss of information is often associated with the downsampling process that is inherent to the end-to-end framework of Mask-RCNN. In particular, a single glomerulus from a WSI can be more than 1000×1000 pixels in image resolution, which yields significant information loss when the corresponding features maps are downsampled to the 28×28 resolution via the end-to-end Mask-RCNN segmentation head,[7] as demonstrated in Fig. 1. Thus the prevalent end-to-end instance segmentation method might not be the best solution for high-resolution WSI. When reimagining glomerular instance segmentation for high-resolution WSI, an intuitive idea of addressing the trade-off between resolution and accuracy would be the fundamental separation of detection and segmentation. In this detect-then-segment manner, detection could be performed on downsampled tiles for computational efficiency, whereas segmentation could be conducted on high-resolution images as unrelated pixels are excluded by detection. Inspired by this rationale, we aim to explore if the end-to-end or detect-then-segment framework is optimal for high-resolution WSI objects in the context of renal pathology.

In this study, we propose a detect-then-segment framework for glomerular instance segmentation in order to more broadly improve current instance segmentation techniques when applied to high-resolution WSI. In our study, we utilize two distinct high-resolution segmentation networks for semantic segmentation, and we use Mask-RCNN for instance glomerular detection. A central focus of our study is to compare our proposed detect-then-segment framework to the performance of the end-to-end Mask-RCNN pipeline on high-resolution WSI. In addition, we conduct extensive analyses to ascertain the best detect-then-segment strategy through two of the most widely used segmentation backbones (U-Net and DeepLab_v3), six unique resolutions
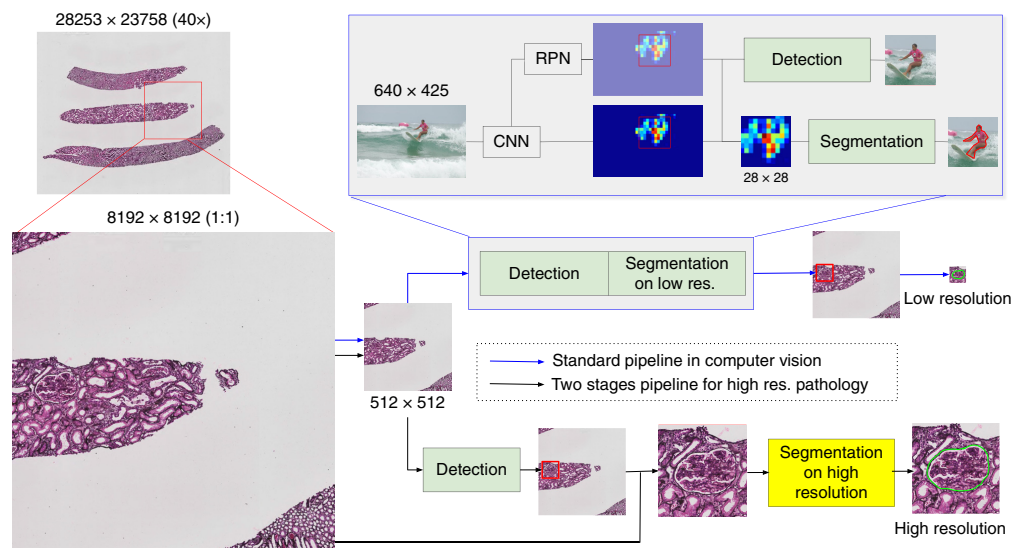
Fig. 1 The end-to-end Mask-RCNN instance segmentation pipeline in blue arrows and the proposed detect-then-segment framework in black arrows. In our proposed method, the two-stage detect-then-segment strategy is used, where detection will first occur on downsampled images, and then segmentation is performed on high-resolution objects.

(512 × 512, 256 × 256, 128 × 128, 64 × 64, 32 × 32, and 28 × 28), and two distinct color spaces (RGB and LAB). To the best of our knowledge, no previous studies have comprehensively evaluated glomerular segmentation performance comparing detect-then-segment and end-to-end strategies.

To evaluate the performance of these two distinct segmentation frameworks, we divided our experiments into two scenarios: (1) manual detection and (2) automatic detection. In the first scenario—labeled as "manual detection"—manual detection results (bounding boxes) were used to evaluate the segmentation performance in our detect-then-segment framework. Then in the second scenario—labeled as "automatic detection"—automatic detection results from the same Mask-RCNN detection head were used to compare segmentation performance across end-to-end and detect-then-segment strategies. The key difference in our automatic detection phase is the use of either an end-to-end Mask-RCNN segmentation head or an additional high-resolution segmentation head for glomerular instance-level segmentation. Decoupling detection and segmentation allows for more freedom in understanding how to improve the segmentation of glomeruli, and more broadly, high-resolution WSI objects of similar character.

For the manual detection phase, we trained the segmentation networks of our proposed detect-then-segment method using 704 manually traced training glomerular images from 42 biopsy samples. Meanwhile, 98 validation glomerular images, 147 internal testing images, and 385 external testing images were manually extracted from 7, 7, and 5 WSI images, respectively, to evaluate segmentation performance. The original resolution of our glomeruli image data was of ≈1000 × 1000 pixels. To be compatible with our GPU memory, we first scaled all input images down to the resolution of 512 × 512. Then according to our procedure, we further scaled down the input images to the resolutions of 512 × 512, 256 × 256, 128 × 128, 64 × 64, 32 × 32, and 28 × 28 to train and evaluate two of the most widely used segmentation backbones, U-Net and DeepLab_v3.

For the automatic detection experiments, we compared performance between the end-to-end Mask-RCNN segmentation and our proposed detect-then-segment approach using automatic detection results. To do so, Mask-RCNN was trained and validated on the same biopsy WSI images as the manual detection experiments. From the 4 internal testing WSI, 120 glomeruli were correctly detected (IOU > 0.5 compared with true bounding box) from the detection head in Mask-RCNN. Then we directly applied the trained Mask-RCNN segmentation head as well as our trained segmentation models (U-Net and DeepLab_v3) from our first phase to the same 120 glomerular detected images to compute the final segmentation performance when the detection is fairly provided. The experiments show that our detect-then-segment framework, under the automatic detection scenarios, achieves a DSC value of 0.953, whereas Mask-RCNN provides a lower DSC value of 0.902.

Our contributions, as listed as follows, do not claim algorithmic novelty over prior arts but rather investigate the problems overlooked in the previous works.

- Proposing a new detect-then-segment glomerular instance segmentation framework by performing instance detection and semantic segmentation on different resolutions with a coarse-to-fine design to avoid extreme downsampling for high-resolution glomerular segmentation in renal pathology WSI.
- Evaluating if the *de facto* end-to-end design or our detect-then-segment approach is optimal for segmenting glomeruli in high-resolution WSI.
- Performing extensive analyses by varying the image resolution, color space, and segmentation framework in segmenting previously detected glomeruli image objects. This comprehensive analysis allows us to provide an extensive quantitative reference for other researchers to select the optimized segmentation approaches for glomeruli, or other biological objects of similar character, on high-resolution WSI.

2 Related Works

The introduction of WSI demonstrates a shift toward computer-aided diagnosis (CAD) techniques to more accurately characterize critical objects. The use of WSI and its associated analysis techniques has been shown to be effective and even expanding in the field of renal pathology.[10]

To properly distinguish and characterize different glomeruli within renal biopsy samples, modern deep learning techniques of detection and segmentation have been utilized. Several studies have shown the great accuracy by which CNNs are able to properly detect and localize glomeruli within sample images.[6,11–13] Similarly, CNNs have also been able to accurately segment glomeruli, allowing for normal and sclerosed glomeruli to be properly distinguished.[14–19] Other studies have further combined the process of the detection and segmentation of glomeruli.[20] Of course, the end-goal of deep learning in renal imaging is its application in CAD. In this regard, several studies have also shown the ability to perform diagnoses based on the preliminary quantification and characterization of glomerular data through deep learning.[21,22] Common to all of the above studies is the application of CNNs to localize, detect, segment, or characterize glomeruli to better understand renal pathology. The uniqueness of our research presents itself by identifying the specific techniques that work best for high-resolution glomeruli data, rather than using common solutions to the niche field of renal imaging. Our paper analyzes several factors that work best in the specific context of instance segmentation of high-resolution glomerular data, and other biological objects of similar character and size. Additionally, we further propose a pipeline that is different than the conventional end-to-end instance segmentation tactics that are often used in medical imaging and computer vision, so as to yield better and more accurate results.

3 Method

Generally, the methodology followed in this study can be broken down into two major steps: (1) detection and (2) segmentation. Within detection, we discuss our approach toward manual and automatic detection of glomeruli; on the other hand, within segmentation, we demonstrate how we comprehensively analyzed our detect-then-segment framework, and the steps taken to compare it to a classic end-to-end Mask-RCNN pipeline. Our detect-then-segment approach can be seen visually in Fig. 2.
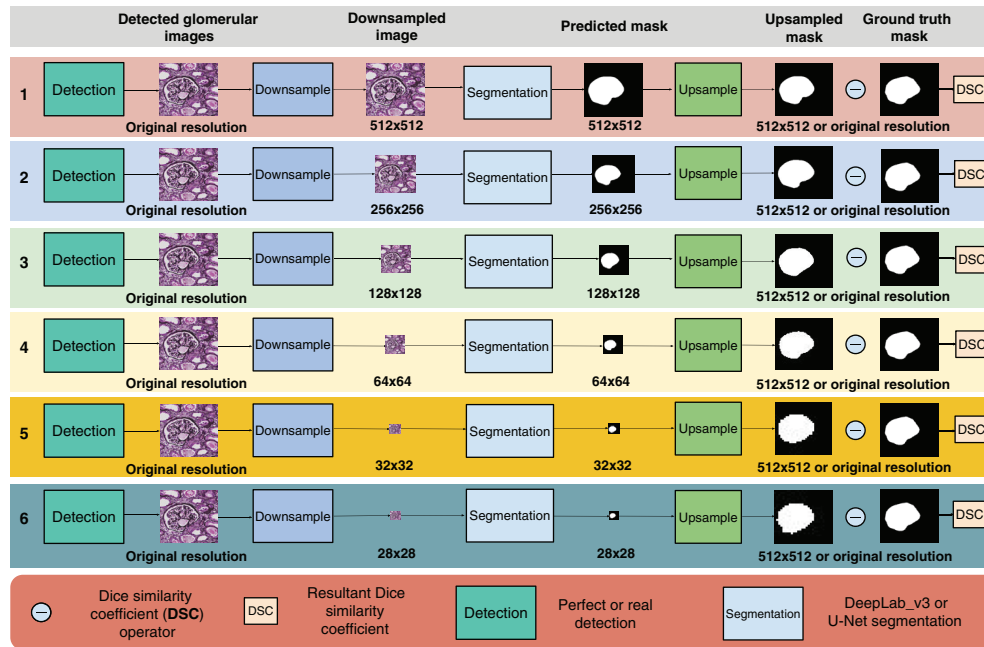
3.1 Detection

In the manual detection portion of our study, the manually traced bounding boxes for glomeruli are used to provide the ideal detection results. In order to introduce more background variation and avoid the problematic situation in which the glomeruli is always in the middle of the detection, we randomly expanded the detection bounding boxes to 1.5 times the longest dimension of the manual boxes with random center shift. We ensured that the image still contained the complete glomerulus.

In the automatic detection portion, Mask-RCNN[9] was employed as the detection method. The feature pyramid network[23] with ResNet-101[24] is used as the feature extraction backbone. The default Mask-RCNN implementation (https://github.com/facebookresearch/maskrcnn-benchmark) was used during training. For all training and testing within detection, the original high-resolution WSI (0.25 μm per pixel) was downsampled to a lower resolution (4 μm per pixel), given the size of a glomerulus[25] as well as its ratio within a patch. Then we randomly tiled the image patches (where each patch contained at least one glomerulus with 512×512 pixels) as experimental images for our detection networks. Eventually, we formed a cohort with 7040 training images with manual segmentation masks for training the Mask-RCNN glomerular detection.

3.2 Segmentation

3.2.1 Manual detection

In this study, a standard implementation of the U-Net and DeepLab_v3 architectures was used to perform segmentation on the glomeruli image data in the manual detection phase of our experiment. In particular, U-Net and DeepLab_v3 were trained with the preprocessed images as described in Sec. 3.1. The input image data for both segmentation frameworks contained three input channels (RGB or LAB), and the output data contained two classes (foreground and background).

Fig. 2 An abstraction of the proposed detect-then-segment methodology used in our experiments. Each row indicates a different trial, where a previously detected glomerulus is downsampled to the distinct dimensions of 512×512, 256×256, 128×128, 64×64, 32×32, and 28×28. Then these downsampled images are passed through a segmentation network. Two of the most prevalent segmentation backbones (U-Net and Deeplab_v3) are used as the segmentation networks in this study. The predicted masks are produced, and then upsampled back to the initial and original resolution—which in our study was 512×512—of the glomerulus for a fair DSC comparison. Additionally, in the trial utilizing a U-Net backbone, we separately evaluated the images in both an RGB and LAB color space so as to understand the effect of color space on segmentation performance. In the Deeplab_v3 trial, we only evaluated the best performing color space from our U-Net experiment. Overall, the segmentation networks evaluated across six different resolutions, two unique color spaces, and two distinct segmentation backbones.

Limited by GPU memory, all original image resolution glomeruli ($>1000 \times 1000$) were initially scaled down to 512×512. Then this input image dataset was further scaled down to the sizes of 512×512, 256×256, 128×128, 64×64, 32×32, and 28×28. Once these images were downsampled, the training images were further represented in either the standard RGB or LAB image space. The LAB image space was evaluated as it was recently shown to confer the best performance for basic image classification tasks by reducing image channel-wise correlation.[26] Data augmentation was also performed for image segmentation, where 50% of the training images were altered through channel shuffling, translation, rotation, sheer, left-to-right flipping, and Gaussian blur.

Two of the most prevalent segmentation backbones (U-Net and DeepLab_v3) were employed in this study. Briefly, the U-Net architecture is an end-to-end fully convolutional network. In terms of general network architectures, U-Net can be divided into two major portions: (1) encoder and (2) decoder. The encoder contains both convolutional and max pooling layers which obtain greater context of the input image through downsampling, allowing for the encoding of the input image into feature representations at multiple different resolutions. The second path is the decoder, which symmetrically expands and upsamples the input image. This allows for precise localization using bilinear interpolation and effectively rescales the feature map to the original image size.[27] Similarly, DeepLab_v3 also has encoder and decoder stages. However, in its encoder phase, DeepLab_v3 utilizes Atrous, or dilated, convolution to obtain greater context of the input image. The decoder phase then follows to create and rescale the feature map of the original image.[9] Through the manually detected glomerular images, we evaluated the performance of U-Net and DeepLab_v3 with the aforementioned designs.

3.2.2 *Automatic detection*

Finally, in the automatic detection phase of our experiment, the trained Mask-RCNN network was performed on all testing images to achieve 120 glomerular detection bounding boxes from 5 WSI biopsies in downsampled images. Then the bounding box coordinates were upscaled to the original image resolution ($>1000 \times 1000$ pixels) to crop the corresponding glomeruli and the masks in the highest resolution. Furthermore, both DeepLab_v3 and U-Net pretrained models were also applied on the same group of 120 images, which were downscaled to each of the tested resolutions (512×512, 256×256, 128×128, 64×64, 32×32, and 28×28). After corresponding predicted masks were generated, they were upsampled to the initial resolution to calculate the mean and median DSC scores from the manual masks.

3.3 *Data Analysis*

The DSC was primarily used to evaluate the performance of segmentation. To begin, in our manual detection experimentation—which comprised of the 704 training, 98 validation, 147 internal testing, and 385 external testing images—we evaluated the performance of U-Net and DeepLab_v3 segmentation across six different resolutions (512×512, 256×256, 128×129, 64×64, 32×32, and 28×28) and two different color spaces (RGB and LAB). In particular, for each epoch within the segmentation process for each resolution, DSC values were developed for the validation and testing data. For each tested resolution, the best epoch was selected via the highest DSC for the validation dataset, and the generated model in that epoch was saved. Then these generated, predicted masks for each tested resolution were upsampled to a 512×512 resolution to be compared against the initial ground truth mask data (which is also of 512×512 resolution) for the validation and testing images. Mean and median DSC, as well as standard deviation, were computed again for these upsampled image sets for each resolution. Figures 3 and 4 visually demonstrate three modes of DSC performance (good, average, and bad) across both U-Net and DeepLab_v3. Namely, the green overlay is the predicted image, and the background of the image is the ground truth input.

Throughout our study, in the specific context of our manual detection phase results, we draw a distinction between the terms of "sample space" and "512 space." We define "sample space" as the evaluation of the predicted images in each of the six tested resolutions (512×512, 256×256, 128×128, 64×64, 32×32, and 28×28) against the corresponding downsampled input images to produce a preliminary DSC value across six distinct resolutions. We similarly define "512 space" as the evaluation of the predicted images across the six tested resolutions
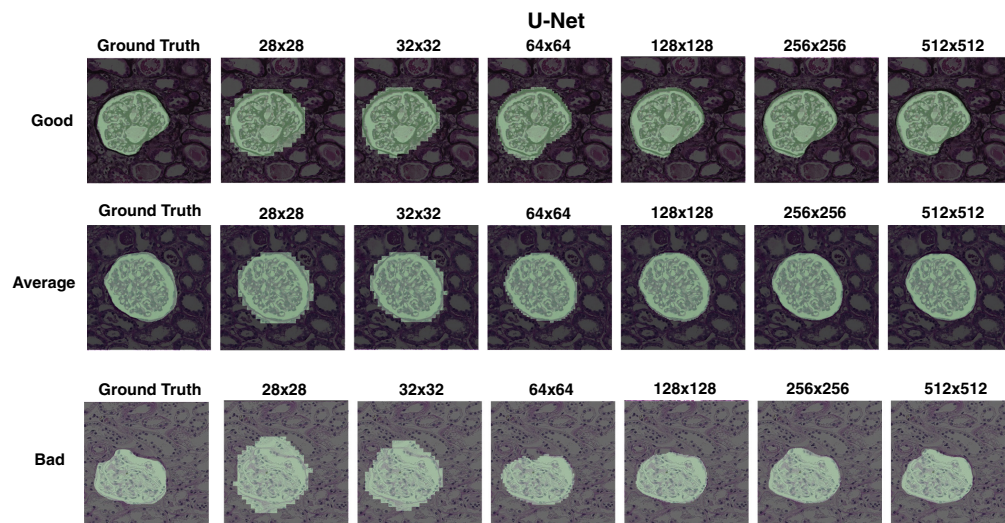
Fig. 3 Three categories of good, average, and bad performance which is defined by DSC. The background of each comparison is the ground truth input, and the overlaid image is the predicted mask. The results of U-Net are shown.
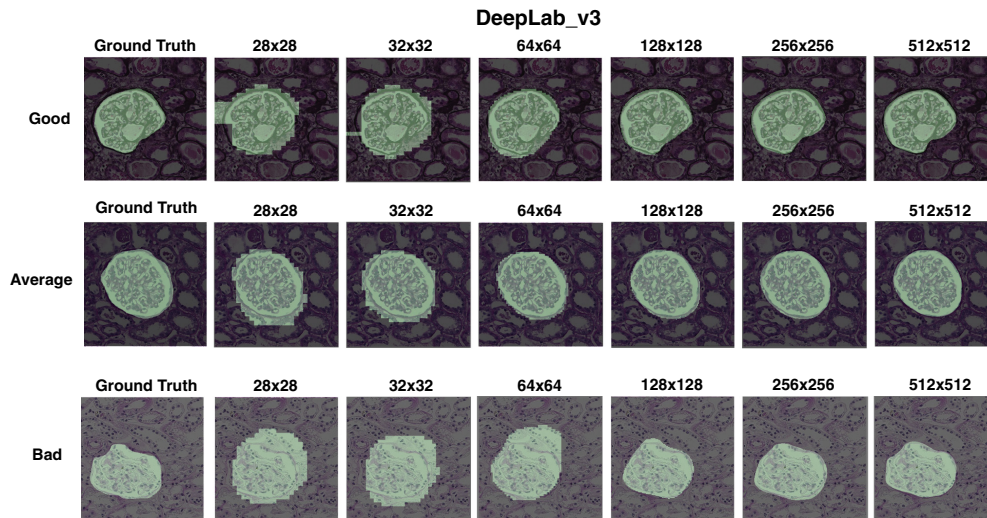
DeepLab_v3

Fig. 4 Three categories of good, average, and bad performance which is defined by DSC. The background of each comparison is the ground truth input, and the overlaid image is the predicted mask. The results of DeepLab_v3 are shown.

which are then upsampled to the original size of the input image—which in our study was 512×512 as established earlier—and then compared to the original resolution 512×512 input images to produce a fair DSC score. This can be seen visually in the latter columns of Fig. 2.

Furthermore, in our automatic detection experimentation, which comprised a cohort of 120 glomerular images in original resolution ($>1000 \times 1000$), we similarly applied the detect-then-segment framework which was directly compared to Mask-RCNN through the use of mean and median DSC, as well as the standard deviation of the DSC data. Through a similar process in the manual detection phase, we applied Mask-RCNN to our cohort of input images and produced relevant statistics for the DSC scores. Additionally, after applying both U-Net and DeepLab_v3 to each of the six tested resolutions of the input image and producing corresponding predicted masks, such predicted masks were then upsampled and compared to the original resolution of the input image ($>1000 \times 1000$) for a fair DSC comparison.

4 Experimental Design

4.1 Dataset

WSI from renal needle biopsies and human kidney nephrectomy tissues were utilized for analysis. The kidney needle biopsy was routinely processed, paraffin embedded, and 2 μm thickness sections cut and stained with hematoxylin and eosin (HE), periodic acid Schiff (PAS) or Jones. The human nephrectomy tissues were acquired from noncancerous tissue from patients with cancer. The tissue was routinely processed, paraffin embedded, and 3 μm thickness sections cut and stained with PAS. The data were deidentified, and studies were approved by the Institutional Review Board. All manual annotations of glomerular detection were performed by a renal pathologist with more than 15 years of clinical and research experience. Then the segmentation masks for the detected glomeruli were either traced by the same renal pathologist or first traced by a research associate and then confirmed by the renal pathologist. For the purposes of training and testing, the high-resolution WSI (0.25 μm per pixel) was downsampled to a lower resolution (4 μm per pixel). Then patches were identified which contained glomeruli in its original resolution ($>1000 \times 1000$ pixels). Images of glomeruli, as well as their manually traced ground truth masks, were then collected. In this study, these input images served as our previously detected glomerular images upon which segmentation then was performed. Eventually, we formed a cohort with 704 training, 98 validation, and 147 internal testing images. Additionally, a group of 385 images was used as external testing data. The training, validation,

and testing data were used in our manual detection experimentation. Finally, a separate cohort of 120 images with Mask-RCNN detected glomeruli was used to directly evaluate the performance of our proposed framework relative to Mask-RCNN. This set of 120 images was derived from 5 different patients with WSI of the kidney tissue and was utilized in our automatic detection experimentation.

4.2 Experimental Design

Our study was split into two distinct phases: manual detection and automatic detection. For our manual detection experimentation, 704 images were randomly chosen as testing images, whereas the remaining 98 images were used for validation. Additionally, 147 internal testing images were utilized alongside 385 external testing images from an independent cohort. On the other hand, for our automatic detection experimentation, 120 images, kept in their original resolutions ($>1000 \times 1000$ pixels), were used to evaluate the performance of our detect-then-segment framework relative to Mask-RCNN.

The U-Net, DeepLab_v3, and Mask-RCNN pipelines were deployed on a typical workstation with Intel Xeon CPU 2.2 GHz, 13 GB RAM, 33 GB Disk Space, 12 GB NVIDIA Tesla K80 GPU, and CUDA 10.1. For our manual detection experimentation, the hyperparameters of the U-Net and DeepLab_v3 pipelines were 150 epochs, a batch size of 4, a learning rate of 0.0001, a color space argument (RGB or LAB, depending on the trial), as well as a scale argument, which was altered to test the performance of segmentation across six distinct resolutions. Additionally, an Adam Optimizer was used to adaptively alter the learning rate, with beta values ranging from 0.9 to 0.999.

4.2.1 Manual detection

Within our manual detection experimentation, manually detected glomerular images were processed with two distinct color spaces, six different image resolutions, and two unique segmentation backbones. We began with a U-Net segmentation framework, where the RGB and LAB color spaces were studied. Within each color space, six resolutions were tested through the U-Net pipeline: (1) 512×512, (2) 256×256, (3) 128×128, (4) 64×64, (5) 32×32, and (6) 28×28. For each trial, 150 epochs were run for all training, validation, and testing images. Additionally, within each epoch, a DSC value was generated for the validation and testing images. For each resolution, the epoch with the highest DSC value for the validation data was recorded and its generated model was saved. With this generated model, predicted masks were created at each of the six resolutions analyzed for the validation and testing data. This process was repeated yet again for the LAB color space. After this experiment was completed for the U-Net framework, the previously described methodology was also repeated for the DeepLab_v3 framework, but the only color space analyzed for DeepLab_v3 was the best performing color space in the U-Net trial. If the difference in performance between the two color spaces was negligible in the U-Net trial, then we defaulted to RGB to analyze the DeepLab_v3 framework. This is because utilizing the RGB color space is standard, and the introduction of the LAB color space in our study was due to recent findings that show that the LAB color space produces better results for image classification tasks by reducing image channel-wise correlation.[26]

Once all predicted masks were generated, the performance of the U-Net and DeepLab_v3 segmentation networks was then analyzed by upsampling the predicted masks for the validation and testing images, and comparing them back to the original 512×512 ground truth mask. Mean and median DSC scores were computed as a result of the upsampling.

4.2.2 Automatic detection

In the automatic detection phase, the performance of our detect-then-segment framework was directly compared against Mask-RCNN through a cohort of 120 glomeruli images that were kept in their original resolution ($>1000 \times 1000$ pixels). In doing so, we utilized the model files that were generated for each of the six resolutions for both U-Net and DeepLab_v3 as described in Sec. 4.2.1. In particular, we first downsampled the 120 glomerular images to the scales of

512×512, 256×256, 128×128, 64×64, 32×32, and 28×28. Then we procured each model file produced at the corresponding resolutions in the U-Net trial with the RGB color space. We then generated the predicted masks for each of the six downsampled resolutions of the 120 glomerular images utilizing the U-Net model file. We repeated this process for DeepLab_v3 in the RGB color space. All predicted image sets were then upsampled back to the original size of the glomeruli images ($>1000 \times 1000$ pixels). Mean and median DSC values were then generated to evaluate performance. Furthermore, we also used a standard Mask-RCNN implementation to generate predicted masks at the corresponding original resolutions ($>1000 \times 1000$ pixels) of the glomeruli images. We then compared mean, median, and standard deviation DSC values to investigate which methodology performed best.

4.3 Evaluation Metrics and Statistical Methods

DSC was the primary statistic used to evaluate segmentation performance. In particular, mean, median, and standard deviation of DSC were generated for analysis. Additionally, to evaluate statistical significance between each resolution and different methods, the Wilcoxon Rank Sum test was used with a significance threshold of either $p < 0.05$ or $p < 0.01$. A notched box plot was generated to visually demonstrate median DSC data, as well as the results of Wilcoxon Rank Sum test. Additionally, bar graphs were generated to show mean and standard deviation DSC data. Similarly, to demonstrate the relation between the performance of segmentation within each resolution in both LAB and RGB color space, DSC values for the U-Net trial were summarized in data tables.

5 Results

Our results presented in this section are divided into two central sections to explore each aspect of our experimentation: (1) manual detection and (2) automatic detection.

5.1 Manual Detection

The first aspect of our study is manual detection, wherein we analyzed the key factors of segmentation, which include: segmentation backbones, input image resolutions, as well as color spaces. This phase of our study allowed us to better assess the conditions, in which a detect-then-segment framework performs most optimally in the context of high-resolution WSI.

5.1.1 U-Net vs. DeepLab_v3

We first present that both U-Net and DeepLab_v3 confer particular advantages over one another across the six tested resolutions in the context of glomerular image data. Tables 1 and 2 present

Table 1 The DSC scores collected for the internal testing data using both U-Net and DeepLab_v3. The DSC is evaluated on images in 512 space. In particular, we define "evaluation of images in 512 space" as the process by which the prediced masks are upsampled from the six tested resolutions to the original resolution of the input images, which is 512×512 in the case of U-Net and DeepLab_v3, and then compared against the 512×512 input images to produce a fair DSC value.

Image resolution		512×512	256×256	128×128	64×64	32×32	28×28
U-Net	Mean Dice ± Std dev	0.909 ± 0.099	0.936 ± 0.073	**0.940 ± 0.051**	0.920 ± 0.047	0.878 ± 0.066	0.853 ± 0.072
	Median Dice	0.948	**0.961**	0.957	0.936	0.897	0.872
DeepLab_v3	Mean Dice ± Std dev	**0.948 ± 0.062**	0.947 ± 0.033	0.935 ± 0.048	0.907 ± 0.059	0.840 ± 0.080	0.817 ± 0.090
	Median Dice	**0.963**	0.959	0.948	0.925	0.861	0.843

Table 2 The DSC scores collected for the external testing data using both U-Net and DeepLab_v3. The DSC is evaluated on images in 512 space, with an RGB color space.

Image resolution		512×512	256×256	128×128	64×64	32×32	28×28
U-Net	Mean Dice ± Std dev	0.816 ± 0.141	0.899 ± 0.063	**0.902 ± 0.072**	0.873 ± 0.101	0.845 ± 0.086	0.815 ± 0.105
	Median Dice	0.853	0.917	**0.918**	0.897	0.867	0.838
DeepLab_v3	Mean Dice ± Std dev	0.918 ± 0.071	**0.922 ± 0.062**	0.911 ± 0.068	0.887 ± 0.087	0.812 ± 0.122	0.810 ± 0.088
	Median Dice	0.931	**0.934**	0.928	0.906	0.845	0.834

DSC values for internal and external data for U-Net and DeepLab_v3 in the RGB color space. Both tables show us that DeepLab_v3 would perform better than U-Net for larger resolutions, such as 512×512, 256×256, and 128×128, but would under-perform relative to U-Net for smaller image resolutions, such as 64×64, 32×32, and 28×28. As shown, there is no clear, consistent framework that achieved the best DSC results for all trials. However, both U-Net and DeepLab_v3 show distinct advantages—DeepLab_v3 tends to perform better for larger resolutions, whereas U-Net confers higher DSC values for smaller resolutions.

5.1.2 *Image resolution*

In the trial utilizing a U-Net framework, the resolution in RGB space with the highest median DSC value of 0.961 was 256×256, which was statistically different relative to 64×64, 32×32, and 28×28. On the other hand, for the external testing dataset in RGB space, the resolution with the highest median DSC value of 0.918 was 128×128, which was significantly different relative to 28×28, 32×32, 64×64, and 512×512. A similar analysis was performed on the mean values and standard deviation of the internal and external testing DSC data of the RGB color space in the U-Net trial. For both internal and external data, 128×128 was the resolution with the highest mean DSC values in both the sample space and the upscaled 512 space. Additionally, Fig. 5 further shows the resolutions of 64×64, 32×32, and 28×28 experience the greatest decline in the performance when comparing the DSC in sample space to the DSC in 512 space.

Considering the trial that used a DeepLab_v3 framework, the highest median DSC value for internal data was 0.963, which occurred in the 512×512 resolution. This particular resolution was significantly greater relative to 128×128, 64×64, 32×32, and 28×28. Similarly, for the external data, the highest median DSC value was 0.934 which occurred in 256×256 space. This resolution was statistically greater relative to the resolutions of 128×128, 64×64, 32×32, and 28×28. Analyzing the mean data for DeepLab_v3, it is clear that the highest mean DSC in the internal data occurred in the 256×256 resolution when evaluating DSC in the sample space, whereas the highest mean DSC in 512 space occurred for the 512×512 resolution. For the external data, the highest mean DSC value occurred in the 256×256 resolution when evaluating Dice in the sample and 512 space. Similar to U-Net, the highest difference in mean DSC between the sample space and 512 space occurred in the resolutions of 64×64, 32×32, and 28×28. The comparison of mean data can be seen visually in Fig. 6.

The aforementioned trend in the median DSC data and the results of the Wilcoxon Rank Sum Test can be seen in Fig. 7.

5.1.3 *Image color space*

The conferred accuracy of using either the LAB or RGB color spaces were found to be negligible in the trial utilizing a U-Net framework. Table 3 shows the results of the RGB and LAB color spaces on the internal dataset using U-Net. Both RGB and LAB give almost the same best DSC values, which is bolded. Therefore, the rest of the analysis in this paper is focused on the effectiveness of segmentation in the RGB color space, as the results are generalizable due to the
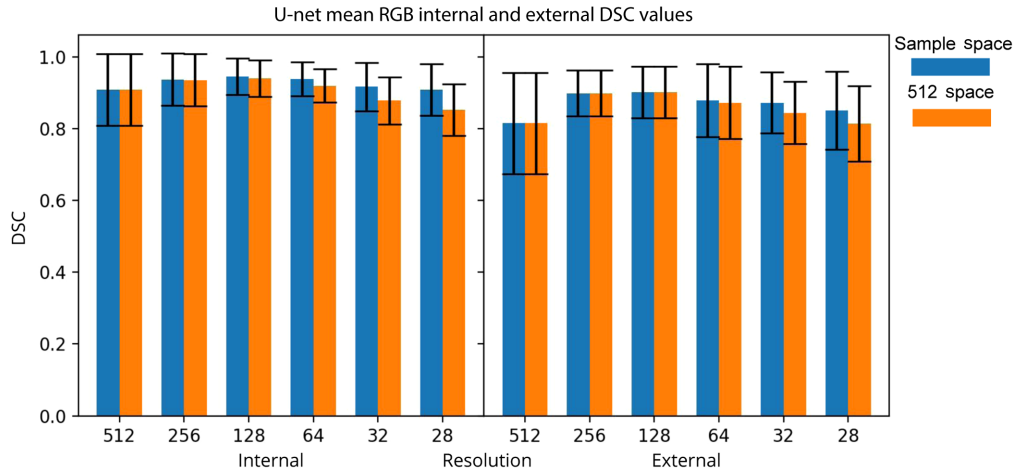
Fig. 5 The mean and standard deviation for both internal and external testing images evaluated in both sample space and 512 space. Similar to how we define "512 space," the term "sample space" refers to the process by which the predicted masks that are produced across six distinct resolutions (512×512, 256×256, 128×128, 64×64, 32×32, and 28×28) are evaluated against the corresponding downsampled input image across the same six resolutions. We re-evaluated such images in 512 space for a fair DSC comparison through upsampling, as discussed earlier. As shown, 64×64, 32×32, and 28×28 declined greatly in accuracy when comparing DSC in 512 space relative to DSC in sample space.
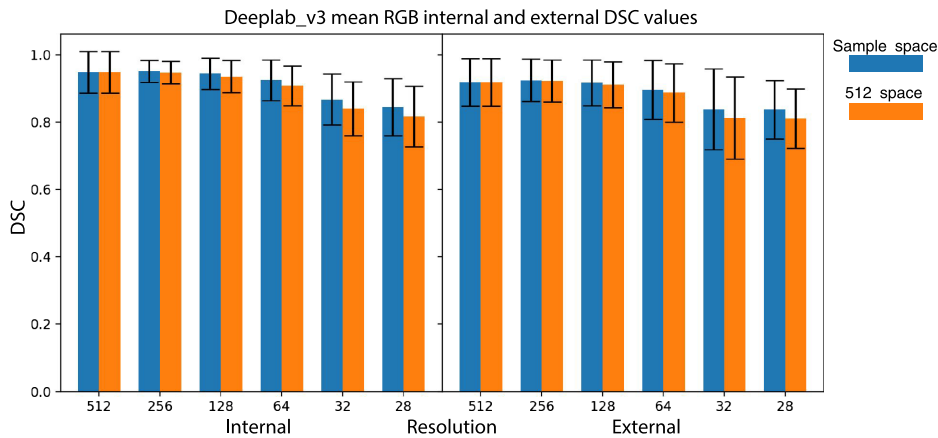

Fig. 6 The mean data and standard deviation for both internal and external testing images evaluated in both sample space and 512 space. As shown, 64×64, 32×32, and 28×28 declined greatly in accuracy when comparing DSC in 512 space relative to DSC in sample space.

similarity of segmentation performance between the RGB and LAB color spaces. In particular, the DeepLab_v3 trial, as stated in the methodology, only performs its segmentation in the RGB color space, due to the results of U-Net.

5.2 Automatic Detection

5.2.1 Detect-then-segment framework vs. end-to-end mask-RCNN

Table 4 demonstrates the results of applying our U-Net and DeepLab_v3 models on the cohort of 120 images during our automatic detection phase, and generating predicted masks by downsampling the input images to six distinct resolutions. By evaluating the difference in the performance between our proposed detect-then-segment pipeline relative to the standard end-to-end Mask-RCNN framework, we found that both U-Net and DeepLab_v3 showed better mean and median
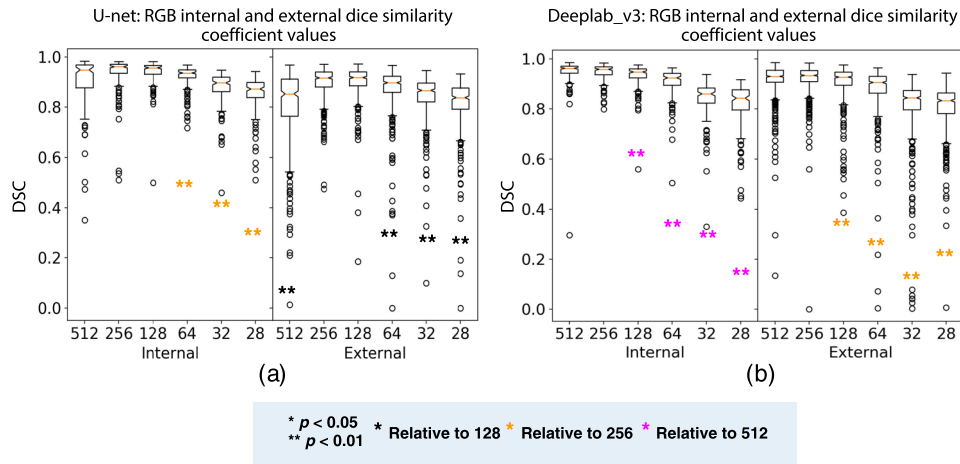
Fig. 7 The notched box plots of each resolution for both internal and external testing data of the RGB color space. (a) The results for the U-Net trial and (b) the results for DeepLab_v3. The legend and asterisks demonstrates results of computing the Wilcoxon Rank Sum Test on the specified resolutions. The median DSC values are derived by evaluating the ground truth mask in 512×512 resolution, relative to the predicted mask in 512×512 resolution for a fair comparison.

Table 3 The performance between RGB and LAB on internal testing data using U-Net in 512 space.

Image resolution		512×512	256×256	128×128	64×64	32×32	28×28
RGB	Mean Dice $\pm$ Std dev	$0.909 \pm$ 0.099	$0.936 \pm$ 0.073	**0.940 $\pm$ 0.051**	$0.920 \pm$ 0.047	$0.878 \pm$ 0.066	$0.853 \pm$ 0.072
	Median Dice	0.948	**0.961**	0.957	0.936	0.897	0.872
LAB	Mean Dice $\pm$ Std dev	$0.917 \pm$ 0.081	$0.936 \pm$ 0.068	**0.940 $\pm$ 0.050**	$0.925 \pm$ 0.045	$0.872 \pm$ 0.069	$0.863 \pm$ 0.066
	Median Dice	0.946	**0.961**	0.957	0.941	0.896	0.882

Table 4 The results of applying our U-Net and DeepLab_v3 models on the cohort of 120 images during our automatic detection phase, and generating predicted masks by downsampling the input images to six distinct resolutions. The predicted masks were then upsampled and compared to the original resolution ground truth masks to generate mean and median DSC values. The results of directly applying Mask-RCNN are also shown.

Resolution	512×512	256×256	128×128	64×64	32×32	28×28	Mask RCNN
U-Net: mean Dice $\pm$ Std dev	$0.935 \pm$ 0.062	**0.947 $\pm$ 0.037**	$0.935 \pm$ 0.035	$0.903 \pm$ 0.041	$0.833 \pm$ 0.060	$0.791 \pm$ 0.067	$0.902 \pm$ 0.038
U-Net: median Dice	0.956	**0.957**	0.945	0.914	0.845	0.798	0.908
DeepLab_v3: mean Dice $\pm$ Std dev	**0.953 $\pm$ 0.027**	$0.941 \pm$ 0.034	$0.919 \pm$ 0.043	$0.876 \pm$ 0.053	$0.771 \pm$ 0.067	$0.750 \pm$ 0.067	$0.902 \pm$ 0.038
DeepLab_v3: median Dice	**0.961**	0.950	0.931	0.891	0.782	0.753	0.908

Table 5 The results of computing an average inference time for each component in the detect-then-segment framework, including the detection, segmentation, and the total inference time for each strategy.

	512×512	256×256	128×128	64×64	32×32	28×28
Average U-Net inference time $\pm$ std dev (s)	2.451 $\pm$ 2.596	1.989 $\pm$ 2.106	1.851 $\pm$ 1.960	1.834 $\pm$ 1.942	1.800 $\pm$ 1.906	1.800 $\pm$ 1.906
Average Mask-RCNN inference time $\pm$ std dev (s)	93.170 $\pm$ 39.813	93.170 $\pm$ 39.813	93.170 $\pm$ 39.813	93.170 $\pm$ 39.813	93.170 $\pm$ 39.813	93.170 $\pm$ 39.813
Average, total inference time $\pm$ std dev (s)	95.621 $\pm$ 39.601	95.159 $\pm$ 39.628	95.021 $\pm$ 39.637	95.004 $\pm$ 39.638	94.970 $\pm$ 39.641	94.970 $\pm$ 39.641
Average DeepLab_v3 inference time $\pm$ std dev (s)	2.994 $\pm$ 3.214	2.074 $\pm$ 2.196	1.954 $\pm$ 2.069	1.920 $\pm$ 2.033	1.937 $\pm$ 2.051	1.080 $\pm$ 1.144
Average Mask-RCNN inference time $\pm$ std dev (s)	93.170 $\pm$ 39.813	93.170 $\pm$ 39.813	93.170 $\pm$ 39.813	93.170 $\pm$ 39.813	93.170 $\pm$ 39.813	93.170 $\pm$ 39.813
Average, total inference time $\pm$ std dev (s)	96.164 $\pm$ 39.555	95.244 $\pm$ 39.623	95.124 $\pm$ 39.630	95.090 $\pm$ 39.633	95.107 $\pm$ 39.632	94.250 $\pm$ 39.699

DSC values for the resolutions of 512×512, 256×256, and 128×128 ($p < 0.01$, Wilcoxon Rank Sum). In particular, at best, our framework provides a mean DSC value of 0.953 via the DeepLab_v3 backbone operating on a previously detected glomerulus of 512×512 resolution, whereas Mask-RCNN produced a mean DSC value of 0.902.

To evaluate the effects of low resolution for segmentation, we compute the upper bound performance of glomerular segmentation at the resolution of 28×28, which is the resolution for the segmentation branch in Mask-RCNN. The upper bound performance is computed by downsampling each manual segmentation glomerular mask image from 512×512 to 28×28 with a nearest neighbor interpolation. The downsampled manual segmentation should outperform any segmentation methods (e.g., Mask-RCNN, U-Net, and DeepLab_v3) at the resolution of 28×28. Next, we upsample the 28×28 manual segmentation masks back to 512×512 with a nearest neighbor interpolation and calculate the DSC between the downsampled-then-upsampled manual segmentation with the original high-resolution 512×512 masks. By doing this, we achieve a mean DSC value of 0.957, with a standard deviation of 0.006, which is the upper bound of a particular segmentation method where the segmentation is performed at 28×28 for an object with original resolution of 512×512. From Table 4, DeepLab_v3, achieved a mean DSC of 0.953 at 512×512, which is very close to the upper bound of performing segmentation at 28×28. This further demonstrates the advantage of performing glomerular segmentation at high resolution, by the fact that the performance of automatic segmentation at 512×512 is comparable to even a manual segmentation at 28×28.

Table 5 shows the average inference time of each component in the detect-then-segment framework for segmenting a high-resolution WSI, rather than an image patch. The detection portion consumes a majority of the time, which is common for all segmentation methods. The computational time of each segmentation method and the total inference time for each overall strategy are also presented. Our results show that the detect-then-segment approaches only introduces, on average, <3 s onto the *de facto* standard end-to-end Mask-RCNN approach, when processing an entire high-resolution whole slide image. In a clinical context, for the total inference time, <3 min of processing, a WSI is acceptable in clinical scenarios, when compared to the time of preparing, scanning, and inspecting the tissue sample.

6 Discussion

First, in our experimentation with manual detection, we comprehensively searched and tested for the best detect-then-segment strategy. The results demonstrate that: (1) utilizing a higher

resolution does not necessarily confer the best segmentation results; (2) the resolutions of 128×128 and 256×256 consistently demonstrated the best segmentation results; (3) lower resolutions (64×64, 32×32, and 28×28) experience the greatest loss of accuracy when comparing DSC in sample space relative to 512 space; (4) DeepLab_v3 yields better results in higher resolutions (512×512, 256×256, and 128×128), whereas U-Net performs most optimally in lower resolutions (64×64, 32×32, and 28×28); and (5) neither the LAB nor RGB color space give rise to better segmentation results relative to one another. Briefly, our results in Sec. 5.1.2 demonstrate that the resolutions of 128×128 and 256×256 consistently demonstrated the best DSC results, and the particular resolution of 512×512 would actually yield relatively lower segmentation results, especially for median DSC. Additionally, the lower resolutions—namely 64×64, 32×32, and 28×28—consistently experienced a great loss in accuracy when analyzing the effectiveness of segmentation in 512 space. Finally, the results demonstrate that DeepLab_v3 and U-Net perform most optimally in different ranges of resolutions. When considering RGB and LAB color spaces, we found there was no discernible effect or advantage of color space on the segmentation of high-resolution glomerular images.

Further, we show that our proposed detect-then-segment pipeline is superior to the conventional end-to-end Mask-RCNN framework. Our summarized results show that the image resolutions of 512×512, 256×256, and 128×128, in both U-Net and DeepLab_v3 in RGB space are significantly better than that of Mask-RCNN. Of course, the most optimal result was achieved through DeepLab_v3, with a mean DSC of 0.953 which occured in 512×512 space. However, both U-Net and DeepLab_v3 showed the same trend in the data. Overall, through our automatic detection trial, we conclude that utilizing a detect-then-segment framework across 512×512, 256×256, or 128×128 will provide better segmentation results compared to the typical Mask-RCNN pipeline. Additionally, through our manual detection trial, we demonstrate that the most optimal detect-then-segment strategy involves utilizing a DeepLab_ v3 framework on larger resolution input images (256×256 and 128×128), in either the RGB or Lab color space. In this study, Mask-RCNN is employed as it has been used as a *de facto* standard end-to-end method for glomerular segmentation in the renal pathology community. Although other pure detection methods (e.g., Faster-RCNN) can be used in the detect-then-segment framework, we directly use the detection results from Mask-RCNN to ensure a direct and fair comparison amongst different segmentation strategies. Moreover, Mask-RCNN, with its new RoIAlign feature, outperformed Faster-RCNN even without the mask heads.[9] In this study, the default weights and segmentation head of Mask-RCNN are employed directly from its official implementation (https://github.com/facebookresearch/maskrcnn-benchmark) without further optimization, since providing the best detection performance is out of the scope of this paper. Note that the detect-then-segment strategy is an adaptable framework such that the Mask-RCNN method could be replaced by other advanced detection methods for potentially better performance.

One key advantage of our research is our analysis of several critical factors of segmentation. In particular, our efforts strive to understand what works best for high-resolution renal WSI, as opposed to practicing the standard end-to-end methods that are popular in computer vision. By studying the effect of color space, resolution, and segmentation backbone on the characterization of WSI through glomeruli data, we are better able to understand how to improve current segmentation networks that operate on high-resolution images so as to yield better results. Another key advantage is how our study definitively shows that our proposed framework can yield a clear advantage in accuracy over the standard end-to-end instance segmentation methods in the context of high-resolution renal WSI. Overall, our data provide a unique and important view toward new methodologies that show better results in high-resolution imaging.

However, there are some important limitations to our study. First, the focus of this study was in the context of renal pathology and glomerular data. However, we expect the findings will be generalizable for other objects in renal pathology as the scaling issues are similar. Another limitation includes the fundamental restraints of the GPU when processing large scale images. In our study, the largest-resolution photo on which we trained our data were 512×512. In particular, we had to downsample the original resolution of the glomerular data due to the memory limitations of the GPU in order to efficiently conduct our study. Furthermore, another limitation of this study is the lack of large-scale annotated testing data, which is important for quantifying the generalizability of different methods and identifying corner cases. In this study, we have

employed testing images from another independent cohort, distinct from the training cohort, to perform an external validation. In the future, it is beneficial to include an even larger-scale and more heterogeneous external validation data.

A central way by which this study may be improved is by diversifying the dataset so as to include other biological objects, such as veins, arteries or tubules. Doing so would help solve the problem of a lack of generalizability with respect to our results and would allow for the significance of our conclusions to be more widespread.

7 Conclusions

Overall, our experimentation through manual and automatic detection phases lead us to a three-fold conclusion: (1) a detect-then-segment framework is more effective than an end-to-end pipeline in the context of high-resolution renal WSI; (2) the performance of a detect-then-segment framework is most optimal with a DeepLab_v3 segmentation backbone operating on a 512×512 resolution for previously detected glomerular input images; and (3) utilizing either RGB or LAB color spaces for previously detected glomerular input images does not yield a particular advantage over the other in a detect-then-segment framework. To conclude, our research paves the way toward further discussion and analysis in understanding effective and more nuanced methodologies that are more accurate than the current framework by which we characterize high-resolution images of glomeruli, and biological objects of similar character, on large-scale WSI.

Disclosures

The authors of the paper have no conflicts of interest to report.

Acknowledgments

References

1. A. Greenberg, *Primer on Kidney Diseases E-Book*, Elsevier Health Sciences, Philadelphia, Pennsylvania (2009).
2. C. S. Rayat et al., "Glomerular morphometry in biopsy evaluation of minimal change disease, membranous glomerulonephritis, thin basement membrane disease and Alport's syndrome," *Anal. Quant. Cytol. Histol.* **29**(3), 173–182 (2007).
3. V. G. Puelles and J. F. Bertram, "Counting glomeruli and podocytes: rationale and methodologies," *Curr. Opin. Nephrol. Hypertension* **24**(3), 224 (2015).
4. M. V. Østergaard et al., "Automated image analyses of glomerular hypertrophy in a mouse model of diabetic nephropathy," *Kidney360* **1**, 469–479 (2020).
5. J. Gallego et al., "Glomerulus classification and detection based on convolutional neural networks," *J. Imaging* **4**(1), 20 (2018).
6. J. D. Bukowy et al., "Region-based convolutional neural nets for localization of glomeruli in trichrome-stained whole kidney sections," *J. Am. Soc. Nephrol.* **29**(8), 2081–2088 (2018).
7. C. Peng et al., "To what extent does downsampling, compression, and data scarcity impact renal image analysis?" in *Digital Image Comput.: Tech. and Appl.*, IEEE, pp. 1–8 (2019).
8. M. MacDonald et al., "Improved automated segmentation of human kidney organoids using deep convolutional neural networks," *Proc. SPIE* **11313**, 113133B (2020).
9. K. He et al., "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vision*, pp. 2961–2969 (2017).
10. S. Sarwar et al., "Physician perspectives on integration of artificial intelligence into diagnostic pathology," *npj Digital Med.* **2**(1), 28 (2019).

11. E. Uchino et al., "Classification of glomerular pathological findings using deep learning and nephrologist-AI collective intelligence approach," *Int. J. Med. Inf.* **141**, 104231 (2020).
12. N. P. Pavinkurve, K. Natarajan, and A. J. Perotte, "Deep vision: learning to identify renal disease with neural networks," *Kidney Int. Rep.* **4**(7), 914 (2019).
13. M. Temerinac-Ott et al., "Detection of glomeruli in renal pathology by mutual comparison of multiple staining modalities," in *Proc. 10th Int. Symp. Image and Signal Process. and Anal.*, IEEE, pp. 19–24 (2017).
14. N. Altini et al., "Semantic segmentation framework for glomeruli detection and classification in kidney histological sections," *Electronics* **9**(3), 503 (2020).
15. S. Kannan et al., "Segmentation of glomeruli within trichrome images using deep learning," *Kidney Int. Rep.* **4**(7), 955–962 (2019).
16. M. Gadermayr et al., "Segmenting renal whole slide images virtually without training data," *Comput. Biol. Med.* **90**, 88–97 (2017).
17. M. Gadermayr et al., "CNN cascades for segmenting whole slide images of the kidney," arXiv:1708.00251 (2017).
18. B. Ginley et al., "Computational segmentation and classification of diabetic glomerulosclerosis," *J. Am. Soc. Nephrol.* **30**(10), 1953–1967 (2019).
19. B. Ginley et al., "Neural network segmentation of interstitial fibrosis, tubular atrophy, and glomerulosclerosis in renal biopsies," arXiv:2002.12868 (2020).
20. G. Bueno et al., "Glomerulosclerosis identification in whole slide images using semantic segmentation," *Comput. Methods Programs Biomed.* **184**, 105273 (2020).
21. G. O. Barros et al., "Pathospotter-k: a computational tool for the automatic identification of glomerular lesions in histological images of kidneys," *Sci. Rep.* **7**, 46769 (2017).
22. J. N. Marsh et al., "Deep learning global glomerulosclerosis in transplant kidney frozen sections," *IEEE Trans. Med. Imaging* **37**(12), 2718–2728 (2018).
23. T.-Y. Lin et al., "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vision and Pattern Recognit.*, pp. 2117–2125 (2017).
24. K. He et al., "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vision and Pattern Recognit.*, pp. 770–778 (2016).
25. V. G. Puelles et al., "Glomerular number and size variability and risk for kidney disease," *Curr. Opin. Nephrol. Hypertension* **20**(1), 7–15 (2011).
26. S. N. Gowda and C. Yuan, "Colornet: investigating the importance of color spaces for image classification," *Lect. Notes Comput. Sci.* **11364**, 581–596 (2018).
27. O. Ronneberger, P. Fischer, and T. Brox, "U-net: convolutional networks for biomedical image segmentation," *Lect. Notes Comput. Sci.* **9351**, 234–241 (2015).

Aadarsh Jha is an undergraduate student in the Department of Electrical Engineering and Computer Science at Vanderbilt University studying computer science.

Biographies of the other authors are not available.