Omni-Seg+: A Scale-aware Dynamic Network for Pathological Image Segmentation

Ruining Deng, Quan Liu, Can Cui, Tianyuan Yao, Jun Long, Zuhayr Asad, R. Michael Womick, Zheyu Zhu, Agnes B. Fogo, Shilin Zhao, Haichun Yang, Yuankai Huo, Senior Member, IEEE

Abstract—Comprehensive semantic segmentation on renal pathological images is challenging due to the heterogeneous scales of the objects. For example, on a whole slide image (WSI), the cross-sectional areas of glomeruli can be 64 times larger than that of the peritubular capillaries, making it impractical to segment both objects on the same patch, at the same scale. To handle this scaling issue, prior studies have typically trained multiple segmentation networks in order to match the optimal pixel resolution of heterogeneous tissue types. This multi-network solution is resource-intensive and fails to model the spatial relationship between tissue types. In this paper, we propose the Omni-Seg+ network, a scale-aware dynamic neural network that achieves multi-object (six tissue types) and multi-scale (5× to 40× scale) pathological image segmentation via a single neural network. The contribution of this paper is three-fold: (1) a novel scale-aware controller is proposed to generalize the dynamic neural network from single-scale to multi-scale; (2) semi-supervised consistency regularization of pseudo-labels is introduced to model the inter-scale correlation of unannotated tissue types into a single end-to-end learning paradigm; and (3) superior scale-aware generalization is evidenced by directly applying a model trained on human kidney images to mouse kidney images, without retraining. By learning from 150,000 human pathological image patches from six tissue types at three different resolutions, our approach achieved superior segmentation performance according to human visual assessment and evaluation of image-omics (i.e., spatial transcriptomics). The official implementation is available at https://github.com/ddrrnn123/Omni-Seg.

Index Terms—Renal pathology, Image segmentation, Multi-label, Multi-scale, Semi-supervised Learning

I. INTRODUCTION

The process of digitizing glass slides using a whole slide image (WSI) scanner-known as “digital pathology” - has led to a paradigm shift in pathology [1]. Digital pathology not only liberates pathologists from local microscopes to remote monitors, but also provides an unprecedented opportunity for computer-assisted quantification [2]–[4]. Many prior arts have developed pathological image segmentation approaches for pixel-level tissue characterization, especially with deep learning methods [5]–[8]. However, comprehensive semantic (multi-label) segmentation on renal histopathological images is challenging due to the heterogeneous scales of the objects. For example, the cross-sectional area of glomeruli can be 64 times larger than that of peritubular capillaries on a 2D WSI section [9]. Thus, human physiologists have to zoom in and out (e.g., between 40× and 5× magnifications) when visually examining a tissue in practice [10]. To handle this scaling issue, prior studies [11]–[13] typically trained multiple segmentation networks that matched the optimal pixel resolution for heterogeneous tissue types. This multi-network solution is resource-intensive and its model fails to consider the spatial relationship between tissue types.

Recent advances in dynamic neural networks shed light on segmenting comprehensive tissue types via a single multi-label segmentation network [14]–[16]. Dynamic neural networks generate the parameters of a neural network (e.g., the last convolutional layer) adaptively in the testing stage, achieving superior segmentation performance via a single network on various applications in natural and radiological image analysis. However, the multi-scale nature of the digitized pathological images (e.g., a WSI pyramid) leads to the unique challenge of adapting the Dynamic Neural Networks to pathology [17]. For instance, Jayapandian et al. [11] showed that the optimal...
resolution for segmenting glomerular units and tufts is 5x, while the optimal resolution for segmenting the much smaller peritubular capillaries is 40x.

In this paper, we propose a single segmentation network, Omni-Seg+, that performs multi-label multi-scale semantic segmentation on WSIs via a single dynamic neural network trained end-to-end. OmniSeg+ explicitly models the scale information as a scale-aware controller to, for the first time, make a single dynamic segmentation network aware of both scale information and tissue types in pathological image segmentation. The design is further generalized by introducing semi-supervised consistency regularization to model the spatial relationships between different tissue types even with different optimal segmentation scales. We evaluate the proposed method using the largest public multi-tissue segmentation benchmark in renal pathology, involving the glomerular tuft (TUFT), glomerular unit (CAP), proximal tubular (PT), distal tubular (DT), peritubular capillaries (PTC), and arteries (VES) with four different stains [Hematoxylin and Eosin (H&E), Periodic-acid-Schiff (PAS), Silver (SIL), and Trichrome (TRI)] at three digital magnifications (5x, 10x, 40x).

The contribution of this paper is three-fold: (1) a novel scale-aware controller is proposed to generalize the dynamic neural network from single-scale to multi-scale; (2) semi-supervised consistency regularization of pseudo-labels is introduced to model the inter-scale correlation of unannotated tissue types; and (3) superior scale-aware generalization of the proposed method is achieved by directly applying a model trained on human kidney images to mouse kidney images, without retraining. The code has been made publicly available at https://github.com/ddrrnn123/Omni-Seg.

II. RELATED WORKS

A. Renal pathology segmentation

With the recent advances in deep learning, Convolutional Neural Networks (CNNs) have become the de facto standard method for image segmentation. Gadermayr et al. [18] proposed two CNN cascades for histological segmentation with sparse tissue-of-interest. Gallego et al. [19] implemented AlexNet for precise classification and detection using pixel-wise analysis. Bueno et al. [20] introduced SegNet-VGG16 to detect glomerular structures through multi-class learning in order to achieve a high Dice Similarity Coefficient (DSC). Bouteldja et al. [21] developed a CNN for the automated multi-class segmentation of renal pathology for different mammalian species and different experimental disease models. Recently, instance segmentation approaches and Vision Transformers (ViTs) have been introduced to pathological image segmentation [22], [23]. However, most of these approaches mainly focus on single tissue segmentation, such as glomerular segmentation.

Our previous network, Omni-Seg [24], utilizes a single residual U-Net as its backbone [25], [26] with a dynamic head design to achieve multi-class pathology segmentation. Omni-Seg+ builds upon our previous work by using a scale-aware vector to describe the scale-specific features and training the model with semi-supervised consistency regularization to understand spatial inferences between multiple tissue types at multiple scales, combining the information that is essential for pathological image segmentation.

B. Multi-label medical image segmentation

Deep learning-based segmentation algorithms have shown the capability of performing multi-label pathological image segmentation [11], [13], [21]. Due to the issue of partial labeling, most approaches rely on an integration strategy to learn single segmentation from one network. This multi-network solution is resource intensive and suboptimal, without explicitly modeling the spatial relationship between tissue types. To address this issue, many methods have been proposed to investigate the partial annotation of a medical image dataset. Chen et al. [27] designed a class-shared encoder and class-specific decoders to learn a partially labeled dataset for eight tasks. Fang et al. [28] proposed target-adaptive loss (TAL) to train the network by treating voxels with unknown labels as the background.

Our proposed method, Omni-Seg+, was inspired by DoDNet [14], which introduced the dynamic filter network to resolve multi-task learning in a partially labeled dataset. As
shown in Fig. 2, we generalized the multi-label DoDNet to a multi-label and multi-scale scenario. An online semi-supervised consistency regularization of pseudo-label learning extended the partially labelled dataset to the densely labelled dataset with non-overlap pseudo-labels.

C. Multi-scale medical image segmentation

Unlike radiological images, pathological images contain multi-resolution images, called image pyramids, that allow different tissue types to be examined at their optimal magnifications or best resolutions [14]. However, modeling scale information for segmentation models is still challenging. Several deep learning-based approaches have been developed to aggregate scale-specific knowledge within the network architecture [29]-[33]. However, such technologies focus on feature aggregation from different scales and fail to learn scale-aware knowledge for heterogeneous tasks.

In our proposed network, we explicitly modeled and controlled pyramid scales (5×, 10×, 20×, 40×) for a U-Net architecture by using a scale-aware controller joined with a class-aware controller by a feature fusion block. A scale-aware vector is proposed to encourage the network to learn distinctive features at different resolutions.

III. METHODS

The overall framework of the proposed Omni-Seg method is presented in Fig. 3. The backbone structure is a residual U-Net, inspired by the existing multi-label segmentation network DoDNet [14] and Omni-Seg [24] methods.

A. Simultaneous multi-label multi-scale modeling

Omni-Seg method was recently proposed to achieve multi-label segmentation using dynamic neural network design [24]. However, such a method is not optimized for the multi-scale image pyramids in digital pathology. Moreover, the context information across different scales is not explicitly utilized in the learning process. To develop a digital pathology optimized dynamic segmentation method, the proposed Omni-Seg+ method generalize the model-aware encoding vectors to a multi-modal multi-scale fashion, with: (1) \(m\)-dimensional one-hot vector for class-aware encoding and (2) \(n\)-dimension one-hot vector for scale-aware encoding, where \(m\) is the number of tissue types, and \(n\) is the number of magnifications for pathological images. The encoding calculation follows the following equation:

\[
T_k = \begin{cases} 
1, & \text{if } k = i \\
0, & \text{otherwise} \\
\end{cases} 
\]

\(k = 1, 2, ..., m \quad (1)\)

\[
S_p = \begin{cases} 
1, & \text{if } p = j \\
0, & \text{otherwise} \\
\end{cases} 
\]

\(p = 1, 2, ..., n \quad (2)\)

where \(T_k\) is a class-aware vector of \(i\)th tissue, and \(S_p\) is a scale-aware vector in \(p\)th scale.

B. Feature fusion block with dynamic head mapping

To provide the multi-class and multi-scale information to the embedded features, we combine two vectors into the low-dimensional feature embedding at the bottom of the residual U-Net architecture. The image feature \(F\) is summarized by a Global Average Pooling (GAP) and receives a feature vector in the shape \(R^{N \times 256 \times 1 \times 1}\), where \(N\) is batch-size. The class-aware vector \(T_k\) and the scale-aware vector \(S_p\) are reformatted to be the same shape as the image features for the next fusion step. Different from Omni-Seg [24] which directly concatenates the feature vectors, a triple outer product is implemented to combine three vectors into a one-dimensional vector by a flatten function, following a single 2D convolutional layer controller, \(\varphi\), as a feature fusion block to refine the feature vector as the final controller for the dynamic head mapping:

\[
\omega = \varphi(GAP(F)||T||S; \Theta_\varphi) \quad (3)
\]

where \(GAP(F)\), \(T_i\), and \(S\) are combined by the fusion operation, \(||\|\), and \(\Theta_\varphi\) is the number of parameters in the dynamic head. The feature-based fusion implementation is shown in Fig. 4.

Inspired by [14], a binary segmentation network is employed to achieve multi-label segmentation via a dynamic
filter. From the multi-label multi-scale modeling above, we derive joint low-dimensional image feature vectors, class-aware vectors, and scale-aware vectors at an optimal segmentation magnification. The information is then mapped to control a light-weight dynamic head, specifying (1) the target tissue type and (2) the corresponding pyramid scale.

The dynamic head concludes with three layers. The first two have eight channels, while the last layer has two channels. We directly map parameters from the fusion-based feature controller to the kernels in the 162-parameter dynamic head to achieve precise segmentation from multi-modal features. Therefore, the filtering process can be expressed by Eq. 4

\[
P = (((M * \omega_1) * \omega_2) * \omega_3)
\]

where * is convolution, \( P \in \mathbb{R}^{N \times 2 \times W \times H} \) is the final prediction, and \( N, W, \) and \( H \) correspond to the batch-size, width, and height of the dataset, respectively.

C. Semi-supervised consistency regularization of pseudo Label learning

An online semi-supervised pseudo-label learning strategy is proposed to generate the “densely labeled” dataset for the learning of spatial correlation. The original large images at 40× magnification are tiled into small patches with multiple sizes and downsampled to a size of 256×256 pixel resolution to rescale their magnifications to the optimal resolutions, respectively. At each scale, the patches are segmented for multiple tissues at their optimal segmentation magnification by using different class-aware vectors and scale-aware vectors. Then, the patches are aggregated back into the original 40× physical space according to their original location and are then rescaled. There are two strategies for collecting the “densely labeled” dataset with pseudo-labels at the patch level. The first one is tiling the large images into different scales with a 256×256 pixel resolution, while the second one uses a similarity score to locate the patches in the supervised training data, matching and cropping the consistent area pseudo-labels. The matching selection is shown in Fig. 6. As a result
of the ablation study in Table III, the matching selection attained a better performance with a better understanding of spatial relationships between supervised labels and pseudo-labels. Fig. 5 demonstrates the online “densely labeled” dataset with extended pseudo-labels. The pseudo-labels expand the dimensional correspondences for multiple tissues at multiple resolutions. Inspired, [34], a semi-supervised constraint is introduced to enforce the similar embedding of two augmentations upon the same images.

IV. DATA

1,751 regions of interest (ROIs) images were captured from 459 WSIs, obtained from 125 patients with Minimal Change Diseases. The images were manually segmented for six structurally normal pathological primitives [11], using the digital renal biopsies from the NEPTUNE study [17]. All of the images had a resolution of 3000×3000 pixels at a 40× magnification (0.25 μm pixel resolution), including TUFT, CAP, PT, DT, PTC, and VES in H&E, PAS, SIL, and TRI stain. Four stain methods were regarded as color augmentations in each type of tissue. We followed [11] to randomly crop and resized them into 256×256 pixels resolution. We kept the same splits as the original release in [11], where the training, validation, and testing samples were separated with a 6:1:3 ratio. The splits were performed at the patient level to avoid data contamination.

V. RESULTS

We compared the proposed Omni-Seg+ network to baseline models, including (1) multiple individual U-Net models (Multi U-Net) [11] and (2) multiple individual DeepLabv3 models (Multi DeepLabv3) [29]. We also compared the proposed network to (3) a multi-head model with target adaptive loss (TAL) for multi-class segmentation [28], (4) a multi-head 3D model (Med3D) for multiple partially labeled datasets [27], (5) a multi-class segmentation model for partially labeled datasets [35], and (6) a single dynamic network [24] from our previous work.

A. Internal validation

Table I and Fig. 7 show the results on the publicly available dataset [11]. The distance metrics are in units of Micron. In Table I, Omni-Seg+ achieved the better performance in most metrics. In Fig. 7, Omni-Seg+ achieved better qualitative results with less false-positive, false-negative, and morphological errors among the best, the median, and the worst Dice cases. The Dice similarity coefficient (Dice: %, the higher, the better), Hausdorff distance (HD, Micron unit: the lower, the better), and Mean Surface Distance (MSD, Micron unit, the lower, the better) were used as performance metrics for evaluating the quantitative performance.

Fig. 8 illustrates the functionality of the multi-class and multi-scale dynamic design in Omni-Seg+, with both intermediate representations and final segmentation masks. First, the shared feature maps are identical before applying the class-aware and scale-aware dynamic control. Then, different segmentation results are achieved for different tissue types (Row 1 to 6) and different scales (Row 7 to 10), from a single deep neural network.

B. External Validation

To validate our proposed method on another application, Omni-Seg+ was evaluated by directly applying the model trained on a human kidney dataset to a murine kidney dataset (without retraining).

1) Data: Four murine kidneys were used as the external validation, with both H&E WSIs (20×) and 10× Visium spatial transcriptomics acquisition.

2) Approach: We applied different segmentation approaches (as shown in Table II) to the whole kidney WSI. We extracted the patches with 55 μm diameter (circle shaped spots) according to the 10× Visium spatial transcriptomics protocol [36]. Then, we compared the proportions of the targeting tissue types in each spot with human labels and genetic labels (Fig. 9).

CAP percentages in spots. One pathologist was asked to label the percentage of CAP area in each spot, rather than performing resource-intensive pixel-level annotation. Then, such percentage can be automatically achieved from different segmentation methods. A Pearson correlation score was computed between the manual labels and automatic estimations, as shown in Table II.

PT percentages in spots. It was difficult to replicate the above evaluation for PT since to visually differentiate PT from DT is challenging even for human pathologists. Fortunately, spatial transcriptomics analytics were able to offer the percentile of PT specific cell counts with in each spot. We believe this was the most unbiased approximation that was available to evaluate the PT segmentation. Briefly, the transcriptomics sequencing data were demultiplexed by “mkfastq” module in SpaceRanger [37], fastQC [38] were used for Quality control. R package Seurat [39] was used for data processing, while the spacexr [40] software was employed to obtain the PT cell percentages via cell deconvolution. We compare such percentages with the ones from different automatic segmentation approaches, as shown in Table II.

3) Experimental Details: PT and CAP were extracted with the diameter of the spots is 55 μm, which is 110 pixels on 20× digital WSIs, following the standard 10× Visium spatial transcriptomics protocol [36].

4) Results: Table II shows the Spearman Correlation scores of CAP and PT percentages with human and spatial transcriptomics labels. Three digital magnifications (5×, 10×, 20×) are generated by downsampling the 20× WSIs for a more comprehensive assessment. As a result, Omni-Seg+ achieved superior performance (in red) for most evaluations. The correlation metric of TAL for the capsule glomerular tissue is nan because of zero predictions for all patches.

C. Ablation Studies

Table III indicates the performance of the six model designs of Omni-Seg and Omni-Seg+ on the external validation dataset. As a result, the Omni-Seg+ approach with matching Selection (MS) and Consistency Regularization (CR) achieved the superior performance.
Fig. 7. This figure shows the qualitative results of different approaches. The red, green, and yellow bounding boxes present the false positive, false negative, and morphological errors in the predicted masks, respectively. The cases with best, median, and worst Dice scores are provided for a qualitative comparison.

TABLE I

| Method                  | DT       | PTC       | CAP       |
|-------------------------|----------|-----------|-----------|
|                         | Dice↑    | HD↓       | MSD↓      |
|                         | Dice↑    | HD↓       | MSD↓      |
|                         | Dice↑    | HD↓       | MSD↓      |
| U-Nets [11]             | 78.51    | 107.63    | 36.05     |
| DeepLabV3s [29]         | 77.92    | 107.61    | 35.45     |
| TAL [28]                | 47.76    | 280.44    | 198.07    |
| Med3D [27]              | 47.73    | 194.55    | 110.09    |
| Multi-class [35]        | 47.76    | 280.44    | 198.07    |
| Omni-Seg [24]           | 81.01    | 97.27     | 24.19     |
| Omni-Seg+ (Ours)        | 81.11    | 97.99     | 22.54     |

TABLE II

| Method                  | CAP      | PT        |
|-------------------------|----------|-----------|
|                         | 5×       | 10×       | 20×       |
|                         | 5×       | 10×       | 20×       |
| U-Nets [11]             | 60.15    | 90.88     | 89.30     |
| DeepLabV3s [29]         | 64.31    | 89.42     | 89.80     |
| TAL [28]                | nan      | nan       | nan       |
| Med3D [27]              | 43.58    | 89.45     | 85.89     |
| Multi-class [35]        | 44.77    | 90.15     | 87.39     |
| Omni-Seg [24]           | 58.37    | 89.38     | 87.14     |
| Omni-Seg+ (Ours)        | 52.10    | 87.33     | 91.73     |

VI. DISCUSSION

In this study, we propose a novel single dynamic segmentation network with scale information for histopathology images. With the consistency regularization of multi-tissues and multi-scale vectors, the proposed algorithm achieved outstanding Dice scores on the human dataset, where the Dice score values on human glomeruli are consistent with those on murine glomeruli (Table III). The Dice score on the multi-scale multi-tissues on the murine dataset outperformed the state-of-the-art methods by a large margin.

Fig. 8. Intermediate representation – This figure shows the unique advantage of the dynamic neural network design. Specifically, the single set of feature maps are shared by all organs and scales, while different segmentation outcomes are achieved with the multi-label multi-scale controllers.
**TABLE III**

| Backbone | MS  | CR  | PT  | CAP |
|----------|-----|-----|-----|-----|
| Omni-Seg | 57.73 | 87.14 |     |     |
| Omni-Seg | ✓   | 64.32 | 87.15 |     |
| Omni-Seg | ✓   | 66.12 | 90.04 |     |
| Omni-Seg+ | 72.80 | 89.64 |     |     |
| Omni-Seg+ | ✓   | 69.80 | 83.52 |     |
| Omni-Seg+ | ✓   | 75.25 | 91.73 |     |

*MS is Matching Selection
*CR is Consistency Regularization

scales on a consistent area of supervised training data, the proposed model can observe and extend the spatial relationship and the scale consistency from originally partially annotated multi-scale pathological images.

Table I demonstrates that the single network design can enhance 3% of the overall DSC of segmentation by aggregating multi-class and multi-scale knowledge in a single backbone. Since the testing images are from the same scale and same distribution as the training images, the Omni-Seg+ just achieves the comparable performance as the Omni-Seg. However, when applying both methods onto another independent dataset with different tissue scales, the Omni-Seg+ achieves overall superior performance compared with Omni-Seg as well as other bench marks (Table II). Table III further assesses the contributions of the proposed scale-aware design as an ablation study.

There are several limitations and potential future improvements for our study. In the current version of the network, each region of the WSIs needs to be resized to the optimal resolution since all the tissues are segmented in different resolutions as a means of binary segmentation. Thus, it is a time consuming process to aggregate the tissue-wise segmentation results into the final multi-label segmentation masks, which increases the computational times during the testing stage.

The network provides morphological quantification for multiple tissues that can efficiently assist to the topography of gene expression in transcriptomics analysis for future genomics examinations. Meanwhile, the current single network with a class-aware vector and scale-aware vector can be simply applied to the additional dataset by fine-tuning the specific tissue types at different scales. Further work is needed to evaluate the proposed method’s applicability to types of digital pathology datasets other than the ones explored here.

**VII. CONCLUSION**

In this paper, we propose a holistic dynamic segmentation network with scale-aware knowledge, Omni-Seg+, that segments multiple tissue types at multiple resolutions using partially labeled images. The dynamic neural network based design achieves superior segmentation performance with less computational resource consumption compared with the previous multi-network or multi-head designs. The dynamic scale controller and the semi-supervised consistency regularization of pseudo-label learning are introduced to improve the segmentation performance on external validations by modeling spatial correlations and consistency between different tissue types. The propose Omni-Seg+ method provides a generalizable solution for multi-scale multi-label segmentation in digital pathology, so as to ultimately leverage the quantitative clinical practice and research for various kidney diseases.

**ACKNOWLEDGMENT**

This work was supported in part by NIH NIDDK DK56942(ABF).
