DualSeg: Unified Multi-Scale Framework With Dual-Stage Encoder For Glomerular Segmentation

Yan Zhang, Wei Yuan, Jiayu Zhang, Jing Zhang, Ling He

***Abstract*— Chronic Kidney Disease (CKD) demands precise histopathological analysis to identify glomerular alterations critical for therapeutic interventions. However, manual segmentation of glomeruli in Whole Slide Images (WSIs) is labor-intensive and error-prone, necessitating automated solutions. Convolutional Neural Networks (CNNs), which exhibit limited adaptability, and Vision Transformers (ViTs), which entail high computational costs, both struggle to address different key issues in glomerular segmentation, including local texture discriminability and spatial heterogeneity. Furthermore, there remains a lack of a unified model framework to integrate the handling of these challenges. To tackle these challenges, we present DualSeg, a dual-stage framework integrating CNN and Vision Recurrent Weighted Key Value (VRWKV). This design leverages CNN for multi-scale local feature extraction and VRWKV for efficient long-range dependency modeling via linear attention mechanisms, achieving robust glomerular segmentation under complex pathological conditions. Additionally, we employ a Plug-and-Play module, named as the Wave-Swin Block. This module builds upon the Wave Vision module by incorporating multi-directional semantic feature extraction capabilities and dynamic adaptation to pathological variations. Evaluated on three 2D medical image datasets, DualSeg demonstrates superior performance to state-of-the-art models, showing exceptional accuracy and scalability. Our work presents a new methodological approach for renal histology analysis by combining local-texture sensitivity with global context modeling in a unified framework. The model code is available in https://github.com/unskye/DualSeg.**

***Index Terms*— Chronic Kidney Disease (CKD), Histopathology Image Analysis, Glomerular Segmentation, Convolutional Neural Network (CNN), Vision Recurrent Weighted Key Value(VRWKV)**

I. Introduction

Chronic Kidney Disease (CKD) currently affects over 9% of the global population, with significant systemic implications that can precipitate the development of secondary conditions such as hypertension [1]–[3]. The primary histological manifestations of CKD include glomerulosclerosis and renal interstitial fibrosis [4], [5], and pathological alterations in the glomerulus represent a critical focal point for CKD therapeutic interventions. Nevertheless, the significant amount of time and resources required for expert manual analysis, annotation, and diagnosis of glomeruli across extensive Whole Slide Images (WSIs) present notable challenges [6]. Therefore, advanced computational tools enabling automated, high-precision glomerulus segmentation in renal pathology WSIs offer a promising solution to address these diagnostic difficulties.

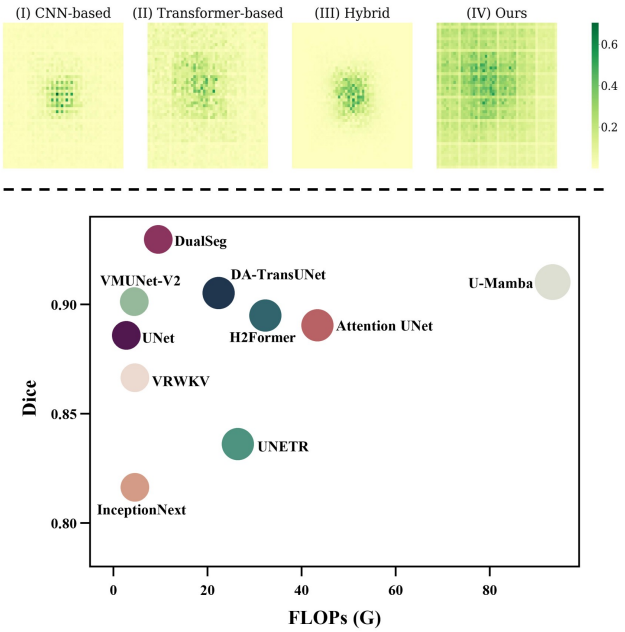

Fig. 1: **Top:** Visualization of the *Effective Receptive Fields* (ERF) for different architectures. **Bottom:** Performance *vs.* FLOPs with recent CNN/Transformer-based/Hybrid methods.

The accurate segmentation of glomeruli in renal histopathological images remains a significant challenge, emerging from the complex interplay between local textural details and global structural complexities. Three core technical challenges define this task: (1) *Local texture discriminability*: glomeruli possess complex morphologies with diverse substructures, resulting in significant intra-class variability in appearance [12], [13]. (2) *Spatial heterogeneity*: glomeruli exhibit irregular spatial distributions, whose accurate decoding necessitates global

Corresponding author: Ling He

Yan Zhang, Wei Yuan, Jiayu Zhang, Jing Zhang and Ling He are with the College of Biomedical Engineering, Sichuan University, Chengdu 610065, China (e-mail: zzzzy@stu.scu.edu.cn; yuanw@stu.scu.edu.cn; zhang_jiayu@stu.scu.edu.cn; jing_zhang@scu.edu.cn; ling.he@scu.edu.cn).

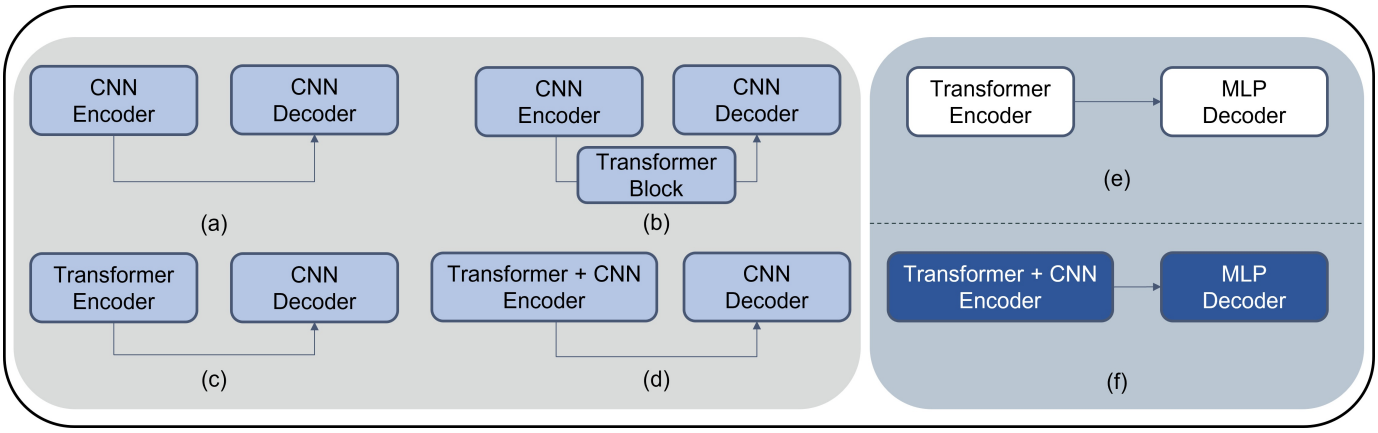

Fig. 2: Comparison of six representative methods for medical image segmentation. The left group (a–d) follows a **U-shaped** architecture with the right group (e–f) adopts a **pyramid-shaped** design. (a): The standard pure convolutional UNet [7]. (b)(c)(d): The cascaded hybrid of CNN and Transformer structure, such as TransUNet [8], UNETR [9] and H2Former [10]. (e): The standard pure SegFormer architecture [11], featuring in the cross-scale feature combination approach. (f): Our proposed efficient hierarchical hybrid structure, combining a hybrid encoder with the multi-scale fusion mechanism.

contextual information to ensure structural continuity [14]. (3) *Multi-scale mapping of morphological priors*: effective multi-scale spatial perception requires the integration of receptive fields at different resolutions to decode fine-grained details and macroscopic structural priors simultaneously [14]–[16]. Collectively, these challenges necessitate advanced model designs capable of balancing pixel-level precision with global contextual understanding through a multi-scale architecture.

As a landmark development in digital pathology, deep learning techniques have achieved widespread adoption and recognition. Convolutional Neural Networks (CNNs) offer a compelling solution to the first challenge of local texture discrimination due to their inductive biases of local receptive fields and parameter sharing [8], [17]. For instance, Kaur et al. [18] utilized a modified U-Net architecture to demonstrate the effectiveness of U-Net in glomerular segmentation by leveraging CNNs' ability to capture hierarchical texture patterns through stacked convolutional layers, which effectively model sub-pixel-level intensity transitions in periodic acid-Schiff (PAS)-stained glomerular boundaries. However, CNN-based approaches inherently struggle to model long-range contextual dependencies due to their local Effective Receptive Field (ERF) [8], [10], [19], as shown in Fig. 1(I), thereby limiting their capacity to capture morphological complexities in glomerular structures, such as irregular shapes and overlapping regions.

In contrast, Vision Transformers (ViTs) offer an effective solution to the second challenge of spatial heterogeneity, which requires robust global contextual integration as shown in Fig. 1(II). By leveraging self-attention mechanisms, ViTs demonstrate enhanced capability in capturing complex spatial patterns across irregular glomerular distributions [20], [21]. Notably, ViTs outperform UNets under identical experimental conditions [21], as their global attention maps more effectively maintain glomerular continuity during the decoding of fibrosis-induced fragmented spatial arrangements. However, ViT-based architectures face significant computational challenges: the quadratic complexity of self-attention mechanisms limits their practicality for high-resolution histology imaging [10].

Complementary research directions have emerged under these frameworks. Simple multi-layer perceptron (MLP) with less inductive bias and theoretically stronger generalization performance has been exploited [22]–[24]. The Wave-MLP framework [25], for example, proposed the concept of wave vision and incorporates spatial-wise convolutions within MLP blocks, thereby preserving fine-grained structural details while mitigating parameter redundancy. For ViTs, the VRWKV (Vision Receptance Weighted Key Value) framework [26] substitutes self-attention of squared complexity with a linear-time recurrent kernel, and VM-UNet [27] pioneered the use of a U-shaped architecture grounded in state space models (SSM), achieving comparable segmentation accuracy with considerable reduction in computational overhead. These innovations underscore the potential of novel architectures to harmonize efficiency and performance.

Beyond stand-alone architectural advancements, researchers are investigating synergistic hybrid frameworks, which are as illustrated in Fig. 9. Prior hybrid architectures, such as TransUNet [8], H2Former [10] and U-mamba [28], integrated convolutional layers with self-attention mechanisms within U-Net paradigms. Conversely, these approaches exhibit limitations: TransUNet's ineffective multi-scale feature capture compromises glomeruli segmentation within complex pathological imagery; H2Former's shallow ViT-CNN integration underutilizes ViT strengths on large datasets and may degrade CNN performance; Mamba-based models have not yet been widely recognized [29]. Furthermore, pyramid-shaped architectures such as SegFormer [11] and SegNext [30] address the third challenge by integrating morphological priors through feature fusion. While computationally efficient, these models rely on unidirectional feature extraction mechanisms. This design choice compromises robustness when applied to large-scale datasets, as the lack of diverse feature aggregation weakens adaptability to complex morphological variations.

This prompts a critical inquiry: *Is it possible to integrate local and global features within a unified multi-scale framework to tackle all three segmentation challenges?*

Answering this question not only addresses the fundamental limitations of current architectures but also offers a unified solution to all three challenges—local texture discrimination, spatiotemporal heterogeneity, and morphological prior. To respond to this question, we propose a noval hybrid framework that synergizes the strengths of Wave Vision and VRWKV.

Specifically, our architecture incorporates a dual-stage encoder: Wave-Swin blocks are employed in the early stages for hierarchical multi-scale local feature extraction, effectively mitigating the challenge of local texture discrimination. Within the CNN branch, we enhance the Wave Vision module by introducing an efficient attention mechanism, which mitigates the inherent limitations of fixed inductive biases and excessive computational overhead. Subsequently, VRWKV blocks are integrated in the later stage to model long-range dependencies via linear attention mechanisms, addressing the issue of spatial heterogeneity. This design preserves the superior texture discrimination capability of convolutions, the strong generalization of MLPs, and the parameter efficiency and scalability of VRWKV, rendering it well-suited for processing large-scale histopathological data. Furthermore, by bridging local and global feature processing, our method achieves multi-scale mapping of morphological priors, enabling robust glomerular segmentation under complex pathological conditions and thus setting a new benchmark for renal histopathological analysis. Additionally, experimental results demonstrate that DualSeg delivers highly competitive performance with moderate floating-point operations (FLOPs), as illustrated in Fig. 1. Collectively, the contributions of this study are fourfold:

- We first identified three key challenges in glomerular segmentation: local texture distinguishability, spatial heterogeneity, and multi-scale mapping of morphological priors. To address these, we propose **DualSeg**, a novel hybrid architecture that integrates convolutional local feature extraction with linear-complexity bidirectional attention within a pyramid design, thereby enhancing glomeruli segmentation.
- We design a plug-and-play attention module based on wave vision to capture multi-directional semantic features, improving feature representation in glomerulus segmentation.
- We validate cross-species and cross-center generalizability by training on mouse data and performing direct inference on a held-out human dataset, demonstrating robust adaptation across species and medical institutions while providing an ethically compliant solution.
- Our method achieves state-of-the-art performance on three 2D glomerulus segmentation datasets, demonstrating superior accuracy and robustness compared to existing approaches.

II. Related Work

A. Development of Medical Image Analysis and Kidney Pathology Image Segmentation

Medical image analysis has been profoundly transformed by deep learning, particularly in pathological image segmentation [31], [32]. The advent of deep learning revolutionized this field by enabling end-to-end hierarchical feature learning from WSIs—delivering enhanced accuracy and robustness in segmenting organs, tissues, and lesions across diverse histopathological scenarios, while overcoming challenges associated with morphological diversity and subtle pathological changes [33].

In kidney pathology, automated glomerular segmentation has emerged as a critical diagnostic tool, given the glomerulus' role as the core functional unit of the kidney [34]. Early approaches combined manual annotations with traditional machine learning for boundary detection [35], [36], extracting morphometric and texture features for classification. However, these methods exhibited poor generalizability across datasets and high susceptibility to staining variations and subjective annotation biases [37]. The transition to deep learning—especially CNNs—marked a pivotal advance: CNNs outperformed recurrent neural networks (RNNs) and artificial neural networks (ANNs) in capturing intricate glomerular morphologies from high-resolution WSIs [38]–[42]. Landmark studies validated this progress: Marsh et al. [43] achieved pathologist-level accuracy in sclerotic glomerulus discrimination, while Bueno et al. [44] pioneered pixel-wise semantic segmentation for glomerulosclerosis detection. These innovations not only improved diagnostic precision for conditions such as glomerulonephritis and diabetic nephropathy but also enabled quantitative renal function assessment, laying the groundwork for standardized disease evaluation in nephrology.

Andreini et al. [45] demonstrated that pre-training on mouse models followed by fine-tuning on human datasets enhances the generalization ability of models for human glomerular segmentation and alleviates the constraints of limited annotated human histopathological data. This finding underscores the practical value of cross-species transfer learning in bridging the gap between preclinical research and clinical applications, providing a feasible paradigm for optimizing model performance when direct access to large-scale human pathological data is restricted.

B. Technological Evolution of Glomerular Segmentation Architectures

Subsequent advances in semantic segmentation architectures have further driven performance improvements in glomerular segmentation. CNNs established a foundational framework, and architectural innovations based on U-Net—such as the Deeplab series, notably Deeplabv3+ [46] with atrous convolution for multi-scale feature extraction, have proven effective in addressing glomerular morphological variability [15], [46]. This underscores that U-Net and its variants, which leverage encoder-decoder architectures with skip connections, enhance pixel-level accuracy by fusing high-level semantic features and low-level spatial details, a capability critical for distinguishing glomeruli from adjacent tubules or interstitial tissues [15].

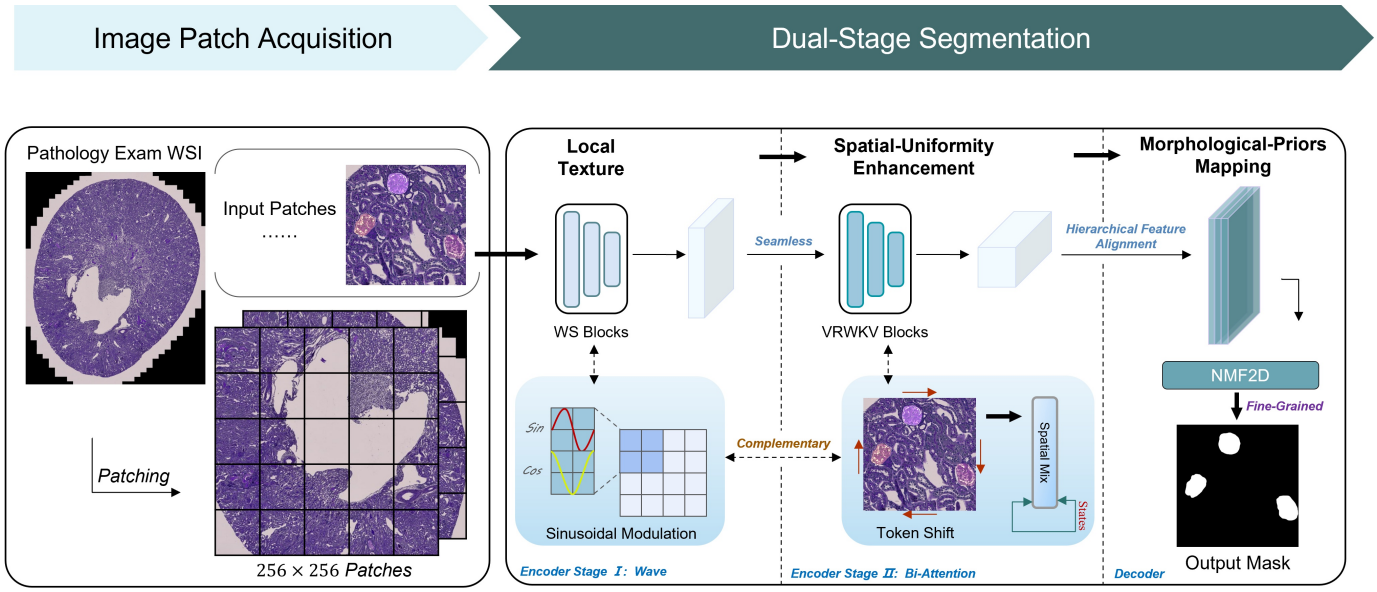

Fig. 3: Overview of the proposed **DualSeg**. The original image is fed into the Dual-Stage encoder composed of Wave-Swin Block and VRWKV Block to extract local texture and patial heterogeneity. Next, the features are merged and refined by the lightweight *Hamburger* decoder to obtain the predicted glomerular mask. The composition of main modules is explained in detail.

However, the computational efficiency of convolution-only models has been questioned. Lightweight architectures such as UNext and MLP UNet have addressed this by introducing MLPs to balance inference speed and parameters, improving generalizability while reducing computational burden [22]–[24]. These models achieved outstanding performance with 70% fewer parameters than traditional U-Net variants.

Frameworks based on self-attention mechanisms, such as SegFormer and U-mamba, have pioneered an alternative paradigm. They demonstrated superior performance in segmenting irregularly shaped glomeruli by capturing long-range spatial correlations, outperforming CNN-based methods on boundary delineation tasks [11], [26], [28], [32], [47]–[49]. Inspired by this, hybrid architectures that integrate convolutional local feature extraction with transformer-based global modeling have further pushed performance boundaries, achieving excellent results on HuBMAP glomerular datasets [13].

Despite these advances, existing frameworks still struggle to achieve precise boundary delineation (e.g., distinguishing sclerotic glomerular capsules from surrounding fibrosis) and ensure generalizability across diverse disease conditions. Studies including [50], [51] and [21] have attempted to enhance accuracy through ensemble modeling; however, this approach inevitably incurs increased computational overhead, limiting its applicability in clinical real-time analysis.

III. Methodology

In this section, we present a comprehensive overview of our proposed architecture, DualSeg, which is designed for the precise extraction and synthesis of multi-scale textural features from glomeruli in renal histopathological images. The architecture, illustrated in the Fig 3, is composed of three core components: (1) a dual-stage hybrid encoder that integrates an initial-stage Wave-Attention Block for local feature refinement, (2) a subsequent-stage VRWKV module for global context modeling, and (3) a lightweight decoder for efficient feature aggregation. This modular design ensures robust representation learning while maintaining computational efficiency, addressing the dual challenges of fine-grained texture discrimination and long-range dependency modeling in complex pathological scenarios.

A. Wave-Swin Block

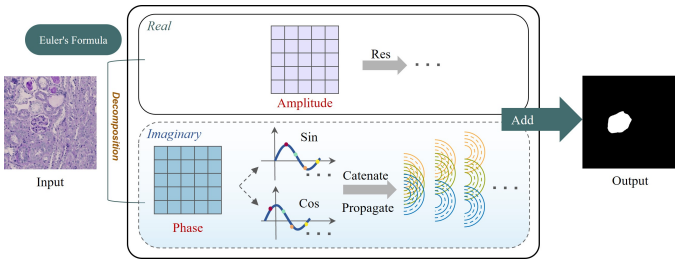

Fig. 4: Illustration of the Token-Mixing mechanism within the **Wave-Swin** Block. The input is initially decomposed into a *real* component and an *imaginary* component via Euler's formula. The modulated phase, undergoes two distinct cosine transformations and window-based propagation, is fused with the amplitude information to generate the output.

To address the dual challenges of fine-grained texture discrimination and global contextual modeling in renal glomerular segmentation, we designed a hybrid Dual-Stage encoder that seamlessly integrates CNNs with VRWKV. By synergistically combining these components, the encoder enables robust and scalable glomerular feature extraction across both local

and global scales, effectively bridging the gap between texture fidelity and contextual awareness in complex histopathological scenarios.

Initially in the first stage, the encoder employs improved Wave-Attention layers for local texture analysis, enabling precise extraction of glomerular features due to their powerful local feature extraction capabilities. Immediately after, we employed VRWKV to address the issue of global irregular spatial arrangement affecting renal tubulointerstitial fibrosis.

To circumvent the aforementioned methodological limitation of local texture discriminability, we propose the Wave-Swin (WS) block, a novel attention mechanism tailored for glomerular segmentation. The WS block is strategically designed to address the challenge through a wave-based representation and dynamic feature fusion. The illustration of its main part, Token-Mixing, is shown in Fig. 4.

Conceptually, we consider the pathology renal image, $P = [p_1, p_2, \ldots, p_n]$ as the input of Wave Attention with p_j standing for the tokens of the input image, the amplitude of w_j is defined by :

$$w_j = CF(p_j, W^C), \quad j = 1, 2, 3, \ldots, n, \tag{1}$$

where CF stands for Channel-Fusion operation and W^C is a learnable weight.

Then modeling each image token as a complex **wave** $w_j^{\sim}$:

$$w_j^{\sim} = |p_j| \Theta e^{i\theta_j}, \quad j = 1, 2, 3, \ldots, n, \tag{2}$$

where the amplitude $|p_j| \in \mathbb{R}^+$ represents the semantic intensity (ground truth context), phase $\theta_j \in [0, 2\pi)$ encodes the token's relative position within the glomerular structure, and Θ denotes the element-wise multiplication.

By invoking Euler's formula $e^{i\theta} = \cos\theta + i\sin\theta$, the wave token can be decomposed into orthogonal spatial components:

$$w_j^{\sim} = |p_j| \Theta \cos\theta_j + i|p_j| \Theta \sin\theta_j, \quad j = 1, 2, 3, \ldots, n, \tag{3}$$

The phase term θ_j in the wave representation is strategically designed to encode the spatial relationships between glomeruli and surrounding interstitial tissues. Specifically, glomeruli in cortical regions demonstrate more compact clustering and smoother boundaries compared to those near the medulla, where fibrotic remodeling disrupts structural continuity. The phase θ_j captures these zonal microanatomical variations by assigning distinctive angular values to tokens based on their relative proximity to key anatomical landmarks, such as the renal capsule.

Resultant complex value output tokens can be aggregated through the token-mixing operation, as expressed by:

$$I_j^{\sim} = TF(W^{\sim}, W_k^T), \quad j = 1, 2, 3, \ldots, n, \tag{4}$$

TF stands for the token fusion operation, W_k^T is the weight with learnable parameters, $W^{\sim} = [w_1^{\sim}, w_2^{\sim}, \ldots, w_n^{\sim}]$ is the set of tokens, and the parameter k in W_k^T is the dynamic size of the wave propagation window.

To overcome the static receptive field limitation of prior wave-propagation windows, we propose a dynamic Swin mechanism that adaptively reshapes the wavefront. Specifically, we define a set $S = \{7, 11, 15, \ldots\}$ of candidate odd-sized windows, informed by both domain-validated baselines and dataset-specific anatomical statistics. First, Wave-MLP has empirically demonstrated that windows smaller than 7 lack the generality necessary to capture spatial dependencies in medical images, thus smaller sizes are a priori excluded; meanwhile, the anchor sizes 7 and 11 align with the kernel sizes employed in SegNeXt. Second, the average glomerular bounding box in our murine dataset measures 154px [45]; after the $4\times$ and $8\times$ downsampling stages, this reduces to 38px and 19px, respectively. Accordingly, selecting 15 instead of 21 prevents the network from capturing extraneous background information while still fully covering the target glomerular structure. These window sizes can be easily tailored to other imaging modalities. For each token, the mapping function $M : S \to \{1, 2, \ldots, |S|\}$ is defined and tailored based on the aforementioned benchmarks and pixel-level statistics to select the optimal window size, ensuring the faithful encoding of diverse spatial patterns.

As output of the wave attention, the real value I_j can be estimated through the sum of real and imaginary parts of the $I_j^{\sim}$ as defined by:

$$I_j = \sum_n (W_{jkn}^T w_n \Theta cos\theta_n + W_{jn}^C w_n \Theta sin\theta_n), \tag{5}$$

By introducing the concept of dynamic wave propagation window, the WS module effectively enhances token-weight correlations. This wave-based representation methodology substantially improves the feature extraction capabilities in glomerular segmentation, enabling the model to comprehensively consider glomerular positional and semantic information within a broader contextual framework while systematically eliminating redundant computational representations.

B. VRWKV Block

After wave feature extraction, VRWKV block establishes non-local constraints by modeling pairwise affinities between glomerular candidates, solving the problem of spatial heterogeneity.

The input image x_0 undergoes four layers of downsampling:

$$x_i = DownSample(x_{i-1}), i = 1, 2, 3, 4, \tag{6}$$

after the first two layers of Wave-Swin blocks and downsampling, the feature map $x_3 \in \mathbb{R}^{\frac{H}{8} \times \frac{W}{8} \times C_2}$ is obtained.

To establish connections between glomeruli, the feature map x_{co} is processed by the third downsampling layer with the VRWKV mechanism. Specifically, x_{co} is serialized via linear projection and patch embedding, with positional encoding incorporated. The serialization process is formulated as:

$$\text{Tokens} = \text{Flatten}\left(\text{LN}(x_{\text{co}})\right), \tag{7}$$

where Flatten denotes the reshape operation that converts 2D feature maps into 1D token sequences, and LN represents linear projection. The resulting 1D token sequence satisfies $\text{Tokens} \in \mathbb{R}^{P^2 \times \text{Dim}}$, where $P^2 = \frac{H}{2^i} \times \frac{W}{2^i}$ ($i \in \{4, 5\}$), H/W denote the height/width of the original feature map, and Dim is the corresponding hidden dimension. These tokens are then fed into the first VRWKV encoder layer for further processing.

TABLE I: PERFORMANCE COMPARISON FOR GLOMERULAR SEGMENTATION ON THE FIVE-FOLD CROSS-VALIDATION OF THE **KPIs** DATASET WITH RESPECT TO EXISTING METHODS

Models	DN			NEP25			Normal			5/6Nx			AVG		
	mDSC↑	HD95↓	IoU↑	mDSC↑	HD95↓	IoU↑	mDSC↑	HD95↓	IoU↑	mDSC↑	HD95↓	IoU↑	mDSC↑	HD95↓	IoU↑
U-Net [7]	0.8973 ±0.0025	74.9936 ±2.3086	0.8737 ±0.0027	0.8737 ±0.0028	76.5603 ±1.8606	0.8469 ±0.0029	0.9091 ±0.0023	77.0467 ±2.4556	0.8860 ±0.0025	0.8332 ±0.0032	160.2147 ±2.7958	0.8052 ±0.0034	0.8859 ±0.2657	93.3440 ±2.4627	0.8628 ±0.0028
Attention U-Net [52]	0.9006 ±0.0025	60.0169 ±1.9383	0.8784 ±0.0026	0.8837 ±0.0027	85.9233 ±2.1759	0.8588 ±0.0028	0.9153 ±0.0023	69.4993 ±2.2980	0.8940 ±0.0024	0.8237 ±0.0033	204.4328 ±3.4987	0.7976 ±0.0035	0.8905 ±0.0026	97.0253 ±2.5783	0.8676 ±0.0028
SegNext [30]	0.9242 ±0.0022	77.6273 ±2.4398	0.9066 ±0.0023	0.9168 ±0.0023	256.4412 ±2.3992	0.8984 ±0.0024	0.9331 ±0.0021	68.0672 ±2.2544	0.9157 ±0.0022	0.8444 ±0.0032	253.6379 ±4.0326	0.8236 ±0.0033	0.9118 ±0.0024	109.2467 ±2.8236	0.8935 ±0.0025
DeepLabv3+ [46]	0.9283 ±0.0022	77.1417 ±2.4108	0.9115 ±0.0023	0.9213 ±0.0023	80.2639 ±2.0792	0.9044 ±0.0024	0.9356 ±0.0020	64.7229 ±2.2177	0.9188 ±0.0021	0.8466 ±0.0032	240.0954 ±3.9110	0.8254 ±0.0033	0.9147 ±0.0024	103.9777 ±2.7477	0.8970 ±0.0025
Wave-MLP [25]	0.9299 ±0.0021	71.4539 ±2.4084	0.9133 ±0.0022	0.9189 ±0.0023	77.6752 ±2.0753	0.9006 ±0.0024	0.9361 ±0.0020	62.6423 ±2.1879	0.9189 ±0.0021	0.8375 ±0.0033	274.2791 ±4.1231	0.8153 ±0.0034	0.9131 ±0.0024	108.5067 ±2.8366	0.8949 ±0.0025
InceptionNext [53]	0.7798 ±0.0037	279.4065 ±3.5819	0.7510 ±0.0038	0.8445 ±0.0030	117.9740 ±2.3975	0.8136 ±0.0032	0.8490 ±0.0030	113.4765 ±2.3526	0.8165 ±0.0031	0.7333 ±0.0041	375.5270 ±4.5353	0.7116 ±0.0042	0.8164 ±0.0033	187.3932 ±3.2660	0.7869 ±0.0035
SegFormer [11]	0.9332 ±0.0021	71.8198 ±2.3607	0.9174 ±0.0022	0.9235 ±0.0023	85.5397 ±2.2958	**0.9064 ±0.0023**	0.9363 ±0.0020	65.9465 ±2.2332	0.9194 ±0.0021	0.8166 ±0.0035	325.5992 ±4.5238	0.7964 ±0.0036	0.9102 ±0.0025	121.5212 ±3.0465	0.8927 ±0.0026
VRWKV [26]	0.9330 ±0.0021	70.7004 ±2.3996	0.9176 ±0.0022	0.9232 ±0.0023	77.5173 ±2.1069	0.9061 ±0.0024	**0.9370 ±0.0020**	61.6787 ±2.1709	**0.9201 ±0.0021**	0.8738 ±0.0029	188.1708 ±3.4140	0.8518 ±0.0030	0.9013 ±0.0028	90.5756 ±2.5482	0.8942 ±0.0027
VM-UNET-V2 [49]	0.8983 ±0.0025	95.4675 ±2.5092	0.8757 ±0.0027	0.8843 ±0.0027	113.2830 ±2.4918	0.8604 ±0.0028	0.9077 ±0.0024	87.9016 ±2.4350	0.8850 ±0.0025	0.7248 ±0.0042	390.3619 ±5.1129	0.7099 ±0.0043	0.8665 ±0.0030	153.0778 ±3.3923	0.8452 ±0.0031
UNETR [9]	0.8667 ±0.0028	93.1881 ±2.3828	0.8382 ±0.0030	0.8187 ±0.0033	146.9859 ±2.5678	0.7874 ±0.0034	0.8745 ±0.0027	79.2260 ±2.1871	0.8450 ±0.0029	0.7243 ±0.0042	394.2254 ±4.8802	0.7059 ±0.0043	0.8361 ±0.0032	153.2457 ±3.2384	0.8088 ±0.0033
Swin UNETR [54]	0.9142 ±0.0023	85.0643 ±2.5062	0.8945 ±0.0025	0.9025 ±0.0025	86.2819 ±2.0911	0.8810 ±0.0026	0.9158 ±0.0024	78.0189 ±2.4094	0.9067 ±0.0023	0.7223 ±0.0043	462.2809 ±5.5391	0.7081 ±0.0043	0.8801 ±0.0029	157.4215 ±3.6115	0.8615 ±0.0030
DA-TransUNet [55]	0.9190 ±0.0023	88.3991 ±2.4660	0.9004 ±0.0024	0.9111 ±0.0024	80.9081 ±2.0693	0.8917 ±0.0025	0.9278 ±0.0021	82.5779 ±2.5860	0.9095 ±0.0022	0.7823 ±0.0038	334.6921 ±4.4253	0.7619 ±0.0039	0.8950 ±0.0027	134.0005 ±3.1449	0.8761 ±0.0028
H2Former [10]	0.9259 ±0.0022	73.2565 ±2.3295	0.9082 ±0.0023	0.9173 ±0.0023	80.9923 ±1.9999	0.8993 ±0.0024	0.9315 ±0.0021	65.9763 ±2.2418	0.9135 ±0.0022	0.8141 ±0.0035	307.6074 ±4.4570	0.7936 ±0.0036	0.9052 ±0.0025	117.6038 ±2.9761	0.8868 ±0.0026
U-mamba [28]	0.8794 ±0.0027	75.3772 ±1.8672	0.8545 ±0.0029	0.8799 ±0.0028	82.9104 ±1.8076	0.8563 ±0.0029	0.9172 ±0.0023	63.2446 ±2.1390	0.8980 ±0.0024	0.7335 ±0.0041	325.2652 ±4.2271	0.7151 ±0.0042	0.8693 ±0.0030	120.3657 ±2.8212	0.8487 ±0.0031
DualSeg(ours)	**0.9603 ±0.0012**	**62.9389 ±1.7681**	**0.9386 ±0.0014**	**0.9349 ±0.0016**	**74.7815 ±1.5739**	0.9055 ±0.0018	0.9187 ±0.0005	**33.2828 ±1.1220**	0.8537 ±0.0008	**0.9007 ±0.0022**	**149.6097 ±2.5757**	**0.8677 ±0.0025**	**0.9298 ±0.0016**	**66.8485 ±1.7330**	**0.8967 ±0.0019**

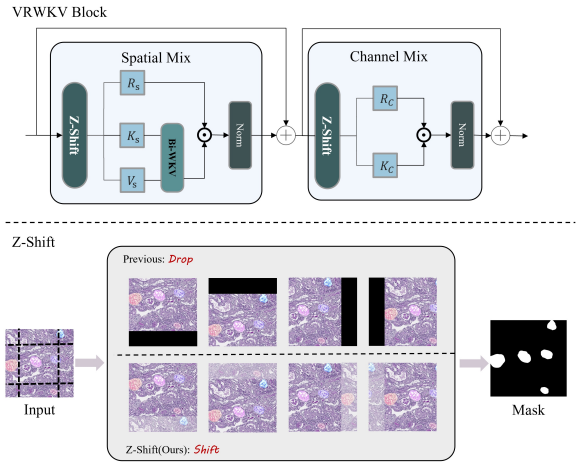

Fig. 5: Structure of the imporoved VRWKV block. Black margins in the lower panel illustrate the prior approach that pixels shifted beyond the boundary are simply dropped, our $Z\text{-}Shift$ instead wraps these pixels to the opposite edge, filling the faded regions and eliminating information loss.

Within this encoder layer, tokens are first processed in the Spatial Mixing module. Specifically, tokens are shifted and input into three parallel linear layers to generate matrices $R_s, K_s, V_s \in \mathbb{R}^{P^2 \times \mathrm{Dim}}$, which serve as the reset gate, key, and value, respectively:

$$N_s = \text{Z-Shift}(\text{Tokens})W^N, \quad N \in \{R, K, V\}, \tag{8}$$

where the Z-Shift operator is a modified version of the original Q-Shift operator in VRWKV [26]. Unlike Q-Shift (which discards a proportion of pixels according to the shift direction), Z-Shift shifts pixels in the opposite direction to mitigate the segmentation accuracy degradation caused by edge pixel loss (see Fig. 5). W^N denotes the learnable weight matrix for each branch ($N = R, K, V$).

The output S of the Spatial Mixing module is calculated as:

$$S = \sigma(R_s) \odot \text{Bi-WKV}(K_s, V_s), \tag{9}$$

where $\sigma(\cdot)$ denotes the sigmoid activation function, $\odot$ represents element-wise multiplication, and Bi-WKV is the linear-complexity bidirectional attention (Bi-Attention) mechanism of VRWKV, which fuses spatial feature extraction and attention computation.

Subsequently, the tokens are passed to the Channel-Mix module for channel-wise feature fusion. Matrices R_c and K_c are generated from the Spatial Mixing output S using the same shift-and-linear-projection process as in Spatial Mixing. The output O_c of the Channel-Mix module is expressed as:

$$O_c = \sigma(R_c) \odot V_c, \tag{10}$$

where $V_c = \text{SquaredReLU}(K_c)W_V$, and $\text{SquaredReLU}(\cdot)$ denotes the squared ReLU activation function.

The output O_c (derived from the first VRWKV block in the initial VRWKV encoder layer) is converted back to a 2D feature map via the PatchMerging operation (the inverse of Flatten). This spatially reconstructed feature map undergoes downsampling and is then input into the subsequent VRWKV encoder layer, where hierarchical feature refinement is performed through sequential processing blocks. After computation in the second VRWKV encoder layer, the top-level feature map $x_{\text{ot}} \in \mathbb{R}^{\frac{H}{32} \times \frac{W}{32} \times C_4}$ is transmitted to the decoder, along with outputs from earlier layers, to enable multi-scale feature integration.

The necessity of combining CNNs and VRWKV stems from their complementary nature in feature extraction. Through

integrated convolution processes and bidirectional attention mechanisms in the encoder, these models synergistically combine local and global glomerular information. Wave-Swin Block excels at extracting intricate local glomerular details, while VRWKV effectively captures global contextual information through Bi-Attention. This integrated approach simultaneously leverages local and global features, maintaining the model's comprehensive perception capabilities while optimizing computational efficiency.

C. The Lightweight Decoder

Due to the fact that the features from the first CNN layer contain excessive low-level information that negatively impacts model performance, we selectively exclude the output features of the first layer [30].

Accordingly, we opted to employ a lightweight decoder, *HamDecoder* [56] , to fuse the final three feature maps. Designed as a key component of our multi-scale interaction framework, HamDecoder specifically addresses the challenge of morphological priors integration by synergizing multi-scale texture features. Non-negative Matrix Factorization(NMF) efficiently improves segmentation accuracy by matrix decomposing feature maps and removing noise:

$$Mask = D \times C + N, \tag{11}$$

$Mask$ is the integrated rough features, and D, C, N represent Dictionary, Codes and Noisy respectively. This yields a purer segmentation mask.

Combined with our Dual-Stage Hybrid encoder, this convolutional and NMF-based decoder provides a complementary enhancement to the overall architecture.

IV. Experiments

In this section, we evaluate the method on three glomeruli datasets. We first introduce the benchmark datasets, implementation details, and evaluation metrics. Then we compare our method with state-of-the-art methods and perform ablation studies to validate the design choices of the proposed architecture.

A. Datasets

Mice Glomeruli dataset. The mice kidney dataset used in this study was shared by the MICCAI 2024 Kidney Pathology Image Segmentation (KPIs) challenge [57], which was available from four groups of mouse models: normal mice, 5/6 nephrectomy (5/6Nx) mice, diabetic nephropathy (DN) mice, and NEP25 mice, stained with PAS. To maintain consistency in patch-level tasks and evaluation metrics across three datasets, we selected Task 1 of the KPIs dataset as our benchmark. The image patches were provided at a fixed resolution of $2,048 \times 2,048$ pixels, each containing FTUs, with dataset splits of 5,213 training images, 1,643 validation images, and 2,305 test images. To convert images to a size compatible with the model, we performed non-overlapping cropping on training and validation images to obtain 256×256 patches and implemented 5-fold cross-validation, designating 80% as the training set and the remainder as the validation set. Finally, the results were derived from the test set for comparative analysis.

Human Glomeruli dataset I. The first human kidney dataset used in this study was obtained from the HuBMAP Kidney Challenge [58], aimed at advancing research in the segmentation of functional tissue units (FTUs) in renal histology. The dataset comprises 20 PAS-stained WSIs in TIFF format, with an average resolution of approximately $36,000 \times 29,000$ pixels. Pixel-level annotations, delineate each glomerulus using polygonal contours that closely follow its boundary. Similarly, we first partitioned them into 256×256 patches, generating a total of 258,956 patches. For the extracted patches, we implemented a 5-fold cross-validation strategy. In each fold, 20% of the patches were randomly selected as the validation set, whereas the remaining 80% were utilized for training [45]. Following the 5-fold cross-validation, we uploaded the weights to the Kaggle platform, performed inference on the test set, and compared the resulting outcomes.

Human Glomeruli Dataset II. The second human kidney dataset utilized in this study was retrived from the Kidney Precision Medicine Project (KPMP) Atlas Repository [59]. Four PAS-stained WSIs in SVS format, alongside their corresponding segmentation masks, were randomly selected; these images were generated in accordance with the KPMP Main Protocol and have an average resolution of approximately $84,000 \times 50{,}000$ pixels. To validate the feasibility and effectiveness of cross-species training, as well as the generalization capability of our proposed model, following their initial training on the mouse KPIs dataset, all models were directly applied to inference on this KPMP human dataset without retraining. For consistency with the preprocessing pipeline of the mouse KPIs dataset, the KPMP WSIs were directly partitioned into 2048×2048 patches.

B. Baselines

We benchmark DualSeg against 14 baseline methods that are representative of three contemporary architectural paradigms, ensuring a comprehensive performance evaluation. (1)**CNN-based models**: U-Net [7] and Attention U-Net [52] are regarded as canonical U-shaped networks with skip connections, the latter appending attention gates to suppress background activations, thereby providing elementary references for glomerular boundary delineation. SegNext [30] re-engineers convolutional attention blocks and furnishes an efficient large-kernel-convolution alternative, testing adaptability to diverse glomerular morphologies. DeepLabV3+ [46] adopts atrous spatial pyramid pooling to enlarge receptive fields without sacrificing spatial resolution, serving as a standard multi-scale CNN baseline. Wave-MLP [25] incorporates wave-like token mixing within an MLP backbone, preserving fine structural details with fewer parameters and thus accommodating size-varying glomerular profiles. InceptionNext [53] incorporates large-kernel convolution within the Inception [60] framework, achieving a favorable balance between computational efficiency and fine-grained feature extraction. (2)**Transformer-based models**: SegFormer [11] employs a hierarchical Transformer encoder coupled with a lightweight decoder to capture long-range dependencies inherent in irregular

TABLE II: PERFORMANCE COMPARISON FOR GLOMERULAR SEGMENTATION ON THE TEST SET OF THE **HUBMAP** DATASET WITH RESPECT TO EXISTING METHODS

Models	mDSC↑											
	Fold1		Fold2		Fold3		Fold4		Fold5		AVG	
	Private	Public	Private	Public	Private	Public	Private	Public	Private	Public	Private	Public
UNet [7]	0.8731	0.7923	0.8629	0.7812	0.8714	0.7899	0.8512	0.7865	0.8711	0.7918	0.8659 ± 0.0082	0.7883 ± 0.0041
Attention-UNet [52]	0.8699	0.7731	0.8661	0.7662	0.8550	0.7770	0.8408	0.7821	0.8551	0.7742	0.8574 ± 0.0102	0.7745 ± 0.0052
SegNext [30]	0.7834	0.7447	0.7608	0.7244	0.7597	0.7445	0.6938	0.7108	0.7416	0.7117	0.7479 ± 0.0301	0.7272 ± 0.0150
DeepLabV3+ [46]	0.8658	0.7901	0.8750	0.7742	0.8726	0.7821	0.8307	0.7805	0.8307	0.7805	0.8550 ± 0.0200	0.7815 ± 0.0051
Wave-MLP [25]	0.8753	0.7996	0.8461	0.7843	0.8669	0.8063	0.8474	0.8002	0.8684	0.7701	0.8608 ± 0.0118	0.7921 ± 0.0132
InceptionNext [53]	0.7902	0.7233	0.7698	0.7532	0.7909	0.7341	0.7102	0.6810	0.7606	0.7225	0.7943 ± 0.0295	0.7228 ± 0.0237
SegFormer [11]	0.8926	0.8120	0.8795	0.7832	0.8855	0.8048	0.8908	0.8098	0.8781	0.8158	0.8853 ± 0.0058	0.8051 ± 0.0115
VRWKV [26]	0.8931	0.8330	0.8814	0.8171	0.8877	0.8006	0.8680	0.7901	0.8886	0.8263	0.8838 ± 0.0087	0.8134 ± 0.0159
VM-UNET-V2 [49]	0.8772	0.7913	0.8488	0.7818	0.8555	0.7917	0.8322	0.7996	0.8661	0.7400	0.8560 ± 0.0153	0.7809 ± 0.0212
UNETR [9]	0.5819	0.5725	0.6329	0.6267	0.5819	0.5916	0.5728	0.5757	0.6457	0.6171	0.6030 ± 0.0301	0.5967 ± 0.0218
Swin-UNETR [54]	0.8559	0.7718	0.8531	0.7546	0.8552	0.7768	0.8571	0.7764	0.8313	0.7666	0.8505 ± 0.0097	0.7692 ± 0.0082
DA-TransUNet [55]	**0.8960**	0.7911	0.8944	0.7692	**0.9019**	0.7906	0.8887	0.7931	0.8866	0.7864	0.8935 ± 0.0054	0.7861 ± 0.0087
H2Former [10]	0.7991	0.7075	0.7644	0.6808	0.8151	0.7547	0.6443	0.5675	0.7825	0.7298	0.7611 ± 0.0608	0.6881 ± 0.0650
U-mamba [28]	0.8667	0.7699	0.8671	0.7595	0.8543	0.7580	0.8312	0.7723	0.8600	0.7701	0.8559 ± 0.0132	0.7660 ± 0.0060
DualSeg(Ours)	0.8925	**0.8366**	**0.8948**	**0.8488**	0.9016	**0.8441**	**0.8972**	**0.8217**	**0.9032**	**0.8354**	$\mathbf{0.8979 \pm 0.0040}$	$\mathbf{0.8373 \pm 0.0092}$

TABLE III: **CROSS-DATASET INFERENCE** PERFORMANCE COMPARISON FOR GLOMERULAR SEGMENTATIO ON THE **KPMP** DATASET USING 5-FOLD MOUSE-TRAINED MODELS WITH RESPECT TO EXISTING METHODS

Models	1			2			3			4			AVG		
	mDSC↑	HD95↓	IoU↑	mDSC↑	HD95↓	IoU↑	mDSC↑	HD95↓	IoU↑	mDSC↑	HD95↓	IoU↑	mDSC↑	HD95↓	IoU↑
U-Net [7]	0.3936 ±0.0045	279.4223 ±2.4890	0.5095 ±0.0045	0.2662 ±0.0042	471.9885 ±2.5648	0.2560 ±0.0042	0.5189 ±0.0047	119.5596 ±1.8510	0.4973 ±0.0047	0.5387 ±0.0046	152.3241 ±2.0956	0.5095 ±0.0045	0.4836 ±0.0046	196.2395 ±2.4062	0.4588 ±0.0046
Attention U-Net [52]	0.7165 ±0.0041	165.1443 ±2.1467	0.6875 ±0.0041	0.4762 ±0.0046	274.6819 ±2.4661	0.4540 ±0.0047	0.8431 ±0.0033	54.8253 ±1.1642	0.8230 ±0.0034	0.6953 ±0.0040	90.7108 ±1.5556	0.6559 ±0.0041	0.7056 ±0.0041	110.6999 ±1.7900	0.6736 ±0.0041
SegNext [30]	0.7926 ±0.0036	164.5795 ±2.2619	0.7656 ±0.0037	0.5235 ±0.0046	86.5286 ±2.3992	0.4978 ±0.0046	0.8488 ±0.0032	36.7357 ±0.5685	0.8277 ±0.0033	0.7287 ±0.0039	103.8195 ±1.7509	0.6948 ±0.0040	0.7410 ±0.0039	112.1102 ±1.8356	0.7116 ±0.0040
DeepLabv3+ [46]	0.7846 ±0.0038	100.0552 ±1.9003	0.7640 ±0.0038	0.6030 ±0.0043	182.8433 ±2.1006	0.5615 ±0.0043	0.8545 ±0.0031	67.7374 ±1.6939	0.8333 ±0.0032	0.6952 ±0.0041	116.7432 ±1.9572	0.6617 ±0.0041	0.7317 ±0.0040	112.6154 ±1.9438	0.7018 ±0.0040
Wave-MLP [25]	0.8152 ±0.0034	149.4421 ±2.1076	0.7899 ±0.0036	0.5158 ±0.0046	299.4355 ±2.7413	0.4929 ±0.0047	0.8505 ±0.0031	49.6731 ±0.8149	0.8253 ±0.0032	0.7675 ±0.0037	92.2304 ±1.6319	0.7342 ±0.0038	0.7646 ±0.0038	106.0041 ±1.7929	0.7353 ±0.0038
InceptionNext [53]	0.6564 ±0.0044	236.3450 ±2.3658	0.6352 ±0.0045	0.4779 ±0.0048	272.6440 ±2.6615	0.4672 ±0.0049	0.7136 ±0.0042	167.7567 ±1.9100	0.6937 ±0.0043	0.4301 ±0.0043	312.1570 ±2.4503	0.3875 ±0.0042	0.5279 ±0.0045	284.8018 ±2.4428	0.4967 ±0.0046
SegFormer [11]	0.8037 ±0.0036	125.7612 ±1.8165	0.7814 ±0.0037	**0.6151 ±0.0044**	206.3655 ±2.4028	**0.5815 ±0.0044**	0.8597 ±0.0030	47.1347 ±1.0667	0.8353 ±0.0031	0.8285 ±0.0033	**49.0393 ±1.1678**	0.7986 ±0.0033	0.8082 ±0.0035	70.7543 ±1.4692	0.7802 ±0.0035
VRWKV [26]	0.7912 ±0.0036	143.4809 ±2.1293	0.7662 ±0.0037	0.5112 ±0.0046	271.6789 ±2.2767	0.4883 ±0.0047	0.8056 ±0.0036	62.1875 ±1.0115	0.7850 ±0.0037	0.7615 ±0.0037	65.9837 ±1.3855	0.7266 ±0.0038	0.7480 ±0.0039	89.0189 ±1.6278	0.7188 ±0.0039
VM-UNET-V2 [49]	0.6521 ±0.0043	181.1196 ±1.9559	0.6233 ±0.0044	0.4833 ±0.0046	332.8376 ±2.5972	0.4575 ±0.0046	0.7880 ±0.0036	104.5297 ±1.5290	0.7620 ±0.0038	0.5410 ±0.0043	232.7710 ±2.4792	0.4963 ±0.0043	0.6028 ±0.0044	216.0202 ±2.3843	0.5664 ±0.0044
UNETR [9]	0.6199 ±0.0046	250.5465 ±2.1782	0.6050 ±0.0047	0.4500 ±0.0048	391.7400 ±2.8152	0.4416 ±0.0049	0.6436 ±0.0047	158.4295 ±1.8608	0.6349 ±0.0047	0.4899 ±0.0046	333.3864 ±2.6834	0.4416 ±0.0049	0.5369 ±0.0047	308.9828 ±2.6276	0.5193 ±0.0047
Swin UNETR [54]	0.5171 ±0.0047	207.6704 ±2.4884	0.4921 ±0.0046	0.3438 ±0.0045	432.7990 ±2.8733	0.3300 ±0.0045	0.7037 ±0.0042	85.6554 ±1.7000	0.6805 ±0.0043	0.6599 ±0.0041	131.7022 ±1.9244	0.6201 ±0.0042	0.6137 ±0.0044	156.2208 ±2.2260	0.5824 ±0.0044
DA-TransUNet [55]	0.7783 ±0.0038	97.7011 ±1.4945	0.7563 ±0.0039	0.5907 ±0.0045	**166.4438 ±2.1574**	0.5610 ±0.0045	0.8319 ±0.0034	63.6652 ±1.4344	0.8143 ±0.0035	0.7948 ±0.0035	81.5234 ±1.6102	0.7594 ±0.0035	0.7780 ±0.0037	87.6505 ±1.6488	0.7489 ±0.0038
H2Former [10]	0.7490 ±0.0039	148.2938 ±2.0519	0.7220 ±0.0040	0.5627 ±0.0045	219.4662 ±2.3340	0.5323 ±0.0045	0.8095 ±0.0035	55.2702 ±1.0098	0.7859 ±0.0036	0.6369 ±0.0042	169.6482 ±2.0895	0.5934 ±0.0042	0.6815 ±0.0041	154.2361 ±2.0343	0.6460 ±0.0042
U-mamba [28]	0.6549 ±0.0043	132.5447 ±1.7257	0.6214 ±0.0043	0.3967 ±0.0045	315.6740 ±1.6164	0.3729 ±0.0045	0.7500 ±0.0040	75.1434 ±1.5002	0.7262 ±0.0040	0.5380 ±0.0043	192.8970 ±2.3236	0.4878 ±0.0041	0.5843 ±0.0044	181.0034 ±2.1725	0.5449 ±0.0043
DualSeg(ours)	**0.8251 ±0.0034**	**81.6201 ±1.5314**	**0.8022 ±0.0035**	0.5888 ±0.0045	235.8327 ±2.6482	0.5573 ±0.0045	**0.8848 ±0.0028**	**21.0608 ±0.2919**	**0.8630 ±0.0029**	**0.8393 ±0.0032**	57.6049 ±1.3786	**0.8123 ±0.0033**	**0.8195 ±0.0034**	**69.6369 ±1.5420**	**0.7938 ±0.0035**

glomerular shapes while maintaining rapid inference. VRWKV [26] substitutes quadratic self-attention with a linear recurrent operator, delivering a memory-efficient Transformer variant suited to high-resolution histological images. VM-UNET-V2 [49] adopts a purely SSM-based architecture, enhanced by semantic and detail injection modules to facilitate effective multi-scale feature fusion. (3)**Hybrid models**: UNETR [9] and Swin UNETR [54] integrate pure or shifted-window Transformer encoders within a U-shaped decoder, the latter's locality-aware windows being particularly amenable to heterogeneous glomerular textures. H2Former [10] balances local texture extraction and global context via hybrid conv-attention blocks embedded in a U-Net, epitomizing current trends in medical image segmentation. DA-TransUNet [55] accentuates ambiguous glomerular contours through spatial–channel dual attention in a TransUNet backbone, exemplifying attention-augmented hybrid Transformer approaches. U-mamba [28] pioneers the integration of the SSM into medical image segmentation, opening a new avenue for high-performance solutions in the field.

C. Evaluation Metrics

To assess segmentation performance, we utilized a widely used metric: the Dice coefficient, the Hausdorff Distance 95% percentile (HD95) and the Intersection over Union (IoU). Given two generic sets A and B, the three indices are defined in Eqs. 12 to 14:

$$Dice = \frac{2 \mid A \cap B \mid}{\mid A \mid + \mid B \mid}. \tag{12}$$

With regard to the HuBMAP dataset, results were obtained by uploading predictions to the official evaluation portal.

For comprehensive evaluation, we additionally measured HD95 and IoU metrics on the KPIs and KPMP datasets:

$$HD95 = \max_{p95} \left\{ \sup_{a \in A} \inf_{b \in B} d(a,b), \sup_{b \in B} \inf_{a \in A} d(b,a) \right\}, \tag{13}$$

where sup represents the supremum operator, inf is the infimum operator, with $\inf_{a \in A} d(b,a)$ quantifying the distance from point $a \in X$ to the subset $B \subseteq X$, $\max_{p95}$ is the 95% percentile of surface distances between segmentation and Ground Truth (GT).

$$IoU = \frac{\mid A \cap B \mid}{\mid A \cup B \mid}, \tag{14}$$

when used for semantic segmentation performance assessments, the two sets A and B are respectively the set of the pixels in the ground truth mask and of the pixels in the segmentation mask produced by the deep neural network.

D. Implementation Details

To mitigate overfitting and enhance model generalization, we incorporated targeted data augmentation strategies encompassing horizontal flips, vertical flips, and random rescaling. The network underwent 20 training epochs using the NAdam optimizer [61] with a weight decay of 0.0001; the initial learning rate was set to 0.0001 and dynamically adjusted via Cosine Annealing [62] for adaptive convergence.

For both HuBMAP and KPIs datasets, we utilized pre-cut 256×256 images paired with corresponding annotations at a batch size of 16. A key innovative adaptation was implemented for cross-dataset inference: given the official KPIs test split consists of 2048×2048 images, we cropped the KPMP dataset into 2048×2048 patches and downscaled them by a factor of 0.4 to specifically addressing the inherent glomerular size discrepancy between mouse (KPMP) and human (KPIs) samples [45]. Critically, this design enabled direct cross-dataset inference using KPIs-pretrained models without retraining, eliminating domain shift-induced performance degradation while streamlining the inference pipeline.

Patches were inferred via *Sliding Window Inference* from MONAI [63] with a fixed window size of 256×256, for consistent high-resolution of both KPIs and KPMP datasets,. Notably, all experiments were conducted under strictly identical parameter settings to ensure rigorous, unbiased comparison across competing models.

V. Results

In this section, we present the performances of the proposed DualSeg on three different renal pathology image segmentation datasets compared to widely applied methods aforementioned. Besides, in order to verify the validity of the proposed blocks and dual-stage encoder, we perform ablation studies to validate the design choices of the proposed architecture. Both quantitative and qualitative results are introduced in this section.

A. Mice Glomeruli Segmentation

The performance of DualSeg and baseline methods on the KPIs dataset, the mice glomeruli, is quantified in Table I, with evaluations based on mDSC, HD95, and IoU. DualSeg outperforms all baselines across all pathological categories, with its advantages becoming more pronounced in complex scenarios.

In DN cases—characterized by mild glomerular hypertrophy—DualSeg achieves an mDSC of 96.03%, surpassing the second-ranked VRWKV by 2.73% and reducing HD95 by 7.76. For NEP25 mice, DualSeg delivers an mDSC of 93.49%, outperforming SegNext by 1.81% and lowering HD95 by 11.75, highlighting its robustness to morphological variations. Most strikingly—which represent the most challenging subset among the four evaluated categories for segmentation—DualSeg maintains an mDSC of 90.07%, the only method to exceed 90% in this difficult subset, with an associated HD95 of 149.61. This performance outperforms VRWKV by 2.69 percentage points in mDSC and by 38.56 units in HD95. Across all evaluation categories, DualSeg not only attains the highest average mDSC of 92.98% and lowest average HD95 of 66.85, outperforming H2Former and DA-TransUNet by 4.30% and 5.37% respectively, but also exhibits the lowest variance among all evaluation metrics.

B. Human Glomeruli Segmentation

1) HuBMAP Dataset: On the HuBMAP dataset, the human glomeruli, DualSeg maintains its dominance, as shown in Table II, with performance metrics including private mDSC and public mDSC.

Notably, DualSeg exhibits exceptional stability across data splits. It achieves private mDSC of 89.79% and public mDSC of 83.73%, outperforming the closest competitors by significant margins: DA-TransUNet by 0.44% and SegFormer by 1.26%, with a standard deviation of 0.0040 for private mDSC and 0.0092 for public mDSC, which contrasts sharply with baselines.

2) KPMP Dataset: On the KPMP dataset, DualSeg maintains its outstanding performance advantages, as shown in Table III. Notably, despite all models being pre-trained solely on the mouse KPIs dataset and directly deployed to this human dataset through **cross-dataset** inference, DualSeg still achieves highly competitive performance. For the average mDSC, DualSeg reaches 0.8195, surpassing SegFormer's 0.8082 and InceptionNext's mere 0.5279. In terms of HD95, DualSeg scores 69.6369, which is substantially reduced to one-third of that of VM-UNet-V2. Even when compared to other hybrid models, its IoU further achieves 0.7938, outperforming them by over 5%, which demonstrates far more robust performance.

In summary, DualSeg demonstrates consistent superior performance across the mouse KPI, human HuBMAP and KPMP datasets, validating its strong generalizability and robustness. During training on the mouse KPI dataset, DualSeg not only

TABLE IV: ABLATION STUDY OF MAJOR COMPONENTS ON THE TEST SET OF THE **KPIs** DATASET

Stage	Models	Layers			mDSC↑				
		Wave	Attention	VRWKV	DN	NEP25	Normal	5/6Nx	AVG
Sole-Stage	**Wave** [25]	✓	-	-	0.9165	0.9236	0.9322	0.8068	0.9036
	Attention [11]	-	✓	-	0.9223	0.9123	0.9249	0.8603	0.9099
	VRWKV [26]	-	-	✓	0.9209	0.9117	0.9268	0.8613	0.9108
Dual-Stage	**Attention-Wave(Ours)**	✓	✓	-	0.	0.	0.	0.	0.
	VRWKV-Wave(Ours)	✓	✓	-	0.	0.	0.	0.	0.
	Wave-Attention(Ours)	✓	✓	-	0.9267	0.9174	0.9334	0.8678	0.9171
	Wave-VRWKV(Ours)	✓	-	✓	0.9298	0.9163	0.9349	0.8972	0.9292

TABLE V: ABLATION STUDY OF THE PROPAGATION WINDOW ON THE TEST SET OF THE **KPIs** DATASET

Models	Propagation Window Size	mDSC↑				
		DN	NEP52	Normal	5/6Nx	AVG
Wave-MLP [25]	7	0.9299	0.9189	0.9361	0.8375	0.9131
	11	0.9228	0.9175	0.9366	0.8430	0.9075
	15	0.9227	0.9189	0.9144	0.8320	0.9012
	7-15	0.9365	0.9236	0.9322	0.8645	0.9146
DualSeg(Ours)	7	0.9544	0.9332	0.9245	0.8737	0.9205
	11	0.9571	0.9123	0.9224	0.8571	0.9132
	15	0.9450	0.9122	0.9111	0.8480	0.9112
	7-15	0.9603	0.9349	0.9187	0.9007	0.9298

TABLE VI: ABLATION STUDY OF THE SHIFT MODE ON THE TEST SET OF THE **KPIs** DATASET

Models	Shift Mode	mDSC↑				
		DN	NEP52	Normal	5/6Nx	AVG
VRWKV [26]	*Q-Shift*	0.9330	0.9232	0.9212	0.8738	0.9013
	Z-Shift	0.9344	0.9377	0.9255	0.8840	0.9108
DualSeg(Ours)	*Q-Shift*	0.9554	0.9220	0.9166	0.8902	0.9177
	Z-Shift	0.9603	0.9349	0.9187	0.9007	0.9298

achieved notable average mDSC improvements over all baselines but also maintained the lowest variance across evaluation metrics, reflecting its stability in intra-dataset learning. When extended to human datasets: it retained dominant, outperforming competitors in both private and public splits on the HuBMAP dataset; on the KPMP dataset, it even delivered highly competitive results via direct cross-dataset inference with no additional adaptation. This seamless transfer across species and datasets, paired with its stable, high-performance in intra-dataset training, underscores DualSeg's exceptional robustness for glomerular segmentation tasks.

C. Ablation Studies

Systematic ablation experiments on the KPIs dataset quantify the impact of key components in DualSeg, including the dual-stage encoder, dynamic propagation window in the Wave-Swin Block, and Z-Shift operator in the VRWKV Block. Results are summarized in Tables IV, V, and VI, with performance evaluated via mDSC across DN, NEP25, Normal, and 5/6Nx categories.

1) Effect of Dual-Stage Encoder: Table IV compares single-stage and dual-stage encoders. Single-stage variants, using only Wave-Swin Block, self-attention from SegFormer or VRWKV Block, achieve average mDSC of 90.36%, 90.99%, and 91.08% respectively. The dual-stage design Wave-VRWKV outperforms all single-stage counterparts with an average mDSC of 92.92%. We also conducted an in-depth exploration of the impact of fusion order on single-part architectures. Specifically, when the sequence of feature extraction modules was reversed, placing the attention mechanism at the beginning, the average mDSC values decreased significantly to 90% and 91%, respectively. These results underscore the critical role of the original module arrangement in single-part designs.

2) Effect of Dynamic Propagation Window: Table V compares fixed propagation window sizes from 7 to 15 and the adaptive 7–15 window range in DualSeg. The dynamic window design achieves the highest average mDSC of 92.98%, outperforming all fixed-window configurations. The advantage is most pronounced in 5/6Nx cases, where dynamic windows improve mDSC by 2.70% compared to the fixed window of 7. Even in Wave-MLP, adopting dynamic windows increases average mDSC by 0.15%.

3) Effect of Z-Shift Operator: Table VI compares the Z-Shift operator with the conventional Q-Shift. For DualSeg, Z-Shift improves average mDSC by 1.21%, with the largest gains in NEP25 (1.29%) and 5/6Nx (1.05%) cases. In VRWKV alone, replacing Q-Shift with Z-Shift also increases average mDSC by 0.95%.

D. Visualization Results

To intuitively verify the superior performance of DualSeg in glomerular segmentation, we present ualitative comparisons of segmentation results on both mice KPIs dataset and human HuBMAP dataset renal histopathological images in Fig. 6 and Fig. 7, respectively. These visualizations focus on challenging scenarios, including glomeruli with irregular shapes, ambiguous boundaries, and fragmented structures—key pathological features that test the model's ability to balance local texture discrimination and global context integration. To enhance clarity, we magnify critical regions (marked with arrows) to highlight differences in segmentation precision among methods.

For the mice KPIs dataset (Fig. 6), DualSeg accurately outlines renal glomeruli, avoiding over-segmentation of adjacent areas. In contrast, baselines such as SegFormer and

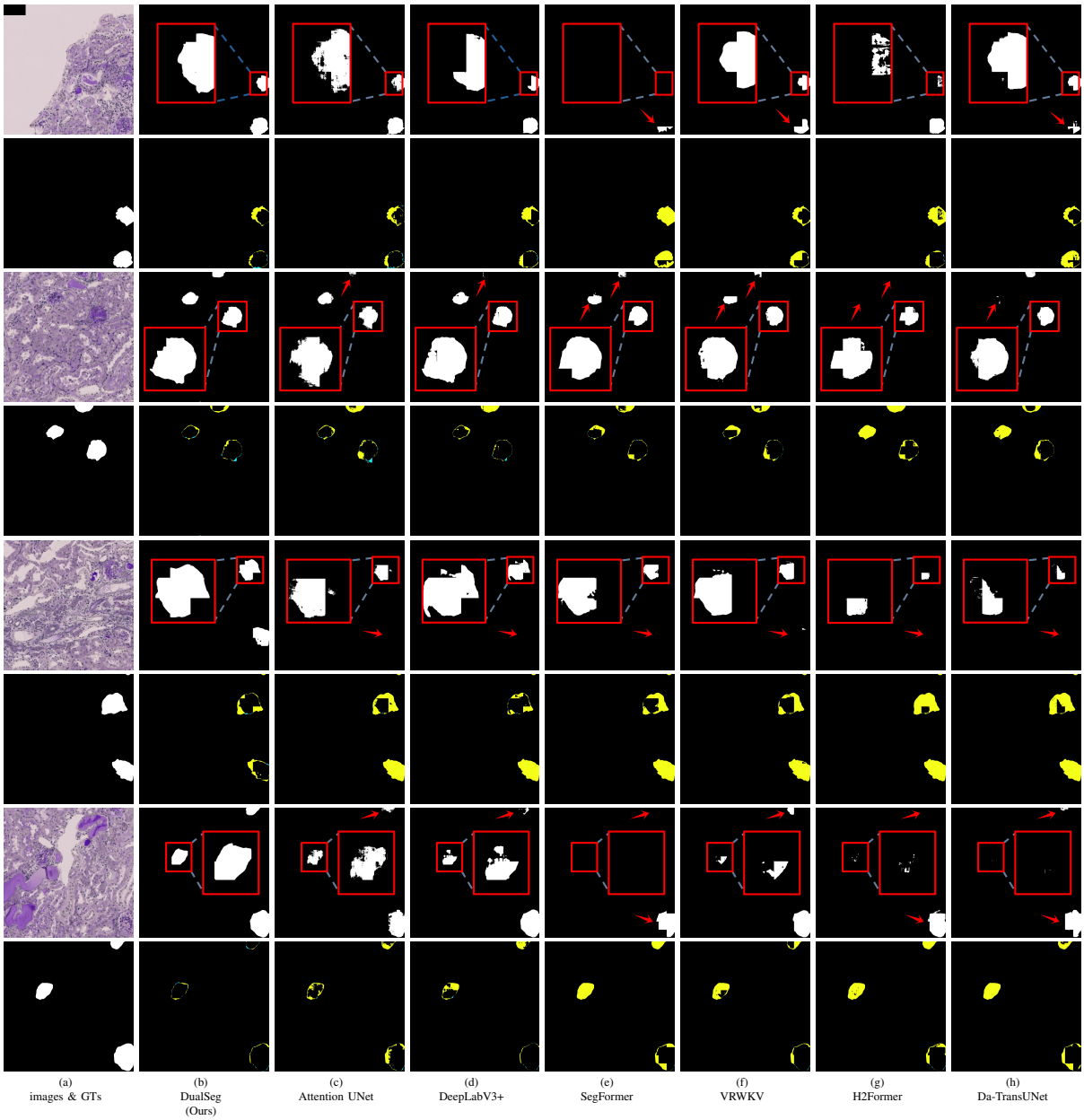

Fig. 6: The visual comparison results of mice glomeruli segmentation on the test set of KPIs dataset, where we emphasize the segmentation challenges by magnifying local details and illustrating them with arrows. It can be seen that our proposed model, **DualSeg**, demonstrates superior performance in segmenting glomeruli with high heterogeneity, particularly for targets exhibiting intricate morphological characteristics or ambiguous boundaries.

DA-TransUNet exhibit difficulties in delineating glomerular regions under abnormal conditions. In areas with subtle staining variations, DualSeg captures abnormal glomerular lesions, while VRWKV shows slight segmentation deficiencies. Notably, DualSeg reconstructs the continuity of damaged glomerular structures, whereas Attention UNet and H2Former fail to connect fragmented regions, resulting in disjointed masks.

For the human HuBMAP dataset (Fig. 7), DualSeg demonstrates superior performance in handling large-scale, complex histopathological backgrounds. Human glomeruli exhibit greater size and spatial variability, with ambiguity between sclerotic glomeruli and surrounding interstitial fibrosis. DualSeg clearly distinguishes these boundaries and captures residual glomerular textures, while DeepLabV3+ and Wave-MLP often blur such boundaries. For dispersed glomerular segments in the medullary region, DualSeg outperforms DA-TransUNet and VRWKV, which tend to overemphasize main structures, leading to overgrowth or incomplete segmentation.

For the human KPMP dataset shown in Fig. 8, our DualSeg model, pre-trained on mouse-derived data, exhibited remarkable stability when handling unseen cross-center and cross-species samples.Across these baseline models, varying degrees of outlier regions are evident in their segmentation

outputs: these regions manifest as spurious or misaligned segments that diverge from the true glomerular anatomical structures. Additionally, these baselines display pronounced uncertainty in both the internal and edge regions of segmented glomeruli: internal areas frequently contain noisy pseudo-segments, whereas edge contours appear blurred or discontinuous, failing to align with the GT boundaries. Notably, VM-UNet-V2 and H2Former exhibit prominent under-detection artifacts: partial glomerular regions—particularly those that are small or possess intricate morphological features—are entirely omitted from their segmentation results.In stark contrast, our DualSeg model's segmentation outputs demonstrate robust performance: its edge contours are sharply delineated to align closely with GT, while the internal regions of segmented glomeruli are consistent and largely free of substantial noise—these characteristics reflect greater confidence and stability in its predictive results.This visual comparison directly underscores DualSeg's superior adaptability to cross-center and cross-species variations, as it outperforms the baselines in both segmentation integrity and boundary precision.

VI. Discussion

As mentioned before, glomerular segmentation is challenged by three core issues: **local texture discriminability**, **spatial heterogeneity**, and **multi-scale mapping of morphological priors**. DualSeg's design directly targets these challenges, with its advantages validated through comparisons against baseline models, as elaborated below.

A. Superiority Over CNN-based Models

CNN models such as U-Net, Attention U-Net, and DeepLabV3+ rely on fixed receptive fields, struggling to adapt to diverse glomerular morphologies—limiting their ability to address **local texture discriminability** and **multi-scale mapping of morphological priors**. For instance, on the KPIs dataset, U-Net achieves an average mDSC of 88.59% with an HD95 of 93.34, while DeepLabV3+ improves to 91.47% mDSC but retains a high HD95 of 103.98, indicating persistent difficulties in capturing fine textures and multi-scale structural details.

DualSeg overcomes these limitations through three key innovations: the Wave-Swin Block's dynamic propagation window (adjusting from 7 to 15, Table V) adapts receptive fields to glomerular size, enhancing **local texture discriminability** in normal glomeruli and capturing broader contextual details in damaged ones; the lightweight decoder's multi-scale fusion integrates features across resolutions, addressing **multi-scale mapping**. These designs explain why DualSeg achieves 92.98% average mDSC and 66.85 HD95 on the KPIs dataset—outperforming CNN baselines by 1.51-4.39% in mDSC and reducing HD95 by 26.49-37.13. On the HuBMAP dataset, DualSeg's private mDSC of 89.79% surpasses Wave-MLP (86.08%) and DeepLabV3+ (85.50%), confirming its robustness in handling human glomerular variability across scales.

B. Superiority Over Transformer-based Models

Transformers like SegFormer and VRWKV excel in global context modeling but face trade-offs that hinder their performance on **local texture discriminability**. SegFormer, despite achieving 91.02% average mDSC on the KPIs dataset, struggles with fragmented 5/6Nx glomeruli (mDSC 81.66%) due to inadequate handling of irregular spatial distributions. VRWKV reduces computational cost but retains 87.38% mDSC in 5/6Nx cases, as Q-Shift operations lose edge details critical for local texture discrimination (Table VI).

The modified VRWKV Block in DualSeg resolves these issues through the Z-Shift operator, which preserves edge features to enhance **local texture discriminability**. This modification improves mDSC by 1.21% across all KPIs categories, with largest gains in NEP25 (1.29%) and 5/6Nx (1.05%) cases—critical for distinguishing sclerotic capsules from surrounding fibrosis. Additionally, VRWKV's linear attention efficiently models **spatial heterogeneity**. Combined with the Wave-Swin Block's local feature extraction, DualSeg outperforms SegFormer by 1.96% in average mDSC and VRWKV by 2.85% in 5/6Nx cases, balancing global context and local precision.

C. Superiority Over Hybrid Models

Hybrid models such as H2Former and DA-TransUNet aim to integrate CNNs and Transformers but fall short of synergistically addressing all three challenges due to overly rigid architectural designs. H2Former, for instance, achieves only 90.52% average mDSC on the KPIs dataset and 76.11% private mDSC on HuBMAP. This suboptimal performance stems from its superficial fusion of convolutional and transformer modules, which treats local and global information as independent streams without sequential refinement. Such parallel fusion fails to prioritize discriminative local textures (e.g., details of the glomerular basement membrane) before modeling global **spatial relationships**, resulting in suboptimal feature alignment. Similarly, DA-TransUNet embeds self-attention modules into a CNN encoder but retains the dominance of fixed convolution operations, thereby restricting its capacity to handle **spatial heterogeneity**. For UNETR and Swin UNETR, their fusion strategies introduce a semantic gap between the encoder and decoder. Both architectural designs either overemphasize local cues in the information flow or fail to highlight critical local cues prior to integrating global context, leading to occasional misalignment of small or weakly contrasted glomeruli—such as in 5/6Nx mouse models with severe atrophy—consistent with the issue of **morphological prior mapping** we focus on.

In contrast, DualSeg's dual-stage encoder addresses these limitations through a sequential *local-to-global* refinement mechanism. Ablation studies confirm this design's critical role: removing either stage reduces average mDSC by 1.20–1.96%, with largest drops in 5/6Nx cases (2.94% when omitting VRWKV, impacting spatial heterogeneity) and normal cases (1.21% when omitting Wave-Swin, impacting local texture). This structured integration explains why DualSeg outperforms H2Former by 2.46% in average mDSC on KPIs and by 13.68%

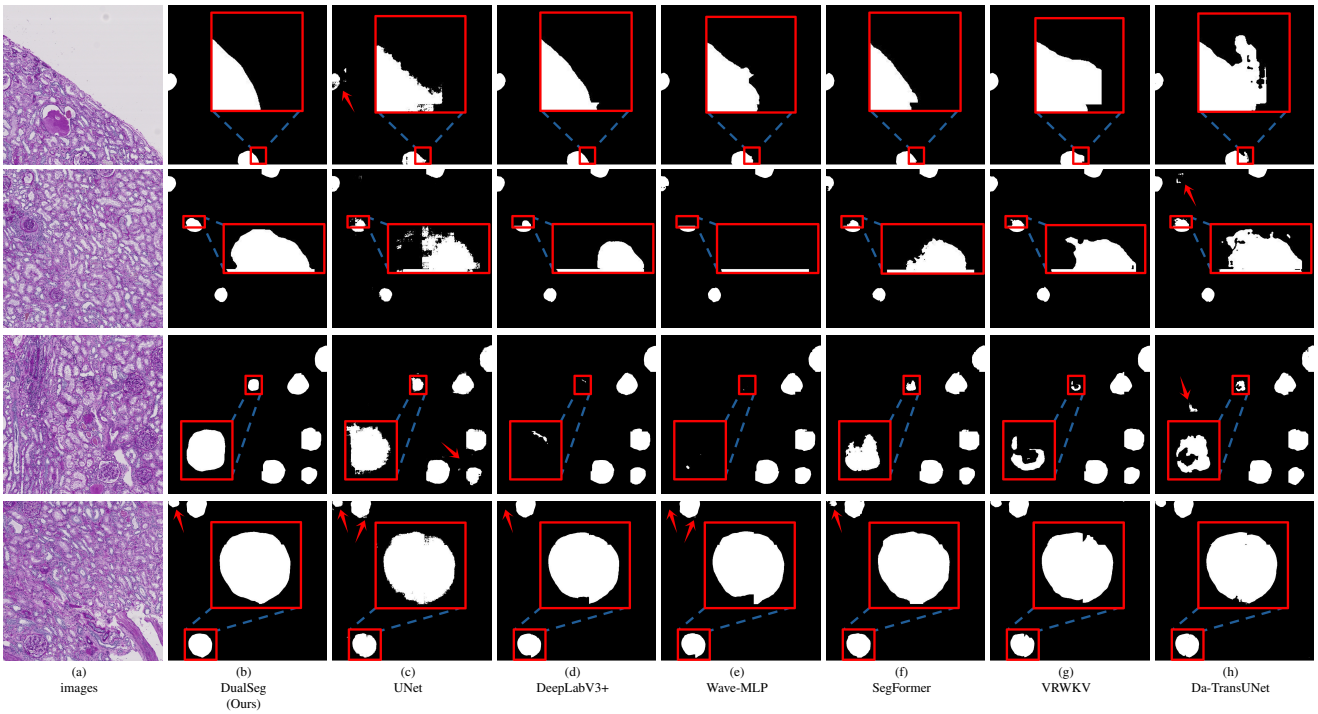

Fig. 7: Visual comparison results of human glomerulus segmentation on the test set of HuBMAP dataset are presented, where we emphasize the segmentation challenges by magnifying local details and illustrating them with arrows. Our proposed model, **DualSeg**, exhibits superior performance, as evidenced by its more comprehensive edge processing capabilities. This is complemented by its enhanced robustness against error-prone artifacts and its improved ability to restore the contours of human glomeruli.

on HuBMAP, and surpasses DA-TransUNet (89.35% private mDSC on HuBMAP) by 0.44%.

In summary, DualSeg's superiority arises from its unified framework that adapts to morphological variability via dynamic windows, preserves edge integrity through Z-Shift, and balances local-global features via dual-stage encoding. These design choices collectively address the limitations of existing models, establishing DualSeg as a robust tool for renal histology analysis.

D. Failure Cases Anlysis

It is also important to analyze the limitations of the model, and thus we show a failure case of fundus lesion segmentation in Fig. 9. We can see that the lesion regions indicated by the white rectangle are highly similar to the surrounding regions with low contrast, the segmentation performance will be decreased. And some lesions in fundus image are very small, only a few tens of pixels, and thus it is easy to be overlooked during downsampling, which makes it difficult to segment. The segmentation results show the misclassification, blurred boundary and the neglect of small lesions, as shown by the yellow arrows. The main reason for this failure case is that features are difficult to be retained and differentiated for this challenging region, resulting in poor segmentation performance for tiny lesions. First, a simple method is to increase the resolution of the input images and improve the perception ability of the model for small lesions, but this will have high requirements for GPU memory and computational cost, which may not be practical. Second, designing more powerful self-attention models could improve the segmentation results, more advanced training piplines like self-supervised learning and semi-supervised learning may enhance the feature diversity and improve the generalization ability of the model.

E. Clinical Relevance

DualSeg's consistent outperformance across mice and human datasets—with stable 5-fold cross-validation results (standard deviation less than 0.005 for private mDSC on HuBMAP)—validates its potential for clinical translation. By accurately segmenting diverse glomerular morphologies (from mild hypertrophy in DN to severe fragmentation in 5/6Nx), it supports quantitative analysis of pathological features like sclerosis extent and tubulointerstitial fibrosis, aiding standardized CKD diagnosis and progression monitoring. Notably, this automation could reduce the time required for expert manual annotation of WSIs, while minimizing inter-observer variability in glomerular counting and sizing—key advantages for multi-center clinical trials and routine diagnostic consistency.

F. Limitations

While DualSeg achieves superior performance in glomerular segmentation, several limitations remain. First, the VRWKV Block, despite its linear complexity, introduces non-negligible inference overhead when processing ultra-high-resolution WSIs with gigapixel dimensions, which may restrict its real-time applicability in clinical workflows. Second,

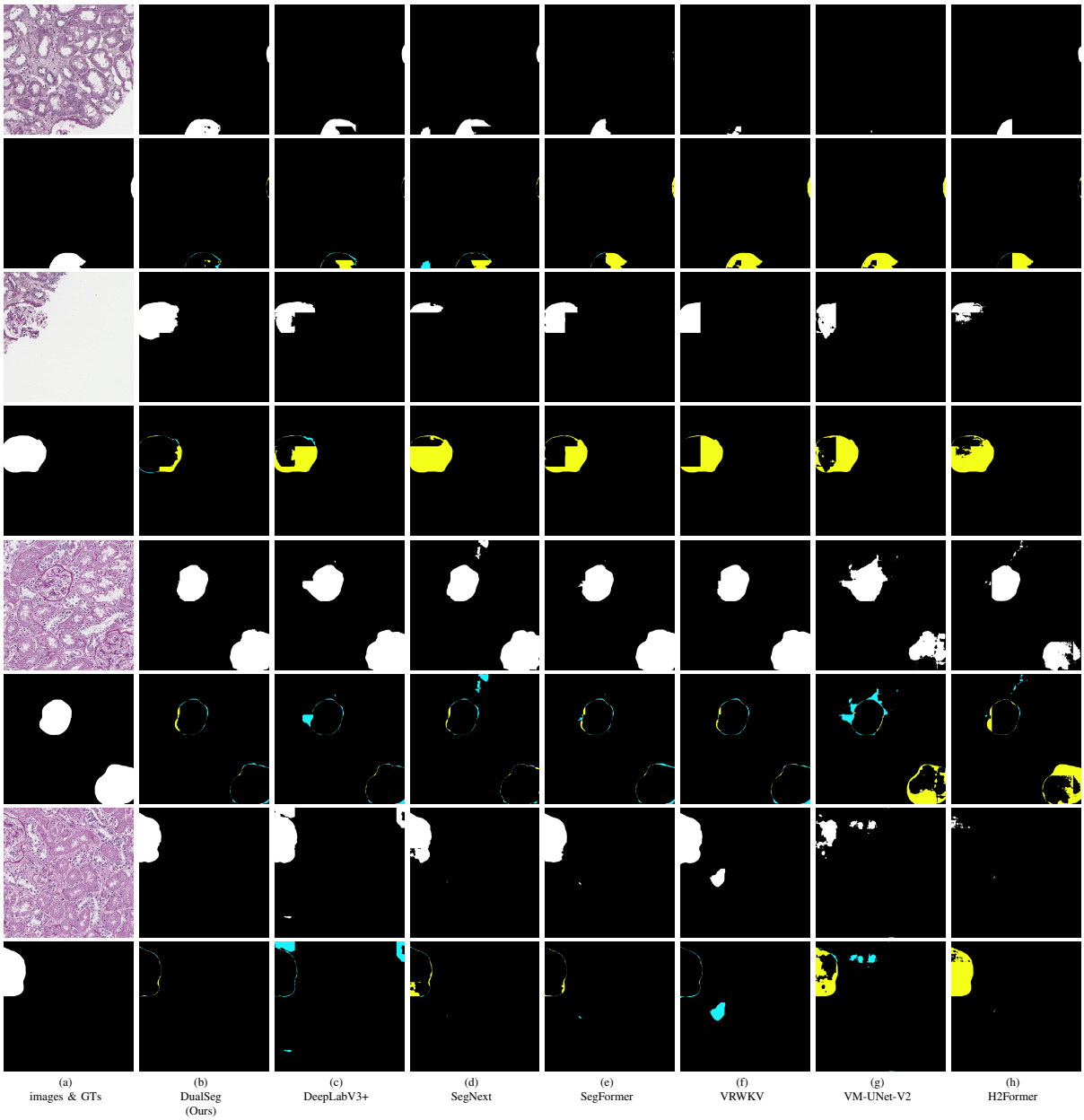

Fig. 8: The visual comparison results of human glomeruli segmentation on the held-out test set of the KPMP dataset are presented. It can be observed that our proposed model **DualSeg** exhibits unprecedented stability when handling cross-center and cross-species data.

the current framework is optimized for 2D histopathological slices, limiting its ability to capture 3D spatial relationships between glomeruli in volumetric renal tissues—a critical aspect for assessing disease progression. Third, although DualSeg demonstrates robustness across mice and human datasets, its generalization to rare pathological subtypes (e.g., focal segmental glomerulosclerosis with unusual morphological patterns) remains underexplored, as these cases were underrepresented in the evaluated datasets.

VII. Conclusion

In this paper, we present DualSeg, a unified dual-stage hybrid framework integrating CNN and VRWKV for robust glomerular segmentation in renal histopathology. Targeting the core challenges of local texture discriminability, spatial heterogeneity, and multi-scale morphological prior integration, DualSeg employs a two-stage encoder: Wave-Swin Blocks capture multi-directional local features via dynamic propagation windows, while VRWKV Blocks model long-range spatial dependencies through linear attention with a Z-Shift operator to preserve edge integrity, complemented by a lightweight decoder for multi-scale feature fusion. Evaluations on KPIs mice and HuBMAP human datasets demonstrate that DualSeg outperforms state-of-the-art methods, achieving superior mDSC and lowest HD95, with strong robustness across diverse pathological subtypes. Ablation studies validate the synergistic

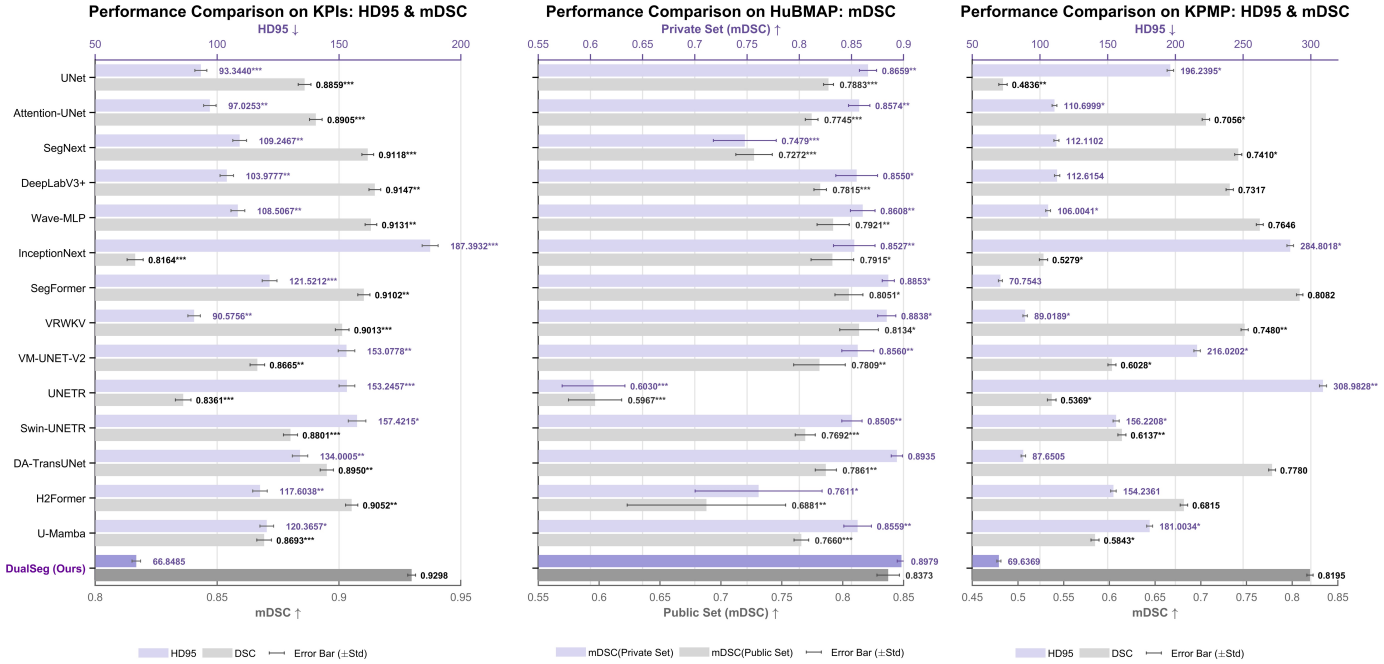

Fig. 9: Comparison of

contributions of its key components, confirming the dual-stage encoder's role in balancing local and global feature learning, dynamic windows' adaptability to morphological variations, and Z-Shift's mitigation of boundary loss. These findings establish DualSeg as a powerful tool for renal histology analysis, bridging local texture sensitivity and global context modeling, with potential as a versatile backbone for automated glomerular segmentation and extensions to other histopathological tasks requiring precise structural delineation.

[1] S. L. James, "A systematic analysis for the global burden of disease study 2017," *The Lancet*, vol. 392, no. 10159, pp. 1789–1858, 2018.

[2] T. K. Chen, D. H. Knicely, and M. E. Grams, "Chronic kidney disease diagnosis and management: a review," *Jama*, vol. 322, no. 13, pp. 1294–1304, 2019.

[3] C. Zoccali, R. Vanholder, Z. A. Massy, A. Ortiz, P. Sarafidis, F. W. Dekker, D. Fliser, D. Fouque, G. H. Heine, and K. J. Jager, "The systemic nature of ckd," *Nat. Rev. Nephrol.*, vol. 13, no. 6, pp. 344–358, 2017.

[4] J. M. Muñoz-Felix, B. Oujo, and J. M. Lopez-Novoa, "The role of endoglin in kidney fibrosis," *Expert Rev. Mol. Med.*, vol. 16, p. e18, 2014.

[5] A. Z. Rosenberg and J. B. Kopp, "Focal segmental glomerulosclerosis," *Clin. J. Am. Soc. Nephrol.*, vol. 12, no. 3, pp. 502–517, 2017.

[6] M. Aljabri, M. AlAmir, M. AlGhamdi, M. Abdel-Mottaleb, and F. Collado-Mesa, "Towards a better understanding of annotation tools for medical imaging: a survey," *Multimed. Tools Appl.*, vol. 81, no. 18, pp. 25 877–25 911, 2022.

[7] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, Eds. Cham: Springer International Publishing, 2015, pp. 234–241.

[8] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou, "Transunet: Transformers make strong encoders for medical image segmentation," *arXiv preprint arXiv:2102.04306*, 2021.

[9] A. Hatamizadeh, Y. Tang, V. Nath, D. Yang, A. Myronenko, B. Landman, H. R. Roth, and D. Xu, "Unetr: Transformers for 3d medical image segmentation," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2022, pp. 574–584.

[10] A. He, K. Wang, T. Li, C. Du, S. Xia, and H. Fu, "H2former: An efficient hierarchical hybrid transformer for medical image segmentation," *IEEE Trans. Med. Imaging*, vol. 42, no. 9, pp. 2763–2775, 2023.

[11] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "Segformer: Simple and efficient design for semantic segmentation with transformers," *Adv. neural inf. proces. syst.*, vol. 34, pp. 12 077–12 090, 2021.

[12] F. Allender, R. Allégre, C. Wemmert, and J.-M. Dischler, "Conditional image synthesis for improved segmentation of glomeruli in renal histopathological images," in *2022 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI)*, 2022, pp. 1–5.

[13] Y. Liu, "A hybrid cnn-transxnet approach for advanced glomerular segmentation in renal histology imaging," *Int. J. Comput. Int. Sys.*, vol. 17, no. 1, p. 126, 2024.

[14] B. Shickel, N. Lucarelli, A. S. Rao, D. Yun, K. C. Moon, S. S. Han, and P. Sarder, "Spatially aware transformer networks for contextual prediction of diabetic nephropathy progression from whole slide images," *medRxiv*, 2023.

[15] G. M. Dimitri, P. Andreini, S. Bonechi, M. Bianchini, A. Mecocci, F. Scarselli, A. Zacchi, G. Garosi, T. Marcuzzo, and S. A. Tripodi, "Deep learning approaches for the segmentation of glomeruli in kidney histopathological images," *Mathematics*, vol. 10, no. 11, 2022. [Online]. Available: https://www.mdpi.com/2227-7390/10/11/1934

[16] B. Ginley, K.-Y. Jen, A. Rosenberg, F. Yen, S. Jain, A. Fogo, and P. Sarder, "Neural network segmentation of interstitial fibrosis, tubular atrophy, and glomerulosclerosis in renal biopsies," *arXiv preprint arXiv:2002.12868*, 2020.

[17] K. M. Hosny, T. Magdy, N. A. Lashin, K. Apostolidis, and G. A. Papakostas, "Refined color texture classification using cnn and local binary pattern," *Mathematical Problems in Engineering*, vol. 2021, no. 1, p. 5567489, 2021.

[18] G. Kaur, M. Garg, S. Gupta, S. Juneja, J. Rashid, D. Gupta, A. Shah, and A. Shaikh, "Automatic identification of glomerular in whole-slide images using a modified unet model," *Diagnostics*, vol. 13, no. 19, p. 3152, 2023.

[19] H. Sun, J. Xu, and Y. Duan, "Paratranscnn: Parallelized transcnn encoder for medical image segmentation," *arXiv preprint arXiv:2401.15307*, 2024.

[20] J. Zhang, J. D. Lu, B. Chen, S. Pan, L. Jin, Y. Zheng, and M. Pan, "Vision transformer introduces a new vitality to the classification of renal pathology," *BMC nephrology*, vol. 25, no. 1, p. 337, 2024.

[21] G. V. Bharadwaj, Y. R. Sree, J. L. Varshita, and S. Chebrolu, "Ensemble model of u-net efficientnet-b3, u-net efficientnet b6, coat, segformer for segmenting functional tissue units in various human organs," in

2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT). IEEE, 2023, pp. 1–8.

[22] J. M. J. Valanarasu and V. M. Patel, "Unext: Mlp-based rapid medical image segmentation network," in *International conference on medical image computing and computer-assisted intervention*. Springer, 2022, pp. 23–33.

[23] I. O. Tolstikhin, N. Houlsby, A. Kolesnikov, L. Beyer, X. Zhai, T. Unterthiner, J. Yung, A. Steiner, D. Keysers, J. Uszkoreit *et al.*, "Mlp-mixer: An all-mlp architecture for vision," *Adv. neural inf. proces. syst.*, vol. 34, pp. 24 261–24 272, 2021.

[24] F. N. Saikia, Y. Iwahori, T. Suzuki, M. K. Bhuyan, A. Wang, and B. Kijsirikul, "Mlp-unet: glomerulus segmentation," *IEEE Access*, vol. 11, pp. 53 034–53 047, 2023.

[25] Y. Tang, K. Han, J. Guo, C. Xu, Y. Li, C. Xu, and Y. Wang, "An image patch is a wave: Phase-aware vision mlp," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 935–10 944.

[26] Y. Duan, W. Wang, Z. Chen, X. Zhu, L. Lu, T. Lu, Y. Qiao, H. Li, J. Dai, and W. Wang, "Vision-rwkv: Efficient and scalable visual perception with rwkv-like architectures," *arXiv preprint arXiv:2403.02308*, 2024.

[27] J. Ruan, J. Li, and S. Xiang, "Vm-unet: Vision mamba unet for medical image segmentation," *ACM Transactions on Multimedia Computing, Communications and Applications*, 2024.

[28] J. Ma, F. Li, and B. Wang, "U-mamba: Enhancing long-range dependency for biomedical image segmentation," *arXiv preprint arXiv:2401.04722*, 2024.

[29] W. Yu and X. Wang, "Mambaout: Do we really need mamba for vision?" in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 4484–4496.

[30] M.-H. Guo, C.-Z. Lu, Q. Hou, Z. Liu, M.-M. Cheng, and S.-M. Hu, "Segnext: Rethinking convolutional attention design for semantic segmentation," *Adv. neural inf. proces. syst.*, vol. 35, pp. 1140–1156, 2022.

[31] A. Fourcade and R. H. Khonsari, "Deep learning in medical image analysis: A third eye for doctors," *Journal of stomatology, oral and maxillofacial surgery*, vol. 120, no. 4, pp. 279–288, 2019.

[32] Q. Pu, Z. Xi, S. Yin, Z. Zhao, and L. Zhao, "Advantages of transformer and its application for medical image segmentation: a survey," *BioMedical engineering online*, vol. 23, no. 1, p. 14, 2024.

[33] F. Renard, S. Guedria, N. D. Palma, and N. Vuillerme, "Variability and reproducibility in deep learning for medical image segmentation," *Scientific Reports*, vol. 10, no. 1, p. 13724, 2020.

[34] G. E. Lees, R. E. Cianciolo, and F. J. Clubb Jr, "Renal biopsy and pathologic evaluation of glomerular disease," *Topics in companion animal medicine*, vol. 26, no. 3, pp. 143–153, 2011.

[35] T. Kato, R. Relator, H. Ngouv, Y. Hirohashi, O. Takaki, T. Kakimoto, and K. Okada, "Segmental hog: new descriptor for glomerulus detection in kidney microscopy image," *Bmc Bioinformatics*, vol. 16, pp. 1–16, 2015.

[36] P. Sarder, B. Ginley, and J. E. Tomaszewski, "Automated renal histopathology: digital extraction and quantification of renal pathology," in *Medical Imaging 2016: Digital Pathology*, vol. 9791. SPIE, 2016, pp. 112–123.

[37] D. Govind, B. Ginley, B. Lutnick, J. E. Tomaszewski, and P. Sarder, "Glomerular detection and segmentation from multimodal microscopy images using a butterworth band-pass filter," in *Medical Imaging 2018: Digital Pathology*, vol. 10581. SPIE, 2018, pp. 297–303.

[38] S. Sheehan, S. Mawe, R. E. Cianciolo, R. Korstanje, and J. M. Mahoney, "Detection and classification of novel renal histologic phenotypes using deep neural networks," *The American Journal of Pathology*, vol. 189, no. 9, pp. 1786–1796, 2019.

[39] Y. Kawazoe, K. Shimamoto, R. Yamaguchi, Y. Shintani-Domoto, H. Uozaki, M. Fukayama, and K. Ohe, "Faster r-cnn-based glomerular detection in multistained human whole slide images," *J. Imaging*, vol. 4, no. 7, p. 91, 2018.

[40] X. Han, G. Zhang, and X. Wang, "Glomerular microscopic image segmentation based on convolutional neural network," in *2019 Chinese Control Conference (CCC)*. IEEE, 2019, pp. 8343–8348.

[41] N. Altini, G. D. Cascarano, A. Brunetti, I. De Feudis, D. Buongiorno, M. Rossini, F. Pesce, L. Gesualdo, and V. Bevilacqua, "A deep learning instance segmentation approach for global glomerulosclerosis assessment in donor kidney biopsies," *Electronics*, vol. 9, no. 11, p. 1768, 2020.

[42] G. D. Cascarano, F. S. Debitonto, R. Lemma, A. Brunetti, D. Buongiorno, I. De Feudis, A. Guerriero, M. Rossini, F. Pesce, L. Gesualdo *et al.*, "An innovative neural network framework for glomerulus classification based on morphological and texture features evaluated in histological images of kidney biopsy," in *Intelligent Computing Methodologies: 15th International Conference, ICIC 2019, Nanchang, China, August 3–6, 2019, Proceedings, Part III 15*. Springer, 2019, pp. 727–738.

[43] J. N. Marsh, M. K. Matlock, S. Kudose, T.-C. Liu, T. S. Stappenbeck, J. P. Gaut, and S. J. Swamidass, "Deep learning global glomerulosclerosis in transplant kidney frozen sections," *IEEE Trans. Med. Imaging*, vol. 37, no. 12, pp. 2718–2728, 2018.

[44] G. Bueno, M. M. Fernandez-Carrobles, L. Gonzalez-Lopez, and O. Deniz, "Glomerulosclerosis identification in whole slide images using semantic segmentation," *Comput. Meth. Programs Biomed.*, vol. 184, p. 105273, 2020.

[45] P. Andreini, S. Bonechi, and G. M. Dimitri, "Enhancing glomeruli segmentation through cross-species pre-training," *Neurocomputing*, vol. 563, p. 126947, 2024.

[46] B. Ginley, K.-Y. Jen, A. Rosenberg, F. Yen, S. Jain, A. Fogo, and P. Sarder, "Neural network segmentation of interstitial fibrosis, tubular atrophy, and glomerulosclerosis in renal biopsies," *arXiv preprint arXiv:2002.12868*, 2020.

[47] F. Yang, Q. He, Y. Wang, S. Zeng, Y. Xu, J. Ye, Y. He, T. Guan, Z. Wang, and J. Li, "Unsupervised stain augmentation enhanced glomerular instance segmentation on pathology images," *International Journal of Computer Assisted Radiology and Surgery*, vol. 20, no. 2, pp. 225–236, 2025.

[48] Q. He, S. Zeng, S. Ge, Y. Wang, J. Ye, Y. He, T. Guan, Z. Wang, and J. Li, "Identifying and matching 12-level multistained glomeruli via deep learning for diagnosis of glomerular diseases," *International Journal of Imaging Systems and Technology*, vol. 34, no. 2, p. e23032, 2024.

[49] M. Zhang, Y. Yu, S. Jin, L. Gu, T. Ling, and X. Tao, "Vm-unet-v2: rethinking vision mamba unet for medical image segmentation," in *International symposium on bioinformatics research and applications*. Springer, 2024, pp. 335–346.

[50] Y. Gu, R. Ruan, Y. Yan, J. Zhao, W. Sheng, L. Liang, and B. Huang, "Glomerulus semantic segmentation using ensemble of deep learning models," *Arab. J. Sci. Eng.*, vol. 47, no. 11, pp. 14 013–14 024, 2022.

[51] M. Hermsen, T. de Bel, M. Den Boer, E. J. Steenbergen, J. Kers, S. Florquin, J. J. Roelofs, M. D. Stegall, M. P. Alexander, B. H. Smith *et al.*, "Deep learning–based histopathologic assessment of kidney tissue," *J. Am. Soc. Nephrol.*, vol. 30, no. 10, pp. 1968–1979, 2019.

[52] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz *et al.*, "Attention u-net: Learning where to look for the pancreas," *arXiv preprint arXiv:1804.03999*, 2018.

[53] W. Yu, P. Zhou, S. Yan, and X. Wang, "Inceptionnext: When inception meets convnext," in *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, 2024, pp. 5672–5683.

[54] A. Hatamizadeh, V. Nath, Y. Tang, D. Yang, H. R. Roth, and D. Xu, "Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images," in *International MICCAI brainlesion workshop*. Springer, 2021, pp. 272–284.

[55] G. Sun, Y. Pan, W. Kong, Z. Xu, J. Ma, T. Racharak, L.-M. Nguyen, and J. Xin, "Da-transunet: integrating spatial and channel dual attention with transformer u-net for medical image segmentation," *Front. Bioeng. Biotechnol.*, vol. 12, p. 1398237, 2024.

[56] Z. Geng, M.-H. Guo, H. Chen, X. Li, K. Wei, and Z. Lin, "Is attention better than matrix decomposition?" *arXiv preprint arXiv:2109.04553*, 2021.

[57] Y. Tang, Y. He, V. Nath, P. Guo, R. Deng, T. Yao, Q. Liu, C. Cui, M. Yin, Z. Xu *et al.*, "Holohisto: end-to-end gigapixel wsi segmentation with 4k resolution sequential tokenization," *arXiv preprint arXiv:2407.03307*, 2024.

[58] H. Addison, L. Andy, S. Bud, T. Eddie, K. Jarek, B. Katy, G. Leah, N. Marcos, C. Phil, H. Richard, W. Rick, and J. Yingnan. (2021) Hubmap - hacking the kidney. [Online]. Available: https://kaggle.com/competitions/hubmap-kidney-segmentation

[59] KPMP Consortium. (2021) Kidney precision medicine project. National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK). Accessed: 2025-12-11. [Online]. Available: https://atlas.kpmp.org/repository

[60] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.

[61] T. Dozat, "Incorporating nesterov momentum into adam," in *Proceedings of 4th International Conference on Learning Representations (ICLR)*, 2016, pp. 1–4.

[62] I. Loshchilov and F. Hutter, "Sgdr: Stochastic gradient descent with warm restarts," *arXiv preprint arXiv:1608.03983*, 2016.
[63] M. J. Cardoso, W. Li, R. Brown, N. Ma, E. Kerfoot, Y. Wang, B. Murrey, A. Myronenko, C. Zhao, D. Yang *et al.*, "Monai: An open-source framework for deep learning in healthcare," *arXiv preprint arXiv:2211.02701*, 2022.