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) requires accurate histopathological analysis of glomeruli, but manual segmentation in Whole Slide Images (WSIs) is labor-intensive and error-prone. Existing methods such as Convolutional Neural Networks (CNNs) suffer from limited adaptability to global context, while Vision Transformers (ViTs) incur high computational costs, failing to simultaneously address local texture discriminability, spatial heterogeneity, and multi-scale morphological prior mapping in glomerular segmentation. To tackle these challenges, we propose DualSeg, a unified dual-stage hybrid framework integrating CNN and Vision Recurrent Weighted Key Value (VRWKV). The framework employs a two-stage encoder: Wave-Swin Blocks with dynamic propagation windows for multi-directional local feature extraction, and VRWKV Blocks with a Z-Shift operator to model long-range dependencies efficiently via linear attention while preserving edge integrity. A lightweight decoder with Non-negative Matrix Factorization (NMF) further enhances multi-scale feature fusion. Evaluated on three datasets (murine KPIs, human HuBMAP, and human KPMP), DualSeg outperforms state-of-the-art (SOTA) models, achieving superior average mDSC (92.98% on KPIs, 89.79% private mDSC on HuBMAP, 81.95% on KPMP) and the lowest HD95. Notably, it demonstrates robust cross-species and cross-center generalizability through direct inference on human datasets without retraining. DualSeg effectively bridges local texture sensitivity with global context modeling, offering a novel methodological approach for renal histology analysis. 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)**

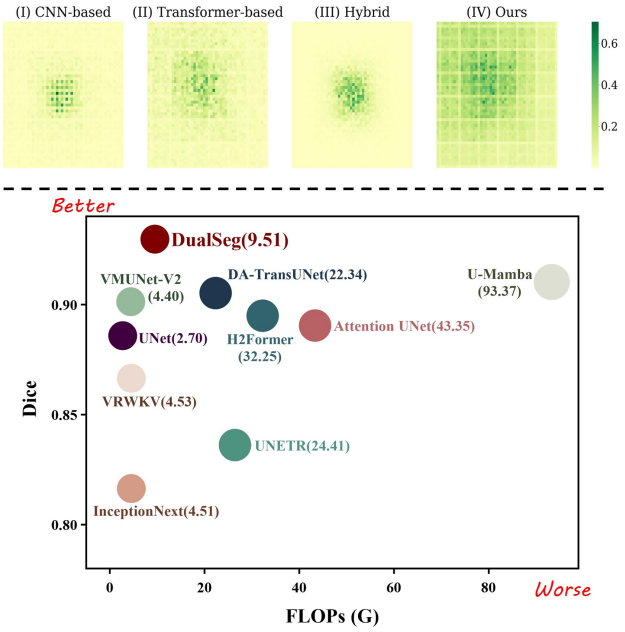

Fig. 1: **Top:** Comparison of *Effective Receptive Fields* (ERF). CNNs (I) show restricted local focus; Transformers (II) display dispersed global patterns; and current Hybrids (III) demonstrate subtimal coverage, failing to form a full context. Our method (IV) overcomes these limitations to generate a dense, global ERF. **Bottom:** Trade-off analysis. DualSeg occupies the optimal top-left position, delivering SOTA accuracy with significantly lower computational complexity (FLOPs) compared to existing methods.

I. Introduction

Chronic Kidney Disease (CKD) affects over % of the global population, precipitating systemic complications such as hypertension [1]–[3]. Since glomerulosclerosis and renal interstitial fibrosis are primary histological manifestations [4], [5], the pathological assessment of glomeruli is pivotal for therapeutic intervention. However, manual analysis of Whole Slide Images (WSIs) is labor-intensive and prone to variability [6]. Consequently, automated computational tools with high precision are essential to resolve these diagnostic bottlenecks.

Accurate glomerular segmentation is impeded by the intricate interplay between local texture and global structure. Three fundamental technical challenges characterize this task: (1) **Local texture discriminability**: Glomeruli display diverse substructures and high intra-class variability, complicating boundary detection [12], [13]. (2) **Spatial heterogeneity**: The irregular distribution of glomeruli demands global contextual modeling to ensure structural continuity [14]. (3) **Multi-scale mapping of morphological priors**: Effective perception

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).

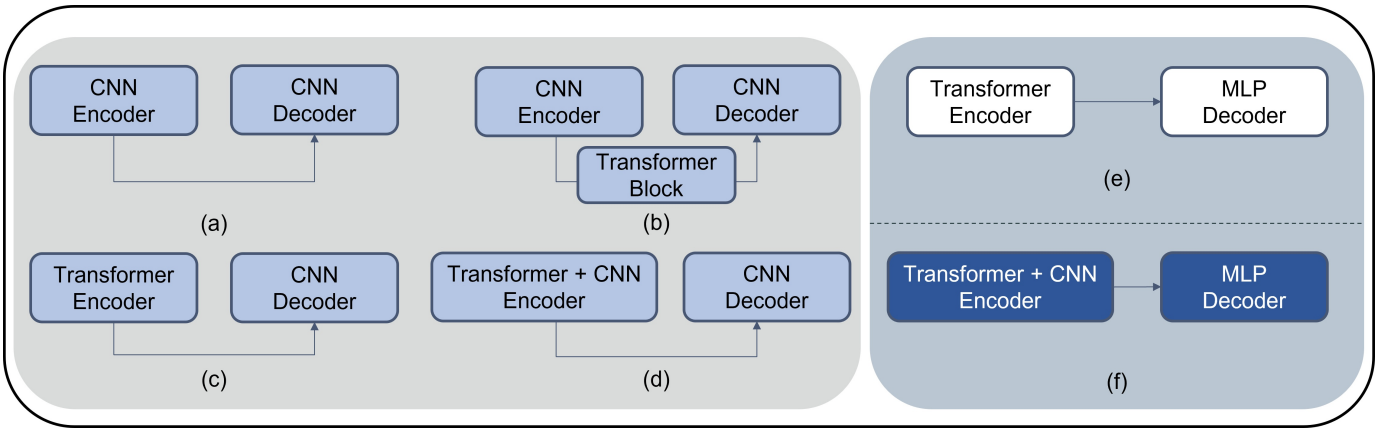

Fig. 2: Comparison of segmentation architectures. (a–d) represent **U-shaped** designs: (a) Pure CNN (UNet [7]); (b-d) Hybrid Cascades (TransUNet [8], UNETR [9] and H2Former [10]). (e–f) represent **Pyramid-shaped** designs: (e) Pure Transformer (SegFormer [11]); (f) Our **DualSeg**, a hierarchical hybrid encoder with multi-scale fusion.

necessitates integrating receptive fields at varying resolutions to decode both fine-grained details and macroscopic priors simultaneously [14]–[16].

Deep learning has revolutionized digital pathology. Convolutional Neural Networks (CNNs) effectively address local texture challenges via inductive biases and weight sharing [8], [17]. For instance, Kaur et al. [18] demonstrated the efficacy of U-Net in modeling sub-pixel intensity transitions. However, CNNs inherently struggle with long-range dependencies due to their limited Effective Receptive Field (ERF) [8], [10], [19], as visualized in Fig. 1(I), which constrains their ability to resolve complex, disjointed morphologies.

Conversely, Vision Transformers (ViTs) tackle spatial heterogeneity by leveraging self-attention for global context modeling [20], [21] (Fig. 1(II)). While ViTs typically outperform CNNs in maintaining structural continuity, the quadratic computational complexity of self-attention incurs high FLOPs, limiting their scalability in high-resolution histology [10]. To mitigate this, computationally efficient alternatives like WaveMLP [22] and VRWKV [23] (employing linear-time recurrent kernels) have emerged. Similarly, VM-UNet [24] introduced State Space Models (SSM) to balance computational cost and segmentation performance.

Beyond standalone architectures, synergistic hybrid frameworks have been explored. As illustrated in Fig. 2(a-d), most prior hybrids (e.g., TransUNet [8], H2Former [10], and U-Mamba [25]) adopt U-shaped paradigms. However, these designs often exhibit limitations: TransUNet compromises multi-scale feature capture; H2Former's shallow integration underutilizes ViTs; and their ERFs remain suboptimal (Fig. 1(III)). Distinct from U-shaped models, pyramid architectures like SegFormer [11] (Fig. 2(e)) and SegNeXt [26] address the third challenge via feature fusion but rely on unidirectional extraction, weakening robustness against morphological variations. This raises a fundamental question: *Is it possible to integrate local and global features within a unified multi-scale framework to tackle all three challenges without incurring excessive computational complexity?*

To address this, we propose **DualSeg**, a novel hybrid framework that synergizes Wave Vision and VRWKV within a pyramid structure (Fig. 2(f)). Our architecture features a dual-stage encoder where early-stage Wave-Swin blocks execute hierarchical local feature extraction to enhance texture discriminability, followed by VRWKV blocks that model long-range dependencies via linear attention to resolve spatial heterogeneity. This design effectively amalgamates the texture sensitivity of CNNs, the generalization of MLPs, and the scalability of VRWKV. By bridging local and global processing, DualSeg achieves robust multi-scale mapping of morphological priors. As shown in Fig. 1(IV) and Bottom, our method attains a global, clean ERF and SOTA segmentation performance. Notably, it achieves this with significantly reduced computational overhead (9.51 G FLOPs), offering an optimal accuracy-complexity trade-off superior to resource-intensive baselines.

The main contributions of this study are:

- We propose **DualSeg**, a novel pyramid hybrid architecture integrating convolutional local extraction with linear-complexity bidirectional attention to address texture, heterogeneity, and multi-scale challenges.
- We design a plug-and-play attention module based on wave vision to capture multi-directional semantic features, enhancing representation learning.
- We validate robust cross-species and cross-center generalizability by training on mouse data and performing direct inference on a held-out human dataset, demonstrating clinical viability.
- Our method achieves SOTA performance across three 2D glomerulus segmentation datasets, outperforming existing approaches in both accuracy and robustness.

II. Related Work

A. Advancements in Glomerular Segmentation

Automated glomerular segmentation serves as a cornerstone for quantitative renal pathology analysis [27]. While early methodologies employed handcrafted features and traditional machine learning for boundary detection [28], [29], these approaches were inherently limited by their susceptibility to staining inconsistencies and poor generalizability [30].

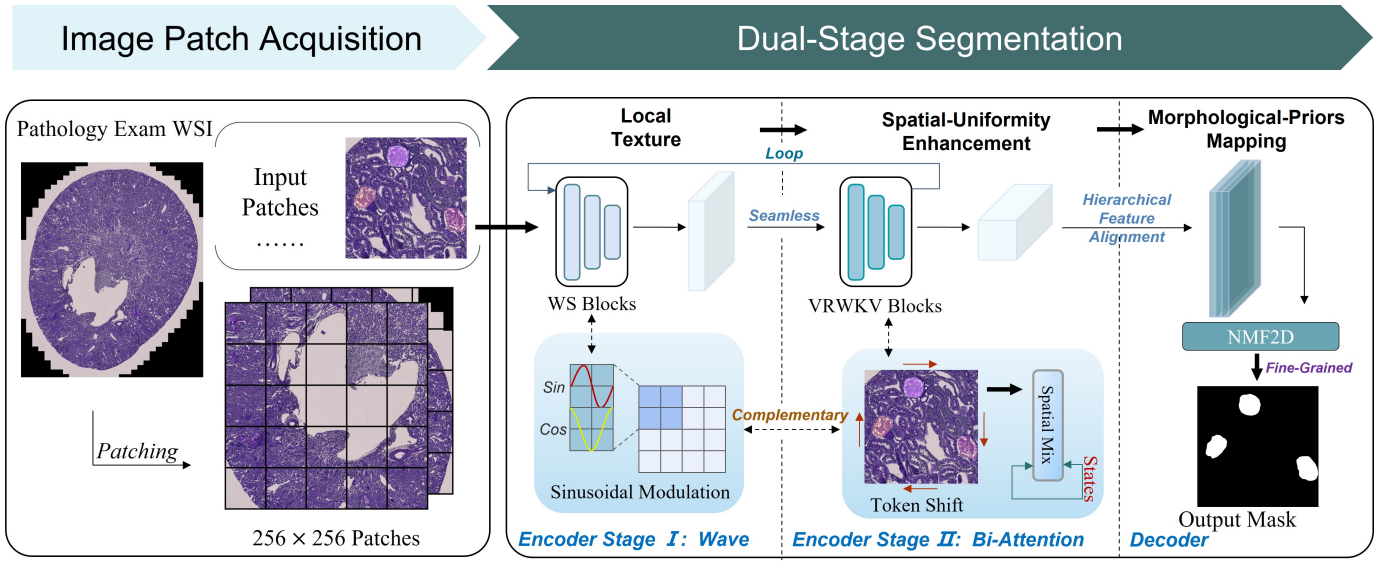

Fig. 3: Overview of the proposed **DualSeg**. The original image is fed into the Dual-Stage encoder composed of Wave-Swin Block and VRWKV Block to extract local texture and patial heterogeneity. Next, the features are merged and refined by the lightweight *Hamburger* decoder to obtain the predicted glomerular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

The integration of Convolutional Neural Networks (CNNs) marked a definitive shift in the field, establishing the standard for decoding intricate glomerular morphologies from high-resolution Whole Slide Images (WSIs). Seminal works by Marsh et al. [31] and Bueno et al. [32] demonstrated that CNN-based pixel-wise segmentation could achieve pathologist-level accuracy, providing a robust foundation for diagnosing conditions such as diabetic nephropathy. Furthermore, addressing the bottleneck of annotated data scarcity, Andreini et al. [33] validated the clinical utility of cross-species transfer learning. Their findings indicate that pre-training on murine models significantly enhances generalization capabilities on human datasets, a paradigm essential for bridging preclinical research with clinical application.

B. Technological Evolution of Glomerular Segmentation Architectures

The architectural landscape of glomerular segmentation has evolved from pure CNNs to advanced hybrid paradigms. Foundational U-Net variants and the DeepLab series established the baseline for capturing local morphological details [15], [34], yet they inherently struggle to model long-range spatial dependencies due to restricted receptive fields. While MLP-based successors sought to simplify architectural dependencies [35], [36], they often compromise feature representation in complex pathological scenarios.

To bridge the gap between local precision and global context, Transformer-based methods like SegFormer [11] introduced self-attention mechanisms, the effectiveness of which has been corroborated in glomerular segmentation studies [37]–[39]. However, standard self-attention faces scalability constraints when processing the extensive spatial dimensions characteristic of Whole Slide Images (WSIs). This limitation has catalyzed the emergence of alternative global modeling paradigms—notably VRWKV [23] and U-Mamba [25], [40]—which utilize recurrent formulations or SSM to achieve effective global receptivity on high-resolution inputs. Although existing hybrid frameworks attempt to synergize the strengths of convolution and attention mechanism [13], achieving seamless multi-scale integration that fully preserves structural continuity remains an ongoing challenge.

Despite these advances, existing frameworks continue to struggle with precise boundary delineation in complex scenarios, such as distinguishing sclerotic capsules from surrounding fibrosis, and lack robust generalizability across diverse pathological conditions. While studies such as [41], [42] and [21] have attempted to mitigate these issues through ensemble modeling, this approach inevitably incurs high computational overhead, limiting its feasibility for real-time clinical analysis.

III. Methodology

In this section, we present a comprehensive overview of our proposed architecture, DualSeg, designed for the precise extraction and synthesis of multi-scale textural features from glomeruli in renal histopathological images. The architecture, illustrated in 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.

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 [43]	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 [26]	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+ [34]	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 [22]	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 [44]	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 [23]	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 [40]	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 [45]	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 [46]	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 [25]	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**

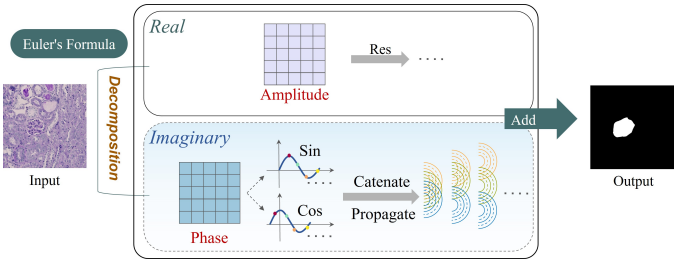

Fig. 4: Illustration of the Token-Mixing mechanism within the **Wave-Swin** Block. The input is initially decomposed into a *real* component and an *imaginary* component via Euler's formula. The modulated phase undergoes two distinct cosine transformations and window-based propagation, and is subsequently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

A. Wave-Swin Block

To address the dual challenges of fine-grained texture discrimination and global contextual modeling in renal glomerular segmentation, we propose a hybrid Dual-Stage encoder that seamlessly integrates an improved CNN-based first stage with a VRWKV-based second stage. Specifically, the initial stage employs our proposed Wave-Swin (WS) Block to capture local glomerular boundaries. By introducing a dynamic wave-based representation shown in Fig. 4, the WS block overcomes the static receptive field limitations of traditional wave methods, enhancing local texture discriminability. Subsequently, the VRWKV module is utilized to model long-range dependencies, effectively handling the global irregular spatial arrangements characteristic of renal tubulointerstitial fibrosis. This synergistic design robustly bridges the gap between texture fidelity and contextual awareness in complex histopathological scenarios.

1) Wave Formulation: Let the input feature map be denoted as $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n]$, where each $\mathbf{x}_j$ represents a token. Instead of standard linear projections, we map these tokens into a complex wave representation to capture both semantic intensity and spatial structural priors.

We define the complex wave form $\mathbf{z}_j$ for the j-th token as:

$$\mathbf{z}_j = \mathcal{A}(\mathbf{x}_j) \odot \exp(i \cdot \mathcal{P}(\mathbf{x}_j)), \tag{1}$$

where $\mathcal{A}(\cdot)$ and $\mathcal{P}(\cdot)$ denote learnable linear transformations that project the input to amplitude and phase terms, respectively. The operator $\odot$ represents element-wise multiplication. Conceptually, the amplitude $\mathcal{A}(\mathbf{x}_j) \in \mathbb{R}^+$ encodes the semantic intensity (e.g., texture presence), while the phase $\mathcal{P}(\mathbf{x}_j) \in [0, 2\pi)$ encodes the token's relative spatial position within the glomerular structure.

The phase term is strategically designed to capture zonal microanatomical variations. Specifically, glomeruli in cortical regions exhibit more compact clustering compared to those in the medulla, where fibrotic remodeling disrupts continuity. By assigning distinctive angular values based on proximity to anatomical landmarks (e.g., the renal capsule), the phase effectively models these spatial relationships.

Utilizing Euler's formula, the complex wave $\mathbf{z}_j$ is implicitly decomposed into orthogonal real and imaginary components. These components are then aggregated via a dynamic token-mixing operation:

$$\mathbf{y}_j = \sum_{k \in \Omega_j} \left(\mathrm{Re}(\mathbf{z}_k)\mathbf{W}_{Re} + \mathrm{Im}(\mathbf{z}_k)\mathbf{W}_{Im} \right), \tag{2}$$

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 [43]	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 [26]	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+ [34]	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 [22]	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 [44]	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 [23]	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 [40]	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 [45]	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 [46]	**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 [25]	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**	$\mathbf{0.8979 \pm 0.0040}$	$\mathbf{0.8373 \pm 0.0092}$

where $\mathbf{y}_j$ is the output token, Ω_j denotes the dynamic propagation window centered at j, and $\mathbf{W}_{Re}, \mathbf{W}_{Im}$ are learnable weights governing the fusion of spatial components.

2) Dynamic Swin Mechanism: To overcome the static receptive field limitation of prior wave-propagation windows, we propose a dynamic Swin mechanism that adaptively reshapes the wavefront Ω_j. We define a candidate set of odd-sized windows $S = \{7, 11, 15, \dots\}$. The selection of these anchors is informed by both domain-validated baselines and dataset-specific anatomical statistics.

First, previous studies on Wave-MLP have empirically demonstrated that windows smaller than 7 lack the generality necessary to capture spatial dependencies in medical images [22]; meanwhile, anchor sizes 7 and 11 align with kernel sizes employed in SOTA encoders like SegNeXt [26]. Second, the average glomerular bounding box in our murine dataset measures approximately 154px [33]. After the $4\times$ and $8\times$ downsampling stages, this dimension reduces to roughly 38px and 19px, respectively. Accordingly, selecting a maximum window of 15 (instead of 21) prevents the network from integrating extraneous background noise while ensuring full coverage of the target glomerular structure.

For each token, a mapping function $M : S \rightarrow \{1, 2, \dots, |S|\}$ is learned to select the optimal window size from S, ensuring the faithful encoding of diverse spatial patterns. This dynamic strategy substantially improves feature extraction by comprehensively balancing glomerular positional semantics with a broader, yet precise, contextual framework.

B. VRWKV Block

Following the local feature extraction, the VRWKV block is employed to establish non-local constraints. This stage specifically addresses the challenge of spatial heterogeneity by capturing long-range dependencies efficiently. The input feature map first undergoes downsampling to facilitate global context modeling at a reduced resolution.

1) Z-Shift and Spatial Mixing: To effectively model global interactions while preserving boundary integrity, we introduce a modified **Z-Shift** mechanism integrated with the Spatial Mixing block.

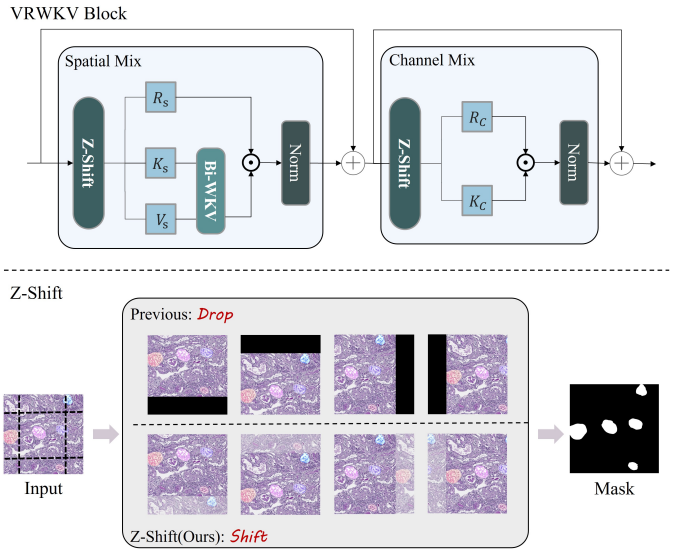

Fig. 5: Structure of the improved VRWKV block. The black margins in the lower panel illustrate the prior approach where pixels shifted beyond the boundary are simply dropped. In contrast, our **Z-Shift** wraps these pixels to the opposite edge, filling the empty regions and eliminating information loss.

Crucially, unlike standard Vision Transformers that immediately flatten inputs, we apply the Z-Shift operator directly on the 2D feature maps before serialization. While the original Q-Shift in VRWKV discards pixels shifted beyond boundaries (zero-padding), our Z-Shift cyclically wraps these pixels to the opposite edge. This *zero-loss* shifting strategy ensures that spatial continuity is maintained, which is vital for delineating the complete contours of glomeruli.

After the spatial shift, the feature maps are flattened into token sequences and projected. The generation of the receptor, key, and value matrices is formulated as:

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 [43]	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 [26]	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+ [34]	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 [22]	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 [44]	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 [23]	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 [40]	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 [45]	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 [46]	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 [25]	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**

$$\mathbf{N}_s = \text{Linear}_N(\text{Flatten}(\text{Z-Shift}(\mathbf{X}))), \quad N \in \{R, K, V\}, \tag{3}$$

where $\mathbf{X}$ denotes the input 2D feature map, and Linear_N represents the learnable projection weights. The flattened tokens then undergo the linear-complexity bidirectional attention aggregation:

$$\mathbf{S} = \sigma(\mathbf{R}_s) \odot \text{Bi-WKV}(\mathbf{K}_s, \mathbf{V}_s), \tag{4}$$

where $\sigma(\cdot)$ is the sigmoid activation, $\odot$ denotes element-wise multiplication, and Bi-WKV is the time-mixing operator that aggregates global context with linear complexity $O(L)$, efficiently capturing pairwise affinities between distant glomerular candidates.

2) Channel Mixing: Following spatial aggregation, the features undergo Channel Mixing to enable inter-channel communication. This module mirrors the gating mechanism of the spatial stage but focuses on feature refinement within the channel dimension. The transition is expressed as:

$$\mathbf{O}_c = \sigma(\mathbf{R}_c) \odot (\text{SqReLU}(\mathbf{K}_c) \cdot \mathbf{W}_v), \tag{5}$$

where $\mathbf{R}_c$ and $\mathbf{K}_c$ are derived from the spatially mixed features via linear projections, and SqReLU denotes the squared ReLU activation. The final output tokens are reshaped back into 2D spatial maps, fusing the global context captured by VRWKV with the local details from the previous Wave-Swin stage. This synergistic design ensures robust segmentation performance across variable glomerular morphologies.

C. The Lightweight Decoder

To balance computational efficiency with segmentation precision, we adopt a lightweight decoding strategy. Following established practices [26], we selectively exclude the high-resolution features from the first CNN stage, as they contain excessive low-level noise that can degrade semantic consistency.

Consequently, we employ the *HamDecoder* [47] to fuse the feature maps from the subsequent three stages. As a pivotal component of our multi-scale interaction framework, this decoder utilizes Non-negative Matrix Factorization (NMF) to address the challenge of morphological prior integration. By modeling feature fusion as a matrix decomposition problem, the decoder effectively separates coherent semantic signals from background noise:

$$\mathbf{M} = \mathbf{D} \times \mathbf{C} + \mathbf{N}, \tag{6}$$

where $\mathbf{M}$ denotes the aggregated feature mask, while $\mathbf{D}$, $\mathbf{C}$, and $\mathbf{N}$ represent the learned *Dictionary*, *Codes*, and *Noise*, respectively. By reconstructing the mask using only the dictionary and codes ($\mathbf{D} \times \mathbf{C}$), the model filters out the noise component $\mathbf{N}$, yielding a refined and purer segmentation output.

This NMF-based decoding mechanism complements our Dual-Stage Hybrid encoder by explicitly enforcing low-rank constraints on the fused features, thereby enhancing boundary delineation without significant computational overhead.

IV. EXPERIMENTS

In this section, we evaluate the proposed method on three glomerular segmentation datasets. We first introduce the benchmark datasets, implementation details, and evaluation

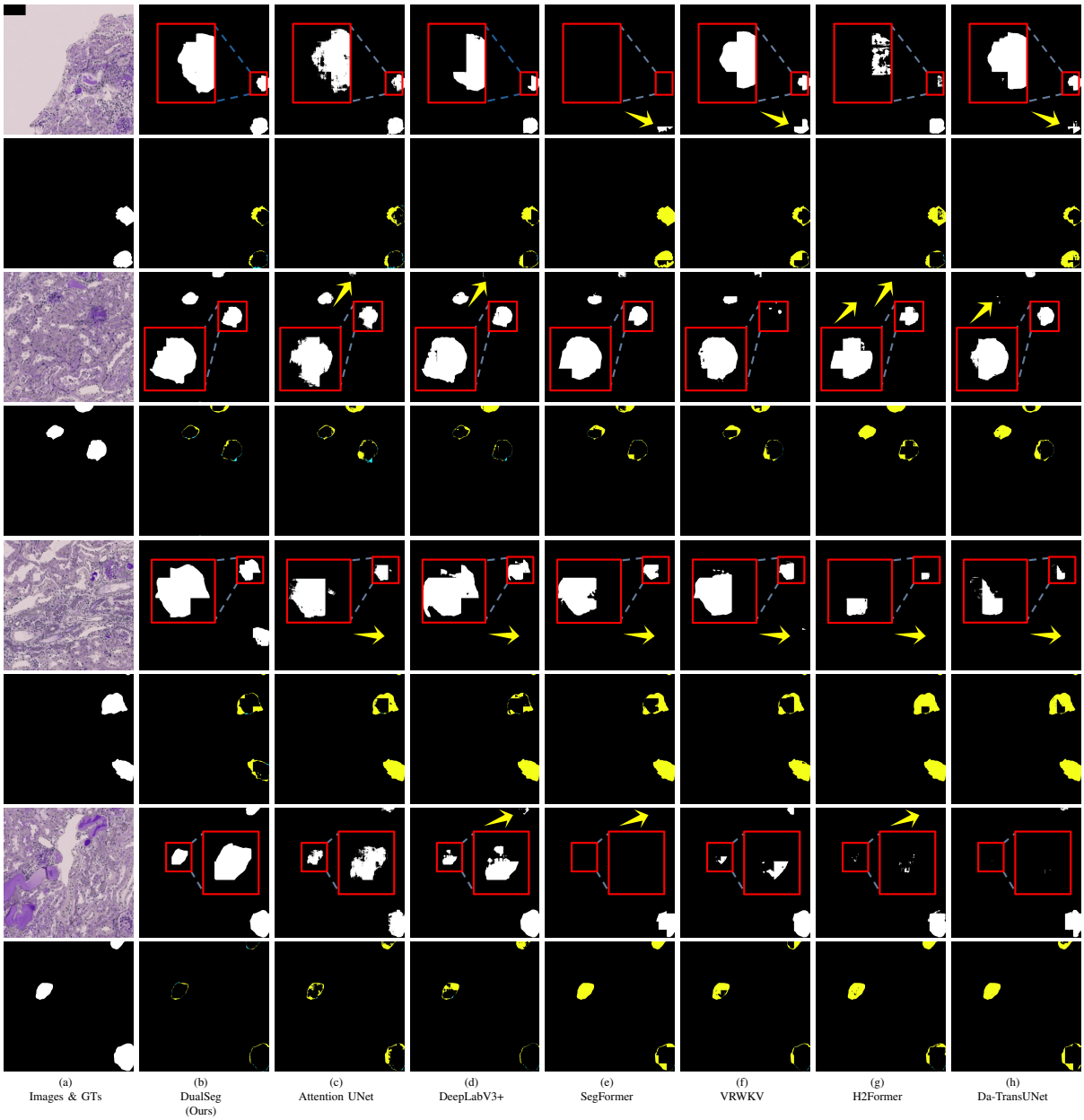

Fig. 6: Visual comparison on the KPIs test set. **Odd rows** display original images and inference masks; **even rows** show GT and error maps, where **yellow** and **green** denote under-segmentation and over-segmentation, respectively. Red boxes and matching arrows highlight magnified details and specific segmentation errors. **DualSeg** demonstrates superior performance in handling heterogeneous glomeruli with intricate boundaries.

metrics. Subsequently, we compare our method against SOTA approaches and perform ablation studies to validate the architectural design choices.

A. Datasets

Dataset I: Mice Glomeruli (KPIs). The murine kidney dataset was sourced from the MICCAI 2024 Kidney Pathology Image Segmentation (KPIs) challenge [48]. It includes PAS-stained images from four mouse models: normal, 5/6 nephrectomy (5/6Nx), diabetic nephropathy (DN), and NEP25 mice. To ensure consistency in patch-level evaluation, we utilized Task 1 data, provided as $2,048 \times 2,048$ pixel patches containing functional tissue units (FTUs). The dataset comprises 5,213 training, 1,643 validation, and 2,305 test images. For model compatibility, we performed non-overlapping cropping to generate 256×256 patches. We employed a 5-fold cross-validation strategy on the training/validation split (80%/20%), with final comparative results derived from the independent test set.

Dataset II: Human Glomeruli (HuBMAP). The first human dataset was obtained from the HuBMAP Kidney Challenge [49], focusing on FTU segmentation. It consists of 20 PAS-stained WSIs in TIFF format, averaging $36,000 \times 29,000$ pixels. Glomeruli are annotated with precise polygonal contours. Following standard protocols [33], we partitioned the

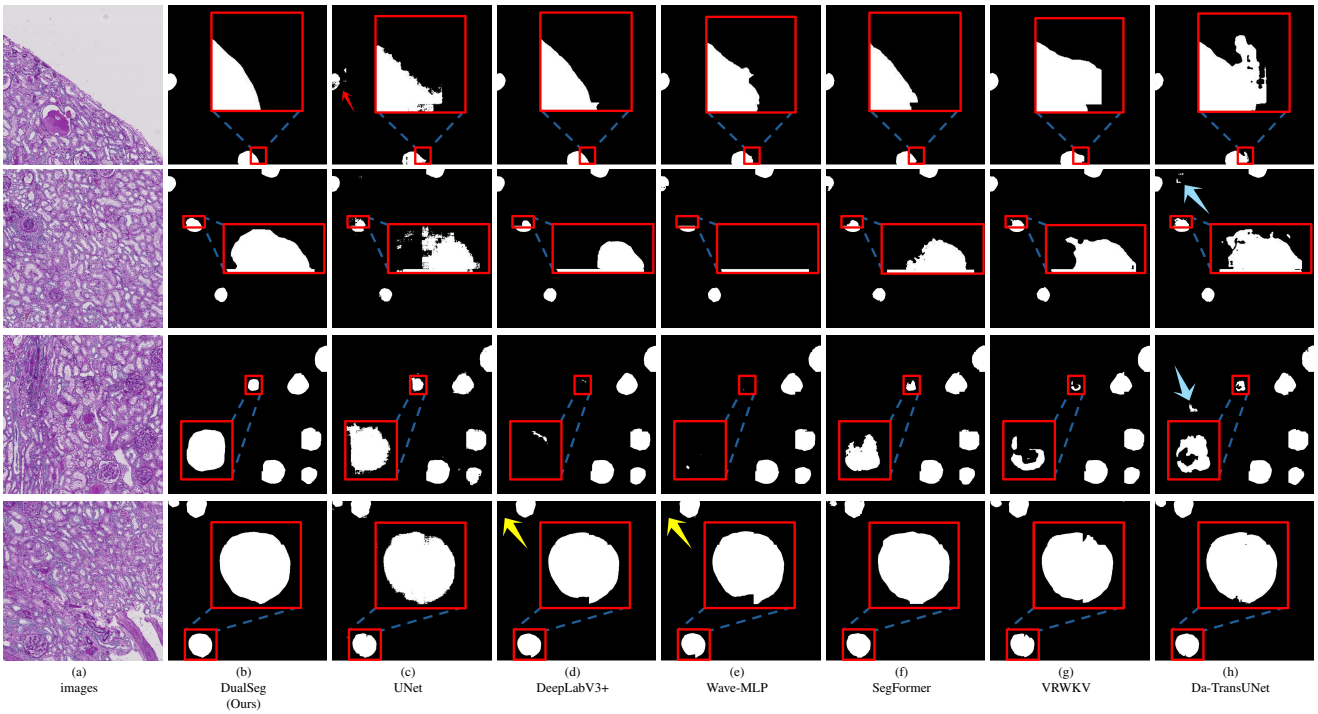

Fig. 7: Visual comparison on the HuBMAP test set. Since the official GT is private, we utilize the consensus of high-confidence predictions across multiple models as a reference. Local details are magnified to highlight segmentation challenges, with **yellow arrows** indicating suboptimal segmentation regions. **DualSeg** exhibits superior performance, evidenced by comprehensive edge processing and enhanced robustness against artifacts compared to baselines.

WSIs into 258,956 patches of size 256×256 and implemented a 5-fold cross-validation (80% training, 20% validation). Final performance metrics were obtained by submitting inference results on the test set to the official Kaggle evaluation portal.

Dataset III: Human Glomeruli (KPMP). To assess cross-species generalization, we retrieved a second human dataset from the Kidney Precision Medicine Project (KPMP) Atlas Repository [50]. Four PAS-stained SVS format WSIs (avg. resolution $84,000 \times 50,000$) were selected with corresponding masks. To rigorously validate generalization, models trained solely on the mouse KPIs dataset were directly applied to this human dataset without retraining. For preprocessing consistency, KPMP WSIs were partitioned into $2,048 \times 2,048$ patches.

B. Baselines

We benchmark DualSeg against 14 representative methods spanning three architectural paradigms. *CNN-based models* include canonical baselines like **U-Net** [7] and **Attention U-Net** [43], the receptive-field-enhanced **DeepLabV3+** [34], and efficient modern architectures such as **SegNeXt** [26], **InceptionNext** [44], and **Wave-MLP** [22]. *Transformer-based models* encompass **SegFormer** [11] for hierarchical encoding, **VRWKV** [23] utilizing linear recurrent operators, and the SSM-integrated **VM-UNet-V2** [40]. Finally, *Hybrid models* feature U-shaped Transformer variants like **UNETR** [9] and **Swin UNETR** [45], alongside advanced fusion frameworks including **H2Former** [10], **DA-TransUNet** [46], and the Mamba-based **U-Mamba** [25].

C. Evaluation Metrics

Performance is evaluated using the Dice Similarity Coefficient (DSC), Hausdorff Distance (95% percentile, HD95), and Intersection over Union (IoU). Let A and B denote the ground truth (GT) and predicted segmentation masks, respectively.

The **Dice coefficient** measures overlap:

$$Dice = \frac{2|A \cap B|}{|A| + |B|}. \tag{7}$$

The **HD95** quantifies boundary adherence by measuring the 95th percentile of the distances between surface points:

$$HD95 = \max_{p95} \left\{ \sup_{a \in A} \inf_{b \in B} d(a,b), \sup_{b \in B} \inf_{a \in A} d(b,a) \right\}, \tag{8}$$

where $d(a,b)$ is the Euclidean distance between points a and b.

The **IoU** assesses the ratio of intersection to union:

$$IoU = \frac{|A \cap B|}{|A \cup B|}. \tag{9}$$

Note that for the HuBMAP dataset, we only have access to the Dice score, which is computed via the official evaluation server.

D. Implementation Details

To mitigate overfitting, we applied data augmentation including horizontal/vertical flips and random rescaling. Models were trained for 20 epochs using the NAdam optimizer [51]. The learning rate was initialized at 10^{-4} and adjusted via Cosine Annealing [52].

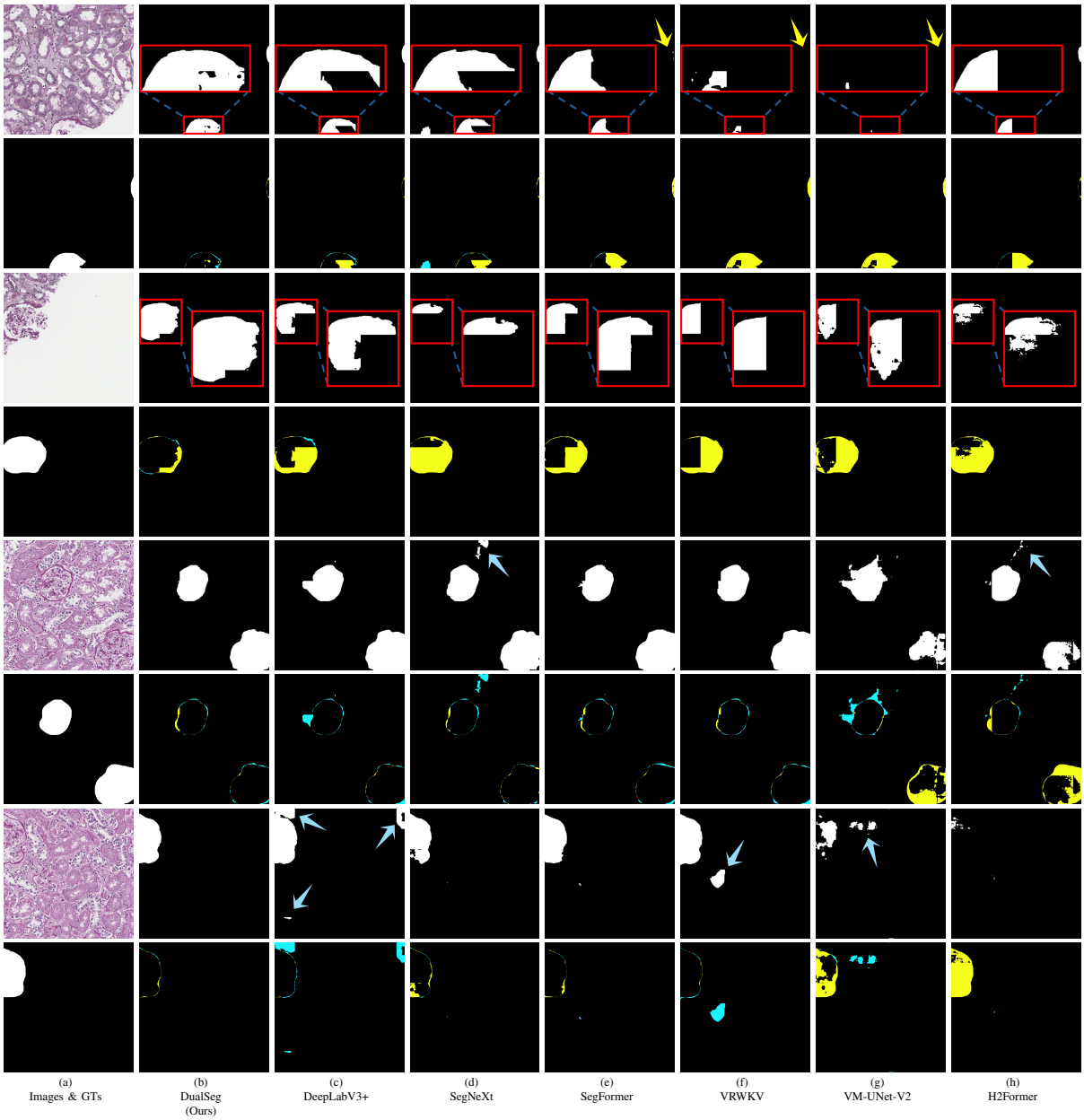

Fig. 8: Visual comparison on the held-out KPMP test set. **Odd rows** display original images and inference masks; **even rows** show GT and error maps, where **yellow** and **green** denote under-segmentation and over-segmentation, respectively. Red boxes and matching arrows highlight magnified details and segmentation errors. **DualSeg** exhibits unprecedented stability when handling cross-center and cross-species data.

Training on both HuBMAP and KPIs datasets utilized 256×256 patches with a batch size of 16. A critical adaptation was implemented for cross-dataset inference on KPMP: its patches were downscaled by a factor of 0.4. This scaling strategy specifically addresses the inherent size discrepancy between murine and human glomeruli [33], enabling direct inference with KPIs-pretrained models without domain adaptation. Inference was performed using MONAI's *Sliding Window Inference* [53] with a window size of 256×256 to ensure resolution consistency. All experiments were conducted under identical settings to guarantee fair comparison.

V. Results

In this section, we evaluate the performance of DualSeg across three distinct renal pathology datasets, benchmarking it against established SOTA methods. Furthermore, we conduct ablation studies to validate the architectural design choices of the proposed dual-stage encoder and associated blocks. Comprehensive quantitative and qualitative analyses are presented below.

A. Glomeruli Segmentation Results

1) KPIs Dataset: Table I quantifies the performance of DualSeg on the KPIs dataset, where it achieved the highest

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** [22]	✓	-	-	0.9165	0.9236	0.9322	0.8068	0.9036
	Attention [11]	-	✓	-	0.9223	0.9123	0.9249	0.8603	0.9099
	VRWKV [23]	-	-	✓	0.9209	0.9117	0.9268	0.8613	0.9108
Dual-Stage	**Attention-Wave(Ours)**	✓	✓	-	0.9192	0.9086	0.9241	0.8275	0.9019
	VRWKV-Wave(Ours)	✓	-	✓	0.8578	0.8168	0.8694	0.7011	0.8271
	Wave-Attention(Ours)	✓	✓	-	0.9267	0.9174	0.9334	0.8678	0.9171
	Wave-VRWKV(Ours)	✓	-	✓	0.9603	0.9349	0.9187	0.9007	0.9298

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 [22]	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 [23]	*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

performance across all pathological categories with superior stability, attaining an average mDSC of 92.98% and the lowest average HD95 of 66.85—outperforming top competing hybrid models like H2Former (90.52% mDSC) and DA-TransUNet (89.50% mDSC). In specific subgroup analyses, DualSeg robustly handled morphological extremes: in DN cases characterized by mild hypertrophy, it achieved an mDSC of 96.03%, surpassing the second-best SegFormer by 2.71% and reducing HD95 by 8.88 units; for NEP25 mice, it yielded an mDSC of 93.49%, significantly reducing the HD95 to 74.78 compared to SegNeXt's 256.44; and most notably, in the challenging 5/6Nx subset characterized by severe fragmentation, DualSeg was the only method to exceed a 90% threshold (90.07% mDSC), surpassing VRWKV by 2.69% while achieving a substantially lower HD95 of 149.61 compared to 188.17.

2) HuBMAP Dataset: On the human HuBMAP dataset, DualSeg demonstrated dominant performance (Table II). The model achieved a Private mDSC of 89.79% and a Public mDSC of 83.73%. It outperformed the closest competitors, DA-TransUNet and SegFormer, by margins of 0.44% and 1.26% on the Private set, respectively. Crucially, DualSeg exhibited exceptional stability, with a standard deviation of only 0.0040 for the Private mDSC, contrasting sharply with the higher variance observed in baseline models.

3) KPMP Dataset (Cross-Species Inference): To assess generalizability, we evaluated models pre-trained solely on the murine KPIs dataset directly on the human KPMP dataset without additional fine-tuning (Table III). DualSeg achieved a highly competitive average mDSC of 81.95%, surpassing SegFormer (80.82%) and significantly outperforming InceptionNext (52.79%). In terms of boundary delineation, DualSeg achieved an HD95 of 69.64, which is approximately one-third of the error recorded by VM-UNet-V2 (216.02 units). Furthermore, its IoU of 79.38% surpassed other hybrid models by over 4%, validating the model's capacity for robust cross-species and cross-center transfer learning.

B. Ablation Studies

We performed systematic ablation experiments on the KPIs dataset to quantify the contributions of key architectural components: the dual-stage encoder, the dynamic propagation window in the Wave-Swin Block, and the Z-Shift operator in the VRWKV Block. Results are summarized in Tables IV, V, and VI.

1) Effect of Dual-Stage Encoder: Table IV contrasts single-stage architectures with the proposed dual-stage design. Single-stage variants employing only Wave-Swin Blocks, SegFormer-style self-attention, or VRWKV Blocks achieved average mDSCs of 90.36%, 90.99%, and 91.08%, respectively. The integrated Dual-Stage (Wave-VRWKV) architecture outperformed all single-stage counterparts with an average mDSC of 92.98%. We further investigated the impact of module sequencing. Reversing the feature extraction order (placing attention mechanisms before wave blocks) resulted in a significant performance decrease, lowering the average mDSC to 85.78% (Attention-Wave) and 91.63% (VRWKV-Wave). These findings corroborate the critical role of the proposed *local-to-global* refinement strategy.

2) Effect of Dynamic Propagation Window: Table V evaluates the impact of the dynamic Swin mechanism. The adaptive window strategy (range 7–15) achieved an average mDSC of 92.98%, outperforming all fixed-window configurations (sizes 7, 11, and 15). This advantage was most pronounced in the heterogeneous 5/6Nx subset, where the dynamic window

improved mDSC by 2.70% compared to a fixed window size of 7.

3) Effect of Z-Shift Operator: Table VI demonstrates the efficacy of the Z-Shift operator over the conventional Q-Shift. Incorporating Z-Shift into DualSeg improved the average mDSC by 1.21%, with maximum gains observed in the NEP25 (1.29%) and 5/6Nx (1.05%) subsets. This confirms that the cyclic pixel wrapping strategy effectively mitigates information loss at patch boundaries.

C. Visualization Results

To intuitively verify the superior performance of DualSeg in glomerular segmentation, we present qualitative comparisons on both the murine KPIs dataset and the human HuBMAP dataset in Fig. 6 and Fig. 7, respectively. These visualizations focus on challenging scenarios—including glomeruli with irregular shapes, ambiguous boundaries, and fragmented structures—which critically test the model's ability to balance local texture discrimination with global context integration. Critical regions are magnified (marked with arrows) to highlight differences in segmentation precision.

For the mice KPIs dataset (Fig. 6), DualSeg accurately outlines renal glomeruli, avoiding the over-segmentation of adjacent areas observed in other methods. While baselines such as SegFormer and DA-TransUNet exhibit difficulties in delineating glomerular regions under abnormal pathological conditions, DualSeg effectively captures lesions even in areas with subtle staining variations. Notably, DualSeg successfully reconstructs the structural continuity of damaged glomeruli, whereas Attention U-Net and H2Former fail to connect fragmented regions, resulting in disjointed masks.

In the human HuBMAP dataset (Fig. 7), DualSeg demonstrates superior robustness in handling large-scale, complex histopathological backgrounds. Human glomeruli exhibit significant size and spatial variability, often presenting ambiguity between sclerotic glomeruli and surrounding interstitial fibrosis. DualSeg clearly distinguishes these boundaries and preserves residual glomerular textures, whereas DeepLabV3+ and Wave-MLP frequently generate blurred boundaries with instances of missed detection. Furthermore, DualSeg outperforms DA-TransUNet and VRWKV, which tend to overemphasize main structures, resulting in either over-segmentation or incomplete segmentation.

Figure 8 illustrates the cross-species generalization on the human KPMP dataset using models pre-trained solely on mouse data. Baseline models display varying degrees of failure, manifesting as spurious outliers and misaligned segments that diverge from anatomical GT. Additionally, these baselines exhibit pronounced uncertainty, with noisy pseudo-segments in internal regions and discontinuous edge contours. Specifically, VM-UNet-V2 and H2Former show prominent under-detection artifacts, entirely omitting small or intricate glomerular regions. In stark contrast, DualSeg maintains sharp boundary delineation and internal consistency, demonstrating remarkable stability and confidence in cross-center, cross-species inference without domain adaptation.

VI. Discussion

Glomerular segmentation faces three fundamental challenges: **local texture discriminability**, **spatial heterogeneity**, and the **multi-scale mapping of morphological priors**. DualSeg directly targets these issues through a unified architecture. Its advantages over state-of-the-art (SOTA) methods are analyzed below.

A. Comparative Analysis with SOTA Methods

DualSeg effectively addresses the limitations of existing architectural paradigms through a unified framework. Unlike conventional CNNs (e.g., DeepLabV3+, InceptionNext) which are constrained by fixed receptive fields, DualSeg utilizes the Wave-Swin Block's dynamic propagation window (Table V). This innovation enhances **local texture discriminability**, reducing HD95 by 26–37 on the KPIs dataset compared to CNN baselines and outperforming InceptionNext by 29.16% mDSC in cross-species tasks.

Furthermore, while Transformers typically excel at global context but compromise local detail due to patch flattening, DualSeg integrates a Z-Shift operator within the VRWKV block to preserve edge integrity. This design mitigates the information loss inherent in standard Q-Shift operations (Table VI) and models **spatial heterogeneity** more effectively, allowing DualSeg to exceed VM-UNet-V2 by 21.67% on the external KPMP dataset.

Finally, in contrast to existing hybrids (e.g., H2Former, DA-TransUNet) that suffer from a "semantic gap" or the high computational costs of U-Mamba (Fig. 1), DualSeg employs a sequential *local-to-global* refinement strategy. This structured integration ensures precise **morphological prior mapping**, enabling the model to outperform H2Former by 13.68% on HuBMAP and surpass U-Mamba by 23.52% on KPMP. By dynamically adapting to morphological variability and seamlessly integrating features, DualSeg establishes a robust and efficient backbone for renal histology analysis.

B. Failure Case Analysis

Fig. 9 reveals that the model occasionally fails to detect globally sclerotic glomeruli in the KPMP dataset. This limitation stems primarily from two factors: the partial truncation of peripheral glomeruli during WSI tiling, which compromises morphological context, and the significant divergence of unseen, extreme pathological variants. To mitigate this, future work could increase patch sizes to preserve boundary information or, more efficiently, integrate uncertainty-guided semi-supervised learning. This strategy aims to enhance robustness against rare phenotypes without incurring excessive computational overhead.

C. Clinical Relevance

DualSeg exhibits statistically significant superiority ($p < 0.05$–0.001; Fig. 10) and exceptional reproducibility, evidenced by a minimal standard deviation (0.004) on the HuBMAP dataset. Its ability to accurately resolve diverse

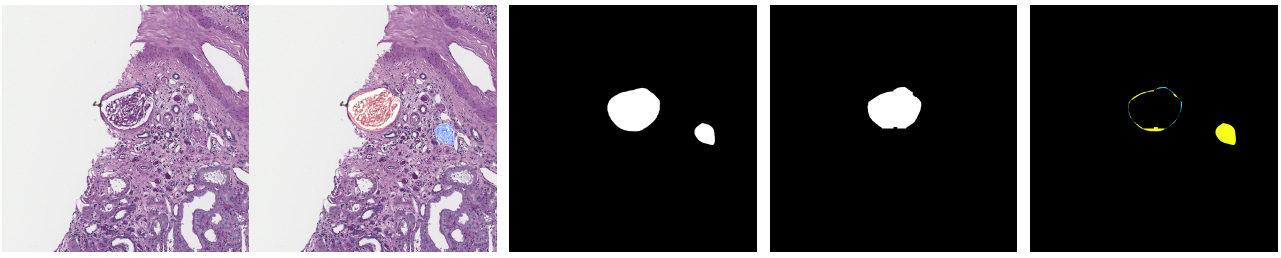

Fig. 9: Visual analysis of segmentation limitations. In the error map (far right), the yellow region highlights a significant false negative (under-segmentation). This failure is primarily attributed to the absence of globally sclerotic samples in the training set, preventing the model from generalizing to the *high heterogeneity* and *distinct morphological* features of this unseen pathology.

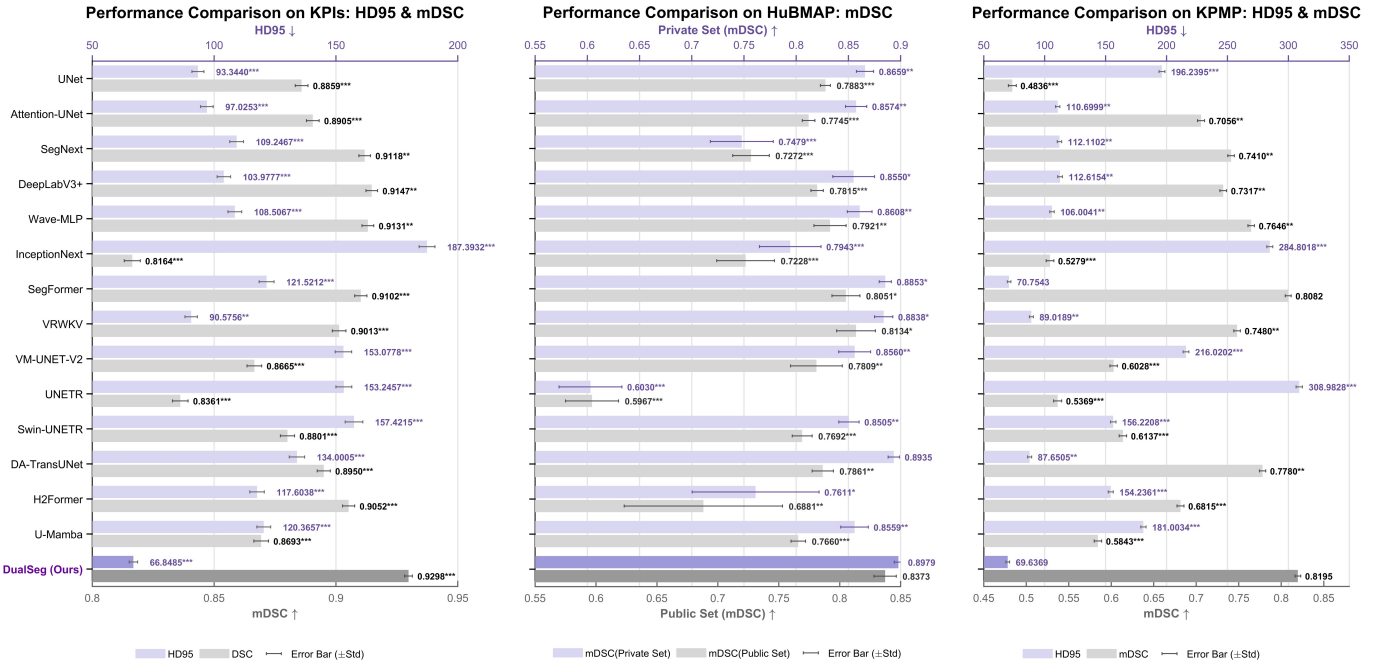

Fig. 10: Performance comparison between our DualSeg model and 14 baseline methods across three datasets. The comparison metrics include mDSC and HD95 for the KPIs and KPMP datasets (left and right panels, respectively), and mDSC for the HuBMAP dataset (middle panel). The error bars represent ± standard deviation. Statistical significance was assessed using paired t-tests, with levels indicated by asterisks: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

morphologies—ranging from mild hypertrophy to severe fragmentation—enables the precise quantification of pathological biomarkers like sclerosis and fibrosis. Furthermore, the model's robust performance on the cross-species KPMP dataset supports standardized CKD monitoring. By mitigating inter-observer variability and reducing manual annotation burdens, DualSeg provides a scalable solution for multi-center clinical trials and routine diagnostic workflows.

D. Limitations and Future Work

While DualSeg demonstrates superior performance in glomerular segmentation, three primary limitations remain to be addressed in future iterations. First, the model's generalization to rare pathological subtypes, such as global glomerulosclerosis, is currently constrained by data scarcity. We plan to mitigate this by employing domain adaptation and transfer learning techniques to enhance feature robustness for these underrepresented classes. Second, our current validation is restricted to PAS-stained images. To ensure broad clinical applicability, we will extend our evaluation to include Hematoxylin-Eosin (HE) stained datasets, verifying the model's adaptability to varying histological protocols. Finally, despite the linear complexity of the VRWKV block, the computational overhead for gigapixel WSI processing remains significant. Future work will focus on model quantization and lightweight optimization to facilitate deployment on resource-constrained platforms and edge devices.

VII. Conclusion

We present DualSeg, a unified hybrid framework designed to resolve the conflicting demands of local texture discriminability, spatial heterogeneity, and multi-scale morphological mapping in renal histopathology. By synergizing a Wave-Swin encoder featuring dynamic propagation windows with a

VRWKV module employing Z-Shift operators, the model effectively balances fine-grained feature extraction with efficient global context modeling. Extensive evaluations across murine (KPIs) and human (HuBMAP, KPMP) datasets demonstrate that DualSeg outperforms state-of-the-art methods, delivering superior mDSC and lowest HD95 scores while exhibiting robust cross-species generalization. Ablation studies further validate the critical role of the dual-stage architecture in mitigating boundary loss and adapting to morphological variations. These findings establish DualSeg as a precise, scalable backbone for automated glomerular segmentation, with broad potential for complex histopathological analysis.

[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] 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.

[23] 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.

[24] 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.

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

[26] 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.

[27] 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.

[28] 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.

[29] 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.

[30] 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.

[31] 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.

[32] 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.

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

[34] 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.

[35] 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.

[36] 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.

[37] 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.

[38] 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.

[39] 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.

[40] 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.

[41] 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.

[42] 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.

[43] 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.

[44] 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.

[45] 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.

[46] 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.

[47] 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.

[48] 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.

[49] 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

[50] 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

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

[52] I. Loshchilov and F. Hutter, "Sgdr: Stochastic gradient descent with warm restarts," *arXiv preprint arXiv:1608.03983*, 2016.

[53] 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.