U-RWKV: Lightweight medical image segmentation with direction-adaptive RWKV

Hongbo Ye[1,2], Fenghe Tang[1,2], Peiang Zhao[1,2], Zhen Huang[1,2], Dexin Zhao[1,2], Minghao Bian[1,2], and S. Kevin Zhou[1,2,3,4]✉

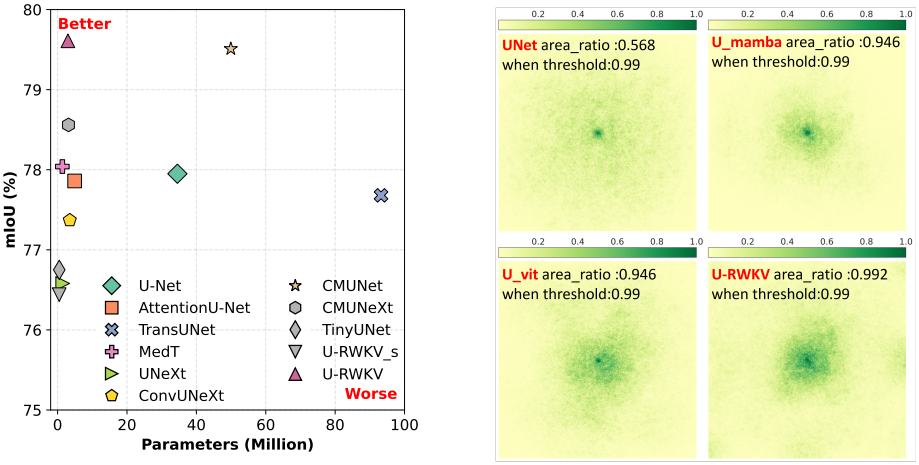
[1] School of Biomedical Engineering, Division of Life Sciences and Medicine, University of Science and Technology of China (USTC), Hefei Anhui, 230026, China
[2] Center for Medical Imaging, Robotics, Analytic Computing & Learning (MIRACLE), Suzhou Institute for Advance Research, USTC
[3] Jiangsu Provincial Key Laboratory of Multimodal Digital Twin Technology, Suzhou
[4] State Key Laboratory of Precision and Intelligent Chemistry, USTC

Abstract. Achieving equity in healthcare accessibility requires lightweight yet high-performance solutions for medical image segmentation, particularly in resource-limited settings. Existing methods like U-Net and its variants often suffer from limited global Effective Receptive Fields (ERFs), hindering their ability to capture long-range dependencies. To address this, we propose U-RWKV, a novel framework leveraging the Recurrent Weighted Key-Value(RWKV) architecture, which achieves efficient long-range modeling at O(N) computational cost. The framework introduces two key innovations: the Direction-Adaptive RWKV Module(DARM) and the Stage-Adaptive Squeeze-and-Excitation Module(SASE). DARM employs Dual-RWKV and QuadScan mechanisms to aggregate contextual cues across images, mitigating directional bias while preserving global context and maintaining high computational efficiency. SASE dynamically adapts its architecture to different feature extraction stages, balancing high-resolution detail preservation and semantic relationship capture. Experiments demonstrate that U-RWKV achieves state-of-the-art segmentation performance with high computational efficiency, offering a practical solution for democratizing advanced medical imaging technologies in resource-constrained environments. The code is available at https://github.com/hbyecoding/U-RWKV.

Keywords: RWKV · Lightweight neural networks · Scanning strategy

1 Introduction

Bridging the gap in healthcare accessibility requires not only breakthroughs in medical technology but also solutions that can be widely deployed across diverse clinical environments, especially in resource-limited settings [19]. In the field of medical image segmentation, while convolutional neural networks (CNNs), such as U-Net [20] and its variants [17,31,12,25,11,29,22,23], have achieved initial success through localized feature extraction, they fundamentally suffer from inadequate global Effective Receptive Fields(ERFs) [16,7], as shown by U-Net's



(a) U-RWKV achieves the highest Avg. Dice with efficient parameters

(b) U-RWKV has a larger ERF than U-Mamba and U-ViT-please see 3.3

Fig. 1: Performance comparison and Effective Receptive Field (ERF).

0.568 high-contribution ratio at 0.99 threshold (Fig. 1(b)). Lightweight yet high-performing models with global ERF thus hold immense potential, offering a viable and equitable pathway to democratize access to advanced medical imaging technologies.

To address the limitations of existing methods [4,3,28,14,15,13,30], we propose **U-RWKV**, a novel lightweight framework that leverages the emerging Recurrent Weighted Key-Value (RWKV) architecture [18]. RWKV achieves long-range modeling at $O(N)$ computational cost, offering a powerful foundation for efficient medical image segmentation through its linear-complexity attention mechanism. At the core of our U-RWKV are two key components: the Direction Adaptive RWKV Module (DARM) and the Stage Adaptive Squeeze and Excitation Module (SASE). These modules work synergistically to model long-range spatial dependencies while maintaining computational efficiency, setting it apart from traditional transformer-based or convolutional models, as show in Fig. 1(a).

The DARM is designed to dynamically aggregate contextual cues across the entire image by introducing two innovative mechanisms: **Dual-RWKV** and **QuadScan**. The core algorithm of RWKV, inspired by RNN-like WKV computations, is inherently designed for processing one-dimensional sequential data. However, this presents a challenge when adapting it to visual data, which lacks an inherent sequential arrangement of components. To address this issue, we propose the Dual-RWKV mechanism, which processes 2D feature maps as dual 1D sequences—one in the original order and the other in reverse order. This bidirectional design ensures cross-orientation context preservation while eliminating directional bias. By propagating information bidirectionally, Dual-RWKV

captures multi-scale contextual dependencies, mitigating information loss in ambiguous regions such as diffuse boundaries or corner-situated lesions.

On the other hand, the complex spatial relationships and diverse modalities (e.g., CT, MRI) in medical imaging demand adaptive mechanisms capable of capturing anisotropic features [26,24,27]. To meet this requirement, we introduce **QuadScan**, a quad-directional scanning strategy that traverses the image through four directional flows: left→right, right→left, top→bottom, and bottom→top. Each image patch acquires contextual knowledge exclusively through a compressed hidden state computed along its corresponding scanning path, reducing computational complexity while preserving global context. This systematic integration of edge semantics from multiple directions achieve global ERF, as shown in Fig. 1(b).

To further enhance the adaptability of U-RWKV, we introduce the Stage-Adaptive Squeeze-and-Excitation Module (SASE). SASE dynamically adjusts its architecture based on the stage of feature extraction. In early stages, SASE employs dilated inverted bottleneck structures to preserve high-resolution features, ensuring detailed spatial information is retained. In deeper layers, SASE transitions to compact bottleneck designs to maintain computational efficiency while capturing high-level semantic relationships. This stage-adaptive design enables U-RWKV to generalize effectively across different datasets, accommodating the intricate spatial correlations and semantic relationships inherent in medical imaging modalities such as CT and MRI.

In summary, our main contributions are as follows: **(I)** We propose **U-RWKV**, a lightweight framework balancing computational efficiency and segmentation performance for resource-constrained settings; **(II)** We introduce two innovations: (a) **DARM**, which uses **Dual-RWKV** and **QuadScan** to model long-range dependencies efficiently while reducing directional bias; and (b) **SASE**, which adapts dynamically to enhance the model's robustness and generalization; **(III)** Comprehensive experiments validate U-RWKV's state-of-the-art performance, efficiency, and adaptability across diverse medical imaging tasks.

2 Method

2.1 Architecture Overview

The proposed architecture introduces a novel U-shaped encoder-decoder framework tailored for medical image segmentation, as depicted in Fig. 2(a). The model is designed to efficiently process input images through a hierarchical structure that captures multi-scale features. The encoder progressively reduces the spatial dimensions of the input while increasing the channels. It uses a series of convolutional layers with 3×3 kernels and stride=2 for downsampling. The decoder path aims to reconstruct the feature maps through a series of upsampling operations, performed by transposed convolutions that gradually restore the spatial resolution. This process is named *ChannelFusion* because it involves two layers of CNNs. Each convolutional layer is followed by batch normalization to ensure robust feature extraction and regularization.

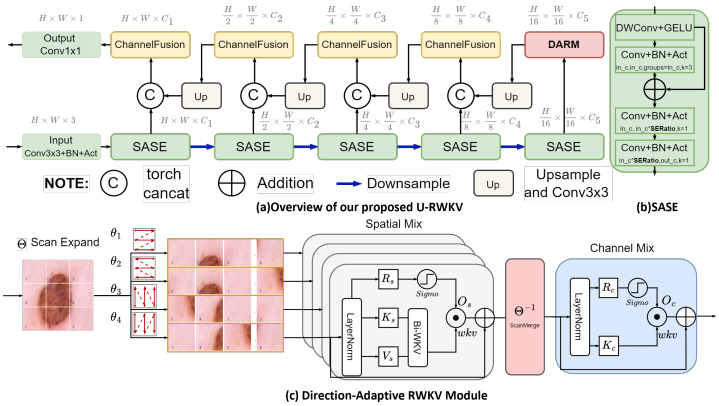


Fig. 2: Overview of the proposed U-RWKV; SASE Module and DARM

2.2 Stage-Adaptive SE Module (SASE)

The *SASE* block dynamically adapts to different stages of the network to complement RWKV. Its architecture enhances the hierarchical feature transmission to DARM. Specifically, this design uses lightweight pointwise convolutions and inverted residual structures in shallow modes. Shallow mode refers to early stages where the resolution is high and the feature channel ratio, calculated as the number of channels divided by the product of height and width, is large. In these stages, the SERatio is set to 4, meaning the output channels are four times the input channels. In deeper stages, where the ratio of channels to H×W is smaller (C/H/W $\geq$ 1), we perform channel-wise splitting into 8 parts and then double the channels, making the SERatio effectively 1/4. We also use depthwise separable convolutions in these deeper modes to improve spatial feature extraction. Using lightweight convolutions and residuals in shallow stages balances efficiency and feature richness. This resolution-aware design enhances the feature informativeness of DARM.

2.3 Direction-Adaptive RWKV Module (DARM)

As discussed earlier, the fine-grained local features extracted by the encoder need to incorporate long-range dependencies to enable effective fusion of local and global information in the decoder. To achieve this, we propose DARM, which refines the encoded features while preserving their spatial and channel-wise relationships, leveraging the temporal and channel mixers from RWKV and Vision RWKV.

Preliminaries: RWKV for vision data RWKV processes an input $s \in \mathbb{R}^{T \times C}$, where C is the number of channels and T is the sequence length. First, layer normalization (LN) stabilizes the features. The normalized features are projected

Algorithm 1 DARM: Direction-Adaptive RWKV Module

Require: $E \in \mathbb{R}^{C\times H\times W}$ (encoder features)
Ensure: $D_{\text{pre}} \in \mathbb{R}^{C\times H\times W}$ (features for decoder)
Step 1: QuadScan Expansion and Spatial Mix
for each direction $i \in \{L \to R,\ R \to L,\ T \to B,\ B \to T\}$ **do**
  $s_i = \theta_i(E)$ ▷ Apply directional scan $\theta_i : \mathbb{R}^{C\times H\times W} \to \mathbb{R}^{T\times C}$
  $s_i' = \text{spa}(s_i)$ ▷ Spatial mix
end for
Step 2: QuadScan Merge
$E_i = \theta_i^{-1}(s_i')$ ▷ Reconstruct 2D features from sequences
$E' = \text{Average}(\text{Stack}(E_1, E_2, E_3, E_4))$ ▷ Pixel-wise averaging
Step 3: Channel Mixing
$D_{\text{pre}} = \text{cha}(\text{Flatten}(E'))$ ▷ Channel-wise enhancement
Return D_{pre}

into three components: receptance R_s, key K_s, and value V_s, via learnable matrices $W_R, W_K, W_V \in \mathbb{R}^{C\times C}$.

Borrowing from [9], we define spatial and channel mixing as follows: **Spatial mixing (spa(·))** is a token-wise aggregation:

$$spa_t = wkv_t = \text{Bi-WKV}(K_s, V_s)_t = \frac{\sum_{i=1,i\neq t}^{T} e^{-\frac{|t-i|-1}{T}\cdot w + k_i} v_i + e^{u+k_t} v_t}{\sum_{i=1,i\neq t}^{T} e^{-\frac{|t-i|-1}{T}\cdot w + k_i} + e^{u+k_t}}, \quad (1)$$

where u and w are learnable parameters controlling local and non-local interactions respectively. This enables dynamic importance adjustment of nearby and distant tokens, ensuring robust feature aggregation.

Channel mixing (cha(·)) is a pointwise feed-forward network applied across the channel dimension.

For vision data $X \in \mathbb{R}^{C\times H\times W}$, we transform spatial features into a sequence using Vision2Seq(·), which is a more sophisticated process than simply reading each row sequentially (Flatten(·)). The VRWKV process then combines spatial and channel mixing:

$$\text{VRWKV}(X) = \text{cha}(\text{Flatten}(\text{spa}(\text{Vision2Seq}(X)))). \quad (2)$$

For the DARM input, let $E \in \mathbb{R}^{C\times H\times W}$ be the encoded features and $D_{\text{pre}} \in \mathbb{R}^{C\times H\times W}$ the refined features. Since H and W are reduced due to encoder downsampling, we set the patch size in DARM's patch embedding to 1, and the sequence length to $H \times W$. As shown in Fig. 2(c), E first passes through the **QuadScan** mechanism, which operates along spatial dimensions while keeping channel information intact.

QuadScan Mechanism: This operation scans the feature map E along four directions: left-to-right, right-to-left, top-to-bottom, and bottom-to-top. Each directional scan produces a 1D sequence s_i via θ_i. These sequences undergo spatial mixing through spa(·) to refine long-range dependencies. After processing, the inverse functions θ_i^{-1} reconstruct spatial features E_i, which are then

averaged pixel-wise into the final feature map E'. The resulting feature map is subsequently flattened into a sequence and sequentially fed into the channel mix module, enabling the model to capture a comprehensive receptive field for subsequent processing.

Dual-RWKV Mechanism. This core feature refinement module processes 2D feature maps as two separate 1D sequences—one in the original order and the other in reverse—without weight sharing between directions. This bidirectional design preserves cross-orientation context while preventing directional bias. By propagating information in both forward and backward passes independently, Dual-RWKV captures richer spatial dependencies. When combined with Quad-Scan, the model can capture complementary directional information, further enhancing its robustness and adaptability.

Let the symbols be defined as above. The unified process of our **Direction-Adaptive RWKV Module (DARM)** can be formulated as:

$$D_{pre} = \mathrm{cha}\left(\mathrm{Flatten}\left(\Theta^{-1}\left(\mathrm{spa}\left(s\right)\right)\right)\right) + \mathrm{cha}\left(\mathrm{Flatten}\left(\Theta^{-1}\left(\mathrm{spa}\left(s^{\leftarrow}\right)\right)\right)\right) \tag{3}$$

3 Experiments and Results

3.1 Settings

Datasets. Our study utilizes diverse datasets. The BUSI dataset [1] consists of breast ultrasound images from 600 female patients, with 780 images in total, classified into normal, benign, and malignant. Kvasir [10] and ClinicDB [2] are polyp-related endoscopic datasets. Kvasir has 1,000 manually-annotated polyp images, and ClinicDB contains 612 static images from colonoscopy videos. The ISIC 2017 and 2018 datasets [6] focus on skin diseases, with different numbers of training and test images.

Metrics. In medical image segmentation, we commonly use the Dice Similarity Coefficient (DSC) and the Intersection over Union (IoU) to evaluate performance. Higher values of DSC and IoU indicate better segmentation accuracy.

Implementation Details. The training procedure follows the settings described in [22,25], with the following modifications: training is conducted for 280 epochs on a single NVIDIA 3090 GPU; the official Synapse dataset is used exclusively, while for other datasets, a 70/30 split is applied for training and validation, respectively; the RWKV model is initialized with weights from [9].

3.2 Comparison with State-of-the-Art Methods

We compare our U-RWKV model against several state-of-the-art methods. Table 1 presents the Dice scores on five datasets, along with comparisons of the number of parameters (in M) and FLOPs (in G) for different models, which reflect computational complexity. Our U-RWKV model achieves competitive performance, attaining the highest average Dice score of **82.27**, surpassing most existing methods. Notably, the lightweight variant U-RWKV-s achieves a Dice score of 80.06 with only 0.46M parameters, highlighting its efficiency.

Table 1: Segmentation performances of competing methods in terms of Dice score. Reported Params, GFLOPs, average Dice (Avg), and Dice scores per dataset. Higher Dice values are better. Maximum values are highlighted in **bold**.

Methods	params	FLOPs	Avg	BUSI	Kvasir	Clinic	ISIC'17	ISIC'18
U-Net [20]	34.53	65.52	81.48	79.58	87.65	90.97	89.87	86.99
TransUnet [4]	93.23	32.23	81.23	79.61	87.13	90.84	90.10	86.58
CMU-Net [25]	49.93	91.25	83.06	81.92	**89.12**	**92.48**	89.70	86.83
SwinUnet [3]	27.15	5.91	76.05	76.46	80.67	84.15	87.71	86.57
UNeXt [29]	1.47	**0.57**	80.18	80.47	85.11	88.76	89.60	86.80
Att-UNet [17]	4.91	9.45	81.48	79.61	87.13	91.77	89.57	86.86
MedT [28]	1.37	2.41	81.81	81.86	88.85	90.38	86.72	87.21
ConvUNeXt [11]	3.51	7.25	81.11	80.37	86.67	90.99	89.35	85.89
CMUNeXt [22]	3.14	7.41	82.13	81.66	87.82	91.21	89.85	86.77
TinyUnet [5]	0.48	1.67	80.45	77.42	87.32	90.37	89.03	86.87
U-RWKV-s	**0.46**	1.02	80.06	79.77	86.15	87.98	89.41	86.94
U-RWKV	2.97	7.28	**82.27**	**82.34**	88.17	90.58	**90.13**	**87.26**

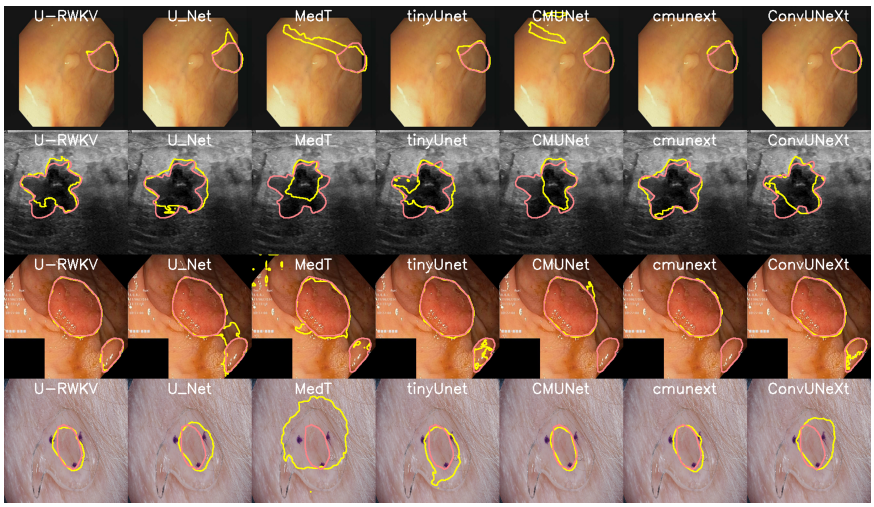


Fig. 3: Visualization of segmentation results on four datasets: CVC-ClinicDB, BUSI, Kvasir-Seg, and ISIC 2017. The orange contour lines represent the ground-truth annotations of lesions, while the yellow contour lines indicate the segmentation results produced by our model.

We evaluate U-RWKV on the Synapse multi-organ segmentation dataset. Table 2 shows the Dice scores and Hausdorff Distance (HD95) for each organ, along with the average scores across all organs. Our U-RWKV model achieves

competitive performance, with an average Dice score of **80.64** and HD95 of 26.61. Notably, U-RWKV outperforms several state-of-the-art methods, including TransUNet and MedT.

The synergy between RWKV's long-range dependency modeling and SASE's stage-adaptive feature refinement is key to this performance. SASE dynamically enhances lesion-specific features at different decoder stages—coarse-grained localization in early stages and fine-grained boundary precision in later ones. This is evident in Fig. 3, where U-RWKV reliably handles diverse challenges: heterogeneous textures in BUSI, mucosal folds in Kvasir, and low-contrast boundaries in ISIC. The ground truth (orange) and predictions (yellow) align closely, particularly for irregular structures, highlighting SASE's role in preserving topological consistency.

Table 2: Comparison of different models on Synapse dataset.

Methods	Dice↑	HD95↓
U-Net	77.10	29.97
TransUNet	77.54	38.78
CMU-Net	76.22	29.65
UNeXt	72.52	39.61
AttUNet	76.10	36.77
MedT	70.09	33.53
ConvUNeXt	<u>78.55</u>	26.89
CMUNeXt	77.95	**24.43**
TinyUnet	75.75	32.23
U-RWKV-s	71.12	45.36
U-RWKV	**80.64**	<u>26.61</u>

Table 3: IoU results of ablation studies on BUSI, Kvasir and ISIC'17.

Ablation	BUSI↑	Kvasir↑	ISIC↑	Avg.↑
Left→**R**ight	69.61	77.91	81.65	76.75
Right→**L**eft	69.08	77.56	82.26	76.61
Top→**B**ottom	69.04	78.87	82.23	77.02
Bottom→**T**op	68.05	78.22	81.77	76.43
L.→R.+R.→L.	69.73	77.22	81.90	76.64
T.→B.+B.→T.	69.73	78.35	81.98	76.98
L.→R.+T.→B.	69.38	77.74	82.26	76.77
R.→L.+B.→T.	68.86	78.75	81.92	76.86
w/o Dual RWKV	70.30	77.31	82.42	76.68
w/o DARM	66.73	76.55	81.62	74.97
w/o SASE	68.57	75.78	81.53	75.29
U-RWKV	**71.01**	**79.58**	**82.27**	**77.62**

3.3 Ablation Studies

We conduct comprehensive ablation studies, summarized in Table 3. The baseline results show that combining multi-directional scans (e.g., L→R + R→L, T→B + B→T, etc.) improves IoU across datasets. Removing components such as Dual RWKV or DARM causes notable performance drops (e.g., without DARM, the average IoU decreases from 77.62 to 74.97). Importantly, the full U-RWKV, which integrates SASE with DARM, achieves the highest average IoU of 77.62, demonstrating that SASE on its own is not merely a variant but works synergistically with DARM to enhance segmentation performance.

We further validate these findings through Effective Receptive Field analysis. As shown in Fig. 1(b), U-RWKV achieves a 0.992 high-contribution area ratio at 0.99 threshold, significantly outperforming both the baseline U-Net (0.568) and our model's variants with ViT [8] or VMUnet's [21]mamba bottlenecks (both

0.946). This 74.6% improvement over U-Net and 4.9% advantage over the backbone variants confirms our architecture's superior ability to focus activation energy on diagnostically relevant regions while maintaining global context awareness, consistent with the IoU improvements observed in the component ablation studies.

4 Conclusion

In summary, we propose U-RWKV, a framework that combines convolutional features with DARM's global dependency modeling, achieving a good balance between efficiency and accuracy. We acknowledge that inference speed is slightly slower than CNNs like UNeXt, and future work will focus on optimizing for high-resolution settings and extending to 3D segmentation.

Acknowledgments. Supported by National Natural Science Foundation of China under Grant 62271465, Suzhou Basic Research Program under Grant SYG202338.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Al-Dhabyani, W., Gomaa, M., Khaled, H., Fahmy, A.: Dataset of breast ultrasound images. Data in brief **28**, 104863 (2020)
2. Bernal, J., Sánchez, F.J., Fernández-Esparrach, G., Gil, D., Rodríguez, C., Vilariño, F.: Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. Computerized medical imaging and graphics **43**, 99–111 (2015)
3. Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., Wang, M.: Swin-unet: Unet-like pure transformer for medical image segmentation. In: European conference on computer vision. pp. 205–218. Springer (2022)
4. Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A.L., Zhou, Y.: Transunet: Transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306 (2021)
5. Chen, J., Chen, R., Wang, W., Cheng, J., Zhang, L., Chen, L.: Tinyu-net: Lighter yet better u-net with cascaded multi-receptive fields. In: MICCAI. pp. 626–635. Springer (2024)
6. Codella, N., Rotemberg, V., Tschandl, P., Celebi, M.E., Dusza, S., Gutman, D., Helba, B., Kalloo, A., Liopyris, K., Marchetti, M., et al.: Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic). arXiv preprint arXiv:1902.03368 (2019)
7. Ding, X., Zhang, X., Han, J., Ding, G.: Scaling up your kernels to 31x31: Revisiting large kernel design in cnns. In: CVPR. pp. 11963–11975 (2022)
8. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)

9. Duan, Y., Wang, W., Chen, Z., Zhu, X., Lu, L., Lu, T., Qiao, Y., Li, H., Dai, J., Wang, W.: Vision-rwkv: Efficient and scalable visual perception with rwkv-like architectures. arXiv preprint arXiv:2403.02308 (2024)
10. Guo, Y., Bernal, J., J Matuszewski, B.: Polyp Segmentation with Fully Convolutional Deep Neural Networks—Extended Evaluation Study. Journal of Imaging **6**(7), 69 (2020)
11. Han, Z., Jian, M., Wang, G.G.: Convunext: An efficient convolution neural network for medical image segmentation. Knowledge-based systems **253**, 109512 (2022)
12. Huang, H., Lin, L., Tong, R., Hu, H., Zhang, Q., Iwamoto, Y., Han, X., Chen, Y.W., Wu, J.: Unet 3+: A full-scale connected unet for medical image segmentation. In: ICASSP. pp. 1055–1059. IEEE (2020)
13. Huang, Z., Li, H., Shao, S., Zhu, H., Hu, H., Cheng, Z., Wang, J., Kevin Zhou, S.: Pele scores: pelvic x-ray landmark detection with pelvis extraction and enhancement. IJCAS **19**(5), 939–950 (2024)
14. Jha, D., Riegler, M., Johansen, D., Halvorsen, P., Johansen, H.: DoubleU-Net: A Deep Convolutional Neural Network for Medical Image Segmentation. In: CBMS (2020)
15. Jha, D., Smedsrud, P.H., Riegler, M.A., Johansen, D., De Lange, T., Halvorsen, P., Johansen, H.D.: ResUNet++: An Advanced Architecture for Medical Image Segmentation. In: ISM. pp. 225–230 (2019)
16. Luo, W., Li, Y., Urtasun, R., Zemel, R.: Understanding the effective receptive field in deep convolutional neural networks. NeurIPS **29** (2016)
17. Oktay, O.: Attention u-net: Learning where to look for the pancreas. arXiv preprint arXiv:1804.03999 (2018)
18. Peng, B., Alcaide, E., Anthony, Q.G., Albalak, A., Arcadinho, S., Biderman, S., Cao, H., Cheng, X., Chung, M.N., Derczynski, L., et al.: Rwkv: Reinventing rnns for the transformer era. In: EMNLP (2023)
19. Richardson, S., Lawrence, K., Schoenthaler, A.M., Mann, D.: A framework for digital health equity. NPJ digital medicine **5**(1), 119 (2022)
20. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18. pp. 234–241. Springer (2015)
21. Ruan, J., Li, J., Xiang, S.: Vm-unet: Vision mamba unet for medical image segmentation (2024), https://arxiv.org/abs/2402.02491
22. Tang, F., Ding, J., Quan, Q., Wang, L., Ning, C., Zhou, S.K.: Cmunext: An efficient medical image segmentation network based on large kernel and skip fusion. In: 2024 IEEE International Symposium on Biomedical Imaging (ISBI). pp. 1–5. IEEE (2024)
23. Tang, F., Nian, B., Ding, J., Quan, Q., Yang, J., Liu, W., Zhou, S.K.: Mobileutr: Revisiting the relationship between light-weight cnn and transformer for efficient medical image segmentation. arXiv preprint arXiv:2312.01740 (2023)
24. Tang, F., Nian, B., Li, Y., Jiang, Z., Yang, J., Liu, W., Zhou, S.K.: Mambamim: Pre-training mamba with state space token interpolation and its application to medical image segmentation. Medical Image Analysis p. 103606 (2025)
25. Tang, F., Wang, L., Ning, C., Xian, M., Ding, J.: Cmu-net: a strong convmixer-based medical ultrasound image segmentation network. In: ISBI. pp. 1–5. IEEE (2023)
26. Tang, F., Xu, R., Yao, Q., Fu, X., Quan, Q., Zhu, H., Liu, Z., Zhou, S.K.: Hyspark: Hybrid sparse masking for large scale medical image pre-training. In: MICCAI. pp. 330–340. Springer (2024)

27. Tang, F., Yao, Q., Ma, W., Wu, C., Jiang, Z., Zhou, S.K.: Hi-end-mae: Hierarchical encoder-driven masked autoencoders are stronger vision learners for medical image segmentation. arXiv preprint arXiv:2502.08347 (2025)
28. Valanarasu, J.M.J., Oza, P., Hacihaliloglu, I., Patel, V.M.: Medical transformer: Gated axial-attention for medical image segmentation. In: MICCAI. pp. 36–46 (2021)
29. Valanarasu, J.M.J., Patel, V.M.: Unext: Mlp-based rapid medical image segmentation network. In: MICCAI. pp. 23–33. Springer (2022)
30. Zhou, X., Huang, Z., Zhu, H., Yao, Q., Zhou, S.K.: Hybrid attention network: An efficient approach for anatomy-free landmark detection. arXiv preprint arXiv:2412.06499 (2024)
31. Zhou, Z., Rahman Siddiquee, M.M., Tajbakhsh, N., Liang, J.: Unet++: A nested u-net architecture for medical image segmentation. In: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4. pp. 3–11. Springer (2018)