

# 基于 Stage 扫描机制的 Vision RWKV 让医学影像分割轻量而高效

sora\*, Zhendong Li\*,

\*Ningxia University

Email: kyochilian@gmail.com

\*Ningxia University

Email: lizhendong@nxu.edu.cn

**Abstract**—基于 Transformer 的分割框架是当前医学影像分割的主流方法，能够构建全局关系。然而，在需要高精度与高分辨率的医学图像分割中，Transformer 计算复杂度高，应用有限。最近的 RWKV 模型降低了空间聚合复杂度并具备全局处理能力，但在视觉领域存在长程建模羸弱的问题，贪吃蛇效应突出，同时循环嵌套机制在分割领域中的解码性能不足。为了解决这些问题，我们提出了一种基于 Vision RWKV 的分割框架 RWKV-SSS，该框架在具有较低复杂度的同时具备了全局建模能力。为了增强 RWKV 的长程空间连续性，我们设计了一种 Stage 扫描机制，该机制将图像分为全局 block 与局部 block，根据不同 Stage 的编码器动态地调整扫描权重，将全局与局部特征融合，大大提升了模型的精度。同时，我们在解码器中构建了上采样模块，该模块改良了 RWKV 的循环嵌套机制，提升了重建多尺度特征的能力。我们还采用多种方法进一步降低 SSS 的计算复杂度，实现模型的轻量化。(LRFormer 低分辨率自注意力，池化 QKV 等等) 实验表明，我们的方法在多个数据集取得优异性能，同时计算复杂度与计算速度显著增强，使其在高精度分割的同时具备轻量化特性。代码开源至：<https://github.com/shepherdxu/SCI-winning>。

**Index Terms**—computer vision, semantic segmentation, RWKV, Transformer, lightweight

## I. INTRODUCTION

Semantic Segmentation 是计算机视觉中的一项重要任务，医学影像分割是关键应用之一。它将复杂的医学图像分割成不同的区域，使器官、病灶及注意区域清晰可见，对于医学辅助诊断与治疗具有重要意义。相较于人工，使用计算机视觉技术进行辅助诊断不仅提高了诊断效率，还提升了诊断精度，因此许多方法被提出用于医学影像分割任务。多年来，传统的 ML（机器学习）方法在研究中被广泛应用，如基于图割（Graph Cut）的方法 [4]、随机森林（Random Forest）[5]、支持向量机（Support Vector Machine, SVM）[6] 以及条件随机场（Conditional Random Field, CRF）[7] 等。这些方法依赖人工设计的特征提取器，在处理复杂医学图像时存在局限性，由于图像存在模糊、噪声、对比度低等问题，传统方法的准确性和鲁棒性受到限制。

近年来，随着 FCN(Full Convolution Network) [9]、CNN(Convolution Natural Network) [8] 的提出，深度学习方法逐渐代替了传统机器学习方法，语义分割方法的网络深度、特征提取均有了较大提高。随着 Unet [10] 的提出，基于深度学习的分割方法占据了主导地位，通过创新性的 U 型架构，使其成为了语义分割的主流框架，涌现了 Unet++ [?]、Attention-Unet [?] 等方法。2017 年，Google 提出的 Transformer [11] 在自然语言处理（NLP）领域 have made 深远影响，which 为视觉领域带来了

ViT(Vision Transformer) [12] in 2020，使得计算机视觉领域拥有了全局注意力（Global Attention）与自注意力机制（Self-Attention）[11]，makes better for 全局语义信息提取，改善了 CNN 的不足 [?]。\*\* 等人提出了 TransUnet [?]，将 Unet 与 CNN 和 Transformer 结合起来，使得医学图像分割方法的精度大大提高。

目前，主流的方法主要遵循 TransUnet 的思路，在上述深度学习方法中作修改 [?]，将 CNN 作为局部注意力提取 [?]，为了进行更高精度的分辨，将 Transformer 作为把握全局特征的模块，采用类似 Unet 的上下采样与 skip connection 的结构，达到期望的分割效果 [?]。然而，医学图像具有低对比度、高分辨率、目标边界模糊等诸多问题 [?]，对于自注意力机制来说缺乏友好，往往需要花费大量的参数来提取低对比度医学特征，牺牲了时间与空间性能 [?]，同时自注意力机制具有平方复杂度，高分辨率的医学图像会使计算复杂度大大增加，高参数网络也不适宜部署在医院等边缘设备，这些限制了 Transformer 在医学图像上的直接应用 [?] [?]

一些研究人员研究改善这种情况。\*\*\*\* 等人提出了 Linear Transformer [?]，\*\* 等人在 20xx 年提出了 Mamba [?] 模型，不同于 Transformer 的框架，他们均通过将自注意力机制的复杂度改为线性，从而缓解计算复杂度问题。然而，事实证明，线性复杂度的 Transformer 变体总会牺牲分割精度与准确率，而 Mamba 模型的长序列建模（SSM）也不适用于视觉问题 [?]。鉴于医学影像中 3D 体数据、MRI 与 CT 等高分辨率数据普遍存在，且存在高精度的像素级分割要求，实践中还需要考虑医院边缘设备的部署问题，如何平衡精度与效率、参数与性能，成为急需解决的问题。

Recurrent Weighted Key-Value(RWKV) [?] 的出现引起了我们的关注。其在 NLP 领域的线性复杂度注意力机制，不同于 Transformer 的结构，使其有价值迁移到视觉任务上。Vision RWKV(VRWKV) [?] 尝试了该问题，并针对图像输入进行了结构改进，展现出优异的计算效率与模型精度，\*\* 等人提出了 BSBP-RWKV [?]，首次将 RWKV 应用于医学图像任务，一些研究如 RWKV-Unet [?] 将 VRWKV 与 Unet 迁移到医学图像分割任务上。在高分辨率的医学图像任务下，RWKV 优于其他 Transformer 变体与 Mamba 模型，处理医学图像输入时的推理速度更高，且效果更好 [?]。

作为一款线性注意力模型，尽管 RWKV 拥有良好的计算效率与精度，但迁移到图像任务上，还是有不可避免的问题。RWKV 本质是沿序列建模的状态空间式结构，针对离散数据（如自然语言处理）的处理过程是固定的，自注意

力的处理方式也是一维的。面对连续且二维的图像，基于 RWKV 的图像处理方法是图像分块，并延展为一维，导致了空间连续性上的破坏 [?]。

在注意力提取过程中，RWKV 使用一维的方式在二维图像上进行特征提取，我们称其为贪吃蛇效应 ??。RWKV 的扫描方式是固定的，在图像任务中，注意力的权重往往动态且多变，固定的扫描方式往往会将注意力分片 [?]。(如图 ??)。若提取到更深层次，图像的特征更容易模糊，RWKV 容易将模糊的特征边缘分离，进行分片扫描，容易导致器官边缘、细小病灶（如微小息肉、早期病变区）分割不精细，而这是医学图像分割的要求之一（如图 ??）。同时，医学方向涉及三维数据，在三维医学图像中，边界分割、注意力分片的现象将会更为突出，这为 RWKV 向医学领域的适配带来了挑战 [?]。

为此，很多工作进行了改进。U-RWKV 提出了方向自适应 RWKV 模块 [?]，改进了 RWKV 的扫描方式，使其不仅仅局限于一维层面的注意力叠加。Zig-RiR [?] 提出了 ZigZag 的扫描方案，试图解决 RWKV 扫描的空间连续性问题。然而，其带来了诸多问题：扫描缺乏对图像的动态调整，导致其训练过程不稳定，且不同器官之间的精度存在较大差异 ??；同时，贪吃蛇效应仍然存在；两者工作的上采样过于简单，导致恢复特征分辨率时的精度不足，无法满足医学图像像素级分割的要求。

我们的研究致力于解决该问题。受到 Vision RWKV [?] 以及 URWKV [?] 的启发，我们提出了一种基于 RWKV 的 U 型架构 Stage Scan Segmentation，该架构具备完备的编码器-解码器，在保持了 RWKV 的长程依赖和线性复杂度的同时，增强了局部细节的表达与分割的精度。具体来说，我们采用了 Bi-WKV，一种线性注意力，作为我们的注意力核心 [?]，并提出了 stage scan，一种基于 encoder 层次的扫描机制，根据不同层次的编码器动态地调整扫描方向与权重。为了解决 RWKV 的空间连续性问题，SSS 将一个维度的模块设置为全局 block 与局部 block，分别提取全局与局部特征。在一开始，输入特征的分辨率与 H、W 较大，特征比较分散，噪声较多，需要我们进行更多的分块扫描。随着编码器的 Stage 增加，输入特征的细节将会更多更集中，图像的分块数也会随着层数的增加而减少 ??，以此把握全局的注意力。在一个模块的前端，我们使用全局 block，较大的 Stage 分块，进行初步的特征提取。随后，采用残差机制 [?] 将提取后的全局权重与原特征加和。局部 block 接收加和后的权重，将 stage 分块进一步增大，进行局部的特征提取。同时，为了缓解贪吃蛇效应 ??，我们改进了四项 token 位移 [?]，根据已确定的扫描机制，动态调整扫描位移的方向。低参数量 ??。其次，我们对 RWKV 的 WKV 进行适当池化 [?]，实现进一步的轻量化。最后，我们在上采样阶段实现动态上采样 [?]，而不仅仅进行反卷积，进一步保留了分割结果的精度与细节。通过这种方式，SSS 获得优于其他方法的特征捕捉能力 ??，增强了对边缘的分割能力和鲁棒性，并发挥了 RWKV 应有的计算效率与低参数量。

总的来说，我们的贡献有：

1. 使用 VRWKV 作为分割的核心方法，达到了线性复杂度与注意力精度的平衡；
2. 提出了基于 Stage 的 Scan 方法，缓解了贪吃蛇效应，and 改进了上采样，有效提取全局特征与局部特征；
3. 在对多个 2d 以及 3d 数据集中，SSS 的效果优于目

前许多最先进的方法，同时参数量降低了 20%，GPU 占用降低了 23% ??。

待写：上采样问题轻量化问题所谓的动态？

加入 RWKV 扫描后的热力图，表示注意力在图像上的衰减

## II. RELATED WORKS

### A. 医学图像分割

目前，基于深度学习的医学图像分割网络是主流，且很多的先进方法均基于 CNN 和 Transformer 的变体得来。

1. 纯净的 CNN。最具代表性的为 Unet [10]。UNet 的优势在于编码器-解码器的结构，并通过跳跃连接有效减少信息在传输过程中的损失，实现对复杂医学图像中细微结构的精准分割。部分研究者对 Unet 进行了改进，例如 UNet++、UNet3+、3D UNET 和 Attention Unet 等等。但是，卷积核的有限感受野难以捕捉远程关系和远程关系，限制了模型的分割性能。同时，在面对大尺寸医学图像或三维图像时，UNet 的性能往往不足，导致细节的丢失和全局能力的不足。

2. 纯净的 Transformer。Transformer 的核心机制 self-attention 使其能够把握全局上下文关系，该机制被大量应用至计算机视觉领域，例如 ViT。Transformer 克服了 CNN 在处理长程依赖关系时的不足，并将每个图像块作为输入序列传递到 Transformer 编码器，以获得图像的全局表示。Liu [?] 等人设计了 SwinTransformer 架构，该架构针对图像分割和目标检测任务设计，并在医学影像分割上取得了不错的成绩。Cao [?] 等人将 Unet 与 Transformer 架构融合，设计了 Swin-Unet，提升了医学影像的分割精度和泛化能力。Lin [?] 等人进行了进一步的改进，采用双尺度编码器子网来提取不同语义尺度的特征，建立不同尺度之间的依赖关系。采用纯净 Transformer 的模型对医学图像的精度和鲁棒性较好，但其复杂度较高，精细局部特征提取能力不足，限制了其在医学图像中的应用。

3. 将 CNN 与 Transformer 结合，利用 CNN 在捕捉局部特征上的优越性以及 Transformer 对全局依赖性的处理能力。具有代表性的是 TransUNet [?]、CoTr [?]、UTNet [?] 和 Transfuse [?] 网络等。这些网络通过各种串行、并行的方式将 CNN 与 Transformer 结合到一起，整合了全局与局部特征，使精度进一步提高。该方法的核心问题是：模型计算复杂度较高。Transformer 模块处理图像任务需要将图像序列化，形成较长的序列。尽管 ViT 提出了图像块序列，但图像切割后计算量仍然较大；混合模型需要同时维护 UNet 的卷积层和高复杂度的自注意力机制，虽然增强了细节的特征提取能力，但模型的参数和推理速度均显著降低，无法部署在医院的大量边缘及低算力设备中，应用有限。

改进 [?]。

### B. 线性注意力

线性注意力致力于打破二次方级的注意力复杂度瓶颈，使注意力机制更加高效，并催生出了一系列 Transformer 变体和非 Transformer 架构。

1. Transformer 变体。一些针对于 Transformer 的改进如 Linear Transformer [?]，提出了特征映射的概念，用以替代 Softmax，将 Attention 机制视为 Linear RNN 的过程。\*\* 等人提出了 Performer [?]，引入了正交随机特征

(Orthogonal Random Features, ORF), 使其接近 Softmax 的效果。二者均通过对 softmax 函数效果的改进, 进而达到线性注意力的目的。以此为基础, Han [?] 等人提出了 Flatten Transformer, 设计了一种聚焦线性注意力 (Focused Linear Attention), 改善了空间特征的完整性, 完善了线性注意力在视觉领域的应用。

2. *the new wave*. 如经典的 mamba [?] 模型, 其引入了选择性扫描机制 (Selective Scan), 让模型能够根据输入内容动态压缩记忆。U-Mamba [?], Swin-UMamba [?], VM-UNet [?] 等方法探索了 Mamba 模型在医学图像分割领域的运用, 将 Mamba 模型与 SSM 的思想整合进了分割领域当中, 并进行了针对性的优化。然而, SSM 的核心机制致使其无法有效建模初窗口之外的全局上下文信息, 基于 SSM 的视觉模型与最先进的卷积模型和基于注意力的模型相比, 表现不佳, 导致其并不适合视觉任务 [?], 且其运行速度往往以准确率为代价。

### C. RWKV

最近提出的 Receptance Weighted Key Value (RWKV) [?] 是不同于 Transformer 的线性架构, 擅长处理长序列关系与自回归任务, 在 NLP 领域效果优良, 并催生出在视觉领域的一系列运用 [?] [?]. 同时, 一些研究人员, 也探索了 RWKV 在医学分割领域的运用 [?] [?] [?].

以此为契机, 我们继续完善专精于医学分割的 RWKV。经过研究, 我们发现, RWKV 不适宜直接运用到分割任务中, 因为原始的因果感受野的机制, 在图像任务中表现不好, 同时, 各种针对图像分割的改进, 没有缓解贪吃蛇效应??, 这导致边界分割模糊, 没有达到像素级分割的要求??, 全局注意力的上下文关系也比较割裂。为了解决该问题, 我们提出了 Stage scan structure, 一种基于 Stage 的 RWKV 扫描方法, 并将一次处理分为全局 block 和局部 block, 以此, 针对特征信息的大小, 进行精准的提取, 保留了空间连续性。同时, 我们大幅改进了 Q-shift, 四向位移机制, 使其动态调整位移方向, 有效缓解贪吃蛇效应。额外地, 在原有的 Bi-WKV 基础上, 我们利用低分辨率的 WKV 实现进一步的轻量化。

## III. METHODS

### A. Stage Scan Structure

Let  $\mathcal{W} = \{W_l\}_{l=1}^L$  denote the weights of an  $L$ -layer network. Our goal is to find a sparse mask  $M_l$  for each layer such that the remaining weights  $W_l \odot M_l$  retain accuracy.

### B. Low-Resolution WKV

We use the gradient magnitude as the saliency score:

$$s_{ij}^{(l)} = \left| \frac{\partial \mathcal{L}}{\partial w_{ij}^{(l)}} \right|. \quad (1)$$

## IV. EXPERIMENTS

We evaluate DST on CIFAR-10/100 and ImageNet with ResNet-50.

## V. CONCLUSION

We presented a simple yet effective DST framework that dynamically adjusts sparse connectivity during training. Future work includes extending DST to transformer architectures.

TABLE I  
TOP-1 ACCURACY (%) ON CIFAR-10 UNDER DIFFERENT SPARSITIES.

Method	90% sparsity	95% sparsity
Baseline	93.5	91.2
Magnitude [1]	92.8	89.7
DST (ours)	<b>94.1</b>	<b>92.3</b>

## ACKNOWLEDGMENT

This work was supported by the National Natural Science Foundation of China under Grant 62XXXXXX.

## REFERENCES

- [1] S. Han, J. Pool, J. Tran, and W. Dally, "Learning both weights and connections for efficient neural network," *Proc. NIPS*, 2015.
- [2] U. Evci, T. Gale, J. Menick, P. S. Castro, and E. Elsen, "Rigging the lottery: Making all tickets winners," *Proc. ICML*, 2020.

[3]  
[4]  
[5]  
[6]  
[7]  
[8]  
[9]  
[10]  
[11]  
[12]