Restore-RWKV: Efficient and Effective Medical Image Restoration with RWKV

Zhiwen Yang[1], Hui Zhang[2], Dan Zhao[3], Bingzheng Wei[4], Yan Xu*[1]

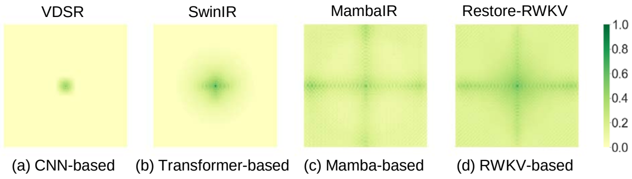
[1]Beihang University [2]Tsinghua University [3]Peking Union Medical College [4]ByteDance Inc.

Abstract

Transformers have revolutionized medical image restoration, but the quadratic complexity still poses limitations for their application to high-resolution medical images. The recent advent of RWKV in the NLP field has attracted much attention as it can process long sequences efficiently. To leverage its advanced design, we propose Restore-RWKV, the first RWKV-based model for medical image restoration. Since the original RWKV model is designed for 1D sequences, we make two necessary modifications for modeling spatial relations in 2D images. First, we present a recurrent WKV (Re-WKV) attention mechanism that captures global dependencies with linear computational complexity. Re-WKV incorporates bidirectional attention as basic for a global receptive field and recurrent attention to effectively model 2D dependencies from various scan directions. Second, we develop an omnidirectional token shift (Omni-Shift) layer that enhances local dependencies by shifting tokens from all directions and across a wide context range. These adaptations make the proposed Restore-RWKV an efficient and effective model for medical image restoration. Extensive experiments demonstrate that Restore-RWKV achieves superior performance across various medical image restoration tasks, including MRI image super-resolution, CT image denoising, PET image synthesis, and all-in-one medical image restoration. Code is available at: https://github.com/Yaziwel/Restore-RWKV.

1. Introduction

Medical image restoration (MedIR) aims at recovering the high-quality (HQ) medical image from its degraded low-quality (LQ) counterpart. It encompasses a variety of tasks such as MRI image super-resolution [1–7], CT image denoising [8–10], PET image synthesis [11–15], and all-in-one medical image restoration [16–19]. This field is particularly challenging due to the ill-posed nature of the problem, where crucial information about the image content is often lost in LQ images. Recent years have witnessed significant progress in MedIR, largely driven by advanced deep learning models such as convolutional neural networks (CNNs) [20, 21], Transformers [22, 23], and Mambas [24–26].



(a) CNN-based (b) Transformer-based (c) Mamba-based (d) RWKV-based

Figure 1. The effective receptive field (ERF) [27] visualization for different models. A more extensively distributed dark area indicates a larger ERF. Our proposed Restore-RWKV achieves the most significant global ERF.

In spite of significant advancements, current CNN-based, Transformer-based, and Mamba-based models still exhibit inherent limitations for accurate medical image restoration. CNN-based models [2] often face constraints in their effective receptive field [27] (ERF, as illustrated in Figure 1 (a)) due to the limited kernel size of the convolution operator. This limitation hampers their ability to capture information across broader ranges to compensate for lost details in degraded pixels. In contrast, Transformer-based models excel in modeling long-range dependencies and achieving a global ERF through self-attention [23]. However, the computational complexity of self-attention grows quadratically with spatial resolution, rendering standard Transformers impractical for high-resolution medical images. To mitigate this, many other Transformer-based models aim to reduce computational demands by restricting attention to smaller windows [28], inevitably diminishing the ERF (as depicted in Figure 1 (b)) and conflicting with the goal of capturing long-range dependencies. Mamba-based models [7, 29] serve as an efficient alternative to Transformers, leveraging state space models (SSMs) [24] to causally capture long-range dependencies with linear computational complexity. Recent studies suggest that Mamba-based models can achieve performance comparable to or even better than Transformer-based models. However, their inherent unidirectional sequence modeling in SSMs

*Corresponding author (xuyan04@gmail.com) still poses challenges in achieving an optimal ERF for 2D images (as shown in Figure 1 (c)). In summary, existing models in MedIR struggle to strike a good balance between computational efficiency and achieving an effective global receptive field.

The **R**eceptance **W**eighted **K**ey **V**alue (RWKV) [30, 31] model, originating from natural language processing (NLP), is emerging as an alternative to the Transformer. RWKV [30] introduces two significant innovations. First, it introduces a WKV attention mechanism that achieves long-range dependencies with linear computational complexity, addressing the expensive computational cost of self-attention in Transformers. Second, it incorporates a token shift layer to enhance the capture of local dependencies, which is typically overlooked by the standard Transformer [22, 23]. With these advanced designs, RWKV demonstrates strong capacity and scalability, performing admirably with large-scale NLP [30, 31] and vision datasets [32–36]. Recent studies [31, 34, 36] have shown that RWKV-based models outperform Transformer-based models and linearly scaled Mamba-based models in both effectiveness and efficiency. Despite pioneering efforts to adopt RWKV for vision tasks, it is still in the initial stages of extending the original RWKV from modeling 1D sequences to modeling 2D images, and the adaptation of RWKV to the field of MedIR still remains unexplored.

In this paper, we propose Restore-RWKV, an efficient and effective model adapted from RWKV for medical image restoration. The original RWKV [30] is designed for handling 1D sequences and has limitations in capturing spatial dependencies in 2D medical images. To address these limitations, we make essential modifications to two key components of RWKV—the attention layer and the token shift layer—to adapt it for processing 2D medical images. (1) We introduce a recurrent WKV (Re-WKV) attention mechanism that effectively captures global dependencies with linear computational complexity. Unlike the causal WKV attention in the original RWKV, which is unidirectional and has a limited receptive field, Re-WKV utilizes a bidirectional attention mechanism to achieve a global receptive field. Additionally, it employs a recurrent attention mechanism that effectively models 2D dependencies across various scan directions. (2) We develop an omnidirectional token shift (Omni-Shift) layer to accurately capture local dependencies. Previous token shift layers only shift adjacent tokens from limited directions using simple interpolation. In contrast, Omni-Shift enhances interaction among neighbors by shifting tokens from all directions and over a large context range through efficient depthwise convolution. Furthermore, during training, Omni-Shift employs a structural re-parameterization strategy to learn to accurately shift tokens from various context ranges, while maintaining the original structure for efficiency during testing. Extensive experimental results indicate that our proposed Restore-RWKV can achieve superior performance with good efficiency, serving as a general restoration backbone for various tasks, including MRI image super-resolution, CT image denoising, PET image synthesis, and all-in-one medical image restoration.

Our main contribution can be summarized as follows:

- We propose Restore-RWKV, which pioneers the adaptation of the RWKV model for medical image restoration. It has proven an efficient and effective alternative for medical image restoration backbones.
- We present a recurrent WKV (Re-WKV) attention mechanism that effectively captures global dependencies in high-resolution medical images with linear computational complexity.
- We develop an omnidirectional token shift (Omni-Shift) layer, enhancing local dependencies by establishing accurate token interactions from all directions.

2. Related Work

2.1. Medical Image Restoration

Medical image restoration (MedIR) aims to recover high-quality (HQ) medical images from their degraded low-quality (LQ) counterparts. Typical MedIR tasks include MRI image super-resolution [1–7], CT image denoising [8–10], PET image synthesis [11–15], and all-in-one medical image restoration for multi-task MedIR [16–19]. Recently, deep learning-based models have become the primary approach for addressing MedIR tasks. These methods can be broadly categorized into CNN-based models, Transformer-based models, and Mamba-based models.

CNN-based Models. Over the past decades, CNN-based models [4, 5, 8, 11, 13] have made extraordinary contributions to the field of MedIR due to their strong representation capabilities. A key advantage of CNN-based models for MedIR is their ability to capture local context and build local dependencies through convolution. This allows the neighborhood of a degraded pixel to be used as a reference to aid in its recovery. However, due to the limited kernel size of convolutions, CNN-based methods often suffer from restricted receptive fields and cannot model long-range dependencies. While some approaches have tried to expand the receptive field by designing deeper networks [2, 3] or utilizing the multi-scale U-Net [8, 13] architecture, these improvements are still limited and struggle to capture long-range dependencies effectively.

Transformer-based Models. The Transformer model [22, 23], originally developed for natural language processing (NLP), has been adapted to achieve significant performance in vision tasks, including MedIR. Compared to CNN-based models, Transformer-based models excel at building long-range dependencies through global self-attention. This capability allows Transformer-based models to better utilize information from the entire image to compensate for the information lost in degraded pixels. However, the standard Transformer architecture suffers from quadratic computational complexity due to its self-attention mechanism, limiting its practicality for MedIR tasks involving high-resolution images. To address this issue, several approaches [28, 37, 38] have explored efficient attention mechanisms by limiting the scope of token interactions in attention calculation, such as window attention [28] and sparse attention [37]. Nevertheless, these efficient attention mechanisms often come at the cost of a reduced global receptive field and struggle to balance the trade-off between accuracy and efficiency.

Mamba-based Models. Mamba [24–26], a sequence model backbone grounded in state-space models (SSMs), has emerged as a prominent alternative to Transformers. Unlike Transformers, which exhibit quadratic computational complexity, Mamba-based models can model long-range dependencies with linear computational complexity. This capability allows Mamba-based models to directly capture global dependencies in high-resolution medical images through SSMs without the need for efficient designs like window partitioning. Recent studies have shown that Mamba-based models [7, 29] outperform Transformer-based models in MedIR tasks. However, due to the unidirectional nature of SSM, Mamba's global dependency modeling capability still has limitations in achieving accurate results for MedIR applications.

2.2. Receptance Weighted Key Value

The **R**eceptance **W**eighted **K**ey **V**alue (RWKV) model [30, 31], which originates from the field of NLP, is an emerging efficient alternative to Transformers. Compared to the standard Transformer [22], RWKV offers two significant advantages. First, RWKV proposes a WKV attention mechanism to build long-range dependencies with linear computational complexity, addressing the quadratic computational complexity of self-attention in Transformers. Second, RWKV introduces a token shift layer to capture the local context, which is often ignored by the standard Transformer. Recent study [31] indicates that RWKV can achieve performance comparable to or even better than that of both Transformers and Mamba in NLP tasks. Recently, Vision-RWKV [32] has transferred RWKV from NLP to vision tasks, demonstrating superior performance compared to vision Transformers with reduced computational complexity. To better utilize spatial information in 2D images, Vision-RWKV proposes a bidirectional WKV attention mechanism to capture global dependencies and a quad-directional token shift mechanism to capture local context from four different directions. Building on RWKV and Vision-RWKV, several RWKV-based models have been developed for various vision-related tasks, such as Diffusion-RWKV [33] for image generation, RWKV-SAM [36] for segment anything, Point-RWKV [34] for 3D point cloud learning, and RWKV-CLIP [35] for vision-language representation learning. However, few studies have validated the performance of RWKV in low-level vision tasks like medical image restoration. In this paper, we reveal that current RWKV-based models still have limitations in building both global and local dependencies in 2D images, hindering their performance in MedIR tasks. To address this, we propose Restore-RWKV, which efficiently and effectively models both global and local dependencies, achieving superior performance in restoring medical images.

3. Method

We aim to build an efficient and effective model for medical image restoration based on the **R**eceptance **W**eighted **K**ey **V**alue (RWKV) [30], termed Restore-RWKV. In Sec. 3.1, we introduce the network architecture of Restore-RWKV, and in Sec. 3.2, we present a novel R-RWKV block as its fundamental component for feature extraction. Since the original RWKV [30] is designed for processing 1D sequences, we incorporate two innovations in the R-RWKV block to adapt it for capturing global and local dependencies in 2D images. First, in Sec. 3.3, we introduce a recurrent WKV (Re-WKV) attention mechanism that effectively builds global dependencies in 2D images with linear complexity. Second, in Sec. 3.4, we introduce an omnidirectional token shift that accurately captures local context to enhance the model performance. We will elaborate on the details of Restore-RWKV in the following section.

3.1. Restore-RWKV Architecture

As shown in Figure 2 (a), the proposed Restore-RWKV is a 4-level U-shaped encoder-decoder architecture with 3 times downsampling and upsampling, which enjoys the advantage of capturing image features at different hierarchies and computational efficiency [39]. Given an input low-quality (LQ) image $I_{LQ} \in \mathbb{R}^{H\times W\times 1}$, Restore-RWKV first employs a 3×3 convolutional layer as an input projection to project the input image to a shallow feature $F_S \in \mathbb{R}^{H\times W\times C}$, where H, W, and C denote the height, width, and channel, respectively. Then the shallow feature F_S undergoes a 4-level hierarchical encoder-decoder and is transformed into a deep feature $F_D \in \mathbb{R}^{H\times W\times 2C}$. Each level of the encoder-decoder comprises N_i, where $i \in \{1, 2, 3, 4\}$, R-RWKV blocks for feature extraction, with an additional $N_{\text{refinement}}$ R-RWKV blocks used for feature refinement in the final decoder level. For feature downsampling, Restore-RWKV utilizes a 1×1 convolutional layer and a pixel-unshuffle operation [40], reducing the spatial size by half

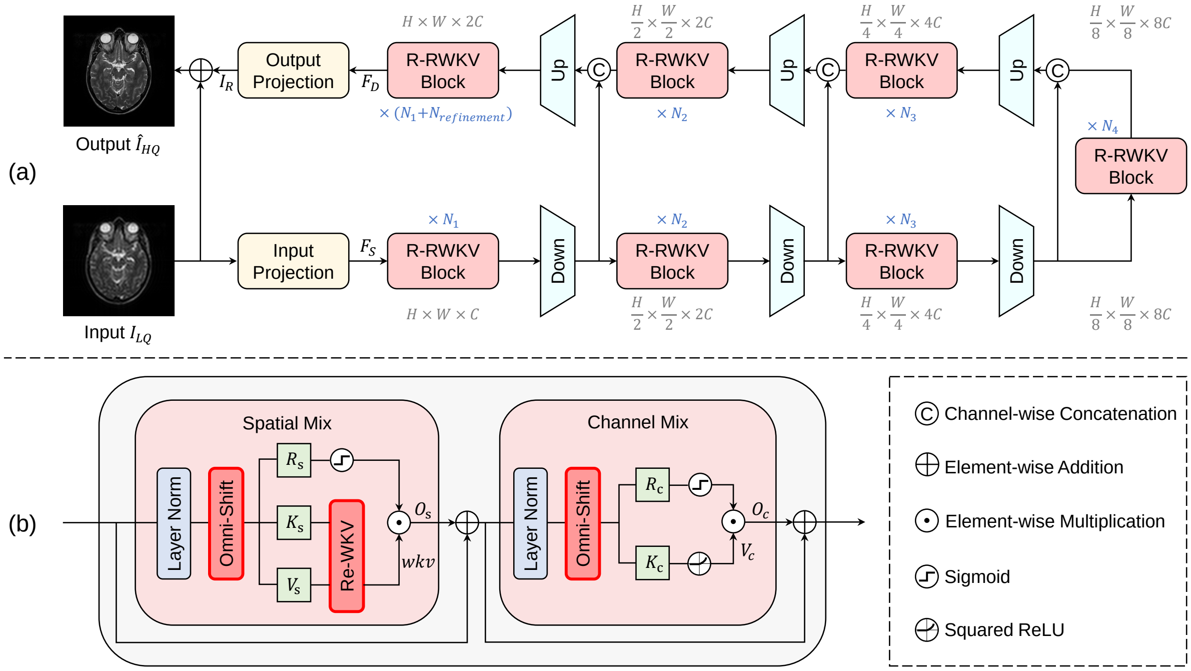


Figure 2. (a) Overview of the Restore-RWKV architecture. (b) Illustration of the R-RWKV block, which incorporates a Re-WKV attention mechanism to model global dependencies with linear complexity, and an Omni-Shift layer to capture local context.

and doubling the channel number. For feature upsampling, Restore-RWKV employs a pixel-shuffle operation [40] and a 1×1 convolutional layer, doubling the spatial size and halving the channel number. To aid in the restoration process, encoder features are concatenated with decoder features via skip connections. The deep feature F_D is then projected out to a residual image $I_R \in \mathbb{R}^{H\times W\times 1}$ by a 3×3 convolutional layer. Finally, the restored image $\hat{I}_{HQ}$ can be obtained by adding the input I_{LQ} and the residual image I_R: $\hat{I}_{HQ} = I_{LQ} + I_R$.

3.2. R-RWKV Block

The R-RWKV blocks play a crucial role in feature extraction across different hierarchical levels within Restore-RWKV. As shown in Fig. 2 (b), our proposed R-RWKV block follows the design of the original RWKV block [30] and integrates a spatial mix module and a channel mix module, enabling spatial-wise token interaction and channel-wise feature fusion, respectively. Since the original RWKV block [30] is tailored for processing 1D sequences and lacks comprehensive capability in capturing global and local contexts in 2D images, our R-RWKV block introduces two innovations to tackle the increased dimensionality of 2D images: recurrent WKV attention (Re-WKV) for capturing global dependencies and omnidirectional token shift (Omni-Shift) for capturing local context. The data flow in R-RWKV is detailed as follows.

Spatial Mix. The spatial mix module is designed to build long-range dependencies within tokens across the spatial dimension. Given an input feature flattened to a one-dimensional sequence $X \in \mathbb{R}^{T\times C}$, where $T = H \times W$ denotes the total number of tokens, the spatial mix module first passes it through a layer normalization (LN) and an Omni-Shift layer (refer to Section 3.4):

$$X_s = \text{Omni-Shift}(\text{LN}(X)). \tag{1}$$

Here, LN is applied to stabilize the training process. Our proposed Omni-Shift is specifically introduced to capture local context and expand the context range of individual tokens. Then X_s is passed through three parallel linear projection layers to obtain the matrices of receptance $R_s \in \mathbb{R}^{T\times C}$, key $K_s \in \mathbb{R}^{T\times C}$, and value $V_s \in \mathbb{R}^{T\times C}$:

$$R_s = X_s W_{R_s}, \quad K_s = X_s W_{K_s}, \quad V_s = X_s W_{V_s}, \tag{2}$$

where W_{R_s}, W_{K_s}, and W_{V_s} represent the three linear projection layers. Subsequently, K_s and V_s are utilized to acquire the global attention result $wkv \in \mathbb{R}^{T\times C}$ by our proposed linear-complexity Re-WKV attention mechanism (as detailed in Section 3.4):

$$wkv = \text{Re-WKV}(K_s, V_s). \tag{3}$$

Finally, the receptance after gating $\sigma(R_s)$ modulates the received probability of the attention result wkv through element-wise multiplication:

$$O_s = (\sigma(R_s) \odot wkv)W_{O_s}, \tag{4}$$

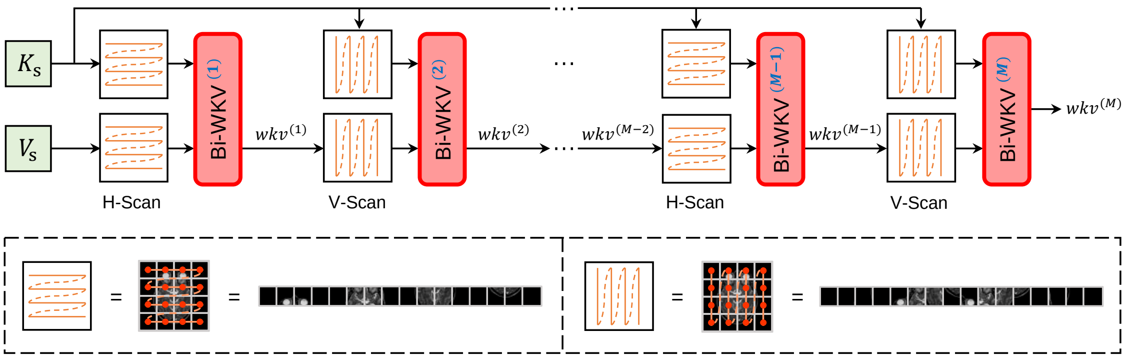


Figure 3. Illustrations of the Re-WKV attention mechanism. Re-WKV employs Bi-WKV [32] as its basic attention operator and applies Bi-WKV attention to 2D images recurrently through various scan directions to better model global dependencies.

where O_s denotes the output, $\sigma(\cdot)$ represents the Sigmoid gating function, and W_{O_s} signifies the linear projection layer for output projection.

Channel Mix. The channel mix module aims at performing feature fusion in the channel dimension. Given an input feature $X \in \mathbb{R}^{T \times C}$, the channel mix passes it through LN and Omni-Shift layer as spatial mix:

$$X_c = \text{Omni-Shift}(\text{LN}(X)). \tag{5}$$

Then the receptance $R_c \in \mathbb{R}^{T \times C}$, key $K_c \in \mathbb{R}^{T \times C}$, and value $V_c \in \mathbb{R}^{T \times C}$ can be acquired as follows:

$$R_c = X_c W_{R_c}, \quad K_c = X_c W_{K_c}, \quad V_c = \gamma(K_c) W_{V_c}. \tag{6}$$

Here, W_{R_c}, W_{K_c}, and W_{V_c} represent the three linear projection layers. $\gamma(\cdot)$ denotes the squared ReLU activation function, known for its enhanced nonlinearity. Notably, V_c is estimated from K_c rather than directly from X_c, which differs from the approach used in the spatial mix module. In fact, the transformation from X_c to K_c to V_c involves a multi-layer perception (MLP) consisting of W_{K_c}, $\gamma(\cdot)$, and W_{V_c}, facilitating channel-wise feature fusion.

Finally, the output O_c is derived by multiplying V_c with $\sigma(R_c)$ to control the received probability of V_c:

$$O_c = (\sigma(R_c) \odot V_c) W_{O_c}, \tag{7}$$

where W_{O_c} signifies the linear projection layer for output projection.

Differences with RWKV Block. Our principle is to retain the advantages of the original RWKV block architecture [30], which is initially designed for processing 1D sequences, while making essential modifications to accommodate modeling spatial relationships in the context of 2D images. Consequently, the proposed R-RWKV block is built by replacing only the attention layer and token shift layer of the original RWKV block with our proposed Re-WKV attention and Omni-Shift, keeping other layers unchanged. This modification allows R-RWKV to effectively capture global (via Re-WKV) and local (via Omni-Shift) dependencies within the spatial dimensions of 2D images.

3.3. Recurrent WKV Attention

The WKV attention mechanism is the core component in the spatial mix to achieve long-range dependencies with linear computational complexity, addressing the quadratic complexity of the standard self-attention in Transformers. However, the original WKV attention in RWKV [30] is unidirectional (Uni-WKV) and has a limited receptive field confined to the scanned part of sequential data. While this characteristic is well-suited for 1D causal sequence modeling in natural language, it faces challenges when applied to non-causal 2D images that require global modeling. To address these challenges, we propose a recurrent WKV attention (Re-WKV) mechanism for processing 2D images. Re-WKV integrates a basic bidirectional attention mechanism to achieve a global receptive field and a recurrent attention mechanism to better adapt the basic bidirectional attention for modeling 2D image dependencies from multiple scan directions. We introduce the proposed Re-WKV by sequentially detailing its bidirectional and recurrent attention mechanisms.

Bidirectional Attention. To address the limited receptive field of Uni-WKV attention in RWKV [30], which spans from the first token to only the current token in a sequence, we follow Vision-RWKV [32] by adopting a bidirectional WKV (Bi-WKV) attention mechanism that ensures a global receptive field spanning from the first token to the last token. Given the input projections of key K_s and value V_s, the attention result of the current t-th token (denoted as $wkv_t \in \mathbb{R}^C$) can be formulated as follows:

$$\begin{aligned} wkv_t &= \text{Bi-WKV}(K_s, V_s)_t \\ &= \frac{\sum_{i=1, i\neq t}^{T} e^{-(|t-i|-1)/T \cdot w + k_i} v_i + e^{u+k_t} v_t}{\sum_{i=1, i\neq t}^{T} e^{-(|t-i|-1)/T \cdot w + k_i} + e^{u+k_t}}. \end{aligned} \tag{8}$$

Here, T denotes the total number of tokens. $k_i \in \mathbb{R}^C$ and

$v_i \in \mathbb{R}^C$ indicate i-th spatial token of K_s and V_s, respectively. $-(|t-i|-1)/T$ represents the relative position bias between the t-th and i-th tokens, with a learnable parameter $w \in \mathbb{R}^C$ controlling its magnitude. The learnable parameter $u \in \mathbb{R}^C$ is a special case of w, giving a bonus to the current t-th token. To sum up, Eq. 8 indicates that the attention result wkv_t of the t-th token is a weighted sum of V_s along the token dimension from 1 to T, with summation weight being collectively determined by relative position bias $-(|t-i|-1)/T$ and key k_i.

Bi-WKV ensures both global receptive field and computational efficiency. On the one hand, since the attention result for each token is determined by all other tokens, Bi-WKV guarantees a global receptive field. On the other hand, it eliminates the query-key matrix multiplication, thus avoiding the quadratic computational complexity inherent in standard self-attention mechanisms. According to Vision-RWKV [32], given the input K_s and V_s with the shape of $T \times C$, the practical computational complexity of Bi-WKV is $O(T \times C)$, which scales linearly with the number of tokens T.

Recurrent Attention. Bi-WKV attention can still be challenging to apply to 2D images due to its direction-sensitive nature. According to Eq. 8, Bi-WKV is partially determined by the relative position bias between tokens, indicating that Bi-WKV can be sensitive to the arrangement order of sequential tokens. However, the sequential order of 2D image tokens can vary with different scan directions. Therefore, previous approaches that use a single-direction scan cannot effectively model the dependencies in 2D images. To overcome this limitation, we propose a recurrent WKV attention (Re-WKV) that applies Bi-WKV attention along different scan directions. The mechanism of our proposed Re-WKV is illustrated in Fig. 3, and its core can be formulated as follows:

$$wkv^{(j)} = \text{Bi-WKV}^{(j)}(\Delta_{dir}(K_s), \Delta_{dir}(wkv^{(j-1)})). \quad (9)$$

Here, $\text{Bi-WKV}^{(j)}(\cdot)$ denotes the j-th Bi-WKV attention. $\Delta_{dir}(\cdot)$ represents the changing direction operation. In two adjacent Bi-WKV attention operations, we alternately use two different scan directions: horizontal scan (H-Scan) and vertical scan (V-Scan). $wkv^{(j)}$ and $wkv^{(j-1)}$ are attention results of the j-th and $(j-1)$-th Bi-WKV, respectively. Note that $wkv^{(0)}$ is the input V_s. Eq. 9 indicates the current j-th Bi-WKV attention takes the previous $(j-1)$-th Bi-WKV attention result as its input value, complementing the $(j-1)$-th attention result from a different scan direction. The final attention result wkv is obtained after applying Bi-WKV attention recurrently M times:

$$wkv = \text{Re-WKV}(K_s, V_s) = wkv^{(M)}. \quad (10)$$

In comparison with Bi-WKV, Re-WKV enhances global token interactions in 2D images by recurrently performing attention along different scan directions. Furthermore, since $M \ll T$ in our implementation, the computational complexity of Re-WKV maintains the linear computational complexity as Bi-WKV.

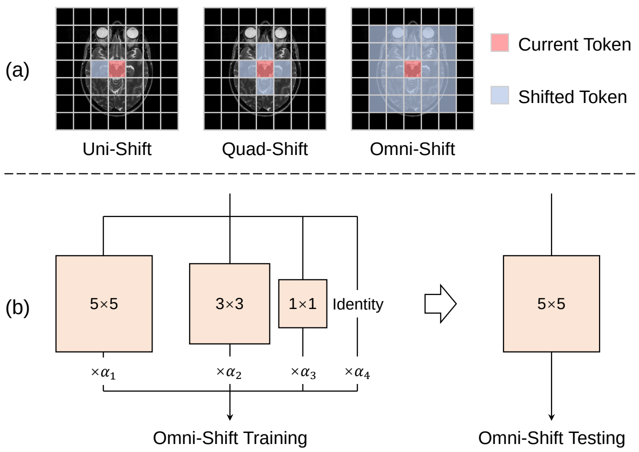


Figure 4. (a) Illustrations of different token shift mechanisms. The Uni-Shift [30] fuses the current token with only the last (left) one by linear interpolation. The Quad-Shift [32] fuses the current token with four adjacent tokens by linear interpolation. Our proposed Omni-Shift fuses the current token with tokens from all directions by convolution. (b) Illustration of the Omni-Shift with structural re-parameterization.

3.4. Omnidirectional Token Shift

The token shift mechanism in RWKV [30] is proposed to capture the local context in a token sequence, where neighboring tokens are assumed to be correlated and share similar context information. It works by shifting neighboring tokens to fuse with individual tokens through simple linear interpolation [30, 32]. However, as shown in Fig. 4 (a), existing token shift mechanisms, such as the unidirectional token shift (Uni-Shift) in RWKV [30] and the quad-directional token shift (Quad-Shift) in Vision-RWKV [32], only shift tokens from limited directions and do not fully exploit the spatial relationships inherent in 2D images, where neighboring tokens in all directions are correlated. To address this issue, we propose an omnidirectional token shift (Omni-Shift) mechanism that shifts and fuses neighboring tokens from all directions using convolution. To achieve an accurate and efficient token shift mechanism for aggregating local context, Omni-Shift employs a structural re-parameterization [41] over convolution, involving a multi-branch structure during the training phase and a single-branch structure during testing. We will elaborate on these details as follows.

Multi-branch Training of Omni-Shift. Considering the differences in importance across various context ranges, Omni-Shift employs a multi-branch structure during training (as shown in Fig. 4 (b)), with each branch responsible for shifting tokens within a specific context range. Given an input feature $X \in \mathbb{R}^{H\times W\times C}$, the mechanism of Omni-Shift can be formulated as follows:

$$\begin{aligned}\text{Omni-Shift}(X) =&\alpha_1\,\text{DConv}_{5\times5}(X) + \alpha_2\,\text{DConv}_{3\times3}(X)\\ &+ \alpha_3\,\text{DConv}_{1\times1}(X) + \alpha_4 X,\end{aligned} \tag{11}$$

where α_i denotes a learnable parameter for scaling the specific branch. $\text{DConv}_{k\times k}(\cdot)$ denotes the efficient depth-wise convolution with kernel size of k. The final result of Omni-Shift is a fusion of four branches, with each branch specializing in a specific context range, resulting in an accurate token shift result.

Single-branch Testing of Omni-Shift. A multi-branch design inevitably introduces additional parameters and computational costs compared to a single branch. To mitigate this issue, Omni-Shift adopts the approach used in RepVGG [41], which involves performing structural re-parameterization during testing. Given that the convolution weights of smaller kernels can be integrated with those of larger kernels by zero-padding the weights of the smaller kernels, the multi-branch design of Omni-Shift can be consolidated into a single branch using a convolution kernel size of 5 during testing (as shown in Fig. 4 (b)). This re-parameterization strategy guarantees both accuracy and testing-time efficiency of Omni-Shift.

4. Experimental Setup

4.1. Dataset

We conduct extensive experiments on various MedIR tasks to demonstrate the performance of our proposed Restore-RWKV, including MRI image super-resolution, CT image denoising, PET image synthesis, and an all-in-one medical image restoration task that combines the aforementioned three tasks for multi-task learning. The corresponding datasets for each task are introduced as follows.

MRI Image Super-Resolution. We use the publicly available IXI dataset [1], which consists of 578 HQ T2-weighted 3D MRI images. These images are divided into 405 for training, 59 for validation, and 114 for testing. Each 3D MRI image has dimensions of 256×256×n, from which we extract the central 100 2D slices sized 256×256 to exclude side slices. This results in 40500 high-quality MRI slices for training, 5828 for validation, and 11400 for testing. LQ images are generated by cropping the k-space of HQ images with a downsampling factor of 4×, retaining only 6.25% of the central data points.

CT Image Denoising. We utilize the publicly available AAPM dataset [1], which comprises paired standard-dose HQ CT images and quarter-dose LQ CT images. These images are collected from 10 patients and we divide them into 8 patients for training, 1 for validation, and 1 for testing. Each 3D CT image has a dimension of 512×512×n. Consequently, we extract 2D slices sized 512×512 and obtain 2039 images for training, 128 images for validation, and 211 images for testing.

PET Image Synthesis. We collect HQ PET images from 159 patients using the PolarStar m660 PET/CT system in list mode, with an injection dose of 293MBq ^{18}F-FDG. These 159 HQ PET images were divided into 120 for training, 10 for validation, and 29 for testing. LQ PET images are generated through random list mode decimation with a dose reduction factor of 12. Both HQ and LQ PET images are reconstructed using the standard OSEM [42] method. Each PET image has 3D shapes of 192×192×400 and is divided into 192 2D slices sized 192×400. Then, slices containing only air are excluded. Consequently, we obtain 8350 2D PET images for training, 684 images for validation, and 2044 images for testing.

All-in-One Medical Image Restoration. The goal of all-in-one medical image restoration [18] is to train a universal model capable of handling multiple MedIR tasks. Therefore, the dataset we use is a combination of the datasets from the aforementioned MRI image super-resolution, CT image denoising, and PET image synthesis tasks.

4.2. Implementation

For the network architecture of Restore-RWKV, the number of R-RWKV blocks is set to $N_1 = N_{\text{refinement}} = 4$, $N_2 = N_3 = 6$, and $N_4 = 8$. The number of input channels is $C = 48$, and the number of attention recurrences in Re-WKV is $M = 2$. For model training, we use image patches sized 128×128 with a batch size of 4. We train our proposed Restore-RWKV using the Adam optimizer and L_1 loss for 30K iterations. The learning rate is initialized at $2e^{-4}$ and gradually reduced to $1e^{-6}$ using cosine annealing. All experiments are conducted in PyTorch, utilizing an NVIDIA A100 GPU with 40 GB of memory.

4.3. Evaluation

Three popular metrics are used for quantitative comparisons: Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM), and Root Mean Squared Error (RMSE). Higher values of PSNR and SSIM indicate better performance, while lower values of RMSE signify better results. Additionally, we provide comparisons of model size and FLOPs. The input image size is set to 128×128 when calculating FLOPs.

[1] https://brain-development.org/ixi-dataset/

[1] https://www.aapm.org/GrandChallenge/LowDoseCT/

Table 1. MRI image super-resolution results. The best results are highlighted in **bold** and the second-best results are underlined.

Method	Type	Params (M)	FLOPs (G)	PSNR↑	SSIM↑	RMSE↓
SRCNN	CNN	0.0573	0.9369	28.8067	0.8919	41.3488
VDSR	CNN	0.6647	10.8905	30.0446	0.9140	36.0508
EDSR	CNN	38.3432	628.2145	31.4168	0.9338	31.1390
SwinIR	Transformer	11.4977	187.9274	31.5549	0.9334	30.5788
Restormer	Transformer	26.1241	35.2052	31.8474	0.9378	29.7005
Xformer	Transformer	25.2197	35.7291	31.5697	0.9338	30.5616
MambaIR	Mamba	31.5035	34.3491	31.7677	0.9369	29.8372
Restore-RWKV	RWKV	27.9140	37.4585	**32.0913**	**0.9408**	**28.9713**

Table 2. CT image denoising results. The best results are highlighted in **bold** and the second-best results are underlined.

Method	Type	Params (M)	FLOPs (G)	PSNR↑	SSIM↑	RMSE↓
REDCNN	CNN	1.8489	24.9226	33.1889	0.9113	8.9427
Eformer	Transformer	0.3441	1.2775	33.3496	0.9175	8.8030
CTformer	Transformer	1.4475	3.4437	33.2506	0.9134	8.8974
SwinIR	Transformer	11.4977	187.9274	33.6844	0.9179	8.4650
Restormer	Transformer	26.1241	35.2052	33.6610	0.9182	8.4909
Xformer	Transformer	25.2197	35.7291	33.6903	0.9183	8.4621
MambaIR	Mamba	31.5035	34.3491	33.7274	0.9188	8.4250
Restore-RWKV	RWKV	27.9140	37.4585	**33.7988**	**0.9198**	**8.3600**

Table 3. PET image synthesis results. The best results are highlighted in **bold** and the second-best results are underlined.

Method	Type	Params (M)	FLOPs (G)	PSNR↑	SSIM↑	RMSE↓
Xiang's	CNN	0.2293	3.7654	35.9268	0.9167	0.0980
DCNN	CNN	1.3331	21.8576	36.2710	0.9243	0.0954
ARGAN	CNN	31.1426	8.4807	36.7272	0.9406	0.0902
SwinIR	Transformer	11.4977	187.9274	37.1559	0.9453	0.0868
Restormer	Transformer	26.1241	35.2052	37.1617	0.9460	0.0869
Xformer	Transformer	25.2197	35.7291	37.0885	0.9461	0.0876
MambaIR	Mamba	31.5035	34.3491	37.1815	0.9455	0.0864
Restore-RWKV	RWKV	27.9140	37.4585	**37.3314**	**0.9474**	**0.0852**

5. Experimental Results

5.1. MRI Image Super-Resolution Results

For MRI image super-resolution, we compare our proposed Restore-RWKV with three specific image super-resolution methods: SRCNN [1], VDSR [2], and EDSR [3], as well as four general image restoration methods: SwinIR [28], Restormer [43], Xformer [44], and MambaIR [29]. Table 1 demonstrates that Restore-RWKV achieves the best performance in MRI image super-resolution while maintaining good efficiency. Restore-RWKV attains an effective global receptive field with linear computational complexity, thereby better balancing computational efficiency and model performance compared to all CNN-based, Transformer-based, and Mamba-based methods presented in Table 1. Notably, Restore-RWKV outperforms the second-best method Restormer, by more than 0.20 dB in PSNR. Figure 5 (a) illustrates visualization results across different methods, showing that our Restore-RWKV excels in recovering structures and details in MRI images.

5.2. CT Image Denoising Results

For CT image denoising, we compare our proposed Restore-RWKV with three specific CT image denoising methods: REDCNN [8], Eformer [9], and CTformer [10], as well as the four general image restoration methods: SwinIR [28], Restormer [43], Xformer [44], and MambaIR [29]. Table 2 indicates that Restore-RWKV achieves better CT image denoising performance with good computational efficiency compared to all other methods. Compared to the second-best method MambaIR, it achieves a 0.07 dB increase in PSNR while maintaining comparable computational complexity. Figure 5 (a) shows that the image recovered by Restore-RWKV is sharper and closer to the corresponding HQ image.

5.3. PET Image Synthesis Results

For PET image synthesis, we compare Restore-RWKV with three specific PET image synthesis methods: Xiang's [11], DCNN [12], and ARGAN [13], as well as the four general image restoration methods: SwinIR [28], Restormer [43], Xformer [44], and MambaIR [29]. Table 3 indicates that Restore-RWKV achieves state-of-the-art performance in PET image synthesis. Compared to MambaIR, which achieves the second-best performance in PSNR and RMSE, and Xformer, which achieves the second-best results in SSIM, Restore-RWKV consistently outperforms them in all three evaluation metrics. Figure 5 (a) presents the synthesis results, showing that Restore-RWKV generates more pleasing images with clearer contrast.

5.4. All-in-One Medical Image Restoration Results

The aforementioned experimental results indicate that Restore-RWKV achieves superior performance in individual MedIR tasks. To further evaluate the model's capacity and generalizability, we assess its performance on the all-in-one medical image restoration task, which involves using a single model to handle multiple MedIR tasks. Given the apparent disparities among different MedIR tasks, this all-in-one task is particularly challenging and serves as a robust test of model capacity and generalizability. We compare our proposed Restore-RWKV with three all-in-one image restoration methods: AirNet [16], DRMC [17], and AMIR [18], as well as four general image restoration methods: SwinIR [28], Restormer [43], Xformer [44], and MambaIR [29]. As shown in Table 4, all-in-one methods generally outperform general image restoration methods due to their specifically designed modules that address task differences. Although our Restore-RWKV does not incorporate any specialized modules for handling task disparities, it achieves the second-best results on average, slightly behind the best method, AMIR. Notably, Restore-RWKV even outperforms AMIR in PET image synthesis. These results indicate that Restore-RWKV has a strong capacity and generalizability, making it a robust backbone for medical image restoration.

Table 4. All-in-one medical image restoration results. The best results are highlighted in **bold** and the second-best results are underlined.

Method	Type	Params (M)	FLOPs (G)	MRI Image Super-Resolution			CT Image Denoising			PET Image Synthesis			Average		
				PSNR↑	SSIM↑	RMSE↓	PSNR↑	SSIM↑	RMSE↓	PSNR↑	SSIM↑	RMSE↓	PSNR↑	SSIM↑	RMSE↓
AirNet	CNN	7.6068	230.4830	31.3921	0.9316	31.1141	33.6222	0.9176	8.5226	37.1721	0.9451	**0.0864**	34.0621	0.9314	13.2410
DRMC	Transformer	0.6236	9.9154	29.5466	0.9032	38.1691	33.2770	0.9153	8.8674	36.1909	0.9376	0.0960	33.0048	0.9187	15.7108
AMIR	Transformer	23.5405	31.7645	**32.0262**	**0.9396**	**29.0988**	**33.7011**	0.9182	**8.4520**	37.1193	**0.9475**	0.0876	**34.2822**	**0.9351**	**12.5461**
SwinIR	Transformer	11.4977	187.9274	30.7966	0.9235	33.1968	33.5414	0.9162	8.5991	37.0070	0.9437	0.0880	33.7817	0.9278	13.9613
Restormer	Transformer	26.1241	35.2052	31.7177	0.9362	30.0549	33.6142	0.9177	8.5329	37.1368	0.9473	0.0872	34.1562	0.9337	12.8917
Xformer	Transformer	25.2197	35.7291	31.3292	0.9305	31.3538	33.5441	0.9168	8.6001	37.1421	0.9458	0.0865	34.0051	0.9310	13.3468
MambaIR	Mamba	31.5035	34.3491	31.3145	0.9305	31.3150	33.5025	0.9165	8.6345	37.1669	0.9458	**0.0864**	33.9946	0.9309	13.3453
Restore-RWKV	RWKV	27.9140	37.4585	31.9404	0.9386	29.3562	33.6935	**0.9188**	8.4609	**37.2091**	**0.9475**	0.0865	34.2810	0.9350	12.6345

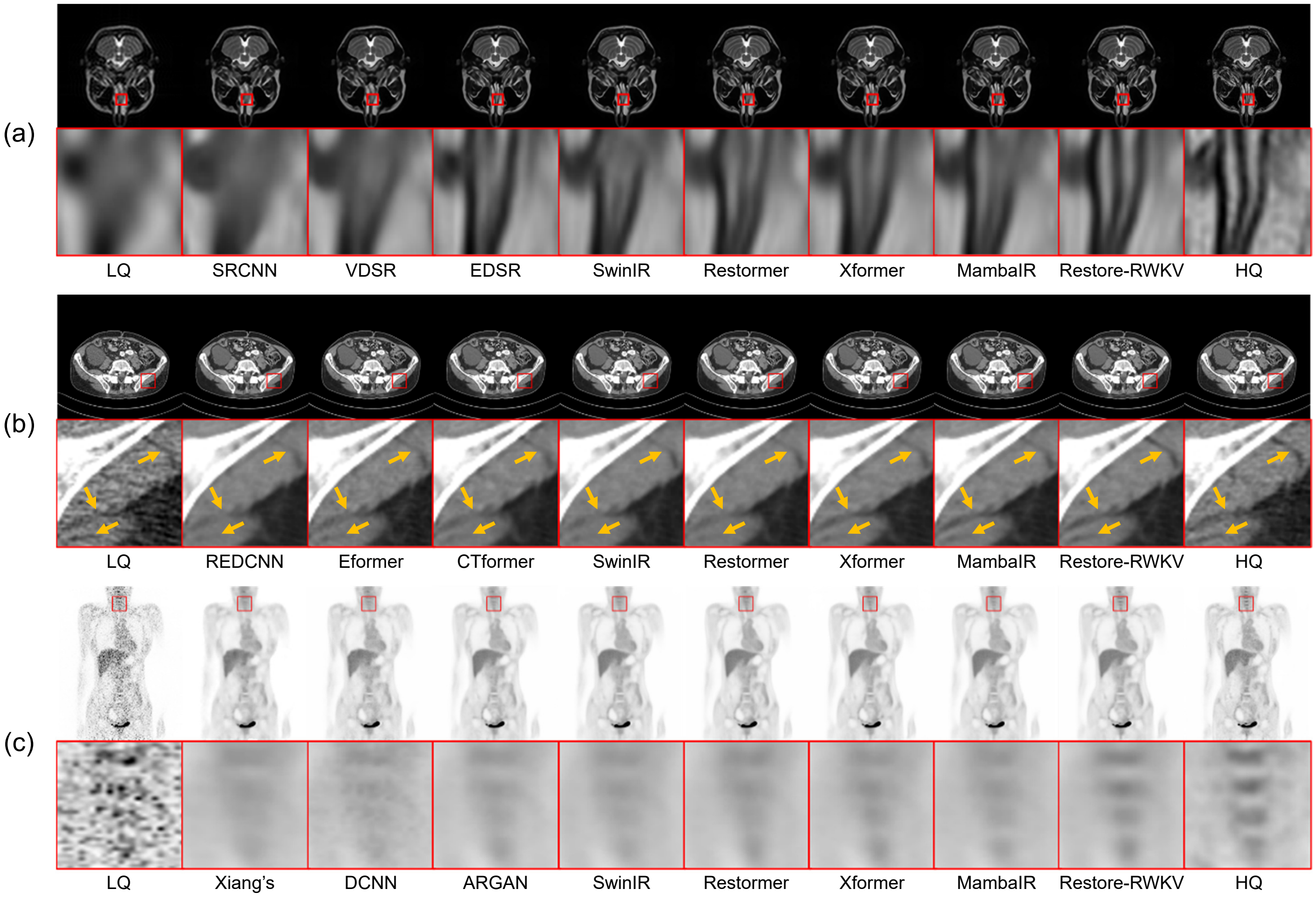


Figure 5. Visual comparison of different methods. (a) MRI image super-resolution. (b) CT image denoising. (c) PET image synthesis. Zoomed ROI of the rectangle region is recommended for better visualization. Arrows in (b) indicate regions with notable differences.

5.5. Ablation Studies

We conduct ablation experiments on the MRI image super-resolution task to investigate the significance of two key innovations in the Restore-RWKV: Re-WKV attention and Omni-Shift mechanisms.

Effect of Re-WKV and Omni-Shift. To evaluate the effectiveness of Re-WKV attention and Omni-Shift, we conduct experiments by replacing them with other WKV attention layers, such as Uni-WKV [30] and Bi-WKV [32], as well as token shift layers, including Uni-Shift [30] and Quad-Shift [32]. Table 5 indicates that both Re-WKV and Omni-Shift can effectively improve model performance, with the combination of both achieving the best results. Notably, compared to the original RWKV's [30] combination of "Uni-WKV + Uni-Shift," our proposed combination "Re-WKV + Omni-Shift" achieves an improvement of over 0.77 dB in PSNR. Figure 6 illustrates the impact of different attention and token shift combinations on the effective receptive field (ERF) of models. Generally, a larger receptive field allows the model to capture information from a wider region, thereby enhancing restoration performance. We can see that the original RWKV's [30] combination of "Uni-

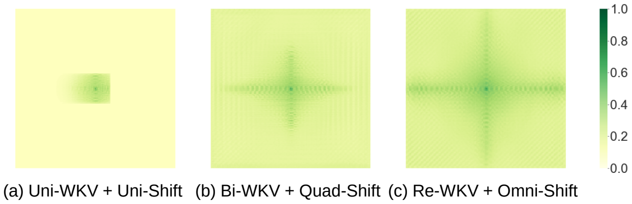


Figure 6. Comparison of the effective receptive field (ERF) in models with various WKV attention and token shift combinations. (a) The combination of ”Uni-WKV + Uni-Shift” in the original RWKV [30]. (b) The combination of ”Bi-WKV + Quad-Shift” in the Vision-RWKV [32]. (c) The combination of ”Re-WKV + Omni-Shift” in the proposed Restore-RWKV. A more extensively distributed dark area indicates a larger ERF.

Table 5. Component analysis of Re-WKV and Omni-Shift. The best results are highlighted in **bold**.

Component	PSNR↑	SSIM↑	RMSE↓
Uni-WKV + Uni-Shift	31.3184	0.9315	31.3278
Bi-WKV + Q-Shift	31.7261	0.9359	30.1125
Re-WKV + Q-Shift	31.9273	0.9386	29.4302
Bi-WKV + Omni-Shift	31.9377	0.9384	29.4318
Re-WKV + Omni-Shift (Ours)	**32.0913**	**0.9408**	**28.9713**

WKV + Uni-Shift” achieves only a local receptive field, while the Vision-RWKV’s [32] combination of ”Bi-WKV + Quad-Shift” achieves a global receptive field. Our proposed Restore-RWKV’s combination of ”Re-WKV + Omni-Shift” achieves the largest global receptive field. This indicates that the combination of ”Re-WKV + Omni-Shift” most effectively ensures the model’s ability to capture global dependencies in 2D images.

Ablation on Re-WKV. We investigate the influence of attention recurrence numbers on Re-WKV. As shown in Table 6, the model performance increases with the number of recurrences M in Re-WKV attention. Especially, $M = 2$ achieves a significant improvement over $M = 1$. This is because conducting attention from two different scan directions ($M = 2$) enables more effective token interaction on 2D images than using only a single direction ($M = 1$). As further increasing the recurrence number does not yield a significant performance improvement, we finally adopt $M = 2$ in Restore-RWKV.

We also explore architectural designs for conducting attention from both horizontal and vertical directions. We consider three designs. The first design alternates the scan direction between two adjacent blocks. The second design conducts attention from both directions within a block individually and then sums the attention results. The third design is our proposed recurrent attention, which recurrently conducts attention from both directions within a block. The results are shown in Table 7 and our proposed design with recurrent attention achieves the best performance.

Table 6. Ablation study on the recurrence number in Re-WKV. The best results are highlighted in **bold**.

Recurrence Number	PSNR↑	SSIM↑	RMSE↓
$M = 1$ (H-Scan)	31.9377	0.9384	29.4318
$M = 1$ (V-Scan)	31.9337	0.9383	29.4371
$M = 2$ (Ours)	32.0913	0.9408	28.9713
$M = 3$	32.1082	**0.9409**	28.8183
$M = 4$	**32.1133**	**0.9409**	**28.7891**

Table 7. Ablation study on architecture designs for conducting attention from both horizontal and vertical scan directions. The best results are highlighted in **bold**.

Architecture	PSNR↑	SSIM↑	RMSE↓
Alternating between Blocks	31.9728	0.9393	29.2942
Sum in a Block	31.8968	0.9386	29.4702
Recurrence in a Block (Ours)	**32.0913**	**0.9408**	**28.9713**

Table 8. Ablation study on Omni-Shift. The best results are highlighted in **bold**.

Context Range	PSNR↑	SSIM↑	RMSE↓
$k = 1$	31.8545	0.9375	29.7197
$k = 3$	31.9865	0.9394	29.2586
$k = 5$ (Ours)	**32.0913**	**0.9408**	**28.9713**
$k = 5$ (w/o Reparam.)	32.0050	0.9396	29.1992
$k = 7$	32.0578	0.9396	28.9602
$k = 9$	32.0290	0.9398	29.1360

Ablation on Omni-Shift. We conduct experiments to investigate two factors that most influence the performance of Omni-Shift: the context range of token shift and the parameterization strategy. As shown in Table 8, the experimental results indicate that Omni-Shift achieves the best performance with the context range set to $k = 5$. Additionally, the reparameterization strategy significantly improves the accuracy of token shift, thereby enhancing MedIR performance.

6. Conclusion

In this paper, we propose Restore-RWKV, leveraging the advanced RWKV architecture for the first time in the field of medical image restoration (MedIR). To achieve superior restoration performance, Restore-RWKV innovatively incorporates a recurrent WKV (Re-WKV) attention mechanism with linear computational complexity and an omnidirectional token shift (Omni-Shift) mechanism to effectively capture global and local dependencies in 2D images, respectively. Extensive experiments on various MedIR tasks demonstrate that our proposed Restore-RWKV is an efficient and effective backbone for medical image restoration.

References

[1] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(2):295–307, 2015. 1, 2, 8

[2] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1646–1654, 2016. 1, 2, 8

[3] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 136–144, 2017. 1, 2, 8

[4] Yuhua Chen, Feng Shi, Anthony G Christodoulou, Yibin Xie, Zhengwei Zhou, and Debiao Li. Efficient and accurate mri super-resolution using a generative adversarial network and 3d multi-level densely connected network. In *International conference on medical image computing and computer-assisted intervention*, pages 91–99. Springer, 2018. 1, 2

[5] Yulun Zhang, Kai Li, Kunpeng Li, and Yun Fu. Mr image super-resolution with squeeze and excitation reasoning attention network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13425–13434, 2021. 1, 2

[6] Zhiyun Song, Xin Wang, Xiangyu Zhao, Sheng Wang, Zhenrong Shen, Zixu Zhuang, Mengjun Liu, Qian Wang, and Lichi Zhang. Alias-free co-modulated network for cross-modality synthesis and super-resolution of mr images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 66–76. Springer, 2023. 1, 2

[7] Zexin Ji, Beiji Zou, Xiaoyan Kui, Pierre Vera, and Su Ruan. Deform-mamba network for mri super-resolution. *arXiv preprint arXiv:2407.05969*, 2024. 1, 2, 3

[8] Hu Chen, Yi Zhang, Mannudeep K Kalra, Feng Lin, Yang Chen, Peixi Liao, Jiliu Zhou, and Ge Wang. Low-dose ct with a residual encoder-decoder convolutional neural network. *IEEE transactions on medical imaging*, 36(12):2524–2535, 2017. 1, 2, 8

[9] Achleshwar Luthra, Harsh Sulakhe, Tanish Mittal, Abhishek Iyer, and Santosh Yadav. Eformer: Edge enhancement based transformer for medical image denoising. *arXiv preprint arXiv:2109.08044*, 2021. 1, 2, 8

[10] Dayang Wang, Fenglei Fan, Zhan Wu, Rui Liu, Fei Wang, and Hengyong Yu. Ctformer: convolution-free token2token dilated vision transformer for low-dose ct denoising. *Physics in Medicine & Biology*, 68(6):065012, 2023. 1, 2, 8

[11] Lei Xiang, Yu Qiao, Dong Nie, Le An, Weili Lin, Qian Wang, and Dinggang Shen. Deep auto-context convolutional neural networks for standard-dose pet image estimation from low-dose pet/mri. *Neurocomputing*, 267:406–416, 2017. 1, 2, 8

[12] Chung Chan, Jian Zhou, Li Yang, Wenyuan Qi, Jeff Kolthammer, and Evren Asma. Noise adaptive deep convolutional neural network for whole-body pet denoising. In *2018 IEEE Nuclear Science Symposium and Medical Imaging Conference Proceedings (NSS/MIC)*, pages 1–4. IEEE, 2018. 1, 2, 8

[13] Yanmei Luo, Luping Zhou, Bo Zhan, Yuchen Fei, Jiliu Zhou, Yan Wang, and Dinggang Shen. Adaptive rectification based adversarial network with spectrum constraint for high-quality pet image synthesis. *Medical Image Analysis*, 77:102335, 2022. 1, 2, 8

[14] Yang Zhou, Zhiwen Yang, Hui Zhang, I Eric, Chao Chang, Yubo Fan, and Yan Xu. 3d segmentation guided style-based generative adversarial networks for pet synthesis. *IEEE Transactions on Medical Imaging*, 41(8):2092–2104, 2022. 1, 2

[15] Se-In Jang, Tinsu Pan, Ye Li, Pedram Heidari, Junyu Chen, Quanzheng Li, and Kuang Gong. Spach transformer: Spatial and channel-wise transformer based on local and global self-attentions for pet image denoising. *IEEE transactions on medical imaging*, 2023. 1, 2

[16] Boyun Li, Xiao Liu, Peng Hu, Zhongqin Wu, Jiancheng Lv, and Xi Peng. All-in-one image restoration for unknown corruption. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17452–17462, 2022. 1, 2, 8

[17] Zhiwen Yang, Yang Zhou, Hui Zhang, Bingzheng Wei, Yubo Fan, and Yan Xu. Drmc: A generalist model with dynamic routing for multi-center pet image synthesis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 36–46. Springer, 2023. 1, 2, 8

[18] Zhiwen Yang, Haowei Chen, Ziniu Qian, Yang Yi, Hui Zhang, Dan Zhao, Bingzheng Wei, and Yan Xu. All-in-one medical image restoration via task-adaptive routing. *arXiv preprint arXiv:2405.19769*, 2024. 1, 2, 7, 8

[19] Xiangtao Kong, Chao Dong, and Lei Zhang. Towards effective multiple-in-one image restoration: A sequential and prompt learning strategy. *arXiv preprint arXiv:2401.03379*, 2024. 1, 2

[20] Yann LeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel. Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4):541–551, 1989. 1

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1

[22] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 1, 2, 3

[23] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 1, 2

[24] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023. 1, 3

[25] Lianghui Zhu, Bencheng Liao, Qian Zhang, Xinlong Wang, Wenyu Liu, and Xinggang Wang. Vision mamba: Efficient visual representation learning with bidirectional state space model. *arXiv preprint arXiv:2401.09417*, 2024. 1, 3

[26] Yue Liu, Yunjie Tian, Yuzhong Zhao, Hongtian Yu, Lingxi Xie, Yaowei Wang, Qixiang Ye, and Yunfan Liu. Vmamba: Visual state space model. *arXiv preprint arXiv:2401.10166*, 2024. 1, 3

[27] Wenjie Luo, Yujia Li, Raquel Urtasun, and Richard Zemel. Understanding the effective receptive field in deep convolutional neural networks. *Advances in neural information processing systems*, 29, 2016. 1

[28] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1833–1844, 2021. 1, 3, 8

[29] Hang Guo, Jinmin Li, Tao Dai, Zhihao Ouyang, Xudong Ren, and Shu-Tao Xia. Mambair: A simple baseline for image restoration with state-space model. *arXiv preprint arXiv:2402.15648*, 2024. 1, 3, 8

[30] Bo Peng, Eric Alcaide, Quentin Anthony, Alon Albalak, Samuel Arcadinho, Huanqi Cao, Xin Cheng, Michael Chung, Matteo Grella, Kranthi Kiran GV, et al. Rwkv: Reinventing rnns for the transformer era. *arXiv preprint arXiv:2305.13048*, 2023. 2, 3, 4, 5, 6, 9, 10

[31] Bo Peng, Daniel Goldstein, Quentin Anthony, Alon Albalak, Eric Alcaide, Stella Biderman, Eugene Cheah, Teddy Ferdinan, Haowen Hou, Przemysław Kazienko, et al. Eagle and finch: Rwkv with matrix-valued states and dynamic recurrence. *arXiv preprint arXiv:2404.05892*, 2024. 2, 3

[32] Yuchen Duan, Weiyun Wang, Zhe Chen, Xizhou Zhu, Lewei Lu, Tong Lu, Yu Qiao, Hongsheng Li, Jifeng Dai, and Wenhai Wang. Vision-rwkv: Efficient and scalable visual perception with rwkv-like architectures. *arXiv preprint arXiv:2403.02308*, 2024. 2, 3, 5, 6, 9, 10

[33] Zhengcong Fei, Mingyuan Fan, Changqian Yu, Debang Li, and Junshi Huang. Diffusion-rwkv: Scaling rwkv-like architectures for diffusion models. *arXiv preprint arXiv:2404.04478*, 2024. 2, 3

[34] Qingdong He, Jiangning Zhang, Jinlong Peng, Haoyang He, Yabiao Wang, and Chengjie Wang. Pointrwkv: Efficient rwkv-like model for hierarchical point cloud learning. *arXiv preprint arXiv:2405.15214*, 2024. 2, 3

[35] Tiancheng Gu, Kaicheng Yang, Xiang An, Ziyong Feng, Dongnan Liu, Weidong Cai, and Jiankang Deng. Rwkv-clip: A robust vision-language representation learner. *arXiv preprint arXiv:2406.06973*, 2024. 2, 3

[36] Haobo Yuan, Xiangtai Li, Lu Qi, Tao Zhang, Ming-Hsuan Yang, Shuicheng Yan, and Chen Change Loy. Mamba or rwkv: Exploring high-quality and high-efficiency segment anything model. *arXiv preprint arXiv:2406.19369*, 2024. 2, 3

[37] Xuanyao Chen, Zhijian Liu, Haotian Tang, Li Yi, Hang Zhao, and Song Han. Sparsevit: Revisiting activation sparsity for efficient high-resolution vision transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2061–2070, 2023. 3

[38] Zhijun Tu, Kunpeng Du, Hanting Chen, Hailing Wang, Wei Li, Jie Hu, and Yunhe Wang. Ipt-v2: Efficient image processing transformer using hierarchical attentions. *arXiv preprint arXiv:2404.00633*, 2024. 3

[39] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 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*, pages 234–241. Springer, 2015. 3

[40] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1874–1883, 2016. 3, 4

[41] Xiaohan Ding, Xiangyu Zhang, Ningning Ma, Jungong Han, Guiguang Ding, and Jian Sun. Repvgg: Making vgg-style convnets great again. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13733–13742, 2021. 6, 7

[42] H Malcolm Hudson and Richard S Larkin. Accelerated image reconstruction using ordered subsets of projection data. *IEEE transactions on medical imaging*, 13(4):601–609, 1994. 7

[43] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5728–5739, 2022. 8

[44] Jiale Zhang, Yulun Zhang, Jinjin Gu, Jiahua Dong, Linghe Kong, and Xiaokang Yang. Xformer: Hybrid x-shaped transformer for image denoising. *arXiv preprint arXiv:2303.06440*, 2023. 8