Multi-View Learning with Context-Guided Receptance for Image Denoising

Binghong Chen[1], **Tingting Chai**[2*], **Wei Jiang**[1], **Yuanrong Xu**[2], **Guanglu Zhou**[2] and **Xiangqian Wu**[2,3]

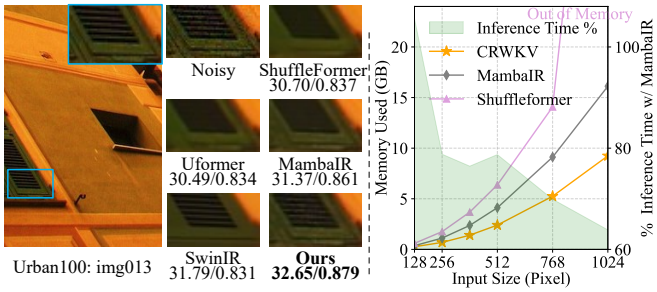
[1]School of Mathematics, Harbin Institute of Technology, China
[2]Faculty of Computing, Harbin Institute of Technology, China
[3]Suzhou Research Institute, Harbin Institute of Technology, China
math.cbh@stu.hit.edu.cn, ttchai@hit.edu.cn

Abstract

Image denoising is essential in low-level vision applications such as photography and automated driving. Existing methods struggle with distinguishing complex noise patterns in real-world scenes and consume significant computational resources due to reliance on Transformer-based models. In this work, the Context-guided Receptance Weighted Key-Value (CRWKV) model is proposed, combining enhanced multi-view feature integration with efficient sequence modeling. The Context-guided Token Shift (CTS) mechanism is introduced to effectively capture local spatial dependencies and enhance the model's ability to model real-world noise distributions. Also, the Frequency Mix (FMix) module extracting frequency-domain features is designed to isolate noise in high-frequency spectra, and is integrated with spatial representations through a multi-view learning process. To improve computational efficiency, the Bidirectional WKV (BiWKV) mechanism is adopted, enabling full pixel-sequence interaction with linear complexity while overcoming the causal selection constraints. The model is validated on multiple real-world image denoising datasets, outperforming the state-of-the-art methods quantitatively and reducing inference time up to 40%. Qualitative results further demonstrate the ability of our model to restore fine details in various scenes. The code is publicly available at https://github.com/Seeker98/CRWKV.

1 Introduction

Images captured in real-world scenes are commonly influenced by noise from a mixture of sources, including optical sensing limitations and electronic functional failures, making image denoising an essential topic in low-level computer vision. Failing to effectively remove noise can significantly reduce the performance of subsequent high-level tasks or scenes. For instance, this may affect content extracted from a low-light, noisy document capture or identification of pedestrians in adverse weather conditions for vision-based autonomous driving systems. Furthermore, recent advancements in computational resources and specialized hardware such as image signal processors (ISPs), combined with the inherent limitations of optical systems, create a timely opportunity to advance denoising methods further.



(a) Selected visual result (b) Model complexity comparison

Figure 1: Comparison of existing image denoising methods. (a) Visual results on Urban100 dataset: our method preserves fine details under complex noise pattern, while others suffer from information losses or artifacts. (b) Computational complexity across input scales for Transformer- and state space-based models.

Various foundational architectures have emerged, combining classical image processing principles with recent deep learning advances, which include convolutional neural networks (CNNs) [Zhang *et al.*, 2017], Transformers [Chen *et al.*, 2021], Mambas [Zhu *et al.*, 2024], and Receptance Weighted Key Values (RWKVs) [Peng *et al.*, 2023]. CNNs, often considered as an evolution of classical methods utilizing hand-crafted priors [Zheng *et al.*, 2019; Laghrib and Afraites, 2024; He *et al.*, 2010], remain their core idea of local window filtering. Despite advanced techniques such as non-local operations [Wang *et al.*, 2018], large kernels [Ding *et al.*, 2022], and architectures like U-Net [Ronneberger *et al.*, 2015], CNNs struggle with oversmoothing in complex scenes and fail to handle distant information. Transformers and Mambas introduce global modeling through self-attention (SA) mechanisms [Dosovitskiy *et al.*, 2020] and state space models [Zhu *et al.*, 2024; Liu *et al.*, 2024], respectively. While these approaches achieve the state-of-the-art (SOTA) performance, they encounter challenges such as quadratic computational complexity in Transformers and causal-style dependencies in Mambas. Alternative approaches [Jin *et al.*, 2024] tried to

*Corresponding author.

tackle these problems, but still lead to inconsistent modeling in local regions.

Reviewing prior work reveals two main issues: feature-level design and effective model architecture development. Consider a scenario where a single convolution layer efficiently reduces noise in detailed regions but struggles in flat areas, leaving residual noise. In such cases, the high-frequency components, often dominated by noise in flat regions, require attenuation rather than selective enhancement. In addition, image information is spatially distributed across all directions. However, causal-style token mixing, where information is derived only from past tokens, introduces asymmetry in both weighting and positioning. While this lack of symmetry may have minimal impact in natural language processing (NLP), it poses a significant challenge in image feature representation, where spatial symmetry is essential for accurate reconstruction and effective denoising.

In this study, we propose Context-guided Receptance Weighted Key-Value (CRWKV), a novel model that addresses challenges in noise modeling and computational efficiency mentioned above. Our model provides superior real-world image denoising performance within limited resources, and the main contributions are as follows:

- We propose CRWKV, a novel model for image denoising that integrates a multi-view learning approach. Our design enhances the RWKV model by introducing the BiWKV mechanism, enabling full pixel-sequence computation superior to causal-style selection with linear complexity relative to sequence length.
- We introduce CTS mechanism to effectively model local noise correlations in image. Besides, we propose FMix module to selectively process frequency-domain information and attenuate noise. They together significantly improve denoising performance.
- Comprehensive experiments on real-world image denoising datasets demonstrate that CRWKV consistently outperforms SOTA methods. Its strong generalization ability is validated through testing on multiple unseen datasets, showcasing its efficiency and effectiveness for practical denoising applications.

2 Related Works

2.1 Real-world Image Denoising

Image denoising is a fundamental problem in image processing, with applications ranging from photography to medical imaging. Classical methods, such as BM3D [Dabov *et al.*, 2007], rely on hand-crafted priors and assumptions about noise characteristics, achieving good results in controlled scenarios but struggle with complex noise patterns. With advancements in deep learning, modern approaches like DnCNN [Zhang *et al.*, 2017] have emerged, focusing primarily on removing Additive White Gaussian Noise (AWGN). While effective for synthetic noise, these methods face significant challenges when applied to real-world noise, which is far more complex than AWGN due to spatial correlation, intensity variation, and limited paired data.

To address these issues, various methods have been proposed. Some rethink the imaging and noise generation processes, as seen in CBDNet and SCUNet [Guo *et al.*, 2019; Zhang *et al.*, 2023], while others adopt self-supervised pixel reconstruction techniques using blind-spot networks [Krull *et al.*, 2019; Lee *et al.*, 2022]. Additionally, Gaussian denoisers have been adapted for real-world denoising by performing noise pattern corruption with shuffling techniques in advance [Zhou *et al.*, 2020; Xiao *et al.*, 2023]. Recent works have also explored novel image restoration architectures. For instance, MambaIR [Guo *et al.*, 2025] leverages state space modeling, while Restormer [Zamir *et al.*, 2022] incorporates windowed self-attention, both demonstrating notable improvements in real-world image denoising. Despite these advancements, limited attention has been given to integrating noise-specific priors into modern architectures while achieving a balance between computational efficiency and restoration quality.

2.2 RWKV Models

RWKV [Peng *et al.*, 2023] was proposed as an efficient alternative to Transformers for NLP, particularly in Large Language Models (LLMs). RWKV introduces two main innovations: token shifting and WKV computation. By incorporating token shifts in the preceding direction and concatenating shifted and non-shifted channels, the model can separate two tasks—predicting the next token, and accumulating and passing information from previous tokens. The WKV computation improves upon the Attention-Free Transformer [Zhai *et al.*, 2021] by employing trainable distance factors and enhanced integration of the current token to model token weights more precisely. Additionally, it adopts an equivalent recurrent form for efficient inference. These features position RWKV as a strong competitor to CNNs and Transformers.

Recent studies, such as Vision-RWKV [Duan *et al.*, 2024], have showed the potential of RWKV-based models as vision backbones [Fei *et al.*, 2024; Zhou and Chen, 2024]. Two major improvements have been developed to adapt RWKV for vision tasks: quad-shift, a token-shifting strategy on 2D planes, and Bi-WKV, an attention mechanism with absolute positional bias to align with the symmetric nature of images. While promising efforts like Restore-RWKV [Yang *et al.*, 2024] have emerged, targeting all-in-one medical image restoration with Re-WKV and omni-shift, there remains limited exploration of how RWKV-based models can enhance performance in low-level vision tasks, such as real-world image denoising.

3 Methodology

In this section, the details of our proposed model CRWKV will be introduced. We will start with the model overview and gradually analyze the structure of CRB and FMix designed specifically for real-world noise modeling.

3.1 Overall Architecture

As illustrated in Figure 2, the proposed CRWKV model adopts a U-shaped hierarchical structure with long-skip connections to effectively capture both local and global features. For a noisy input image $x \in \mathbb{R}^{H\times W\times 3}$, a 3×3 convolution

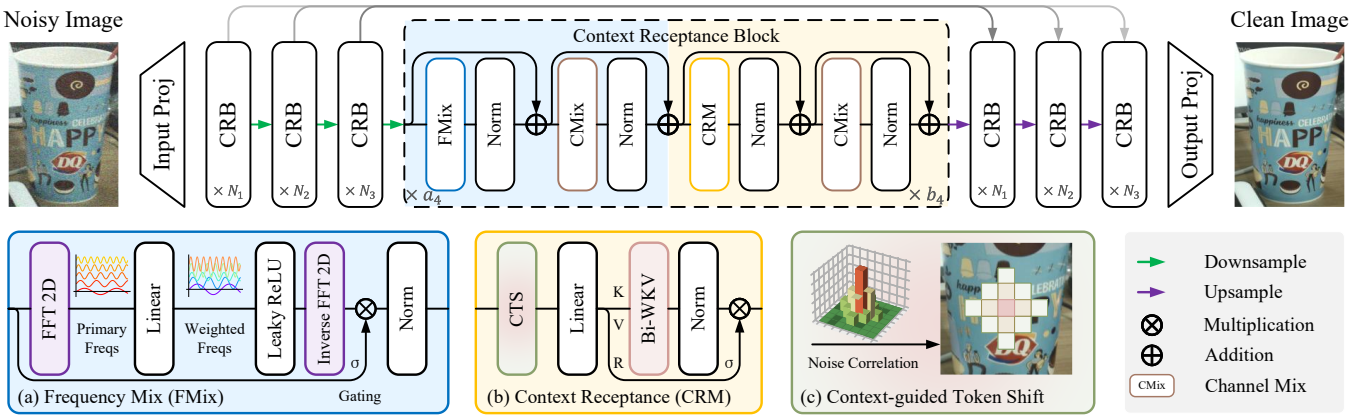


Figure 2: Architecture of the proposed CRWKV model, and (a) Frequency Mix Module (FMix), (b) Context Receptance Module (CRM) and (c) Context-guided Token Shift (CTS).

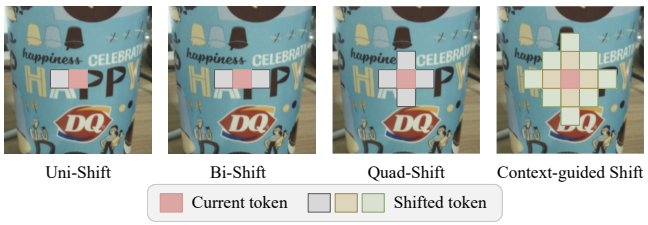


Figure 3: Illustration of different token shift mechanisms.

is first applied to extract low-level features. The resulting feature map is then processed through four distinct stages of encoders and decoders following the U-shaped structure. After this process, the refined features are passed through a final projection layer to reconstruct the denoised output image.

Each encoder and decoder stage contains N_k Context Receptance Blocks (CRBs). Each CRB consists of two types of modules for multi-view learning: a_k Frequency Mix (FMix) Modules and b_k Context Receptance Modules (CRMs), such that $a_k + b_k = N_k$ for $k = 1, 2, 3, 4$. The computation in a single CRB can be expressed as:

$$\begin{aligned} z_1 &= \text{Norm}(\text{FMix}(x)) + \alpha_1 x\,, \\ z &= \text{Norm}(\text{CMix}(z_1)) + \alpha_2 z_1, \end{aligned} \tag{1}$$

and

$$\begin{aligned} y_1 &= \text{Norm}(\text{CRM}(z)) + \beta_1 z\,, \\ y &= \text{Norm}(\text{CMix}(y_1)) + \beta_2 y_1\,, \end{aligned} \tag{2}$$

where x and y denote the input and output features, respectively. The Channel Mix (CMix) module is computed as follows:

$$\begin{aligned} r_c, k_c &= \text{CTS}(z)\,, \\ \text{CMix}(z) &= \sigma(L(r_c)) \odot \text{Norm}(\max(0, k_c^2))\,, \end{aligned} \tag{3}$$

where $\text{CTS}(\cdot)$ represents the Context-guided Token Shift mechanism, further detailed in Section 3.2.

3.2 Context Receptance Module

Algorithm 1 Context-guided Token Shift

Input: input x, offset dictionary D, learnable weight ω
Parameter: channel C
Output: shifted output $\text{CTS}(x)$

Let $p_{\text{sum}} = 0, c = 0, o = \text{zeros}(x.\text{shape})$
for offset p in D **do**
  calculate p's Manhattan distance $d_p = d_m(p, 0)$
  calculate offset p's weight $w_p = 1/d_p$
  $p_{\text{sum}} \mathrel{+}= w_p$
end for
calculate channel expansion factor $k = C/p_{\text{sum}}$
for offset p in D **do**
  fill o w/ shifted x: $o[c : c + k \cdot w_p] = x_p[c : c + k \cdot w_p]$
  $c = c + k \cdot w_p$
end for
return $\text{CTS}(x) = \omega \cdot o + (1 - \omega) \cdot x$

Context-guided Token Shift. The shifting operations in the mixing processes enable selective accumulation of information from previous tokens to the current token. Unlike NLP problems, pixels in an image often exhibit correlations with their neighbors in a centrosymmetric manner. Real-world noise, in particular, correlate with a specific neighborhood structure. Therefore, it is essential to identify this neighborhood shape to ensure that all significant pixels are fully considered while avoiding shifts over excessively large areas, which may introduce extra complexity in the implementation and reduce efficiency.

To address this issue, we propose a CTS mechanism, which assigns predefined weights to the most correlated pixels, as illustrated in Figure 2(c), and a visual comparison with existing token shift methods is provided in Figure 3. The colored masks in the figure depict individual pixels rather than image patches. Starting with the red central pixel, a context-guided region is defined as the equivalent reception field. This region aligns with the specific neighborhood identified in [Wang *et al.*, 2023] through Pearson's correlation analysis, which highlights the most correlated pixels with the central noise. The detailed algorithm of CTS is presented in Algorithm 1.

Characteristics	Vanilla Attention	Window Attention	Linear Attention	State Space Model	BiWKV
Operator Type	SA	SA	Linearized SA	Selective scan	Attention-free
Inductive Bias	-	Locality	Low-rank (LR) approx.	Causality	Spatial symmetry
Token Mixer	Dense	Blockwise sense	Lower-triangle (LT)	LT, blockwise LR	Almost dense
Token Type	Patch (typ.)	Patch (typ.)	Patch (typ.)	Pixel (typ.)	Pixel (typ.)
Local Interaction	-	Shifting window	-	Scan strategy	Token shifting
Global Interaction	Direct	Arch-specific	Kernel-dependent	Hidden state	Recurrent state
Complexity	Quadratic	Quadratic to win_size	Linear	Linear	Linear

Table 1: Comparison of different operators.

Bidirectional WKV operation. Unlike the self-attention mechanism, BiWKV operation employs a token-shift operation to achieve a weighted fusion of the feature map with its context-guided shifted version, generated using the CTS operation. The fused feature map, denoted as $\mathrm{CTS}(x)$, is then used to produce r_1, k_1, and v_1 through three linear projection layers:

$$r_1 = \mathrm{CTS}(x)W_r\,, k_1 = \mathrm{CTS}(x)W_k\,, v_1 = \mathrm{CTS}(x)W_v, \quad (4)$$

where W_r, W_k, and W_v are the weight matrices of the respective linear projection layers. Among these, k_1 and v_1 are utilized for the BiWKV computation, while r_1 serves as a gating mechanism after passing through a sigmoid activation function. The output of a single CRM block is computed as:

$$\mathrm{CRM}(x) = \sigma(r_1) \odot \mathrm{Norm}(\mathrm{BiWKV}(k_1, v_1))\,, \quad (5)$$

where σ represents the sigmoid function and $\odot$ is the Hadamard element-wise product. Denote the relative position bias between the t-th and i-th token as:

$$b_{t,i} = -(|t-i|-1)/T\,, \quad (6)$$

the computation of the BiWKV operation is given as follows:

$$\begin{aligned}&\mathrm{BiWKV}(k_1, v_1)_t\\ =&\frac{\sum_{i\neq t}\exp\left(b_{t,i}w + k_{1,i}\right)v_{1,i} + \exp\left(u + k_{1,t}\right)v_{1,t}}{\sum_{i\neq t}\exp\left(b_{t,i}w + k_{1,i}\right) + \exp\left(u + k_{1,t}\right)}\,,\end{aligned} \quad (7)$$

where T represents the length of the processed sequence, $k_{1,i}$ and $v_{1,i}$ correspond to the i-th token of k_1 and v_1, respectively, and u is a learnable parameter representing bonus of the current token. The absolute valued design in $b_{t,i}$ ensures that tokens equidistant from the current token in both forward and backward directions are weighted equally, preserving the spatial symmetry of image planes. Overall, the BiWKV operation computes a weighted sum of all tokens in v_1 along the token dimension. The weights are determined by a combination of symmetric relative position bias, the vector k_1, and the current token bonus controlled by the parameter u.

To justify the BiWKV operation for real-world image denoising, we compare it to several attention mechanisms and sequence models relevant to low-level vision. These include the vanilla attention [Dosovitskiy *et al.*, 2020], window attention [Liang *et al.*, 2021], linear attention [Cai *et al.*, 2023], state space model [Guo *et al.*, 2025], and BiWKV. Table 1 summarizes a detailed comparison of these operations, focusing on key characteristics. Specifically, the 'Token Mixer' row in the table reflects the shape of L in the following equation, based on the structured masked attention framework [Dao and Gu, 2024]:

$$y = f(L \circ (QK^\top)) \cdot V. \quad (8)$$

3.3 Frequency Mix Module

Real-world images often contain spatially correlated noise, which complicates detail-rich areas. This noise spreads information across all frequencies, causing overlap in high-frequency components and resulting in either over-smoothing or incomplete noise removal.

To address this issue, we propose the FMix module, illustrated in Figure 2(a), designed to extract detailed frequency representations from high-level features using FFT. Given a feature map $x \in \mathbb{R}^{h\times w\times c}$, a 2D FFT is applied:

$$x^F_{u,v,c} = \sum_{m=0}^{h-1}\sum_{n=0}^{w-1} x_{m,n,c}\exp\left(-2\pi i\left(\frac{ux}{h} + \frac{vy}{w}\right)\right). \quad (9)$$

The extracted frequencies are linearly weighted, followed by activation with Leaky ReLU. The weighted frequencies are then passed through an inverse FFT (iFFT) to return to the spatial domain. The resulting feature map is combined with the input via an element-wise product and normalized to produce the filtered feature map:

$$\begin{aligned} x^F &= \mathrm{FFT}(x)\,,\\ z &= \mathrm{LReLU}(\mathrm{Linear}(x^F))\,,\\ \mathrm{FMix}(x) &= \mathrm{Norm}(\mathrm{iFFT}(z)\cdot x).\end{aligned} \quad (10)$$

3.4 Loss Function

The proposed CRWKV model is trained by minimizing the L_1 loss, which can be written as follows:

$$L_1(y, x^*) = \|y - x^*\|_1\,, \quad (11)$$

where x^* is the ground truth and y is the model output. The L_1 loss is capable of retaining fine details such as edges and textures with its linear penalty that other losses such as L_2 may smooth out, ensuring robustness to outliers and accurate reconstruction of the denoised image.

4 Experiments

4.1 Experiment Setup

Datasets and metrics. We use the SIDD dataset [Abdelhamed *et al.*, 2018], consisting of 320 real-world images, as the primary training set. From each high-resolution image (256×256), we crop 300 non-overlapping slices, generating a total of 96,000 training samples. For testing, we evaluate on the SIDD, ccnoise [Nam *et al.*, 2016], and PolyU [Xu *et al.*, 2018] datasets, all containing realistic noise. Additionally, we create a synthetic dataset, Urban100GP, by introducing

Methods	Params (M)	SIDD		ccnoise		PolyU		Urban100GP	
		↑PSNR	↑SSIM	↑PSNR	↑SSIM	↑PSNR	↑SSIM	↑PSNR	↑SSIM
BM3D [Dabov *et al.*, 2007]	-	29.97	0.679	36.15	0.947	37.40	0.957	25.02	0.813
AP-BSN [Lee *et al.*, 2022]	3.10	36.74	0.889	33.30	0.918	36.46	0.947	24.25	0.716
B2U [Wang *et al.*, 2022a]	1.96	32.37	0.727	35.72	0.938	35.71	0.947	27.39	0.847
DnCNN [Zhang *et al.*, 2017]	0.56	26.21	0.604	33.88	0.959	36.11	0.960	24.20	0.866
SwinIR [Liang *et al.*, 2021]	11.75	33.70	0.864	35.26	0.978	37.14	0.977	28.21	0.896
Uformer [Wang *et al.*, 2022b]	50.88	39.68	0.958	36.02	0.979	37.48	0.979	26.88	0.885
ShuffleFormer [Xiao *et al.*, 2023]	50.53	39.60	0.958	35.88	0.978	37.50	0.979	27.89	0.900
Restormer [Zamir *et al.*, 2022]	26.10	**40.01**	0.960	36.33	0.981	37.56	0.979	28.18	0.904
MambaIR [Guo *et al.*, 2025]	26.78	39.88	0.960	36.20	0.981	37.58	0.980	27.42	0.898
CRWKV (Ours)	20.19	39.87	**0.960**	**36.69**	**0.983**	**37.60**	**0.980**	**28.30**	**0.908**

Table 2: Quantitative results of real-world image denoising on SIDD, ccnoise, PolyU and Urban100GP.

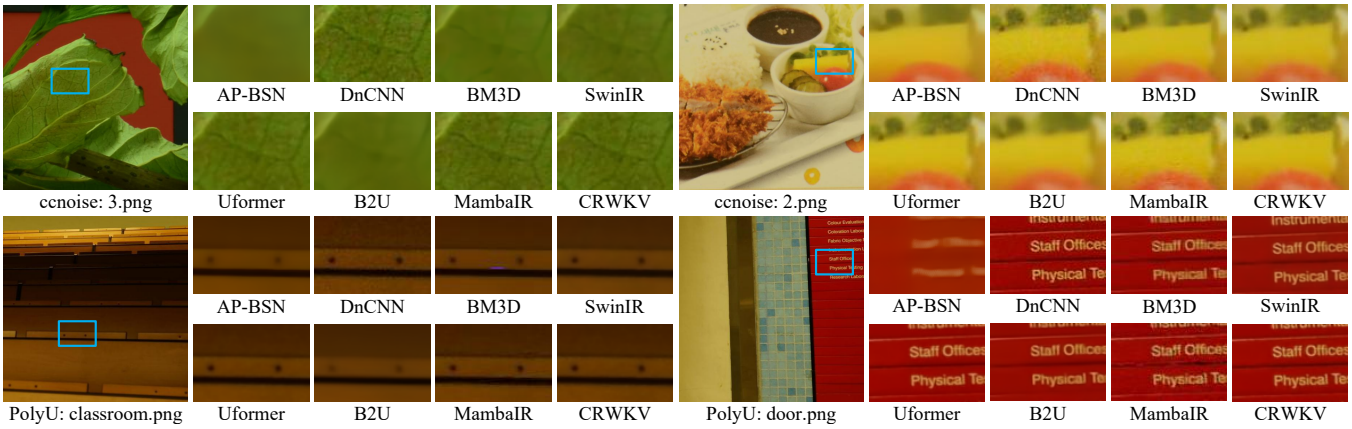


Figure 4: Selected visual results on ccnoise and PolyU dataset.

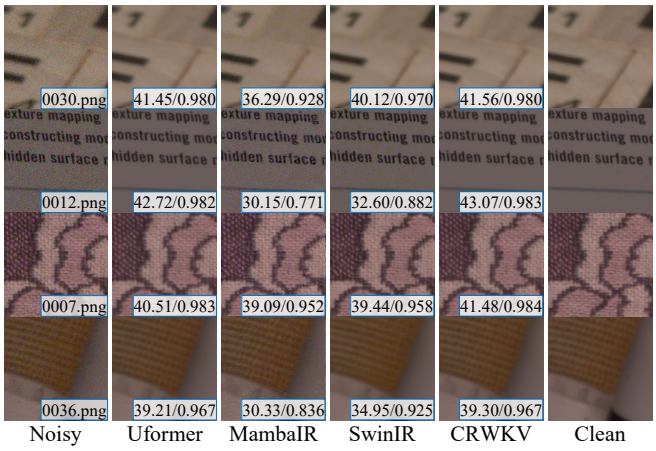


Figure 5: Selected visual results on SIDD dataset.

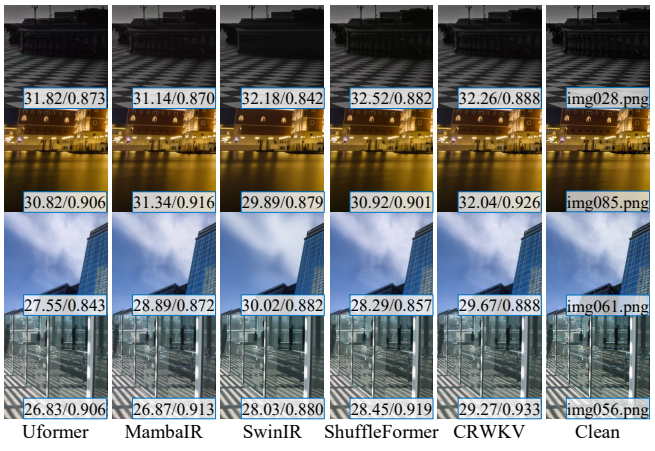


Figure 6: Selected visual results on Urban100GP dataset.

mixed Additive White Gaussian Noise (AWGN) with $\sigma = 10$ and Poisson noise to Urban100 [Huang *et al.*, 2015] to simulate real-world noise. To quantify model performance, we use Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index Measure (SSIM) as evaluation metrics.

Implementation details. During training, images are cropped to 128×128, and data augmentation techniques such as rotations ($90°$, $180°$, $270°$) and random flipping are applied to enhance model robustness. The training process is carried out with a batch size of 4 for a total of $288,000$ iterations. We use the AdamW optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The learning rate starts at 3×10^{-4} and is gradually reduced to 1×10^{-6} after the $192,000$-th iteration. For model-specific configurations, the output channel size of the input projection is set to 48. The depths of the four stages are empirically chosen as $L_1 = 3$, $L_2 = L_3 = 4$, and $L_4 = 6$. All experiments are conducted on a single NVIDIA RTX 4090 GPU, running

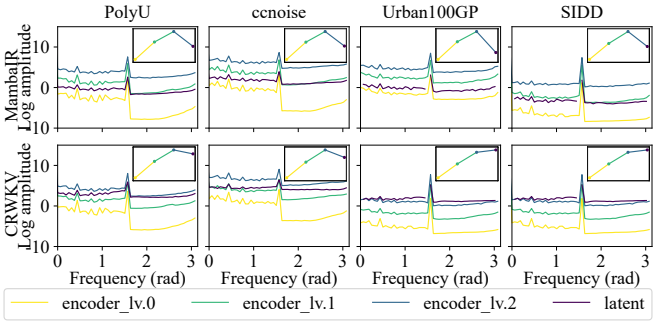


Figure 7: Power spectrum of feature maps at different depths of MambaIR and CRWKV across various datasets. Lines with deeper colors represent deeper layers.

Ubuntu 22.04 with PyTorch 2.5 as the software environment.

4.2 Comparison on Real-world Image Denoising

The proposed FRWKV method was evaluated on real-world image denoising tasks against several SOTA methods, including BM3D [Dabov *et al.*, 2007], AP-BSN [Lee *et al.*, 2022], B2U [Wang *et al.*, 2022a], DnCNN [Zhang *et al.*, 2017], Uformer [Wang *et al.*, 2022b], SwinIR [Liang *et al.*, 2021], ShuffleFormer [Xiao *et al.*, 2023], Restormer [Zamir *et al.*, 2022], and MambaIR [Guo *et al.*, 2025]. The compared methods cover backbones including CNNs, Transformers and Mambas, functioning paradigms including supervised, self-supervised, and non-learning based.

Quantitative comparison. Quantitative results on four datasets—SIDD, ccnoise, PolyU, and Urban100GP—are summarized in Table 2, along with the parameter count for each model. Our model achieves superior performance across nearly all datasets and metrics. The only exception is the PSNR metric on the SIDD dataset, where Restormer slightly surpasses our model by a margin of 0.1 dB. However, this minor advantage comes at the cost of a significantly larger model size, with Restormer requiring 30% more parameters than CRWKV. On the ccnoise and PolyU datasets, the differences between the models' metrics are less significant compared to the SIDD dataset. This can be attributed to the complex noise pattern in SIDD images, including noise introduced from multiple stages of the image processing pipeline.

Figure 7 compares the power spectrum of feature maps between MambaIR and CRWKV. A significant amplitude drop is observed in MambaIR between encoder_layer.2 and latent, which is essential in semantic-level feature reconstruction. In contrast, CRWKV exhibits a much smaller amplitude reduction on PolyU and ccnoise while maintaining superior performance on Urban100GP and SIDD. This indicates that CRWKV is able to retain amplitude stably across layers, preserving and leveraging high-frequency information throughout the denoising process effectively.

Qualitative comparison. On ccnoise and PolyU datasets (Figure 4), our model excels in preserving both fine details (leaf veins in ccnoise 3.png) and flat regions (ccnoise 2.png), and maintaining high-quality text fidelity (PolyU door.png). Other methods, such as B2U, struggle with edge preservation (PolyU classroom.png), while models like MambaIR leave visible noise residuals in flat regions (ccnoise 2.png).

Methods	Params (M)	FLOPs (G)	Time (ms)	SIDD PSNR	SIDD SSIM
SwinIR [Liang *et al.*, 2021]	11.75	253.46	170.33	33.70	0.864
ShuffleFormer [Xiao *et al.*, 2023]	50.53	120.67	98.96	39.60	0.958
Uformer [Wang *et al.*, 2022b]	50.88	41.44	62.50	39.68	0.958
Restormer [Zamir *et al.*, 2022]	26.10	35.24	47.40	40.01	0.960
MambaIR [Guo *et al.*, 2025]	26.78	34.39	79.61	39.88	0.960
CRWKV (Ours)	20.19	28.78	62.74	39.87	0.960

Table 3: Efficiency comparison with the SOTA methods.

For SIDD dataset (Figure 5), visual comparisons showcase CRWKV's superior ability to restore text and textures. In the first two examples, our model successfully reconstructs text at varying scales, while competing methods struggle. In the latter two examples, CRWKV recovers intricate textures such as those in the 0036.png, where other models like SwinIR fail to reproduce fine details. The PSNR and SSIM values displayed in the lower-right corners further confirm the model's effectiveness. For the Urban100GP dataset (Figure 6), CRWKV produces denoised images that strike a superior balance between preserving structural details and reducing noise artifacts. For instance, the grid pattern and fine features in img056.png can be jointly restored with CRWKV, outperforming other models. Additionally, CRWKV produces the most realistic reconstruction of the reflective water surface, showcasing its ability to handle challenging scenarios.

Compared to other methods, CRWKV demonstrates significant improvements in restoring both flat regions and fine details, achieving smoother textures and sharper edges. Furthermore, evaluations on unseen datasets, such as ccnoise, PolyU, and Urban100GP, suggest strong generalizability of CRWKV, as they were excluded from the training process.

Computational complexity. The model efficiency comparison results are summarized in Table 3. On the ccnoise dataset, CRWKV outperforms Restormer by over 0.3 dB PSNR and achieves competitive performance with MambaIR, while utilizing only 83% of the FLOPs. Figure 1(b) illustrates GPU memory usage during inference for varying input image sizes, comparing CRWKV with state-space and full-attention architectures. Notably, the Transformer-based Shuffleformer encounters memory limitations when the input size reaches 1024, whereas CRWKV requires only 40% of the GPU memory. Even with a comparable parameter count, CRWKV uses approximately 60% of the memory consumed by state-space-based MambaIR. This demonstrates that CRWKV strikes an optimal balance between denoising performance and computational efficiency, offering a practical solution for high-resolution and resource-constrained applications.

4.3 Ablation Study

Effectiveness of CRB module designs. To evaluate the effectiveness of the FMix module and CTS mechanism, we perform experiments on different configurations within the CRB. In the first configuration, we remove FMix and replace it with a basic spatial-mix while omitting the CTS in the CRM and CMix modules. In the second configuration, FMix is retained,

but CTS is excluded. Finally, we evaluate the model's performance with partial (CRM-only or CMix-only) and full insertion of CTS. The results suggest that removing FMix leads to a substantial performance degradation, with a PSNR drop of at least 0.40 dB on the SIDD dataset compared to the second configuration. Similarly, omitting CTS from CRM and CMix results in suboptimal performance. Introducing CTS yields consistent improvements, increasing PSNR by at least 0.15 dB on the ccnoise dataset and further boosting results on the SIDD dataset. These findings validate the importance of FMix and CTS in enhancing denoising performance.

Effectiveness of shifting mechanisms. As a guidance to the model's vast search space, shifting mechanisms include Uni-shift, Bi-shift, and Quad-shift are adopted previously. To further evaluate the effectiveness of shifting window in CTS, we implement CTS(+), an extended version including pixels at a Manhattan distance of 3 from the central pixel additionally, covering a total of 16 neighboring pixels. Table 5 shows CTS outperforms CTS(+), suggesting the additional context introduces complexity without significant benefits. Although learning a dynamic offset dictionary with deformable convolution is feasible, we choose a fixed offset dictionary to avoid overfitting to noise patterns in the training set and to reduce computational costs. While the results show that a fixed offset dictionary achieves comparable performance, the need for learnable offsets for optimal results remains an open question.

Effectiveness of parameter settings. To evaluate the roles of FMix and CRM in the model, we analyze the impact of parameters a_k , b_k (where each CRB contains a FMix modules and b CRM modules), as shown in Table 6. The findings can be summarized as follows: (1) A small number of FMix modules is optimal for low-level feature extraction and excessive usage disrupts early-stage modeling. (2) Applying frequency domain analysis at the middle layers negatively impacts performance. This may be attributed to the limited ability of FMix to model mid-level features, which are less semantically structured. (3) Incorporating frequency selection at deeper network layers significantly enhances performance. Deeper layers deal with semantically rich features that benefit more from frequency-domain processing. Based on these findings, the optimal configuration for the CRB is $b_1 = 3$, $b_2 = b_3 = 4$ and $a_4 = 6$.

Effectiveness of loss functions. Table 7 presents a comparison of different loss functions on SIDD and ccnoise datasets. L_1 loss achieves the best performance, offering a strong balance between pixel-level fidelity and perceptual quality. In contrast, MSE loss, while converging faster, suffers from over-smoothing, leading to unsatisfied results. Charbonnier loss achieves comparable but slightly weaker results, likely due to its more complex convergence dynamics. PSNR loss, however, shows a significant performance drop on the ccnoise dataset, suggesting its limitations in generalizability.

FMix	CTS		SIDD		ccnoise	
	CRM	CMix	PSNR	SSIM	PSNR	SSIM
✘			39.28	0.954	36.45	0.981
✔			39.70	0.957	36.50	0.982
✔	✔		39.75	0.957	36.64	0.983
✔		✔	39.74	0.957	36.62	0.982
✔	✔	✔	39.87	0.960	36.69	0.983

Table 4: The effectiveness of CRB module designs.

Shifting	SIDD		ccnoise	
	PSNR	SSIM	PSNR	SSIM
Uni-Shift	39.57	0.957	36.46	0.982
Bi-Shift	39.58	0.958	36.54	0.982
Quad-Shift	39.74	0.958	36.50	0.982
CTS(+)	39.79	0.958	36.60	0.982
CTS	39.87	0.960	36.69	0.983

Table 5: The effectiveness of shifting mechanisms.

k	L_k	(a_k, b_k)	SIDD		ccnoise	
			PSNR	SSIM	PSNR	SSIM
1	3	$(3,0)$	39.72	0.957	36.13	0.935
		$(2,1)$	39.82	0.958	36.60	0.980
		$(1,2)$	39.83	0.958	36.57	0.925
		$(0,3)$	39.87	0.960	36.69	0.983
2, 3	4	$(4,0)$	32.77	0.779	22.91	0.727
		$(2,1)$	34.63	0.789	30.71	0.824
		$(1,2)$	38.56	0.908	34.90	0.884
		$(0,4)$	39.87	0.960	36.69	0.983
4	6	$(6,0)$	39.87	0.960	36.69	0.983
		$(3,3)$	39.84	0.958	36.63	0.982
		$(2,4)$	39.85	0.957	36.59	0.982
		$(0,6)$	39.84	0.957	36.65	0.983

Table 6: The effectiveness of parameter settings.

Loss function	SIDD		ccnoise	
	PSNR	SSIM	PSNR	SSIM
Charbonnier Loss	39.86	0.959	36.67	0.983
MSE Loss	39.66	0.957	36.53	0.983
PSNR Loss	39.82	0.958	36.31	0.982
L_1 Loss	39.87	0.960	36.69	0.983

Table 7: The effectiveness of loss functions.

5 Conclusion

This work introduces the CRWKV model, a novel approach designed to tackle the challenges of noise modeling and computational inefficiency in real-world image denoising tasks. Key contributions include the CTS mechanism, which effectively captures local spatial contexts affected by noise, and the FMix module, which integrates semantic-level frequency-domain information through a multi-view learning process. By incorporating the BiWKV mechanism into the RWKV backbone, CRWKV achieves efficient pixel-sequence computation with linear complexity, overcoming the limitations of causal-style computation. These advancements enable CRWKV to effectively differentiate noise from complex scenes and separate high-frequency noise from structural details. Extensive experimental results demonstrate CRWKV's superior performance, both quantitatively and qualitatively, highlighting its robustness and practicality for real-world image denoising applications.

Ethical Statement

There are no ethical issues.

Acknowledgments

This work is supported by Shandong Provincial Natural Science Foundation (Grant No. ZR2023QF030).

References

[Abdelhamed *et al.*, 2018] Abdelrahman Abdelhamed, Stephen Lin, and Michael S Brown. A high-quality denoising dataset for smartphone cameras. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1692–1700, 2018.

[Cai *et al.*, 2023] Han Cai, Junyan Li, Muyan Hu, Chuang Gan, and Song Han. Efficientvit: Lightweight multi-scale attention for high-resolution dense prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 17302–17313, 2023.

[Chen *et al.*, 2021] Hanting Chen, Yunhe Wang, Tianyu Guo, Chang Xu, Yiping Deng, Zhenhua Liu, Siwei Ma, Chunjing Xu, Chao Xu, and Wen Gao. Pre-trained image processing transformer. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12294–12305. IEEE Computer Society, 2021.

[Dabov *et al.*, 2007] Kostadin Dabov, Alessandro Foi, Vladimir Katkovnik, and Karen Egiazarian. Image denoising by sparse 3-d transform-domain collaborative filtering. *IEEE Transactions on image processing*, 16(8):2080–2095, 2007.

[Dao and Gu, 2024] Tri Dao and Albert Gu. Transformers are ssms: Generalized models and efficient algorithms through structured state space duality. *arXiv preprint arXiv:2405.21060*, 2024.

[Ding *et al.*, 2022] Xiaohan Ding, Xiangyu Zhang, Jungong Han, and Guiguang Ding. Scaling up your kernels to 31x31: Revisiting large kernel design in cnns. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11963–11975, 2022.

[Dosovitskiy *et al.*, 2020] 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.

[Duan *et al.*, 2024] 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.

[Fei *et al.*, 2024] 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.

[Guo *et al.*, 2019] Shi Guo, Zifei Yan, Kai Zhang, Wangmeng Zuo, and Lei Zhang. Toward convolutional blind denoising of real photographs. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1712–1722, 2019.

[Guo *et al.*, 2025] 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. In *European conference on computer vision*, pages 222–241. Springer, 2025.

[He *et al.*, 2010] Kaiming He, Jian Sun, and Xiaoou Tang. Single image haze removal using dark channel prior. *IEEE transactions on pattern analysis and machine intelligence*, 33(12):2341–2353, 2010.

[Huang *et al.*, 2015] Jia-Bin Huang, Abhishek Singh, and Narendra Ahuja. Single image super-resolution from transformed self-exemplars. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5197–5206, 2015.

[Jin *et al.*, 2024] Yanliang Jin, Yifan Wu, Yuan Gao, Shunqing Zhang, Shugong Xu, and Cheng-Xiang Wang. Linformer: A linear-based lightweight transformer architecture for time-aware mimo channel prediction. *arXiv preprint arXiv:2410.21351*, 2024.

[Krull *et al.*, 2019] Alexander Krull, Tim-Oliver Buchholz, and Florian Jug. Noise2void-learning denoising from single noisy images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2129–2137, 2019.

[Laghrib and Afraites, 2024] Amine Laghrib and Lekbir Afraites. Image denoising based on a variable spatially exponent pde. *Applied and Computational Harmonic Analysis*, 68:101608, 2024.

[Lee *et al.*, 2022] Wooseok Lee, Sanghyun Son, and Kyoung Mu Lee. Ap-bsn: Self-supervised denoising for real-world images via asymmetric pd and blind-spot network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17725–17734, 2022.

[Liang *et al.*, 2021] 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.

[Liu *et al.*, 2024] 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.

[Nam *et al.*, 2016] Seonghyeon Nam, Youngbae Hwang, Yasuyuki Matsushita, and Seon Joo Kim. A holistic approach to cross-channel image noise modeling and its application to image denoising. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1683–1691, 2016.

[Peng *et al.*, 2023] Bo Peng, Eric Alcaide, Quentin Anthony, Alon Albalak, Samuel Arcadinho, Stella Biderman, Huanqi Cao, Xin Cheng, Michael Chung, Matteo Grella, et al. Rwkv: Reinventing rnns for the transformer era. *arXiv preprint arXiv:2305.13048*, 2023.

[Ronneberger *et al.*, 2015] 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.

[Wang *et al.*, 2018] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7794–7803, 2018.

[Wang *et al.*, 2022a] Zejin Wang, Jiazheng Liu, Guoqing Li, and Hua Han. Blind2unblind: Self-supervised image denoising with visible blind spots. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2027–2036, 2022.

[Wang *et al.*, 2022b] Zhendong Wang, Xiaodong Cun, Jianmin Bao, Wengang Zhou, Jianzhuang Liu, and Houqiang Li. Uformer: A general u-shaped transformer for image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 17683–17693, 2022.

[Wang *et al.*, 2023] Zichun Wang, Ying Fu, Ji Liu, and Yulun Zhang. Lg-bpn: Local and global blind-patch network for self-supervised real-world denoising. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18156–18165, 2023.

[Xiao *et al.*, 2023] Jie Xiao, Xueyang Fu, Man Zhou, Hongjian Liu, and Zheng-Jun Zha. Random shuffle transformer for image restoration. In *International Conference on Machine Learning*, pages 38039–38058. PMLR, 2023.

[Xu *et al.*, 2018] Jun Xu, Hui Li, Zhetong Liang, David Zhang, and Lei Zhang. Real-world noisy image denoising: A new benchmark. *arXiv preprint arXiv:1804.02603*, 2018.

[Yang *et al.*, 2024] Zhiwen Yang, Hui Zhang, Dan Zhao, Bingzheng Wei, and Yan Xu. Restore-rwkv: Efficient and effective medical image restoration with rwkv. *arXiv preprint arXiv:2407.11087*, 2024.

[Zamir *et al.*, 2022] 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.

[Zhai *et al.*, 2021] Shuangfei Zhai, Walter Talbott, Nitish Srivastava, Chen Huang, Hanlin Goh, Ruixiang Zhang, and Josh Susskind. An attention free transformer. *arXiv preprint arXiv:2105.14103*, 2021.

[Zhang *et al.*, 2017] Kai Zhang, Wangmeng Zuo, Yunjin Chen, Deyu Meng, and Lei Zhang. Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. *IEEE transactions on image processing*, 26(7):3142–3155, 2017.

[Zhang *et al.*, 2023] Kai Zhang, Yawei Li, Jingyun Liang, Jiezhang Cao, Yulun Zhang, Hao Tang, Deng-Ping Fan, Radu Timofte, and Luc Van Gool. Practical blind image denoising via swin-conv-unet and data synthesis. *Machine Intelligence Research*, 20(6):822–836, 2023.

[Zheng *et al.*, 2019] Yu-Bang Zheng, Ting-Zhu Huang, Xi-Le Zhao, Tai-Xiang Jiang, Tian-Hui Ma, and Teng-Yu Ji. Mixed noise removal in hyperspectral image via low-fibered-rank regularization. *IEEE Transactions on Geoscience and Remote Sensing*, 58(1):734–749, 2019.

[Zhou and Chen, 2024] Xudong Zhou and Tianxiang Chen. Bsbp-rwkv: Background suppression with boundary preservation for efficient medical image segmentation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 4938–4946, 2024.

[Zhou *et al.*, 2020] Yuqian Zhou, Jianbo Jiao, Haibin Huang, Yang Wang, Jue Wang, Honghui Shi, and Thomas Huang. When awgn-based denoiser meets real noises. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 13074–13081, 2020.

[Zhu *et al.*, 2024] 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.