DRWKV: Focusing on Object Edges for Low-Light Image Enhancement (Appendices & Supplementary Material)

Xuecheng Bai[1*] Yuxiang Wang[2*] Boyu Hu[3] Qinyuan Jie[1]
Chuanzhi Xu[2†] Kechen Li[4] Hongru Xiao[5] Vera Chung[2]

1. Supplementary Experiments

In this appendix, we first present a comprehensive exposition of the core computational pipeline of the DRWKV model introduced in the main text, together with the detailed implementation specifics of each key module. Subsequently, to rigorously validate the enhanced capabilities claimed by the proposed DRWKV model, we designed a series of extended experiments; the experimental protocol and the corresponding analyses of the results are elaborated in the ensuing sections.

2. Architecture Details of DRWKV

In this section, we provide a detailed exposition of the Light Preprocessing pipeline introduced in Section 3, the ES-RWKV computational flow, and the constituent fine-grained modules.

2.1. Light Preprocessing

Light Preprocessing takes a low-light image $I \in \mathbb{R}^{3\times H\times W}$ as input. First, under the gray-world hypothesis, it computes the global illumination component $L_{\text{ill}} \in \mathbb{R}^{C\times H\times W}$. This hypothesis assumes that, in a color-balanced scene, the average reflectance of all pixels tends toward neutral gray. Guided by this assumption, the global illumination map L_{ill}, which characterizes the spatial distribution of overall illumination intensity, is derived from the input image I and serves as the basis for the final enhancement.

The input image I is initially processed by a depthwise-separable 3×3 convolution layer with ReLU activation. The resulting feature map is then fed in parallel into three independent 1×1 convolution modules, each with a distinct activation function to produce specific components: a module with Sigmoid activation outputs the structured artifact component $S \in \mathbb{R}^{3\times H\times W}$, intended to capture potential structured interference patterns; a module with Sigmoid activation outputs the noise estimation component $N \in \mathbb{R}^{3\times H\times W}$, quantifying the noise contamination level; and a module with Tanh activation outputs the local illumination estimation component $L \in \mathbb{R}^{1\times H\times W}$, providing a finer, spatially adaptive representation of illumination intensity. The procedure is as follows:

$$B(I) = \mathrm{ReLU}\big(\mathrm{DepthConv}_{3\times3}(I)\big), \tag{1}$$

$$S = \mathrm{Sigmoid}\big(\mathrm{Conv}_{1\times1}(B^3(I))\big), \tag{2}$$

$$L = \mathrm{Sigmoid}\big(\mathrm{Conv}_{1\times1}(B^3(I))\big), \tag{3}$$

$$N = \mathrm{Tanh}\big(\mathrm{Conv}_{1\times1}(B^3(I))\big), \tag{4}$$

where $\mathrm{DepthConv}_{3\times3}(\cdot)$ denotes the 3×3 depthwise convolution, $\mathrm{Conv}_{1\times1}(\cdot)$ denotes the 1×1 convolution, $\mathrm{ReLU}(\cdot)$, $\mathrm{Sigmoid}(\cdot)$, and $\mathrm{Tanh}(\cdot)$ are the respective activation functions, and $B^3(\cdot)$ denotes three consecutive applications of $B(\cdot)$.

The noise estimation N, illumination estimation L, and the original input I are combined via

$$\hat{R} = \frac{I - N}{L}, \tag{5}$$

yielding the reconstructed reflectance $\hat{R} \in \mathbb{R}^{3\times H\times W}$. Then, I and $\hat{R}$ are combined to obtain the enhanced image $\hat{I} \in \mathbb{R}^{3\times H\times W}$:

$$\hat{I} = I \odot \hat{R}, \tag{6}$$

where $\odot$ denotes element-wise multiplication.

2.2. Evolving Scanning RWKV

Unlike traditional spatial modelling approaches, the proposed ES-RWKV breaks through spatial-size constraints, preserves structural integrity, enhances modelling capacity, and enables cross-space information interaction while markedly reducing the computational overhead of handling complex spatial features, thereby offering an efficient solution for spatial feature extraction and fusion in vision tasks.

*Equal contribution.
†Corresponding author: Chuanzhi Xu (chuanzhi.xu@sydney.edu.au).
Affiliations:
[1]Shenyang Ligong University, Shenyang, China
[2]The University of Sydney, NSW, Australia
[3]University of International Business and Economics, Beijing, China
[4]Nanjing University of Aeronautics and Astronautics, Nanjing, China
[5]Tongji University, Shanghai, China

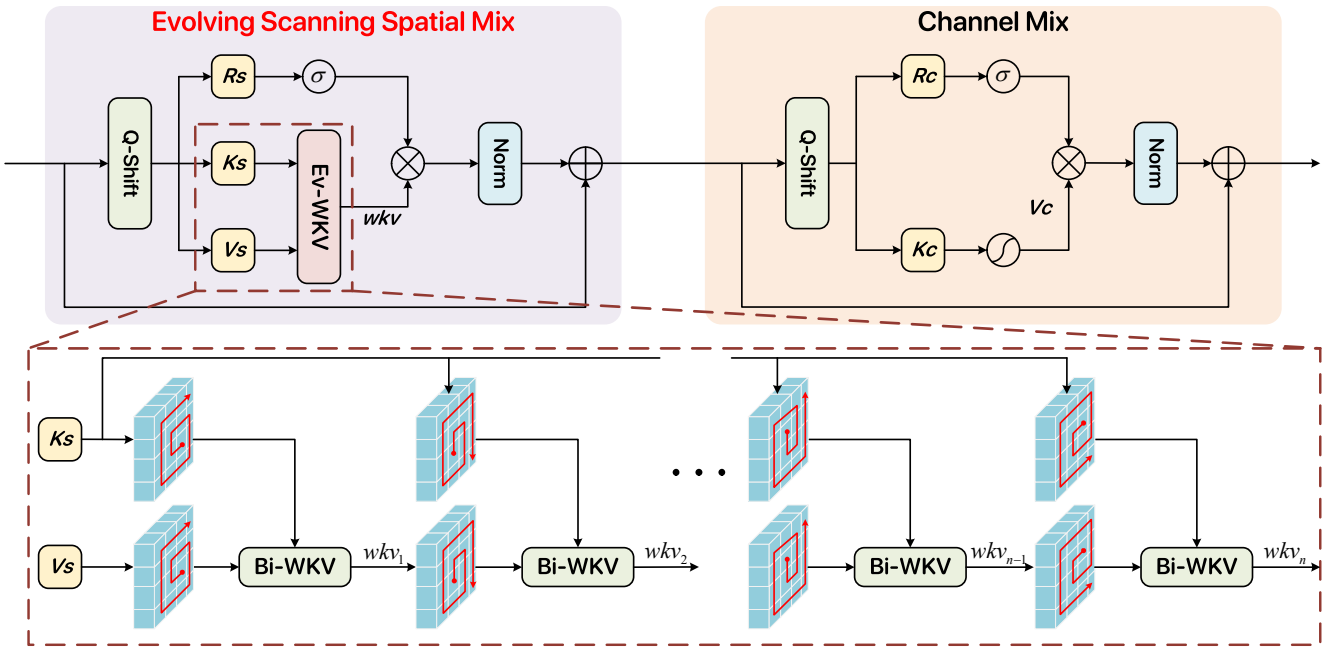

Figure 1. Evolving Scanning RWKV's structure, it consists of two parts: Evolving Scanning Spatial Mix and Channel Mix.

The ES-RWKV block is decomposed into two sequential stages: Evolving Scanning Spatial Mix and Channel Mix. The Evolving Scanning RWKV structure is shown in Figure.1

2.2.1 Evolving Scanning Spatial Mix

They tackles the dilemma between spatial-size limits and structural integrity. A channel-wise Q-Shift allows the model to adapt flexibly to different tasks without extra computation. Coupled with the EV-WKV mechanism, an embedded eight-spiral scanning pattern is introduced to capture spatial dependencies more comprehensively and efficiently, suppressing redundancy irrelevant to spatial extraction. Spatial information is then consolidated by a LayerNorm, enabling the model to automatically emphasize salient features and attenuate noise, thereby facilitating downstream layers in learning effective spatial patterns.

For an input feature $X \in \mathbb{R}^{T\times C}$, we augment the standard sequential processing framework with a channel-partitioned Q-Shift mechanism.By slicing and concatenating along the channel axis, we construct a shifted feature $X^{\dagger}$.Incorporating the learnable vector $\mu_{(\star)}$, we dynamically interpolate and fuse the original feature X and the shifted feature $X^{\dagger}$ to further transcend the spatial size limitations of traditional methods, preserve the integrity of spatial structures, and enhance the model's capability to capture cross-spatial positional information. This process yields feature components tailored for spatial mixing computations:

$$\textit{Q-Shift}_{(\star)}(X) = X + (1-\mu_{(\star)})X^{\dagger} \tag{7}$$

$$\begin{aligned} X^{\dagger}[h,w] = \textit{Concat}\big(&X[h-1,w,0:C/4],\\ &X[h+1,w,C/4:C/2],\\ &X[h,w-1,C/2:3C/4],\\ &X[h,w+1,3C/4:C]\big) \end{aligned} \tag{8}$$

In this context, $(*) \in \{R,K,V\}$ denotes three types of interpolation operations performed on X and $X^{\dagger}$. These operations are governed by μ and are utilized for the subsequent computation of R, K, and V. Here, h, w, and C represent the height, width, and number of channels, respectively.

Subsequently, after the Q-Shift operation X is decomposed into three independent components $R_s, K_s, V_s \in R^{T\times C}$

$$(*)_s = \textit{Q-Shift}_{(*)}(X)W_{(*)} = (X + (1-\mu_{(*)})X^{\dagger})W_{(*)} \tag{9}$$

$W_{(*)}$ denotes the weight. To ensure numerical stability, prevent training divergence, and accelerate training convergence, we introduce the LN layer. The total output of the Evolving Scanning Spatial Mix block is as follows:

$$O_s = X + LN((\sigma(R_s)\odot wkv)W_{O_s}), \tag{10}$$

$$\textit{where } wkv = \textit{EV-WKV}(K_s, V_s), \tag{11}$$

Here, σ denotes the sigmoid activation function, $\odot$ represents the element-wise multiplication operation, LN denotes layer normalization, and W_{O_s} denotes the weight. EV-WKV is an Evolving Scanning Mechanism proposed based on the linear scanning formula wkvt of Bi-WKV.

2.2.2 Channel Mix

We reintroduce the Q-Shift mechanism and combine it with the SquaredReLU activation function and linear projection. This is aimed at addressing the shortcomings that may exist after the spatial mixing in the first stage, such as insufficient exploration of correlations between channel features and oversimplified feature expression. It ensures more adequate feature fusion in the channel dimension, enhances the nonlinear expression ability and discriminability of features, and further maintains the stability of numerical distribution through Layer Normalization (LN), thereby providing richer and more robust inputs for feature learning in the deep layers of the model.

First, we take the feature Os output from the first stage as input, and construct the query feature Rc and key feature Kc for channel interaction respectively through the Q-Shift mechanism. The calculation process is as follows:

$$\begin{aligned} R_c &= Q - Shift_R(O_s)W_R, \\ K_c &= Q - Shift_K(O_s)W_K \end{aligned} \tag{12}$$

Here, *Q-Shift*$_R$ and *Q-Shift*$_K$ dynamically interpolate between the original feature O_s and its shifted version $O_s^{\dagger}$ using learnable parameters $\mu_{(R)}$ and $\mu_{(K)}$. This design not only preserves the spatial structural information in the original features but also introduces cross-position channel dependencies, enabling the model to adaptively focus on long-range correlations among different channels.

Subsequently, to enhance the nonlinear expression capability of features and alleviate the gradient vanishing problem, the *SquaredReLU* activation function is applied to for Kc transformation.

Further, the transformed features are projected into the value space via linear projection, and the output of the Channel Mix module is generated through residual connections and layer normalization operations. The detailed process is as follows:

$$O_c = O_s + LN((\sigma(R_c) \odot V_c)W_{O_c}), \tag{13}$$

$$where\ V_c = SquaredReLU(K_c) \tag{14}$$

2.3. Bilateral Spectrum Aligner Block

This subsection focuses on the three core components of the Bilateral Spectrum Aligner: Spectral Alignment Enhancer (SAE), Cross Attention (CA), and Scharr Edge Enhancement (SEE).

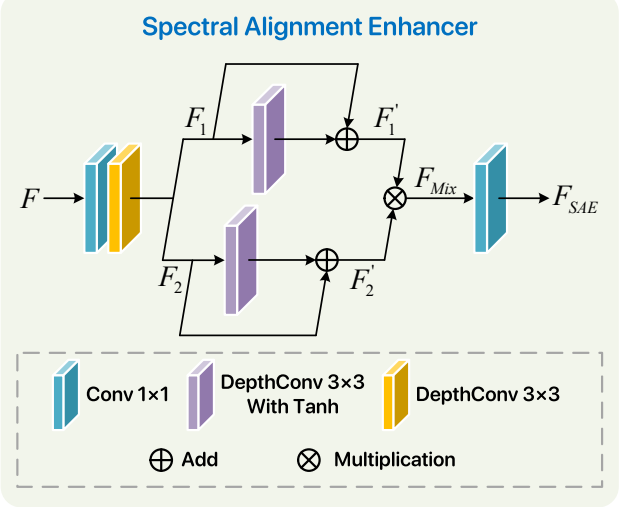

Figure 2. Spectral Alignment Enhancer's structure.

2.3.1 Spectral Alignment Enhancer (SAE)

To enhance the low-light image enhancement model's perception and processing of spectral features under varying illumination conditions, we propose an adaptive feature-enhancement module termed the Spectral Alignment Enhancer (SAE), illustrated in Figure 2. Specifically, the module achieves precise optimization of low-light features through three sequential stages: feature expansion, separated enhancement, and interactive fusion.

Initially, the input feature $F \in \mathbb{R}^{C\times H\times W}$ is channel-wise expanded via a 1×1 convolution and then split into two branches, thereby preserving sufficient expressive capacity for capturing faint spectral details in low-light environments.

$$F_1, F_2 = Split(DepthConv_{3\times3}(Conv_{1\times1}(F))) \tag{15}$$

Here, $Split(\cdot)$ denotes an even channel-wise halving,.

Subsequently, the two feature streams $F_1, F_2 \in \mathbb{R}^{C_1/2\times H\times W}$ are each processed by a depthwise-separable convolution to focus on local spatial patterns, followed by a nonlinear Tanh activation and a residual connection, resulting in refined features $F_1', F_2' \in \mathbb{R}^{C_1/2\times H\times W}$. This design allows the SAE to enhance the extraction of critical low-light cues, such as edges and textures, while preventing detail loss caused by excessive transformations. The procedure can be expressed as:

$$F_1' = F_1 + Tanh(DepthConv_{3\times3}(F_1)) \tag{16}$$

$$F_2' = F_2 + Tanh(DepthConv_{3\times3}(F_2)) \tag{17}$$

Finally, the two refined feature streams are multiplied element-wise to achieve dynamic interaction, adaptively reinforcing informative cues. A subsequent 1×1 convolution

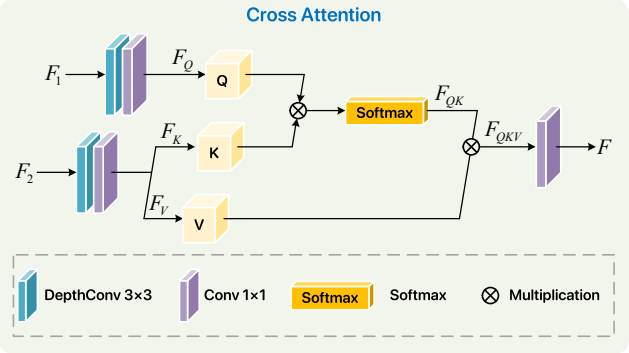

Figure 3. Cross Attention's structure.

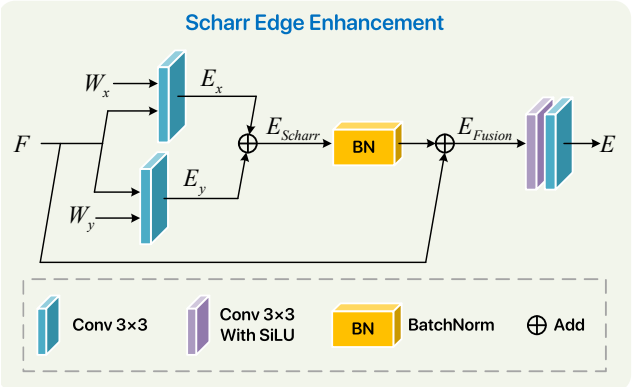

Figure 4. Scharr Edge Enhancement's structure.

then compresses the result back to the original channel dimension, yielding the optimized feature $F_{\text{SAE}} \in \mathbb{R}^{C\times H\times W}$. This process is formulated as:

$$F_{SAE} = Conv_{1\times1}(F_1' \times F_2') \tag{18}$$

2.3.2 Cross Attention (CA)

In low-light enhancement, the luminance stream x and the chrominance stream y are usually mis-aligned in both spatial statistics and semantic content. We propose a lightweight Cross-Attention (CA) block that lets x query relevant information from y through a multi-head, L2-normalized cross-covariance attention implemented with depth-wise convolutions only. The entire procedure is summarized below with strict correspondence to the released code, illustrated in Figure 3.

First, Given two feature maps:

$$x, y \in R^{C\times H\times W} \tag{19}$$

, where x originates from the brightness subnet and y from the color subnet.

Query / Key / Value Generation

All linear projections are instantiated as 1×1 convolutions followed by 3×3 depth-wise convolutions to retain local context while being parameter-efficient.

$$Q = DWConv_{3\times3}(Conv_{1\times1}(x)) \in R^{C\times H\times W} \tag{20}$$

$$KV = DWConv_{3\times3}(Conv_{1\times1}(y)) \in R^{2C\times H\times W} \tag{21}$$

The KV tensor is then channel-split into key and value:

$$K, V = split(KV, dim = 1), K, V \in R^{C\times H\times W} \tag{22}$$

Multi-Head Re-Shape and L2-Normalization

To exploit complementary attention patterns, we reshape each tensor into h heads:

$$Q_{head} = rearrange(Q) \in R^{n_h\times c_h\times(H\cdot W)} \tag{23}$$

$$K_{head} = rearrange(K) \in R^{n_h\times c_h\times(H\cdot W)} \tag{24}$$

$$V_{head} = rearrange(V) \in R^{n_h\times c_h\times(H\cdot W)} \tag{25}$$

To ensure training stability, L2-normalization is applied along the last dimension:

$$\hat{Q} = \frac{Q_{head}}{\|Q_{head}\|_2}, \hat{K} = \frac{K_{head}}{\|K_{head}\|_2} \tag{26}$$

Cross-Covariance Attention

We compute the cross-covariance attention matrix between x-query and y-key:

$$attn = softmax(\frac{\hat{Q}\hat{K}^T}{\tau}) \in R^{n_h\times c_h\times c_h} \tag{27}$$

where the temperature τ is a learnable scalar initialized to d The attended feature is then aggregated via matrix multiplication with y-value:

$$out = attn \cdot V_{head} \in R^{n_h\times c_h\times(H\cdot W)} \tag{28}$$

Output Projection

Finally, the multi-head output is re-arranged and projected back to the original channel dimension via a 1×1 convolution:

$$Output = Conv_{1\times1}(rearrange(out)) \in R^{C\times H\times W} \tag{29}$$

The resulting tensor is element-wise added to the luminance branch, yielding an enhanced feature that is both color-aware and noise-suppressed.

2.3.3 Scharr Edge Enhancement (SEE)

In the field of low-light image enhancement, the accurate preservation of edge contours and the complete recovery of texture details have always been core challenges that restrict the performance of algorithms. In low-light environments,

the image Signal-to-Noise Ratio (PSNR) decreases significantly, causing the gradient information in edge regions to be submerged by noise and thus become weak. Meanwhile, traditional enhancement methods, in the process of improving brightness, often further exacerbate edge blurring and texture loss due to excessive smoothing operations or noise amplification effects, ultimately leading to visual distortion in the enhanced results characterized by "improved brightness but lacking details", illustrated in Figure 4.

2.4. User Subjective Evaluation Experiment

To address the above issues, we propose a Scharr Edge Enhancement (SEE) module. By integrating the prior knowledge of traditional image processing with the feature learning capability of deep learning, it provides an effective solution for structure-aware enhancement in low-light scenarios. Specifically, given the input feature $F \in \mathbb{R}^{C\times H\times W}$, we first perform channel-wise filtering operations using prefixed Scharr convolution kernels and convolution weights W_x,W_y to simulate the gradient extraction process in traditional image processing, obtaining the x-direction gradient feature $E_x \in \mathbb{R}^{C\times H\times W}$ and y-direction gradient feature $E_y \in \mathbb{R}^{C\times H\times W}$ respectively.

Subsequently, to suppress noise interference and aggregate multi-directional edge information, we fuse E_x and E_y using the L1 norm to compute the Scharr edge strength feature $E_{Scharr} \in \mathbb{R}^{C\times H\times W}$.

Then, the edge strength feature is integrated with the original input feature through a residual mechanism to obtain the fused feature $E_{Fusion} \in \mathbb{R}^{C\times H\times W}$. This operation, on one hand, preserves the structural prior in the original feature; on the other hand, it avoids "edge overexposure" caused by excessive feature transformation under the guidance of edge information, thereby ensuring the naturalness of edge enhancement.

Finally, to achieve adaptive adjustment of feature dimensions and deep fusion of semantic information, 1×1 convolution and 3×3 convolution are used for dimension reduction and recovery, resulting in the final output $SEE_{out} \in \mathbb{R}^{C\times H\times W}$.

The specific calculation process can be expressed as follows:

$$E_x = Conv_{3\times3}(F, W_x), E_y = Conv_{3\times3}(F, W_y) \quad (30)$$

$$W_x = \begin{bmatrix} -3 & 0 & 3 \\ -10 & 0 & 10 \\ -3 & 0 & 3 \end{bmatrix}, W_y = \begin{bmatrix} -3 & -10 & -3 \\ 0 & 0 & 0 \\ 3 & 10 & 3 \end{bmatrix} \quad (31)$$

$$E_{Scharr} = \|E_x\|_1 + \|E_y\|_1 \quad (32)$$

$$E_{Fusion} = F + BN(E_{Scharr}) \quad (33)$$

$$SEE_{out} = Conv_{3\times3}(SiLU(Conv_{1\times1}(E_{Fusion}))) \quad (34)$$

Here, $Conv_{3\times3}(\cdot,\cdot)$ denotes a 3×3 convolution with fixed weights, $BN(\cdot)$ represents the BatchNorm operation, and $SiLU(\cdot)$ denotes the SiLU activation function.

3. The input of the Global Edge Retinex theory

In the main text, we elaborate and derive the proposed GER, and present the input structure in the main text figures: edge features E, artifacts S, and processed global illumination L_{ill}. As the artifacts S and processed global illumination L_{ill} have been thoroughly explained, in the appendix, we discuss the extraction and integration of edge features E.

To reduce the computational load and complexity of the model itself, we further utilize the SEE based on the output $I_{DRWKV} \in \mathbb{R}^{3\times H\times W}$ of the DRWKV model, thereby obtaining the integration of input edge features E suitable for Global Edge Retinex theory.

The specific calculation process can be expressed as follows:

$$E = SEE(I_{DRWKV}) \quad (35)$$

4. Extended experiment

4.1. User Subjective Evaluation Experiment

To rigorously assess the perceptual quality of enhancement algorithms under complex real world conditions we designed and conducted a user subjective evaluation experiment. Ten low light images were randomly selected from the publicly available LSRW Huawei benchmark and processed by seven enhancement models: RetinexNet, Zero-DCE, RRDNet, RetinexFormer, RetinexMamba, MambaIR and the proposed DRWKV to generate corresponding enhanced results. One hundred participants spanning diverse ages and professional backgrounds were then recruited. In an independent double blind protocol participants assigned scores on a five point Likert scale from one worst to five best along three key perceptual dimensions overall visual quality local detail restoration and global color fidelity. The mean opinion scores are reported in Figure 5. Across all three dimensions the proposed DRWKV model achieved the highest ratings indicating superior perceptual enhancement performance from the human visual perspective.

4.2. Low-light Object Tracking Experiment

The aforementioned experiments have demonstrated that DRWKV already exhibits superior performance in basic experimental scenarios. To investigate its generalization ability, we designed a low-light multi-object tracking task. For the dataset, we adopted the Drone low-light tracking benchmark, UAVDark135. It contains 135 video sequences in total, encompassing various tracking scenarios such as intersections and highways, as well as tracking targets including

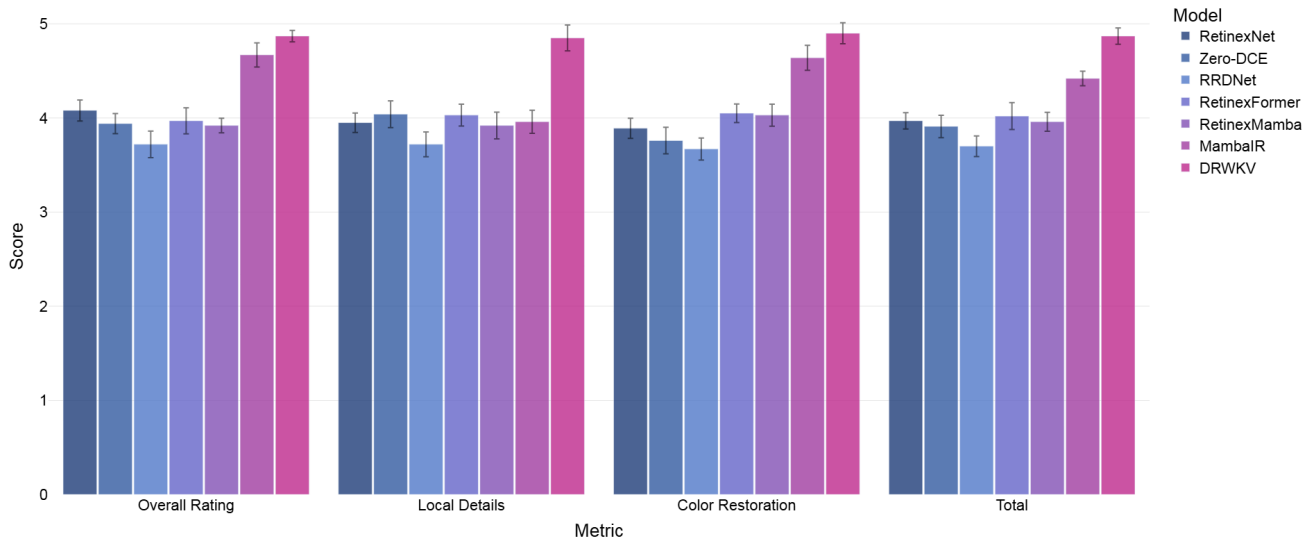

Figure 5. User subjective evaluation experiment.

pedestrians, boats, and vehicles. The videos are captured at a frame rate of 30 FPS with a resolution of 1920×1080.

In terms of methodology, we compared the performance of mainstream object tracking algorithms with and without the integration of the DRWKV low-light enhancement module, using MOTA (multi-object tracking accuracy), IDF1 (identity recognition F1 score), and HOTA (high-order tracking accuracy) as evaluation metrics.

The experimental conclusions show that DRWKV generally improved accuracy across multiple mainstream tracking algorithms, with particularly prominent performance in multi-object tracking and identity preservation. Meanwhile, the limitations of this module in optimizing Deep OC-SORT warn us that in the face of low adaptability and weak tuning of network structures, it may not be possible to achieve simultaneous improvements in detection accuracy and association capability.

4.3. α, β and γ Parameter Detail Experiment

To obtain the optimal configuration of the three parameters (α, β and γ), we designed a series of parameter detail experiments. For the dataset, we adopted the LOLv2-Real dataset. During the implementation of the parameter detail experiments, we only adjusted one of the three parameters (α, β and γ) while keeping the other two unchanged. The experimental results are presented in Table 1, Table 2, and Table 3. These results indicate that the optimal parameter configuration for DRWKV is $\alpha = 0.2$, $\beta = -0.05$, and $\gamma = 0.2$.

Table 1. The impact of different α values on the performance of the LOLv2-Real dataset.

α	0.05	0.10	0.15	**0.20**	0.25	0.30	0.35
PSNR↑	23.56	23.89	24.02	**24.12**	23.84	23.97	23.94
SSIM↑	0.798	0.815	0.814	**0.832**	0.792	0.821	0.814
NIQE↓	4.012	3.967	3.941	**3.926**	3.951	3.989	4.173

Table 2. The impact of different β values on the performance of the LOLv2-Real dataset.

β	-0.100	-0.075	**-0.050**	-0.025	0
PSNR↑	23.72	23.98	**24.12**	23.86	23.61
SSIM↑	0.803	0.821	**0.832**	0.817	0.804
NIQE↓	3.995	3.948	**3.926**	3.973	4.003

Table 3. The impact of different γ values on the performance of the LOLv2-Real dataset.

γ	0.05	0.10	0.15	**0.20**	0.25	0.30	0.35
PSNR↑	22.74	23.65	23.91	**24.12**	23.99	23.78	23.84
SSIM↑	0.804	0.801	0.823	**0.832**	0.826	0.812	0.814
NIQE↓	4.112	4.023	3.957	**3.926**	3.962	3.998	4.472