GFRRN: Explore the Gaps in Single Image Reflection Removal

Supplementary Material

A. The details of Mona Layer

Fig. 1 illustrates the insert position of the Mona layer [7] within the SwinBlock [6] and presents its detailed structure. The Mona layer is inserted after the Attention and Feed-Forward Network to fine-tune their outputs. The core of the Mona layer is a set of visual filters composed of three multi-scale depthwise convolutions (3×3, 5×5, 7×7) and a pointwise convolution (1×1), which collectively capture multi-scale visual information. More sophisticated convolution group can be designed to enhance the performance, which is not the focus of this paper.

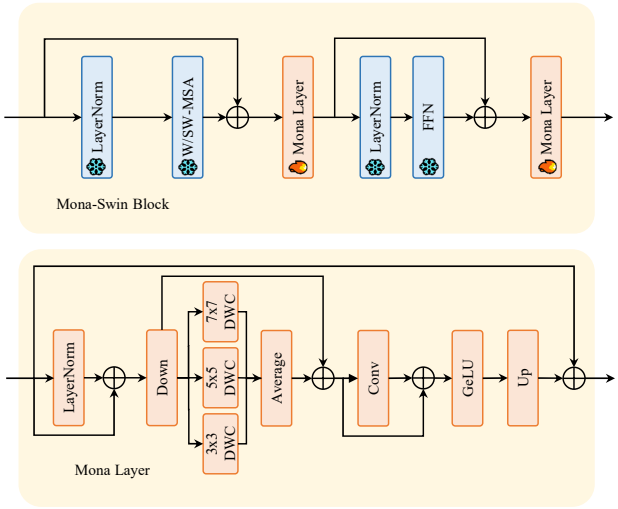

Figure 1. Top: The insert position of the Mona layer within the SwinBlock. Bottom: The detail of Mona layer.

B. The overall structure of decoder

Alg. 1 outlines the overall workflow of a certain level in the decoder. Each level consists of the G-AFLB and the Dual-stream Dynamic Interaction Block (DDIB). The DDIB is a dual-stream Transformer block that incorporates DAA and Layer-wise DAA (i.e., LDAA).

C. The details of G-AFLB

The Adaptive Frequency Learning Block (AFLB) [1] can be divided into Frequency Mining Module (FMiM) and Frequency Modulation Module (FMoM). Its function is to modulate the feature maps using the frequency information of the input image $\mathbf{I}$. Specifically, the FMiM adaptively separates the high-low frequency components of the enhanced input image, while the FMoM modulates the input feature map with the separated frequency components, enabling the feature map to explicitly carry frequency information. G-AFLB's key is the FMiM, which separates high-low frequency through Gaussian low-pass filter. The details are shown in Fig. 2. This process can be formally represented as follows:

$$\begin{aligned} F_{low}, F_{high} &= \text{FMiM}(\mathbf{I}), \\ X_{low} &= \text{CrossAttention}\left(F_{low}, X_{in}\right), \\ X_{high} &= \text{CrossAttention}\left(F_{high}, X_{in}\right), \\ X_{out} &= \text{FMoM}\left(F_{low}, F_{high}, X_{in}\right). \end{aligned} \tag{1}$$

The original FMiM used rectangular masks in the fre-

Algorithm 1 Structure of Certain Level of Decoder

Require: Input features $F_{\mathbf{T}}^0$, $F_{\mathbf{R}}^0$ and degraded image $\mathbf{I}$
Ensure: Output features $F_{\mathbf{T}}^K$ and $F_{\mathbf{R}}^K$

Step1: Apply G-AFLB
  $F_{\mathbf{T}}^1 = \text{G-AFLB}(F_{\mathbf{T}}^0, \mathbf{I})$
  $F_{\mathbf{R}}^1 = \text{G-AFLB}(F_{\mathbf{R}}^0, \mathbf{I})$
Step2: Apply K DDIB
  $F_{\mathbf{T}}^{current} = F_{\mathbf{T}}^1$, $F_{\mathbf{R}}^{current} = F_{\mathbf{R}}^1$
for $i = 1$ **to** K **do**
    LayerNorm:
      $F_{\mathbf{T}}^{LN} = LN(F_{\mathbf{T}}^{current})$
      $F_{\mathbf{R}}^{LN} = LN(F_{\mathbf{R}}^{current})$
    Combine tokens:
      $X_0^{LN} = \text{Concat}([F_{\mathbf{T}}^{LN}, F_{\mathbf{R}}^{LN}], dim = 0)$
      $X_1^{LN} = \text{Concat}([F_{\mathbf{T}}^{LN}, F_{\mathbf{R}}^{LN}], dim = 1)$
    Apply attention:
      $X_0^{SA} = DAA(X_0^{LN})$
      $X_1^{CA} = LDAA(X_1^{LN})$
    Split back:
      $F_{\mathbf{T}}^{SA}, F_{\mathbf{R}}^{SA} = \text{Split}(X_0^{SA}, dim = 0)$
      $F_{\mathbf{T}}^{CA}, F_{\mathbf{R}}^{CA} = \text{Split}(X_1^{CA}, dim = 1)$
    Combine the dual-attention results:
      $F_{\mathbf{T}}^{DA} = F_{\mathbf{T}}^{current} + F_{\mathbf{T}}^{SA} + F_{\mathbf{T}}^{CA}$
      $F_{\mathbf{R}}^{DA} = F_{\mathbf{R}}^{current} + F_{\mathbf{R}}^{SA} + F_{\mathbf{R}}^{CA}$
    Apply FFN:
      $F_{\mathbf{T}}^{LN'} = LN(F_{\mathbf{T}}^{DA})$
      $F_{\mathbf{R}}^{LN'} = LN(F_{\mathbf{R}}^{DA})$
      $F_{\mathbf{T}}^{FFN}, F_{\mathbf{R}}^{FFN} = \text{DSLPBlock}(F_{\mathbf{T}}^{LN'}, F_{\mathbf{R}}^{LN'})$
    Output:
      $F_{\mathbf{T}}^{current} = F_{\mathbf{T}}^{DA} + F_{\mathbf{T}}^{FFN}$
      $F_{\mathbf{R}}^{current} = F_{\mathbf{R}}^{DA} + F_{\mathbf{R}}^{FFN}$
end for
Output: $F_{\mathbf{T}}^K = F_{\mathbf{T}}^{current}$, $F_{\mathbf{R}}^K = F_{\mathbf{R}}^{current}$

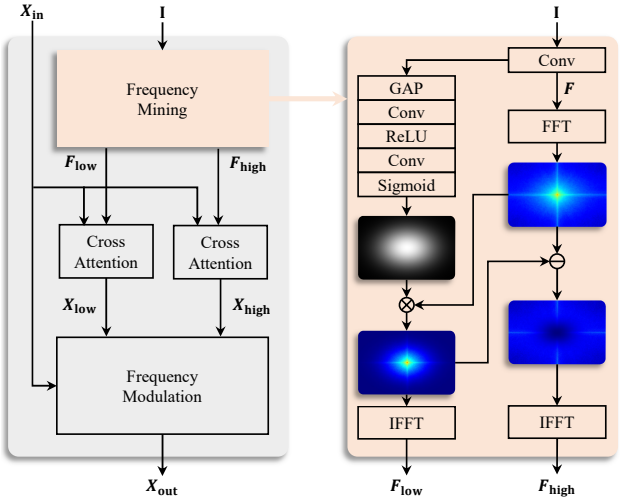

Figure 2. Overall structure of G-AFLB and details of FMiM.

quency domain to separate the high-low frequency information of the input image **I**. However, this can lead to ringing artifacts in the corresponding spatial domain image. The specific derivation is provided below:

Let (x, y) denote spatial coordinates and (ω_x, ω_y) denote the corresponding frequency domain coordinates. Let $f(x, y)$ be a function in the spatial domain, and $F(\omega_x, \omega_y)$ be its frequency domain response. The formulas for the Fourier transform and its inverse are:

$$\begin{aligned} F(\omega_x, \omega_y) &= \iint_{\mathbb{R}^2} f(x, y) e^{-i(\omega_x x + \omega_y y)} dx dy, \\ f(x, y) &= \frac{1}{(2\pi)^2} \iint_{\mathbb{R}^2} F(\omega_x, \omega_y) e^{i(\omega_x x + \omega_y y)} d\omega_x d\omega_y. \end{aligned} \tag{2}$$

According to the convolution theorem, multiplication in the frequency domain is equivalent to convolution in the spatial domain. Therefore, multiplying by a mask in the frequency domain is equivalent to convolving with the spatial domain response of that mask in the spatial domain:

$$\mathcal{F}\{f * g\} = F(\omega_x, \omega_y) \cdot G(\omega_x, \omega_y). \tag{3}$$

A frequency domain mask is essentially a low-pass filter. The frequency domain expression for an ideal rectangular low-pass filter is:

$$H_{Rec}(\omega_x, \omega_y) = \begin{cases} 1, & |\omega_x| \le \omega_{c_x} \text{ and } |\omega_y| \le \omega_{c_y} \\ 0, & \text{else} \end{cases}. \tag{4}$$

Substituting H_{Rec} into the inverse Fourier transform formula yields its spatial domain impulse response:

$$\begin{aligned} h_{Rec}(x, y) &= \frac{1}{(2\pi)^2} \iint_{|\omega_x| \le \omega_{c_x}, |\omega_y| \le \omega_{c_y}} e^{i(\omega_x x + \omega_y y)} d\omega_x d\omega_y \\ &= \frac{1}{(2\pi)^2} \left(\int_{-\omega_{c_x}}^{\omega_{c_x}} e^{i\omega_x x} d\omega_x \right) \left(\int_{-\omega_{c_y}}^{\omega_{c_y}} e^{i\omega_y y} d\omega_y \right) \\ &= \frac{1}{\pi^2} \frac{\sin(\omega_{c_x} x)}{x} \cdot \frac{\sin(\omega_{c_y} y)}{y}. \end{aligned} \tag{5}$$

This result shows that the spatial domain response of the ideal rectangular low-pass filter is a two-dimensional Sinc function. This function is not single-peaked and oscillates between positive and negative values. When convolved with image edges, this oscillatory response causes oscillations in the image, manifesting as the ringing artifact.

The frequency domain representation of a Gaussian low-pass filter is:

$$H_G(\omega_x, \omega_y) = e^{-\frac{1}{2}\left(\frac{\omega_x^2}{\sigma_x^2} + \frac{\omega_y^2}{\sigma_y^2}\right)}, \quad \sigma_x > 0, \sigma_y > 0. \tag{6}$$

Substituting this into the inverse transform formula:

$$\begin{aligned} h_G(x, y) &= \frac{1}{(2\pi)^2} \iint_{\mathbb{R}^2} e^{-\frac{1}{2}\left(\frac{\omega_x^2}{\sigma_x^2} + \frac{\omega_y^2}{\sigma_y^2}\right)} e^{i(\omega_x x + \omega_y y)} d\omega_x d\omega_y \\ &= \frac{1}{(2\pi)^2} \left(\int_{\mathbb{R}} e^{-\frac{\omega_x^2}{2\sigma_x^2}} e^{i\omega_x x} d\omega_x \right) \left(\int_{\mathbb{R}} e^{-\frac{\omega_y^2}{2\sigma_y^2}} e^{i\omega_y y} d\omega_y \right). \end{aligned} \tag{7}$$

Using the Gaussian integral formula:

$$\int_{-\infty}^{\infty} e^{-a\omega^2} e^{j\omega x} d\omega = \sqrt{\frac{\pi}{a}} e^{-\frac{x^2}{4a}}. \tag{8}$$

For the x-direction integral, let $a = \frac{1}{2\sigma_x^2}$. For the y-direction integral, let $a = \frac{1}{2\sigma_y^2}$. This yields:

$$\int_{\mathbb{R}} e^{-\frac{\omega_x^2}{2\sigma_x^2}} e^{j\omega_x x} d\omega_x = \sqrt{2\pi\sigma_x^2} e^{-\frac{\sigma_x^2 x^2}{2}}, \tag{9}$$

$$\int_{\mathbb{R}} e^{-\frac{\omega_y^2}{2\sigma_y^2}} e^{j\omega_y y} d\omega_y = \sqrt{2\pi\sigma_y^2} e^{-\frac{\sigma_y^2 y^2}{2}}. \tag{10}$$

Combining these results:

$$\begin{aligned} h_G(x, y) &= \frac{1}{(2\pi)^2} \cdot \sqrt{2\pi\sigma_x^2} e^{-\frac{\sigma_x^2 x^2}{2}} \cdot \sqrt{2\pi\sigma_y^2} e^{-\frac{\sigma_y^2 y^2}{2}} \\ &= \frac{\sigma_x \sigma_y}{2\pi} e^{-\frac{1}{2}(\sigma_x^2 x^2 + \sigma_y^2 y^2)}. \end{aligned} \tag{11}$$

Therefore, the spatial domain response of the Gaussian frequency domain filter is still a two-dimensional Gaussian function. The Gaussian function is single-peaked and always positive. Convolving this response with a spatial image results in blurring without introducing ringing artifacts.

D. The details of DAA and LDAA

We propose the DAA mechanism, a novel self-attention variant that accounts for content differences across windows. The overall procedure is outlined in Alg. 2, where DWC denotes depthwise convolution.

A lightweight WIE module is introduced to predict the difference in reflection intensity between windows. This design can be generalized to all window-based attention mechanisms, with detailed structure shown in Fig. 3. The Q vector is reshaped to dimensions (N_w, C, H_w, W_w) and fed into the WIE module, where N_w represents the number of windows, and H_w and W_w denote the height and width of each window, respectively. The WIE module outputs a vector of shape $(N_w, 1)$, assigning a weight to each window. For visualization purposes, the resulting weights are remapped to the spatial dimensions $(1, 1, H, W)$, where H and W are the height and width of the feature map.

We further extend the DAA to a method termed LDAA, which enables dual-stream interaction between features of the transmissive layer and the reflective layer. The complete procedure is described in Alg. 3.

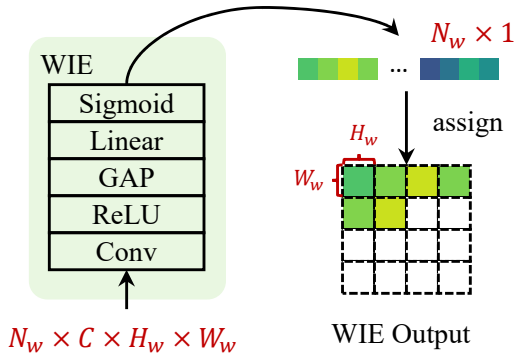

Figure 3. The structure of our proposed WIE.

E. New benchmark

We have additionally captured a new testing dataset named GF40. It consists of 40 image pairs. Each pair includes a superimposed image **I** and its transmission layer image **T**.

The capturing process is illustrated in Fig. 4. Following [9], we first capturing **T** by blocking the light source on the same side of the camera, and then capturing **I** without any obstruction. This approach can prevent pixel misalignment caused by glass refraction.

We test and compare the state-of-the-art methods on the GF40 dataset. As shown in Tab. 1, the quantitative comparisons are provided, while qualitative results are visualized in Fig. 5.

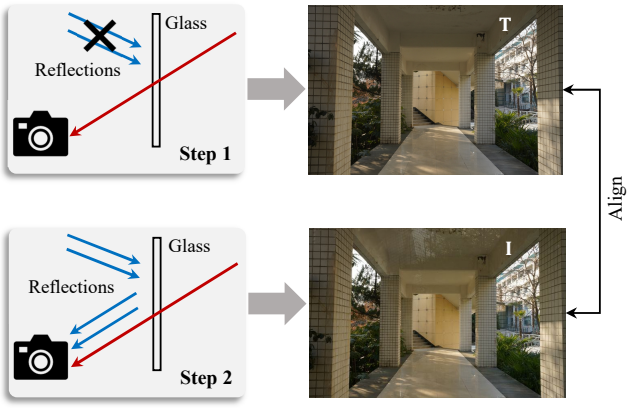

Figure 4. The capturing process of GF40.

Table 1. Performance comparison of our GFRRN and current state-of-the-art methods on GF40. We use the pre-trained models they provided for evaluation. The best results are displayed in **bold**, while the second-best are underlined.

Methods	Venues	PSNR	SSIM
DSRNet [2]	ICCV'23	23.79	0.860
DURRNet [4]	ICASSP'24	22.43	0.815
RRW [9]	CVPR'24	23.65	0.863
DSIT [3]	NeurIPS'24	24.98	0.868
DExNet [5]	TPAMI'25	23.16	0.845
RDNet [8]	CVPR'25	24.35	0.856
GFRRN	-	**25.95**	**0.876**

Algorithm 2 Dynamic Agent Attention

Require: Input tensor $x_{in} \in \mathbb{R}^{2N_w \times H_w W_w \times C}$
Ensure: Output tensor $x_{out} \in \mathbb{R}^{2N_w \times H_w W_w \times C}$
Step1: Compute Q, K, V
  $Q, K, V = \text{Linear}(x_{in})$
Step2: Agent Generation
  $A = \text{AgentGenerate}(Q)$
Step3: Window Importance Estimation
  $score = \text{WIE}(Q)$
  $A_w = A \cdot score$
Step4: Agent Aggregation
  $V_A = \text{softmax}(A_w \cdot K^\top + bias) \cdot V$
Step5: Agent Broadcast
  $F^{attn} = \text{softmax}(Q \cdot A_w^\top + bias) \cdot V_A$
Step6: Feature Enhancement
  $F^{dwc} = \text{DWC}(V)$
Step7: Output Projection
  $x_{out} = \text{Linear}(F^{attn} + F^{dwc})$
Output: x_{out}

Algorithm 3 Layer-wise Dynamic Agent Attention

Require: Input tensor $x_{in} \in \mathbb{R}^{N_w \times 2H_w W_w \times C}$
Ensure: Output tensors $x_{out} \in \mathbb{R}^{N_w \times 2H_w W_w \times C}$

Step1: Compute Q, K, V
  $Q, K, V = \text{Linear}(x_{in})$
Step2: Layer-wise Agent Generation
  $Q_{\mathbf{T}}, Q_{\mathbf{R}} = \text{LayerSeparate}(Q)$
  $A_{\mathbf{T}} = \text{AgentGenerate}(Q_{\mathbf{T}})$
  $A_{\mathbf{R}} = \text{AgentGenerate}(Q_{\mathbf{R}})$
  $A = \text{LayerCombine}(A_{\mathbf{T}}, A_{\mathbf{R}})$
Step3: Layer-wise Window Importance Estimation
  $score_{\mathbf{T}} = \text{WIE}(Q_{\mathbf{T}}), score_{\mathbf{R}} = \text{WIE}(Q_{\mathbf{R}})$
  $score = \text{Average}(score_{\mathbf{T}}, score_{\mathbf{R}})$
  $A_w = A \cdot score$
Step4: Layer-wise Agent Aggregation
  $V_A = \text{softmax}(A_w K^\top + bias_{layered})V$
Step5: Layer-wise Agent Broadcast
  $F^{attn} = \text{softmax}(QA_w^\top + bias_{layered})V_A$
Step6: Layer-wise Feature Enhancement
  $V_{\mathbf{T}}, V_{\mathbf{R}} = \text{LayerSeparate}(V)$
  $F_{\mathbf{T}}^{dwc} = \text{DWC}(V_{\mathbf{T}})$
  $F_{\mathbf{R}}^{dwc} = \text{DWC}(V_{\mathbf{R}})$
  $F^{dwc} = \text{LayerCombine}(F_{\mathbf{R}}^{dwc}, F_{\mathbf{T}}^{dwc})$
Step7: Output Projection
  $x_{out} = \text{Linear}(F^{attn} + F^{dwc})$
Output: x_{out}

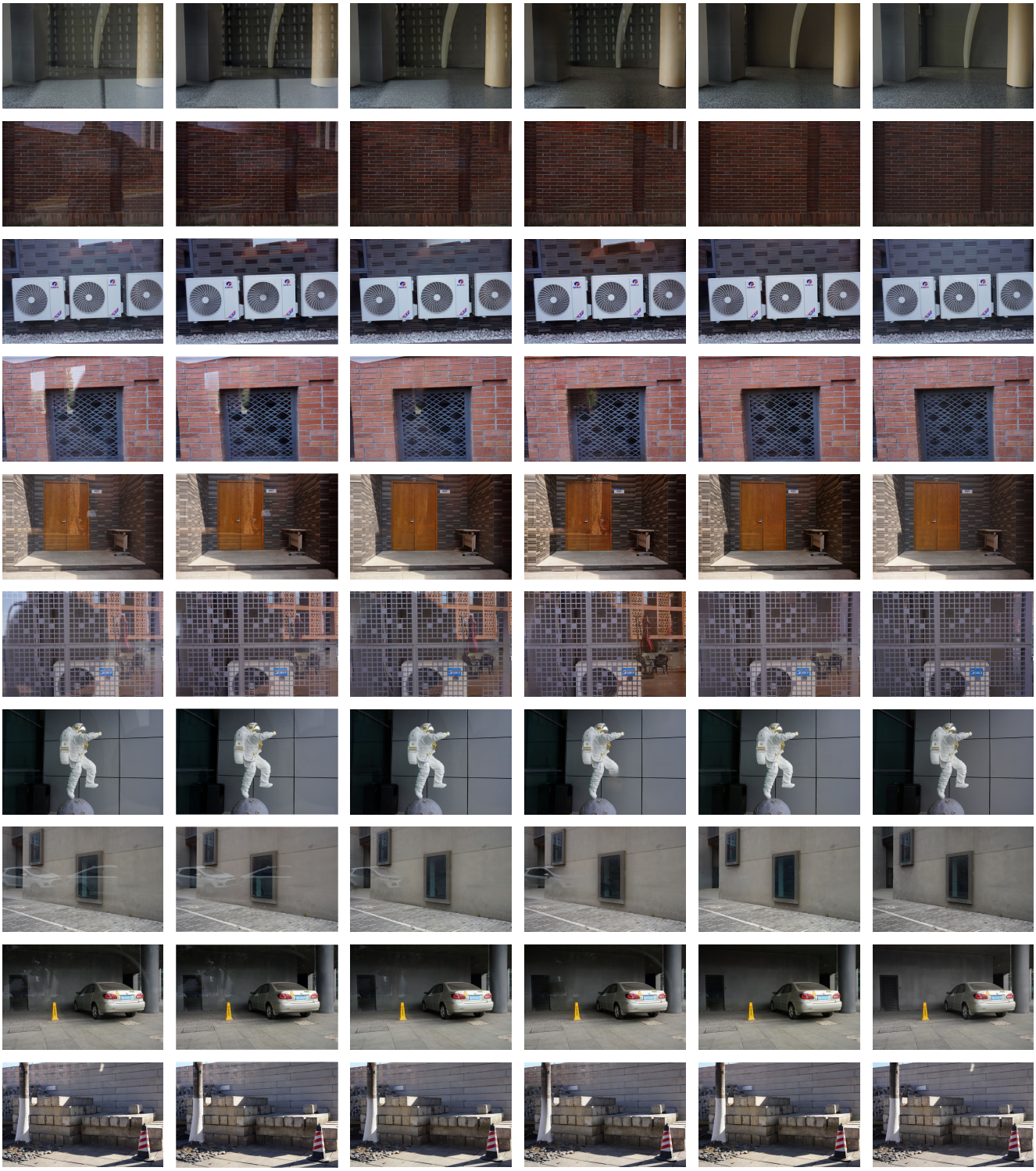

Figure 5. The visual results of RRW, DSIT, RDNet and our GFRRN on samples from GF40.

References

[1] Yuning Cui, Syed Waqas Zamir, Salman Khan, Alois Knoll, Mubarak Shah, and Fahad Shahbaz Khan. Adair: Adaptive all-in-one image restoration via frequency mining and modulation. In *ICLR*, pages 57335–57356, 2025. 1

[2] Qiming Hu and Xiaojie Guo. Single Image Reflection Separation via Component Synergy. In *ICCV*, pages 13138–13147, 2023. 3

[3] Qiming Hu, Hainuo Wang, and Xiaojie Guo. Single Image Reflection Separation via Dual-Stream Interactive Transformers. In *NeurIPS*, pages 55228–55248, 2024. 3

[4] Jun-Jie Huang, Tianrui Liu, Jingyuan Xia, Meng Wang, and Pier Luigi Dragotti. Durrnet: Deep unfolded single image reflection removal network with joint prior. In *ICASSP*, pages 5235–5239. IEEE, 2024. 3

[5] Jun-Jie Huang, Tianrui Liu, Zihan Chen, Xinwang Liu, Meng Wang, and Pier Luigi Dragotti. A lightweight deep exclusion unfolding network for single image reflection removal. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(6):4957–4973, 2025. 3

[6] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, pages 10012–10022, 2021. 1

[7] Dongshuo Yin, Leiyi Hu, Bin Li, Youqun Zhang, and Xue Yang. 5%>100%: Breaking Performance Shackles of Full Fine-Tuning on Visual Recognition Tasks. In *CVPR*, pages 20071–20081, 2025. 1

[8] Hao Zhao, Mingjia Li, Qiming Hu, and Xiaojie Guo. Reversible decoupling network for single image reflection removal. In *CVPR*, pages 26430–26439, 2025. 3

[9] Yurui Zhu, Xueyang Fu, Peng-Tao Jiang, Hao Zhang, Qibin Sun, Jinwei Chen, Zheng-Jun Zha, and Bo Li. Revisiting Single Image Reflection Removal In the Wild. In *CVPR*, pages 25468–25478, 2024. 3