BiMaCoSR: Binary One-Step Diffusion Model Leveraging Flexible Matrix Compression for Real Super-Resolution

Kai Liu [*1] **Kaicheng Yang** [*1] **Zheng Chen** [1] **Zhiteng Li** [1]
Yong Guo [2] **Wenbo Li** [3] **Linghe Kong** [†1] **Yulun Zhang** [†1]

A. One Step Diffusion Distillation

A.1. Diffusion Process Formulation

Following ResShift and SinSR(Yue et al., 2024; Wang et al., 2024), the diffusion process incorporates the low-resolution (LR) image y into the noisy step x_t. The forward diffusion process is formulated as:

$$q(x_t|x_0, y) = \mathcal{N}(x_t; x_0 + \eta_t(y - x_0), \kappa^2\eta_t\mathbf{I}), \quad (1)$$

where η_t is a time-dependent hyperparameter. The initial state x_T integrates information from y:

$$x_T = y + \kappa\sqrt{\eta_T}\epsilon, \quad \epsilon \sim \mathcal{N}(0, \mathbf{I}). \quad (2)$$

The reverse process follows:

$$\hat{x}_0 = f_\theta(x_t, y, t) \quad (3)$$

where the coefficients k_t, m_t, j_t are defined as:

$$\begin{cases} m_t = \sqrt{\frac{\eta_{t-1}}{\eta_t}}, \\ j_t = \eta_{t-1} - \sqrt{\eta_{t-1}\eta_t}, \\ k_t = 1 - \eta_{t-1} + \sqrt{\eta_{t-1}\eta_t} - \sqrt{\frac{\eta_{t-1}}{\eta_t}}. \end{cases} \quad (4)$$

Here, $\hat{x}_0 = f_\theta(x_t, y, t)$ represents the estimated HR image obtained from a pre-trained ResShift model. By iteratively sampling from the reformulated process in Eq. 3, a deterministic multi-steps mapping process between x_T (or ϵ) and $\hat{x}_0$ can be obtained, which is denoted as $F_\theta(x_T, y)$.

A.2. Distillation

Following SinSR(Wang et al., 2024), we utilize a student network $\hat{f}_\theta$ to learn the deterministic mapping F_θ between the randomly initialized state x_T and its deterministic output $F_\theta(x_T, y)$ from a teacher diffusion model. The distillation loss is defined as:

$$\mathcal{L}_{distill} = L_{MSE}(\hat{f}_\theta(x_T, y, T), F_\theta(x_T, y)). \quad (5)$$

[*]Equal contribution [1]Shanghai Jiao Tong University [2]South China University of Technology [3]Huawei Noah's Ark Lab. Correspondence to: Linghe Kong <linghe.kong@sjtu.edu.cn>, Yulun Zhang <yulun100@gmail.com>.

B. More Visualizations on Three Branches

We provide more visual comparison result in this section to demonstrate the role of each branch in this section.

B.1. Rank of Paramter Matrix

Firstly, we provide the rank numbers of the parameter matrices in the LRMB and SMB branches to demonstrate that SMB effectively enhances the diversity of transmitted information in Fig 1.When only LRMB is used, the rank of the parameter matrix is very low, which means that a significant amount of activation information is lost. However, after combining SMB, we can transmit high-rank FP activation information. This implies that LRMB serves as a mechanism for integrating the primary information, while SMB preserves the diversity of information.

Second, we provide a mathematical estimation of the rank of the parameter matrix in the BMB branch.For an $n \times n$ fully random binary matrix A, where each entry independently takes the value -1 with probability p and 1 with probability $1 - p$, the expected rank satisfies:

$$\mathbb{E}[\text{rank}(A)] = n - O(\log n). \quad (6)$$

This indicates that such matrices are **almost full rank** with high probability, as the number of linearly dependent rows is typically small. This further implies that BMB is responsible for extracting information from the full-precision activation stream for computation and generation.

B.2. Frequency-Domain Analysis of Three Branches

Frequency-domain analysis is crucial for understanding the roles of different layers, especially in super-resolution tasks. In this context, low-frequency information corresponds to the low-resolution image, while the primary goal of SR is to generate appropriate high-frequency details. By analyzing different frequency components, we can better interpret how each branch contributes to the reconstruction process.

The high and low-frequency proportions are computed by first summing the activations from BMB, LRMB, and SMB to obtain the total activation. A 2D Fast Fourier Transform (FFT) is applied, and the frequency components are shifted to center the low-frequency region. A binary mask is then used, where the central $\frac{1}{4}$ region represents low frequencies,

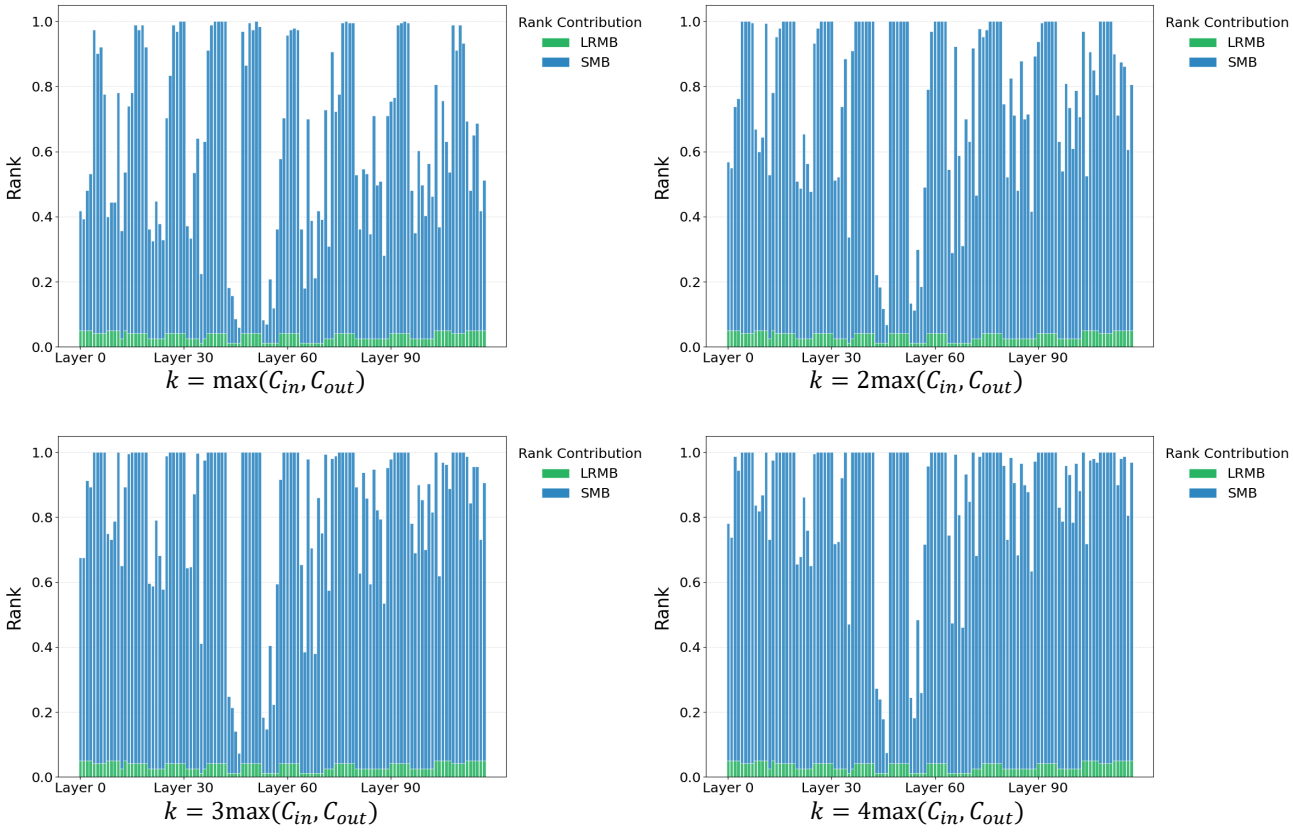

Figure 1: $\frac{r}{r_{\max}}$ of LRMB and SMB. $r_{\max} = \max(C_{in}, C_{out})$.

and the surrounding $\frac{3}{4}$ represents high frequencies. The proportion is determined by summing the spectral energy within each region, providing insight into the frequency contributions of different branches.

As illustrated in Fig 2, the contributions of the three branches vary across different layer types and frequency components. From the perspective of layer types, BMB exhibits a significantly higher contribution in convolutional layers, regardless of low- or high-frequency components. This aligns with prior studies indicating that convolutional layers are inherently more quantization-friendly. In convolutional layers, BMB alone accounts for the majority of the contribution. In contrast, in MLP (linear) layers, our proposed auxiliary branches, LRMB and SMB, contribute a larger proportion of activations, suggesting their role in compensating for the limitations of BMB in such layers. This confirms that LRMB and SMB effectively address ***Issue I***.

From the perspective of frequency components, BMB consistently exhibits stronger contributions to high-frequency components across both convolutional and MLP layers. This indicates that BMB primarily functions as an **information generator**, producing high-frequency details. Meanwhile, LRMB and SMB primarily facilitate **information delivery**, ensuring critical details are preserved and refined, particularly when BMB alone struggles to maintain performance.

C. Training Details

C.1. Training Time

Our model, BiMaCoSR, requires only 20GB of VRAM to train with a batch size of 8. The training time is provided in Table 1a, conducted on one NVIDIA RTX A6000.

C.2. Training Curve Details

We provide training curve plots for the ablation experiments in Fig 3, highlighting their potential applications. All experiments were conducted for 100K steps, but some data was lost, so partial trimming was applied.

D. Visual Results

We provide more visual comparison result in this section, shown in Fig 4 and Fig 5. All results demonstrate the outstanding performance of our proposed BiMaCoSR.

E. Limitations and Future Work

E.1. Limitation

Our model has the potential to implement layer-wise adaptive quantization strategies, leveraging the distinct characteristics of different network layers. However, in the current design, we adopt a uniform configuration across all layers, which may limit the model's ability to fully exploit these layer-specific characteristics. Determining critical hyperparameters, such as the rank r in LRMB and the number of selected elements k in SMB, remains a challenging task due to the large search space. Moreover, our binarization approach has not yet been evaluated on larger-scale pre-trained diffusion models, and the proposed method has not been tested on hardware for practical deployment.

E.2. Future Work

In future research, we will explore adaptive quantization strategies tailored to different layers to achieve more efficient compression. We also plan to extend our approach to larger-scale diffusion models and conduct comprehensive evaluations on hardware to assess its real-world applicability.

References

Wang, Y., Yang, W., Chen, X., Wang, Y., Guo, L., Chau, L.-P., Liu, Z., Qiao, Y., Kot, A. C., and Wen, B. Sinsr: diffusion-based image super-resolution in a single step. In *CVPR*, 2024.

Yue, Z., Wang, J., and Loy, C. C. Resshift: Efficient diffusion model for image super-resolution by residual shifting. *NeurIPS*, 2024.

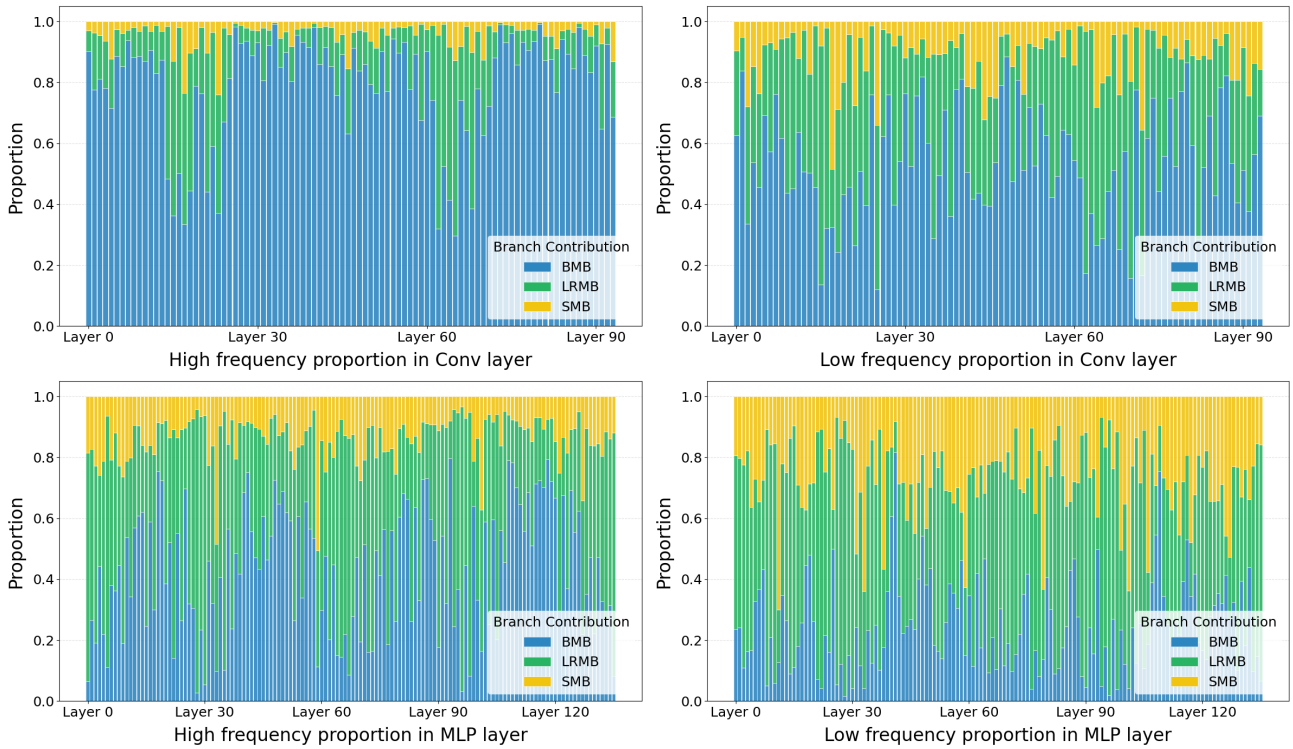

Figure 2: Branch contributions in different frequency domains across layers. The figure show the proportion of contributions from BMB, LRMB, and SMB in both high and low frequency components within convolutional and MLP layers.

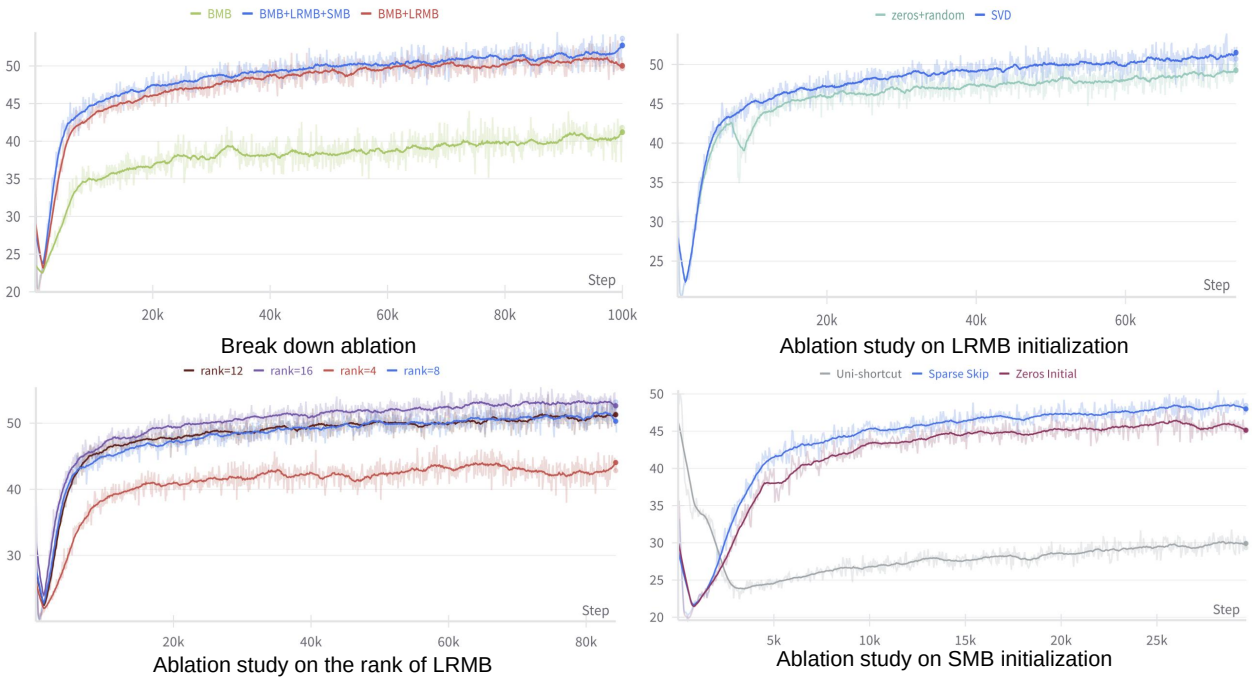

Figure 3: The details of the training curves for the ablation experiments in the main text.

Model	Num of Iters	s/Iters	Training Time
ResShift	500k	1.32s	7.64days
SinSR	30k	7.41s	2.57days
BiMaCoSR	100k	3.86s	4.47dyas

(a) Training Time Comparison.

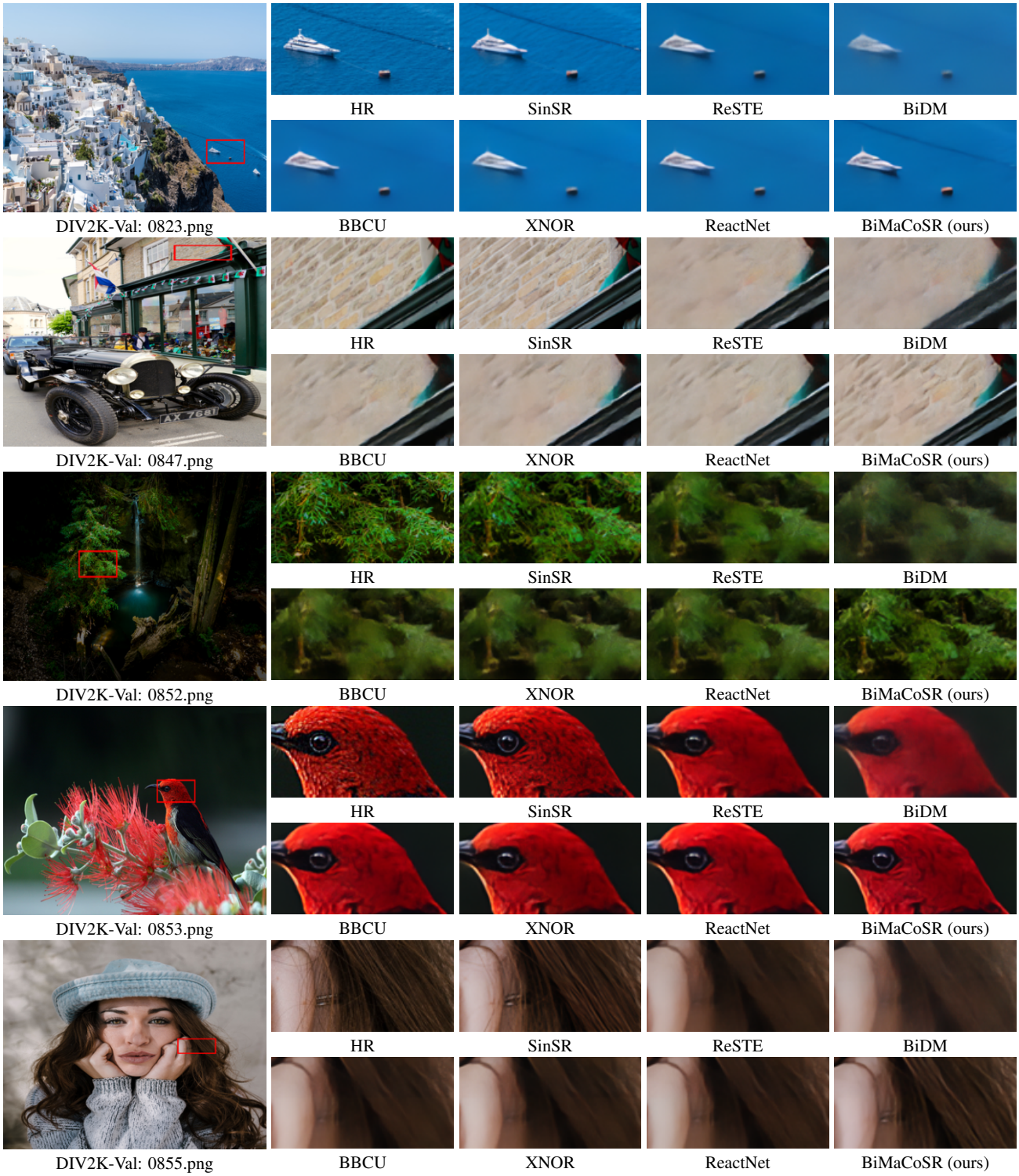

Figure 4: Visual comparison for image SR in some challenging cases.

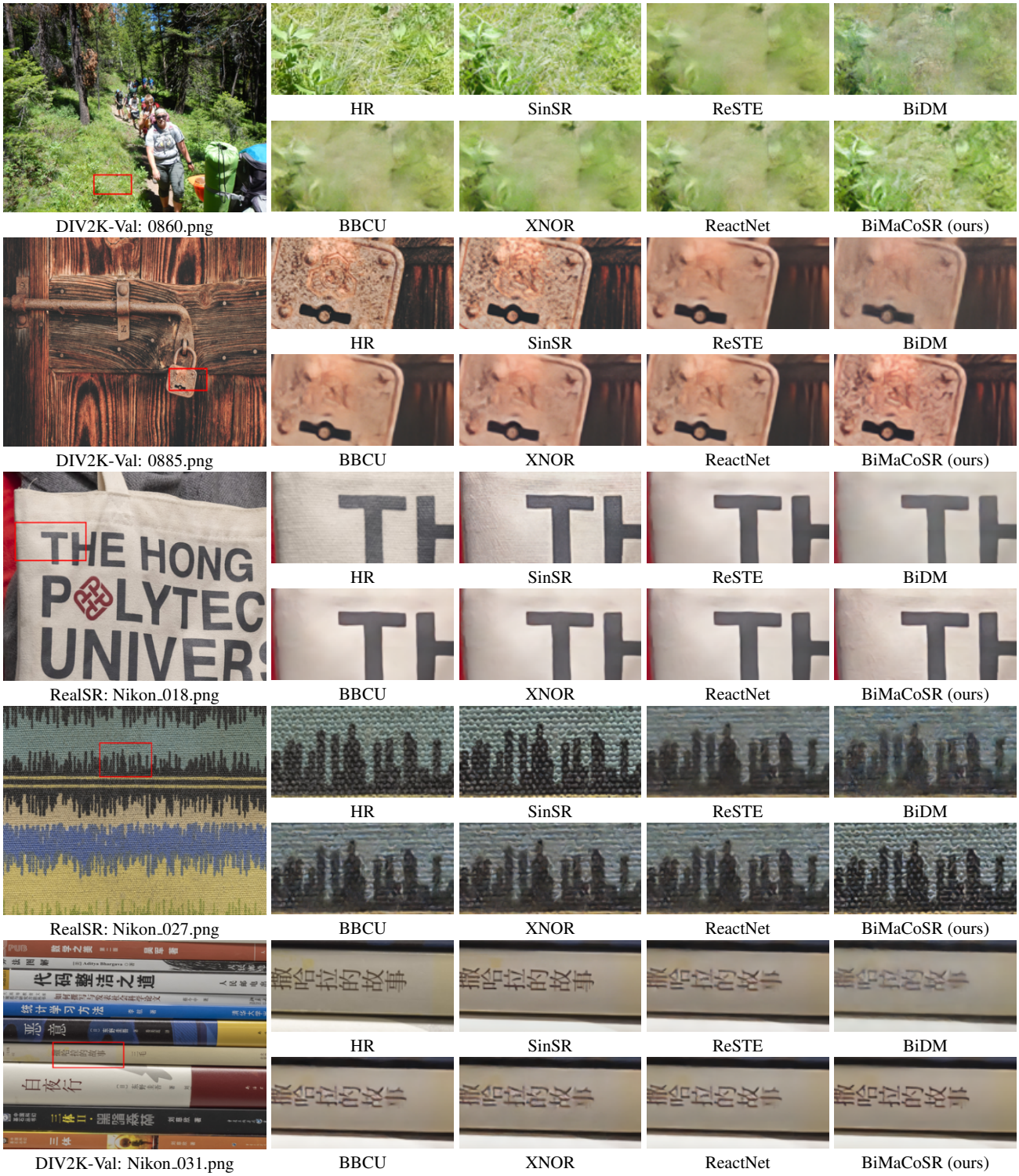

Figure 5: Visual comparison for image SR in some challenging cases.