LightRR: A Lightweight Network for Single Image Reflection Removal

Supplementary Material

6. Additional Qualitative Results

Due to space constraints in the main manuscript, only a subset of visual comparisons was provided. In this section, we present extensive qualitative comparisons against state-of-the-art (SOTA) methods on four real-world benchmarks: Real20 [33], SIR^2 Objects, PostCard, and Wild [26].

6.1. Comparison on Benchmarks

Figure 8 illustrates the visual performance of LightRR compared to competing methods. As observed, our method effectively removes reflection artifacts while preserving high-frequency transmission details, even in challenging scenarios characterized by strong reflections or complex backgrounds.

Limitations and Failure Cases. While LightRR achieves an excellent balance between efficiency and performance, it may encounter difficulties under extreme conditions. As illustrated in Figure 8, when the reflection intensity is excessively high (see the first column) or when the reflection texture is indistinguishable from the background texture (see the fourth column), our model may exhibit residual artifacts.

6.2. Visual Analysis of Ablation Studies

To visually demonstrate the contribution of each proposed component, we provide comparisons of the ablation variants discussed in the main paper (Tables 4, 5, and 6). Figure 9 presents the results of: (1) removing the AF-SSM, (2) replacing WDD/WSU with sub-pixel convolution, and (3) training without Knowledge Distillation.

- (c) **Ours (LightRR):** The proposed method achieves the best visual quality, effectively removing the white grid reflections while preserving the sharp edges of the blue dome and windows.
- (d) **Impact of AF-SSM:** Replacing our AF-SSM with a standard 2D-SSM results in incomplete reflection removal. As seen in (d), the standard SSM fails to capture the frequency-specific characteristics of the reflection, leaving visible residual streaks in the sky.
- (e) **Impact of Wavelet Sampling:** Using standard Sub-pixel Convolution instead of our Wavelet-based sampling (WDD/WSU) leads to a loss of high-frequency details. The resulting image (e) appears slightly blurred compared to Ours, validating the effectiveness of reversible wavelet transforms for detail preservation.
- (f) **Direct VGG Injection (vs. KD):** Simply injecting VGG19 features into the encoder without distillation (Variant B in the main paper) yields suboptimal results. As discussed in our paper, these generic classification features are not aligned with the reflection removal task. Directly using them restricts the lightweight model's flexibility, leading to color shifts or residual artifacts.
- (g) **w/o Pre-trained Features:** Training the lightweight network from scratch (Variant A) results in the poorest performance. Without the semantic guidance from the Teacher network (Knowledge Distillation) or pre-trained features, the shallow encoder struggles to distinguish the reflection layer.

6.3. Lightweight Network Configuration for SOTA Methods

In the main manuscript, we adjusted the architectural hyperparameters of several SOTA methods to align their parameter counts with those of our proposed LightRR, ensuring a fair comparison under lightweight constraints. We summarize the specific architectural modifications for these lightweight variants in Table 7.

Visual comparisons of these lightweight variants are presented in Figure 10. It can be observed that, despite having a comparable number of parameters, our model effectively eliminates artifacts caused by highlights and strong semantic reflections. In contrast, the performance of other competing models degrades significantly when reduced to a lightweight configuration.

7. Detailed Network Architecture

In this section, we provide the detailed specification of the proposed LightRR architecture. The network adopts a four-stage U-shaped structure.

7.1. Layer Configuration

We set the base channel width to $C = 32$. Consequently, the channel dimensions expand as $[C, 2C, 4C, 8C]$ across the four stages. The specific number of AFM Blocks (N) and the configuration of the Low-Frequency Semantic Fusion (LSF) heads are detailed in Table 8.

7.2. Detailed Hyperparameters

The specific hyperparameters for the core modules are configured as follows:

- **AF-SSM (Mamba):** The state dimension (d_{state}) is set to 16, and the expansion ratio (mlp_ratio) is set to 2.0. Bias terms in the linear layers are disabled.
- **Intermediate Processing Layers (Middlers):** The Low-Frequency (LL) features extracted from the encoder are processed by a lightweight "BasicLayer" (a stack of AFM

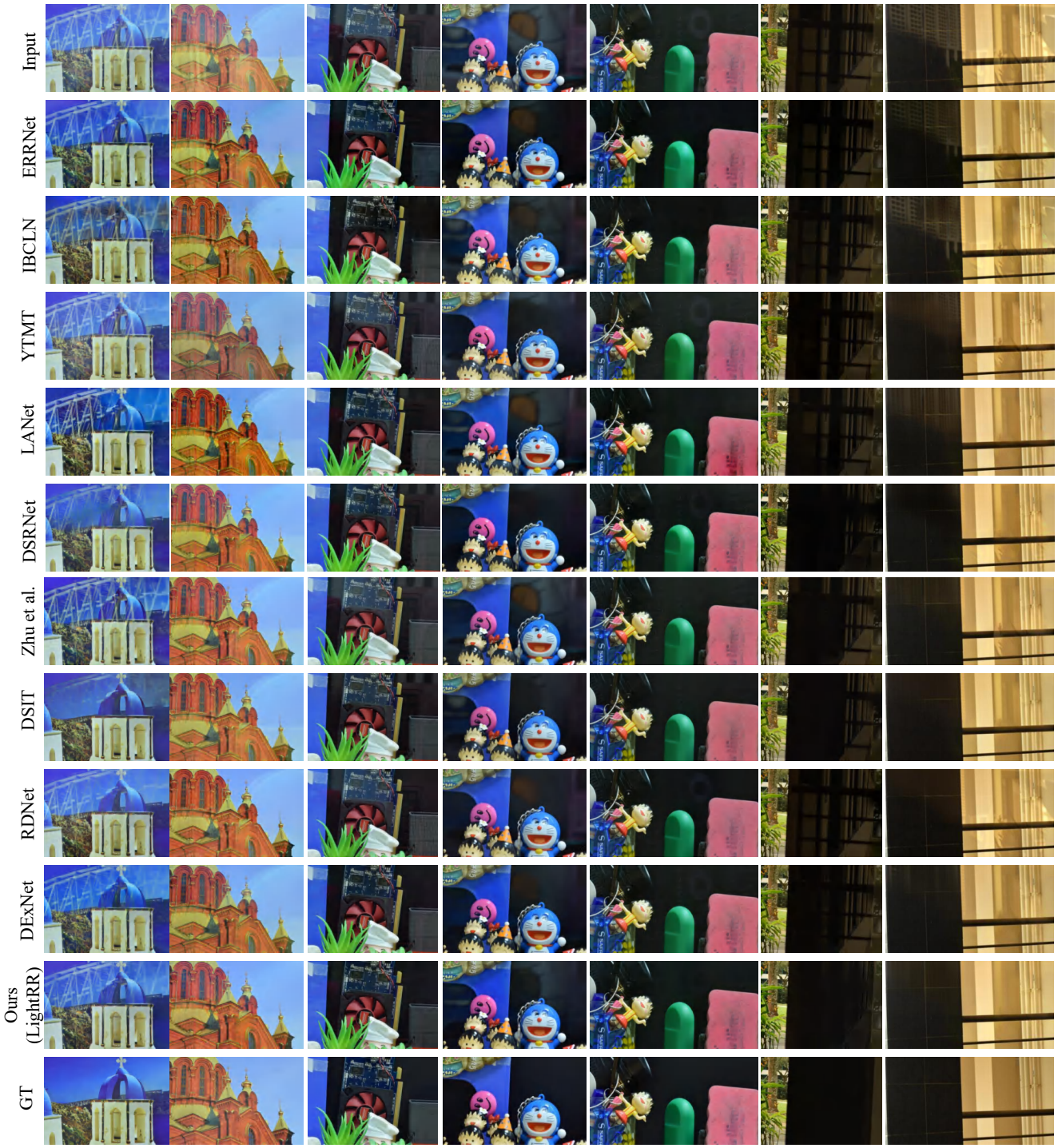

Figure 8. Visual comparisons on the **Real20** dataset. Our LightRR effectively separates reflection layers while maintaining the color fidelity of the transmission layer.

blocks) before entering the LSF attention module. The number of blocks for these intermediate layers corresponds to the encoder stages: $N = [2, 3, 3]$ for Stages 1, 2, and 3, respectively.

7.3. Knowledge Distillation Setup

To compensate for the limited capacity of the lightweight encoder, we incorporate a Knowledge Distillation (KD)

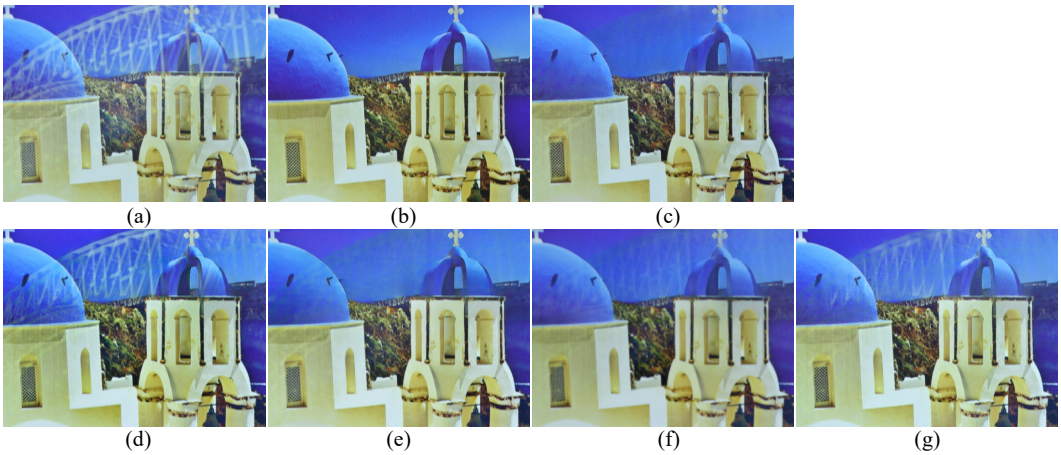

Figure 9. Visual comparison of ablation studies. (a) Input image. (b) Ground Truth. (c) **Ours (LightRR)**. (d) w/o AF-SSM (using standard 2D-SSM). (e) w/o Wavelet Sampling (using Sub-pixel Conv). (f) Direct VGG Injection (using VGG19 features directly as encoder input without distillation). (g) w/o Pre-trained Features (training from scratch without any VGG prior). Zoom in for best view.

Table 7. Detailed Architecture Configurations for SOTA Methods and their Lightweight Variants. The "Light" variants are modified to have parameter counts comparable to LightRR.

Method	Variant	Channels	Encoder Blocks	Decoder Blocks	Middle/Aux. Blocks
DSRNet	Original	[64, 128, 256, 512]	[2, 2, 4, 8]	[2, 2, 2, 2]	12
	Light	[32, 64, 128, 256]	[2, 2, 2]	[2, 2, 2]	4
DSIT	Original	[48, 96, 192, 384, 768, 1536]	[2, 2, 4, 8, 12]	[2, 2, 2, 2, 2]	[2, 2, 2, 2, 2]
	Light	[32, 64, 128, 256, 512]	[2, 2, 4, 8]	[2, 2, 2, 2]	[2, 2, 2, 2]
RDNet	Original	[64, 128, 256, 512]	[2, 2, 4, 2]	-	-
	Light	[32, 64, 128, 256]	[1, 1, 2, 1]	-	-

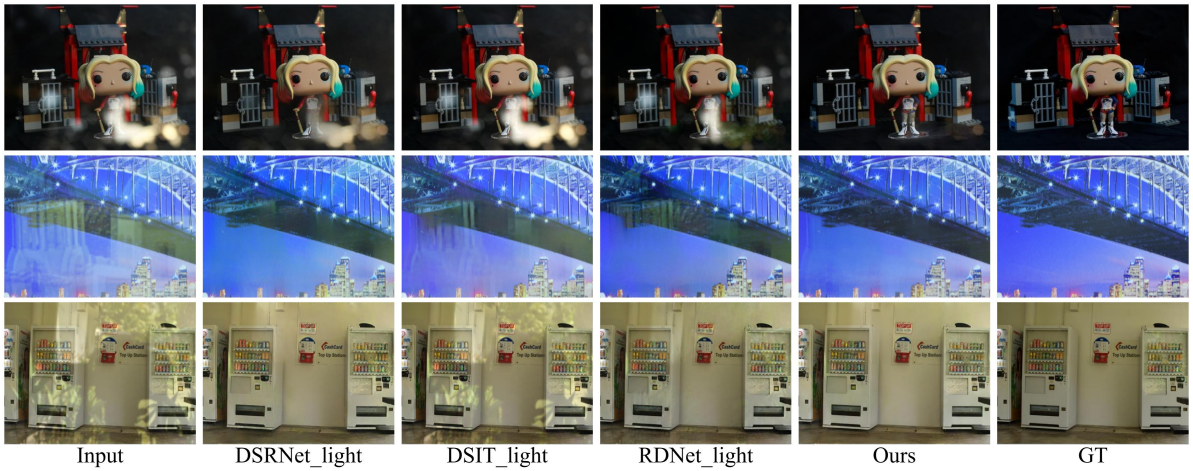

Figure 10. Visual Comparison of Lightweight Models against SOTA Methods. Under similar parameter constraints, LightRR demonstrates superior removal capabilities compared to the lightweight versions of DSRNet, DSIT, and RDNet. Zoom in for best view.

strategy during training. Specifically, we employ a pre-trained VGG19 network as the Teacher to provide multi-scale semantic guidance. The distillation loss, denoted as $\mathcal{L}_{\text{distill}}$, is minimized to align the intermediate feature representations of the Student (LightRR) with those of the Teacher. This forces the lightweight student to capture rich

Table 8. Detailed architecture configuration of LightRR. The input resolution is denoted as $H \times W$, and the base channel number is $C = 32$. Note that "Middlers" refers to the intermediate processing layers for the skipped low-frequency (LL) features before they enter the LSF module.

Stage	**Input Resolution**	**Channels**	**Block Type**	**Num. Blocks (N)**	**LSF Heads**
Encoder Path					
Tokenizer	$H \times W$	$3 \to 32$	Conv 3×3	-	-
Stage 1	$H \times W$	32	AFM Block	2	-
Downsample 1	$H \times W \to H/2 \times W/2$	$32 \to 64$	WDD	-	-
Stage 2	$H/2 \times W/2$	64	AFM Block	3	-
Downsample 2	$H/2 \times W/2 \to H/4 \times W/4$	$64 \to 128$	WDD	-	-
Stage 3	$H/4 \times W/4$	128	AFM Block	3	-
Downsample 3	$H/4 \times W/4 \to H/8 \times W/8$	$128 \to 256$	WDD	-	-
Bottleneck					
Stage 4	$H/8 \times W/8$	256	AFM Block	4	-
Decoder Path & Semantic Fusion					
Decoder 1	$H/8 \times W/8$	256	AFM Block	4	-
Fusion 1 (LSF)	$H/8 \times W/8$	128	Middle Layer	-	4 Heads
Upsample 1	$H/8 \times W/8 \to H/4 \times W/4$	128	WSU	-	-
Decoder 2	$H/4 \times W/4$	128	AFM Block	3	-
Fusion 2 (LSF)	$H/4 \times W/4$	64	Middle Layer	-	2 Heads
Upsample 2	$H/4 \times W/4 \to H/2 \times W/2$	64	WSU	-	-
Decoder 3	$H/2 \times W/2$	64	AFM Block	3	-
Fusion 3 (LSF)	$H/2 \times W/2$	32	Middle Layer	-	1 Head
Upsample 3	$H/2 \times W/2 \to H \times W$	32	WSU	-	-
Final Output	$H \times W$	$64 \to 3$	Conv 3×3	-	-

structural and semantic information inherent in the teacher's deep features.

Table 9 details the correspondence between the LightRR encoder stages and the specific feature maps extracted from the VGG19 teacher. Note that since the channel widths of LightRR are smaller than those of VGG19, we align the channel dimensions before computing the loss.

Table 9. Correspondence between LightRR Encoder Stages and VGG19 Layers for Knowledge Distillation.

Student Stage (LightRR)	Teacher Feature Map Size(VGG19)
Encoder Stage 1 Output	$64 \times H \times W$
Encoder Stage 2 Output	$128 \times H/2 \times W/2$
Encoder Stage 3 Output	$256 \times H/4 \times W/4$
Bottleneck Output	$512 \times H/8 \times W/8$

References

[1] Ya-Chu Chang, Chia-Ni Lu, Chia-Chi Cheng, and Wei-Chen Chiu. Single image reflection removal with edge guidance, reflection classifier, and recurrent decomposition. In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 2032–2041, 2021. 2

[2] Zheng Dong, Ke Xu, Yin Yang, Hujun Bao, Weiwei Xu, and Rynson W.H. Lau. Location-aware single image reflection removal. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4997–5006, 2021. 5

[3] Mark Everingham, Luc Gool, Christopher K. Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *Int. J. Comput. Vision*, 88(2): 303–338, 2010. 5

[4] Qingnan Fan, Jiaolong Yang, Gang Hua, Baoquan Chen, and David Wipf. A generic deep architecture for single image reflection removal and image smoothing. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3238–3247, 2017. 2, 5

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

[6] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. In *First conference on language modeling*, 2024. 1

[7] Albert Gu, Tri Dao, Stefano Ermon, Atri Rudra, and Christopher Ré. Hippo: Recurrent memory with optimal polynomial projections. *Advances in neural information processing systems*, 33:1474–1487, 2020. 2

[8] Albert Gu, Karan Goel, and Christopher Ré. Efficiently modeling long sequences with structured state spaces. *arXiv preprint arXiv:2111.00396*, 2021. 2

[9] Albert Gu, Isys Johnson, Karan Goel, Khaled Saab, Tri Dao, Atri Rudra, and Christopher Ré. Combining recurrent, convolutional, and continuous-time models with linear state space layers. *Advances in neural information processing systems*, 34:572–585, 2021. 1, 2

[10] 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 *ECCV*, 2024. 1, 3, 8

[11] Qiming Hu and Xiaojie Guo. Trash or treasure? an interactive dual-stream strategy for single image reflection separation. *Advances in Neural Information Processing Systems*, 34:24683–24694, 2021. 2, 5

[12] Qiming Hu and Xiaojie Guo. Single image reflection separation via component synergy. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13138–13147, 2023. 1, 2, 5

[13] Qiming Hu, Hainuo Wang, and Xiaojie Guo. Single image reflection separation via dual-stream interactive transformers. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, 2024. Curran Associates Inc. 1, 2, 5

[14] 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 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5235–5239, 2024. 2

[15] 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. 2, 5

[16] Anat Levin, Assaf Zomet, and Yair Weiss. Learning to perceive transparency from the statistics of natural scenes. *Advances in Neural Information Processing Systems*, 15, 2002. 2

[17] Chao Li, Yixiao Yang, Kun He, Stephen Lin, and John E Hopcroft. Single image reflection removal through cascaded refinement. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3565–3574, 2020. 2, 5

[18] Yu Li and Michael S. Brown. Single image layer separation using relative smoothness. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2752–2759, 2014. 2

[19] 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. 3

[20] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 5

[21] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115 (3):211–252, 2015. 5

[22] Yuan Shi, Bin Xia, Xiaoyu Jin, Xing Wang, Tianyu Zhao, Xin Xia, Xuefeng Xiao, and Wenming Yang. Vmambair: Visual state space model for image restoration. *arXiv preprint arXiv:2403.11423*, 2024. 1, 3

[23] YiChang Shih, Dilip Krishnan, Fredo Durand, and William T Freeman. Reflection removal using ghosting cues. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3193–3201, 2015. 2

[24] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014. 5

[25] Jimmy TH Smith, Andrew Warrington, and Scott W Linderman. Simplified state space layers for sequence modeling. *arXiv preprint arXiv:2208.04933*, 2022. 1

[26] Renjie Wan, Boxin Shi, Ling-Yu Duan, Ah-Hwee Tan, and Alex C Kot. Benchmarking single-image reflection removal algorithms. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3922–3930, 2017. 5, 1

[27] Renjie Wan, Boxin Shi, Ling-Yu Duan, Ah-Hwee Tan, Wen Gao, and Alex C Kot. Region-aware reflection removal with unified content and gradient priors. *IEEE Transactions on Image Processing*, 27(6):2927–2941, 2018. 2

[28] Renjie Wan, Boxin Shi, Haoliang Li, Ling-Yu Duan, Ah-Hwee Tan, and Alex C. Kot. Corrn: Cooperative reflection removal network. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(12):2969–2982, 2020. 2

[29] Kaixuan Wei, Jiaolong Yang, Ying Fu, David Wipf, and Hua Huang. Single image reflection removal exploiting misaligned training data and network enhancements. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8178–8187, 2019. 2, 5

[30] Jianwei Yang, Chunyuan Li, Xiyang Dai, and Jianfeng Gao. Focal modulation networks. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, 2022. Curran Associates Inc. 5

[31] Yang Yang, Wenye Ma, Yin Zheng, Jian-Feng Cai, and Weiyu Xu. Fast single image reflection suppression via convex optimization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8141–8149, 2019. 2

[32] 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 *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5718–5729, 2022. 4

[33] Xuaner Zhang, Ren Ng, and Qifeng Chen. Single image reflection separation with perceptual losses. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4786–4794, 2018. 2, 5, 1

[34] Hao Zhao, Mingjia Li, Qiming Hu, and Xiaojie Guo. Reversible Decoupling Network for Single Image Reflection Removal . In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26430–26439, Los Alamitos, CA, USA, 2025. IEEE Computer Society. 1, 2, 5

[35] Lianghui Zhu, Bencheng Liao, Qian Zhang, Xinlong Wang, Wenyu Liu, and Xinggang Wang. Vision mamba: Efficient visual representation learning with bidirectional state space model. In *Forty-first International Conference on Machine Learning*, 2024. 3

[36] Yurui Zhu, Xueyang Fu, Peng-Tao Jiang, Hao Zhang, Qibin Sun, Jinwei Chen, Zheng-Jun Zha, and Bo Li. Revisiting singlelmage reflection removal in the wild. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2, 5

[37] Wenbin Zou, Hongxia Gao, Weipeng Yang, and Tongtong Liu. Wave-mamba: Wavelet state space model for ultra-high-definition low-light image enhancement. In *ACM Multimedia 2024*, 2024. 1, 3