Depth-Synergized Mamba Meets Memory Experts for All-Day Image Reflection Separation

Siyan Fang[1], Long Peng[2], Yuntao Wang[1], Ruonan Wei[1], Yuehuan Wang[1]*

[1]Huazhong University of Science and Technology
[2]University of Science and Technology of China
siyanfang@hust.edu.cn, longp2001@mail.ustc.edu.cn, yuntaowang@hust.edu.cn, ruonan2765@gmail.com, yuehwang@hust.edu.cn

Abstract

Image reflection separation aims to disentangle the transmission layer and the reflection layer from a blended image. Existing methods rely on limited information from a single image, tending to confuse the two layers when their contrasts are similar, a challenge more severe at night. To address this issue, we propose the Depth-Memory Decoupling Network (DMDNet). It employs the Depth-Aware Scanning (DAScan) to guide Mamba toward salient structures, promoting information flow along semantic coherence to construct stable states. Working in synergy with DAScan, the Depth-Synergized State-Space Model (DS-SSM) modulates the sensitivity of state activations by depth, suppressing the spread of ambiguous features that interfere with layer disentanglement. Furthermore, we introduce the Memory Expert Compensation Module (MECM), leveraging cross-image historical knowledge to guide experts in providing layer-specific compensation. To address the lack of datasets for nighttime reflection separation, we construct the Nighttime Image Reflection Separation (NightIRS) dataset. Extensive experiments demonstrate that DMDNet outperforms state-of-the-art methods in both daytime and nighttime.

Project Page — https://github.com/fashyon/DMDNet

Introduction

Reflection artifacts often occur when capturing images through transparent media such as glass, not only compromising visual quality but also degrading the performance of downstream vision tasks (Kirillov et al. 2023; Wang et al. 2024). The task of image reflection separation aims to decompose a blended image $\boldsymbol{I}$ into a transmission layer $\boldsymbol{T}$ and a reflection layer $\boldsymbol{R}$, where $\boldsymbol{T}$ represents the scene behind the glass and $\boldsymbol{R}$ represents the reflected content on the glass surface. Early studies mainly rely on physical priors such as gradient sparsity (Levin and Weiss 2007) and reflection blurriness (Fan et al. 2017; Yang et al. 2019), using handcrafted constraints based on physical assumptions. However, these methods are only effective in constrained scenarios. With the development of deep learning, methods such as Zhang et al. (Zhang, Ng, and Chen 2018) and DSIT (Hu, Wang, and Guo 2024) learn implicit priors of $\boldsymbol{T}$ and $\boldsymbol{R}$ from data to achieve separation. However, due to the limited information in a single image, these methods often encounter bottlenecks when $\boldsymbol{T}$ and $\boldsymbol{R}$ exhibit similar contrast. The challenge becomes particularly severe in nighttime scenes. In the daytime, abundant natural illumination strengthens $\boldsymbol{T}$ while suppressing $\boldsymbol{R}$, resulting in a clear contrast between the two layers. At night, illumination comes from artificial light sources that are randomly distributed, leading to uneven lighting conditions. Consequently, $\boldsymbol{T}$ appears darker due to insufficient global illumination, while localized strong lights incident on the glass surface produce glare and scattered highlights. As a result, $\boldsymbol{T}$ and $\boldsymbol{R}$ exhibit similar contrast levels, making their separation more challenging.

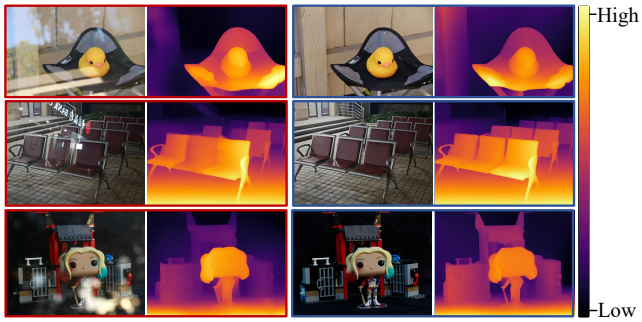


Figure 1: Proximity maps obtained by depth estimation across daytime, nighttime, and indoor scenes. Depth estimation sees through reflection occlusion to capture the underlying structures of $\boldsymbol{T}$.

Although these difficulties are not directly addressed, some studies attempt to compensate by introducing additional physical cues, such as multi-view images (Xue et al. 2015; Niklaus et al. 2021), polarizing filters (Li et al. 2020b; Lei et al. 2020), infrared cameras (Sun et al. 2019; Hong et al. 2022), and flash illumination (Lei and Chen 2021; Wang et al. 2025). However, such methods require controlled environments and extra devices, limiting their flexibility in applications. To eliminate reliance on external hardware, some studies incorporate human interaction, such as language prompts (Zhong et al. 2024; Hong et al. 2024) and

*Corresponding author.

manual region annotation (Zhang et al. 2019; Chen et al. 2025). Nevertheless, these approaches are time-consuming and labor-intensive.

Depth estimation offers physical cues without additional hardware or manual intervention. By performing depth estimation (Ranftl et al. 2022) on blended images, we observe that the resulting proximity map highlights coherent and sharp structures corresponding to $\boldsymbol{T}$, while blurry and transparent overlays associated with $\boldsymbol{R}$ are naturally suppressed, as shown in Figure 1. This indicates that high proximity values tend to carry salient structures. These structures often span large spatial ranges, such as the outline of a building or a row of chairs, fully exploiting these cues requires a model capable of capturing long-range dependencies.

Mamba (Gu and Dao 2023) has achieved impressive results in various fields (Zhu et al. 2024a; Peng et al. 2025), enabled by the efficient long-range modeling of its State-Space Model (SSM). VMamba (Liu et al. 2024) brings this capability to the vision domain through four-directional scanning. However, this scanning strategy has two limitations for image reflection separation:

(1) Disruption of Structural Continuity. The transmission scene is typically defined by coherent contours, shapes, and textures, such as the edges of windows or the curves of human faces. The fixed sequential scanning fragments this content, leading to distorted structural cues while hindering the perception of these semantic entities as a whole.

(2) Error Propagation. In SSM, the state of earlier-scanned regions continuously influences subsequent ones. If ambiguous features are propagated first, their uncertainty spreads throughout the entire image, amplifying separation errors.

To address these issues, we propose the Depth-Synergized Decoupling Mamba (DSMamba). Its Depth-Aware Scanning Strategy (DA-Scan) customizes scanning strategies separately for $\boldsymbol{T}$ and $\boldsymbol{R}$, allowing the model to encounter salient structures at early stages of modeling, helping to establish semantic continuity. In synergy with DA-Scan, we design the Depth-Synergized State-Space Model (DS-SSM) to modulate the activity of state evolution while suppressing activations in ambiguous areas, preventing the spread of erroneous information.

To overcome the limited information of a single image, we introduce the Memory Expert Compensation Module (MECM) to leverage cross-image historical knowledge. Each expert is equipped with a memory bank that stores feature patterns, and MECM dynamically activates the most relevant experts to provide targeted compensation. For example, experts specialized in texture details and structural contours can be activated for $\boldsymbol{T}$, while those handling sparse highlights and blurred ghosting can be used for $\boldsymbol{R}$.

To address the scarcity of datasets for nighttime image reflection separation, we construct the Nighttime Image Reflection Separation (NightIRS) dataset. It comprises 1,000 image triplets obtained under nighttime reflection conditions. This dataset captures the unique complexities of nighttime imaging, including uneven illumination, strong artificial light sources, and diverse reflection artifacts, which are often overlooked in existing public datasets.

Overall, the contributions of this work are as follows:

- We propose DSMamba, with DA-Scan and DS-SSM working in synergy to guide Mamba toward structural saliency and suppress erroneous propagation.
- We introduce MECM to leverage cross-image historical memory for targeted compensation.
- We construct the NightIRS dataset for evaluating nighttime reflection separation.
- Experimental results demonstrate that DMDNet outperforms State-of-the-Art Methods (SOTAs).

Related Work

Image Reflection Separation. Early studies (Levin and Weiss 2007; Yang et al. 2019) rely on handcrafted priors, which only work in simple cases. Deep learning methods (Zhang, Ng, and Chen 2018; Hu, Wang, and Guo 2024) learn mappings from contaminated to clean images using large-scale data, but often struggle with complex scenes due to limited information in a single image. To incorporate physical cues, some approaches leverage multi-view images, polarization (Lei et al. 2020), flash (Lei and Chen 2021), or infrared cameras (Hong et al. 2022), but these require extra hardware, making them unsuitable for internet images. To avoid this, Zhong et al. (Zhong et al. 2024) introduce language prompts, while FIRM (Chen et al. 2025) relies on manual region annotations. However, these methods need human intervention and thus limit automation. In contrast, depth estimation offers physical cues without external sensors. Elnenaey et al. (Elnenaey and Torki 2024) coarsely quantize the depth map into four levels and concatenate it with the input image for guidance. DGR^2-Net (He et al. 2025) applies global pooling on the depth map and then concatenates it with the input for binocular reflection removal. However, these methods lack fine-grained depth guidance, resulting in inadequate effectiveness. More importantly, they overlook the structural saliency embedded in depth maps for image reflection separation.

Visual Mamba. Due to Mamba's strong performance in long-sequence modeling, it has recently been widely adopted in various vision tasks (Zhu et al. 2024a; Liu et al. 2024). MambaIR (Guo et al. 2024) and VMambaIR (Shi et al. 2025) are among the earliest works to introduce the Mamba into the field of image restoration. Subsequently, MambaIRv2 (Guo et al. 2025) proposes a semantics-guided neighborhood interaction mechanism to facilitate information transfer. TAMambaIR (Peng et al. 2025) introduces a multi-directional receptive field expansion scheme to enhance modeling capability. However, these methods lack dynamic state modeling strategies sensitive to geometric structures, limiting their ability to distinguish between layers in reflection separation.

Mixture of Experts (MoE). MoE enables adaptive computation by employing multiple experts, and has been widely applied to image restoration tasks. MoCE-IR (Zamfir et al. 2025) designs expert modules with varying computational complexity to match different degradation. FAME (He et al. 2024) adopts a frequency-adaptive MoE architecture, applying different dynamic processing strategies to low- and high-frequency components. However, these methods lack cross-

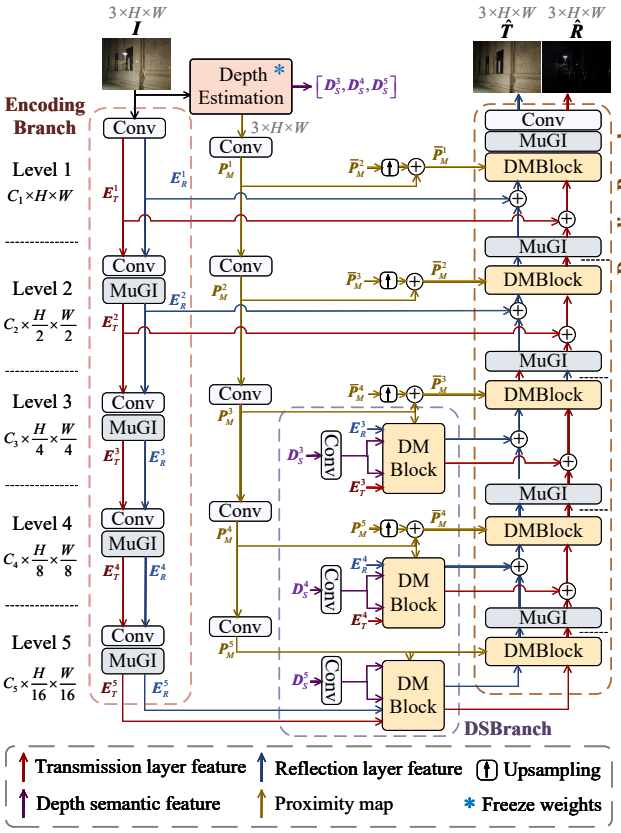


Figure 2: Depth-Memory Decoupling Network (DMDNet). DMDNet employs the DMBlock to decouple $\boldsymbol{T}$ and $\boldsymbol{R}$ using depth and memory cues.

image memory, limiting their ability to compensate for contaminated information within a single image.

Memory-Augmented Methods. Several studies explore memory mechanisms for image restoration. For instance, Xu et al. (Xu et al. 2021) propose a texture memory that stores patch samples to guide texture synthesis. ER^{2}Net (Zou et al. 2024) leverages a memory module to inpaint eyeglass reflection regions. However, the high computational cost restricts it to one-off usage, making it unsuitable for the deployment of multiple experts. Moreover, they are limited to either global matching or local modeling, without a unified mechanism to enable adaptive expert behavior.

Methodology

Depth-Memory Decoupling Network

The Depth-Memory Decoupling Network (DMDNet) consists of the Encoding Branch, the Depth Semantic Modulation Branch (DSBranch), and the Decoding Branch, as shown in Figure 2. The Encoding Branch adopts the Mutually-Gated Interactive Block (MuGI) (Hu and Guo 2023) to extract the features of $\boldsymbol{T}$ and $\boldsymbol{R}$, where $\boldsymbol{E}_T^i,\ \boldsymbol{E}_R^i \in \mathbb{R}^{C_i \times \frac{H}{2^{i-1}} \times \frac{W}{2^{i-1}}},\quad i \in \{1,2,3,4,5\}$. Here, C_i denotes the number of channels at the i-th level, and H and W are the height and width of the input image $\boldsymbol{I}$, respectively. The DSBranch leverages depth semantic features $\boldsymbol{D}_S^3 \in \mathbb{R}^{96 \times \frac{H}{4} \times \frac{W}{4}},\ \boldsymbol{D}_S^4 \in \mathbb{R}^{256 \times \frac{H}{8} \times \frac{W}{8}},\ \boldsymbol{D}_S^5 \in \mathbb{R}^{512 \times \frac{H}{16} \times \frac{W}{16}}$, modulating the encoded features for the Decoding Branch. The Decoding Branch performs the separation of $\boldsymbol{T}$ and $\boldsymbol{R}$ through the Depth-Memory Decoupling Block (DMBlock) and the proximity maps $\boldsymbol{P}_M^i \in \mathbb{R}^{C \times \frac{H}{2^{i-1}} \times \frac{W}{2^{i-1}}},\ i \in \{1,2,3,4,5\}$. As shown in Figure 3(a), the DMBlock consists of DSMamba, MECM, and EFFN (Shi et al. 2025).

Depth-Synergized Decoupling Mamba

To address the limitation of Mamba's fixed scanning strategy, we propose Depth-Synergized Decoupling Mamba (DSMamba). As illustrated in Figure 3(b), DSMamba consists of the Depth-Aware Scanning (DAScan) and the Depth-Synergized State-Space Model (DS-SSM). The DAScan adopts Depth-Aware Regional Scanning (DA-RScan) for $\boldsymbol{T}$, and Depth-Aware Global Scanning (DA-GScan) for $\boldsymbol{R}$.

DA-RScan follows a "large-area-first + near-to-far" scheme. Specifically, the proximity map is partitioned into a region scanning map $\boldsymbol{M}_{reg}$. Regions are scanned from the largest to the smallest, as larger regions indicate more salient semantics, with the background region scanned at the end to ensure completeness. This region-based scheme preserves the semantic continuity of pixels within the same object. Inside each region, pixels are scanned in a near-to-far order, prioritizing structurally salient structures.

DA-GScan follows a "global near-to-far" scheme, scanning from the globally nearest pixels to the farthest. This scheme emphasizes global structural saliency, which matches the sparse and discontinuous distribution of $\boldsymbol{R}$ to enhance the modeling of reflection features. Finally, inverse DAScan is applied in the opposite order to complement structural cues.

The vanilla State Space Model (SSM) in Mamba adopts a uniform state update mechanism for all regions, formulated as:

$$\boldsymbol{h}_t = \boldsymbol{A}\boldsymbol{h}_{t-1} + \boldsymbol{B}\boldsymbol{x}_t, \quad \boldsymbol{y}_t = \boldsymbol{C}\boldsymbol{h}_t + \boldsymbol{D}\boldsymbol{x}_t \tag{1}$$

where $\boldsymbol{A} \in \mathbb{R}^{N \times N}$, $\boldsymbol{B} \in \mathbb{R}^{N \times 1}$, $\boldsymbol{C} \in \mathbb{R}^{1 \times N}$, and $\boldsymbol{D} \in \mathbb{R}$. N is the state size. This mechanism lacks structural awareness, making it difficult to disentangle regions where $\boldsymbol{T}$ and $\boldsymbol{R}$ are intricately intertwined. To overcome this constraint while synergizing with DAScan, we design the DS-SSM, whose state update is defined as:

$$\begin{aligned} \boldsymbol{h}_t &= \boldsymbol{A}\boldsymbol{h}_{t-1} + \boldsymbol{B}_{aware}\boldsymbol{x}_t, \\ \boldsymbol{y}_t &= \boldsymbol{C}_{aware}\boldsymbol{h}_t + \boldsymbol{D}\boldsymbol{x}_t, \\ \boldsymbol{B}_{aware} &= (1-\boldsymbol{\gamma}) \cdot \boldsymbol{B} + \boldsymbol{\gamma} \cdot \boldsymbol{B}_{depth}, \\ \boldsymbol{C}_{aware} &= (1-\boldsymbol{\gamma}) \cdot \boldsymbol{C} + \boldsymbol{\gamma} \cdot \boldsymbol{C}_{depth} \end{aligned} \tag{2}$$

Here, $\boldsymbol{\gamma}$ is a weighting map between 0–1, derived from the proximity map. $\boldsymbol{B}_{depth}$ and $\boldsymbol{C}_{depth}$ are depth-guided state matrices that respectively control the magnitude of state updates and the contribution of the state to the output.

In structurally salient regions, a larger $\boldsymbol{\gamma}$ strengthens the influence of $\boldsymbol{B}_{depth}$ and $\boldsymbol{C}_{depth}$, accelerates the integration of clear structures, and reinforces their guidance on the output. Conversely, in structurally ambiguous regions, the intervention is suppressed to prevent the propagation of ambiguous features.

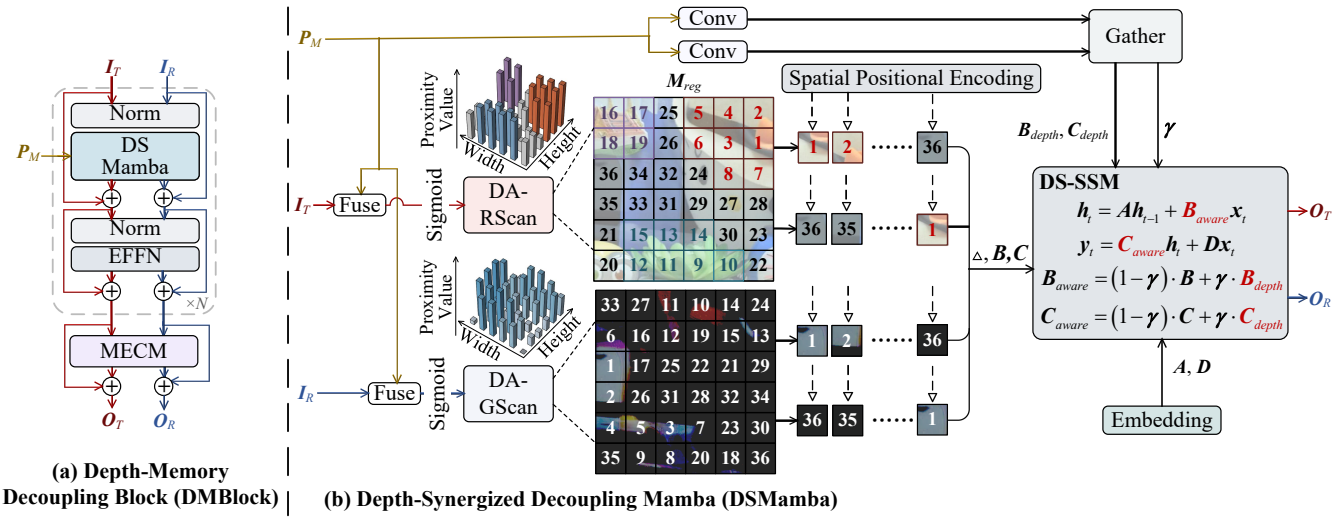


Figure 3: DMBlock and DSMamba. DSMamba prioritizes salient structures via DAScan and synergistically modulates state activations through DS-SSM. The numbers indicate the forward scanning order.

Spatial Positional Encoding. To reinforce positional specificity during the scanning, DSMamba employs a Spatial Positional Encoding (SPE) based on 2D sine and cosine functions:

$$\begin{aligned} \boldsymbol{PE}_x &= [\sin(x \cdot f_i),\ \cos(x \cdot f_i)], \\ \boldsymbol{PE}_y &= [\sin(y \cdot f_i),\ \cos(y \cdot f_i)] \end{aligned} \tag{3}$$

where x and y denote the normalized spatial coordinates, and f_i represents different frequency bands.

By combining the horizontal and vertical encodings, a positional embedding $\boldsymbol{PE} \in \mathbb{R}^{H \times W \times d_{inner}}$ is obtained, where d_{inner} is the channel dimension of the state-space model. The embedding is realigned with the scanning order and added to the state features, providing positional cues for state modeling.

Memory Expert Compensation Module

To leverage cross-image accumulated knowledge for targeted compensation, we introduce the Memory Expert Compensation Module (MECM), as illustrated in Figure 4. MECM consists of the Expert Gate (He et al. 2024) and Memory Experts. The Expert Gate is responsible for selecting the most relevant N_{Exp}^K experts from N_{Exp} candidates. The Memory Experts perform feature retrieval and evolution, enabling adaptive compensation with historical knowledge.

The Memory Expert comprises the Global-Pattern Interaction Stream (GPStream) and the Spatial-Context Refinement Stream (SCStream). The GPStream is further divided into Global-Pattern Adjustment and Memory Evolution.

For the Global-Pattern Adjustment, the input image $\boldsymbol{I} \in \mathbb{R}^{B \times C \times H \times W}$ is first pooled into a global representation $\boldsymbol{I}_G \in \mathbb{R}^{B \times C}$, which is used to compute similarity with the memory bank $\boldsymbol{Mem} \in \mathbb{R}^{M \times C}$, yielding a similarity score matrix $\boldsymbol{S} \in \mathbb{R}^{B \times M}$, where B is the batch size and M is the number of memory items. We apply softmax along the memory and image dimensions to obtain $\boldsymbol{S}_I$ and $\boldsymbol{S}_M$, respectively. Here, $\boldsymbol{S}_I$ denotes matching distribution of each image over all memory items, while $\boldsymbol{S}_M$ represents the contribution of each memory item to the image. Next, $\boldsymbol{S}_I$ is used to perform weighted aggregation of $\boldsymbol{Mem}$, producing the memory response feature $\boldsymbol{F}_M \in \mathbb{R}^{B \times C}$. Finally, $\boldsymbol{F}_M$ interacts with $\boldsymbol{I}_G$ to generate an attention mask that modulates the input $\boldsymbol{I}$, producing the global compensation $\boldsymbol{O}_G$.

Memory Evolution aims to provide feedback and update the memory bank. For each image sample $b \in [1, B]$, the most responsive memory index $j_b \in [1, M]$ is selected from the matching matrix $\boldsymbol{S}_I$. The corresponding score $\boldsymbol{S}_M[b, j_b]$ is used as a weight to perform multiplication with the global representation $\boldsymbol{I}_G[b]$, resulting in an update vector $\boldsymbol{U}_b \in \mathbb{R}^C$. All $\boldsymbol{U}_b$ vectors are aggregated along their associated index j_b to form a memory increment $\Delta\boldsymbol{Mem} \in \mathbb{R}^{M \times C}$:

$$\Delta\boldsymbol{Mem}[m] = \sum_{b \in [1,B],\ j_b = m} \boldsymbol{U}_b, \quad m \in [1, M] \tag{4}$$

Finally, the memory bank is updated in a residual manner to obtain the updated memory $\boldsymbol{O}_{Mem}$.

SCStream focuses on spatial contextual compensation. First, the memory bank $\boldsymbol{Mem}$ is reshaped as convolutional kernels and convolved with the input image $\boldsymbol{I}$ to obtain the similarity map $\boldsymbol{S}_S \in \mathbb{R}^{B \times M \times H \times W}$. $\boldsymbol{S}_S[b, m, h, w]$ denotes the similarity between location (h, w) and the m-th memory item. Next, for each spatial position, the Top-k most relevant memory items are selected. Specifically, $\boldsymbol{Idx}_K$, $\boldsymbol{S}_K \in \mathbb{R}^{B \times K \times HW}$ (where $HW = H \times W$) denote the indices and similarity scores of the Top-k memory items for each pixel. The similarity scores $\boldsymbol{S}_K$ are normalized using softmax to obtain the attention weights $\boldsymbol{W}_A \in \mathbb{R}^{B \times K \times HW}$, representing the degree of matching between each pixel and the Top-k memory items. Then, the corresponding memory features are retrieved from the memory bank using $\boldsymbol{Idx}_K$. The retrieved memory tensor is denoted

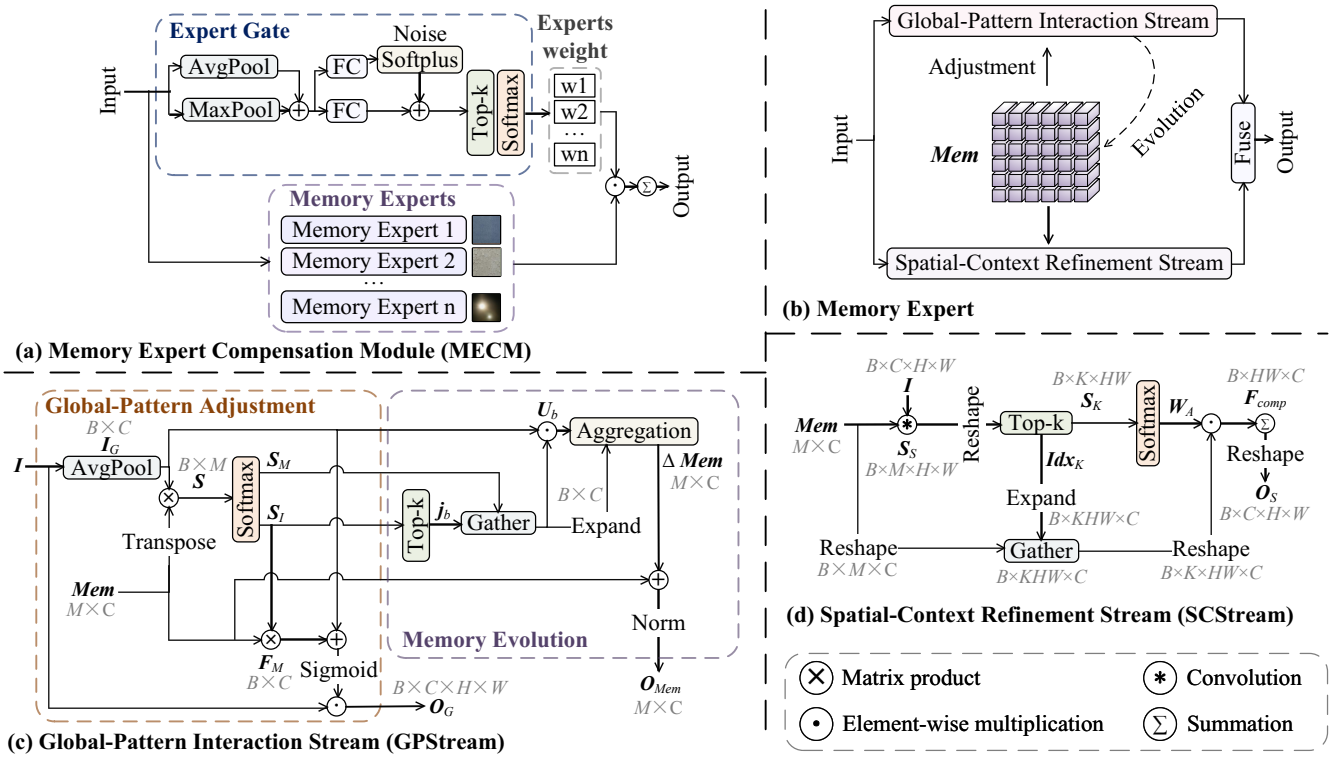


Figure 4: Memory Expert Compensation Module (MECM) and its components, which leverage cross-image historical knowledge to guide the decoupling. Each memory expert consists of the GPStream for global adjustment and memory evolution, and the SCStream for spatial-level refinement.

as $\boldsymbol{Mem}_K \in \mathbb{R}^{B\times K\times HW\times D}$, which contains the features of the Top-k memory items associated with each pixel position. The weighted sum yields the compensation feature $\boldsymbol{F}_{comp} \in \mathbb{R}^{B\times HW\times D}$, and the final output $\boldsymbol{O}_S$ is obtained by reshaping. The weighted sum is computed as:

$$\boldsymbol{F}_{comp}[b, hw, d] = \sum_{k=1}^{K} \boldsymbol{W}_A[b, k, hw] \cdot \boldsymbol{Mem}_K[b, k, hw, d] \tag{5}$$

Each expert employs distinct convolutions to fuse the features from GPStream and SCStream, capturing specific semantic relations and enabling adaptive refinement.

Nighttime Image Reflection Separation Dataset

The Nighttime Image Reflection Separation (NightIRS) dataset contains 1,000 nighttime reflection image triplets. Each triplet consists of $\boldsymbol{I}$, $\boldsymbol{T}$, and $\boldsymbol{R}$, as shown in Figure 5. Reflection interference is introduced using glass and acrylic sheets of varying thicknesses. To ensure illumination diversity, the dataset is collected under various nighttime conditions, such as street lights, neon signs, illuminated buildings, and low-light natural environments. To capture geometric variations of reflections, different camera-to-glass distances and viewing angles are considered. The dataset also provides a high-resolution version (NightIRS-HR), offering scalable benchmarks for nighttime reflection separation.

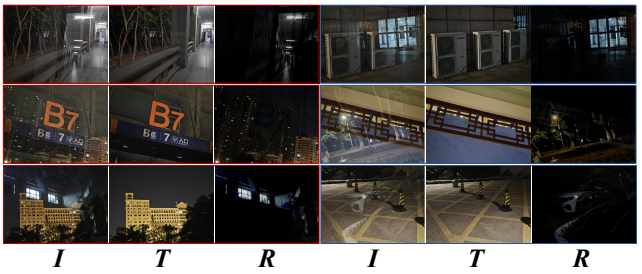


Figure 5: Examples from the NightIRS dataset. $\boldsymbol{I}$, $\boldsymbol{T}$, and $\boldsymbol{R}$ denote the blended image, transmission layer, and reflection layer, respectively.

Experiments

Implementation Details

The channel dimensions are set as $C_1, C_2, C_3, C_4, C_5 = [48, 96, 192, 384, 768]$. In MECM, $N_{Exp} = 4$ and $N^K_{Exp} = 2$. We adopt a batch size of 1 and crop images into 352×352 patches. Random horizontal flipping is adopted for data augmentation during training. The model is optimized using the Adam optimizer (Kingma and Ba 2014) with an initial learning rate of 10^{-4}. We train for 60 epochs, and reduce the learning rate to 5×10^{-5} and 10^{-5} at the 30th and 50th epochs, respectively. All experiments are conducted on a single NVIDIA RTX 4090 GPU. See supplementary mate-

Methods	Nature (20)			Real (20)			Wild (55)			Postcard (199)			Solid (200)			Average		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
BDN (ECCV'18)	18.83	0.737	0.242	18.68	0.728	0.284	22.02	0.822	0.181	20.54	0.857	0.177	22.68	0.856	0.125	20.55	0.800	0.202
ERRNet (CVPR'19)	20.43	0.756	0.172	23.03	0.810	0.156	23.87	0.848	0.132	21.81	0.874	0.152	24.72	0.896	0.095	22.77	0.837	0.141
IBCLN (CVPR'20)	23.78	0.784	0.145	21.59	0.764	0.210	24.46	0.885	0.134	22.95	0.875	0.155	24.74	0.893	0.097	23.50	0.840	0.148
LANet (ICCV'21)	23.55	0.811	0.115	22.51	0.815	0.145	26.06	0.900	0.109	24.14	0.907	0.106	24.30	0.898	0.087	24.11	0.866	0.112
YTMT (NIPS'21)	20.77	0.769	0.178	22.86	0.807	0.158	25.07	0.892	0.116	22.40	0.881	0.147	24.70	0.899	0.092	23.16	0.850	0.138
DMGN (TIP'21)	20.63	0.764	0.167	20.28	0.763	0.215	21.34	0.774	0.152	22.65	0.879	0.151	23.27	0.872	0.102	21.63	0.810	0.157
HGNet (TNNLS'23)	25.23	0.824	0.111	23.65	0.818	0.155	26.88	0.897	0.109	23.56	0.900	0.124	25.00	0.900	0.092	24.86	0.868	0.118
DSRNet (ICCV'23)	21.62	0.781	0.149	23.41	0.805	0.147	24.35	0.893	0.117	24.66	0.911	0.111	26.10	0.914	0.071	24.03	0.861	0.119
RDRNet (CVPR'24)	24.44	0.820	0.107	21.29	0.769	0.190	26.48	0.905	0.101	23.65	0.891	0.146	25.93	0.912	0.080	24.36	0.860	0.125
DSIT (NIPS'24)	26.05	0.830	0.128	24.34	0.823	0.136	27.55	**0.920**	**0.081**	**26.01**	**0.921**	0.103	26.62	0.922	0.075	26.11	0.883	0.105
RDNet (CVPR'25)	25.77	0.828	0.108	**25.13**	**0.838**	**0.117**	27.59	0.915	0.085	25.95	**0.921**	**0.088**	26.59	0.922	0.069	26.21	0.885	0.094
DMDNet (Ours)	**26.68**	**0.838**	**0.097**	24.60	0.836	0.130	**27.70**	**0.920**	0.083	25.32	**0.921**	0.093	**27.07**	**0.929**	**0.064**	**26.27**	**0.889**	**0.093**

Table 1: Quantitative comparison of the transmission layer on public datasets. DMDNet achieves the best average performance. **Bold** and underline denote Top-1 and Top-2 results, respectively. ↑ indicates higher is better, while ↓ indicates lower is better.

Methods	Transmission Layer			Reflection Layer			Param	FLOPs
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	(M)↓	(G)↓
BDN (ECCV'18)	20.52	0.680	0.293	8.79	0.082	0.843	75.16	12.70
ERRNet (CVPR'19)	22.43	0.767	0.180	N/A	N/A	N/A	18.95	116.72
IBCLN (CVPR'20)	23.16	0.803	0.196	20.54	0.292	0.701	21.61	98.16
LANet (ICCV'21)	23.68	0.817	0.171	21.61	0.280	0.472	10.93	83.81
YTMT (NIPS'21)	23.03	0.799	0.186	24.96	0.500	0.503	76.90	110.98
DMGN (TIP'21)	22.88	0.799	0.174	24.77	0.488	0.508	45.49	116.85
HGNet (TNNLS'23)	23.60	0.817	0.170	N/A	N/A	N/A	14.51	82.08
DSRNet (ICCV'23)	23.39	0.813	0.175	24.80	0.404	0.499	124.6	90.21
RDRNet (CVPR'24)	24.04	0.824	0.185	N/A	N/A	N/A	29.09	5.14
DSIT (NIPS'24)	24.61	0.827	0.168	27.18	0.569	0.372	131.76	74.18
RDNet (CVPR'25)	25.08	0.831	0.149	27.93	**0.636**	0.309	266.43	66.10
DMDNet (Ours)	**25.24**	**0.832**	**0.144**	**28.37**	0.633	**0.286**	87.22	39.33

Table 2: Quantitative comparison with SOTAs on the NightIRS dataset. FLOPs for a 128×128 RGB image.

rial for more details.

Dataset and Evaluation Metrics

Following previous works (Zhao et al. 2025; Hu, Wang, and Guo 2024; Hu and Guo 2023; Dong et al. 2021), we train our model on 7,643 image pairs from the PASCAL VOC dataset (Everingham et al. 2010), 200 image pairs from the Nature dataset (Li et al. 2020a), and 89 image pairs from the Real dataset (Zhang, Ng, and Chen 2018). The remaining images from the Nature and Real datasets, together with the Wild, Postcard, and Solid subsets from the SIR^2 dataset (Wan et al. 2017), as well as the NightIRS dataset, are used for testing. To avoid GPU memory overflow, images from the Real dataset are resized by scaling the longer side to 420 pixels while preserving the original aspect ratio.

To ensure fairness, all output images are saved in lossless PNG format, and evaluation metrics are computed in the RGB color space, including PSNR (Huynh-Thu and Ghanbari 2008), SSIM (Wang et al. 2004), and LPIPS (Zhang et al. 2018), which assess image quality from pixel-wise, structural, and perceptual perspectives, respectively.

Performance Evaluation

We compare our DMDNet with 11 methods, including BDN (Yang et al. 2018), ERRNet (Wei et al. 2019), IBCLN (Li et al. 2020a), LANet (Dong et al. 2021), YTMT (Hu and Guo 2021), DMGN (Feng et al. 2021), HGNet (Zhu et al. 2023), DSRNet (Hu and Guo 2023), RDRNet (Zhu et al. 2024b), DSIT (Hu, Wang, and Guo 2024), and RDNet(Zhao et al. 2025). Table 1 presents a quantitative comparison on public datasets, which primarily consist of daytime scenes, demonstrating that DMDNet achieves the best average performance. Table 2 presents a quantitative comparison on the NightIRS dataset. DMDNet attains the largest number of top-ranking metrics on both the transmission and reflection layers, demonstrating its adaptability to nighttime reflections, while maintaining a reasonable number of parameters and Floating-Point Operations (FLOPs).

Methods	Transmission Layer			Reflection Layer			Param	FLOPs
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	(M)↓	(G)↓
MambaIR	25.56	0.880	0.106	22.09	0.500	0.420	103.61	42.43
VMambaIR	25.89	0.884	0.100	22.06	0.490	0.414	83.76	38.05
MambaIRv2	24.84	0.868	0.118	21.66	0.490	0.445	88.38	40.38
DSMamba (Ours)	**26.27**	**0.889**	**0.093**	**22.31**	**0.522**	**0.403**	87.22	39.33

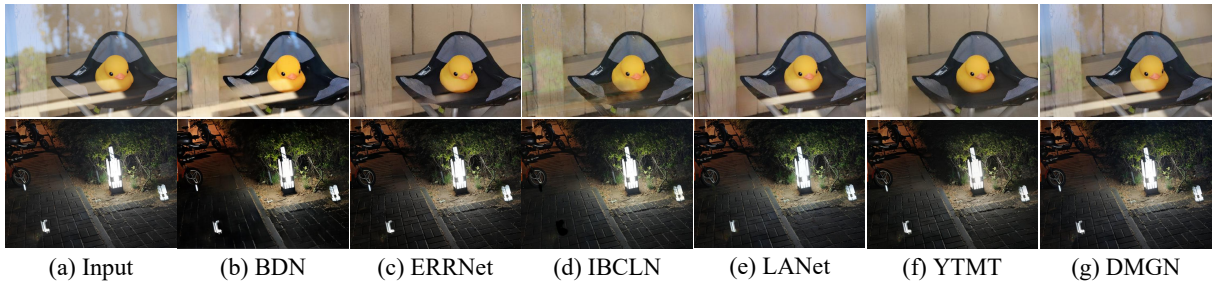
Table 3: Comparison with Mamba variants on public datasets. FLOPs are calculated for a 128×128 RGB image.

Figure 6 presents qualitative comparisons on the transmission layer. Our DMDNet achieves the most effective recovery, preserving structural details and suppressing residual reflections in daytime scenes. Even under nighttime conditions, where reflections closely resemble scene content, DMDNet effectively removes reflections.

Ablation Studies

DSMamba Visualization Analysis Figure 7 (b) visualizes the scanning region map $\boldsymbol{M}_{reg}$ generated by DA-RScan. The partitioned regions align well with the structural layout. Figures 7 (c)-(d) show that the original state-space matrices $\boldsymbol{B}$ and $\boldsymbol{C}$ exhibit a uniform distribution of activations, lacking discriminative focus. In contrast, $\boldsymbol{B}_{depth}$ and $\boldsymbol{C}_{depth}$ amplify activations in salient structural regions while suppressing responses in ambiguous areas, improving the structural awareness of the state evolution. Notably, $\boldsymbol{B}_{depth}$ appears darker than $\boldsymbol{C}_{depth}$, as it more strictly regulates the influence of inputs on the state, resulting in generally lower activation values.

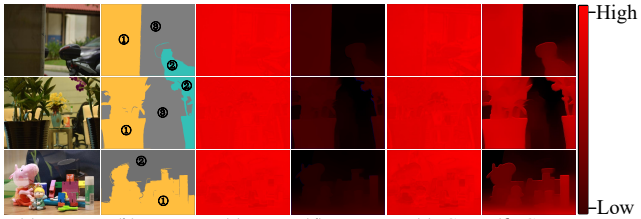
Comparison with Mamba Variants For a fair comparison and to adapt these methods to reflection separation, we replace our DSMamba with MambaIR (Guo et al. 2024), VMambaIR (Shi et al. 2025), and MambaIRv2 (Guo et al.



(a) Input (b) BDN (c) ERRNet (d) IBCLN (e) LANet (f) YTMT (g) DMGN

(h) HGNet (i) DSRNet (j) RDRNet (k) DSIT (l) RDNet (m) DMDNet* (n) Ground truth

Figure 6: Qualitative comparison with SOTAs on the transmission layer. Our DMDNet removes reflections most effectively in both daytime and nighttime scenes. The nighttime image is taken from the NightIRS dataset.



(a) Input (b) $\boldsymbol{M}_{reg}$ (c) $\boldsymbol{B}$ (d) $\boldsymbol{B}_{depth}$ (e) $\boldsymbol{C}$ (f) $\boldsymbol{C}_{depth}$

Figure 7: DSMamba visualization. $\boldsymbol{M}_{reg}$ shows the region-wise scanning order. (c)–(f) show the state-space matrices. $\boldsymbol{B}_{depth}$ and $\boldsymbol{C}_{depth}$ focus more on salient structures.

2025) while keeping the training strategy identical. As shown in Table 3, MambaIR and VMambaIR are constrained by fixed scanning orders, limiting their ability to disentangle overlapping layers. MambaIRv2's attentive state-space design is easily disturbed by reflections with similar semantics to scene content. By contrast, our DSMamba outperforms these variants on both $\boldsymbol{T}$ and $\boldsymbol{R}$ restoration.

Ablation Study on DSMamba As shown in Table 4, the best performance is achieved when DA-RScan is used for $\boldsymbol{T}$ and DA-GScan for $\boldsymbol{R}$, outperforming the original four-directional scanning strategy in Vmamba. The results also demonstrate the superiority of DS-SSM over the original SSM, and validate the effectiveness of SPE.

Ablation Study on MECM As shown in Table 5, both GPStream and SCStream are beneficial for performance. Furthermore, increasing the total number of memory experts N_{Exp} offers more diverse feature priors, while selecting an appropriate number of top-k experts N_{Exp}^{K} enables effective expert routing and reduces computational cost. The setting $N_{Exp} = 4$, $N_{Exp}^{K} = 2$ achieves a satisfactory balance.

Scanning Strategy		State-Space Model	SPE	Average			Param	FLOPs
$\boldsymbol{T}$	$\boldsymbol{R}$			PSNR ↑	SSIM ↑	LPIPS ↓	(M)↓	(G)↓
DA-RScan	DA-GScan	DS-SSM	✓	**26.27**	**0.889**	**0.093**	87.22	39.33
DA-RScan	DA-RScan	DS-SSM	✓	25.99	0.886	0.098	87.22	39.33
DA-GScan	DA-GScan	DS-SSM	✓	25.87	0.886	0.100	87.22	39.33
DA-GScan	DA-RScan	DS-SSM	✓	<u>26.09</u>	<u>0.887</u>	<u>0.096</u>	87.22	39.33
DA-RScan	DA-GScan	DS-SSM	×	25.66	0.882	0.105	87.22	39.33
DA-RScan	DA-GScan	Original	✓	25.78	0.884	0.098	83.29	38.55
Original	Original	DS-SSM	✓	25.69	0.884	<u>0.096</u>	89.36	39.21

Table 4: Ablation study on DSMamba. Results are reported on the transmission layer of public datasets.

GP-Stream	SC-Stream	N_{Exp}^{K}	N_{Exp}	Average			Param	FLOPs
				PSNR ↑	SSIM ↑	LPIPS ↓	(M)↓	(G)↓
✓	✓	2	4	**26.27**	**0.889**	**0.093**	87.22	39.33
×	×	0	0	24.93	0.882	0.100	55.92	35.85
×	✓	2	4	25.91	0.884	0.100	80.98	37.66
✓	×	2	4	<u>26.07</u>	<u>0.887</u>	<u>0.096</u>	80.98	37.67
✓	✓	1	3	25.93	0.884	0.098	79.39	37.60

Table 5: Ablation study on MECM. Results are reported on the transmission layer of public datasets.

Conclusion

We propose DMDNet to address the challenge of separating transmission and reflection layers when they exhibit similar contrast, especially in nighttime scenes. We present DSMamba, employing DAScan to prioritize structurally salient regions, and DS-SSM to enhance their influence on state evolution while suppressing the diffusion of interference. We introduce MECM, enabling experts to adaptively leverage cross-image knowledge to compensate for layer recovery. In addition, we construct the NightIRS dataset for evaluating nighttime reflection separation. Experimental results show that DMDNet outperforms SOTAs across all-day scenarios. One limitation is its reliance on supervised training data, and future work will explore unsupervised approaches.

References

Chen, X.; Jiang, X.; Tao, Y.; Lei, Z.; Li, Q.; Lei, C.; and Zhang, Z. 2025. FIRM: Flexible Interactive Reflection ReMoval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 2230–2238.

Dong, Z.; Xu, K.; Yang, Y.; Bao, H.; Xu, W.; and Lau, R. W. 2021. Location-aware single image reflection removal. In *Proceedings of the IEEE/CVF international conference on computer vision*, 5017–5026.

Elnenaey, A.; and Torki, M. 2024. Utilizing Multi-Step Loss for Single Image Reflection Removal. In *2024 IEEE/ACS 21st International Conference on Computer Systems and Applications (AICCSA)*, 1–6. IEEE.

Everingham, M.; Van Gool, L.; Williams, C. K.; Winn, J.; and Zisserman, A. 2010. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2): 303–338.

Fan, Q.; Yang, J.; Hua, G.; Chen, B.; and Wipf, D. 2017. A generic deep architecture for single image reflection removal and image smoothing. In *Proceedings of the IEEE International Conference on Computer Vision*, 3238–3247.

Feng, X.; Pei, W.; Jia, Z.; Chen, F.; Zhang, D.; and Lu, G. 2021. Deep-masking generative network: A unified framework for background restoration from superimposed images. *IEEE Transactions on Image Processing*, 30: 4867–4882.

Gu, A.; and Dao, T. 2023. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*.

Guo, H.; Guo, Y.; Zha, Y.; Zhang, Y.; Li, W.; Dai, T.; Xia, S.-T.; and Li, Y. 2025. Mambairv2: Attentive state space restoration. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 28124–28133.

Guo, H.; Li, J.; Dai, T.; Ouyang, Z.; Ren, X.; and Xia, S.-T. 2024. Mambair: A simple baseline for image restoration with state-space model. In *European conference on computer vision*, 222–241. Springer.

He, L.; Chang, Y.; Cong, R.; Liu, H.; Huang, S.; Tao, R.; and Zhao, Y. 2025. Rethinking depth guided reflection removal. *IEEE Transactions on Multimedia*.

He, X.; Yan, K.; Li, R.; Xie, C.; Zhang, J.; and Zhou, M. 2024. Frequency-adaptive pan-sharpening with mixture of experts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 2121–2129.

Hong, Y.; Lyu, Y.; Li, S.; Cao, G.; and Shi, B. 2022. Reflection removal with NIR and RGB image feature fusion. *IEEE Transactions on Multimedia*, 25: 7101–7112.

Hong, Y.; Zhong, H.; Weng, S.; Liang, J.; and Shi, B. 2024. L-differ: Single image reflection removal with language-based diffusion model. In *European Conference on Computer Vision*, 58–76. Springer.

Hu, Q.; and Guo, X. 2021. Trash or treasure? an interactive dual-stream strategy for single image reflection separation. *Advances in Neural Information Processing Systems*, 34: 24683–24694.

Hu, Q.; and Guo, X. 2023. Single image reflection separation via component synergy. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 13138–13147.

Hu, Q.; Wang, H.; and Guo, X. 2024. Single image reflection separation via dual-stream interactive transformers. *Advances in Neural Information Processing Systems*, 37: 55228–55248.

Huynh-Thu, Q.; and Ghanbari, M. 2008. Scope of validity of PSNR in image/video quality assessment. *Electronics letters*, 44(13): 800–801.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; et al. 2023. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, 4015–4026.

Lei, C.; and Chen, Q. 2021. Robust reflection removal with reflection-free flash-only cues. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14811–14820.

Lei, C.; Huang, X.; Zhang, M.; Yan, Q.; Sun, W.; and Chen, Q. 2020. Polarized reflection removal with perfect alignment in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 1750–1758.

Levin, A.; and Weiss, Y. 2007. User assisted separation of reflections from a single image using a sparsity prior. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(9): 1647–1654.

Li, C.; Yang, Y.; He, K.; Lin, S.; and Hopcroft, J. E. 2020a. Single image reflection removal through cascaded refinement. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 3565–3574.

Li, R.; Qiu, S.; Zang, G.; and Heidrich, W. 2020b. Reflection separation via multi-bounce polarization state tracing. In *European Conference on Computer Vision*, 781–796. Springer.

Liu, Y.; Tian, Y.; Zhao, Y.; Yu, H.; Xie, L.; Wang, Y.; Ye, Q.; Jiao, J.; and Liu, Y. 2024. Vmamba: Visual state space model. *Advances in neural information processing systems*, 37: 103031–103063.

Niklaus, S.; Zhang, X. C.; Barron, J. T.; Wadhwa, N.; Garg, R.; Liu, F.; and Xue, T. 2021. Learned dual-view reflection removal. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 3713–3722.

Peng, L.; Di, X.; Feng, Z.; Li, W.; Pei, R.; Wang, Y.; Fu, X.; Cao, Y.; and Zha, Z.-J. 2025. Directing mamba to complex textures: An efficient texture-aware state space model for image restoration. *arXiv preprint arXiv:2501.16583*.

Ranftl, R.; Lasinger, K.; Hafner, D.; Schindler, K.; and Koltun, V. 2022. Towards Robust Monocular Depth Estimation: Mixing Datasets for Zero-shot Cross-dataset Transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(3): 1623–1637.

Shi, Y.; Xia, B.; Jin, X.; Wang, X.; Zhao, T.; Xia, X.; Xiao, X.; and Yang, W. 2025. Vmambair: Visual state space model for image restoration. *IEEE Transactions on Circuits and Systems for Video Technology*.

Sun, J.; Chang, Y.; Jung, C.; and Feng, J. 2019. Multimodal reflection removal using convolutional neural networks. *IEEE Signal Processing Letters*, 26(7): 1011–1015.

Wan, R.; Shi, B.; Duan, L.-Y.; Tan, A.-H.; and Kot, A. C. 2017. Benchmarking single-image reflection removal algorithms. In *Proceedings of the IEEE International Conference on Computer Vision*, 3922–3930.

Wang, T.; Xie, M.; Cai, H.; Shah, S.; and Metzler, C. A. 2025. Flash-split: 2d reflection removal with flash cues and latent diffusion separation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 5688–5698.

Wang, Y.; Zhang, J.; Wei, R.; Gao, W.; and Wang, Y. 2024. Mfrgn: Multi-scale feature representation generalization network for ground-to-aerial geo-localization. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 2574–2583.

Wang, Z.; Bovik, A. C.; Sheikh, H. R.; and Simoncelli, E. P. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4): 600–612.

Wei, K.; Yang, J.; Fu, Y.; Wipf, D.; and Huang, H. 2019. Single image reflection removal exploiting misaligned training data and network enhancements. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8178–8187.

Xu, R.; Guo, M.; Wang, J.; Li, X.; Zhou, B.; and Loy, C. C. 2021. Texture memory-augmented deep patch-based image inpainting. *IEEE Transactions on Image Processing*, 30: 9112–9124.

Xue, T.; Rubinstein, M.; Liu, C.; and Freeman, W. T. 2015. A computational approach for obstruction-free photography. *ACM Transactions on Graphics (TOG)*, 34(4): 1–11.

Yang, J.; Gong, D.; Liu, L.; and Shi, Q. 2018. Seeing deeply and bidirectionally: A deep learning approach for single image reflection removal. In *Proceedings of the european conference on computer vision (ECCV)*, 654–669.

Yang, Y.; Ma, W.; Zheng, Y.; Cai, J.-F.; and Xu, W. 2019. Fast single image reflection suppression via convex optimization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8141–8149.

Zamfir, E.; Wu, Z.; Mehta, N.; Tan, Y.; Paudel, D. P.; Zhang, Y.; and Timofte, R. 2025. Complexity experts are task-discriminative learners for any image restoration. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 12753–12763.

Zhang, H.; Xu, X.; He, H.; He, S.; Han, G.; Qin, J.; and Wu, D. 2019. Fast user-guided single image reflection removal via edge-aware cascaded networks. *IEEE Transactions on Multimedia*, 22(8): 2012–2023.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 586–595.

Zhang, X.; Ng, R.; and Chen, Q. 2018. Single image reflection separation with perceptual losses. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4786–4794.

Zhao, H.; Li, M.; Hu, Q.; and Guo, X. 2025. Reversible decoupling network for single image reflection removal. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 26430–26439.

Zhong, H.; Hong, Y.; Weng, S.; Liang, J.; and Shi, B. 2024. Language-guided image reflection separation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 24913–24922.

Zhu, L.; Liao, B.; Zhang, Q.; Wang, X.; Liu, W.; and Wang, X. 2024a. Vision mamba: Efficient visual representation learning with bidirectional state space model. *arXiv preprint arXiv:2401.09417*.

Zhu, Y.; Fu, X.; Jiang, P.-T.; Zhang, H.; Sun, Q.; Chen, J.; Zha, Z.-J.; and Li, B. 2024b. Revisiting Single Image Reflection Removal In the Wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 25468–25478.

Zhu, Y.; Fu, X.; Zhang, Z.; Liu, A.; Xiong, Z.; and Zha, Z.-J. 2023. Hue guidance network for single image reflection removal. *IEEE Transactions on Neural Networks and Learning Systems*.

Zou, W.; Lu, X.; Yi, Z.; Zhang, L.; Fu, G.; Li, P.; and Xiao, C. 2024. Eyeglass reflection removal with joint learning of reflection elimination and content inpainting. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(10): 10266–10280.