Learning Latent Transmission and Glare Maps for Lens Veiling Glare Removal

(Supplementary Material)

Xiaolong Qian[1,*] Qi Jiang[1,*] Lei Sun[2,†] Zongxi Yu[1] Kailun Yang[3] Peixuan Wu[1]
Jiacheng Zhou[1] Yao Gao[1] Yaoguang Ma[1] Ming-Hsuan Yang[4] Kaiwei Wang[1,†]
[1]Zhejiang University [2]INSAIT, Sofia University "St. Kliment Ohridski"
[3]Hunan University [4]University of California, Merced

This supplementary material is organized as follows. First, §1 details the procedures for data acquisition and usage across different domains. Next, §2 presents a detailed motivation analysis regarding the learning of latent veiling glare maps. §4 provides further implementation details, including training configurations and hyperparameters. In §5, we discuss the broader societal impacts of our work, along with its limitations and potential future directions. Finally, §6 demonstrates the high data efficiency of our domain adaptation framework and presents additional visual comparisons.

1. Detailed Data Collection and Usage

This section details the three-domain dataset structure. We evaluate our method on two distinct optical systems: the large-aperture single lens (SL) and the metasurface-refractive hybrid lens (MRL), as shown in Fig. 1(a)-(b) and Fig. 2. A summary of acquisition setups, illumination conditions, and dataset splits is provided in Tab. 1. To illustrate the degradation components, Fig. 3 shows a comparison among the clear image, the aberration-only capture (Source domain), and the compound case with additional veiling glare (`Sreen-Compound` domain). This reveals that veiling glare degrades the image through global contrast reduction and color shifts, compounding the intrinsic optical blur. While identical scenes are presented here for direct comparison, we ensure strictly disjoint scene contents across all dataset splits in our experiments to prevent data leakage. All procedures described below are conducted independently for each system, resulting in two parallel datasets.

1.1. Data Acquisition Setup

We employ two distinct data acquisition setups: a controlled monitor-based setup for the Source and `Screen-Compound` domains, and an in-the-wild capture setup for the `Realworld-Compound` domain.

Monitor-based Setup. We use a screen-capture setup [2, 13] for both the Source and `Screen-Compound` domains. This approach is chosen because precise lens parameters for accurate simulation are not available. In this setup, a high-resolution monitor displays the GT images, and the optical system is placed on a controlled mounting setup to capture the screen. The illumination conditions are carefully controlled to distinguish the two domains:

- **Source Domain:** The environment is kept completely dark to avoid any ambient light, ensuring that the captured images contain only intrinsic optical aberrations without veiling glare.
- **`Screen-Compound` Domain:** Under the same geometric alignment, we introduce an external light source (Fig. 1(c)) to generate compound degradation. This source is placed off the optical axis and remains outside the field of view. This configuration simulates the common real-world scenario where invisible external light induces diffuse veiling glare. We employ this standardized lighting setup to establish a controlled benchmark. Despite lacking the variability of natural environments, this setup provides the essential paired data required for full-reference validation. In contrast, the `Realworld-Compound` domain (detailed below) encompasses random, diverse lighting conditions to verify the method's generalization capability.

In-the-wild Capture Setup. We collect diverse real-world scenes in uncontrolled environments to form the `Realworld-Compound` domain. These images are captured under naturally varying illumination, where the intensity and distribution of light are not regulated. As a result, the collected data reflect the actual imaging behavior of the optical system in everyday scenarios in which glare arises naturally.

[*] Equal contribution. [†] Corresponding authors.

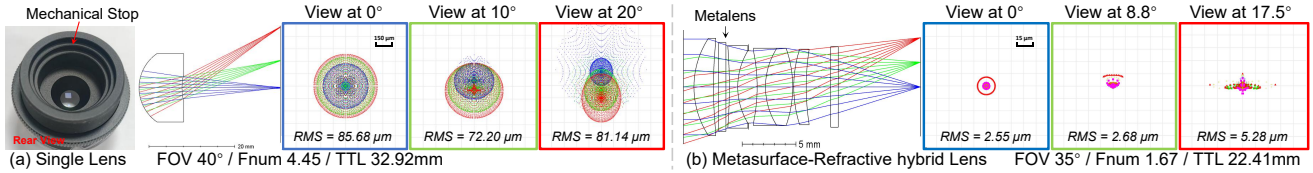
Domain	Acquisition Setup	Lighting Condition	GT Avail.	Data Quantity (SL / MRL)		Usage Setting
				Training	Testing	
Source	Monitor-based	Dark	✓	170 / 125	—	**Train:** Supervised (Aberration Only)
`Screen-Compound`	Monitor-based	Artificial light	✓†	50 / 50	42 / 25	**Train:** Unpaired Adaptation (GT Discarded) **Test:** Full-Ref. Metrics & Visual Comp.
`Realworld-Compound`	In-the-wild	Uncontrolled	×	50 / 50	51 / 11	**Train:** Unpaired Adaptation **Test:** No-Ref. Metrics & Visual Comp.

† GTs are available but explicitly discarded during training to ensure strictly unpaired adaptation.

Table 1. **Summary of Datasets and Usage Setting.** We evaluate our method on two independent optical systems: Large-aperture Single Lens (SL) and Metasurface-Refractive Lens (MRL). The table details the acquisition setup, lighting conditions, data splits, and specific usage for each domain.

(a) SL System (b) MRL System (c) Screen-Compound Domain Capture Setup

Figure 1. (a) A large-aperture single-lens (SL) system. (b) A metasurface–refractive hybrid-lens (MRL) system. (c) The capture setup for the `Screen-Compound` domain.



(a) Single Lens FOV 40° / Fnum 4.45 / TTL 32.92mm (b) Metasurface-Refractive hybrid Lens FOV 35° / Fnum 1.67 / TTL 22.41mm

Figure 2. **Optical specifications and simulated performance.** Simulated ray tracing layouts and corresponding RMS spot diagrams for (a) the SL imaging system and (b) the MRL imaging system.

Geometric Alignment. To ensure precise pixel-level correspondence for paired data, a two-stage geometric correction process is applied independently for each system.

- **Lens Distortion Correction:** Intrinsic parameters are calibrated using the MATLAB Camera Calibrator [5]. Chessboard images captured from multiple viewpoints allow computation of a distortion map that removes nonlinear distortions such as barrel or pincushion effects.
- **Affine Alignment:** Residual planar misalignment between the sensor and the monitor is corrected using an affine transformation. We extract N corner correspondences $\{(x_i, y_i) \rightarrow (x'_i, y'_i)\}$ from a high-definition chessboard displayed on the monitor. The transformation is modeled directly in homogeneous coordinates as:

$$\begin{bmatrix} x'_i \\ y'_i \\ 1 \end{bmatrix} = \mathbf{M}_{\text{aff}} \begin{bmatrix} x_i \\ y_i \\ 1 \end{bmatrix} = \begin{bmatrix} m_{11} & m_{12} & t_x \\ m_{21} & m_{22} & t_y \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_i \\ y_i \\ 1 \end{bmatrix}$$

The affine parameters are estimated by minimizing the sum of squared reprojection errors:

$$\min_{m_{ij},t} \sum_{i=1}^{N} \left\| \begin{bmatrix} m_{11} & m_{12} \\ m_{21} & m_{22} \end{bmatrix} \begin{bmatrix} x_i \\ y_i \end{bmatrix} + \begin{bmatrix} t_x \\ t_y \end{bmatrix} - \begin{bmatrix} x'_i \\ y'_i \end{bmatrix} \right\|^2$$

This correction is strictly applied to the Source and `Screen-Compound` domains to ensure high-fidelity alignment for quantitative metrics. Conversely, for the `Realworld-Compound` domain, geometric correction is

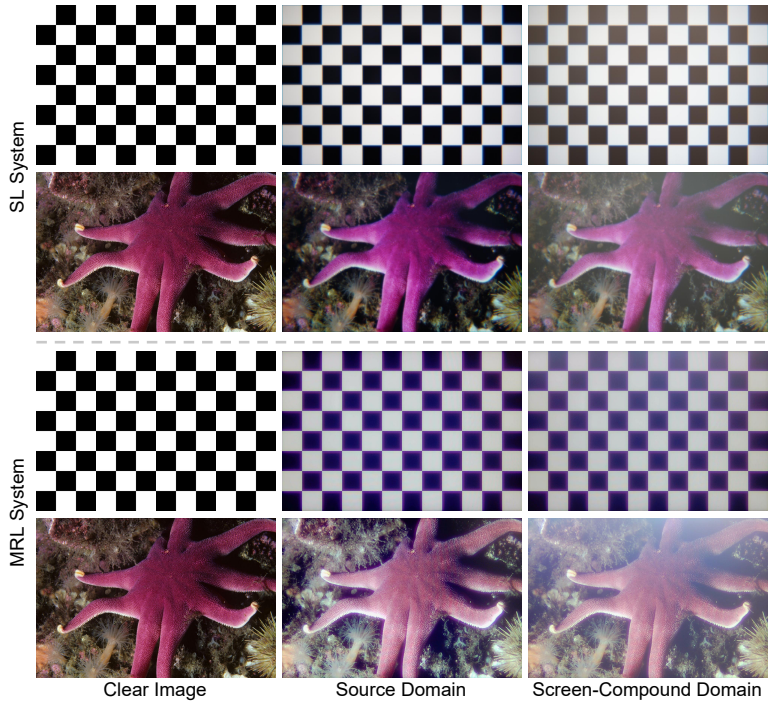


Figure 3. Visual Decomposition of Compound Degradation.

intentionally omitted. By operating directly on uncorrected RGB images containing native distortion, we explicitly validate the robustness of the model in practical, calibration-free deployment scenarios.

1.2. Domain Definitions and Usage

We define three domains. The Source domain establishes baseline capabilities for aberration synthesis and correction. For unpaired adaptation, each target domain (`Screen-Compound` and `Realworld-Compound`) is treated separately, ensuring the model adapts to each domain independently.

Source Domain (Paired, Aberration-only).

- **Acquisition:** Using the aligned screen-capture setup, clean images from the DIV2K dataset [11] are re-captured in a dark environment. This ensures that only the intrinsic spatial-varying aberrations are recorded.
- **Data Split:** This domain yields 170 pairs for the SL system and 125 pairs for the MRL system.
- **Usage:** These pairs provide fully supervised supervision for modeling and correcting spatial-varying aberrations.

Target Domain: `Screen-Compound` (Unpaired Training, Paired Testing). This domain introduces controlled veiling glare and serves a dual purpose for training and evaluation.

- **Acquisition:** The process follows the Source domain setup, except that a single external light source is directed toward the lens (Fig. 1), inducing compound degradation from both aberrations and veiling glare.
- **Data Split:** We collect 50 training images and 42 test images for the SL system (50 training, 25 test for MRL). All images are captured with corresponding GT pairs.
- **Usage:** The dataset is strictly split:
  - **Training (Unpaired):** The 50 degraded images form the target domain, with GTs explicitly discarded to ensure unpaired adaptation.
  - **Testing (Paired):** Held-out pairs are reserved exclusively for full-reference quantitative evaluation.

Target Domain: `Realworld-Compound` (Unpaired). This domain serves to adapt the model to uncontrolled, realistic scenarios and evaluate its generalization capabilities.

- **Acquisition:** Diverse real-world scenes featuring natural, complex compound degradation are captured using independent optical systems.
- **Data Split:** No GT is available for these images. The dataset comprises 50 unpaired training images and 51 test images for the SL system (50 training, 11 test for MRL).
- **Usage:** Unpaired training images are used for adaptation; test images are used for qualitative comparison and quantitative evaluation via no-reference IQA metrics.

2. Motivation Analysis

Our framework addresses compound degradation arising from optical aberrations and veiling glare. Although we aim to mitigate both artifacts, our methodology prioritizes the modeling of the veiling glare component. This section outlines the rationale for this focus.

Baseline for Aberration Correction. Existing literature demonstrates that supervised networks trained on paired datasets effectively handle spatially varying aberrations [3, 13]. We follow this data-driven strategy by constructing a paired source domain to train the backbone network. This approach yields robust performance for aberration correction. Consequently, our work focuses on the removal of veiling glare, a more challenging and under-constrained problem, rather than the incremental improvements of aberration correction.

Data Scarcity in Veiling Glare. Unlike optical aberrations, acquiring paired data for real-world veiling glare is infeasible due to the sensitivity to dynamic lighting conditions. Moreover, high-fidelity simulation is restricted by the inaccessibility of proprietary opto-mechanical structures and the prohibitive computational cost of non-sequential ray tracing. This lack of GT results in an under-constrained degradation model (Eq. 1 in the main text). While the source domain constrains the PSF term K^p, the transmission map T^p and the glare map I_g^p remain unsupervised. Consequently, general domain adaptation methods yield suboptimal performance without explicit physical guidance.

Targeted Design for Glare Modeling and Removal. To bridge the supervision gap, we introduce VeilGen for data synthesis and DeVeiler for image restoration. Both components leverage the proposed LOTGMP prior to ensure physical consistency. In VeilGen, the prior guides the generation of plausible transmission and glare maps in latent space. This synthesis is validated in two ways: the resulting degradations are more realistic (Fig. 6 in the main text), and a network trained on this synthetic data achieves superior restoration performance (Tab. 3 in the main text). Similarly, DeVeiler incorporates the prior during restoration. The VGEnc module estimates latent glare and transmission, which the VGCM module then uses to selectively suppress activations in glare-affected regions, a process confirmed by our feature analysis (Fig. 8 in the main text). This dual application of the prior effectively addresses the challenges of data scarcity and model ambiguity. Future work could explore a unified architecture that jointly learns to correct both aberration and glare.

3. Generalization Analysis

While retraining a model for a specific new lens is a common practice in computational aberration correction [1, 13], our framework features a highly streamlined workflow that enables rapid and cost-effective generalization to novel optical systems. The generalization capability of our pipeline is supported by three key aspects:

Flexible Source Data Construction. The acquisition of source domain data in our framework is highly adaptable. If the optical design parameters are known, paired source data can be efficiently generated via optical simulation. Conversely, in strictly blind scenarios where lens parameters are completely unknown, the source data can be easily acquired through a simple screen capture setup, eliminating the need for complex calibration.

High Target Data Efficiency. Bridging the domain gap to a new lens typically requires extensive target data. However, our unsupervised adaptation process is remarkably efficient. As detailed in our subsequent data efficiency analysis (Fig. 4), our model requires only $\sim$25 unpaired degraded images from the target system to achieve stable and high-quality restoration. This drastically reduces the real-world deployment cost.

Empirical Validation on Diverse Optics. We have rigorously validated this streamlined pipeline on two fundamentally distinct hardware prototypes: the large-aperture single lens (SL) and the metasurface-refractive hybrid lens (MRL). Notably, both evaluations were conducted under a completely blind setting, confirming that our framework is robust and readily applicable to diverse, unknown optical systems in practice.

4. More Implementation Details

VeilGen Training. In stage I, the VeilGen is initialized from the pre-trained Stable Diffusion v2-1 [10]. We employ AdamW [9] with a constant learning rate of 1×10^{-5} for both the main diffusion backbone and the LOTGMP module. LOTGMP utilizes a ResBlock-Attention backbone, with Trans/Glare Heads implemented as convolutional layers. Physical consistency is enforced explicitly by Eq. 1, which constrains the prediction of Trans-Head to be multiplicative (modulation) and the Glare-Head to be additive (bias). This structural prior compels the heads to disentangle distinct physical roles The model is trained for $9k$ iterations with a batch size of 16. Following standard practices in hybrid-domain diffusion learning [12], we set the probability parameter p to 0.3 during training and the mixture coefficient w to 0.85 during inference. We set the reference timestep to $t^* = 0$ to extract latent maps from the fully denoised state, ensuring estimation precision. To balance generation quality and efficiency, we employ a 10-step sampling strategy. Using the Flickr2K [11] dataset as the source, we synthesize 500 high-resolution images (1280×1920). This process requires approximately 3 hours on a single NVIDIA A100 GPU.

Distillation to DDN. In stage II, we distill the trained VeilGen for $25k$ iterations using the Adam optimizer [7] with

a cosine annealing learning rate schedule, starting from 2×10^{-4} and decaying to 1×10^{-7}. The batch size is 8, and the input patch size is 256×256. This distillation preserves VeilGen's physically grounded degradation behavior while producing an efficient forward model suitable for supervision in Stage III.

DeVeiler Training. In stage III, DeVeiler is trained end-to-end in two phases. In phase I (pre-training), we train on paired source-domain data for $100k$ iterations using Adam [7], with a learning rate that decays from 2×10^{-4} to 1×10^{-7} via cosine annealing. In phase II (fine-tuning), we adapt the model on a mixed dataset of 500 generated pseudo-pairs and the original source pairs for $5k$ iterations. The learning rate decays from 5×10^{-5} to 1×10^{-7} with cosine annealing. Both phases use a batch size of 8, a 256×256 patch size, and random horizontal/vertical flips for augmentation. The loss weight λ_{rev} is set to 1.0. The fine-tuning phase is highly efficient, completing in approximately 40 minutes on an NVIDIA A100 GPU.

Text prompts. We utilize distinct text prompts for the source and target domains. The source domain prompt focuses solely on aberration: *a photograph with spatial-varying PSF blur, optical aberrations, defocus, and chromatic fringing.* For the target domains, which include the `Screen-Compound` and `Realworld-Compound`, the prompt is expanded to describe the full compound degradation: *a photograph with spatial-varying PSF blur, optical aberrations, defocus, chromatic fringing, and noticeable stray light with veiling glare.*

Baseline Implementations. For fair comparison, we use official implementations, retraining learning-based models on our dataset with default settings.

5. Discussion

5.1. Societal Impacts

This work alleviates the inherent trade-off between optical miniaturization and image quality. While aberration correction is established, veiling glare remains a bottleneck for compact optical systems. By addressing this compound degradation, we facilitate high-performance imaging in hardware with strict spatial limits, such as medical endoscopes, autonomous drones, and mobile devices. Beyond specific applications, our framework demonstrates a strategy for inverse problems in data-scarce domains. Instead of purely black-box approaches, we incorporate physical models into the latent space to synthesize realistic training pairs from unpaired data. This physics-aware generation strategy offers a scalable solution for other fields lacking GT, including underwater imaging and astronomy.

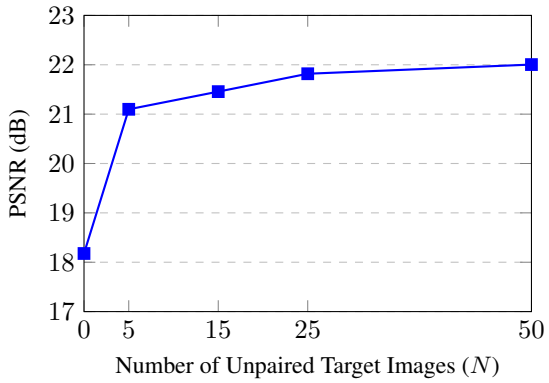


Figure 4. **Data Efficiency Analysis.** PSNR performance on the `Screen-Compound` test set with varying numbers of unpaired target training images (N). Note that $N = 0$ denotes the non-adapted Source-only baseline.

5.2. Limitations and Future Work

Limitations. Although our framework delivers favorable restoration results, slight color deviations may persist under intense veiling glare (see Fig. 7). This is fundamentally due to the ambiguity in overexposed regions, where the additive glare merges with the saturated scene signal. Regarding data acquisition, the source domain is constructed using images with a fixed scene depth. Although aberrations are theoretically depth-dependent, real-world performance indicates that this approximation is acceptable for compact optical systems. From a modeling perspective, the current framework learns implicit light representations rather than explicit parameters (*e.g.*, 3D position). While sufficient for effective restoration, explicit parameterization could offer additional controllability in future iterations.

Future Work. To address these limitations, future research could incorporate explicit light-source modeling for color recovery and extend the source domain to include multiple scene depths. Additionally, we plan to investigate frequency-domain strategies for improved glare separation and adapt advanced generative backbones for efficient one-step synthesis.

6. More Results

Analysis of Target Data Efficiency We analyze the impact of target training set size ($N \in \{0, 5, 15, 25, 50\}$) on restoration performance. To decouple the adaptation strategy from the specific architecture of DeVeiler, we employ a standard SwinIR [8] backbone. Figure 4 indicates that a subset of $N = 5$ yields substantial improvements over the baseline ($N = 0$). Furthermore, performance converges at $N = 15$, reaching parity with the full dataset ($N = 50$). These results verify the capability of the framework for robust few-shot adaptation with limited data.

Physical Interpretability of LOTGMP. We validated this via a "black screen" experiment (capturing $I_c \approx 0$ with side-illumination to isolate pure glare). The predicted glare map (Fig. 8(c)) exhibits strong structural consistency with physical measurements (Fig. 5(a)), confirming that LOTGMP accurately learns the glare distribution.

Robustness in Challenging Scenes. We further demonstrate the robustness of our method on specific challenging scenarios. As shown in Fig. 5(b), our method effectively recovers high-frequency details in the line pairs scene captured by SL. Furthermore, it successfully removes veiling glare and preserves underlying structures in the high-contrast HDR scene captured by MRL (Fig. 5(c)).

More Visual Results. To further verify the effectiveness of our method, we present more visual comparison results between the proposed DeVeiler and other advanced methods across both the `Screen-Compound` and `Realworld-Compound` domains. Specifically, the results captured by the SL and MRL systems are shown for the `Screen-Compound` domain in Fig. 6 and Fig. 7, and for the `Realworld-Compound` domain in Fig. 8 and Fig. 9, respectively.

In these scenarios, the input images exhibit compound degradation, where light sources introduce veiling glare, leading to a noticeable reduction in contrast and color shifts alongside intrinsic optical aberrations. Baseline aberration correction methods [8], trained on the source domain, show limited generalization to the unseen veiling glare. Cascaded pipelines employing dehazing models [12] can improve global contrast but may result in the smoothing of fine textures. Similarly, cascaded flare removal approaches [4], typically designed for localized artifacts, are less effective in addressing the spatially diffusive nature of veiling glare. Furthermore, general domain adaptation methods [6, 12], without explicit physical modeling for glare, may exhibit color deviations. In contrast, DeVeiler leverages the latent veiling glare maps to achieve favorable results across diverse scenarios, preserving structural details and recovering color fidelity. In regions of intense illumination, the input signal approaches saturation (Fig. 7). While signal loss hinders full recovery, DeVeiler minimizes color deviations compared to competing methods.

References

[1] Liqun Chen, Yuxuan Li, Jun Dai, Jinwei Gu, and Tianfan Xue. A physics-informed blur learning framework for imaging systems. In *CVPR*, 2025. 4

[2] Qikai Chen, Jiacheng Zhou, Sijie Pian, Jingang Xu, Xingyi Li, Bihua Li, Chentao Lu, Zhuning Wang, Qi Jiang, Shanhe Qin, Hantao Zhan, Benhao Zhang, Xu Liu, Kaiwei Wang, and Yaoguang Ma. Hybrid meta-optics enabled compact augmented reality display with computational image reinforcement. *ACS Photonics*, 2024. 1

[3] Shiqi Chen, Huajun Feng, Dexin Pan, Zhihai Xu, Qi Li, and Yueting Chen. Optical aberrations correction in postprocessing using imaging simulation. *ACM Transactions on Graphics*, 2021. 4

[4] Yuekun Dai, Chongyi Li, Shangchen Zhou, Ruicheng Feng, Yihang Luo, and Chen Change Loy. Flare7K++: Mixing synthetic and real datasets for nighttime flare removal and beyond. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 6

[5] The MathWorks Inc. MATLAB version: 9.13.0 (R2022b), 2022. 2

[6] Qi Jiang, Zhonghua Yi, Shaohua Gao, Yao Gao, Xiaolong Qian, Hao Shi, Lei Sun, JinXing Niu, Kaiwei Wang, Kailun Yang, and Jian Bai. Representing domain-mixing optical degradation for real-world computational aberration correction via vector quantization. *Optics & Laser Technology*, 2025. 6

[7] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 4, 5

[8] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. SwinIR: Image restoration using swin transformer. In *ICCVW*, 2021. 5, 6

[9] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019. 4

[10] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 4

[11] Radu Timofte, Eirikur Agustsson, Luc Van Gool, Ming-Hsuan Yang, Lei Zhang, Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, Kyoung Mu Lee, et al. NTIRE 2017 challenge on single image super-resolution: Methods and results. In *CVPRW*, 2017. 3, 4

[12] Ruiyi Wang, Yushuo Zheng, Zicheng Zhang, Chunyi Li, Shuaicheng Liu, Guangtao Zhai, and Xiaohong Liu. Learning hazing to dehazing: Towards realistic haze generation for real-world image dehazing. In *CVPR*, 2025. 4, 6

[13] Jianing Zhang, Jiayi Zhu, Feiyu Ji, Xiaokang Yang, and Xiaoyun Yuan. Degradation-modeled multipath diffusion for tunable metalens photography. *ICCV*, 2025. 1, 4

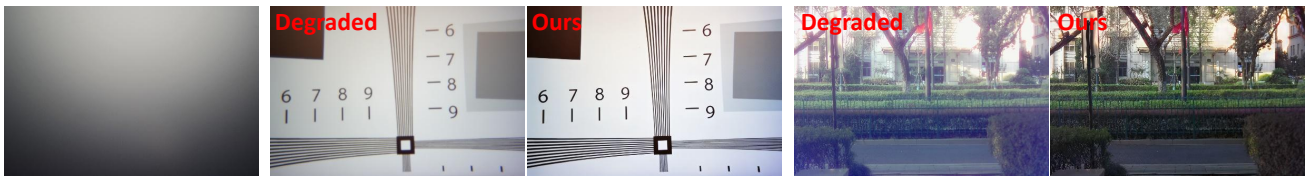


Figure 5. (a) Physically measured glare map ($I_c \approx 0$). (b) High-frequency detail recovery on an ISO resolution chart (SL). (c) Robust veiling glare removal in a high-contrast HDR environment (MRL).

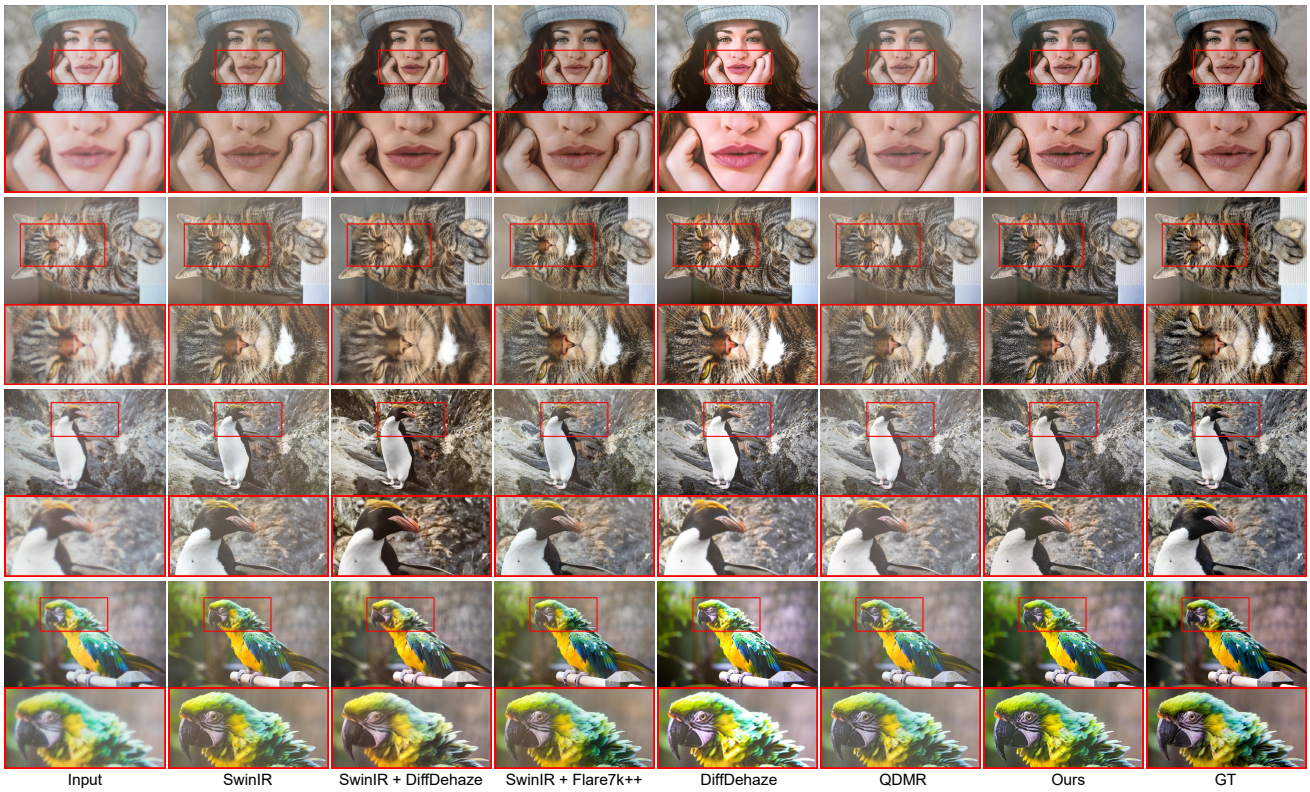


Figure 6. **Visual results on the `Screen-Compound` domain captured by the SL system.** The method is shown at the bottom of each case. Zoom in for the best view.

Figure 7. **Visual results on the `Screen-Compound` domain captured by the MRL system.** The method is shown at the bottom of each case. Zoom in for the best view.

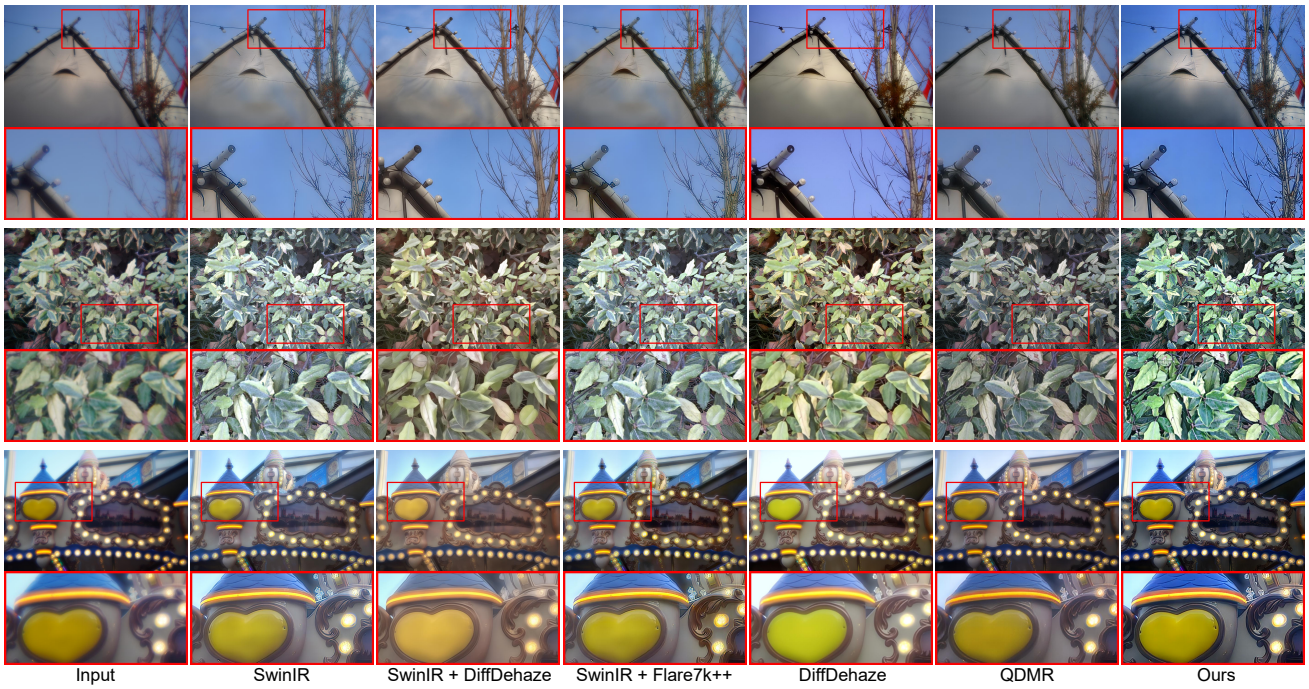


Figure 8. **Visual results on `Realworld-Compound` domain captured by the SL system.** The method is shown at the bottom of each case. Zoom in for the best view.

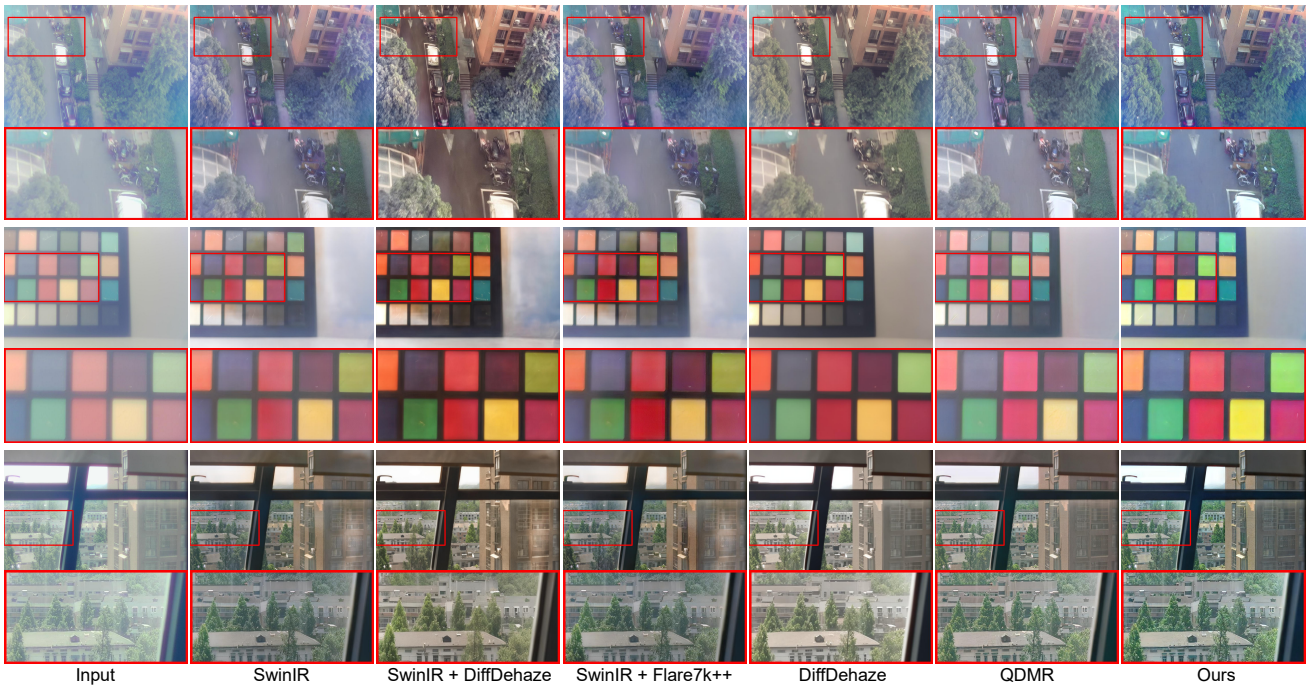


Figure 9. **Visual results on `Realworld-Compound` domain captured by the MRL system.** The method is shown at the bottom of each case. Zoom in for the best view.