Supplementary Material for Degradation-Modeled Multipath Diffusion for Tunable Metalens Photography

Jianing Zhang[2,3] Jiayi Zhu[1] Feiyu Ji[1] Xiaokang Yang[1] Xiaoyun Yuan[1,*]
[1]MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University
[2]Fudan University [3]Tsinghua University
22110720080@m.fudan.edu.cn {zjy-bixi, jownr25, xkyang, yuanxiaoyun}@sjtu.edu.cn

1. MetaCamera Details

1.1. Fabrication

To achieve an ultra-compact imaging system, a nano-lens metasurface optical design is employed and integrated with a miniaturized image sensor (OV6946, OmniVision Technologies). Traditional imaging systems rely on bulky refractive optics, which limit miniaturization. In contrast, the use of a planar nano-lens metasurface enables ultra-thin imaging, significantly reducing the system's size while maintaining effective wavefront control.

As shown in Fig. 2, the nano-lens consist of a periodic array of subwavelength-scale nano-pillars, where the local phase response is determined by the geometric parameters of each pillar. The phase modulation follows a differentiable polynomial function[1]:

$$\begin{aligned}\varphi =& 2\pi(6.051 - 0.0203\lambda + 2.26\eta \\ &+ 1.371 \times 10^{-5}\lambda^2 - 0.00295\lambda\eta + 0.797\eta^2),\end{aligned} \tag{1}$$

where λ is the incident light wavelength (in nm), and η represents the duty cycle of the nano-pillars. The height of the nano-lens is $705\,\mathrm{nm}$, and the diameter of the whole nano-lens is $200\,\mu\mathrm{m}$, approximately comprising 571×571 units.

To optimize the optical performance, Finite-Difference Time-Domain (FDTD) simulations are conducted, and a Neural Nano-Optics approach is employed to refine the nano-lens structure. The fabrication process involves Electron-Beam Lithography (E-beam Lithography) on a 705 nm-thick silicon nitride (Si_3N_4) layer deposited on a fused silica substrate.

The processed optical signals are captured by the OV6946 image sensor, which provides digital readout through an ADC chip (OV426, OmniVision Technologies). By leveraging this metasurface optical approach, the entire imaging module achieves a compact form factor of approximately 1 mm in size.

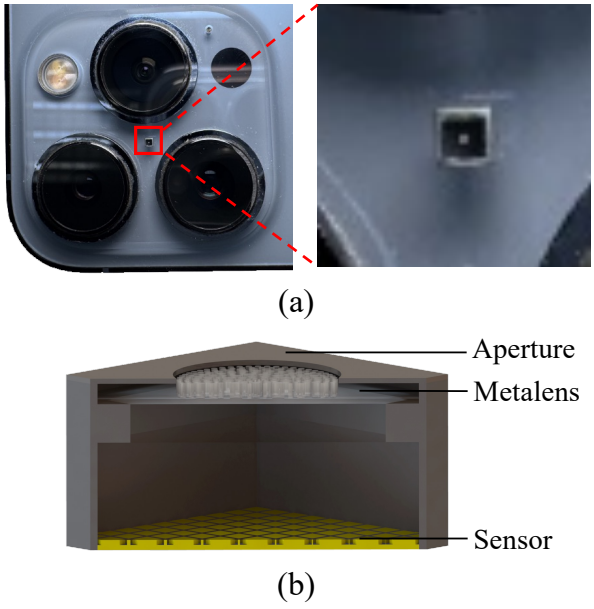

Figure 1. (a) Comparison between our MetaCamera system and the iPhone camera module. (b) Longitudinal cross-sectional schematic.

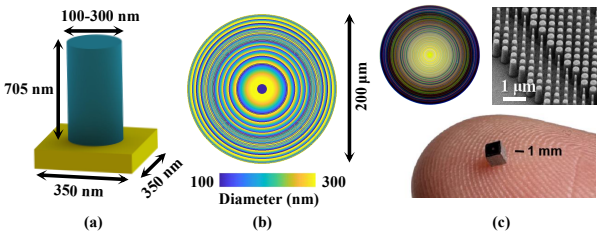

Figure 2. (a) Schematic of the metalens unit cell, consisting of a Si_3N_4 nano-pillar on a SiO_2 substrate. The pillar diameter varies radially to modulate the optical phase. (b) Diameter distribution map of the metalens. (c) Top: Optical and SEM images of the fabricated metalens. Bottom: Optical image of the fully integrated MetaCamera, demonstrating its ultra-compact size.

1.2. PSF Evaluation

To enable comparative algorithms that rely on point spread function (PSF) measurements and to validate the agreement between our fabricated metalens and its simulated design, we conducted direct PSF evaluations. As illustrated in Fig. 3, we employed an RGB LED module (CREE-XML-5050 RGBD-10W), with each color channel featuring a pixel size of approximately 1 mm. The LED was placed at a distance of 1 m from the metalens, thereby approximating a point source. We recorded the PSFs by sequentially activating each color channel while maintaining a fixed LED position. The measured PSFs are shown in Fig. 3, from which it can be observed that our fabricated metalens are consistent with the designed ones.

2. Data Collection

To obtain paired data, we employed a setup where a camera directly captures a high-definition display. Initially, mechanical micro-adjustments were made to align the captured images with the displayed content roughly. To achieve precise pixel-level correspondence, an affine transformation was applied.

At the start of data collection, a high-definition chessboard pattern was displayed on the monitor. Corner detection was then used to extract the coordinates of the chessboard corners from the captured images. Given N point correspondences $\{(x_i, y_i) \rightarrow (x_i', y_i')\}_{i=1}^{N}$, where for each i, the pair (x_i, y_i) denotes a point in the reference image, and (x_i', y_i') is the corresponding point in the captured image, the affine transformation is modeled as

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} + \begin{bmatrix} e \\ f \end{bmatrix}. \tag{2}$$

The parameters a, b, c, d, e, and f are estimated by minimizing the sum of squared errors:

$$\min_{a,b,c,d,e,f} \sum_{i=1}^{N} \left\| \begin{bmatrix} x_i' \\ y_i' \end{bmatrix} - \left(\begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} x_i \\ y_i \end{bmatrix} + \begin{bmatrix} e \\ f \end{bmatrix} \right) \right\|^2. \tag{3}$$

This process yields high-quality paired images for training and testing.

3. Implementation details

We used the following prompts in our work.

- Negative prompt: *oil painting, cartoon, blur, dirty, messy, low quality, deformation, low resolution, oversmooth*.
- Neutral prompt: *A well-balanced image with acceptable sharpness average detail and natural colors presenting the scene in a simple and straightforward manne*.
- Positive prompt: *A high-resolution 8K ultra-realistic image with sharp focus vibrant colors and natural lighting.*

Our approach relies solely on imaging quality descriptions in the prompt, avoiding dependence on scene content and ensuring consistent performance.

We set the rank of LoRA r to 32 for the UNet and 16 for the VAE encoder, which strikes a good balance between model complexity and super-resolution performance. To obtain the low-frequency components of high-resolution images as the training target for the neutral condition, we employed both bilateral and Gaussian filters. The bilateral filter was used to preserve edges while smoothing the image, whereas the Gaussian filter helped in reducing noise and detail, focusing on capturing the overall structure and color information.

Regarding the attention network($\mathcal{N}_A$), given the relatively low dimensions of the the FWHM score (S_f) and NR-IQA score (S_i), we employ an MLP-based structure. Specifically, we first multiply the two score matrices, flatten the result into a vector, and pass it through the MLP to generate the output Q matrix. This Q matrix is then incorporated into different UNet blocks via LoRA to enhance attention mechanisms.

Regarding the probabilistic selection of different training paths, the initial phase primarily focuses on degradation learning, enabling the rapid generation of a large number of pseudo images. Consequently, during the first 1000 iterations, the probability of selecting the negative path is set to 1, while the other paths are not considered.

After 1000 iterations, the probability of the negative path is gradually reduced. Between 1000 and 1200 iterations, its probability decreases to 0.7, while the positive path is introduced with a probability of 0.3. The neutral path is not included at this stage, as learning from the positive path is more challenging. Once the model sufficiently learns from the positive path, it can later adapt more effectively to the neutral path.

As training progresses, the probability of the negative path continues to decrease, while the positive path gains more emphasis. By 2000 iterations, the probability of the negative path stabilizes at 0.1, while the positive path dominates at 0.9, as the negative path has reached stable performance and no longer requires intensive learning.

Beyond 10,000 iterations, training on the neutral path is introduced. At this stage, the probabilities are set to 0.1 for the negative path, 0.15 for the positive path, and 0.75 for the neutral path. Since prior training on the positive path has established a solid foundation, the model can quickly adapt to the neutral path. After 12,000 iterations, the probabilities are adjusted to 0.1 (negative path), 0.7 (positive path), and 0.2 (neutral path). This distribution is maintained for the remainder of the training process until convergence at 15,000 iterations.

In practice, strict adherence to this probability schedule is not required. Instead, the general training strategy should be followed: prioritize learning the negative path first, then

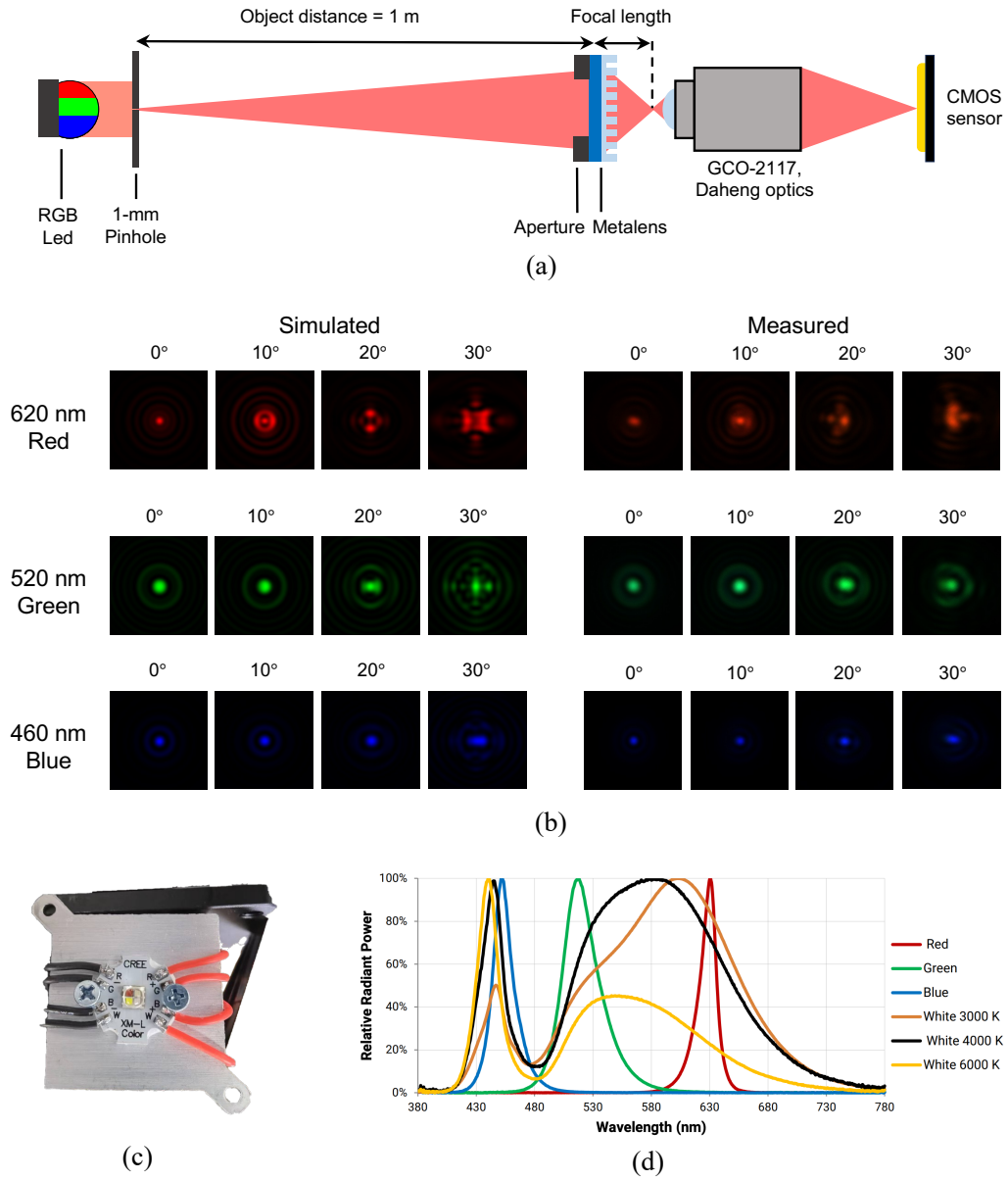

Figure 3. (a) System configuration for point spread function (PSF) calibration, utilizing a color LED and a pinhole as the point light source. (b) Simulated and experimentally measured PSFs for red, green, and blue point light sources at incidence angles of 0°, 10°, 20°, and 30°. Owing to the circular symmetry of the nano-optics, the resulting PSFs are symmetric, and there is strong agreement between simulation and measurement. Note that at 30°, the PSFs extend partially beyond the microscope objective's NA, resulting in incomplete capture. (c) Optical image of the RGBW LED (XLamp XM-L Color LEDs, Cree LED). (d) The relative spectral power distribution of the LED. We only used the red, green, and blue dies for PSF calibration.

focus on the positive path, and finally introduce the neutral path. This progressive approach ensures efficient learning while allowing the model to adapt dynamically throughout training.

4. More Qualitative results

We present additional results on real-world images in Fig. 4 as well as our dataset in Fig. 5. Our method significantly improves image quality by producing results with accurate color tones, rich details, and enhanced realism.

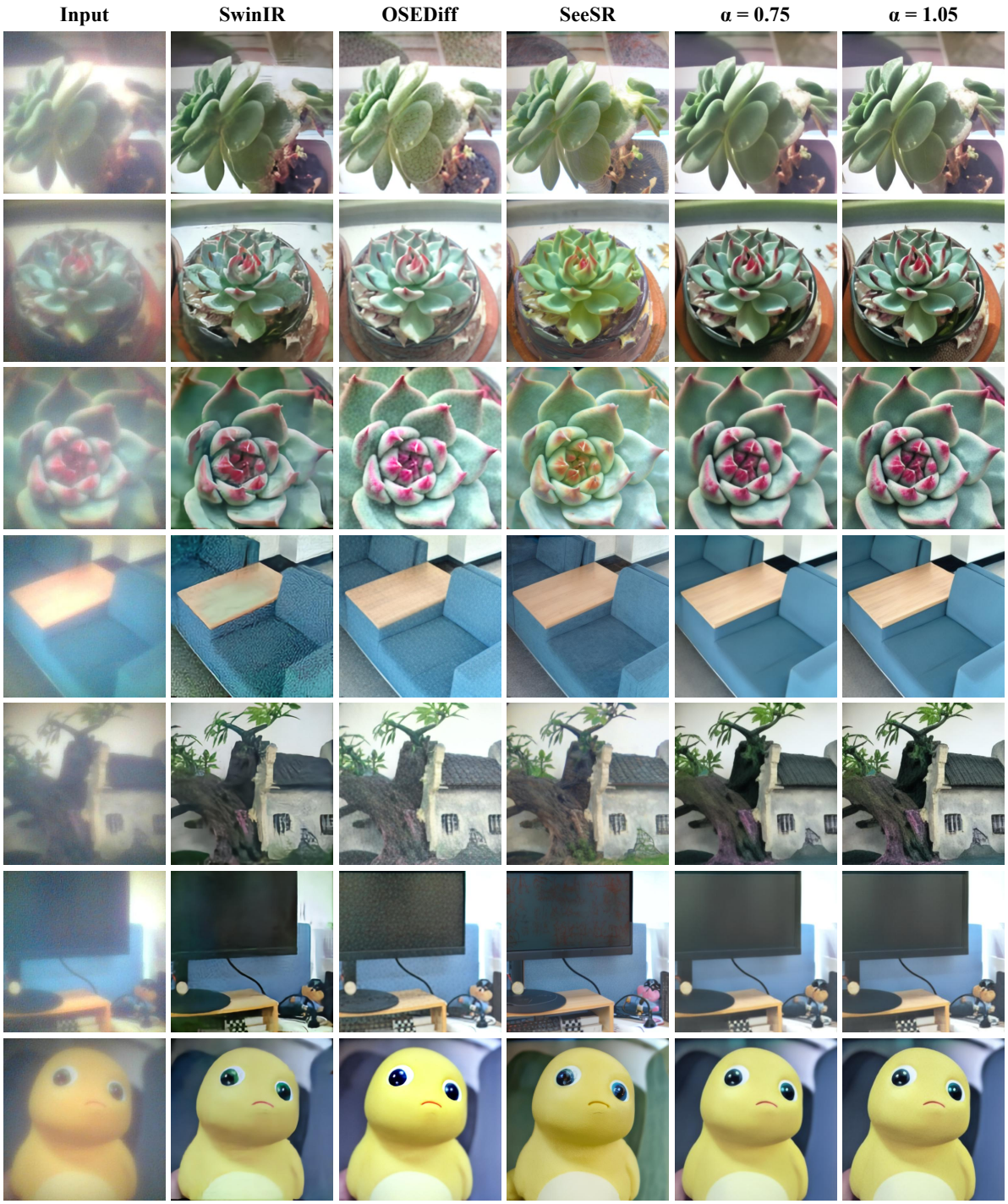

Figure 4. Qualitative comparisons of different methods on real-world images captured by our system.

5. Comparison with Pretrained Models

In this paper, we compared diffusion models trained on our own dataset rather than utilizing pretrained models trained on large-scale datasets. This decision was driven by the observation that pretrained models fail to adequately adapt to our specific dataset characteristics. In Fig. 6, we demonstrate the performance of pretrained super-resolution diffusion models, clearly showing that such models are incapable of addressing chromatic aberrations and generate the results of significantly lower quality.

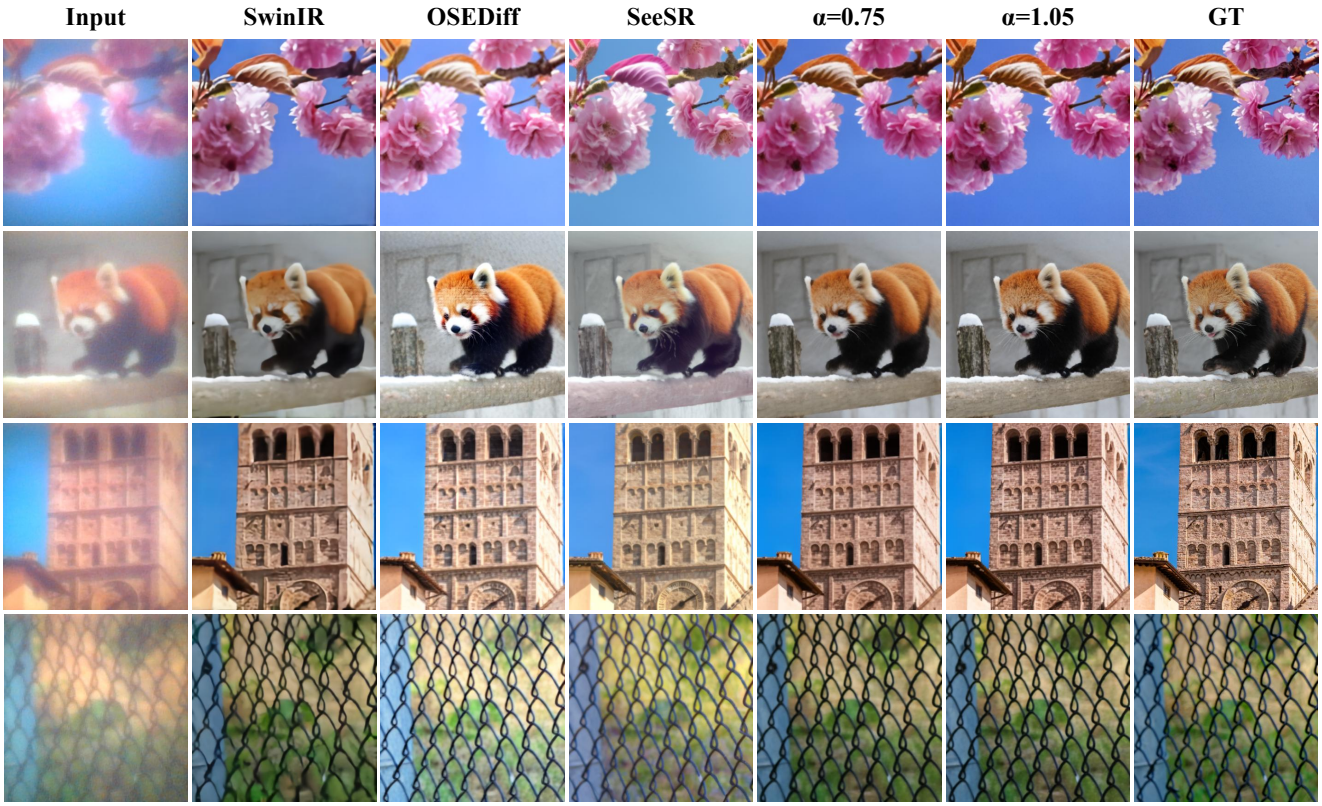

Figure 5. Qualitative comparisons of different methods on our **unseen** test dataset, zoom in for details.

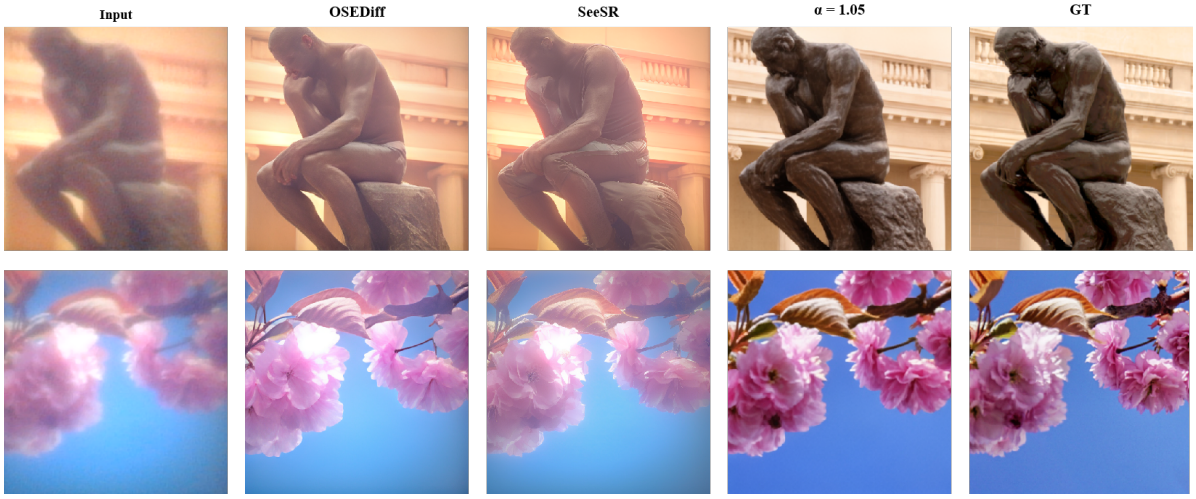

Figure 6. Comparison with pretrained diffusion models

6. More Ablation Results

In Fig. 7, we show more instantly tunable decoding results. In Fig. 8, we show more degradation learning qualitative results

7. More Interactive Results

Please visit the anonymous link https://dmdiff.github.io/.

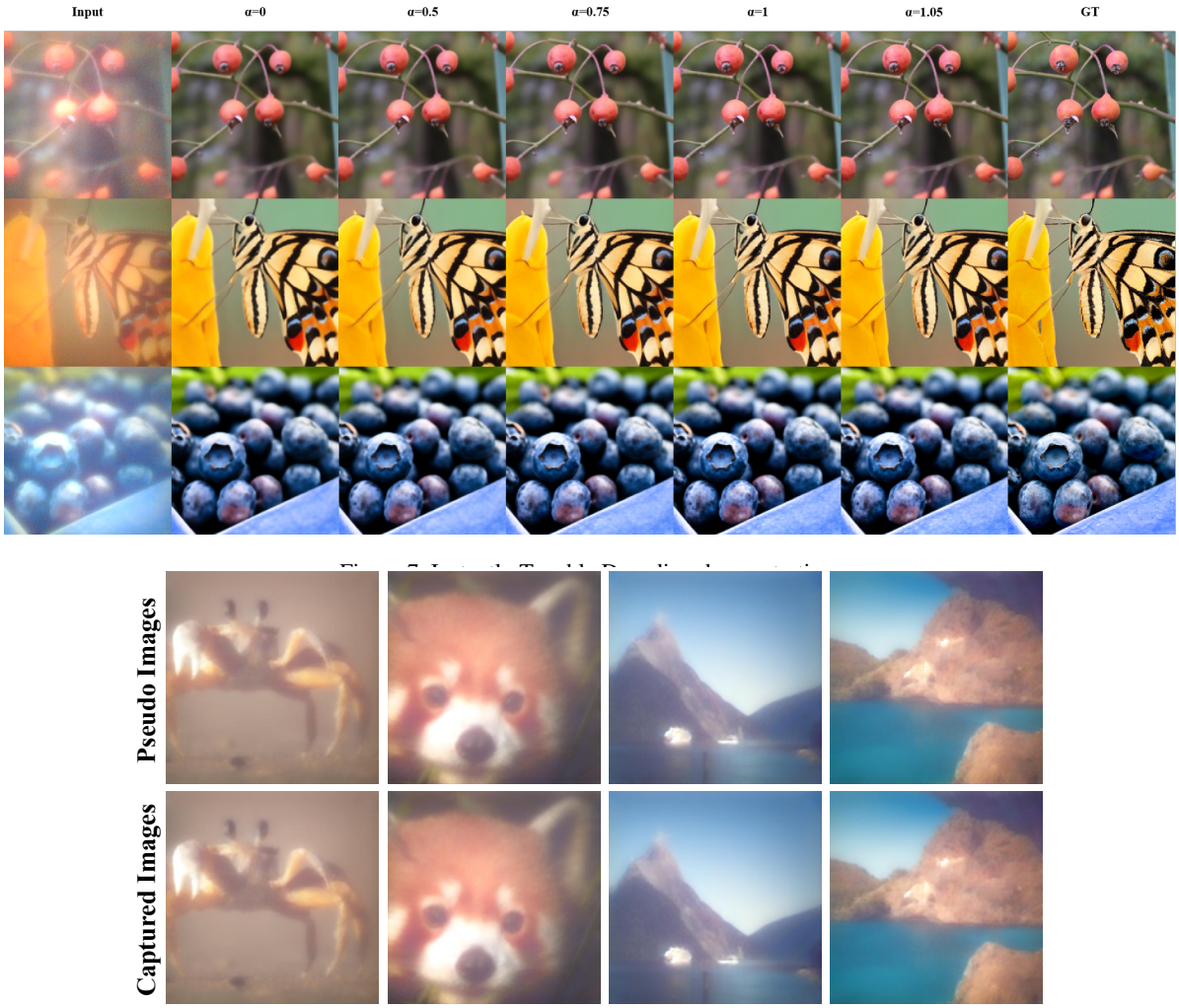

Figure 8. Degradation learning qualitative results. Our method effectively simulates the imaging effects of metalenses.

References

[1] Ethan Tseng, Shane Colburn, James Whitehead, Luocheng Huang, Seung-Hwan Baek, Arka Majumdar, and Felix Heide. Neural nano-optics for high-quality thin lens imaging. *Nature communications*, 12(1):6493, 2021. 1