# Supplementary Material : Deep-learning-driven end-to-end metalens imaging

Joonhyuk Seo[1,†], Jaegang Jo[2,†], Joohoon Kim[3,†], Joonho Kang[4,†], Chanik Kang[1], Seongwon Moon[3], Eunji Lee[5], Jehyeong Hong[1,2,4], Junsuk Rho[3,5,6,7,8,*,1], and Haejun Chung[1,2,4,*,2]

[1]Department of Artificial Intelligence, Hanyang University, Seoul, 04763, Republic of Korea
[2]Department of Electronic Engineering, Hanyang University, Seoul, 04763, Republic of Korea
[3]Department of Mechanical Engineering, Pohang University of Science and Technology (POSTECH), Pohang, 37673, Republic of Korea
[4]Department of Artificial Intelligence Semiconductor Engineering, Hanyang University, Seoul, 04763, Republic of Korea
[5]Department of Chemical Engineering, Pohang University of Science and Technology (POSTECH), Pohang, 37673, Republic of Korea
[6]Department of Electrical Engineering, Pohang University of Science and Technology (POSTECH), Pohang, 37673, Republic of Korea
[7]POSCO-POSTECH-RIST Convergence Research Center for Flat Optics and Metaphotonics, Pohang, 37673, Republic of Korea
[8]National Institute of Nanomaterials Technology (NINT), Pohang, 37673, Republic of Korea
[†]These authors contributed equally to this work.
[*]These authors are corresponding authors.

October 8, 2024

## Contents

# 1 Metalens fabrication

The metalens developed in this work comprises a two-dimensional rectangular array of meta-atoms, each featuring a nano-slab and an arbitrary rotation angle, forming a Pancharatnam-Berry phase-based metalens. The dimensions of the nano-slabs, which are 70 nm wide, 380 nm long, and 900 nm high, were established following the deposition of a 23 nm thick titanium dioxide thin film on the imprinted resin structures. The design parameters, including the width, length, height of the slab, and thickness of the titanium dioxide thin film, were systematically optimized to maximize the focusing efficiency. The metalens focuses light with the polarization-dependent transmittance of the meta-atom, the Jones matrix of which can be expressed as:

$$\mathbf{J} = \begin{bmatrix} t_l & 0 \\ 0 & t_t \end{bmatrix} \tag{1}$$

where $t_l$ and $t_t$ are the complex transmission coefficients of the meta-atoms when the electric field polarization is aligned along the longitudinal or transverse directions of the slab, respectively.

The Jones matrix for the rotation angle $\theta$ of the nano-slab is

$$\mathbf{T} = \mathbf{R}(-\theta)\mathbf{J}\mathbf{R}(+\theta) = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} t_l & 0 \\ 0 & t_t \end{bmatrix} \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \tag{2}$$

where $\mathbf{R}$ is the rotation matrix of the rotation angle $\theta$, and $\mathbf{T}$ is the transfer matrix of the meta-atom. For incident Left-handed Circularly Polarized (LCP) light, the transmitted light can be derived as a linear combination of LCP and Right-handed Circularly Polarized (RCP) light:

$$\mathbf{T}\frac{1}{\sqrt{2}}\begin{bmatrix} 1 \\ i \end{bmatrix} = \frac{t_l + t_t}{2}\begin{bmatrix} 1 \\ i \end{bmatrix} + \frac{t_l - t_t}{2}e^{i2\theta}\begin{bmatrix} 1 \\ -i \end{bmatrix} \tag{3}$$

The phase retardation of the RCP light can be precisely controlled by $\theta$. The ideal phase distribution ($\phi_{\text{Ideal}}$) at the exit plane of the metalens is defined as

$$\phi_{\text{Ideal}}(x, y) = -\frac{2\pi}{\lambda}\left(\sqrt{x^2 + y^2 + f^2} - f\right) \tag{4}$$

where $\lambda$ is the target wavelength of 532 nm; $x$ and $y$ are the spatial coordinates on the metalens; and $f$ is the focal length, which was set to 24.5 mm for a numerical aperture of 0.2. Therefore, the orientation angle of each meta-atom at a point $(x, y)$ is $(\phi_{\text{Ideal}}/2)$.

To mass-produce the designed metalens on a wafer scale, we sequentially applied high-speed electron-beam lithography, ArF immersion scanning, nanoimprint lithography, and atomic layer deposition. The high-speed electron-beam lithography process (JBX Series, JEOL) was instrumental in patterning the photomasks of the metalenses. At this stage, the photomask contained a mask pattern for a singular metalens, which was insufficient for mass production. To rectify this, we transferred the photomask pattern onto a 12″ Si wafer in an array format using an ArF immersion scanner (XT-1900Gi, ASML), thereby creating a master stamp with metalens arrays.

Following this, we employed nanoimprint lithography to replicate the fabricated 12″ master stamp at a significantly reduced cost. In the initial phase of nanoimprint lithography, we coated the prepared master stamp with hard-polydimethylsiloxane (h-PDMS) to achieve a high-resolution replication of the metalens, and subsequently coated it with PDMS to create a replica mold. After baking the replica mold at 80°C for two hours, we separated the cured replica mold from the master stamp.

We then applied a conventional imprint resin (MINS-311RM) onto the replica mold and covered it with an 4″ glass wafer. Following the curing of the imprint resin under a pressure of 2 bar and UV-light irradiation, we formed a metalens pattern on the glass wafer by detaching the replica mold. To enhance the effective refractive index and thereby increase the efficiency further, we thinly coated the imprinted metalens with a high-index titanium dioxide film using atomic layer deposition.

## 2 Metalens imaging system and data acquisition setup

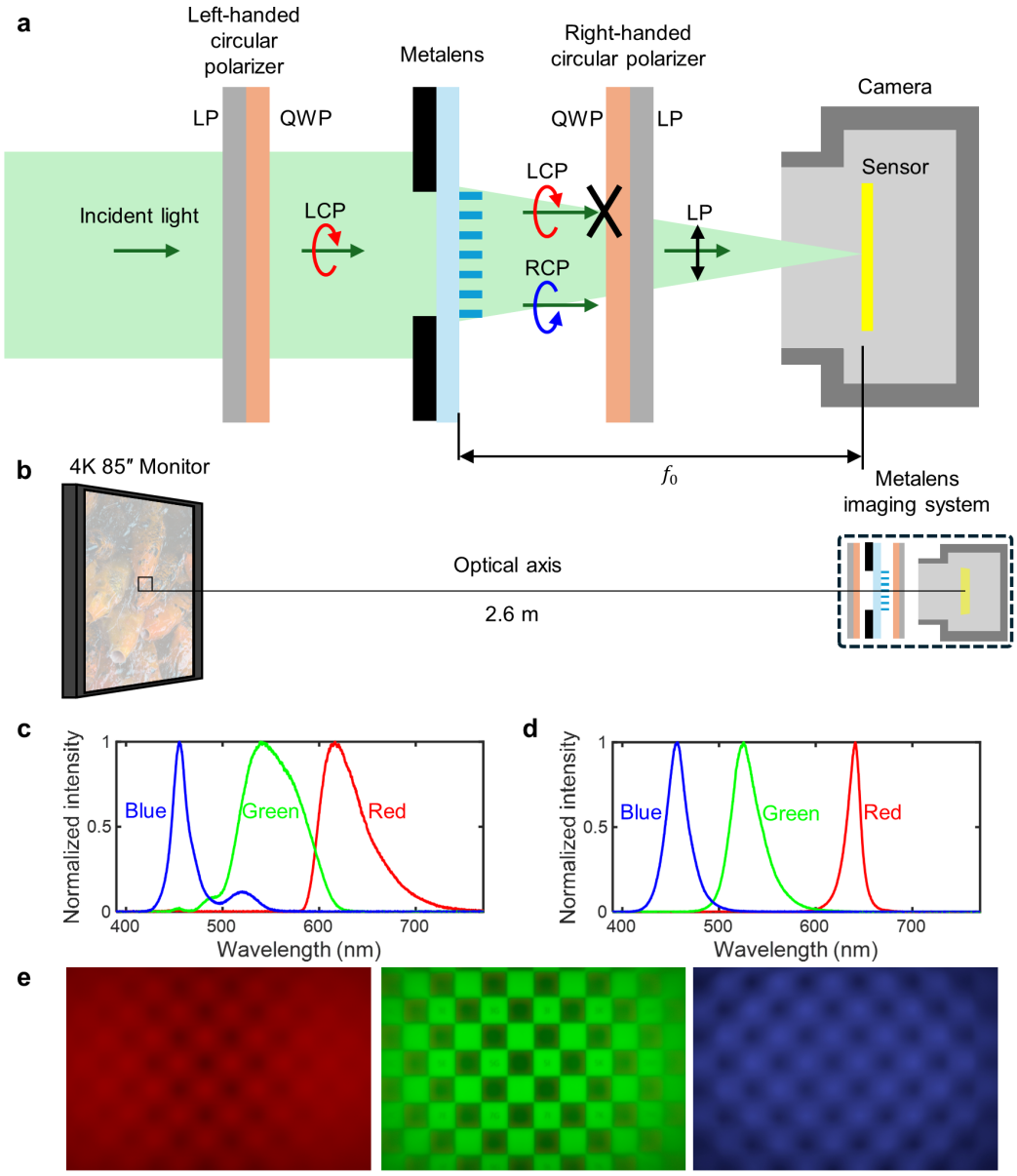


**Figure S1**: (a) Schematics of the metalens imaging system. The metalens imaging system consists of the metalens, the camera, and left- and right-handed circular polarizers. (b) Schematics of data acquisition setup. (c) Spectra of the red (616 nm), green (541 nm), and blue (455 nm) pixels of the 85″ monitor. (d) Spectra of the red (peaks at 640 nm), green (525 nm), and blue (457 nm) light-emitting diodes (LEDs) used for the PSF measurements. (e) Metalens images obtained by capturing the red, green, and blue mosaic patterns on the monitor.

The metalens imaging system incorporates our mass-produced metalens, a commercial camera (Basler acA5472-17uc) with an exposed imaging sensor and two left- and right-handed circular polarizers (Edmund Optics CP42HE and CP42HER). In addition, the polarizers can be expressed as stacks of a Linear Polarizer (LP) and a Quarter Wave Plate (QWP) because they are fabricated by laminating an XP42 LP sheet to a WP140HE QWP sheet. The roles of each optical component are shown in Fig. S1(a). The left-handed circular polarizer only transmits the LCP light to the metalens. As described in the metalens fabrication section, the RCP light transmitted from the metalens has the wavefront propagating to the focal point, while the LCP light has the same wavefront as the incident

light. The right-handed circular polarizer transmits only RCP light and blocks LCP light to remove unfocused lights. Finally, the linearly polarized light transmitted from the right-handed circular polarizer focuses on the image sensor.

The planes of the circular polarizers and image sensor are aligned perpendicular with the optical axis of the metalens, which point towards the center of the image sensor. Figure S1(b) shows the data acquisition setup. The metalens imaging system is distant 2.6 m from the 4K 85″ monitor (Samsung KU85UA7050F) with the optical axis perpendicularly pointing toward the center of the monitor. In addition, the distance between the metalens and the image sensor $f_0$ is tuned to focus the center of the green pattern on the monitor as shown in Fig. S1(e). The spectra of the red, green, and blue pixels are shown in Fig. S1(c), which have peak wavelengths at 616, 541, and 455 nm, respectively. The raw metalens images were obtained by capturing the ground truth images displayed on the monitor.

# 3 PSF measurement setup and MTF calculation

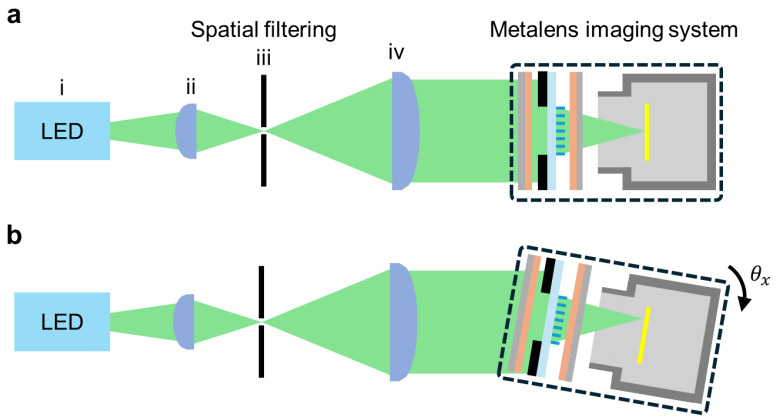


**Figure S2**: (a) Optical setup for PSF measurement with zero viewing angle. (b) PSF measurement setup with horizontal viewing angle $\theta_x$.

The PSF was measured by capturing images of the collimated lights using the metalens imaging system. As described in Fig. S2(a), the red, green, or blue light from the LED with the peak intensities at 640, 525, 457 nm (i) is spatially filtered and collimated by the 0.5″ aspherical lens (ii, Thorlabs AC127-019-A), 20 µm pinhole (iii, Thorlabs P20CB), and 1″ spherical lens (iv, LA1461-A). Before capturing the PSFs, the focal length $f_0$ was set to focus the collimated light from the green LED, and the spectra of the LEDs are shown in Fig. S1(e). The PSFs with zero viewing angle were measured by matching the optical axes of the spatial filter and the metalens imaging system. The PSFs with non-zero viewing angles were measured by tilting the metalens imaging system in the horizontal direction as shown in Fig. S2(b).

The MTFs are subsequently calculated using measured PSFs as [41],

$$\mathrm{MTF} \equiv \left| \frac{\int\int I(x,y) \exp\left[-2\pi i (f_x x + f_y y)\right] dxdy}{\int\int I(x,y) dxdy} \right| \tag{5}$$

where $x$ and $y$ are the horizontal and vertical positions on the image sensor; $I(x,y)$ is the PSF; and $f_x$ and $f_y$ are the spatial-frequencies along the $x$ and $y$ axes, respectively. Defining the viewing angle as the angle between the optical axis of the metalens and the line towards the point light source from the center of the metalens, Fig. 2(d) describes the MTFs depending on $f_x$ with the zero $f_y$.

In Fig. 2(f), the metalens image's blue channel is more blurry than the red channel, while the MTF of the blue channel is higher than the red channel in Fig. 2(d). The peak wavelength of the green light of the monitor is closer to 541 nm rather than 525 nm. Furthermore, the distance between the metalens and the monitor is 2.6 m. Thus, the focal length is set to focus 541 nm green light from a point 2.6 m away along the optical axis of the metalens. Additionally, the intensity distribution of the red light from the monitor has a peak around the 620 nm wavelength. As a result, the red channel

experienced less defocusing compared to the blue channel, which had more significant defocusing due to its spectral characteristics and the focal point adjustment.

# 4 Statistical train-test inconsistency

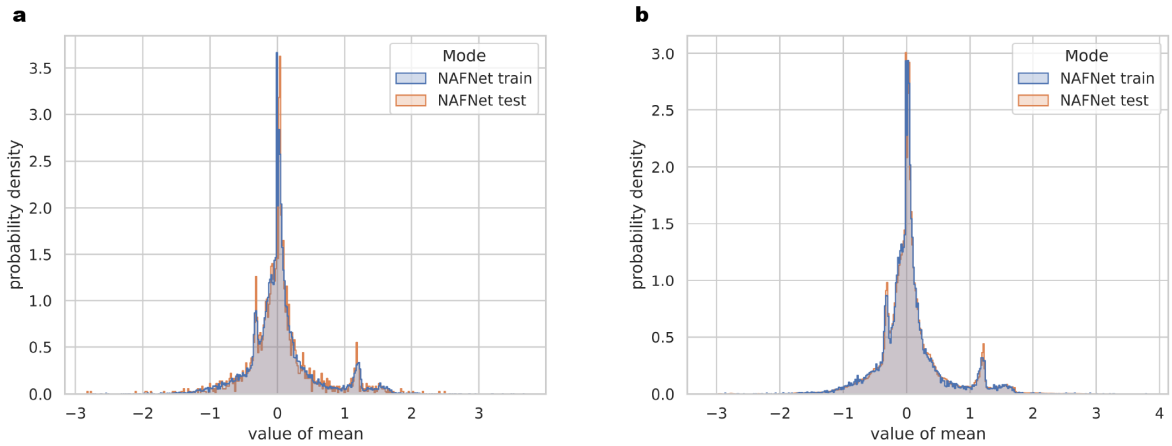


**Figure S3**: **Statistical train-test inconsistency**. (a-b) Probability density functions of global average pooled features of training(red) and testing(blue) (a) without and (b) with TLC.

| | TLC | PSNR | SSIM | LPIPS |
|---|---|---|---|---|
| HINet | | 19.378 | 0.609 | 0.493 |
| | ✓ | 21.364 | 0.641 | 0.456 |
| NAFNet | | 18.447 | 0.559 | 0.519 |
| | ✓ | 21.689 | 0.642 | 0.440 |
| Our framework | | 20.868 | 0.641 | 0.448 |
| | ✓ | 22.095 | 0.656 | 0.432 |

**Table S1**: Comparison of quantitative results from various models without and with TLC.

| | AP | $AP_{50}$ | $AP_{75}$ |
|---|---|---|---|
| Ground truth | 0.451 | 0.730 | 0.482 |
| Metalens image | 0.051 | 0.087 | 0.054 |
| Our framework | 0.386 | 0.646 | 0.397 |

**Table S2**: Comparison of quantitative results of object detection on the PASCAL VOC2007 between ground truth images, metalens images, and restored images.

In natural image restoration studies based on deep neural networks (DNNs) [57, 58, 66], the channel attention module [67] leads to the exceptional performance of the restoration. Global average pooling (GAP) shrinks the given features in spatial dimensions to generate channel-wise statistics. However, GAP significantly degraded the train-test consistency because we used randomly cropped image patches as training data and full-resolution images (non-cropped) as test data, as mentioned in the main text. Figure S3 shows this statistical inconsistency at the training and test phases of NAFNet (baseline model). Specifically, Figures S3(a) and (b) show the probability density functions of the outputs of the GAP of the first channel attention module at the second encoder of the model for the metalens images.

The test-time local converter (TLC) [53] can significantly contribute to overcoming these issues. After applying TLC, the statistics in the test phase are similar to those in the training phase (Fig. S3(b)). Furthermore, TLC ensures this statistical consistency and significantly improves the image quality (Table S2).

# 5 Pattern artifact resulting from adversarial learning scheme in RGB space

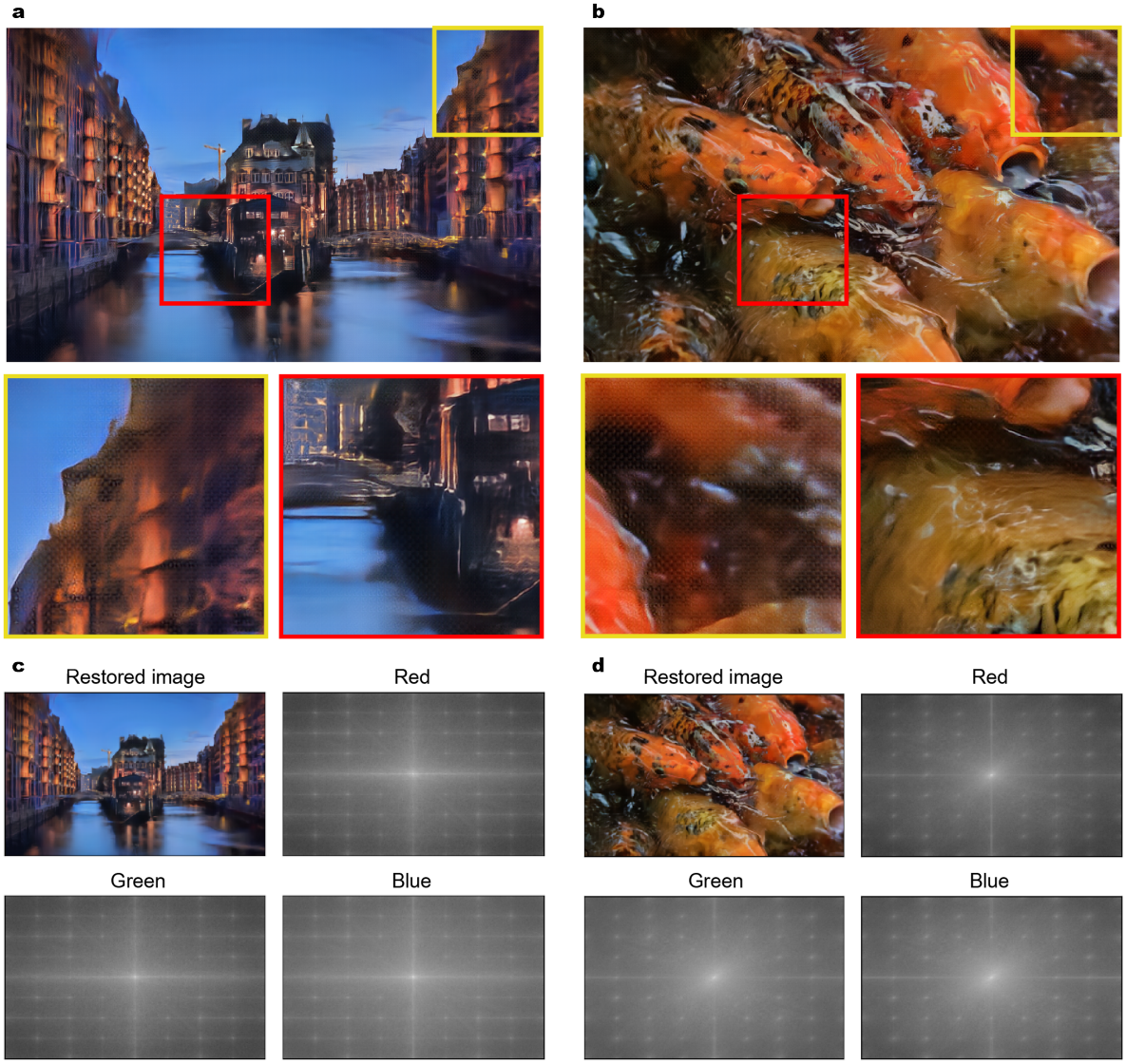


**Figure S4**: Pattern artifacts resulting from the adversarial learning scheme in RGB space. (a-b) Results of image restoration on the test set using adversarial learning regularization in RGB space. (c-d) Visualization of (a-b) in the frequency domain.

As mentioned in the main text, the implementation of an adversarial learning scheme in the RGB space for regularization results in restored images that exhibit pattern artifacts in the RGB space across the entire area (Fig. S4(a) and (b)) and the frequency domain (Fig. S4(c) and (d)). In particular, applying adversarial learning in the RGB space significantly improved LPIPS but drastically reduced PSNR and SSIM (PSNR : 21.48 versus 21.69; SSIM : 0.64 versus 0.68; LPIPS : 0.39 versus 0.44).

# 6 Details of Architectures

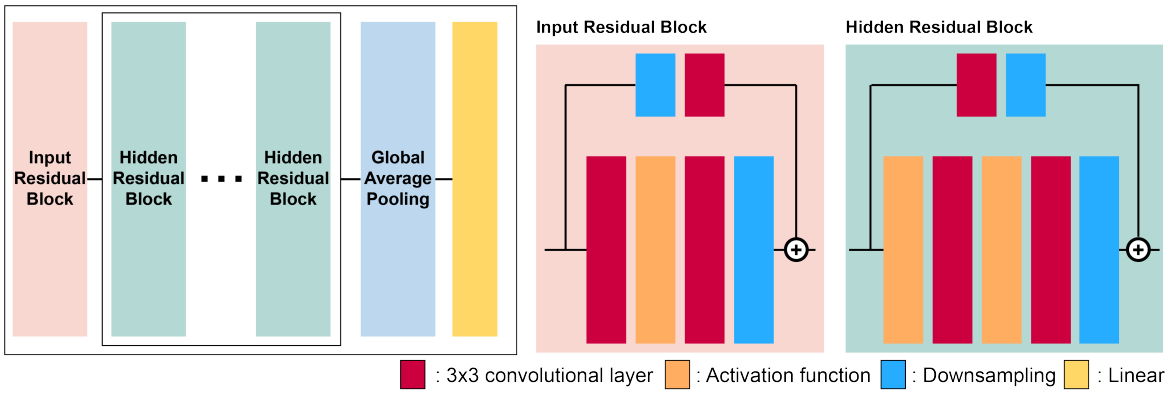


**Figure S5**: Illustration of the discriminator in our study.

| | level 1 | level 2 | level 3 | level 4 |
|---|---|---|---|---|
| Encoder | 2 | 2 | 4 | 8 |
| Decoder | 2 | 2 | 2 | 2 |
| Bottleneck | | | | 12 |

**Table S3**: The number of blocks for each level of the proposed restoration architecture.

The proposed framework consists of a restoration model responsible for practical image restoration and a discriminator for adversarial learning. We employed [38] for the restoration model, which has an encoder-decoder structure. The encoder and decoder of the network are each composed of four levels. Table S3 shows each level's number of blocks in the encoder, decoder, and bottleneck. Additionally, as shown in Fig S5, our discriminator consists of an input residual block and hidden residual blocks, with the number of hidden residual blocks set to 5 in our study.

# 7 Details of Performance Metrics

## 7.1 PSNR and SSIM

We use PSNR and SSIM, representative image quality metrics, to evaluate the images reconstructed by the proposed model. The PSNR is :

$$\mathrm{PSNR}(x, \hat{x}) = 10 \log_{10}(\frac{R^2}{\mathrm{MSE}(x, \hat{x})}). \tag{6}$$

Since MSE is a metric indicating the difference of pixel values, higher PSNR values indicate a better restoration performance. $R$ is the maximum signal value of the ground truth image. However, PSNR is only an approximation to the human visual perception of the reconstruction quality because MSE is the pixel-level intensity difference between the reconstructed and ground truth images. Therefore, we use SSIM to accurately evaluate the visually perceived quality of the restored images. SSIM utilizes three elements in its evaluation: luminance, contrast, and structure between the reconstructed and ground truth images. The SSIM is :

$$\mathrm{SSIM}(\hat{x}, x) = \frac{(2\mu_{\hat{x}}\mu_x + c_1)(\sigma_{\hat{x}x} + c_2)}{(\mu_{\hat{x}}^2 + \mu_x^2 + c_1)(\sigma_{\hat{x}}^2 + \sigma_x^2 + c_2)} \tag{7}$$

In this equation, $\mu_x$, $\sigma_x^2$ and $\sigma_{xy}^2$ indicate mean of $x$, variance of $x$ and covariance of $\hat{x}$, $x$, respectively.

## 7.2 Assessment in Spatial Frequency Domain

We also introduce metrics to evaluate the restoration of lost spatial frequency information due to intense degradation in the Fourier domain. Spatial frequency information is entirely characterized by the magnitude and phase angle. Thus, we employ magnitudes' MAE and cosine similarity as metrics in the spatial frequency domain. First, we transform the given ground truth and restored images from RGB space into Fourier space. Then, we derive the MAE and cosine similarity between the magnitudes of the transformed data. The MAE and cosine similarity are:

$$\mathrm{MAE}_{\mathcal{F}}(\hat{x}, x) = \frac{1}{N}\sum_{n=1}^{N} ||\mathcal{F}(\hat{x})| - |\mathcal{F}(x)|| \tag{8}$$

$$CS_{\mathcal{F}}(\hat{x}, x) = \frac{1}{N}\sum_{n=1}^{N} \frac{\mathcal{F}(\hat{x}_n) \cdot \mathcal{F}(x_n)}{||\mathcal{F}(\hat{x}_n)||_2 ||\mathcal{F}(x_n)||_2} \tag{9}$$

In this equation, $\mathcal{F}$ is Fourier transformation.

# 8 Additional Experiments for Restored Image Quality

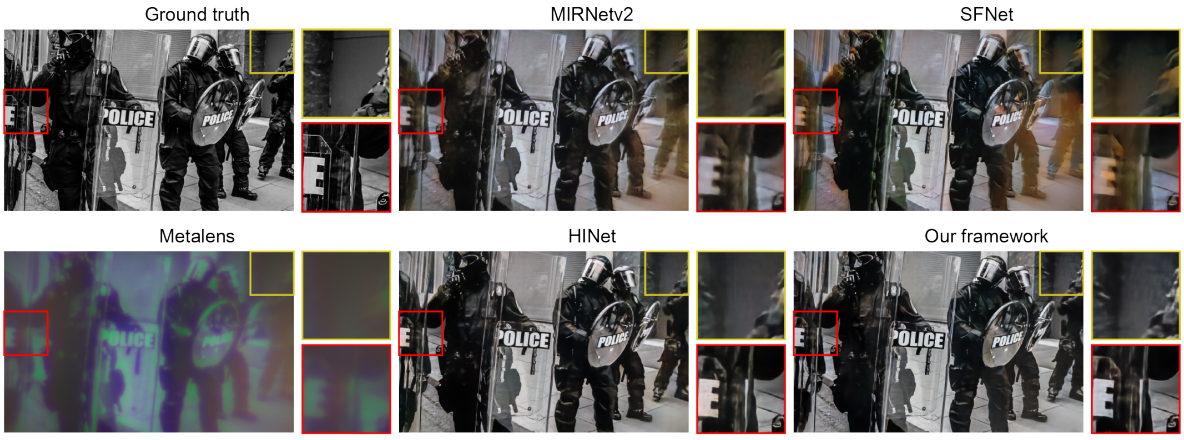


**Figure S6**: Qualitative comparison between various methods including our integrated imaging system.

| | Red | | | Green | | | Blue | | |
|---|---|---|---|---|---|---|---|---|---|
| **Model** | PSNR | SSIM | LPIPS | PSNR | SSIM | LPIPS | PSNR | SSIM | LPIPS |
| Metalens image | 13.476 | 0.405 | 0.804 | 16.083 | 0.472 | 0.732 | 15.257 | 0.416 | 0.786 |
| MIRNetv2 | 17.269 | 0.535 | 0.526 | 19.684 | 0.586 | 0.489 | 19.267 | 0.548 | 0.504 |
| SFNet | 16.963 | 0.545 | 0.497 | 19.345 | 0.596 | 0.461 | 18.982 | 0.560 | 0.475 |
| HINet | 20.197 | 0.624 | 0.439 | 22.887 | 0.674 | 0.408 | 21.935 | 0.625 | 0.426 |
| NAFNet | 20.690 | 0.626 | 0.433 | 23.146 | 0.677 | 0.4 | 22.031 | 0.623 | 0.418 |
| Our framework | **20.93** | **0.639** | **0.423** | **23.831** | **0.692** | **0.387** | **22.481** | **0.636** | **0.407** |

**Table S4**: Comparison of channel-wise quantitative assessments of various models.

| | Normal incidence | | 10.3°-13.1° | | 13.1°-15.9° | |
|---|---|---|---|---|---|---|
| **Model** | PSNR | SSIM | PSNR | SSIM | PSNR | SSIM |
| Metalens image | 16.124 | 0.492 | 15.317 | 0.376 | 14.868 | 0.534 |
| MIRNetv2 | 18.834 | 0.527 | 21.578 | 0.638 | 16.139 | 0.545 |
| SFNet | 18.389 | 0.534 | 21.459 | 0.652 | 15.015 | 0.534 |
| HINet | 20.841 | 0.63 | **23.317** | **0.684** | 19.724 | 0.597 |
| NAFNet | 22.748 | 0.638 | 23.183 | 0.679 | 19.824 | 0.606 |
| Our framework | **23.547** | **0.662** | 23.078 | 0.682 | **22.554** | **0.631** |

**Table S5**: Comparison of quantitative assessments of various image restoration models using upper-right regions (100 × 100 pixels) corresponding to the specified incidence angles (10.3° - 13.1° and 13.1° - 15.9°). Specifically, spatially dependent degradation is the most severe at angle range 13.1° - 15.9°. Additionally, we evaluate the center region (100×100 pixels) for assessment of normal incidence cases.

We conducted additional experiments to thoroughly analyze the performance of the proposed framework. First, we performed a qualitative comparison between the metalens images and the results restored by various baseline models, including our framework. As illustrated in Fig. S6, our framework more effectively captures fine details and produces more accurate color reproduction compared to other models. In particular, as shown in Table S4, our method consistently outperforms others

across all RGB channels, each corresponding to different wavelength regions. Moreover, both Fig. S6 and Table S5 demonstrate that our framework exhibits significantly superior performance, particularly in the outer regions of the image, which suffer from severe degradation due to oblique incidence. Notably, our framework yielded substantial improvements in image quality within the upper-right outer region (13.1°-15.9°), with an increase in PSNR by 7.7 dB, an enhancement in SSIM by 9.7%p, and a reduction in LPIPS by 20%p compared to the original metalens images. Furthermore, in comparison to NAFNet, which serves as the baseline for our framework, the proposed method achieved a 2.7 dB increase in PSNR, a 2.5%p improvement in SSIM. Although our framework exhibits marginally inferior performance at the specific incidence angle (10.3°-13.1°) than HINet [58], we highlight that the proposed framework maintains consistent performance across varying incidence angles.

# 9 Outdoor Image Restoration

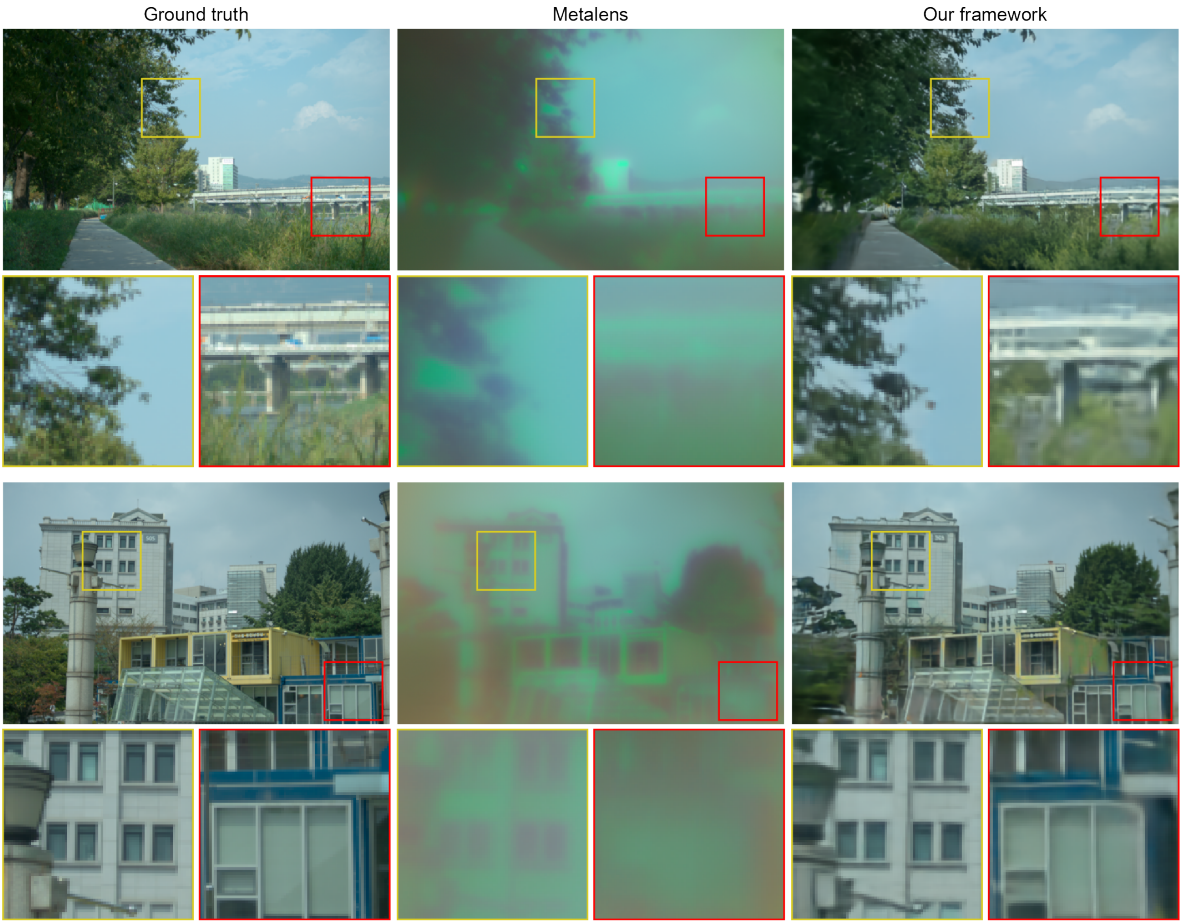


**Figure S7**: Ground truth outdoor images, metalens images, and images restored by our model. The images are affiliated with the test set data. The red and yellow boxes indicate local regions having sharp edges, which demonstrate the restoration quality in high frequency.

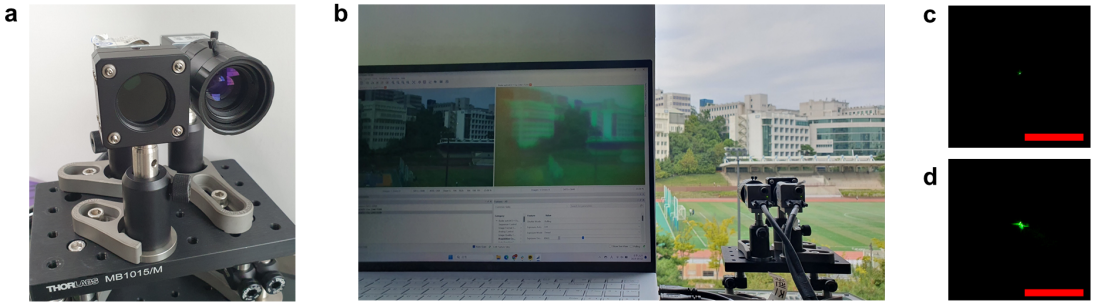


**Figure S8**: (a) The metalens imaging system is at the left while the conventional imaging system is at the right, looking at the whole set. (b) Instance of data collection for outdoor scenes. (c-d) PSFs of the conventional and metalens imaging systems, respectively. The scalebar is 500 µm.

We conducted training and inference on newly collected outdoor images to verify our framework's learning capability. First, we composed a conventional imaging system using a commercial bulky lens system (HP Techspec 25mm fixed focal length lens) and a camera (Basler acA5472-17uc). Then, we positioned the conventional and metalens imaging systems in parallel, with their optical axes aligned and approximately 4 cm apart, as shown in Fig. S8(a). The focal lengths of the imaging systems are set to focus on a collimated 538 nm laser beam; the collimated beam is obtained by the PSF measurement setup in Fig. S3(a) where the LED is replaced with a 538 nm diode laser. The PSFs of the imaging systems are shown in Fig. S8(c) and (d).

We acquired 109 pairs of images by simultaneously capturing outdoor scenes using the two imaging systems as shown in Fig. S8(b). We obtained ground truth images by cropping the raw images from the conventional imaging system with 5472×3648 resolution to 5472×3420 resolution and resizing them to 1280×800 resolution. Then, we gained corresponding metalens images by rotating the raw images, cropping them to 5118×3228 resolution, and resizing them to 1280×800 resolution. The parameters for rotation and cropping for the metalens images are optimized to maximize the SSIM between the metalens and ground truth images. We constructed a dataset with 98 training and 11 test samples from the 105 pairs of ground truth and metalens images and trained our framework.

Figure S7 shows the metalens, ground truth, and restored images. Our framework exhibits significant restoration quality for the outdoor images. However, the outer regions of the restored images exhibit blurs, color distortions, and incorrect edge details, showing lower restoration quality than the restored image from the monitor. We expect that the restoration quality of the outdoor images can be significantly enhanced by solid alignment between two imaging systems (conventional and metalens) and also by capturing diverse and numerous outdoor scenes.

The relatively low restoration quality of the outdoor images may be attributed to the limitations of the training dataset. The accuracy of the training dataset may be diminished due to the slight changes in the optical alignment during the imaging system's movement between shots. Furthermore, the insufficient diversity and quantity of the dataset may further reduce the restoration quality.