Bifocal Neural Metalens for Broadband Depth Estimation

Fenghao Xu, Johan Carlstrom, and JP Berenguer

Abstract—Conventional monocular depth estimation techniques rely on depth cues and impose stringent conditions for accurate depth prediction. Recently, a growing body of literature on neural network architectures have demonstrated the possibility of extracting depth information from contextual, pictorial cues in images. However, on the hardware side, most approaches still employ bulky optical elements, which are not suited for the next-generation of miniature devices. Advances in nanophotonics may bridge this gap with the advent of metasurfaces, a flat 2D array of meta-atoms with an arbitrarily designed phase profile. Inspired by works on bifocal metalenses and deep optics for the joint optmization of the latter with neural networks, we propose an end-to-end fully differentiable imaging pipeline that robustly estimates absolute depth over a range of 10-60mm.

Index Terms—Computational Optics, Deep Optics, Metalens

1 Introduction

Reliable depth estimation is a central problem in computer vision with applications ranging from 3D scene understanding (e.g. in autonomous driving) to robotics [1]. Depth perception from a single image without priors is generally an ill-posed problem, since a 2D image can represent multiple 3D scenes in the real world [2]. More rigorously, the sensor image is given by the convolution of the in-focus image with the imaging system's point spread function (PSF), which for a single scene without extra knowledge are both unknowns.

From the several strategies that have been implemented to tackle this problem, solutions that consist of monocular configurations are more broadly desirable for their efficiency (i.e. single vs multi-shot images) and lack of constraints on the optics such as moving sensors/apertures or multiple lenses. Historically, such constraints were necessary in order to infer depth from defocus; however, with the advent of convolutional neural networks (CNNs), algorithms attempt to extract depth information from contextual pictorial cues, such as texture, occlusions, and learned object sizes [1], [2].

A breakthrough in the field of nanophotonics has been the invention of metasurfaces. These are flat 2D arrays of meta-atoms, made of dielectric or metallic materials, whose individual amplitudes and phases are carefully engineered so as to scatter incident light in a specific way [3]. For instance, in order to construct a metalens, one can perform electromagnetic full-field simulations of a unit cell consisting of a lossless nanopillar on a substrate and obtain its scattered phase and amplitude to a given incident field. Then, by sweeping through a geometric parameter, e.g. the nanopillar diameter, a structure-to-phase "library" can be built which spans 0-2π at constant amplitude. Finally, the pixel-by-pixel distribution is done by mapping the desirable (parabolic) phase profile to the corresponding meta-atom [3].

Metalenses have the potential of replacing bulky, multi-element optical systems with flat and ultra-small nanostructures in the next-generation devices. However, a primary obstacle to its widespread adoption has been the inherent presence of chromatic aberrations, resulting in impractically low numerical apertures (NA) [4]. Thus, it is plausible that leveraging the power of neural networks and incorporating it to the entire imaging pipeline may significantly outperform metalenses that solely rely on physical intuition to address such aberrations.

In this work, we combine efforts from the nanophotonics literature, where a metalens was designed to infer depth from defocus [5], with a novel deep optics algorithm that jointly optimizes the metasurface parameters and an image deconvolution neural network in an end-to-end imaging pipeline [4]. We show that robust, absolute depth estimates can be obtained from single images by simultaneously training a broadband metasurface composed of two interviewing metalenses and a U-Net style network architecture [1], [6].

2 Related Work

2.1 Depth sensors inspired by the jumping spider

In the natural world, the ability to accurately infer the depth of a scene has had obvious adaptive advantages over eons for both predator and prey. It is tempting, therefore, to examine how the optical systems of creatures with very limited brain computing power have evolved to maximize depth cues by the propagation of incident fields on them. One such example is the jumping spider (Salticidae), whose stacked, multilayered retinae effectively form two images of a single scene with different degrees of defocus. These images can then be analyzed to gauge depth by leveraging the defocus cues present within them. Interestingly, the photosensors on these retinae are primarily sensitive to green light, but also slightly responsive to red wavelengths, such that when the spiders hunt under red light in an artificial environment, they miss their prey by the exact

- *J. Carlstrom and JP. Berenguer are with the Department of Materials Science and Engineering, Stanford University.*
- *F. Xu is with the Department of Applied Physics, Stanford University*

same distance predicted by the chromatic aberration in their corneal lenses [7], [8].

Guo et al. have demonstrated a metasurface that mimics the jumping spider optical system in order to obtain relative and absolute depth estimations of a single scene [5]. The metasurface consists of an interviewing phase profile of two individual, off-axis metalenses with different focal lengths, such that an incident monochromatic field focuses on the sensor at two different locations with different degrees of defocus. The proposed metasurface possesses a 3mm diameter and is able to reliably extract absolute depth information from a scene over a 10cm range, though only for a target green wavelength (λ = 532nm).

2.2 Neural Nano-Optics

The field of deep optics studies the implementation of algorithms that jointly optimize hardware (physical layer) and software (digital layer) for better performance at specific tasks. The optical layers can be interpreted as encoders of information present in the incident light field and the neural network layers as the decoders of such information. Thus, the parameters that define the optical system must be fully differentiable in order be fed into an end-to-end pipeline that allows forward propagation, loss calculation, and backpropagation with updated parameters.

Tseng et al. [4] designed such an end-to-end network, where the optical layer was a metalens with a fully differentiable phase profile and the digital layer was a convolutional neural network. The novel insight by the authors was to fit a polynomial – thus differentiable – function to the phase library of a metasurface, which allows it to be integrated to the training algorithm and hence jointly optimized with the neural network. The construction of a wavelength and angle-dependent set of PSFs carried the necessary information on the dispersion of the meta-atoms in order for the learning steps to minimize the chromatic aberration and maximize the system's field-of-view. The authors propose a figure of merit, the Diffractive Lens Achromatic Capacity (DLAC), defined as the product of the Fresnel Number and the fractional bandwidth, which captures the difficulty of meta-optics designs in achieving both broadband and high NA metalenses. It was found that their end-to-end architecture outperform dispersion engineered achromatic metalenses by orders of magnitude.

3 Proposed Method

3.1 Metasurface parameterization

The electric field phase-front impinging the metasurface at position (x,y), emitted from an object on the optical axis a distance z away from the lens is,

$$\phi(x, y, z, \lambda) = \frac{i\pi}{\lambda z}(x^2 + y^2) \tag{1}$$

Importantly, the curvature of the phase-front at the metasurface is inversely proportional to the distance z. As such, the fields are uniquely diffracted by the metasurface, leading to a depth-dependent point spread function.

In order to construct a fully differentiable network, the optical response of the metasurface has to be parameterized. This is accomplished with the following steps: 1) Find a differentiable and injective phase-to-structure response at a nominal wavelength, using effective medium RCWA calculations. 2) Find the structure-to-phase response of other wavelengths by an expansion in wavelength and the phase-to-structure response of the nominal wavelength. 3) Parameterize the spatial phase distribution of the metasurface in a differentiable form (e.g., polynomial).

For the first step, only the first order Fourier term is considered to fit the calculated transmission phase vs. duty cycle relationship. The duty cycle vs phase response is fitted with a polynomial on the form

$$d = \sum_{i=0}^{N} b_i \left(\frac{\phi}{2\pi}\right)^{2i} \tag{2}$$

where d is the duty cycle and b_i are fitting coefficients. Once this relationship is established, the phase response for wavelengths other than the nominal wavelength are parameterized by fitting the their RCWA calculated phase response to a proxy function on the form,

$$\phi(\lambda) = \sum_{n=0}^{N} \sum_{m=0}^{M} c_{nm} d^n \lambda^m \tag{3}$$

where λ are non-nominal wavelengths, d is the duty cycle response from eq 2 and c_{mn} are fitting coefficients. The set of coefficients $\{b_i\}$ and $\{c_{nm}\}$ are fixed parameters of the metasurface and are used to determine how incident fields are diffracted for a given duty cycle distribution and wavelength.

The only things remaining is to parameterize the spatial phase profile across the metasurface. This is achieved with a circularly symmetric polynomial expansion on the form

$$\phi(r) = \sum_{i=0}^{n} a_i \left(\frac{r}{R}\right)^{2i} \tag{4}$$

where r is the distance from the optical axis, R is the size of the metalens and a_i are fitting coefficients. These coefficients are the free parameters for the optical response of the metasurface. For this project, we used pre-fitted values for all the above-mentioned coefficients [4]. One limitation of this approach is the challenge of fitting high spatial frequencies of the phase to a polynomial function, which limits the type of phase profiles can be accurately represented. For instance, the two interleaved metalenses with displaced foci inspired by the jumping spider [5] have substantial high frequency components close to the metasurface edge. There may be better bases functions to represent such phase profiles, but in this work we simplify by considering the two phase profiles independently.

3.2 Point Spread Functions and Image Convolution

First, a set of point spread functions (PSFs) are calculated for a range of z-values by propagating the electric field distribution to the sensor after passing through the metsurface. To simulate a true captured image from our optical system, we combine RGB intensity images with corresponding disparity maps to create a 3D scene. Each depth-plane's RGB intensity distribution is then convolved with the PSF

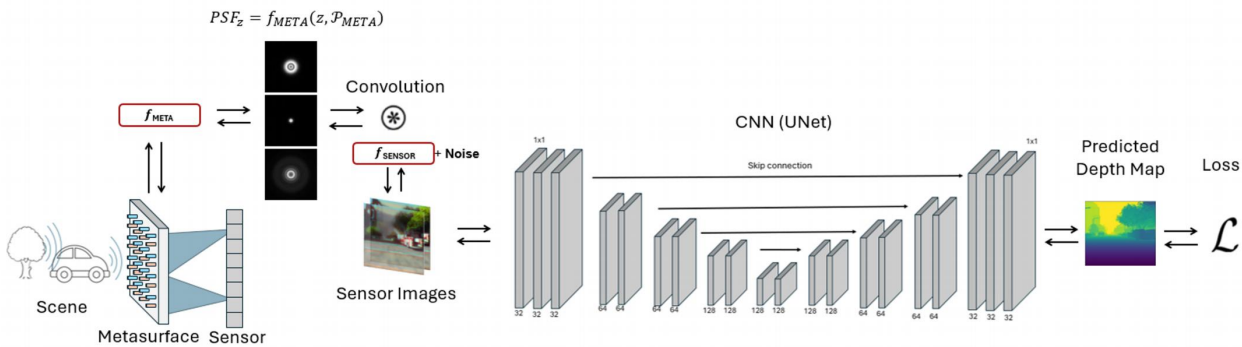

Fig. 1. End-to-end optimization architecture.

associated with that depth. The resulting sensor image is described by,

$$I_{sensor}(x, y) = \sum_{z_i} PSF_{z_i} * I_{z_i}(x, y) \tag{5}$$

where z_i are discretely sampled depth planes and $'*'$ represents a convolution.

3.3 Sensor Noise

We follow reference [4] to model the noise on our sensor. The noise contains both Gaussian and Poisson components and is introduced on a per-pixel basis. More specifically, if the normalized input to a sensor pixel is $x \in [0, 1]$ at a given sensor location, then the noisy measurement at that pixel is given by

$$f_{SENSOR}(x) = \eta_g(x, \sigma_g) + \eta_p(x, a_p) \tag{6}$$

where $\eta_g(x, \sigma_g) \sim \mathcal{N}(x, \sigma_g^2)$ is the Gaussian noise component and $\eta_p(x, p) \sim \mathcal{P}(x/a_p)$ is the Poisson noise component. We kept the empirically estimated parameters, used in [4], $\sigma_g = 1 \times 10^{-5}$ and $a_p = 4 \times 10^{-5}$.

A key feature of this network is the expression of sensor noise as a differentiable step within the imaging pipeline. It is found [4] that by allowing the end-to-end network to adapt for noisy measurements is preferable to conventional TV regularizers, which tend to blur out high frequency components.

3.4 UNet

A UNet convolutional neural network is a type of architecture commonly used for image segmentation tasks. It consists of a contracting path meant to capture context and a symmetric expanding path to precisely localize objects in images. By utilizing skip connections between corresponding layers in the contracting and expanding paths, UNet efficiently combines high-resolution features with low-resolution semantic information. As such, UNET architectures have been used for, among other things, 3D reconstruction from monocular images [1]. We utilize the same architecture in our end-to-end model.

3.5 Loss calculation and Back propagation

Our endpoint loss function is designed to achieve good depth accuracy with good

The loss function is given by,

$$\mathcal{L} = \lambda_1 \mathcal{L}_1 + \lambda_{grad} \mathcal{L}_{grad} + \lambda_{perc} \mathcal{L}_{perc} \tag{7}$$

where $\mathcal{L}_1$ is the per pixel $\|L_1\|$-norm error, $\mathcal{L}_{grad}$ is the spatial gradient loss and $\mathcal{L}_{perc}$ is a VGG19 architecture based perceptual loss [9]. The relative weights $\{\lambda_1, \lambda_{grad}, \lambda_{perc}\}$ are finely tuned to give the best predictions. We add a depth-dependent weight to the per-pixel loss, such that $\lambda_1(z)$, to bias the loss to a desired depth range.

The VGG19 perceptual loss model utilizes the VGG19 convolutional neural network architecture, pre-trained on large-scale image classification tasks. Instead of using traditional pixel-wise loss functions, it employs feature reconstruction loss, where features extracted from deep layers of the VGG19 network are compared between the generated and ground truth images. By emphasizing perceptually meaningful features, such as textures and object shapes, the model can produce visually pleasing results.

3.6 Model Architecture

The suggested architecture for this model is shown in figure 1. In summary, the image formation model can be described by

$$\mathbf{O} = \mathbf{f}_{\text{UNET}}\left(\mathcal{P}_{\text{UNET}}, \mathbf{f}_{\text{SENSOR}}\left(\mathbf{I} * \mathbf{f}_{\text{META}}\left(\mathcal{P}_{\text{META}}\right)\right)\right) \tag{8}$$

where $\mathbf{O}$ is the estimated depth, $\mathcal{P}_{\text{UNET}}$ ($\mathcal{P}_{\text{META}}$) are free variables of the UNet (metasurface) and, f_{SENSOR}, f_{META} and f_{UNET} are the forward pass functions. The free parameters are updated by minimizing the loss as defined above,

$$\{\mathcal{P}^*_{\text{META}}, \mathcal{P}^*_{\text{UNET}}\} = \underset{\mathcal{P}_{\text{META}}, \mathcal{P}_{\text{UNET}}}{\arg\min} \sum_{i=1}^{M} \mathcal{L}(\mathbf{O}^{(i)}, \mathbf{I}^{(i)}). \tag{9}$$

We define two metalenses initialized with different parameters (and thus focal length). After convolution with their respective PSFs, the two images are fed through the sensor forward pass function, adding noise. Then, they are passed as a 6-channel input into UNet (two RGB images). The UNet predicts a depth and the loss is calculated with respect to the ground truth depth map.

4 Experimental Results

4.1 Depth Estimation Results

Figure 2 shows the best depth estimate produced by our model. The top row shows the RGB image and corresponding depth map. In order to train our model, we need to discretize the depth-map. Although there are methods to extrapolate discretely convoluted depth maps to a continuous convolution, it is beyond the scope of this project. The bottom row shows the segmented ground truth depth map used to create the sensor image and the concomitant depth estimation. The estimation is made by a forward pass through the final model and is representative of an average test batch. The final PSNR is 56.50 dB and SSIM = 0.9998. All the final focal lengths of the two metasurfaces are summarized in Table 1.

4.2 Effect of Learning Rate

Throughout training the model, we noticed that the final prediction was highly sensitive to the learning rate of the metasurface parameters. By varying the learning rate we get dramatically different phase profiles for the two lenses. Fig 3 shows the focal length of three wavelengths for both metalenses as the model is being trained. For the highest learning rate (left), the focal lengths all converge to small values outside of the depth range. Clearly, this is a sub-optimal solution, and the resulting depth prediction confirms this. The intermediate learning rate converges to extremely unintuitive phase profiles (middle), where one of the lenses is weakly diverging for the blue and green wavelengths. The slowest learning rate (right) produces the best depth estimates. The final focal lengths are distributed across the depth-range, such that information is acquired throughout. For instance, the relative blur of the blue and green channel from one of the metalens images can estimate depth on long range. An evidence of the need for small learning rates in this case can be seen in rightmost plot, where initially the solid lines appear to be decreasing toward the same local minima as the largest learning rate. However, due to the smaller steps in the negative gradient direction, it is still able to self correct toward more optimal minima.

4.3 Relative Training Time Allocation

In the first iterations of training the model, the metasurface phase profile parameters were very unstable. This ultimately lead to poor final results, where the metasurface generally found a local minima amounting to blurring to depth prediction to an average gray value. Because of this, we effectively pre-trained the UNet with 5000 iterations before introducing any corrections to the metasurfaces. At that point, we ran 100 iteration of the UNet for every metasurface parameter update. This gave a slow and stable parameter evolution of the metalenses and the resulting phase profiles were much better at predicting depth. Fig 4 shows the typical evolution of the training, where every steps amount to 100 iterations of UNet or UNet+metasurface.

4.4 Point Spread Function Visualization

Fig 5 shows the final point spread functions as a function of depth for both metalenses and wavelengths from Table 1.

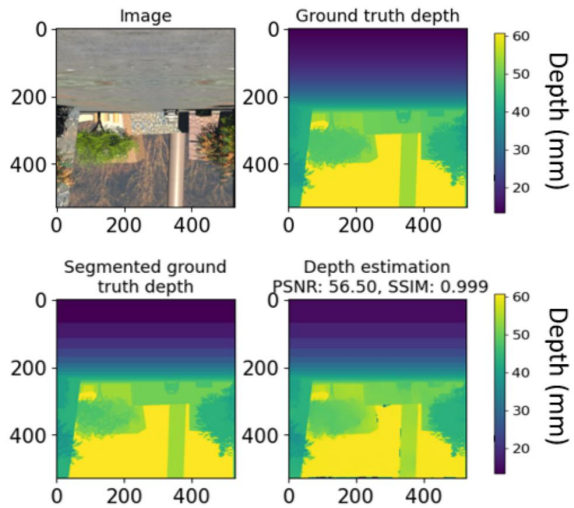

Fig. 2. Depth estimate for an arbitrary scene. We use the Driving datasets from 3DSceneFlow for the training and testing of our model. We feed the segmented ground truth depth maps with 17 z depths. The depth estimation shows a great PSNR and structure similarity (SSIM) to the segmented depth map.

	Focal length (mm)		
	Red (606nm)	Green (511nm)	Blue (462nm)
Metalens 1	14.8	30.0	63.3
Metalens 2	12.9	24.0	42.9

TABLE 1

5 Conclusion

Next-generation miniature optical devices for task-specific applications can greatly benefit from end-to-end deep optics networks, such as the one presented here. This joint optimization approach allows for symbiotic information encoding + decoding by the optical and digital layers, respectively. In addition, extra color channels, traditionally associated with unwanted aberrations in metasurfaces, can not only be corrected, but also leveraged by neural networks as they carry extra information correlating the input fields with the desirable outputs.

Acknowledgments

The authors would like to thank Gun-Yeal Lee for his guidance and mentoring throughout this project.

References

[1] H. Ikoma, C. M. Nguyen, C. A. Metzler, Y. Peng, and G. Wetzstein, "Depth from defocus with learned optics for imaging and occlusion-aware depth estimation," in *2021 IEEE International Conference on Computational Photography (ICCP)*. IEEE, 2021, pp. 1–12.

[2] F. Liu, C. Shen, and G. Lin, "Deep convolutional neural fields for depth estimation from a single image," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 5162–5170.

[3] N. Yu and F. Capasso, "Flat optics with designer metasurfaces," *Nature materials*, vol. 13, no. 2, pp. 139–150, 2014.

[4] E. Tseng, S. Colburn, J. Whitehead, L. Huang, S.-H. Baek, A. Majumdar, and F. Heide, "Neural nano-optics for high-quality thin lens imaging," *Nature communications*, vol. 12, no. 1, p. 6493, 2021.

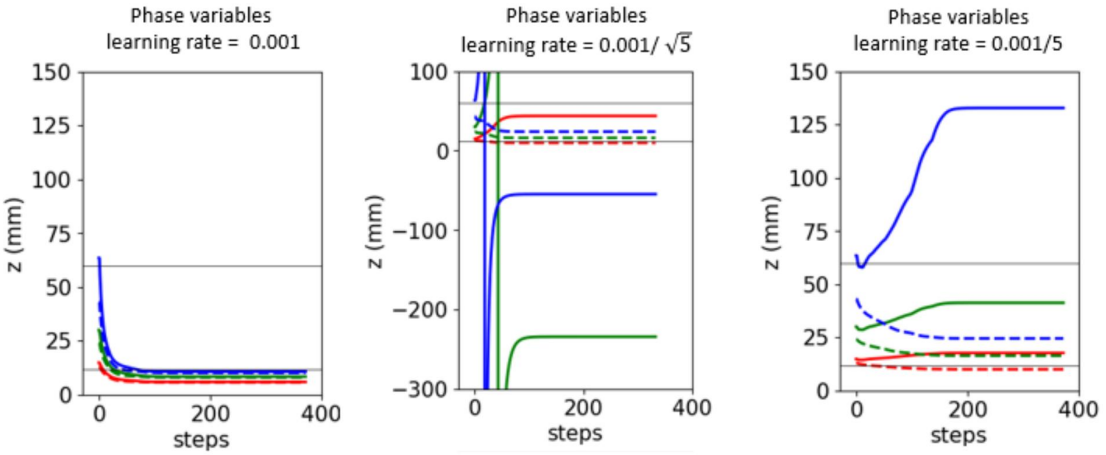

Fig. 3. The focal depths of two metalens evolve during the training for three different learning rate of the metasurface phase variables. The solid color lines are the focal depths for three color channels in one metalens, and the dash lines for the second metalens.

[5] Q. Guo, Z. Shi, Y.-W. Huang, E. Alexander, C.-W. Qiu, F. Capasso, and T. Zickler, "Compact single-shot metalens depth sensors inspired by eyes of jumping spiders," *Proceedings of the National Academy of Sciences*, vol. 116, no. 46, pp. 22 959–22 965, 2019.

[6] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*. Springer, 2015, pp. 234–241.

[7] E. Alexander, "A theory of depth from differential defocus," Ph.D. dissertation, Harvard University, 2019.

[8] T. Nagata, M. Koyanagi, H. Tsukamoto, S. Saeki, K. Isono, Y. Shichida, F. Tokunaga, M. Kinoshita, K. Arikawa, and A. Terakita, "Depth perception from image defocus in a jumping spider," *Science*, vol. 335, no. 6067, pp. 469–471, 2012.

[9] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*. Springer, 2016, pp. 694–711.

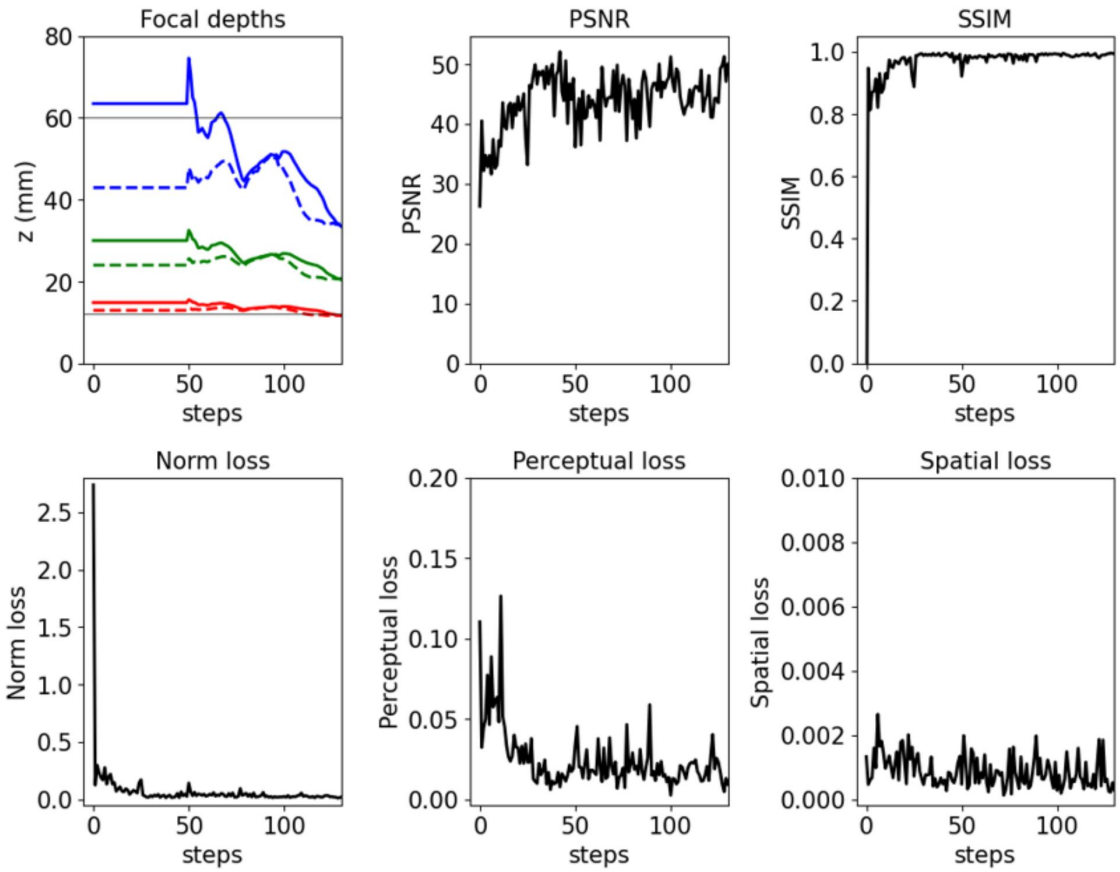

Fig. 4. Training evolution. We start by training the UNet for 5000 iterations (50 steps). During this time the focal lengths remain constant.

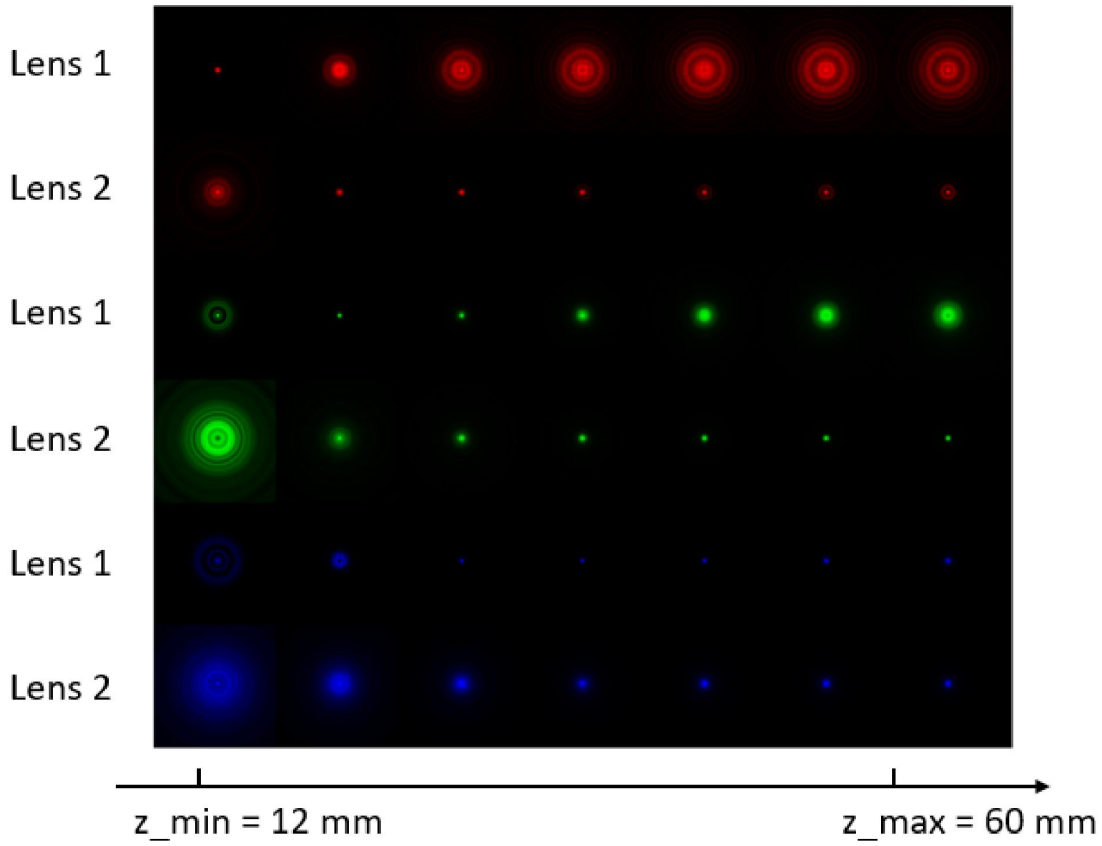

Fig. 5. PSF as a function of depth (mm) for different wavelengths and both metalenses