

End-to-end Learning for Joint Depth and Image Reconstruction from Diffracted Rotation: *Supplementary Material*

Mazen Mel^a, Muhammad Siddiqui^b, Pietro Zanuttigh^a

^a*Department of Information Engineering, University of Padova, Italy.*

^b*Sony Europe B.V., Zweigniederlassung Deutschland Stuttgart Technology Center, Stuttgart, Germany.*

Summary

In this supplementary material document, we start by analyzing the focusing error in Out Of Focus (OOF) imaging in Section 1. Then in Section 2 we briefly present the wave-optics based analysis of light field propagation and the phase transformation introduced by the lens and the phase mask. We also show the effects of the phase mask design parameters on the shape of the resulting RPSF in Section 3.

Then, we present some additional qualitative results for the tasks of monocular depth estimation and image deblurring (Sections 4 and 5) that were not possible to fit in the main paper due to space limitations. Finally, a visual example of the results of the various tests made in the ablation study is also presented in Section 6.

*Corresponding author

Email address: `mazen.mel@phd.unipd.it` (Mazen Mel)

1. Out of Focus Imaging

Consider a simple imaging system consisting of a thin lens with focal length f . The light reflected from an object at distance z_o in front of the lens is focused into an image plane at a distance z_i behind it, i.e., all rays coming from a point-like object are
 5 focused into a point in the image plane that satisfy the well-known thin lens equation $(\frac{1}{z_o} + \frac{1}{z_i} = \frac{1}{f})$.

Away from the in-focus plane, where the thin lens equation is no longer satisfied, a quadratic phase error at the pupil plane is introduced to model OOF objects. Such error is measured by the defocus value [1]:

$$\Psi = \frac{\pi R^2}{\lambda} \left(\frac{1}{z_o} + \frac{1}{z_i} - \frac{1}{f} \right) \quad (1)$$

10 Where R is the pupil radius and λ is the wavelength of incident light waves. Ψ indicates the severity of the focusing error and it increases in absolute value as objects move away from the in-focus plane.

2. Field Propagation and the Point Spread Function

In this section, we discuss how a Fresnel diffraction model can be used to simulate
 15 near-field wave propagation. Consider a thin lens with a thickness profile h situated at the pupil plane of an imaging system (as depicted in Fig. 1), notice that h is a function of the spatial coordinates (x, y) , and let h_0 be the thickest section of the lens. Generally the lens is made of glass with a refractive index $n = 1.5$. For a convex thin lens with focal length f , $h(x, y)$ is defined as:

$$h(x, y) = h_0 - \frac{x^2 + y^2}{2f(n - 1)} \quad (2)$$

20 Light wave fields passing through the lens are delayed by an amount proportional to the thickness of the lens h at each point. Thus, a thin lens applies a phase shift to the incident wavefront given by:

$$\Phi_{lens}(x, y) = \frac{2\pi(n - 1)}{\lambda} h(x, y) \quad (3)$$

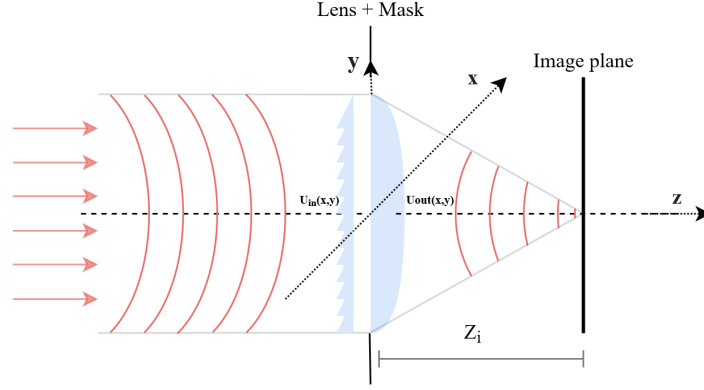


Figure 1: Light emitted by an object propagates through free space and its phase is delayed when it passes through the camera optics.

Notice that the phase shift is wavelength-dependent hence using the same lens with polychromatic light results in chromatic aberrations.

25 By defining the phase shift at the pupil plane, the generalized pupil function, which is a complex function, can be written as:

$$P(x, y) = A(x, y)e^{i\Phi_{lens}(x, y)} \quad (4)$$

Where $A(x, y)$ is the circular aperture mask simulating a finite aperture area.

In the case of OOF imaging, the pupil function would also introduce a phase error expressed as a quadratic phase term in the generalized pupil function showing the effects of defocus aberrations in the captured image. The new generalized pupil function
30 would have the following form:

$$P(x, y) = A(x, y)e^{i(\Phi_{lens}(x, y) + \Psi \frac{(x^2 + y^2)}{R^2})} \quad (5)$$

Where R is the pupil radius and Ψ is the defocus parameter. In the case of a wave field generated by an ideal point source object, the PSF of the imaging system is expressed as:

$$PSF(x, y) \propto |\mathcal{F}(P(x, y))|^2 \quad (6)$$

35 Where $\mathcal{F}$ denotes the Fourier Transform.

Concerning the wave field, let the complex wave field U_{in} propagating in free space just before the entrance pupil be:

$$U_{in}(x, y) = C(x, y)e^{i\Phi_{in}(x, y)} \quad (7)$$

Where $C(x, y)$ is the field's amplitude and Φ_{in} is its phase just before entering the pupil.

40 In the presence of a phase mask with a phase delay Φ_{mask} (as shown in Fig. 1), the phase delay introduced by the combination of both the lens and mask is expressed as:

$$\Phi_{optics}(x, y) = \Phi_{lens}(x, y) + \Phi_{mask}(x, y) \quad (8)$$

Thus, the expression of the wave field just after the lens system U_{out} in this case is:

$$\begin{aligned} U_{out}(x, y) &= P(x, y)U_{in}(x, y) \\ &= A(x, y)C(x, y)e^{i(\Phi_{in}(x, y) + \Phi_{optics}(x, y))} \end{aligned} \quad (9)$$

Using Fresnel propagation, U_{out} can be further propagated of a distance z_i until it reaches the image sensor (we use a coordinate system (u, v) for the sensor).

$$\begin{aligned} U_{sensor}(u, v) &= \frac{e^{ikz_i}}{i\lambda z_i} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} U_{out}(x, y) \times \\ &\quad e^{\frac{ik}{2z_i}[(u-x)^2 + (v-y)^2]} dx dy \end{aligned} \quad (10)$$

45 Where $k = \frac{2\pi}{\lambda}$ is the wave number.

3. Visual Evaluation of the Impact of RPSF Design Parameters

In this section we show some visual examples of RPSF shapes obtained by using different values of the design parameters.

Recall that we denoted with N the number of rotating peaks within the RPSF: in 50 the first row of Fig. 2 we show the resulting RPSF shape for different values of N .

The design parameter ϵ controls the peak separation as well as the confinement of each peak as illustrated in the second row of Fig. 2. It was empirically hypothesized by [2] that the depth range increases as ϵ decreases.

The number of Fresnel zones L within the phase mask also controls the peak separation as illustrated in the last row of Fig. 2. Additionally, the rate of rotation which is $\frac{1}{L}$ in the basic case of a single helix RPSF, can be easily extended to $\frac{1}{NL}$ in the more general case of multiple peaks. This indicates that the rate of rotation would decrease with higher L and/or N values leading to an increase in the practical depth range with no peak rotation ambiguity.

In summary, $[N, L, \epsilon]$ can be jointly optimized depending on the target task: they all influence the shape of the RPSF and the practical depth range in which the PSF can be used to encode unambiguous depth information within the captured 2D images.

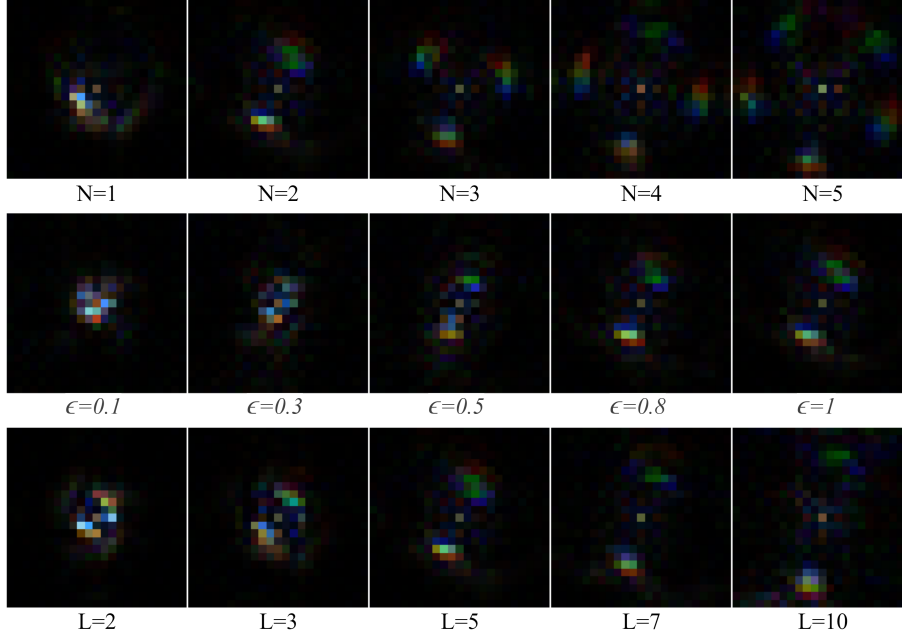


Figure 2: The effects of $[N, L, \epsilon]$ design parameters on the shape of the RPSF: (first row) only the value of N is changed, i.e., $[N = 1, 2, 3, 4, 5; L = 5; \epsilon = 0.9]$, (second row) only the value of ϵ is changed, i.e., $[N = 2; L = 5; \epsilon = 0.1, 0.3, 0.5, 0.8, 1]$, (third row) only the value of L is changed with $[N = 2; L = 2, 3, 5, 7, 10; \epsilon = 0.9]$.

4. Additional Visual Results for Image Deblurring

Fig. 3 shows some additional examples of the proposed image deblurring algorithm on the FlyingThings3D subset [3]: the network achieves overall satisfactory reconstruction with minor artifacts (a bit of ringing and a few texture patterns not properly recovered are visible upon closer inspection). Some challenging complex structures are restored (e.g. the sample shown in the third row) while other fine details, such as the periodic pattern of the “floor” shown in the sample of the last row, proved difficult to be accurately recovered.

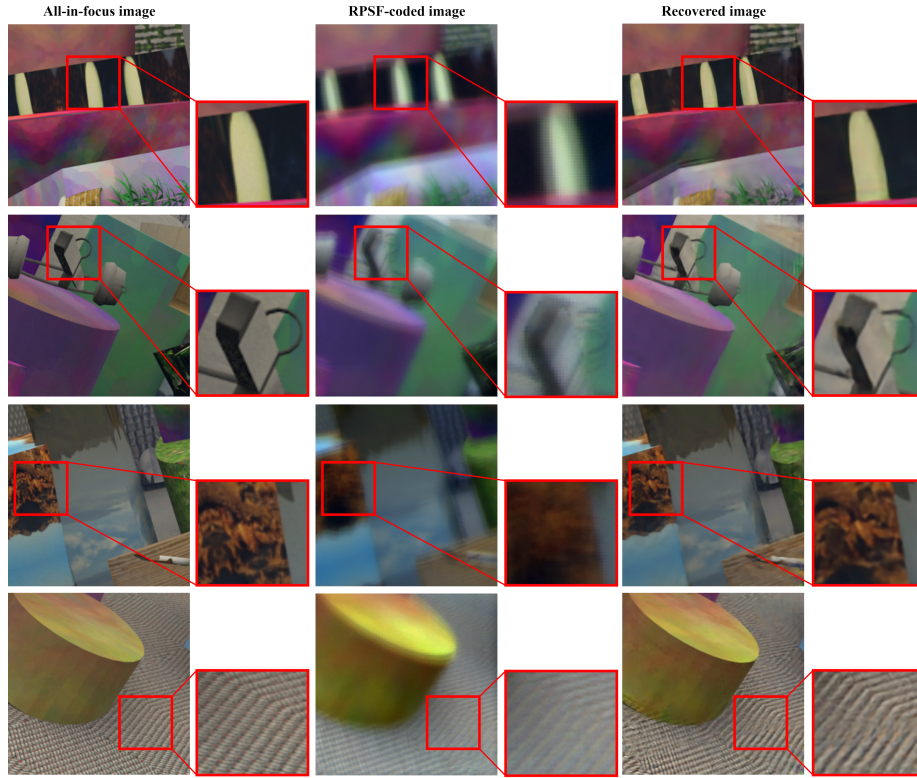


Figure 3: Additional qualitative results of the image deblurring model on the test set of FlyingThings3D [3] subset.

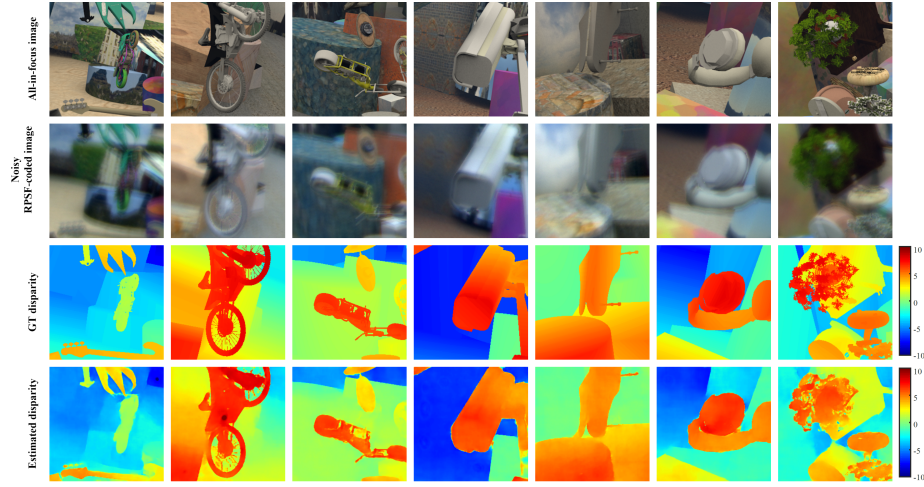


Figure 4: Additional qualitative results on RPSF-blurred images from the test set of FlyingThings3D [3] subset.

5. Additional Visual Results for Monocular Depth Estimation

Fig. 4 shows some additional examples of the proposed depth estimation algorithm on sample images from the FlyingThings3D test set [3]. These results correspond to the simulated noisy input images where sensor noise, CFA, and quantization noise are added. The network was able to produce accurate disparity maps even for objects with very fine details (e.g. the motorcycles in the second and third columns in addition to the plant in the last column). On the other side, edge pixels are not always accurately estimated (e.g., see the fourth column) as discussed in Section VI in the paper.

Then, Fig. 5 shows some additional examples of the proposed depth estimation algorithm on sample images from the NYUV2 test set [4] and compares them with those from AdaBins [5] competing approach. Our approach proved capable of producing superior results for small and large depths ranges alike with no effects or ambiguities whatsoever originating from the scene’s semantics.

Finally, Fig. 6 shows some additional examples of the proposed depth estimation algorithm on sample images from the SUNRGBD test set [6]. Even though the network was not fine-tuned for this specific dataset, it performed very well on this larger and challenging dataset.

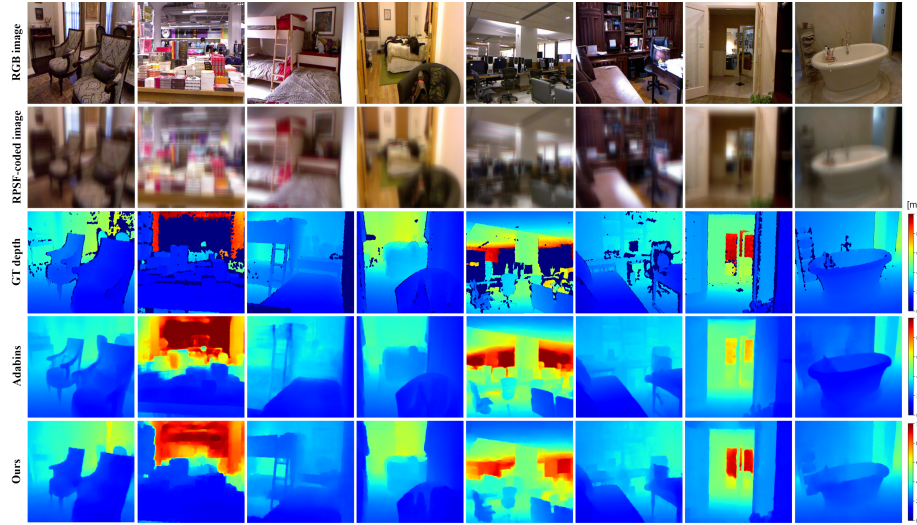


Figure 5: Additional qualitative results from the proposed approach as well as those from AdaBins [5] on the test set of NYUV2 [4].

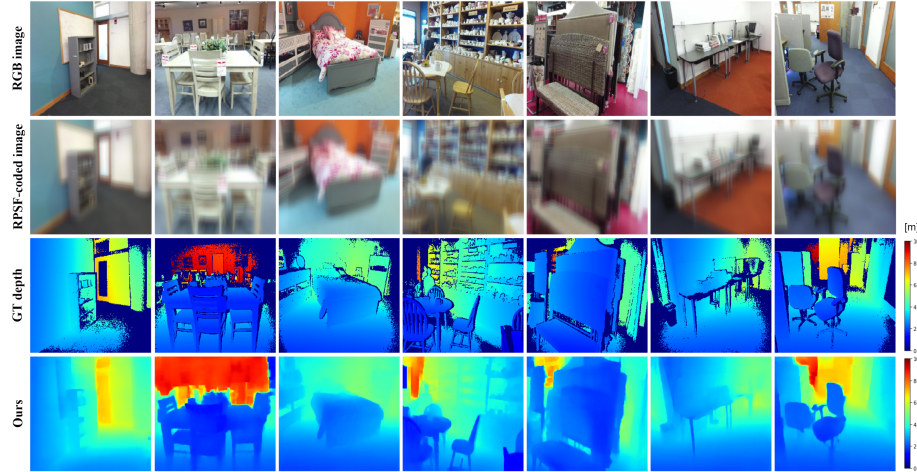


Figure 6: Additional qualitative results from our proposed method on the test set of SUNRGBD [6]

6. Qualitative Results for the Ablation Study

Fig. 7 shows a sample of the predicted depth maps for the four conducted ablation
 90 experiments discussed in the paper. Quantitatively, the network trained using all-in-
 focus images as input fails to discriminate between positive and negative defocus val-

ues, i.e., it is very difficult for the network to accurately predict whether objects are close or far from the camera because of the lack of reliable depth cues within the clear aperture PSF. While the rest of the performed experiments significantly improve the depth estimation accuracy, the learned mask proved capable not only to predict accurate depth maps but also to preserve fine details around object edges at different depth planes as shown, for example, in Fig. 7 around the plant edges.

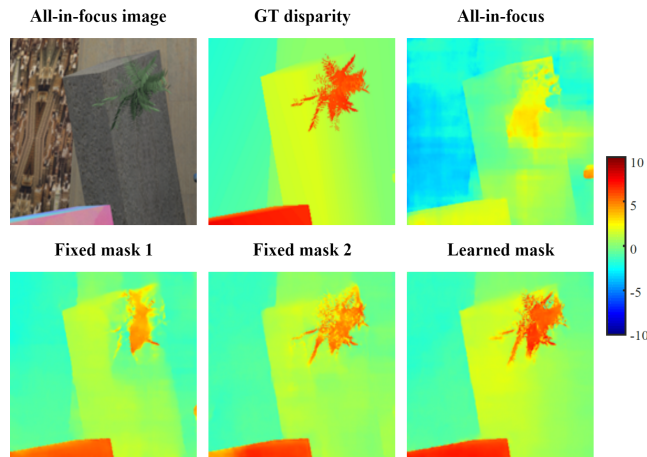


Figure 7: A sample from the qualitative results for the four conducted ablation experiments on the test set of the subset of FlyingThings3D [3].

References

- [1] J. W. Goodman, Introduction to Fourier optics, Roberts and Company Publishers, 2005.
- [2] R. Berlich, S. Stallinga, High-order-helix point spread functions for monocular three-dimensional imaging with superior aberration robustness, Optics express 26 (4) (2018) 4873–4891.
- [3] N. Mayer, E. Ilg, P. Häusser, P. Fischer, D. Cremers, A. Dosovitskiy, T. Brox, A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation, in: IEEE International Conference on Computer Vision and

Pattern Recognition (CVPR), 2016, arXiv:1512.02134.

URL [http://lmb.informatik.uni-freiburg.de/
Publications/2016/MIFDB16](http://lmb.informatik.uni-freiburg.de/Publications/2016/MIFDB16)

- 110 [4] D. Eigen, R. Fergus, Predicting depth, surface normals and semantic labels with a
common multi-scale convolutional architecture, in: Proceedings of the IEEE inter-
national conference on computer vision, 2015, pp. 2650–2658.
- [5] S. F. Bhat, I. Alhashim, P. Wonka, Adabins: Depth estimation using adaptive bins,
arXiv preprint arXiv:2011.14141.
- 115 [6] S. Song, S. P. Lichtenberg, J. Xiao, Sun rgb-d: A rgb-d scene understanding bench-
mark suite, in: Proceedings of the IEEE conference on computer vision and pattern
recognition, 2015, pp. 567–576.