



Deep learning estimation of modified Zernike coefficients and recovery of point spread functions in turbulence

ABU BUCKER SIDDIK,* STEVEN SANDOVAL, DAVID VOELZ,
LAURA E. BOUCHERON, AND LUIS VARELA

Klipsch School of Electrical and Computer Engineering, New Mexico State University, Las Cruces, NM 88003, USA

*siddik@nmsu.edu

Abstract: Recovering the turbulence-degraded point spread function from a single intensity image is important for a variety of imaging applications. Here, a deep learning model based on a convolutional neural network is applied to intensity images to predict a modified set of Zernike polynomial coefficients corresponding to wavefront aberrations in the pupil due to turbulence. The modified set assigns an absolute value to coefficients of even radial orders due to a sign ambiguity associated with this problem and is shown to be sufficient for specifying the intensity point spread function. Simulated image data of a point object and simple extended objects over a range of turbulence and detection noise levels are created for the learning model. The MSE results for the learning model show that the best prediction is found when observing a point object, but it is possible to recover a useful set of modified Zernike coefficients from an extended object image that is subject to detection noise and turbulence.

© 2023 Optica Publishing Group under the terms of the [Optica Open Access Publishing Agreement](#)

1. Introduction

The point spread function (PSF) refers to the impulse response of an imaging system [1]. When the PSF is known, it can be used for the correction of blur and other artifacts in images that are due to the system's response. For example, in a space-invariant imaging situation, a PSF correction might be applied in a deconvolution step. Additionally, the propagation of light through a medium such as the atmosphere introduces wavefront aberrations at the aperture plane that further degrade the images. Therefore, the estimation of the combined system and medium PSF can potentially quantify the aberrations so that image correction can be performed.

1.1. Imaging model

Incoherent imaging of an object can be modeled using a linear space-invariant forward model

$$I(x, y) = h(x, y) * I_0(x, y), \quad (1)$$

where $I(x, y)$ is the intensity image of a source object $I_0(x, y)$, $*$ is the convolution operator, and $h(x, y)$ is the intensity PSF given by

$$h(x, y) \propto |\mathcal{F}\{p(x, y)e^{j\phi(x, y)}\}|^2, \quad (2)$$

where \mathcal{F} is the Fourier transform, $|\cdot|$ is the modulus operator, $p(x, y)$ is the pupil function, and $\phi(x, y)$ is the wavefront phase distortion applied at the pupil plane [2].

1.2. Representing wavefront distortions

Although the PSF can be parameterized in various ways, to leverage results from disciplines such as adaptive optics, we consider the wavefront distortion $\phi(x, y)$ related to the PSF through Eq. (2).

The phase distortion can be represented by a linear superposition of Zernike polynomials

$$\phi(x, y) = \sum_{n,m} a_{nm} Z_n^m(x, y), \quad (3)$$

where Z_n^m are the Zernike polynomials and a_{nm} are the Zernike coefficients [3–6]. Zernike polynomials form a set of orthogonal functions defined over the unit circle and may be expressed in a double (m, n) indexing scheme [7] as

$$Z_n^m(\rho, \varphi) = \begin{cases} R_n^{|m|}(\rho) \cos(m\varphi), & \text{if } m \geq 0 \\ R_n^{|m|}(\rho) \sin(m\varphi), & \text{if } m < 0 \end{cases} \quad (4)$$

where ρ is the normalized radial distance ($0 \leq \rho \leq 1$), φ is the azimuthal angle, n is the radial order, m is the angular frequency with $n \geq m \geq 0$, and $R_n^{|m|}(\rho)$ are radial polynomials given by

$$R_n^{|m|}(\rho) = \begin{cases} \sum_{p=0}^{\frac{n-|m|}{2}} \frac{(-1)^p (n-p)!}{p! \left(\frac{n+|m|}{2} - p\right)! \left(\frac{n-|m|}{2} - p\right)!} \rho^{n-2p}, & \text{if } n - |m| = 0 \pmod{2} \\ 0, & \text{otherwise.} \end{cases} \quad (5)$$

For our work, it is more convenient to utilize the single-indexed Zernike polynomials Z_q where the index q is related (OSA/ANSI standard) to the double-index by

$$q = \frac{n(n+2) + m}{2}. \quad (6)$$

Zernike polynomials can be illustrated in a methodical way using a pyramid structure as given in Fig. 1. In the figure, the indexing starts with $q = 0$ on the top, then moves down the pyramid while scanning left to right. The rows of the pyramid correspond to the radial order n and the columns correspond to the angular frequency m . Using single-indexed Zernike polynomials, we rewrite Eq. (3) in polar coordinates as

$$\phi(\rho, \varphi) = \sum_q a_q Z_q(\rho, \varphi). \quad (7)$$

Due to the relationships between the intensity image $I(x, y)$ in Eq. (1), the PSF $h(x, y)$ in Eq. (2), and the Zernike coefficients a_q in Eq. (7), the Zernike coefficients a_q correspond to a particular intensity result and can thus be used to parameterize and model the associated PSF of an aberrated imaging system.

Unfortunately, there is an ambiguity associated with determining the Zernike coefficients from an intensity image alone without knowledge of the field phase. In particular, there are some Zernike polynomials Z_q which generate the same PSF intensity image for oppositely-signed Zernike coefficients (details in Supplement 1). Due to this ambiguity, neural networks and other learning machines which have been extensively used to estimate PSFs for aberrated imaging systems [5,8–11] face difficulty in directly predicting Zernike coefficients from intensity images [2,9]. In [9], the deep learning model completely fails to predict Zernike coefficients for the in-focus setup from an extended source object due to this intensity image ambiguity (although the ambiguity is not pointed out or identified as the source of failure by the authors). In an attempt to improve performance, the authors in [9] analyze different preconditioners: overexposure, defocus, and scatter, and assess the impact on Zernike coefficient prediction. In [2], the authors recognize the PSF intensity image ambiguity and address it by performing multiple measurements (focused and defocused intensity images) of the source object, passing them through a feature extractor block followed by a neural network (Resnet50).

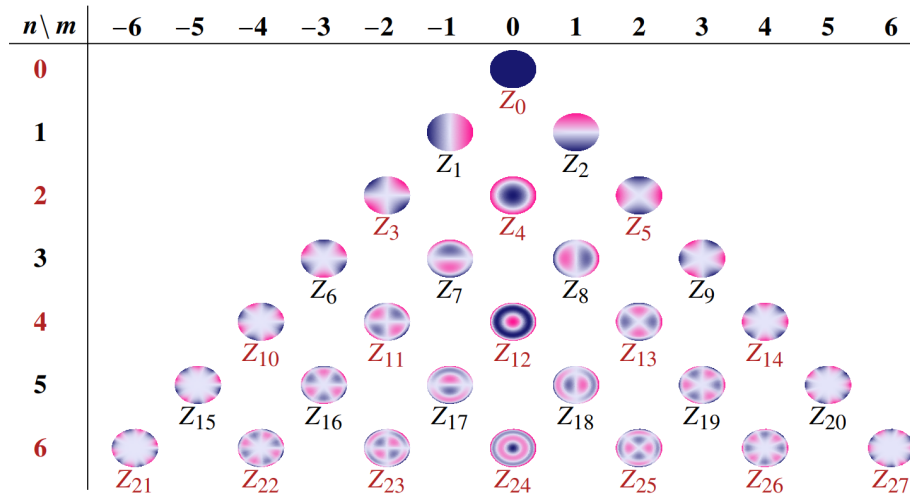


Fig. 1. The first 28 Zernike polynomials arranged in a pyramid structure. The rows follow radial order n and columns follow angular frequency m [7]. The angularly even polynomials which are subject to ambiguity in intensity images correspond to even values of n (corresponding labels are colored red). In the colormap, positive values are indicated as blue, zero values are indicated as white, and negative value are indicated as red.

In our work, we propose a novel approach to address this ambiguity. We begin by identifying the Zernike polynomials which are susceptible to the ambiguity and for these polynomials we only consider the prediction of the absolute value of the associated Zernike coefficients. For the remainder of this work, we will refer to the signless Zernike coefficients for polynomials susceptible to the ambiguity in addition to the signed Zernike coefficients not susceptible to the ambiguity collectively as the *modified Zernike coefficients*. We use a deep neural network to predict the modified Zernike coefficients directly from intensity images. We show that these predicted coefficients can then be used to model the PSF of the aberrated imaging system. Finally, we consider the performance of our network in predicting the modified Zernike coefficients for both point source and extended source objects using incoherent imaging scenarios with both low and high noise levels (Poisson noise and read-out noise). Note that the proposed method is not a complete wavefront sensing approach, particularly because it does not consider the sign of the Zernike coefficients for polynomials that are susceptible to ambiguity. However, it works unambiguously for recovering the point spread function and provides a new construct and insight for the development of novel wavefront sensing methods [12–16].

The remainder of the work is organized as follows. In Section 2, we describe the methodology, data generation process, and deep learning architecture to predict modified Zernike coefficients. In Section 3, we illustrate our results and discuss the logical interpretation of these results. In Section 4, we conclude our work.

2. Methods

2.1. Methodology

As has been pointed out in Section 1.2, there is an ambiguity associated with predicting the Zernike coefficients from intensity images. In particular, we have found via symmetry properties of the Fourier transform [17] (details in Supplement 1) that angularly even Zernike polynomials which correspond to even values of radial order n and even values of angular frequency m generate the same PSF and thus the same intensity image for oppositely signed Zernike coefficients. In

other words, the signs of these coefficients are immaterial for the description of the intensity PSF. Figure 1 shows the first 28 polynomials where labels of angularly even polynomials are colored red. To address the ambiguity, we predict the signless Zernike coefficients for angularly even polynomials and the signed Zernike coefficients for angularly odd polynomials using a deep neural network. This approach might not be sufficient for applications such as wavefront sensing where the coefficient signs are important for specifying the actual shape of the wavefront, but it is useful for characterizing and correcting intensity images of an aberrated imaging system.

2.2. Data generation

A dataset was generated using the imaging model and Zernike polynomial representation of the phase aberration given in Section 1. More specifically, the wavefront phase distortion due to the aberrated system was introduced in the pupil plane by fitting a Kolmogorov turbulence phase screen that is applied to the pupil. The strength of the phase screen was parameterized by the ratio D/r_0 , where D is the diameter of the pupil and r_0 is the Fried parameter (the coherence diameter of the transverse turbulence field) [18,19]. Other simulation values include the wavelength $\lambda = 0.5 \mu\text{m}$ and pupil diameter $D = 0.4 \text{ m}$.

A block diagram of our training methodology is given in Fig. 2. The process begins with the generation of a random phase screen $\phi(x, y)$ that obeys the Kolmogorov spectrum for a selected D/r_0 value [20]. The first 28 Zernike coefficients of the phase screen $\phi(x, y)$ are fitted using matrix inversion [21]. Other authors have demonstrated the effective estimation of this number of coefficients for turbulence-degraded wavefronts [2]. Then, the coefficients are modified by discarding the signs of the coefficients corresponding to angularly even Zernike polynomials resulting in the modified Zernike coefficients $\{a_q\}$. Additionally, the first three Zernike coefficients are considered as zero because there is no relationship between the Z_0 (piston) polynomial and the PSF intensity image [5], and the Z_1 (tip) and Z_2 (tilt) terms correspond to simple offsets in the image plane that can be easily found using centroiding algorithms or other registration methods [8,22]. The modified phase screen $\tilde{\phi}(x, y)$ is reconstructed and used along with the pupil function $p(x, y)$ and source object $I_0(x, y)$ in the imaging model described by Eq. (1) to generate the intensity image $I(x, y)$ for neural network input. Finally, the modified Zernike coefficients $\{a_q\}$ are compared to the output of the neural network $\{\hat{a}_q\}$ to determine the prediction error $\{\text{Error}_q\}$ associated with each Zernike coefficient. As an example, Fig. 3 shows a source object $I_0(x, y)$ and three intensity images $I(x, y)$ corresponding to three different D/r_0 values. It is noted that it is possible to skip the phase screen generation step and use random draws of Zernike coefficient values subject to known variances corresponding to the Kolmogorov spectrum [23] but we find it useful to have a high spatial resolution version of the original phase screen for comparison purposes.

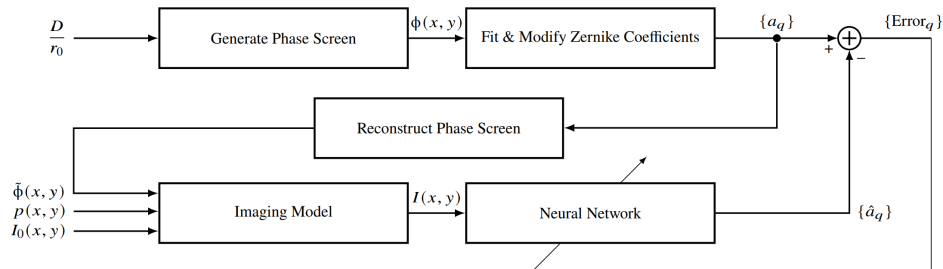


Fig. 2. Block diagram of the proposed training methodology.

Four different scenarios at nine atmospheric turbulence strengths $D/r_0 = 2, 3, \dots, 10$ were considered. For each scenario, 54000 intensity images (512×512 pixels, 6000 images per

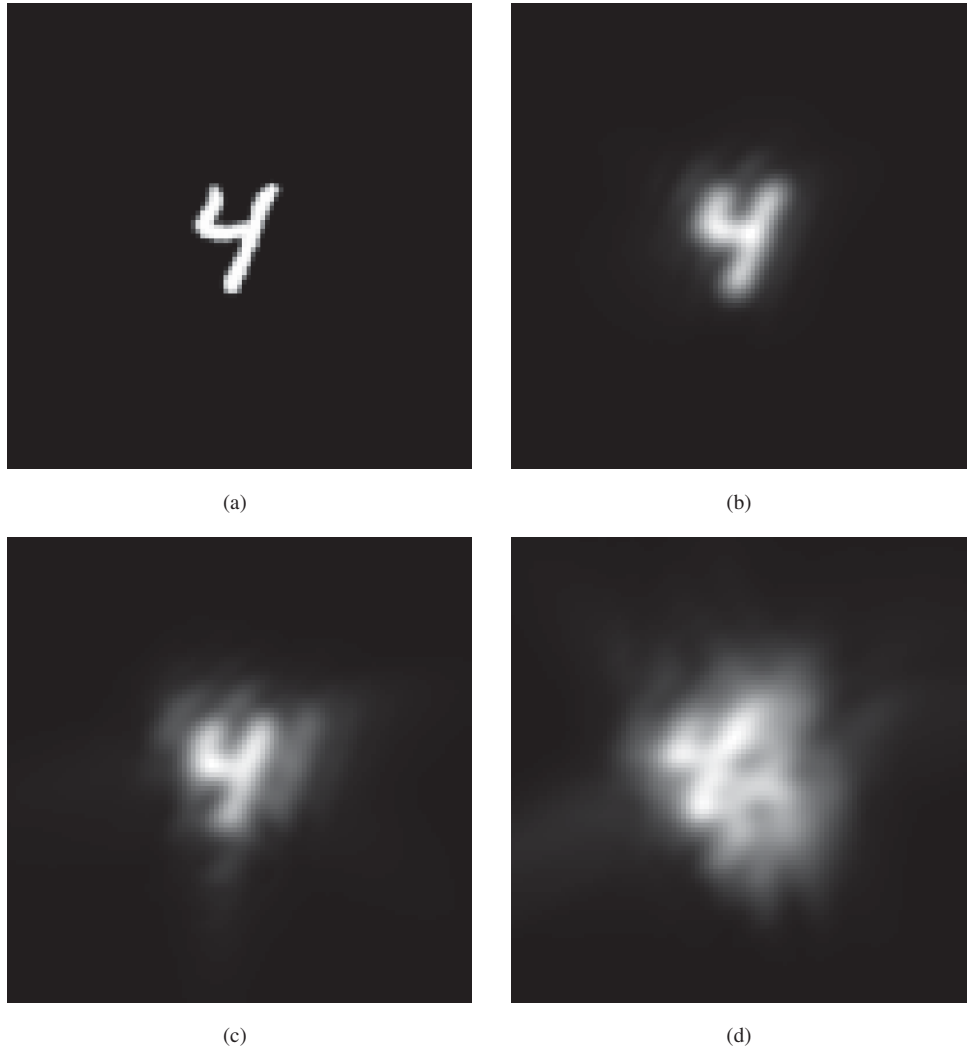


Fig. 3. Example intensity images of an extended source object for different $\frac{D}{r_0}$ ratios: (a) source object (b)-(d) intensity images for $\frac{D}{r_0} = 2, 5, 10$ respectively. The frames shown are 100×100 pixels.

D/r_0 value) were generated and then partitioned 45000/4500/4500 for training/validation/testing purposes. The first scenario considers only a point source object with zero noise. The remaining three scenarios consider extended source objects at zero/low/high noise levels, respectively. More specifically, 6000 different extended source objects (5000/500/500 for training/validation/testing purposes) were used for each D/r_0 . The extended source objects (28×28 pixels) were taken from the EMNIST database [24] and were zero-padded to 512×512 pixels. These objects have relatively simple spatial features, which provide the opportunity to extract the PSF even in the presence of wavefront components due to the object. For the low noise scenario, Poisson noise with peak photon levels of 4000 and read out noise with zero mean, standard deviation 10 was used. For the high noise scenario, Poisson noise with peak photon levels of 15000 and read out noise with zero mean, standard deviation 100 was used. These noise levels are similar to those used in [8].

2.3. Model architecture

We use a deep neural network architecture based on AlexNet [25,26] to predict the modified Zernike coefficients from PSF intensity images. The model structure is illustrated in Fig. 4. The model is implemented in the Julia language [27] using the Flux package [28]. The 512×512 pixel PSF intensity images for the data set are centrally cropped down to 100×100 pixels and the amplitude is normalized using min-max normalization [25] prior to input to the neural network. The cropping procedure is carried out to exclude the outer zero-value pixels that do not contain any significant information for further analysis. The model takes cropped (100×100 pixels) and amplitude normalized (in range $[0, 1]$) PSF images as input and outputs predictions for the modified Zernike coefficients that model the PSF of the aberrated imaging system. The network consists of 5 convolution layers, 3 max-pooling layers, and 3 fully-connected layers. Convolution layers 1 and 2 use 32 kernels with kernel size 5×5 and convolution layers 3, 4 and 5 use 64 kernels with kernel size 3×3 . Each convolution layer in the model is followed by a RELU activation. Max pooling layer 1 performs a 5×5 kernel operation and max pooling layers 2 and 3 perform 3×3 kernel operations. The fully-connected layers 1 and 2 are followed by a RELU activation and have 2000 and 512 nodes, respectively. The final fully-connected layer has 25 nodes followed by a linear activation. We use mean squared error (MSE) as the loss function and the ADAM optimizer [29] with default settings (learning rate $\eta = 0.001$ and decay rates $\mathcal{B} = (0.9, 0.999)$) as an optimizer for the model. The neural network is trained for 20 epochs with a batch size of 100 for each scenario described earlier as in Section 2.2. During the training process, the validation MSE typically levels off but does not diverge and the neural network with the lowest MSE for the validation data is saved for each scenario.

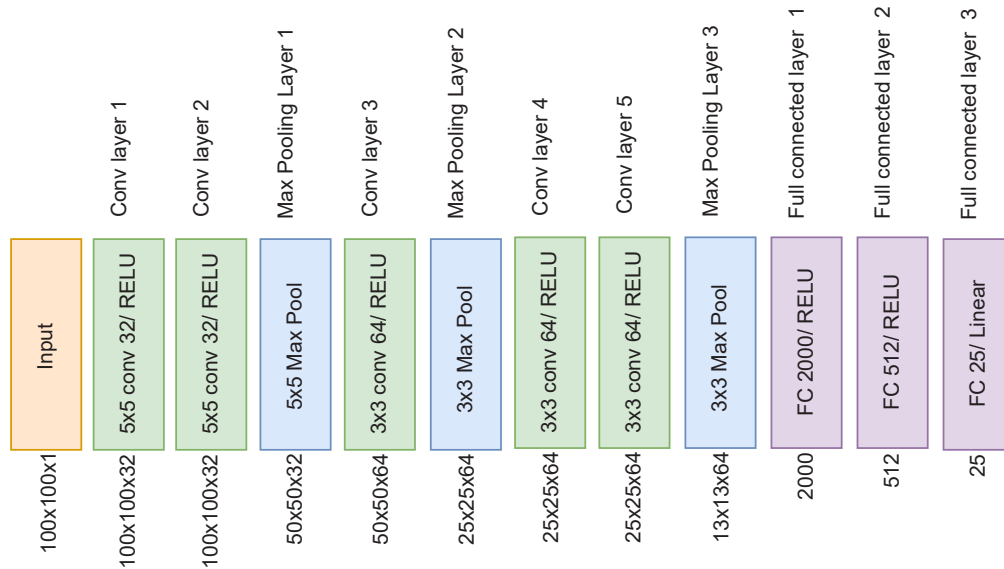


Fig. 4. The neural network architecture used for the estimation consists of several convolutional and pooling layers followed by three fully connected layers. The architecture is based on AlexNet [25,26].

3. Results and discussion

Table 1 shows prediction results for each of the four scenarios described in Section 2.2. Each of these results provides an average MSE for predicting the modified Zernike coefficients from the testing PSF intensity images. Additionally, when we repeated scenario 1, but instead of

Table 1. Average MSE for the four scenarios described in Section 2.2.

Scenario	Average MSE	Description
1	0.0218	Point Source Object
2	0.0599	Extended Source Objects
3	0.0792	Extended Source Objects with Low Noise
4	0.0974	Extended Source Objects with High Noise

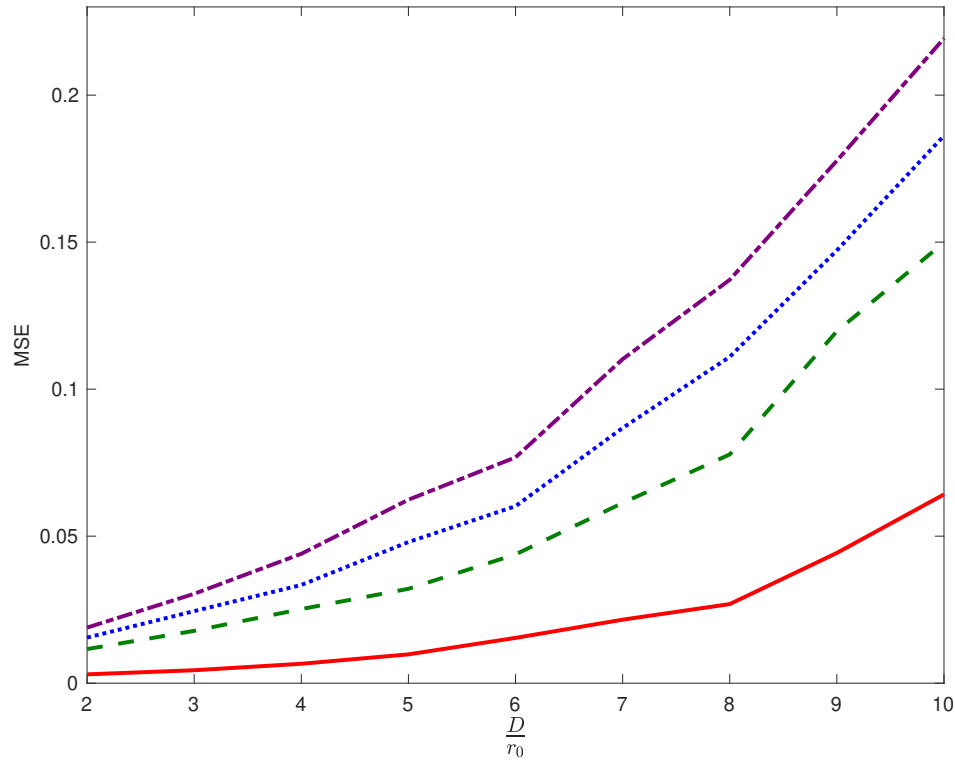


Fig. 5. The MSE as a function of D/r_0 for (—) a point source object with no noise, (---) extended source objects with no noise, (.....) extended source objects with low noise, and (-.-.-) extended source objects with high noise

the modified Zernike coefficients we tried to predict the signed Zernike coefficients for all polynomials, we only achieved an MSE of 0.5673. The results from Table 1 show that scenario 1 with a point source object and no noise produces the lowest average MSE whereas scenario 4 with the extended source objects set and high noise gives the largest MSE. This is expected because of the simplicity of the object and the lack of noise in scenario 1. The introduction of extended source objects without noise in scenario 2 causes an increase in the MSE compared to scenario 1. Scenario 3 uses the same extended source objects as scenario 2 but with low noise. The average MSE increases about 20% compared to scenario 2. The presence of high noise in scenario 4 with the extended source objects leads to a further increase of about 20% in MSE. Figure 5 shows the average MSE for predicting modified Zernike coefficients from the testing PSF intensity images as a function of D/r_0 ratios for the four scenarios. The MSE grows exponentially as a function of D/r_0 for all scenarios, and the ratios between the MSE values of the different scenarios remain relatively constant. Figure 6 illustrates the estimation results for a point source object with zero noise. The diffraction limited point source object is

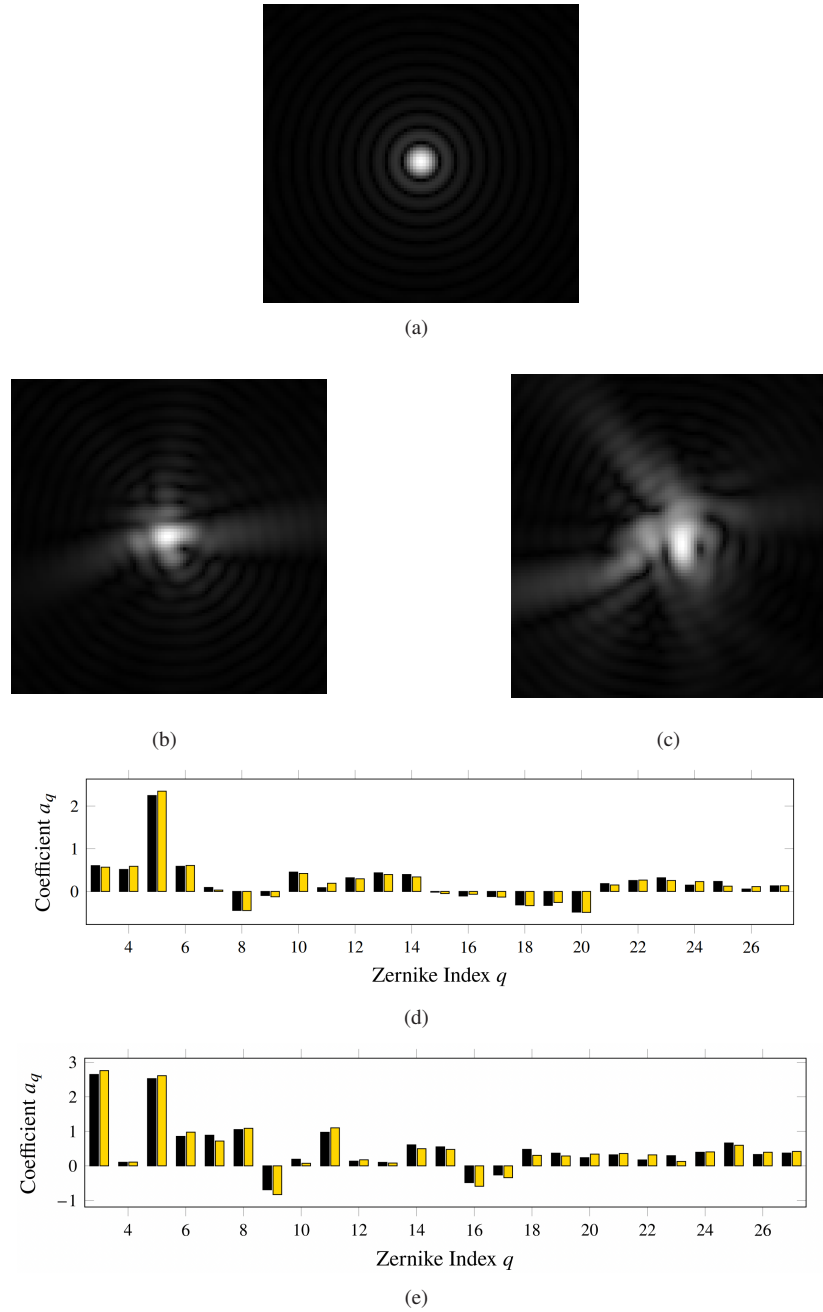


Fig. 6. Demonstration of the proposed estimation for a point source object with zero noise. (a) Diffraction limited PSF for the point source, (b) $I(x, y)$ for $D/r_0 = 2$, (c) $I(x, y)$ for $D/r_0 = 5$, (d) the actual (■) and predicted (■) Zernike coefficients corresponding to the intensity image in (b), and (e) the actual (■) and predicted (■) Zernike coefficients corresponding to the intensity image in (c). The contrast was enhanced for display in (a)-(c).

shown in Fig. 6(a) and the PSF intensity images $I(x, y)$ for that point source with $D/r_0 = 2, 5$ are shown in Fig. 6(b) and Fig. 6(c) respectively. Figure 6(d) and Fig. 6(e) display the actual and predicted Zernike coefficients corresponding to the intensity images in Fig. 6(b) and Fig. 6(c),

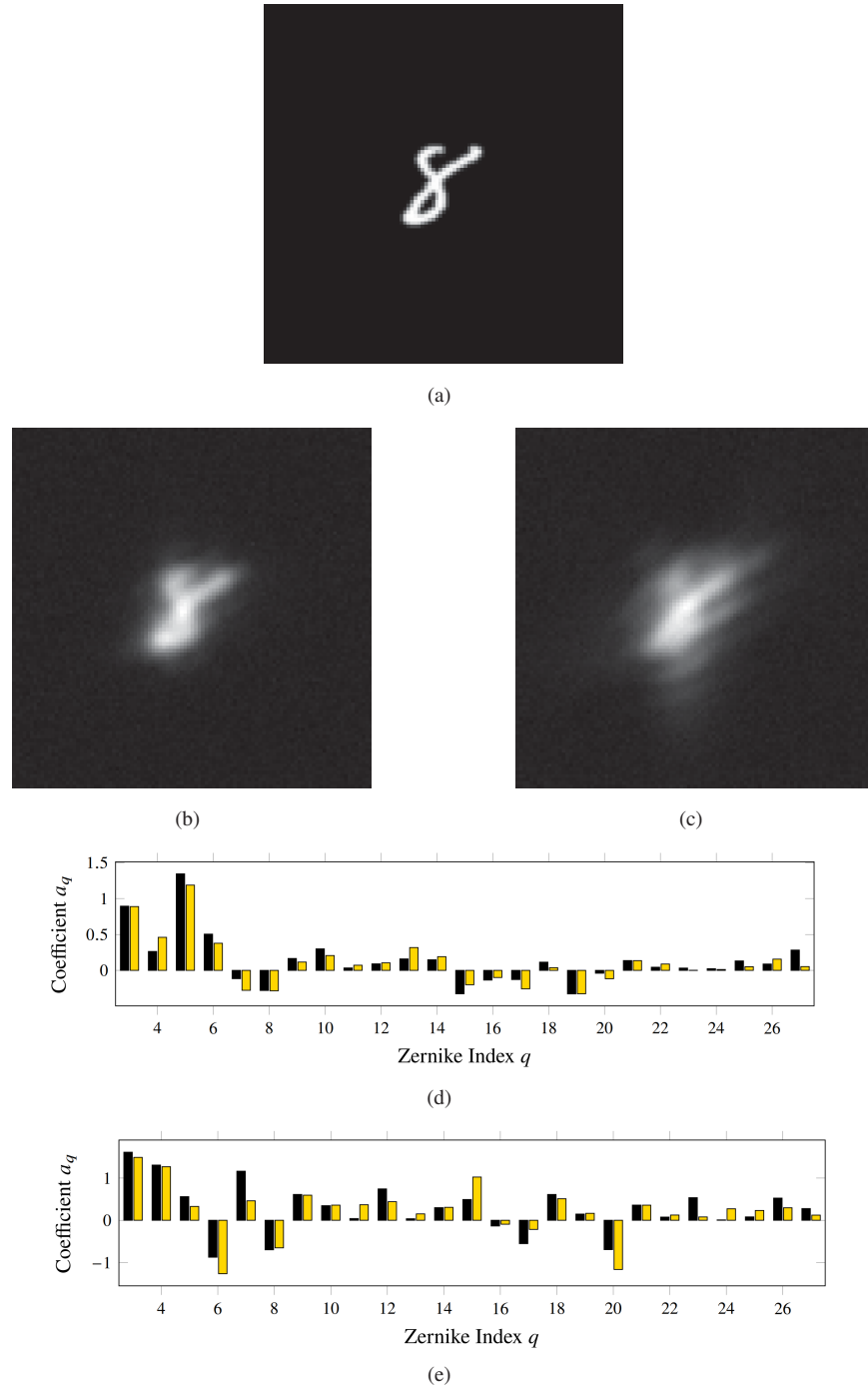


Fig. 7. Demonstration of the proposed estimation for an extended source object with high noise. (a) $I_0(x, y)$ for the the extended source, (b) $I(x, y)$ for $D/r_0 = 2$, (c) $I(x, y)$ for $D/r_0 = 5$, (d) the actual (■) and predicted (■) Zernike coefficients corresponding to the intensity image in (b), and (e) the actual (■) and predicted (■) Zernike coefficients corresponding to the intensity image in (c).

respectively. As can be seen in Fig. 6(d) and Fig. 6(e), the difference between the actual and predicted Zernike coefficients is small for both D/r_0 cases for the point source object at zero noise. Similarly, Fig. 7 illustrates the estimation results for an extended source object with high noise. From Fig. 7(d) and Fig. 7(e), it may be observed that the deviation between actual and predicted Zernike coefficients increases significantly for both D/r_0 cases for the extended source with high noise, compared to Fig. 6(d) and Fig. 6(e).

4. Conclusion

A deep learning model based on a convolutional neural network was trained to predict modified Zernike coefficients in the pupil of an imaging system from a single turbulence-degraded intensity image. The modified Zernike coefficient set differs from a conventional set in that an absolute value is assigned to coefficients of even radial orders due to a sign ambiguity associated with using the intensity image. The modified set was shown to be sufficient for specifying the intensity PSF. Data for the learning model was created with an image simulation of a point object and simple extended objects for a range of turbulence and detection noise levels. The prediction MSE for the learning model shows that it is possible to recover a useful set of modified Zernike coefficients from an extended object intensity image subject to noise and turbulence. As expected, the results show that the point source object with no noise produces the lowest average MSE whereas the extended source objects with high noise give the largest MSE. In all cases, the MSE increases in a predictable way with turbulence strength (D/r_0). Future work could explore the prediction of higher order Zernike terms and the use of more varied source objects. The quality and utility of the PSFs derived from the predicted Zernike coefficients were not investigated in this work and are essential topics for future efforts.

Funding. Office of Naval Research (N00014-21-1-2430).

Disclosures. The authors declare no conflicts of interest.

Data availability. Data underlying the results presented in this paper are not publicly available at this time but may be obtained from the authors upon reasonable request.

Supplemental document. See [Supplement 1](#) for supporting content.

References

1. J. Li, F. Xue, F. Qu, Y.-P. Ho, and T. Blu, "On-the-fly estimation of a microscopy point spread function," *Opt. Express* **26**(20), 26120–26133 (2018).
2. C. Lu, Q. Tian, L. Zhu, R. Gao, H. Yao, F. Tian, Q. Zhang, and X. Xin, "Mitigating the ambiguity problem in the cnn-based wavefront correction," *Opt. Lett.* **47**(13), 3251–3254 (2022).
3. P. Janout, P. Páta, P. Skala, and J. Bednář, "Psf estimation of space-variant ultra-wide field of view imaging systems," *Appl. Sci.* **7**(2), 151 (2017).
4. S. Jing and M. Dongmei, "A new method for comparing zernike circular polynomials with zernike annular polynomials in annular pupils," in *2010 International Conference on Computer, Mechatronics, Control and Electronic Engineering*, vol. 1 (IEEE, 2010), pp. 229–232.
5. Y. Jin, Y. Zhang, L. Hu, H. Huang, Q. Xu, X. Zhu, L. Huang, Y. Zheng, H.-L. Shen, W. Gong, and K. Si, "Machine learning guided rapid focusing with sensor-less aberration corrections," *Opt. Express* **26**(23), 30162–30171 (2018).
6. J. Schwiegerling, *Optical specification, fabrication, and testing* (SPIE, 2014).
7. V. Lakshminarayanan and A. Fleck, "Zernike polynomials: a guide," *J. Mod. Opt.* **58**(7), 545–561 (2011).
8. S. W. Paine and J. R. Fienup, "Machine learning for improved image-based wavefront sensing," *Opt. Lett.* **43**(6), 1235–1238 (2018).
9. Y. Nishizaki, M. Valdivia, R. Horisaki, K. Kitaguchi, M. Saito, J. Tanida, and E. Vera, "Deep learning wavefront sensing," *Opt. Express* **27**(1), 240–251 (2019).
10. Y. Zhang, C. Wu, Y. Song, K. Si, Y. Zheng, L. Hu, J. Chen, L. Tang, and W. Gong, "Machine learning based adaptive optics for doughnut-shaped beam," *Opt. Express* **27**(12), 16871–16881 (2019).
11. Q. Tian, C. Lu, B. Liu, L. Zhu, X. Pan, Q. Zhang, L. Yang, F. Tian, and X. Xin, "Dnn-based aberration correction in a wavefront sensorless adaptive optics system," *Opt. Express* **27**(8), 10765–10776 (2019).
12. L. Zhu, H. Yao, H. Chang, Q. Tian, Q. Zhang, X. Xin, and F. R. Yu, "Adaptive optics for orbital angular momentum-based internet of underwater things applications," *IEEE Internet Things J.* **9**(23), 24281–24299 (2022).
13. H. Guo, Y. Xu, Q. Li, S. Du, D. He, Q. Wang, and Y. Huang, "Improved machine learning approach for wavefront sensing," *Sensors* **19**(16), 3533 (2019).

14. Y. Li, D. Yue, and Y. He, "Prediction of wavefront distortion for wavefront sensorless adaptive optics based on deep learning," *Appl. Opt.* **61**(14), 4168–4176 (2022).
15. J. B. Shohani, M. Hajimahmoodzadeh, and H. Fallah, "Using a deep learning algorithm in image-based wavefront sensing: determining the optimum number of zernike terms," *Opt. Continuum* **2**(3), 632–645 (2023).
16. Y. Guo, L. Zhong, L. Min, J. Wang, Y. Wu, K. Chen, K. Wei, and C. Rao, "Adaptive optics based on machine learning: a review," *Opto-Electron. Adv.* **5**(7), 200082 (2022).
17. R. N. Bracewell, *The Fourier transform and its applications*, vol. 31999 (McGraw-Hill, 1986).
18. T. A. Underwood and D. G. Voelz, "Wave optics approach for incoherent imaging simulation through distributed turbulence," in *Unconventional Imaging and Wavefront Sensing 2013*, vol. 8877 (SPIE, 2013), pp. 112–119.
19. H. Zhan, E. Wijerathna, and D. Voelz, "Wave optics simulation studies of the fried parameter for weak to strong atmospheric turbulent fluctuations," in *Propagation Through and Characterization of Atmospheric and Oceanic Phenomena*, (Optica Publishing Group, 2019), pp. PM1C–3.
20. J. Schmidt, "Numerical simulation of optical wave propagation with examples in matlab," vol. PM199 (SPIE, 2010) (2010).
21. C. Wilcox, "Zernike polynomial coefficients for a given wavefront using matrix inversion in matlab," <https://www.mathworks.com/matlabcentral/fileexchange/27072-zernike-polynomial-coefficients-for-a-given-wavefront-using-matrix-inversion-in-matlab> (2023). MATLAB Central File Exchange [retrieved February 6, 2023].
22. T. Delabie, J. D. Schutter, and B. Vandenbussche, "An accurate and efficient gaussian fit centroiding algorithm for star trackers," *J. Astronaut. Sci.* **61**(1), 60–84 (2014).
23. R. J. Noll, "Zernike polynomials and atmospheric turbulence," *J. Opt. Soc. Am.* **66**(3), 207–211 (1976).
24. G. Cohen, S. Afshar, J. Tapson, and A. Van Schaik, "Emnist: Extending mnist to handwritten letters," in *2017 international joint conference on neural networks (IJCNN)*, (IEEE, 2017), pp. 2921–2926.
25. S. Hu, L. Hu, B. Zhang, W. Gong, and K. Si, "Simplifying the detection of optical distortions by machine learning," *J. Innovative Opt. Health Sci.* **13**(03), 2040001 (2020).
26. A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems* **25**, 1 (2012).
27. J. Bezanson, S. Karpinski, V. B. Shah, and A. Edelman, "Julia: A fast dynamic language for technical computing," *arXiv*, arXiv:1209.5145 (2012).
28. M. Innes, "Flux: Elegant machine learning with julia," *J. Open Source Softw.* **3**(25), 602 (2018).
29. D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv*, arXiv:1412.6980 (2014).