

# Machine learning recognition of light orbital-angular-momentum superpositions

B. Pinheiro da Silva,<sup>1,\*</sup> B. A. D. Marques,<sup>2,†</sup> R. B. Rodrigues,<sup>1,‡</sup> P. H. Souto Ribeiro,<sup>3,§</sup> and A. Z. Khoury<sup>1,¶</sup>

<sup>1</sup>*Instituto de Física, Universidade Federal Fluminense, 24210-346 Niterói, RJ, Brazil*

<sup>2</sup>*Universidade Federal Rural do Rio de Janeiro, 26285-060 Nova Iguaçu, RJ, Brazil*

<sup>3</sup>*Departamento de Física, Universidade Federal de Santa Catarina, 88040-900 Florianópolis, SC, Brazil*

(Dated: December 10, 2020)

We develop a method to characterize arbitrary superpositions of light orbital angular momentum (OAM) with high fidelity by using astigmatic transformation and machine learning processing. In order to define each superposition unequivocally, we combine two intensity measurements. The first one is the direct image of the input beam, which cannot distinguish between opposite OAM components. The second one is an image obtained using an astigmatic transformation that removes this ambiguity. Samples of these image pairs are used to train a convolution neural network and achieve high fidelity recognition of arbitrary OAM superpositions with dimension up to five.

It is well known that orbital angular momentum of light (OAM) [1] has many applications in different areas such as optical manipulation [2, 3], communications [4–8] and simulation [9–11]. Along with this broad range of applications, there is the requirement of being able to prepare and identify OAM single and superposition modes. To this end, interesting schemes can be realized with machine learning algorithms that have been employed to improve the state-of-the-art in computer vision and pattern recognition [12]. In particular, some applications concern structured light, including recognition and correction of OAM beams distorted by propagation through a turbulent medium [13–24], direct recognition of OAM modes [25, 26] and classification of vector vortex beams [27]. However, these previous investigations could not resolve OAM superpositions including positive and negative topological charges [28], which is crucial for quantum information applications.

Pure OAM beams are characterized by rotationally symmetric intensity distributions and an azimuthal variation of the phase given by a topological charge  $\ell \in \mathbb{Z}$ . Opposite OAM values cannot be distinguished by a single direct intensity measurement and one must resort to interferometric techniques [29, 30] or astigmatic transformations [31, 32] in order to distinguish between left- and right-handed beams. Moreover, determining the coefficients of arbitrary OAM superpositions, their weights and relative phases, is a difficult task. In fact, astigmatic transformations can be used to perform tomography of OAM qubits [33]. However, three distinct intensity measurements were required and the method was limited to OAM spaces of dimension two.

In this work, we combine astigmatic transformation and machine learning techniques to achieve high fidelity in the characterization of arbitrary superpositions in OAM spaces with dimensions up to five. Our method is based on two distinct intensity measurements: *i*) direct image of the input beam, *ii*) image of the converted beam after astigmatic transformation. We use a convolution neural network (CNN) to recover the weights and relative phases of the coefficients in the OAM superpo-

sition. The database used for training the CNN consists of theoretical and experimental images.

The transverse structure of paraxial beams propagating in free space can be described by Laguerre-Gaussian (LG) functions. For a beam with wave-number  $k$ , propagating along the  $z$  axis, the LG function reads

$$\text{LG}_{\ell,p}(r, \theta) = \frac{\mathcal{N}_{\ell,p}}{w} \tilde{r}^{|\ell|} L_p^{|\ell|}(\tilde{r}) e^{-\frac{\tilde{r}^2}{2}} e^{i\ell\theta} e^{-i\Phi_N}, \quad (1)$$

$$\Phi_N = \frac{k r^2}{2R} + (N + 1) \arctan(z/z_0), \quad \tilde{r} = \sqrt{2} r/w,$$

where  $N = 2p + |\ell|$  is the mode order,  $\ell$  is the topological charge,  $p$  the radial number,  $L_p^{|\ell|}$  are generalized Laguerre polynomials and  $\mathcal{N}_{\ell,p}$  is a normalization constant. The beam parameters are the wave-front radius  $R$ , the width  $w$  and the Rayleigh length  $z_0$ , which also characterize the Hermite-Gaussian modes

$$\text{HG}_{m,n}(x, y) = \frac{\mathcal{N}_{mn}}{w} H_m(\tilde{x}) H_n(\tilde{y}) e^{-\frac{\tilde{x}^2 + \tilde{y}^2}{2}} e^{-i\Phi_N}, \quad (2)$$

$$\tilde{x} = \sqrt{2} x/w, \quad \tilde{y} = \sqrt{2} y/w,$$

where  $\mathcal{N}_{mn}$  is the proper normalization constant and the HG mode order is  $N = m + n$ .

Both the LG and HG modes constitute orthonormal and complete bases of the transverse mode vector space. This space can be cast as a direct sum of subspaces related to the different mode orders. For a given order  $N$ , it is possible to define a subspace of dimension  $D = N + 1$  analogous to the Hilbert space of a qudit, which is a quantum  $D$ -level system. LG and HG modes up to order  $N$  are vectors in this space and are connected to each other by unitary transformations. In this sense, an arbitrary transverse mode qudit can be written as a superposition of LG modes of order  $N$  according to

$$|\psi\rangle_D = \sum_{\ell,p} c_{\ell,p} |\text{LG}_{\ell,p}\rangle, \quad (3)$$

where the summation runs over indices  $\ell$  and  $p$  restricted by  $2p + |\ell| = D - 1$  and  $c_{\ell,p}$  is a complex weight.

In view of the potential use of these superpositions in applications, their recognition is an essential task that can be approached in many ways, starting from the analysis of their intensity patterns. However, pure LG modes with the same values of  $p$  and  $|\ell|$  have identical intensity distributions, so it is impossible to distinguish them from a direct intensity measurement only. In fact, this problem is more general because any superposition of the type defined by Eq.(3) is subjected to the following symmetry condition

$$\left| \sum_{\ell,p} c_{\ell,p} \text{LG}_{\ell,p}(\mathbf{r}) \right|^2 = \left| \sum_{\ell,p} c_{-\ell,p} \text{LG}_{\ell,p}(\mathbf{r}) \right|^2. \quad (4)$$

This degeneracy can be lifted by supplementing the direct measurement with a second image obtained from astigmatic mode conversion of the input beam [31, 33]. The mode converter acts as a unitary transformation restricted to each mode order subspace. Therefore, it can be written as the direct sum of  $SU(D)$  operators:

$$MC = \bigoplus_D MC_D, \quad (5)$$

$$MC_D = \sum_{m=0}^{D-1} e^{i(m-n)\frac{\pi}{4}} |\text{HG}_{m,n}\rangle \langle \text{HG}_{m,n}|,$$

with  $n = D - m - 1$ . In Eq.(5) we made use of the Hermite-Gaussian base vectors  $\{|\text{HG}_{m,n}\rangle\}$  which are the eigenmodes of the astigmatic transformation. In Fig.1 the astigmatic method is illustrated for 4 different degenerate patterns. Images a<sub>1</sub>) and a<sub>2</sub>) display two degenerate intensity plots associated with the following superpositions

$$\begin{aligned} |\psi_1\rangle &= 0.92 |\text{LG}_{+1,0}\rangle + 0.38 |\text{LG}_{-1,0}\rangle, \\ |\psi_2\rangle &= 0.38 |\text{LG}_{+1,0}\rangle + 0.92 |\text{LG}_{-1,0}\rangle. \end{aligned} \quad (6)$$

The degeneracy is lifted by the mode converted images b<sub>1</sub>) and b<sub>2</sub>). We also illustrate the method with more complex patterns, such as those exhibited in images a<sub>3</sub>) and a<sub>4</sub>), corresponding to

$$\begin{aligned} |\psi_3\rangle &= \frac{1}{2} (|\text{LG}_{+3,0}\rangle - |\text{LG}_{-3,0}\rangle + |\text{LG}_{+1,1}\rangle + |\text{LG}_{-1,1}\rangle), \\ |\psi_4\rangle &= \frac{1}{2} (-|\text{LG}_{+3,0}\rangle + |\text{LG}_{-3,0}\rangle + |\text{LG}_{+1,1}\rangle + |\text{LG}_{-1,1}\rangle). \end{aligned} \quad (7)$$

Although they exhibit the same direct image pattern, they can be resolved by the mode converted images shown in b<sub>3</sub>) and b<sub>4</sub>). Therefore, the two images are sufficient to define the mode superposition unequivocally, since they are capable of resolving opposite OAM states.

The brute-force method for identifying a given superposition of modes, would be the following: i) perform the measurement of the intensity patterns of direct and mode-converted image; ii) generate theoretical intensity

patterns with tentative mode superpositions and compare with the measured patterns; iii) use some optimization procedure to obtain the superposition that best approaches the measurement. However, more efficient strategies are available and we will demonstrate the use of machine learning to improve the recognition method.

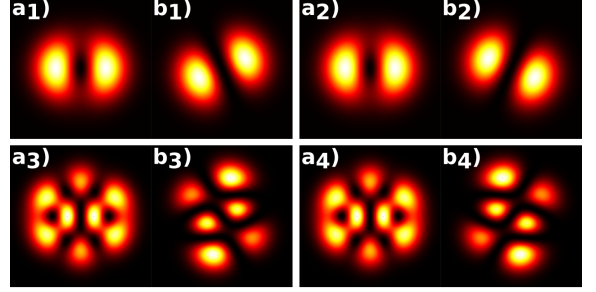


FIG. 1. Theoretical intensity images of four different superpositions ( $|\psi_1\rangle, |\psi_2\rangle, |\psi_3\rangle, |\psi_4\rangle$ ): a) are the direct images of the superpositions. b) are the images after the tilted lens.

We test these ideas experimentally. The experimental setup is shown in Fig. 2. A Gaussian beam from a Nd:YAG laser ( $\lambda = 1064 \text{ nm}$ ) is sent to a spatial light modulator (SLM) programmed to produce an arbitrary mode superposition within the subspace of order  $N$ . The beam splitter (BS) is used to split the incoming beam, transmitting half of the intensity to the spherical lens  $L_a$  ( $f = 1 \text{ m}$ ) and forming the direct image in the CCD (charge-coupled device) camera. The reflected beam having the other half of the intensity is sent to mirror ( $M$ ) and then to the tilted lens  $L_b$  ( $f = 1 \text{ m}$ ) that performs the astigmatic transformation. The beams are acquired in a single frame by the camera positioned at a distance of  $0.74 \text{ m}$  from both lenses.

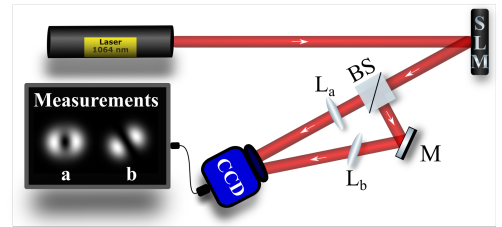


FIG. 2. Experimental setup.

To recover the complex weights  $c_{\ell,p}$  of a given superposition, we employ a deep learning method denominated convolutional neural network (CNN), which is appropriated for image processing. Unlike traditional machine learning algorithms, the CNN can automatically select and extract key-features of images to solve pattern recognition tasks. This representation-learning ability [34], combined with the available processing power of modern graphics processing units (GPUs), allows the usage of a large number of images to construct a robust recognition

system.

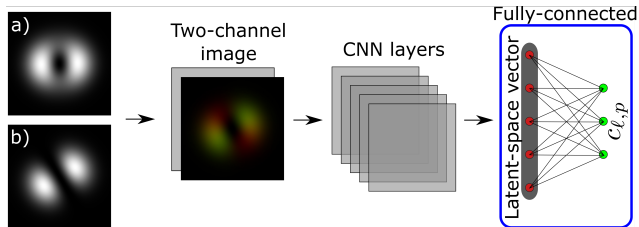


FIG. 3. Superposition recognition system. The system receives two intensity images (a, b) and combines them in a two-channel image fed to the CNN. The CNN extracts a latent-space feature vector that is used to estimate the weights  $c_{l,p}$  of the superposition.

The recognition system developed in this work takes the two input images of the mode superposition (direct and transformed), and outputs the coefficients  $c_{l,p}$  of Eq. (3). We employ a 34-layer CNN that uses a series of convolutions and non-linear functions to extract the images' features. Then, these features are stored into a latent-space vector representing the most relevant features for recognition of the mode superposition. The latent-space vector is fed to the estimator (a fully-connected layer) that outputs a vector representing the values of the  $c_{l,p}$ . Fig. 3 shows the overall representation of our system.

The architecture of our CNN is based on the residual neural network [35]. We use 16 residual blocks totaling 32 convolution layers with  $(3 \times 3)$  kernel-size, an initial convolution layer with  $(7 \times 7)$  kernel-size, and a fully connected layer with 512 units. We state the recognition problem as a regression task, in which the CNN estimates a numerical value for the  $c_{l,p}$ .

Typically, CNNs are trained using big datasets [36, 37]. In a recent work [28] superpositions of LG modes with  $p = 0$  and  $0 \leq \ell \leq 9$  used  $10^5$  theoretical samples, which is a relatively small dataset. However, all topological charges have the same sign, which facilitates the mode recognition. We choose to use the same order of magnitude for the datasets in our experiments. For each dimension  $D$ , we produce a dataset with  $(D - 1) \times 10^4$  experimental samples of arbitrary superpositions. Moreover, we include both theoretical and experimental samples. The dataset is split into three parts: 75% for the training, 15% for the validation, and 10% for the test.

In Machine learning, the CNN model is a function composed of convolution operations that tries to fit the data based on previous examples during the training process. The CNN training process consists of providing examples from the training dataset and adjusting the CNN's parameters regarding a loss function. A training epoch is defined by the processing of all the examples in the training dataset. We validate the training epoch by evaluating the examples from the validation dataset. To ensure the CNN model's generality, we test the trained model with

the testing dataset, which consists of images never used before during previous training and validation steps.

Our CNN initializes with randomly sampled parameters. The network is trained for 100 epochs or until convergence using the Adam optimizer [38] with a learning rate of 0.001. We consider that the model converges if the validation loss does not improve after 40 epochs. The loss function employed for training the CNN is  $1 - \mathcal{F}$ , where  $\mathcal{F} = |\langle \psi_{D_G} | \psi_{D_E} \rangle|^2$  is the fidelity between the ground truth state  $|\psi_{D_G}\rangle$  and the estimated state  $|\psi_{D_E}\rangle$ .

Figure 4 shows the training evolution through the epochs for the model  $D = 2$ . The value of validation and training mean fidelity improves consistently across epochs, without indication of overfitting [39]. The model maintains a relatively stable value after epoch 40. This behavior is similar across all the trained models.

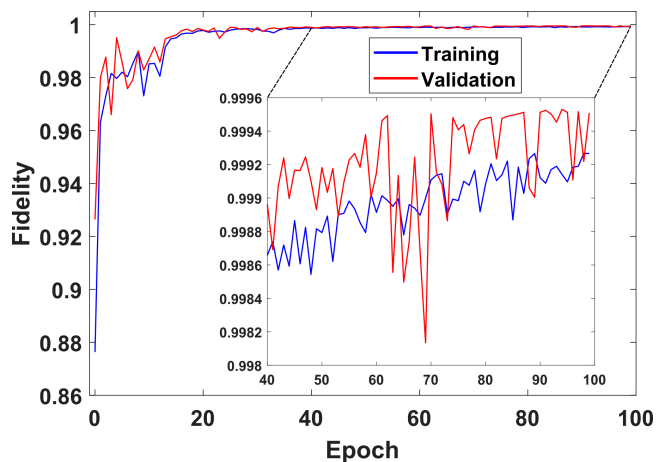


FIG. 4. Mean fidelity in each epoch for training and validation process in the case  $D = 2$ .

Our experimental setup allows the acquisition of a large number of images for the training dataset. However, this fact does not always hold true for other experiments. Hence it is desirable to estimate the minimum number of experimental images, in order to train the network while the model still attains a prediction with acceptable mean fidelity. To do so, we train models for  $D = 2$ , with training and validation datasets composed of 9000 samples. For this approach, we generate theoretical samples through a computer-generated simulation.

Initially, we populate the entire dataset with theoretical samples. Then, we incrementally add experimental samples, maintaining a fixed dataset size. The mean fidelity of the superpositions inferred by each CNN models, as we increase the proportion of experimental samples, is shown in Fig. 5. We tested all the models against the same testing dataset of 1000 experimental samples.

If we use only theoretical images in the CNN's training and validation process, we obtain low mean fidelity and a significant standard deviation ( $0.5 \pm 0.3$ ). Therefore, we observed that it is essential to use experimental images in

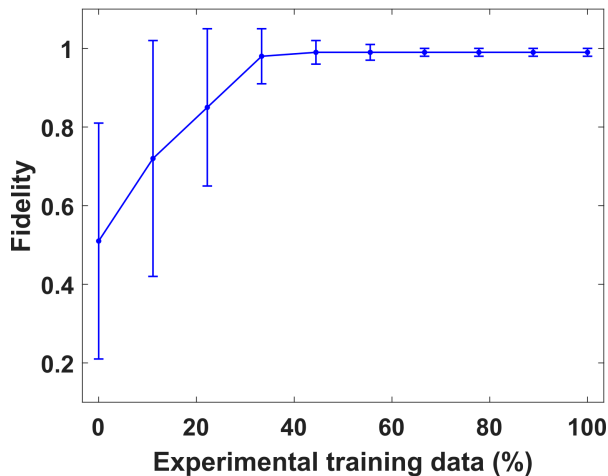


FIG. 5. Mean fidelity as a function of the percentage of experimental train data.

the training of the network. We found that the minimum proportion of experimental data is 44% for this specific problem. With this proportion, the model achieves a mean fidelity of  $0.99 \pm 0.03$ .

At a proportion of 66% experimental images, the model achieve the fidelity of  $0.99 \pm 0.01$ . Above this percentage we have minimal gains in the performance of the model. We speculate that a the amount of data in our experimental dataset contains sufficient information for training the network even when only 66% of the experimental samples are used. The result of mixing experimental and theoretical samples shows that it is possible to reach a reasonable estimation even for cases where it is not feasible to obtain a higher number of experimental samples. Since our experimental setup allows fast acquisition of samples, we choose to use exclusively experimental images in this work.

To verify our method, we train and test several CNN models, one for each dimension  $D = 2, 3, 4, 5$ , as shown in table I. In this test, we inform the dimension of the mode superposition in order to select the correct CNN model. The mean fidelity concerning all superpositions in the testing dataset is calculated and the standard deviation gives the estimation error. We obtained a high mean fidelity value for all the dimensions analyzed. The error slightly increases for dimensions higher than 3. The model is capable of performing the recognition in real-time, with an inference time of  $0.9\text{ms}$  for a single superposition using a consumer-grade GPU (NVIDIA<sup>®</sup> Geforce<sup>®</sup> RTX 2080 Super<sup>™</sup>).

Quite remarkably, our system is also capable to determine the dimension of the mode superposition. The recognition of a superposition with an arbitrary, unknown order ( $D \leq 5$ ) is demonstrated. First, the system tests the input experimental image against all the trained models. Then, the value of the estimated  $c_{l,p}$

Superposition dimension	Testing dataset	Mean fidelity
2	1000	$0.99 \pm 0.01$
3	2000	$0.99 \pm 0.01$
4	3000	$0.99 \pm 0.02$
5	4000	$0.99 \pm 0.03$

TABLE I. Testing dataset size and mean fidelity for each dimension.

for each model is employed to generate the theoretical images. The superposition order is determined by comparing the theoretical images with the input image and selecting the model for which the inference produces the theoretical image that is most similar to the experimental input. We perform a blind test of the system by not informing the superposition order. The testing dataset contains 250 superpositions of each order ( $D = 2, 3, 4, 5$ ), amounting to 1000 samples. In this test, the recognition system gives the superposition coefficients and the dimension as outputs. The system achieved an accuracy of 99.7% for the dimension estimation, and a mean fidelity of  $0.99 \pm 0.02$ .

In conclusion, we developed a tomographic method for the characterization of OAM superpositions based on two measurements and processed via machine learning. To define each superposition unequivocally, we perform two intensity measurements; the first is the direct image and the second is the image after applying an astigmatic transformation with a tilted lens. Once we have the two images, we use a convolutional neural network to recover the superposition coefficients. As we have shown, in cases where the experimental setup has limitations, it is possible to use theoretical images to increase the dataset. Nevertheless, to obtain a reasonable mean fidelity with a tolerable error, the minimum percentage of experimental images in the total training dataset is 44%. Our method was tested for  $D = 2, 3, 4, 5$  using the experimental dataset. The results exhibit a high mean fidelity and low error, demonstrating that our model is reliable in different OAM space dimensions. In the last test, we did not inform the superposition dimension for the recognition system. Still, the method proved to be highly accurate, providing outstanding fidelity values and precise estimation of the dimension. Our method has a fast inference time, effectively enabling real-time recognition of superpositions with arbitrary order.

Finally, our method can be adapted to the quantum regime with a single-photon sensitive camera (such as an intensified or electron multiplying charge coupled device). Although image reconstruction requires a large number of photons to be gathered, the method can be supplemented by compression techniques to reconstruct images from a small number of detected photons per pixel. Moreover, the use of heralded single-photon sources can further improve the signal-to-noise ratio.

## ACKNOWLEDGMENTS

Funding was provided by Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Fundação Carlos Chagas Filho de Amparo à Pesquisa do Estado do Rio de Janeiro (FAPERJ), Fundação de Amparo à Pesquisa do Estado de Santa Catarina (FAPESC), Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), and Instituto Nacional de Ciência e Tecnologia de Informação Quântica (INCT-IQ 465469/2014-0).

\* braianps@gmail.com

† brunodortamarques@gmail.com

‡ rafaelbellasrodrigues@gmail.com

§ p.h.s.ribeiro@ufsc.br

¶ azkhoury@id.uff.br

- [1] M. J. Padgett, *Opt. Express* **25**, 11265 (2017).
- [2] M. Padgett and R. Bowman, *Nature Photonics* **5**, 343 (2011).
- [3] M. Gecevičius, R. Drevinskas, M. Beresna, and P. G. Kazansky, *Applied Physics Letters* **104**, 231110 (2014).
- [4] C. E. R. Souza, C. V. S. Borges, A. Z. Khoury, J. A. O. Huguenin, L. Aolita, and S. P. Walborn, *Phys. Rev. A* **77**, 032345 (2008).
- [5] V. D'Ambrosio, E. Nagali, S. P. Walborn, L. Aolita, S. Slussarenko, L. Marrucci, and F. Sciarrino, *Nature Communications* **3**, 961 (2012).
- [6] F. Tamburini, E. Mari, A. Sponselli, B. Thidé, A. Bianchini, and F. Romanato, *New Journal of Physics* **14**, 033001 (2012).
- [7] B. P. da Silva, M. A. Leal, C. E. R. Souza, E. F. Galvão, and A. Z. Khoury, *Journal of Physics B: Atomic, Molecular and Optical Physics* **49**, 055501 (2016).
- [8] C. E. R. Souza and A. Z. Khoury, *Opt. Express* **18**, 9207 (2010).
- [9] R. M. de Araújo, T. Häffner, R. Bernardi, D. S. Tasca, M. P. J. Lavery, M. J. Padgett, A. Kanaan, L. C. Céleri, and P. H. S. Ribeiro, *Journal of Physics Communications* **2**, 035012 (2018).
- [10] P. H. S. Ribeiro, T. Häffner, G. L. Zanin, N. R. da Silva, R. M. de Araújo, W. C. Soares, R. J. de Assis, L. C. Céleri, and A. Forbes, *Phys. Rev. A* **101**, 052113 (2020).
- [11] T. Häffner, G. L. Zanin, R. M. Gomes, L. C. Céleri, and P. H. S. Ribeiro, *The European Physical Journal Plus* **135**, 601 (2020).
- [12] Y. LeCun, Y. Bengio, and G. Hinton, *Nature* **521**, 436 (2015).
- [13] M. Krenn, J. Handsteiner, M. Fink, R. Fickler, R. Ursin, M. Malik, and A. Zeilinger, *Proceedings of the National Academy of Sciences* **113**, 13648 (2016).
- [14] M. Krenn, R. Fickler, M. Fink, J. Handsteiner, M. Malik, T. Scheidl, R. Ursin, and A. Zeilinger, *New Journal of Physics* **16**, 113028 (2014).
- [15] S. Lohani and R. T. Glasser, *Opt. Lett.* **43**, 2611 (2018).
- [16] J. Li, M. Zhang, and D. Wang, *IEEE Photonics Technology Letters* **29**, 1455 (2017).
- [17] J. Li, M. Zhang, D. Wang, S. Wu, and Y. Zhan, *Opt. Express* **26**, 10494 (2018).
- [18] S. R. Park, L. Cattell, J. M. Nichols, A. Watnik, T. Doster, and G. K. Rohde, *Opt. Express* **26**, 4004 (2018).
- [19] T. Doster and A. T. Watnik, *Appl. Opt.* **56**, 3386 (2017).
- [20] N. Bhusal, S. Lohani, C. You, M. Hong, J. Fabre, P. Zhao, E. M. Knutson, R. T. Glasser, and O. S. Magana-Loaiza, (2020), arXiv:2006.07760.
- [21] M. I. Dedo, Z. Wang, K. Guo, and Z. Guo, *Optics Communications* **456**, 124696 (2020).
- [22] Z. Wang, M. I. Dedo, K. Guo, K. Zhou, F. Shen, Y. Sun, S. Liu, and Z. Guo, *IEEE Photonics Journal* **11**, 1 (2019).
- [23] Z. Huang, P. Wang, J. Liu, W. Xiong, Y. He, X. Zhou, J. Xiao, Y. Li, S. Chen, and D. Fan, *Results in Physics* **15**, 102790 (2019).
- [24] S. Lohani, E. M. Knutson, M. O'Donnell, S. D. Huver, and R. T. Glasser, *Appl. Opt.* **57**, 4180 (2018).
- [25] S. Sharifi, Y. M. Banadaki, G. Veronis, and J. P. Dowling, *Optical Engineering* **59**, 1 (2020).
- [26] A. M. Palmieri, E. Kovlakov, F. Bianchi, D. Yudin, S. Straupe, J. D. Biamonte, and S. Kulik, *npj Quantum Information* **6**, 20 (2020).
- [27] T. Giordani, A. Suprano, E. Polino, F. Acanfora, L. Innocenti, A. Ferraro, M. Paternostro, N. Spagnolo, and F. Sciarrino, *Phys. Rev. Lett.* **124**, 160401 (2020).
- [28] M. Sheikh, *Spatial optical mode decomposition using deep learning (Applications of Machine Learning 2020, 2020)* pp. 105 – 111.
- [29] A. D'Errico, R. D'Amelio, B. Piccirillo, F. Cardano, and L. Marrucci, *Optica* **4**, 1350 (2017).
- [30] J. M. Knudsen, S. N. Alperin, A. A. Voitiv, W. G. Holtzmann, J. T. Gopinath, and M. E. Siemens, (2017), arXiv:https://arxiv.org/abs/1710.02912.
- [31] M. W. Beijersbergen, L. Allen, H. Van der Veen, and J. Woerdman, *Optics Communications* **96**, 123 (1993).
- [32] P. Vaity, J. Banerji, and R. Singh, *Physics letters a* **377**, 1154 (2013).
- [33] B. P. da Silva, D. S. Tasca, E. F. Galvão, and A. Z. Khoury, *Phys. Rev. A* **99**, 043820 (2019).
- [34] Y. Bengio, A. Courville, and P. Vincent, *IEEE transactions on pattern analysis and machine intelligence* **35**, 1798 (2013).
- [35] K. He, X. Zhang, S. Ren, and J. Sun, *Deep residual learning for image recognition (Proceedings of the IEEE conference on computer vision and pattern recognition, Las Vegas, 2016)* pp. 770–778.
- [36] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, *Imagenet: A large-scale hierarchical image database (2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, 2009)* pp. 248–255.
- [37] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, *Microsoft coco: Common objects in context (Computer Vision – ECCV 2014, Cham, 2014)* pp. 740–755.
- [38] D. P. Kingma and J. Ba, *Adam: A method for stochastic optimization (3rd International Conference on Learning Representations, San Diego, 2015)*.
- [39] Y. B. Ian Goodfellow and A. Courville, *Deep Learning* (MIT Press, Cambridge, 2016).