Single-Pixel, Single-Frequency Hand Gesture Recognition with a Dynamic Metasurface Aperture and Deep Neural Network

Oren S. Mizrahi
Department of Electrical & Computer Engineering
Duke University
Durham, NC 27708
oren.mizrahi@duke.edu

Aaron V. Diebold
Department of Electrical & Computer Engineering
Duke University
Durham, NC 27708
aaron.diebold@duke.edu

Abstract

In this paper, we demonstrate a novel system for classifying static hand gestures using single-pixel, single-frequency signals obtained with a dynamic metasurface aperture (DMA). DMAs are printed cavities excited by a single port with tunable radiating metamaterial elements. By altering the binary pattern of active elements (called masks), we generate a large number of diverse radiation patterns which serve as a quasi-orthogonal basis for single-pixel, single-frequency sensing. We took measurements of 10 hand gestures using 50 random masks. Instead of solving the inverse system, we trained a deep neural network (DNN) to classify hand gestures. Using a random subset of only 28 masks, we were able to classify hand gestures with >97% accuracy. To further reduce the number of masks necessary to obtain high classification accuracy, we attempted to optimize the physical layer by specifically designing masks in simulation, instead of picking them at random. We discuss simplified scenarios where this technique is applicable, as well as justification for the lack of significant improvement. The success of this system indicates the potential for low-cost, low-power sensing systems which can read human motion, hand movements, or static patterns for the purposes of human-device interaction.

1 Introduction

Computational imaging and sensing schemes have gained traction recently as methods of producing images or signals which multiplex the scene without introducing complex and costly hardware [1–9]. Unlike traditional imaging schemes, where the relationship between points in the scene and detection pixels is bijective, computational schemes can multiplex various points in the scene to a single detection pixel. Many solutions have been devised to achieve such a system, among them being the use of frequency [7, 10, 11], spatial diversity [12, 13], and pattern diversity [14, 15].

The use of bandwidth requires complex and costly hardware and is heavily limited by FCC regulations which set limits on the amount of bandwidth and the frequency of operation. Spatial diversity does not suffer from the same frequency regulations and may be capable of the required multiplexing, but usually requires a large, expensive, and cumbersome array of antennas.

In light of these constraints, an alternative approach to multiplexing the scene is achieving spatial pattern diversity with aperture modulation. This is a low-cost, single-frequency method which has had much success recently for the purposes of computational imaging [16, 17] and sensing at microwave frequencies. In this scheme, an antenna varies its radiation pattern such that different elements in the scene are multiplexed into a single return signal at a simple dipole receiver. By switching between diverse radiation patterns, we are able to capture different linear combinations of scatterers in the scene.

In this paper, we propose the use of a dynamic metasurface aperture (DMA) as the antenna which generates spatial pattern diversity for a hand gesture classification system. The aperture is a simple and low-cost printed circuit board (PCB) with a single feed and 96 radiating metamaterial elements which can be individually tuned [18]. The elements are placed "randomly" such that any binary tuning pattern (referred to in this paper as a mask) corresponds to a distinct and spatially diverse radiation pattern. By cycling through a number of masks, we generate a basis with which to multiplex the scene, a method which has seen much success in recent years [16, 17, 19–22].

We leveraged the multiplexing advantages [21] of a DMA to classify static hand gestures. Our antenna system consisted of a DMA transmitter beside a dipole antenna receiver. We began by generating our data set: we placed a model hand on a rotation stage, 20 cm from the Tx-Rx pair. By altering the hand gesture, the exact finger position, and the angle of the hand relative to the antennae, we generated a data set of (12810) measurements. For each measurement, we cycled through 50 randomly generated masks and recorded the response of the hand at the receiver.

We trained a DNN using differently sized random subsets of the original 50 masks from $2 - 50$ masks and achieved accuracies of $> 90\%$; with 28 masks, 97% accuracy was achieved. We additionally demonstrate magnitude-only training and classification, achieving 96% classification accuracy using 42 masks. The success of this system demonstrates the potential for the DMA to serve as an antenna for sensing schemes which rely on neural network classification.

2 System Design

2.1 Physical

Our physical setup begins with a Tx-Rx pair connected to a vector network analyzer (VNA) set at 19.21GHz. We used a DMA as the transmitter and a normal dipole antenna mounted beside the DMA as the receiver. The DMA is controlled independent of the VNA using an Arduino back-end to tune the metamaterial elements and switch masks.

The Tx-Rx pair is aimed at a flexible wooden hand with the necessary finger joints and wrist movement to replicate most of the full range of human hand movement. The wooden hand was mounted to a rotation stage to enable data collection at a range of hand orientations relative to the antennae. To replicate the scattering properties of a human hand at k-Band, we spray-painted the hand with electrically reflective nickel-based paint. We avoided painting joint areas so as not to limit the hand's flexibility. The full setup is depicted in Figure 1(a).

2.2 Data Collection

We began by defining a set of $N_{\text{gest}} = 10$ "normal" hand gestures (index finger point, rock-on, fist, etc.), shown in Figure 1(b). For each hand-gesture, we slightly altered the hand orientation and the relative position of each finger $N_{\text{alter}} = 21$ times so as to generate a variety of gestures for use in training and testing. For each alteration, we cycled the measurement procedure over $N_\theta = 61$ angles $\left(\theta \in \left[-\frac{\pi}{6}, \frac{\pi}{6}\right]\right)$ using the rotation stage. We did not take data for $|\theta| > \frac{\pi}{6}$ because these angles would be too extreme and it may not be possible for our system to distinguish between hand gestures when parts of the hand are shielded from view. At each θ, we took single-frequency measurements at 19.21 GHz for $N_{\text{masks}} = 50$ masks, considered here as a single measurement. Thus, the quantity of data with which to train and test is calculated below:

$$N_{\text{meas}} = N_{\text{gest}} * N_{\text{alter}} * N_\theta = 10 * 21 * 61 = 12810 \tag{1}$$

The complex values for 50 masks for a single measurement is shown in Figure 2(a).

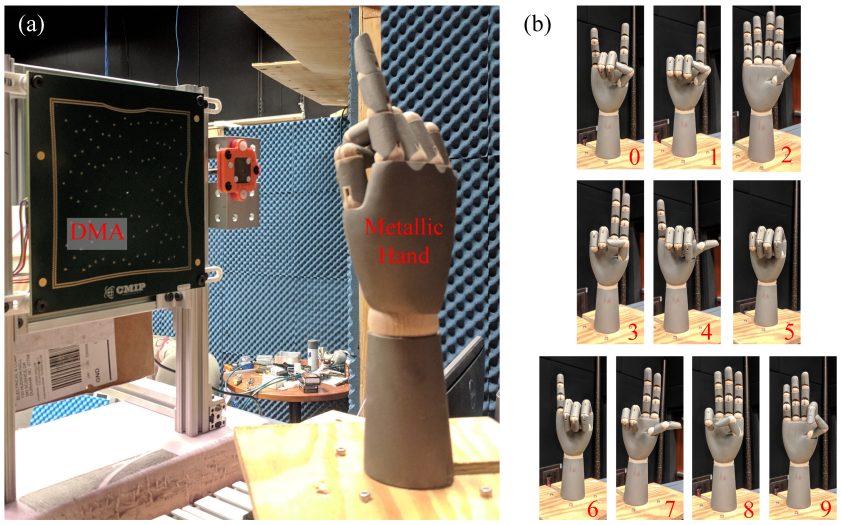

Figure 1: (a) Our experimental setup. (b) The hand gestures included in our classification experiment.

Table 1: DNN Architecture

Type	Output Size	Activation
Flat	$2N_{\text{masks}}$	-
Dense	128	ReLU
Dense	128	ReLU
Dense	128	ReLU
Dense	N_{gest}	SoftMax

2.3 Network Architecture

The DNN used for this system is described in Table 1. Before feeding into the network, our complex data was split into real and imaginary components; these data were stacked into a single vector. Thus, the flat layer has size $2N_{\text{masks}}$.

In total, there were 47242 parameters for $N_{\text{masks}} = 50, N_{\text{gest}} = 10$ gestures. In cases where we trained the network with fewer masks, there were slightly fewer parameters because of a smaller first layer.

3 Experimental Results

We trained the network shown in Table 1 multiple times using N_{used} masks, where N_{used} was an even integer between 2 and 50. Masks were randomly ordered; for each iteration the top N_{used} were used to train the network. For the most part, as N_{used} increased, the validation accuracy did as well. Using only 18, we achieved 94%; using 10 additional masks, we were able to achieve 97% accuracy. A full plot of accuracy vs. number of masks is given in Figure 2(b); the confusion matrix is shown in Figure 2(c).

4 Physical Layer Optimization

The experimental results presented above were obtained using pseudorandom patterns, corresponding to randomly-selected binary tuning states of the 96 metamaterial elements. This physical configuration paired with the above DNN architecture may result in unsatisfactory classification accuracy for more complicated tasks. In addition, efficiency in the acquisition and post-processing computational stages is generally desirable in order to realize practical and real-time performance, motivating the improvement of classification accuracy for fewer measurements. Often, this is achieved through

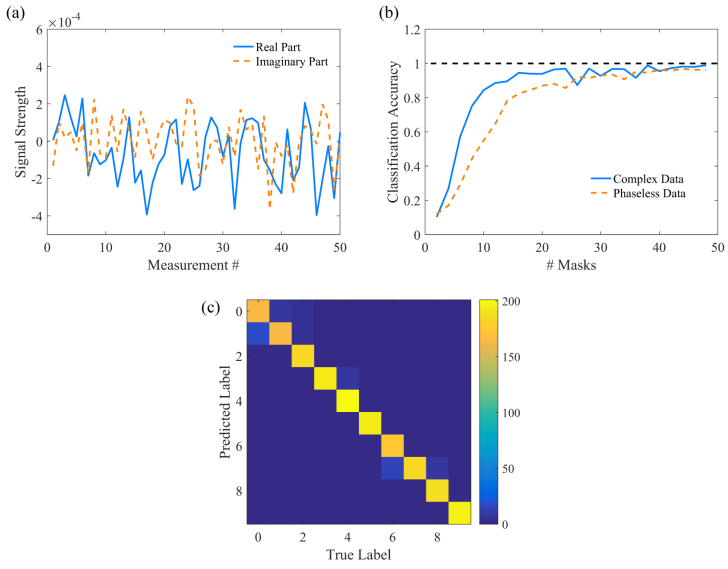

Figure 2: (a) Experimentally-acquired complex signal used as input to neural network. (b) Experimental classification accuracy of 10 different hand gestures, using complex (phase and magnitude) and phaseless (magnitude only) data. (c) Confusion matrix for gesture classification.

optimization of the NN architecture [23]. Alternatively, as discussed in [24, 25], we can seek to optimize the physical acquisition system, preferably directly in the NN itself. This approach achieves task-driven optimization of the combined physical-computational system. In the following, we outline how optimization of the DMA illumination patterns can be prescribed from the measured data under certain assumptions, and demonstrate the procedure in simulation.

A single complex measurement g_n corresponding to the n^{th} DMA tuning state, assuming scalar fields and the first Born approximation, results from the coherent superposition of the illuminating DMA fields $E_n^T(\vec{r})$ with the scattering density $\sigma(\vec{r})$, propagated to our (approximately) point receiver:

$$g_n = \int E_n^T(\vec{r})\sigma(\vec{r})G(\vec{r})dV \tag{2}$$

where $G(\vec{r})$ represents propagation from scene position $\vec{r}$ to the receiver. The illuminating fields consist of the superposition of fields radiated by the "on" metamaterial aperture elements, treated as magnetic dipoles. If the scalar illumination fields represent the y-component of the fields, and the element magnetic dipoles are oriented primarily in the z-direction, then these resulting fields are:

$$E_n^T(\vec{r}) = A\sum_{p=1}^{P} b_{np}\hat{R}_{p,x}m_{p,z}(\vec{r}_p;\vec{b}_n)\left(\frac{jk}{R_p} - \frac{1}{R_p^2}\right)e^{-jkR_p} \tag{3}$$

where A is a constant, $m_{p,z}(\vec{r}_p;\vec{b}_n)$ is the z component of the p^{th} dipole at position $\vec{r}_p$, $\hat{R}_{p,x}$ is the x component of the unit vector from the p^{th} dipole to scene position $\vec{r}$, R_p is the associated distance, and b_{np} is a binary indicator representing the on/off state of the p^{th} element for tuning state n. The p^{th} aperture element dipole moment, given by $m_{p,z} = \alpha H_{cav}(\vec{r}_p, \vec{b}_n)$, generally depends on the on/off configuration (mask), here represented by a vector $\vec{b}_n$. This dependence arises from inter-element coupling and altered boundary conditions that modify the local cavity field H_{cav} according to the tuning state [26].

To optimize our illumination patterns, we wish to use our NN to learn a set of optimized $P \times 1$ tuning vectors $\{\vec{w}_m\}_{m=1,\ldots,M}$ that can be applied to the aperture elements for improved accuracy and efficiency. To this end, we assume we can ignore coupling effects, and treat the cavity fields as independent of tuning state, so that $H_{cav}(\vec{r}_p;\vec{b}_n) = H_{cav}(\vec{r}_p)$. In addition, we consider a

measurement set $\vec{g}$ corresponding to one element on at a time, so $b_{np} = \delta_{np}$. Then

$$E_n^T(\vec{r}) = A\hat{R}_{n,x} m_{n,x}(\vec{r}_n) \left(\frac{jk}{R_n} - \frac{1}{R_n^2} \right) e^{-jkR_n}. \tag{4}$$

We then define an $M \times N$ matrix with elements w_{mn} defining a linear transformation of the obtained data,

$$g_{opt,m} = \sum_n w_{mn} g_n. \tag{5}$$

Given a measurement set corresponding to one-on tuning states, we can thus use the NN to learn the elements w_{mn} leading to an optimally-transformed data set $\vec{g}_{opt}$. Finally, the linearity of the measurement model implies that

$$\begin{aligned} g_{opt,m} &= \sum_n w_{mn} \int E_n^T(\vec{r}) \sigma(\vec{r}) G(\vec{r}) dV \\ &= \int \left[A \sum_n w_{mn} \hat{R}_{n,x} m_{n,z}(\vec{r}_n) \left(\frac{jk}{R_n} - \frac{1}{R_n^2} \right) e^{-jkR_n} \right] \sigma(\vec{r}) G(\vec{r}) dV. \end{aligned} \tag{6}$$

We can see that, under the weak-coupling approximation, the expression in brackets defines optimum illumination patterns resulting from elements tuned with weights w_{mn}. These weights can be continuous if grayscale tuning is realizable and negative weights can be synthesized. Otherwise, the weights can be driven to $\{1, 0\}$ binary values using a temperature parameter [25].

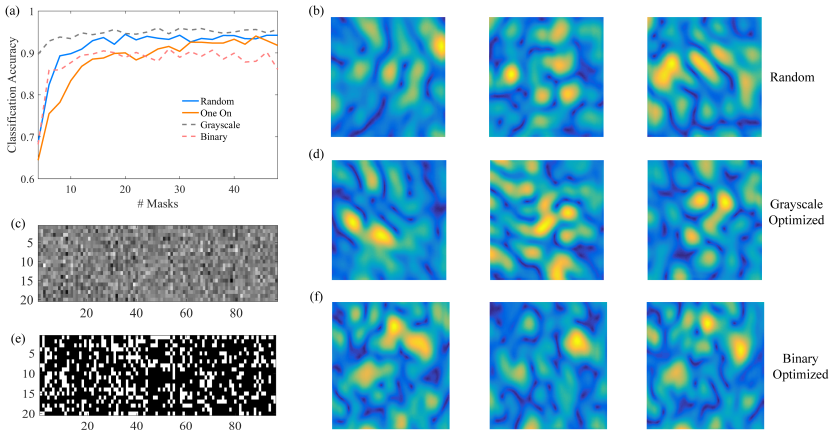

Figure 3: (a) Simulated classification accuracy for different illumination schemes. (b) Three example random illumination patterns using 45 "on" DMA elements. (c) Grayscale weights obtained using NN. (d) Corresponding grayscale-tuned DMA illumination patterns. (e) Binary weights obtained using NN. (f) Corresponding binary-tuned DMA illumination patterns.

Figure 3(a) illustrates how the classification improves with number of measurements for different simulated illumination strategies. The data set was simulated by treating the MNIST digit data set as a real-valued scattering density, and computing the measured signal according to a discrete form of Eq. (2). We employed a different NN architecture in these simulations, consisting of a total of 6 dense layers with ReLU activations. DMA illumination patterns were experimentally obtained using a near field scanning technique, and these illumination patterns were computationally propagated according to the Huygens-Fresnel principle [27]. "Random" illumination patterns correspond to DMA radiation using 45 randomly-selected "on" elements, while "One On" patters result from a single "on" element at a time. These "One On" illumination patterns were used to obtain the optimized patterns according to Eq. (6). Grayscale weights allowed arbitrary tuning of the magnitude of the aperture magnetic dipoles, and these optimized patterns demonstrate improved accuracy over both the random and "One On" illumination results. In practice, grayscale tuning can be enabled using different modulation strategies, and negative weights can be synthesized from a post processing subtraction operation [24]. Nevertheless, our current DMA design supports only binary tuning of the metamaterial elements. We attain binary weights with our NN using a temperature parameter strategy [25], in which the weights

matrix is multiplied by a constant, which gradually increases with each iteration, before performing a SoftMax. To accommodate multiple binary elements for each mask, we make an $M \times N \times 2$ array, perform SoftMax over the final dimension, then multiply one of these two subarrays by our raw input data set. To suppress excessive pattern correlation present in the "One On" patterns due to a strong, constant radiative contribution from the aperture feed location, the mean over all patterns is subtracted from each. Unfortunately, an improvement in classification accuracy over random or "One On" illumination is not generally observed for the resulting binary weighting strategy. This may be due simply to insufficient freedom in pattern synthesis using binary weights. The slight degradation, though, in classification accuracy with more masks relative to the binary "One On" case indicates that the optimization routine can be improved. That is, binary performance may benefit from a different temperature parameter, as well as a different NN architecture and number of iterations. These factors will be considered in the near future.

5 Conclusion

In this work we demonstrated experimental classification of hand gestures using single-frequency, spatially diverse illumination generated by a microwave DMA. Classification accuracy 97% was achieved directly from raw, single-pixel measurements using only ≈ 28 complex measurements or 44 magnitude-only measurements. In addition, we described approximate conditions under which illumination patterns can be optimized from experimental data, and demonstrated this optimization procedure in simulation.

6 Future Steps

The experimental results presented above used a simple, static NN architecture. As the NN itself was not the focus of this work, we expended little effort in optimizing the architecture. The classification accuracies reported may improve with further investigation of an optimal NN.

We propose continuing to search for a method which will effectively sort the masks based on their contribution to detection so that we can further minimize the number of masks needed to classify with high accuracy. We have explored non-linear clustering methods (such as t-SNE and topological data analysis) as a means of quantifying the relative contribution of each mask, but had limited success. With time, we may be able to devise a mask selection scheme which outperforms random selection. In addition, we hope to devise a method for experimental aperture or mask design using an NN, which would involve minimizing or accounting for DMA radiation nonlinearities.

Moreover, we would like to extend this scheme of classification using a DMA to dynamic hand gestures. In such a system, an individual can perform one of a number of predetermined motions that the system will correctly identify. The applications for dynamic gesture recognition are vast and include device control for televisions or computers, car control, and potentially HVAC and lighting control in residential settings. Such a system would require a radio which can sample fast enough to capture the rate of human motion. We are currently designing such a radio and hope to test this system soon.

Acknowledgments

We would like to thank Mohammadreza F. Imani and Philipp del Hougne for their assistance in this project. We would also like to thank our research advisor, David R. Smith, who provided support in the form of facilities, funding, and helpful comments along the way. Moreover, we would like to thank Ouwen Huang and Kevin Zhou for their help with Python, TensorFlow, and Keras. Finally, we would like to thank Dr. Roarke Horstmeyer for engaging lectures and for allowing us the freedom to pursue our ideas as a final project for BME590-Machine Learning in Imaging.

References

[1] R. Fergus, A. Torralba, and W. T. Freeman, "Random lens imaging," 2006.

[2] D. J. Brady, K. Choi, D. L. Marks, R. Horisaki, and S. Lim, "Compressive holography," *Opt. Exp.*, vol. 17, no. 15, pp. 13040–13049, 2009.

[3] D. J. Brady, *Optical imaging and spectroscopy*. John Wiley & Sons, 2009.

[4] J. Hunt, T. Driscoll, A. Mrozack, G. Lipworth, M. Reynolds, D. Brady, and D. R. Smith, "Metamaterial apertures for computational imaging," *Science*, vol. 339, no. 6117, pp. 310–313, 2013.

[5] A. Liutkus, D. Martina, S. Popoff, G. Chardon, O. Katz, G. Lerosey, S. Gigan, L. Daudet, and I. Carron, "Imaging with nature: Compressive imaging using a multiply scattering medium," *Scientific reports*, vol. 4, 2014.

[6] O. Katz, P. Heidmann, M. Fink, and S. Gigan, "Non-invasive single-shot imaging through scattering layers and around corners via speckle correlations," *Nature photonics*, vol. 8, no. 10, pp. 784–790, 2014.

[7] J. Gollub, O. Yurduseven, K. P. Trofatter, M. F. Imani, H. Odabasi, T. Sleasman, M. Boyarsky, T. Zvolensky, D. Arnitz, A. Pedross-Engel, G. Lipworth, A. Rose, D. R. Smith, M. Reynold, and D. Brady, "Large metasurface aperture for millimeter wave computational imaging at the human-scale," *Scientific reports*, vol. 7, p. 42650, 2017.

[8] L.-H. Yeh, L. Tian, and L. Waller, "Structured illumination microscopy with unknown patterns and a statistical prior," *Biomedical optics express*, vol. 8, no. 2, pp. 695–711, 2017.

[9] J. N. Mait, G. W. Euliss, and R. A. Athale, "Computational imaging," *Advances in Optics and Photonics*, vol. 10, no. 2, pp. 409–483, 2018.

[10] C. Wu, Z. Yang, Z. Zhou, X. Liu, Y. Liu, and J. Cao, "Non-invasive detection of moving and stationary human with wifi," *IEEE Journal on Selected Areas in Communications*, vol. 33, no. 11, pp. 2329–2342, 2015.

[11] B. Wang, Q. Xu, C. Chen, F. Zhang, and K. R. Liu, "The promise of radio analytics: A future paradigm of wireless positioning, tracking, and sensing," *IEEE Signal Processing Magazine*, vol. 35, no. 3, pp. 59–80, 2018.

[12] G. Krieger, "Mimo-sar: Opportunities and pitfalls," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 52, pp. 2628–2645, May 2014.

[13] S. S. Ahmed, "Electronic microwave imaging with planar multistatic arrays," 2014.

[14] Y. Bromberg, O. Katz, and Y. Silberberg, "Ghost imaging with a single detector," *Physical Review A*, vol. 79, no. 5, p. 053840, 2009.

[15] B. Sun, M. P. Edgar, R. Bowman, L. E. Vittert, S. Welsh, A. Bowman, and M. Padgett, "3d computational imaging with single-pixel detectors," *Science*, vol. 340, no. 6134, pp. 844–847, 2013.

[16] T. Sleasman, M. F. Imani, J. N. Gollub, and D. R. Smith, "Dynamic metamaterial aperture for microwave imaging," *Appl. Phys. Lett.*, vol. 107, no. 20, 2015.

[17] T. Sleasman, M. F. Imani, J. N. Gollub, and D. R. Smith, "Microwave imaging using a disordered cavity with a dynamically tunable impedance surface," *Physical Review Applied*, 2016.

[18] T. Sleasman, M. F. Imani, W. Xu, J. Hunt, T. Driscoll, M. S. Reynolds, and D. R. Smith, "Waveguide-fed tunable metamaterial element for dynamic apertures," *IEEE Antennas and Wireless Propag. Lett.*, vol. 15, pp. 606–609, 2016.

[19] T. Sleasman, M. Boyarsky, M. Imani, J. Gollub, and D. Smith, "Design considerations for a dynamic metamaterial aperture for computational imaging at microwave frequencies," *JOSA B*, vol. 33, no. 6, pp. 1098–1111, 2016.

[20] T. Sleasman, M. Boyarsky, L. Pulido-Mancera, T. Fromenteze, M. F. Imani, M. S. Reynolds, and D. R. Smith, "Experimental synthetic aperture radar with dynamic metasurfaces," *IEEE Transactions on Antennas and Propagation*, vol. 65, no. 12, pp. 6864–6877, 2017.

[21] T. Sleasman, M. Boyarsky, M. F. Imani, T. Fromenteze, J. N. Gollub, and D. R. Smith, "Single-frequency microwave imaging with dynamic metasurface apertures," *J. Opt. Soc. Am. B*, vol. 34, pp. 1713–1726, Aug 2017.

[22] A. V. Diebold, M. F. Imani, T. Sleasman, and D. R. Smith, "Phaseless computational ghost imaging at microwave frequencies using a dynamic metasurface aperture," *Applied Optics*, vol. 57, no. 9, pp. 2142–2149, 2018.

[23] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[24] R. Horstmeyer, R. Y. Chen, B. Kappes, and B. Judkewitz, "Convolutional neural networks that teach microscopes how to image," *arXiv preprint arXiv:1709.07223*, 2017.

[25] A. Chakrabarti, "Learning sensor multiplexing design through back-propagation," in *Advances in Neural Information Processing Systems*, pp. 3081–3089, 2016.

[26] L. Pulido-Mancera, M. F. Imani, P. T. Bowen, N. Kundtz, and D. R. Smith, "Analytical modeling of a two-dimensional waveguide-fed metasurface," *arXiv preprint arXiv:1807.11592*, 2018.

[27] J. W. Goodman, *Introduction to Fourier optics*. Roberts and Company Publishers, 2005.