Compact computational cameras by joint metasurface and network design

Kelsey Lee

Abstract—Optical computing offers a promising approach for reducing the digital computation required for computer vision tasks by performing feature extraction during image formation. In this project, we investigate metasurface-based optics as a platform for hybrid optical–digital neural networks, using metalenses to encode convolution-like operations directly into the imaging system. We implement a classification pipeline in which a metasurface front end performs optical feature extraction and a digital fully connected layer performs classification. Two physical configurations are explored: a metalens array and a spatially multiplexed singlet metalens, both optimized end-to-end by jointly updating the lens phase profile and classifier weights. The optimized system achieves accuracies of 98.72% on MNIST and 89.62% on Fashion-MNIST while reducing the required digital computation by 97.8%.

Index Terms—optical computing, metasurfaces, metalenses, computational imaging, optical neural networks

1 Introduction

Recent advances in machine learning have dramatically increased the computational demands of computer vision systems. Modern vision pipelines require substantial digital processing to perform tasks such as feature extraction and classification, creating challenges for applications that require low latency, low power consumption, and compact sensor form factors. These constraints are particularly relevant for embedded platforms in robotics, autonomous systems, and edge computing. As a result, there has been growing interest in alternative computing paradigms that can reduce the reliance on digital hardware.

Optical computing provides a promising path towards more efficient inference by performing key linear operations directly in the optical domain. Free-space optical systems naturally implement operations such as convolution and Fourier transforms in parallel and at the speed of light, allowing computation to occur before the signal is captured. By transferring a portion of the computational workload from the digital backend to an analog optical frontend, hybrid optical–digital systems can significantly reduce the computational burden of conventional vision pipelines.

Among the available optical platforms, metasurface-based optics is well-suited for compact and programmable optical computing hardware. Metalenses are planar optical elements composed of subwavelength nanostructures that provide fine spatial control over the phase of incident light, allowing complex wavefront transformations to be encoded within a thin optical layer. As a result, metalenses can be used to implement optical neural networks (ONNs) within imaging systems.

In this project, we investigate hybrid optical–digital computational cameras based on metalens optics for image classification tasks. We study two optical architectures: a metalens array and a spatially multiplexed singlet metalens. The multiplexing configuration specifically is evaluated under two forward models: a conventional shift-invariant point spread function (PSF) model and a spatially varying PSF model that more accurately captures off-axis imaging behavior. Using these models, we briefly discuss the choice of optical and digital design parameters, such as field of view (FOV) and pixel sampling, and their impact on classification performance.

The results will provide insight into the trade-offs between different metalens architectures and modeling assumptions, and demonstrate the potential of metasurface-based computational cameras as lightweight optical frontends for machine learning inference.

2 Related Work

Early demonstrations of optical neural networks for image classification were implemented using $4f$ optical systems with passive diffractive elements placed at the Fourier plane. For example, Chang et al. [1] demonstrated optical feature extraction using a diffractive mask in a $4f$ system, achieving 44.4% accuracy on monochromatic CIFAR-10, compared to roughly 30% accuracy from a fully connected classifier operating on the same features.

More recent work has explored metasurface-based optical encoders, which enable compact implementations of optical computing systems. Two primary hardware configurations are used: metalens arrays, where each lens corresponds to a convolutional kernel, and multiplexed singlet metalenses, where multiple kernels are encoded within a single lens and separated at the sensor. Zheng et al. [2] demonstrated a compound doublet metasurface in which the first lens splits incident light into kernel-weighted channels while the second lens focuses and corrects aberrations. Using polarization multiplexing to encode positive and negative channels, their system achieved 98.6% and 88.8% accuracy on MNIST and Fashion-MNIST, respectively. Liang et al. [3] later showed that comparable performance can be achieved using a single optimized metalens, reporting 98.59%, 92.63%, and 68.87% accuracy on MNIST, Fashion-MNIST, and monochromatic CIFAR-10.

Most prior work chose systems that are well-approximated as shift invariant, such that a PSF describes the imaging process across the field of view. Wei et al. [4]

demonstrated that explicitly optimizing spatially varying PSFs can improve classification performance by introducing additional degrees of freedom, achieving 72.76% accuracy on monochromatic CIFAR-10.

Due to strong chromatic dispersion in metasurface optics, most optical classification systems operate under monochromatic illumination and are evaluated on grayscale datasets such as MNIST and Fashion-MNIST [2], [3], [4]. Recently, Choi et al. [5] demonstrated a polychromatic metasurface classifier by jointly designing wavelength-dependent PSFs for the red, green, and blue channels, achieving state-of-the-art 73.2% accuracy on full-color CIFAR-10.

Finally, we note a distinct but important class of optical neural networks are implemented using photonic integrated circuits (PICs). PIC-based systems offer advantages such as reconfigurability and potential nonlinear optical operations, but they typically rely on coherent illumination and are therefore less compatible with conventional incoherent imaging systems [6]. As a result, free-space optical encoders remain attractive for computational imaging applications, although their performance is fundamentally limited by the linearity of optical propagation, which prevents direct implementation of nonlinear activations such as ReLU.

3 Proposed Method

The goal of this project is to implement and compare two metasurface-based optical encoders: a metalens array and a multiplexed singlet metalens. For both configurations, we perform end-to-end optimization in which the metalens phase profile and the backend fully connected (FC) classifier weights are jointly optimized to improve classification performance.

The metasurface nano-cell period is set to 350 nm, which determines the spatial sampling resolution of the phase mask. In practical implementations, metasurface lenses are typically on the order of millimeters in diameter. However, modeling such systems would require millions of trainable phase parameters. To keep the optimization realistic for this project, we instead use a scaled optical system while maintaining reasonable optical parameters such as the effective $F/\#$ and overall shape factors. This scaling implicitly assumes an idealized sensor capable of resolving 350 nm pixel spacing.

For the multiplexed singlet configuration, we use a metalens with an 80 μm diameter, while the array configuration consists of a 4×4 grid of lenses, each with a diameter of 20 μm. The focal length is fixed at 100 μm for all configurations, and the half field of view is set to 7°.

3.1 Optical Image Formation Model

Images formed by the computational camera are simulated by first computing the system point spread functions (PSFs) and then convolving these PSFs with an ideal sensor image. Here, the ideal sensor image refers to the input image resampled to the sensor pixel grid according to the system magnification, rather than an image optimized for classification.

The PSFs are computed using the angular spectrum method (ASM), which propagates a complex optical field from the lens plane to the sensor plane. Given an incident field $U_0(x, y)$ at the lens plane, the propagated field $U_z(x, y)$ at propagation distance $z = f$ is given in the Fourier domain by

$$U_z(x, y) = \mathcal{F}^{-1}\left\{\mathcal{F}\{U_0(x, y)\}\, H(f_x, f_y)\right\}, \tag{1}$$

where $\mathcal{F}$ denotes the Fourier transform and $H(f_x, f_y)$ is the ASM transfer function

$$H(f_x, f_y) = \exp\left(i\, 2\pi z \sqrt{\frac{1}{\lambda^2} - f_x^2 - f_y^2}\right). \tag{2}$$

The PSF is then obtained from the propagated field as

$$\mathrm{PSF}(x, y) = |U_z(x, y)|^2. \tag{3}$$

For shift-invariant optical systems, the PSF is computed by propagating a single on-axis plane wave input field U_0.

For spatially varying systems, the PSF depends on field position. We choose a sparse 4×4 grid of field points to be sampled across the field of view, and a PSF is computed at each location. To synthesize the final image, overlapping windows are used to interpolate between these field-dependent PSFs. The resulting image can be written as

$$y_i = \sum_{r=0}^{p-1} \sum_{j=0}^{k-1} a_j^{(r)}\, w_{i-j}^{(r)}\, x_{i-j}, \qquad 0 \le i < m, \tag{4}$$

where y denotes the output image, x is the input image, $a^{(r)}$ represents the PSF associated with the r-th patch, and $w^{(r)}$ denotes the window function used to interpolate between overlapping PSF patches. The system uses p spatial patches across the image to approximate the spatially varying convolution [7]. Fig. 1 shows a grid of spatially varying PSFs for a large field-of-view focusing lens with a hyperbolic phase profile. The red box highlights an example PSF window used during spatially varying image synthesis.

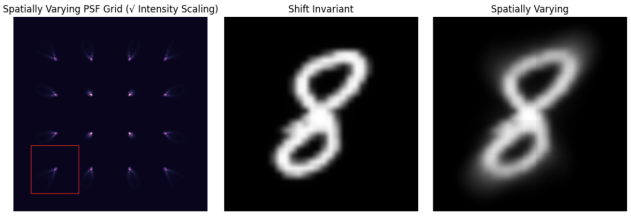

Fig. 1. Left: Example spatially varying PSF grid for a large field-of-view focusing lens with a hyperbolic phase profile; the red box shows a PSF patch. Middle: image formation under a shift-invariant model. Right: image formation under a spatially varying model.

3.2 End-to-End Optimization

The optimization procedure consists of two stages: (1) constructing an initialization solution and (2) end-to-end training. In the initialization stage, a target point spread function (PSF) is obtained from a trained digital model and an optical phase profile is computed to approximate this PSF. In the second stage, the phase profile and fully connected (FC) weights are jointly optimized.

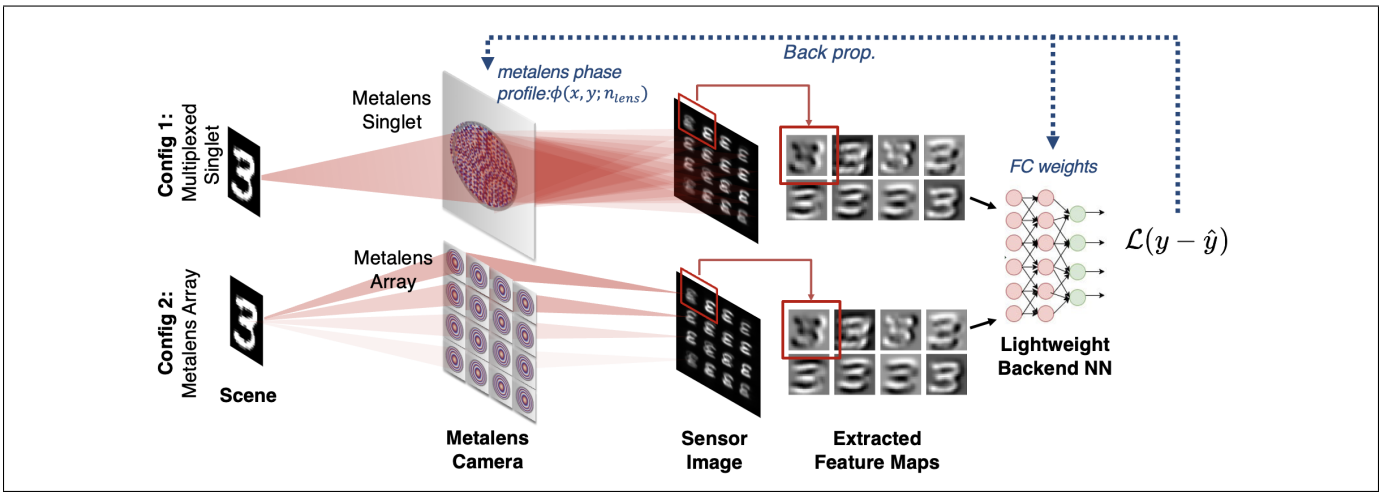

Fig. 2. Optical–digital classification pipeline. Two metasurface implementations: a multiplexed singlet metalens (top) and a metalens array (bottom). The optical system produces kernel-weighted responses at the sensor, from which feature maps are extracted and passed to a lightweight fully connected backend. The metalens phase profile and network weights are jointly optimized using backpropagation.

3.2.1 Phase Initialization

The optical system is designed to perform convolutions with a chosen number of kernels and kernel size. To obtain these kernels, a two-layer digital CNN consisting of a convolutional layer followed by a fully connected layer is first trained. The learned convolutional kernels serve as target kernels for the optical system.

Because optical PSFs are non-negative while digital kernels contain both positive and negative values, each kernel is decomposed into positive and negative components. These components are implemented as separate PSF channels and later subtracted during digital post-processing before the FC layer. For the grid of 16 PSFs described earlier, this corresponds to 8 positive channels and 8 negative channels and thus representing 8 convolution kernels.

The digital kernels are converted into target PSFs by upsampling them to the sensor resolution. An upsampling factor of 2–4 is used depending on how well the phase profile can reproduce the desired PSF. The phase profile is then computed using a Gerchberg–Saxton phase retrieval algorithm with Angular Spectrum Method (ASM) propagation [8].

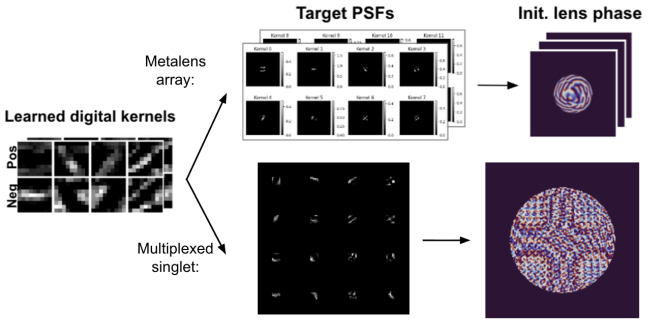

Fig. 3. Initialization pipeline. Digital kernels are converted to target PSFs and used to compute an initial phase via Gerchberg–Saxton retrieval. Two implementations are shown: a metalens array (top) and a multiplexed singlet metalens (bottom).

3.2.2 Optimization

The system is optimized end-to-end using a cross-entropy classification loss:

$$\mathcal{L} = -\sum_{c} y_c \log(\hat{y}_c). \tag{5}$$

Since the optical forward model is differentiable, gradients are propagated through the imaging model to update the lens phase parameters. As shown in Fig. 3, optical responses from positive and negative PSF channels are subtracted to produce feature maps corresponding to the convolution kernels. These feature maps are max-pooled, flattened, and passed to a digital fully connected classifier.

For the lens array configuration, all lens phases are optimized jointly, whereas the singlet configuration optimizes a single phase profile. The optical phase parameters are co-optimized with the FC weights using different learning rates. Because the optical image formation model must be evaluated at every iteration, training is computationally expensive.

4 Experimental Results

4.1 Hybrid System Performance

Across all experiments, we show that a significant portion of the inference computation can be performed in the optical domain. As a baseline, we consider a purely digital fully connected (FC) classifier operating directly on the input images. If the hybrid optical–digital system achieves higher classification accuracy than this FC baseline, the optical frontend is effectively performing useful feature extraction. The FC baseline achieves accuracies of 92.70% on MNIST and 84.29% on Fashion-MNIST (Table 1).

Because optical propagation is inherently linear, the practical upper bound for classification performance is given by a purely digital two-layer CNN consisting of a single convolutional layer followed by a fully connected classifier, using the same kernel size and number of kernels as the optical system. This digital CNN achieves accuracies

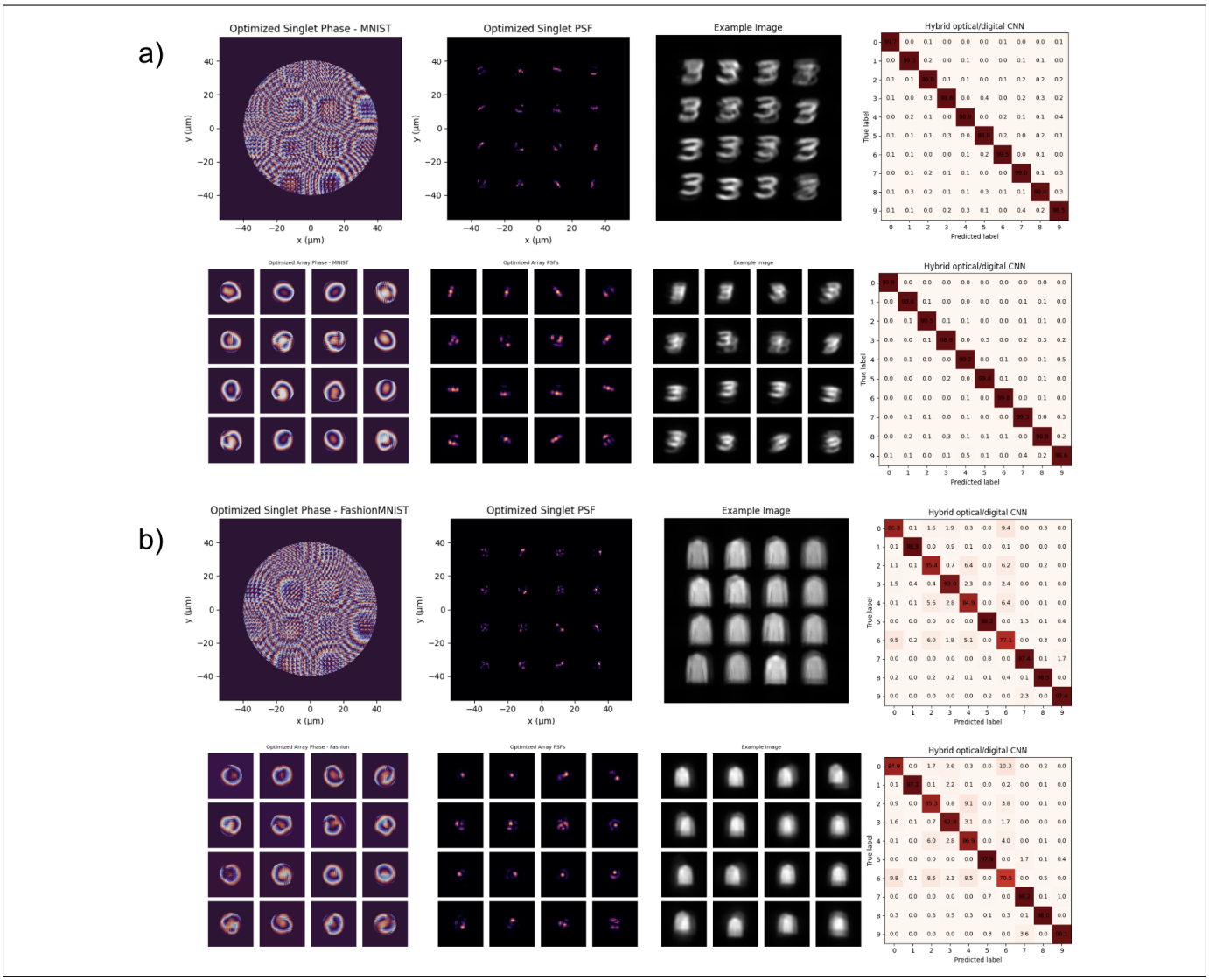

Fig. 4. Hybrid optical–digital classification with optimized metasurfaces. (a) MNIST and (b) Fashion-MNIST. Top: optimized singlet phase, resulting PSF distribution, example sensor image, and confusion matrix. Bottom: optimized array phases, PSFs, example sensor images, and confusion matrix.

of 98.58% and 90.11% on MNIST and Fashion-MNIST, respectively.

After end-to-end optimization of the lens phase profiles and backend FC weights, the hybrid optical system achieves peak accuracies of 98.72% on MNIST and 89.62% on Fashion-MNIST. Notably, the MNIST performance slightly exceeds that of the reference digital CNN. Although the digital CNN represents a loose upper bound, the optimization process does not constrain the phase masks to maintain their initial target PSFs. As a result, the learned optical transformations may deviate from the original digital kernels, and the final performance need not converge exactly to the digital baseline. The slight improvement is therefore likely attributable to suboptimal training of the reference CNN rather than a fundamental violation of the linearity constraint.

Figure 4 and Table 1 provide a detailed comparison of the array and multiplexed singlet configurations, including the optimized phase profiles, resulting PSFs, and example output images. For a given dataset, both configurations begin with identical target PSFs derived from the trained CNN kernels. However, during optimization the PSFs evolve differently depending on the physical configuration.

Qualitatively, the singlet configuration produces noticeably sharper output images than the array configuration. This behavior comes from differences in the PSF parameterization: the singlet PSFs are generated using a $2\times$ upsampling of the kernel grid, while the array PSFs use $4\times$ upsampling. Because the array configuration contains fewer trainable phase parameters per lens, smaller PSFs encourage the optimization to converge toward a simple focusing lens, which provides little useful feature extraction for classification.

Across the experiments, the multiplexed singlet configuration consistently achieves slightly higher performance than an array occupying the same total aperture area. This advantage likely arises from the larger number of controllable phase parameters in the singlet design, allowing the system to better balance spatial encoding across the sensor. However, the lenses used in these simulations are intention-

TABLE 1: Classification performance.

		MNIST $8 \times 7 \times 7$		Fashion MNIST $8 \times 7 \times 7$	
		Arr.	Singl.	Arr.	Singl.
Digital	FC only	92.70		84.29	
	2 layer CNN	98.58		90.11	
Hybrid	GS phase + FC (retrained)	98.14	98.66	88.73	**89.62**%
	End-to-end opt	**98.31**%	**98.72**%	**88.92**%	89.50

TABLE 2: Digital compute comparison.

Operation	2-layer CNN	Hybrid optical/digital
Input crop	229×229	229×229
Processing		$8 \times 229 \times 229$
Convolutions	$223 \times 223 \times 8 \times 7 \times 7$	
FC layer	1568×10	1568×10
Total ops	19,509,448	435,208
Reduction	**97.8% fewer operations**	

ally much smaller than realistic metasurface devices in order to keep the optimization manageable. For larger physical lenses, the parameter limitation of the array configuration would be significantly reduced, and this performance gap may diminish.

Finally, the primary motivation for optical computing is the reduction of digital computation required for inference. As shown in Table 2, the hybrid optical system eliminates the need for digital convolution operations, since the convolution is performed optically during image formation. The only remaining digital operations are channel subtraction used to form feature maps and the final fully connected classification layer. In contrast, performing the same convolution digitally requires a large number of multiply–accumulate operations. Under our system parameters, the hybrid architecture reduces the required digital computation by approximately 97.8%.

4.2 Comparison of Physical Configurations

As noted earlier, the multiplexed singlet configuration performed slightly better than the array configuration in these experiments. One possible explanation is simply the number of optimization parameters available. In the large singlet lens, each phase element contributes to all kernels simultaneously at the sensor, allowing the optimization to balance how light is distributed across the entire set of PSFs. In contrast, the array lenses are more localized, with each lens primarily contributing to a single kernel.

However, this does not necessarily mean that the singlet configuration is always preferable. If the goal is to reproduce a predefined target PSF as accurately as possible—such as when matching convolution kernels learned from a digital CNN—the array configuration may be advantageous. Each lens can be designed more independently, which can make it easier to approximate the desired PSF shape.

On the other hand, a large multiplexed metalens allows every optimization parameter to influence all kernels at the sensor. This global coupling may be beneficial during end-to-end optimization, where the system is free to move away from the original target PSFs and instead learn an optical encoding that improves overall classification performance.

From an optical design perspective, the two configurations also differ in their physical parameters. In our system, the singlet operates at $F/1.25$, while the lenses in the array correspond to $F/5$. The singlet therefore has a much larger aperture and collects more light, but it must also split and redirect that light to form multiple kernels, potentially requiring steeper phase gradients and larger deflection angles. Which configuration is more efficient in practice may depend on the specific metasurface implementation.

Finally, field-of-view considerations may also influence the design choice. For larger fields of view, a single large-aperture metalens may become more susceptible to off-axis aberrations. In such cases, an array configuration—where each lens operates over a smaller angular range—could again become advantageous. In practice, the relative benefits of the two configurations are likely to depend on the specific system requirements.

4.3 Spatially Varying Model

TABLE 3: Spatially varying performance.

Method	Fashion MNIST (%)
Shift-invariant model	88.90
SV model + FC retrained	88.48
SV end-to-end opt	88.53

Only limited experiments were performed using the spatially varying optical model due to the significantly increased computational cost. In this model, each training iteration requires propagating every image in the dataset through the optical simulation, which substantially increases training time compared to the shift-invariant approximation.

Nevertheless, we report some preliminary results. To emphasize spatially varying effects, the half field of view was increased from 7° to 15°. These experiments were conducted only with the multiplexed singlet configuration. In the array configuration, each lens forms images within a relatively localized region of the sensor, making off-axis aberrations less pronounced and reducing the benefit of a spatially varying model. Additionally, only the Fashion-MNIST dataset was used in order to highlight potential differences in performance, since it presents a more challenging classification task than MNIST.

First, a lens optimized under the shift-invariant model achieved a classification accuracy of 88.9% on Fashion-MNIST. When this same lens was evaluated using the spatially varying model, and the fully connected (FC) weights were retrained on the resulting images, the accuracy decreased slightly to 88.48%, as expected due to the increased modeling realism. When the system was subsequently optimized end-to-end using the spatially varying model (with a limited number of training epochs), part of the performance was recovered, reaching 88.53% accuracy. However, this result did not exceed the performance obtained with the original shift-invariant optimization.

Based on this experiment alone, the spatially varying model does not appear to provide a clear advantage in terms of classification accuracy. It is important to note, however, that the larger field of view also introduces additional optical aberrations, making the overall problem more

difficult. Further experiments with longer optimization runs and different system configurations would be required to more conclusively evaluate the potential benefits of spatially varying optical models.

5 Conclusion

We investigated hybrid optical–digital computational cameras that perform part of the inference process directly in the optical domain. Using metalenses, we implemented optical neural networks capable of encoding convolution-like operations during image formation. Two physical configurations were explored: a metalens array and a spatially multiplexed singlet metalens. Both systems were optimized end-to-end by jointly updating the lens phase profile and the digital classifier weights. The resulting systems achieved classification accuracies of 98.72% on MNIST and 89.62% on Fashion-MNIST while reducing the required digital computation by 97.8%.

We additionally explored a spatially varying optical model to better capture off-axis imaging effects. Preliminary experiments suggest that the increased modeling fidelity does not immediately translate to improved classification performance, although the larger field of view also makes the optimization problem more difficult. Further investigation is needed to determine whether more efficient implementations of spatially varying convolution or longer optimization schedules could better leverage these additional degrees of freedom.

Several directions remain for future work. One promising avenue is the development of methods to compress deeper neural networks into representations that can be implemented with a single optical convolution layer, which could provide stronger initialization targets for optical optimization. Finally, the linear nature of conventional imaging optics remains a fundamental limitation of free-space optical neural networks. Developing practical approaches for implementing nonlinear optical operations, though outside of the scope of this catergory of projects, would represent a significant step forward for the field and could enable more powerful fully optical computing architectures.

References

[1] J. Chang, V. Sitzmann, X. Dun, W. Heidrich, and G. Wetzstein, "Hybrid optical-electronic convolutional neural networks with optimized diffractive optics for image classification," *Scientific Reports*, vol. 8, p. 12324, 2018.

[2] H. Zheng, Q. Liu, I. I. Kravchenko *et al.*, "Multichannel meta-imagers for accelerating machine vision," *Nature Nanotechnology*, vol. 19, pp. 471–478, 2024.

[3] R. Liang, S. Wang, Y. Dong, L. Li, Y. Kuang, B. Zhang, and Y. Yang, "Metasurface-generated large and arbitrary analog convolution kernels for accelerated machine vision," *ACS Photonics*, vol. 11, no. 12, pp. 5430–5438, 2024.

[4] K. Wei, X. Li, J. Fröch, P. Chakravarthula, J. Whitehead, E. Tseng, and F. Heide, "Spatially varying nanophotonic neural networks," *Science Advances*, vol. 10, no. 45, p. eadp0391, 2024.

[5] M. Choi, J. Xiang, A. Wirth-Singh *et al.*, "Transferable polychromatic optical encoder for neural networks," *Nature Communications*, vol. 16, p. 5623, 2025.

[6] M. Choi and A. Majumdar, "Free-space optical encoder for computer vision," *npj Nanophotonics*, vol. 2, no. 1, p. 36, 2025.

[7] M. Hirsch, S. Sra, B. Schölkopf, and S. Harmeling, "Efficient filter flow for space-variant multiframe blind deconvolution," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010, pp. 607–614.

[8] R. W. Gerchberg and W. O. Saxton, "A practical algorithm for the determination of phase from image and diffraction plane pictures," *Optik*, vol. 35, no. 2, pp. 237–246, 1972.