OPTO-ELECTRONIC CONVOLUTIONAL NEURAL NETWORK DESIGN VIA DIRECT KERNEL OPTIMIZATION

Ali Almuallem[1], *Harshana Weligampola*[1], *Abhiram Gnanasambandam*[2], *Wei Xu*[1], *Dilshan Godaliyadda*[2], *Hamid R. Sheikh*[2], *Stanley H. Chan*[1], *Qi Guo*[1]

[1]Elmore Family School of Electrical and Computer Engineering, Purdue University
[2]Samsung Research America

ABSTRACT

Opto-electronic neural networks integrate optical front-ends with electronic back-ends to enable fast and energy-efficient vision. However, conventional end-to-end optimization of both modules is limited by costly optical simulations and large parameter spaces. We introduce a two-stage strategy for designing opto-electronic convolutional neural networks (CNNs): first, train a standard electronic CNN, then realize the optical front-end—implemented as a metasurface array—through direct kernel optimization of its first convolutional layer. This approach reduces computational and memory demands by hundreds of times and improves training stability compared to end-to-end optimization. On monocular depth estimation, the proposed two-stage design achieves twice the accuracy of end-to-end training under the same training time and resource constraints.

Index Terms— Opto-electronic neural networks, metasurfaces, depth estimation

1. INTRODUCTION

Opto-electronic neural networks integrate optical front-ends—such as transmission masks [1], diffractive optical elements [2], and metasurfaces [3]—with electronic back-ends based on conventional neural architectures to perform vision and imaging tasks. By leveraging optics to preprocess signals before electronic inference, such systems offer the potential for low-latency [4] and energy-efficient [1] computation. However, most existing approaches rely on an end-to-end training paradigm in which both the optical components and the electronic layers are optimized jointly [2, 3, 4, 5, 6]. In practice, this end-to-end scheme requires excessive computational resources as the optical simulators are expensive to evaluate and the search space has a much higher dimension than purely optimizing computational models [6].

In this work, we propose an alternative strategy for designing opto-electronic *convolutional* neural networks (CNNs) that alleviates the challenges of end-to-end training. Instead of optimizing the hybrid system jointly, we first train a conventional electronic CNN (or employ a pre-trained model) and then design the optical front-end—implemented as a metasurface array—to replicate its first convolutional layer through *direct kernel optimization* (DKO). Compared to end-to-end optimization, this two-stage approach substantially simplifies the design process: the dimension of the variables to be optimized simultaneously is greatly reduced.

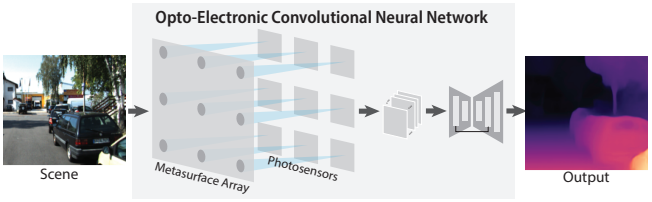

Fig. 1: We consider an opto-electronic convolutional neural network (CNN) that integrates a metasurface array with an electronic backend. The metasurface, a flat nanophotonic device, encodes the incident light from a scene into optical feature maps. As light propagates through the metasurface, it undergoes a phase modulation equivalent to convolving the common photograph of the scene with an engineered kernel. These optically generated feature maps are then processed electronically by a conventional CNN architecture.

We demonstrate these advantages by designing and simulating opto-electronic CNNs for an exemplar task: monocular depth estimation. According to our analysis, the proposed two-stage method achieves two-fold higher accuracy than the end-to-end scheme under identical hardware and training-time constraints. Furthermore, the dimension of the parameter space and computational cost of end-to-end training are hundreds of times higher, whereas the proposed approach maintains a significantly smaller computational footprint. In summary, the key contributions of this paper are:

- A two-stage strategy to design opto-electronic CNNs for vision and imaging;
- An exemplar opto-electronic CNN designed using the two-stage strategy for monocular depth estimation;

The work was supported by Samsung Research America.

- A comprehensive simulation study that demonstrates the accuracy, efficiency, and stability benefits of the proposed two-stage strategy over traditional end-to-end optimization.

2. RELATED WORK

Incorporating optics into artificial vision and imaging systems has emerged as a vibrant discipline, fueled by recent advances in optical fabrication technologies that now enable the accessibility of custom devices such as diffractive optical elements and metasurfaces. Collectively referred to as computational optics, these devices form feature maps—rather than conventional photographs—on the photosensor. Such feature maps can be understood as scene embeddings, generated according to engineered sensitivities of the optical devices [7].

Based on their functionalities, these systems can be broadly divided into two categories. The first category exploits the optics' intrinsic sensitivity to scene properties—such as depth [5, 8], spectrum [3, 9], and polarization [10, 11]—to encode this information into the feature map through point spread functions (PSFs) that vary with the underlying scene attributes. The second category seeks to emulate part [6, 3, 12] or all [4, 13] of a deep neural network architecture in the optical domain. Platforms in this class harness optics' inherent speed and parallelism, employing one or more layers of optical arrays in which each element performs a linear transformation, such as a convolution, on the output of the preceding layer. Such fully or partially optical neural networks have been experimentally demonstrated on basic vision tasks, including image classification [3].

These computational optics systems are often designed in an end-to-end (E2E) manner, where the optical elements and computational parameters are jointly optimized under a unified objective function. Such co-optimization has been shown to yield superior local optima compared to separately designing the optics and the computation [5, 2]. Nonetheless, implementing end-to-end optimization is challenging: the optical module requires differentiable solvers for light propagation—whether wave-based [14], ray-based [15], or hybrid approaches [16]—all of which are computationally intensive and significantly enlarge the design search space.

3. SYSTEM DESIGN

The proposed opto-electronic CNN, illustrated in Fig. 1, employs a 2D metasurface array to simultaneously encode the scene into $M \times N$-channel optical feature maps on a shared photosensor. This metasurface layer functions as an optical approximation of the first convolutional layer of a pre-trained CNN. The resulting features are then fed into the subsequent electronic layers of the CNN, which process the features to generate the final output.

3.1. Optical Model

Consider an incoherent scene located at a distance much larger than the spatial extent of the metasurface array. The incident environmental light can be modeled as a superposition of incoherent plane waves with amplitude distribution $J(\mathbf{k})$ as a function of the wave vector $\mathbf{k} = [k_x, k_y, k_z]$. A single plane-wave component is expressed as:

$$U(x, y; \mathbf{k}) \approx A_0(\mathbf{k}) \exp\left[j\left(k_x x + k_y y\right)\right], \tag{1}$$

where (x, y) denotes the coordinates on the metasurface array, and $A_0(\mathbf{k})$ is the amplitude of the plane wave.

Each metasurface element (m, n) is characterized by a modulation profile $C_{m,n}(x, y)$, which can be written as

$$C_{m,n}(x, y) = T_{m,n}(x, y) \exp\left[j\varphi_{m,n}(x, y)\right], \tag{2}$$

with $T_{m,n}(x, y)$ and $\varphi_{m,n}(x, y)$ representing the amplitude and phase modulation, respectively.

The resulting power distribution generated by metasurface (m, n) under an incident plane wave $\mathbf{k}$ is determined by free-space propagation of the modulated wavefront [17]:

$$P_{m,n}(u, v; \mathbf{k}) \propto A_0^2(\mathbf{k}) \left| \widetilde{C}_{m,n}\left(\frac{u - k_x s}{\lambda s}, \frac{v - k_y s}{\lambda s}\right) \right|^2, \tag{3}$$

where $\widetilde{C}_{m,n}$ denotes the Fresnel diffraction pattern of the modulated wavefront produced when a front-parallel plane wave propagates through the metasurface $C_{m,n}$ [14]. Eq. 3 indicates that the measurement formed on the photosensor, $I_{m,n}$, is a convolution of the pinhole image of the scene with an engineered kernel determined by the metasurface:

$$\begin{aligned} I_{m,n}(u, v) &= \int_{\mathbf{k}} P_{m,n}(u, v; \mathbf{k}) d\mathbf{k} \\ &= I(u, v) * h_{m,n}(u, v), \\ \text{where } I(u, v) &= \int_{\mathbf{k}} A_0^2(\mathbf{k}) d\mathbf{k} \quad \text{(pinhole image)}, \\ h_{m,n}(u, v) &= \left| \widetilde{C}_{m,n}\left(\frac{u}{\lambda s}, \frac{v}{\lambda s}\right) \right|^2 \quad \text{(kernel)}. \end{aligned} \tag{4}$$

This property enables the usage of metasurfaces to perform convolutional operations by designing the modulation profiles $C_{m,n}$ that generate the specialized kernel $h_{m,n}$ that is desired [3, 10].

For convolutional kernels $h_{m,n}$ that contain negative values, we design two metasurfaces with modulation profiles $C_{m,n,+}$ and $C_{m,n,-}$, and approximate the kernel response by subtracting the two corresponding measurements. In addition, because metasurfaces are generally dispersive, the effective kernels vary with wavelength and are only partially correlated across the spectrum. To simplify the analysis, we restrict each metasurface to operate at a single wavelength of

incident light. This can be practically achieved by placing a narrow bandpass filter in front of each metasurface.

To extend the design to CNNs that process RGB images, where each kernel $h_{m,n}$ consists of three channels, we construct three independent pairs of metasurfaces $C_{m,n,\pm,R}$, $C_{m,n,\pm,G}$, and $C_{m,n,\pm,B}$. Each pair transmits only a narrow spectral band (red, green, or blue) from the scene and is assumed to implement a kernel that remains constant within that band. Consequently, to approximate the first convolutional layer with L output channels for RGB inputs, the metasurface array requires $6L$ elements.

3.2. Direct Kernel Optimization

We optimize a pair of metasurface phase modulation profiles, $\varphi_{m,n,+}(x,y)$ and $\varphi_{m,n,-}(x,y)$, assuming uniform transmittance profiles $T_{m,n,\pm}(x,y)$ within a predefined circular aperture, to approximate a given single-channel target kernel $h_{m,n}$. The optimization is formulated as

$$\underset{\varphi_{m,n,\pm}(x,y)}{\arg\min} \; \|\text{Simulator}(\varphi_{m,n,+}(x,y)) - h_{m,n,\pm}(u,v)\|^2, \quad (5)$$

where

$$h_{m,n,\pm}(u,v) = \frac{\pm h_{m,n}(u,v) + |\pm h_{m,n}(u,v)|}{2}.$$

We adopted the D-Flat differentiable simulator to generate the kernel given the phase modulation profiles [14]. After determining the optimal phase modulation, we translate it into a metasurface geometry by performing a standard cell-based library search [18].

4. RESULTS AND ANALYSIS

In this paper, we focus on analyzing the proposed two-step strategy for designing opto-electronic CNNs and compare it with the traditional end-to-end strategy in simulation. The accuracy of the employed simulation process has been validated in our prior work [14], which showed that the simulated kernels closely matched those measured from fabricated metasurfaces designed with the same framework [10].

To facilitate analysis, we select monocular depth estimation as the target application for our study, and design the opto-electronic CNN based on a pre-defined architecture, Monodepth2 [19]. This architecture takes a single RGB image as input, and its first convolutional layer contains 64 channels. Consequently, a total of $384 = 64 \times 6$ metasurfaces need to be optimized to carry out the first layer operation optically. For each metasurface (m,n), the phase modulation profile $\varphi_{m,n}$ is parameterized as a 1025×1025 discrete 2D matrix with a pixel pitch of $2.5\,\mu\text{m} \times 2.5\,\mu\text{m}$. The spacing between the metasurface array and the photosensor, i.e., the sensor distance, is set to 10 mm. The resolution of the feature maps generated by each metasurface is 320×96. We adopt the KITTI dataset [20] for training, evaluation, and testing.

Computational cost of training. Under our parameter settings, the optical layer contains 403M trainable parameters. By comparison, the first convolutional layer implemented electronically would require only 9k parameters, and the entire Monodepth2 architecture has just 14M parameters in total. Thus, an end-to-end training strategy would necessitate optimizing more than 400M parameters jointly. In contrast, the proposed two-step strategy decouples the optimization of the optical layer from that of the electronic layers, and further breaks down the optical optimization into independent metasurface-level subproblems, substantially reducing the dimensionality of the search space (Table 2.)

Moreover, end-to-end optimization requires rendering the full feature map of the scene using variants of Eq. 3, whereas our two-step approach only evaluates the kernels (Eq. 5) without the rendering step. This greatly reduces the computational burden during backpropagation. The computational advantages of the proposed two-step strategy are summarized in Table. 2

Method	Parameters (M)	Time (ms)
DKO	1.1	100
Computational Training	14.84	250
E2E	418	73,000

Table 2: The number of trainable parameters (in millions), and the computational time (in milliseconds) for one forward pass and backward propagation for our DKO method (first row), the computational training of Monodepth2 (second row), and the E2E method (third row).

Direct kernel optimization. Our DKO approach successfully generates metasurfaces whose kernels closely match the target kernels. In Fig. 3, we show the learned kernels against the Monodepth2 target kernels, which show a close match. In practice, that means our metasurfaces would produce feature maps that are very close to those produced by the original model. Note that we show the final kernel with the negative kernels already subtracted from the positive kernels.

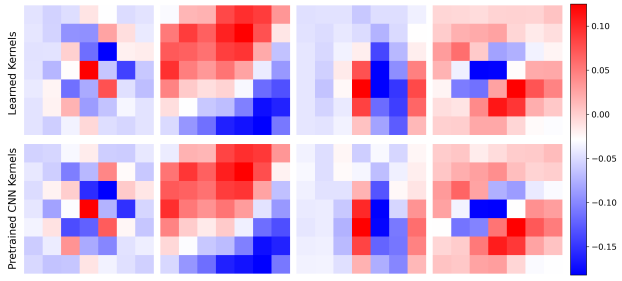

Fig. 3: **Top row:** A sample of our metasurface-learned kernels, and **bottom row:** the corresponding kernels from the pretrained Monodepth2 model. Our optimized metasurfaces learn PSFs that closely match the original model's kernels.

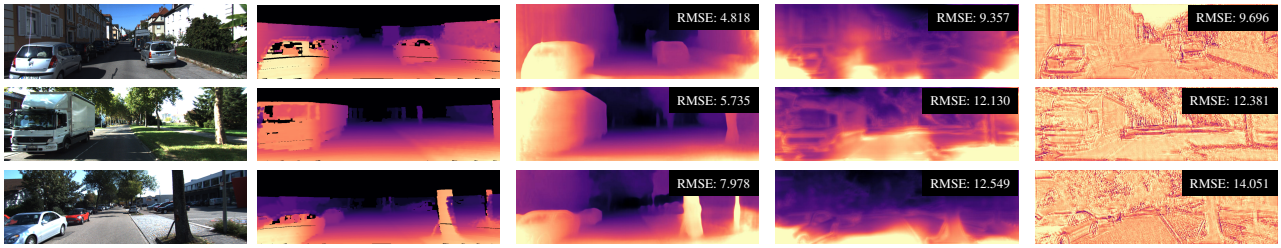

Fig. 2: Qualitative results on the KITTI dataset. First column: input image, second column: sparse ground truth depth map, third column: our method with metasurfaces replacing the first layer, fourth column: E2E with pretrained initialization, fifth column: E2E without pretraining. RMSE results of these individual predictions are reported in the insets. *Note* that the ground truth KITTI dataset has sparse depth maps collected from lidar, so we densified the ground truth images for visualization only.

Experiment	AbsRel	SqRel	RMSE	RMS$_{log}$	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$	Time
Ours	**0.199**	**1.674**	**6.996**	**0.305**	**0.688**	**0.879**	**0.944**	12h
E2E$_{\text{pretrained}}$	0.346	3.618	11.013	0.494	0.401	0.675	0.835	12h
E2E$_{\text{w/o pretraining}}$	0.443	4.758	12.083	0.587	0.303	0.561	0.766	12h

Table 1: Quantitative results using our method with our metasurfaces replacing the first layer (first row), the E2E approach with a pretrained Monodepth2 model with frozen weights (second row), and the E2E without a pretrained model (third row). All methods were given 12 hours to train, including the training for the depth model and optimizing the metasurfaces.

NCC↑	RMSE ↓	MAE ↓
0.9840	0.012909	0.007555

Table 3: Average normalized cross correlation (NCC), root mean square error (RMSE), and mean absolute error (MAE) for all 64×6 kernels produced by our optimized metasurface against the ground truth Monodepth2 first layer kernels.

In Table. 3, we report quantitative metrics over all the learned kernels, which validates our claim.

Depth estimation. For a high-dimensional task like depth estimation, our two-stage strategy demonstrates a significant advantage in computational efficiency. We successfully trained the depth model from scratch and optimized the metasurfaces kernels in less than 12 hours on a single Nvidia A100 GPU. The E2E approach, however, failed to yield meaningful results in the same timeframe due to its immense computational demands. We report the qualitative and quantitative results in Fig. 2 and Table. 1, respectively.

The E2E process requires optimizing hundreds of millions of optical parameters simultaneously, involving costly convolutions and backpropagation steps for every batch. We estimate that for the E2E method to converge, it would require about 30 days of training on the same hardware.

The prohibitive cost explains why others resorted to using a cluster of more than 30 GPUs to make E2E tractable [21]. Furthermore, even when simplifying the task by initializing with a pretrained depth model, the E2E method's computational burden remained too high to produce results comparable to our DKO approach (Table. 1).

5. CONCLUSION

In this work, we introduced a two-stage Direct Kernel Optimization (DKO) strategy to overcome the prohibitive computational cost of end-to-end training for hybrid opto-electronic networks. Our approach trains a CNN or uses a pretrained one, and then directly optimizes a metasurface to replicate the kernels of its first layer, effectively replacing the first digital convolutional layer with an optical one.

While previous methods have applied opto-electronic systems for simpler tasks such as object or hand-written digit classifications, we demonstrated our approach on monocular depth estimation, a higher-dimensional task where the E2E training is intractable within a reasonable time.

By decoupling the optical design from the network training and then further breaking down the optical optimization into individual metasurface-level subproblems, our strategy significantly reduces the computational complexity. Although we demonstrated our DKO for depth estimation, it is generalizable to other tasks and hybrid opto-electronic systems. This efficient paradigm makes the development of hybrid vision systems more accessible and paves the way for leveraging optical computing across a wider range of applications.

6. REFERENCES

[1] Jeremy Klotz and Shree K Nayar, "Minimalist vision with freeform pixels," in *European Conference on Computer Vision*. Springer, 2024, pp. 329–346.

[2] Vincent Sitzmann, Steven Diamond, Yifan Peng, Xiong Dun, Stephen Boyd, Wolfgang Heidrich, Felix Heide, and Gordon Wetzstein, "End-to-end optimization of optics and image processing for achromatic extended depth of field and super-resolution imaging," *ACM Transactions on Graphics (TOG)*, vol. 37, no. 4, pp. 1–13, 2018.

[3] Hanyu Zheng, Quan Liu, You Zhou, Ivan I. Kravchenko, Yuankai Huo, and Jason Valentine, "Meta-optic accelerators for object classifiers," *Science Advances*, vol. 8, no. 30, pp. eabo6410, July 2022.

[4] Xing Lin, Yair Rivenson, Nezih T. Yardimci, Muhammed Veli, Yi Luo, Mona Jarrahi, and Aydogan Ozcan, "All-optical machine learning using diffractive deep neural networks," *Science*, vol. 361, no. 6406, pp. 1004–1008, September 2018.

[5] Yicheng Wu, Vivek Boominathan, Huaijin Chen, Aswin Sankaranarayanan, and Ashok Veeraraghavan, "Phasecam3d—learning phase masks for passive single view depth estimation," in *2019 IEEE International Conference on Computational Photography (ICCP)*. IEEE, 2019, pp. 1–12.

[6] Kaixuan Wei, Xiao Li, Johannes Froech, Praneeth Chakravarthula, James Whitehead, Ethan Tseng, Arka Majumdar, and Felix Heide, "Spatially varying nanophotonic neural networks," November 2024.

[7] Gordon Wetzstein, Aydogan Ozcan, Sylvain Gigan, Shanhui Fan, Dirk Englund, Marin Soljačić, Cornelia Denz, David AB Miller, and Demetri Psaltis, "Inference in artificial intelligence with deep optics and photonics," *Nature*, vol. 588, no. 7836, pp. 39–47, 2020.

[8] Qi Guo, Zhujun Shi, Yao-Wei Huang, Emma Alexander, Cheng-Wei Qiu, Federico Capasso, and Todd Zickler, "Compact single-shot metalens depth sensors inspired by eyes of jumping spiders," *Proceedings of the National Academy of Sciences*, vol. 116, no. 46, pp. 22959–22965, 2019.

[9] Yuxuan Liu and Qi Guo, "Metah2: A snapshot metasurface hdr hyperspectral camera," in *2025 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2025, pp. 1918–1923.

[10] Dean Hazineh, Soon Wei Daniel Lim, Qi Guo, Federico Capasso, and Todd Zickler, "Polarization multi-image synthesis with birefringent metasurfaces," in *2023 IEEE International Conference on Computational Photography (ICCP)*. IEEE, 2023, pp. 1–12.

[11] Noah A Rubin, Gabriele D'Aversa, Paul Chevalier, Zhujun Shi, Wei Ting Chen, and Federico Capasso, "Matrix fourier optics enables a compact full-stokes polarization camera," *Science*, vol. 365, no. 6448, pp. eaax1839, 2019.

[12] Wanxin Shi, Zheng Huang, Honghao Huang, Chengyang Hu, Minghua Chen, Sigang Yang, and Hongwei Chen, "LOEN: Lensless opto-electronic neural network empowered machine vision," *Light: Science & Applications*, vol. 11, no. 1, pp. 121, May 2022.

[13] Zhiwei Xue, Tiankuang Zhou, Zhihao Xu, Shaoliang Yu, Qionghai Dai, and Lu Fang, "Fully forward mode training for optical neural networks," *Nature*, vol. 632, no. 8024, pp. 280–286, 2024.

[14] Dean S. Hazineh, Soon Wei Daniel Lim, Zhujun Shi, Federico Capasso, Todd Zickler, and Qi Guo, "D-flat: A differentiable flat-optics framework for end-to-end metasurface visual sensor design," *arXiv preprint arXiv:2207.14780*, August 2022.

[15] Qi Guo, Huixuan Tang, Aaron Schmitz, Wenqi Zhang, Yang Lou, Alexander Fix, Steven Lovegrove, and Hauke Malte Strasdat, "Raycast calibration for augmented reality hmds with off-axis reflective combiners," in *2020 IEEE International Conference on Computational Photography (ICCP)*. IEEE, 2020, pp. 1–12.

[16] Zheng Ren, Jingwen Zhou, Wenguan Zhang, Jiapu Yan, Bingkun Chen, Huajun Feng, and Shiqi Chen, "Successive optimization of optics and post-processing with differentiable coherent psf operator and field information," *IEEE Transactions on Computational Imaging*, 2025.

[17] Joseph W Goodman, *Introduction to Fourier optics*, Roberts and Company publishers, 2005.

[18] Charles Brookshire, Yuxuan Liu, Yuanrui Chen, Wei Ting Chen, and Qi Guo, "MetaHDR: single shot high-dynamic range imaging and sensing using a multifunctional metasurface," *Optics Express*, vol. 32, no. 15, pp. 26690–26707, July 2024.

[19] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J Brostow, "Digging into self-supervised monocular depth estimation," in *Proceedings of the IEEE/CVF international conference on computer vision*, October 2019, pp. 3828–3838.

[20] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun, "Vision meets robotics: The kitti dataset," *The international journal of robotics research*, vol. 32, no. 11, pp. 1231–1237, 2013.

[21] Jipeng Sun, Kaixuan Wei, Thomas Eboli, Congli Wang, Cheng Zheng, Zhihao Zhou, Arka Majumdar, Wolfgang Heidrich, and Felix Heide, "Collaborative on-sensor array cameras," *ACM Trans. Graph.*, vol. 44, no. 4, July 2025.