Neural Tangent Knowledge Distillation for Optical Convolutional Networks

Jinlin Xiang[*‡] **Minho Choi**[*‡¶] **Yubo Zhang**[‡] **Zhihao Zhou**[‡] **Arka Majumdar**[‡§‖]

Eli Shlizerman[†‡‖]

Abstract

Hybrid Optical Neural Networks (ONNs, typically consisting of an optical frontend and a digital backend) offer an energy-efficient alternative to fully digital deep networks for real-time, power-constrained systems. However, their adoption is limited by **two main challenges**: the accuracy gap compared to large-scale networks during training, and discrepancies between simulated and fabricated systems that further degrade accuracy. While previous work has proposed end-to-end optimizations for specific datasets (e.g., MNIST) and optical systems, these approaches typically lack generalization across tasks and hardware designs. To address these limitations, we propose a task-agnostic and hardware-agnostic pipeline that supports image classification and segmentation across diverse optical systems. To assist optical system design before training, we design the metasurface layout based on fabrication constraints. For training, we introduce Neural Tangent Knowledge Distillation (NTKD), which aligns optical models with electronic teacher networks, thereby narrowing the accuracy gap. After fabrication, NTKD also guides fine-tuning of the digital backend to compensate for implementation errors. Experiments on multiple datasets (e.g., MNIST, CIFAR, Carvana Image Masking Dataset) and hardware configurations show that our pipeline consistently improves ONN performance and enables practical deployment in both pre-fabrication simulations and physical implementations.

1 Introduction

Optical Neural Networks (ONNs) offer a promising approach to achieve efficient computation and energy use compared to digital implementations such as Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs), making them well-suited for resource-constrained, real-time physical systems [1]. For example, ONNs have been proposed for power-limited applications (illustrated in Figure 1.a), including satellites [2], unmanned aerial vehicles [3], smart home devices [4], autonomous driving systems [5], wearable electronics [6], and medical devices [7].

Among different ONN implementations, hybrid optical-electronic architectures are practical options under current hardware constraints [8]. In such systems, the optical frontend accelerates computation at the speed of light, while the digital backend refines predictions to improve robustness [9]. The

[*]These authors contributed equally to this work.

[†]Department of Applied Mathematics, University of Washington, Seattle, WA 98195, USA.

[‡]Department of Electrical and Computer Engineering, University of Washington, Seattle, WA 98195, USA.

[§]Department of Physics, University of Washington, Seattle, WA 98195, USA.

[¶]Department of Electrical Engineering, Ulsan National Institute of Science and Technology, Ulsan 44919, South Korea

[‖]Corresponding authors: arka@uw.edu, shlizee@uw.edu

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

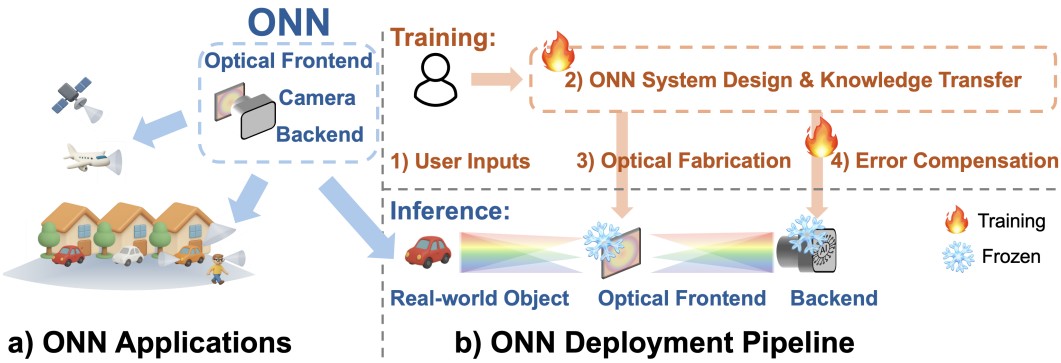

Figure 1: Overview of potential applications for ONNs and our proposed deployment pipeline. (a) ONNs for real-time decision-making in power-constrained scenarios. (b) Our proposed pipeline includes user-driven design, knowledge transfer training, fabrication, and error compensation.

optical frontend generally consists of a single linear layer (as shown in Figure 1.a), since: (1) implementing nonlinear activation functions in physical optics remains extremely challenging due to material and device limitations; and (2) without nonlinearity, multiple linear transformations can be mathematically compressed into a single linear transformation. Moreover, from a theoretical standpoint, the universal approximation property could still be satisfied by shallow networks, suggesting that hybrid ONNs retain sufficient expressive power for a wide range of tasks when appropriately optimized [10–12].

Despite their promises, ONNs remain difficult to design and train due to both **architectural limitations** and **fabrication-related challenges**. First, existing ONN architectures are typically significantly simpler than modern deep CNNs or ViTs. These simplifications cannot be directly obtained through pruning or quantization [13–15]. Second, physical fabrication and experimental deployment inevitably introduce various sources of noise, such as optical misalignment, material variability, and measurement noise, further degrading performance [16]. While some end-to-end optimization strategies have been proposed to address these challenges, they are typically designed for a specific dataset (e.g., MNIST) and tailored to a particular optical system, rather than providing a generalized solution (as also summarized in Related Works). In contrast, we aim to develop a task-agnostic and hardware-agnostic pipeline that can generalize across different datasets and optical hardware setups.

To address these challenges, knowledge transfer, particularly Knowledge Distillation (KD), offers a promising solution by transferring knowledge from pre-trained digital networks to optical models [17, 18]. Moreover, recent work shows that successful KD implicitly leads to student-teacher Neural Tangent Kernel (NTK) similarity, where NTK captures how the network's predictions change with respect to small changes in its parameters [19]. As we show here, utilizing NTK for matching is particularly effective for ONNs, as the NTK provides a linear approximation of network behavior, naturally aligning with the linear operations performed by optical systems.

Thus, we propose a Neural Tangent Knowledge Distillation (NTKD) pipeline that generalizes across different optical network designs and datasets to support multiple tasks such as classification and segmentation (also shown in Figure 1.b). The pipeline starts with specifying the task, the dataset, and the optical structure. Then, NTKD optimization transfers knowledge from digital teacher models to hybrid ONNs by matching their NTKs, effectively transferring the relational structure between classes rather than just matching final predictions. Furthermore, the pipeline compensates for errors introduced during fabrication and experimental deployment by aligning the student's and teacher's NTKs through a small fraction (e.g., 10%) of real experimental data.

In summary, our contributions are as follows:

- We introduce a Neural Tangent Knowledge Distillation (NTKD) pipeline that supports diverse tasks and optical structures, addressing the challenges of shallow architectures and physical imperfections.
- We experimentally validate our pipeline with different ONN implementations on both classification and segmentation tasks, demonstrating its effectiveness through both simulations and fabrications.

Table 1: Summary of previous ONN works categorized by task type (classification, segmentation) and implementation level—either simulation only (denoted as **Sim**) or with physical fabrication and experimental validation (denoted as **Fab**).

ONN Capability	Works	Classification (Sim)	Classification (Fab)	Segmentation (Sim)	Segmentation (Fab)
Monochromatic	2018–2025: [9, 16, 20–27, 30–38]	✓	✓	✗	✗
Polychromatic	2023–2025: [16, 28, 29, 39, 40]	✓	✓	✗	✗
	ExtremeMETA (2025) [5]	✗	✗	✓	✗
	Ours (NTKD)	✓	✓	✓	✓

• We leverage NTK analysis to estimate the achievable accuracy of given hybrid ONNs, providing theoretical guidance on their design and optimization.

2 Related Works

ONN Tasks and Implementations: Table 1 categorizes ONN applications into two tasks (classification and segmentation) and two optical implementations (monochromatic and polychromatic systems). Most previous work focused on monochromatic ONNs for MNIST image classification, including fully optical systems that performed linear transformations [20, 21], physically nonlinear ONNs that used atomic vapors or intensifiers [22–24], and hybrid architectures that combined an optical frontend with a digital backend [25–27]. Previous polychromatic ONNs for classification were limited to small datasets such as CIFAR-10, as ONN architectures faced challenges in scaling to complex benchmarks [28, 29]. Segmentation tasks are still in the early stages, with a previous study based only on simulation [5]. Our work considers both classification and segmentation tasks. In our pipeline, image reconstruction is implicitly incorporated by encouraging the optical frontend output to align with the simulated result, as the physical output deviates from simulation and requires correction.

Transfer learning for ONNs: Transfer learning, particularly Knowledge Distillation (KD), offers a promising solution for transferring knowledge from pre-trained digital networks to optical models [17, 41]. KD minimizes the divergence between a compact student model's predictions and those of a pre-trained teacher model, thereby encouraging the student to actively mimic the teacher's behavior [17, 42]. Beyond conventional KD, recent studies have shown NTK-based approaches to understand knowledge transfer [43]. For example, theoretical insights into KD transfer risk and data efficiency in wide networks have been established through NTK analysis [44]. Subsequent work demonstrated that successful knowledge distillation implicitly led to student-teacher NTK alignment [19], and NTK similarity was further applied to quantify task affinities in multi-task learning [45–47]. In contrast, our work targets physically constrained ONNs, and introduces an explicit NTK matching strategy to directly guide the distillation process from a digital teacher to an optical student.

Compensation Strategies for Practical ONNs: Fabrication imperfections and system noise in physical ONNs often lead to significant performance drops compared to simulations. Some approaches used deep learning to model the system directly in a data-driven manner [48], including ONN auto-learning [49,50], where ONNs were trained to fit experimental input-output mappings. Other methods introduced physical information via hardware-in-the-loop training. For example, physics-constrained frameworks embedded fabrication-aware models and losses to better align learning with optical behavior [16]. Moreover, some approaches avoided simulation entirely by randomly fabricating optical kernels and training a digital backend to adapt to the fixed frontend structure [51–55]. This simplified fabrication but placed the learning burden entirely on the backend. It is also not clear if such random surfaces perform better than an ordinary lens. In contrast, our work identifies sources of physical errors and introduces an NTK alignment strategy for effective compensation.

3 Methods

3.1 Optical Frontend Design

At the initialization of the pipeline, user inputs are required to define the optical system (also shown in Figure 2.1). Specifically, the user specifies (1) the physical size of the optical frontend (e.g., the

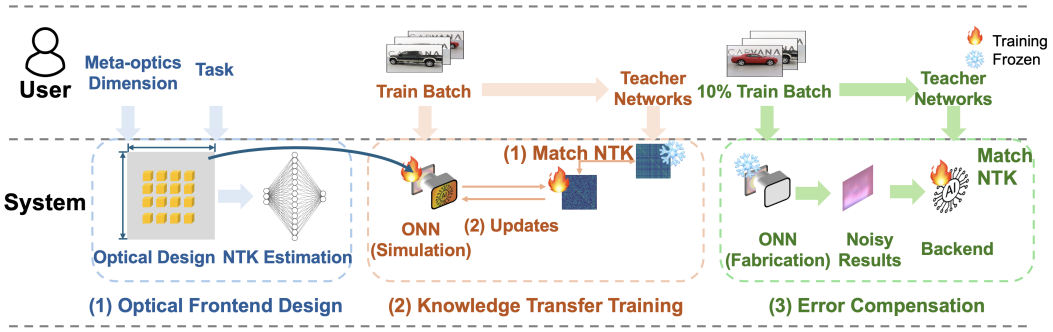

Figure 2: Overview of the pipeline. It consists of three steps: (1) Optical Frontend Design based on user-specified inputs, (2) Knowledge Transfer Training using Neural Tangent Kernel (NTK) matching, and (3) Error Compensation for fabricated optical frontends.

number of meta-optic kernels), (2) the target dataset for the task, and (3) the desired network structure, such as the number of layers and channels. Prior works have demonstrated that optical convolution can be physically realized using either a 4f system or a PSF-based free-space propagation system [18, 30, 38, 56]. In this work, we practically implement a PSF-based metasurface design due to its advantages in compactness, alignment robustness, and ease of fabrication [29, 30].

Optical Frontend Layout: We consider a metasurface of size (h, w), onto which we aim to place $n_{kernels}$ square optical kernels, each of size k (in mm), with a minimum edge-to-edge spacing d to satisfy fabrication constraints. To compute the maximum number of kernels that can be placed while preserving symmetry, we define

$$n_{\text{cols}} = \left\lfloor \frac{w-d}{k+d} \right\rfloor, \quad n_{\text{rows}} = \left\lfloor \frac{h-d}{k+d} \right\rfloor, \quad n_{kernels} = n_{\text{cols}} \times n_{\text{rows}}. \tag{1}$$

Performance Estimation: Once the physical layout of the ONN is determined, we aim to estimate its expected performance without empirical training. We adopt the Neural Tangent Kernel (NTK) framework, which captures the training dynamics of **infinitely wide neural networks** under gradient descent. In particular, we introduce a reference network that shares the same architecture as the designed ONN (e.g., number of layers and connectivity) but has infinite width at each layer. Under this assumption, the predictions of the reference network correspond to NTK regression [57, 58]. Let the reference network $f(x;\theta)$ be a neural network parameterized by θ, which maps an input x to an output $f(x;\theta)$. The NTK is defined as

$$\Theta(x, x') = \nabla_\theta f(x;\theta)^\top \nabla_\theta f(x';\theta), \tag{2}$$

where $\nabla_\theta f(x;\theta)$ is the Jacobian of the network output with respect to its parameters. Let $\{x_i^{\text{train}}, y_i^{\text{train}}\}_{i=1}^{n_{\text{train}}}$ be the training data and $\{x_i^{\text{test}}\}_{i=1}^{n_{\text{test}}}$ be the test data. We compute

$$\Theta_{\text{train,train}} = \Theta(x^{\text{train}}, x^{\text{train}}) \in \mathbb{R}^{n_{\text{train}} \times n_{\text{train}}}, \quad \Theta_{\text{test,train}} = \Theta(x^{\text{test}}, x^{\text{train}}) \in \mathbb{R}^{n_{\text{test}} \times n_{\text{train}}}. \tag{3}$$

and use kernel regression to predict outputs on the test set

$$f(x^{\text{test}};\theta) = \Theta_{\text{test,train}} \left(\Theta_{\text{train,train}} + \lambda I\right)^{-1} y^{\text{train}}, \tag{4}$$

where λ is a regularization parameter, which is selected via grid search on a validation set.

The NTK-based performance estimation serves as a diagnostic tool to evaluate whether the specified ONN architecture is expressive enough for the given task. While this estimation is not used for training or loss computation, it provides an early signal to guide architectural decisions and allows users to iteratively refine the optical design before full training and fabrication. For example, if the estimated test accuracy is much lower than the expected performance, it may suggest a mismatch between the ONN's capacity (e.g., depth) and task complexity.

3.2 Knowledge Transfer Training

After the user specifies the ONN architecture, we train the system for the target task. We define a supervised learning problem with input-output pairs (x, y), where x represents the input samples and y denotes the corresponding ground-truth labels. The network parameters (θ), including the optical frontend and the digital backend, are initialized and optimized jointly (shown in Figure 2.2).

End-to-end loss: The first loss term we consider is a standard end-to-end supervision loss, which directly minimizes the discrepancy between the network's predictions and the ground-truth labels. Formally, we optimize the following objective

$$\mathcal{L}_{\text{E2E}} = \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[\ell(f_{ONN}(x;\theta), y)\right], \tag{5}$$

where $\ell(\cdot,\cdot)$ is a standard loss function such as cross-entropy for classification tasks, $f_{ONN}(x;\theta)$ denotes the output of the network with parameters θ, and $\mathcal{D}$ represents the training dataset.

Neural Tangent Knowledge Distillation (NTKD) Loss: In addition to the standard end-to-end loss, we introduce a knowledge transfer loss based on Neural Tangent Kernel (NTK). Specifically, we assume access to a pretrained teacher network, such as LeNet for MNIST, AlexNet for CIFAR-10, or U-Net for image segmentation tasks.

Given a minibatch of input samples $\{x_i\}_{i=1}^{n_{\text{batch}}}$, we compute the Jacobian matrices of both the teacher network (f_{teacher}) and student ONN (f_{ONN}) with respect to their parameters ($\theta_{\text{teacher}}, \theta_{\text{ONN}}$)

$$J_{\text{teacher}} = \left[\frac{\partial f_{\text{teacher}}(x_i)}{\partial \theta_{\text{teacher}}}\right]_{i=1}^{n_{\text{batch}}}, \quad J_{\text{ONN}} = \left[\frac{\partial f_{\text{ONN}}(x_i)}{\partial \theta_{\text{ONN}}}\right]_{i=1}^{n_{\text{batch}}}. \tag{6}$$

The Jacobians $J_{\text{teacher}} \in \mathbb{R}^{n_{\text{batch}}\times p_{\text{teacher}}\times n_{\text{class}}}$ and $J_{\text{ONN}} \in \mathbb{R}^{n_{\text{batch}}\times p_{\text{ONN}}\times n_{\text{class}}}$ may differ in width depending on the number of trainable parameters in each network. Here, p_{teacher} and p_{ONN} denote the number of parameters in the teacher and ONN, respectively. Their corresponding NTK matrices,

$$\Theta_{\text{teacher}} = J_{\text{teacher}} J_{\text{teacher}}^{\top}, \quad \Theta_{\text{ONN}} = J_{\text{ONN}} J_{\text{ONN}}^{\top}, \tag{7}$$

are both of size $n_{\text{batch}} \times n_{\text{batch}}$. We define the NTKD loss by minimizing the discrepancy between the NTK matrices of the teacher network and the ONN (e.g., MSE)

$$\mathcal{L}_{\text{NTKD}} = \mathbb{E}_{\{x_i\}_{i=1}^{n_{\text{batch}}}\sim\mathcal{D}} \left[\ell\left(\Theta_{\text{teacher}}, \Theta_{\text{ONN}}\right)\right]. \tag{8}$$

Then, we minimize a weighted sum of two losses, controlled by hyperparameters α and β

$$\min_{\theta} \left(\alpha\mathcal{L}_{\text{E2E}} + \beta\mathcal{L}_{\text{NTKD}}\right). \tag{9}$$

3.3 Error Compensation

The physical fabrication uses the optimized simulation parameters obtained through the process described in Section 3.2. The fabrication fixes the optical frontend, and only the digital backend remains tunable. Due to unavoidable fabrication and experimental errors, discrepancies arise between the designed and realized optical system. Given an input image a and convolution kernel k, the ideal output is $y = a * k$. The fabricated optical convolution output with noise at location (i, j) is

$$\tilde{y}_{i,j} = \alpha\beta \sum_{m=1}^{k_{size}} \sum_{n=1}^{k_{size}} a_{i+m-1,j+n-1} \left(k_{m,n} + \delta_{m,n}\right) + \epsilon_{i,j}. \tag{10}$$

Here, the scaling factors α (image brightness) and β (image–kernel misalignment) can be experimentally calibrated to match the designed system, while the sensor noise ϵ is primarily determined by the imaging device characteristics. We further quantify the impact of fabrication noise δ (with proof provided in the Supplementary). In particular, we show that perturbations in the NTK caused by kernel fabrication errors (δ_{ij}) scale as

$$\|\Delta\Theta_{\text{ONN}}\| \sim O\left(\frac{\|\delta\|}{m}\right), \tag{11}$$

where $\Delta\Theta_{\text{ONN}}$ denotes the NTK perturbation, and m is the number of kernels (i.e., the network width). This result indicates that networks with more kernels are inherently more robust to fabrication

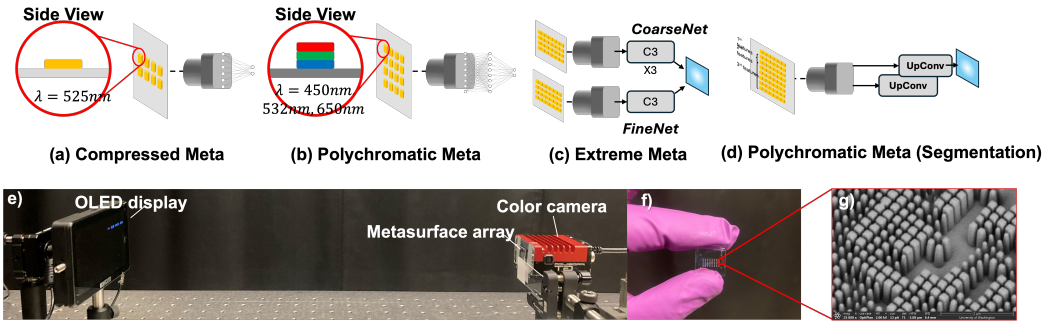

Figure 3: Optical systems. (a) Compressed Meta ONN on MNIST [37]; (b) Polychromatic Meta ONN on CIFAR-10 [29]; (c) ExtremeMETA for segmentation with dual optical frontends [5]; (d) Customized Polychromatic Meta ONN for segmentation (Ours); (e) Optical measurement setup; (f) Fabricated PSF-engineered meta-optics; (g) Scanning electron microscopy image of the meta-optics.

Table 2: Teacher and student network details.

Student Architecture	Teacher Model	Teacher Accuracy	Teacher Size
Compressed Meta (0.05M)	LeNet	99.1%	∼0.18M
Polychromatic Meta (1.62M)	AlexNet	85.4%	233.29M
Polychromatic Meta (1.06M)	U-Net	95.4%	196.97M

noise. Indeed, if $m \gg \|\delta\|$, the impact of fabrication noise on the prediction can be considered negligible. If δ is interpreted as a gradient descent step, Eq. 11 is consistent with NTK theory: as $m \to \infty$, the NTK (Θ) remains constant during training, implying $\Delta\Theta \to 0$ [57].

To correct these errors, we re-apply minimization in Eq. 9, but restrict optimization to the unfrozen backend parameters (shown in Figure 2.3). The teacher network takes the raw input images and computes its NTK matrix over a batch of samples, as defined in Eq. 7. The student network receives feature maps from the fixed optical frontend, which processes the same batch of input images and produces an NTK matrix of size $n_{\text{batch}} \times n_{\text{batch}}$.

4 Results

4.1 Implementation Details, Pre-trained Teachers, Datasets and Evaluation Metrics

Implementation Details: Figure 3 demonstrates four different optical systems used in the experiments. For monochromatic image classification, we conducted experiments on the Compressed Meta ONN architecture [37] using the MNIST dataset [59]. This system consists of a single optical frontend with 8 kernels (7×7) and a compact digital backend composed of two fully connected layers. For polychromatic image classification, we evaluated the Polychromatic Meta ONN [29] on the CIFAR-10 dataset [60]. This model consists of a single optical frontend with 16 kernels (7×7) and a digital backend consisting of three fully connected layers.

For image segmentation, we performed experiments on both the Extreme Meta ONN [5] and a modified version of the Polychromatic Meta ONN [29], using Kaggle's Carvana Image Masking dataset. ExtremeMETA consists of two parallel polychromatic optical frontends, followed by a dual-path digital backend composed of CoarseNet for global feature extraction and FineNet for detail enhancement, with their outputs fused to produce the final segmentation. Our customized polychromatic segmentation system extends the Polychromatic Meta ONN by incorporating a single optical frontend with 56 kernels (8, 16, and 32 kernels to capture hierarchical depth representations, each of size 3×3) and a backend composed of upconvolutional layers to support dense prediction tasks. For a fair comparison between the two systems, we matched the number of optical kernels. Optical implementation details are in the Supplementary.

In experiments, we manipulate polychromatic−red, green, and blue−point spread functions (PSFs) of the meta-optics using a gradient descent algorithm. Physical shapes and dimensions of the PSF-

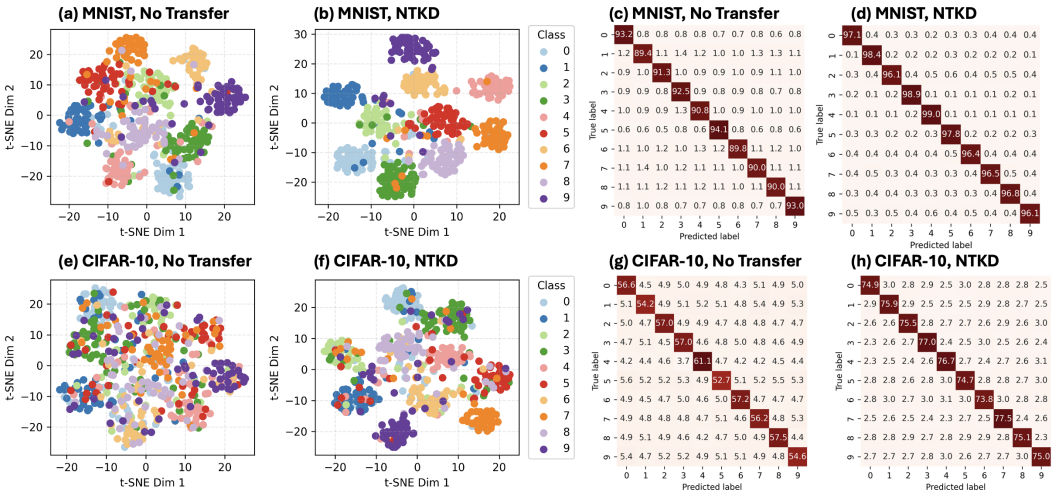

Figure 4: Simulation: t-SNE and confusion matrices in MNIST (a-d) and CIFAR-10 (e-h).

engineered meta-optics are shown in Figure 3. Since the meta-optics are designed with PSFs that function as convolutional kernels, optical convolution occurs naturally during image capture. Our experimental setup is straightforward: we replace a conventional imaging lens with PSF-engineered meta-optics. The meta-optics are carefully positioned and aligned with the color camera, enabling the capture of PSFs and convolved images using a laser pointer and an OLED display, respectively. A schematic representation and a photograph of the setup are provided in Figure 3.

Computing the NTK explicitly via Jacobian–Jacobian products is memory-intensive and infeasible at scale. To address this, we adopt the approximation strategy of NTK-SAP [61–63], which estimates the NTK trace rather than constructing the full matrix. We use batch size of 128 for MNIST and CIFAR-10/100, batch size of 64 for ImageNet-100, and batch size of 8 for segmentation tasks.

Pre-trained teachers and Datasets: The MNIST and CIFAR-10 datasets each consist of $50,000$ training images and $10,000$ testing images. The Carvana dataset, originally introduced in Kaggle's Carvana Image Masking Challenge, contains 5,088 high-resolution 1920×1280 car images. We adopt LeNet (99.1% accuracy on MNIST), AlexNet (84.5% on CIFAR-10), and a full U-Net (95.4% mIoU on Segmentation) as teacher models for their respective tasks. Table 2 summarizes the teacher and student networks.

Evaluation Metrics: For classification tasks, we ran each experiment five times with different random seeds and reported the mean and standard deviation (std) of the classification accuracy. For segmentation tasks, we similarly conducted five independent runs and reported the mean Intersection over Union (mIoU) along with the standard deviation.

4.2 Main Results

Simulation Results: Table 3 summarizes the accuracy of different ONN training strategies in classification and segmentation tasks. For monochromatic classification, NTKD achieved 97.3% accuracy and outperformed both KD-based transfer (95.9%) and the non-transfer baseline (91.4%). Similar trends were observed in the more challenging polychromatic classification setting, where NTKD achieved 75.6%, surpassing KD (72.5%) and baseline (56.4%). For segmentation tasks, we evaluated models on both the Extreme Meta and our proposed Polychromatic Meta datasets. NTKD

Table 3: Performance comparison of ONN methods across different tasks and training strategies.

Methods	**Monochromatic Classification (%)** Compressed Meta (2025) [37]	**Polychromatic Classification (%)** Polychromatic Meta (2025) [29]	**Polychromatic Segmentation (mIoU)** Extreme Meta (2025) [5]	Polychromatic Meta (Ours)
Simulation, No Transfer	$91.4 \pm 0.8\%$	$56.4 \pm 1.9\%$	$68.3 \pm 0.5\%$	$74.3 \pm 0.4\%$
Simulation, KD	$95.9 \pm 0.6\%$	$72.5 \pm 2.1\%$	$75.3 \pm 0.2\%$	$86.7 \pm 0.4\%$
Simulation, NTKD	$97.3 \pm 0.6\%$	$75.6 \pm 0.9\%$	$80.1 \pm 0.2\%$	$91.5 \pm 0.4\%$

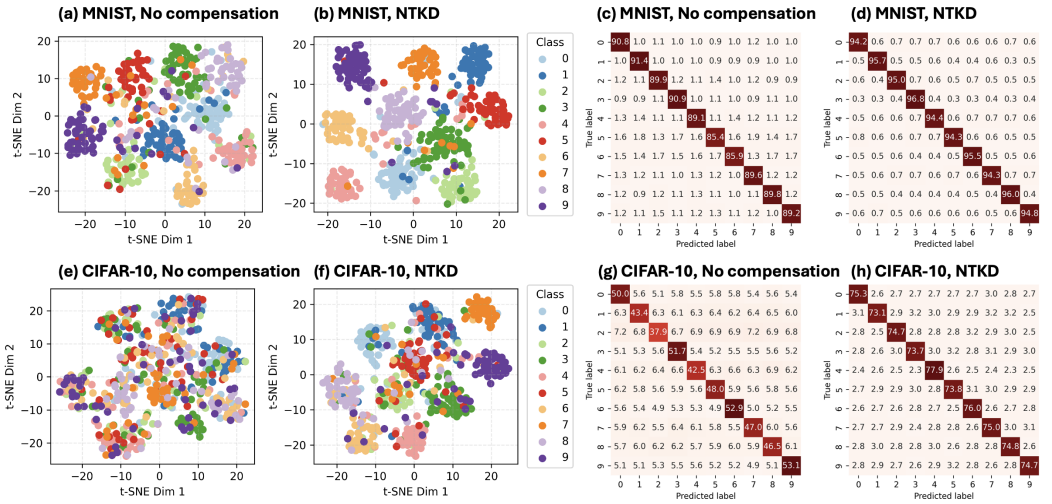

Figure 5: Fabrication: t-SNE and confusion matrices on MNIST (a–d) and CIFAR-10 (e–h).

consistently achieved higher mIoU scores across both optical systems, surpassing KD-based transfer and end-to-end training without transfer (75.3% and 86.7%, respectively).

These results indicate that training ONNs end-to-end without transfer often leads to suboptimal performance. Incorporating knowledge transfer through KD or NTKD improves learning efficacy and overall segmentation quality. In particular, the NTKD approach outperformed KD in different tasks, demonstrating its ability to guide representation learning in optical neural networks.

Figure 4 demonstrates knowledge transfer strategies on MNIST and CIFAR-10 representations and classification performance. Compared to the no-transfer baseline, these strategies improve class separability in t-distributed Stochastic Neighbor Embedding (t-SNE) visualizations and reduce noise in confusion matrices. NTKD transfer shows improved clustering and accuracy in experiments.

Fabrication and Compensation Results: Table 4 summarizes the impact of error compensation strategies on ONN performance in classification and segmentation tasks. Due to unavoidable fabrication and experimental errors, optical frontends suffer significant accuracy drops, especially in the polychromatic setting. Without compensation, the monochromatic system exhibits an 8.1% drop, and the more fabrication-sensitive polychromatic system shows a 28.3% drop on CIFAR-10 and a 41.8% reduction in mIoU on the Carvana segmentation task. This performance gap highlights the increased challenge of fabricating RGB-sensitive kernels in polychromatic ONNs compared to monochromatic ONNs. Our results demonstrate that knowledge transfer methods are able to assist with denoising that gap. NTKD compensation yields higher accuracy in both cases (95.1% for monochromatic, 74.9% for polychromatic and 81.2% for image segmentation task), outperforming end-to-end deep learning compensation, and validating its effectiveness in robust corrections for fabrication.

Figure 5 compares the t-SNE visualizations and confusion matrices of MNIST and CIFAR-10 representations under different compensation strategies. Without compensation (Figures 5 a, c, e, g), both optical systems exhibit class overlap in the feature space and reduced classification accuracy, primarily due to fabrication errors and optical misalignments. The NTKD correction (Figures 5 b, d, f, h) compensates for these errors and improves the clustering structure and classification accuracy, demonstrating robustness and generalization across datasets and optical systems.

Table 4: Evaluation of ONN error compensation methods on fabricated optical systems across both classification and segmentation tasks.

Method	**Monochromatic Classification (%)** Compressed Meta (2025) [37]	**Polychromatic Classification (%)** Polychromatic Meta (2025) [29]	**Polychromatic Segmentation (mIoU)** Polychromatic Meta (Ours)
No compensation (baseline)	89.2%	47.3%	49.7%
Error compensation (End-to-End)	93.2 ± 0.1%	70.4 ± 2.1%	62.7 ± 0.9%
Error compensation (NTKD)	95.1 ± 0.1%	74.9 ± 1.3%	81.2 ± 0.6%

Table 5: Random PSF kernel design across different tasks.

	Monochromatic Classification (%)	Polychromatic Classification (%)	Polychromatic Segmentation (mIoU)
8 kernels	96.12%	36.73%	49.3%
500 kernels	96.81%	49.32%	64.1%
1000 kernels	97.24%	56.23%	69.9%
∞ kernels	97.72%	67.33%	72.1%

4.3 Discussion and Ablation Study

Backend Complexity of ONNs: The complexity of the backend plays a critical role in the performance of optical systems, particularly in hybrid optical-digital architectures. When a strong digital backend is employed, it can effectively denoise and recover accurate outputs, even when the optical frontend introduces significant noise. However, a strong backend not only diminishes the contribution of the optical frontend but also significantly increases power consumption—undermining the core motivation for adopting optical computing in resource-constrained environments. In such scenarios, digital networks (e.g., ViT or U-Net) are often a more practical and effective choice than hybrid ONNs with disproportionately strong backends. For practical deployment, we need to carefully balance the computational load between optics and computational backend and find a tradeoff between acceptable energy consumption/ latency and acceptable accuracy, which will be an application dependent trade-off.

Random vs. Designed Parameters: Another possible direction for designing and training ONNs is to use randomly initialized optical parameters while training only the digital backends. This approach aims to avoid the need for extensive simulation and hardware-in-the-loop optimization of the optical frontends. We conducted experiments using a single optical convolutional layer and a lightweight backend—consisting of a single fully connected layer for classification, or a single upsampling layer for segmentation. As shown in Table 5, increasing the number of random kernels consistently improves performance across tasks. For example, in polychromatic classification, accuracy improves from 36.73% (8 kernels) to 56.23% (1000 kernels), and further to 67.33% in the NTK regime, which approximates an infinite number of random kernels. Similarly, in polychromatic segmentation, mIoU rises from 49.3% to 69.9% and reaches 72.1% under NTK estimation. These results demonstrate that while increasing the number of random PSFs improves performance, they still underperform our approach (designed kernels with knowledge transfer).

Scalability of ONNs: Scaling current ONNs remains challenging, as most designs rely on shallow structures with limited linear computational capacity. Implementing nonlinear operations in ONNs is especially difficult due to physical constraints, such as the limited pixel size. These hardware limitations make it hard for ONNs to support deep and expressive architectures like those used in digital networks. We observed that using different kernels to simulate multiple layers of a digital network leads to better performance, compared to simply compressing a deep CNN into a single-layer ONN.

Table 6: Impact of Teacher Complexity on NTK Distillation Performance for Classification and Segmentation Tasks.

Dataset	Task	Teacher	Student	Teacher Accuracy	Student Accuracy (with / without NTKD)
ImageNet-100	Classification	ResNet-18	Polychromatic Meta	78.43%	46.32% / 33.45%
ImageNet-100	Classification	ResNet-50	Polychromatic Meta	88.32%	47.86% / 33.45%
COCO-Stuff 10k	Segmentation	U-Net	Polychromatic Meta	61.89%	41.43% / 35.03%
COCO-Stuff 10k	Segmentation	ResU-Net	Polychromatic Meta	69.23%	42.73% / 35.03%

Table 6 examines the impact of stronger teachers on the student ONN under optical physical limitations, with additional experiments conducted using ResNet variants and more complex datasets such as ImageNet-100 and COCO-Stuff 10k. For the classification task, we employed both ResNet-18 and ResNet-50 to train a Polychromatic Meta student network. While both teacher models improved performance through NTK distillation, the gain from ResNet-18 to ResNet-50 was marginal with only an increase of 1.54%. Similarly, for segmentation, we used U-Net and ResU-Net as teacher networks to train a Polychromatic Meta-optical student on COCO-Stuff 10k for binary foreground-background segmentation. In this setting, the foreground includes all semantic object classes, and the background consists of non-object regions. Again, while both stronger teacher models provided improvements, the performance gain from a more complex teacher was limited.

These results indicate that bottleneck in performance is primarily due to the physical modeling limitations of current ONN hardware, rather than the complexity of the teacher model. In the case of ONN technology being improved in the future, as one would expect, in such a case, a stronger teacher network can provide additional gains in performance through the NTKD approach that will transfer knowledge from stronger teachers to enhanced students. In summary, scalable and expressive ONNs will ultimately rely on physical advances enabling deep and nonlinear optical computations.

Fabrication Analysis: Several factors may help explain the discrepancy between the designed and measured kernels. First, the local periodic approximation, which simplified the metasurface optics design by assuming the scatterers were arranged periodically, neglected the coupling between adjacent dissimilar scatterers and could introduce phase errors. Second, unavoidable fabrication errors further contributed to the observed discrepancy. Third, the pre-designed kernel needed to be properly matched to the sensor's pixel array; any misalignment between the metasurface optics and the camera could also lead to deformation of the measured kernels. Regarding the polychromatic versus single-wavelength kernels, metasurfaces inherently suffered from strong chromatic aberrations, a universal characteristic of diffractive optics. While we co-optimized the kernel across multiple wavelengths during the design process to ensure consistent behavior, the performance remained limited by the intrinsic material properties.

MACs and Power Consumption: We estimated the multiply–accumulate operations (MACs) and power consumption of hybrid ONNs using our polychromatic ONN as an example (details in the Supplementary). The total number of MAC operations in the simulated digital network is approximately 239 MMACs (while the full U-Net requires 65.9 GMACs and Efficient U-Net reaches 1.37 GMACs), which is reduced to 65 MMACs after incorporating optical frontends [64]. The total energy consumption includes both image capture and digital computation. The full U-Net (pre-trained teacher) consumes 2.03 J for computation and 2.36 mJ for image capture, totaling **2.04 J** per image. The compact digital network requires 7.37 mJ for computation and 2.36 mJ for image capture, totaling **9.73 mJ** per image. In contrast, our hybrid ONN consumes 3.82 mJ for image capture and 2.01 mJ for backend processing, totaling **5.83 mJ**—representing over a 40% reduction in system-level energy consumption compared to the simulated digital network, and over 300$\times$ energy compression compared to the pre-trained teacher U-Net.

5 Conclusion

We propose a comprehensive NTKD pipeline that addresses multiple tasks and multiple optical systems in ONN design, training, and compensation. By incorporating knowledge transfer, particularly NTK-based knowledge distillation, our framework consistently improves the accuracy of different optical systems across both classification and segmentation tasks. Based on extensive experiments, we observe that the current ONN performance is primarily limited by the shallow and linear nature of existing optical architectures. Future advances in deeper, nonlinear ONNs may help narrow the performance gap between optical and electronic neural networks, improving scalability and accuracy.

Acknowledgement

The research is supported by the National Science Foundation (EFRI-BRAID-2223495) and partial support from HDR Institute: Accelerated AI Algorithms for Data-Driven Discovery (A3D3) National Science Foundation grant PHY-2117997 (JX,ES). Part of this work was conducted at the Washington Nanofabrication Facility/ Molecular Analysis Facility, a National Nanotechnology Coordinated Infrastructure (NNCI) site at the University of Washington with partial support from the National Science Foundation via awards NNCI-1542101 and NNCI-2025489. The authors also acknowledge the partial support by the Departments of Electrical Computer Engineering, Applied Mathematics and Physics at the University of Washington. The authors are also thankful to the eScience Center at the University of Washington.

References

[1] Guibin Zhao, Pengfei Li, Zhibo Zhang, Fusen Guo, Xueting Huang, Wei Xu, Jinyin Wang, and Jianlong Chen. Towards sar automatic target recognition: Multi-category sar image classification

based on light weight vision transformer. In *2024 21st Annual International Conference on Privacy, Security and Trust (PST)*, pages 1–6. IEEE, 2024.

[2] Tingzhao Fu, Jianfa Zhang, Run Sun, Yuyao Huang, Wei Xu, Sigang Yang, Zhihong Zhu, and Hongwei Chen. Optical neural networks: progress and challenges. *Light: Science & Applications*, 13(1):263, 2024.

[3] Yingzhang Wu, Wenbo Li, Jie Zhang, Bangbei Tang, Jinlin Xiang, Shen Li, and Gang Guo. Driver's hand-foot coordination and global-regional brain functional connectivity under fatigue: Via graph theory and explainable artificial intelligence. *IEEE Transactions on Intelligent Vehicles*, 9(2):3493–3508, 2023.

[4] Adith Boloor, Weikai Lin, Tianrui Ma, Yu Feng, Yuhao Zhu, and Xuan Zhang. Privateeye: In-sensor privacy preservation through optical feature separation. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2357–2367. IEEE, 2025.

[5] Quan Liu, Brandon T Swartz, Ivan Kravchenko, Jason G Valentine, and Yuankai Huo. Extrememeta: High-speed lightweight image segmentation model by remodeling multi-channel metamaterial imagers. *Journal of Imaging Science and Technology*, pages 1–10, 2025.

[6] Baiheng Zhao, Junwei Cheng, Bo Wu, Dingshan Gao, Hailong Zhou, and Jianji Dong. Integrated photonic convolution acceleration core for wearable devices. *Opto-Electronic Science*, 2(12):230017–1, 2023.

[7] Jinlin Xiang, Hillol Sarker, Bozhao Qi, Ruisu Zhang, Roger Trullo, Salvatore Badalamenti, Maria Wiekowski, Annie Kruger, Etienne Pochet, Qi Tang, et al. Endoscopic scoring and localization in unconstrained clinical trial videos. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 4006–4015. IEEE, 2025.

[8] Runqin Xu, Pin Lv, Fanjiang Xu, and Yishi Shi. A survey of approaches for implementing optical neural networks. *Optics & Laser Technology*, 136:106787, 2021.

[9] Shane Colburn, Yi Chu, Eli Shilzerman, and Arka Majumdar. Optical frontend for a convolutional neural network. *Applied optics*, 58(12):3179–3186, 2019.

[10] Zhiyuan Lu, Huan Pu, Feicheng Wang, Zhiqiang Hu, and Liwei Wang. The expressive power of neural networks: A view from the width. In *NeurIPS*, 2017.

[11] Ding-Xuan Zhou. Universality of deep convolutional neural networks. *Applied and Computational Harmonic Analysis*, 48(2):787–794, 2020.

[12] Changshuo Bao, Qianxiao Li, and Haizhao Yang. Approximation analysis of convolutional neural networks from a feature extraction view. *Applied and Computational Harmonic Analysis*, 54:47–84, 2021.

[13] Song Han, Jeff Pool, John Tran, and William Dally. Learning both weights and connections for efficient neural network. *Advances in neural information processing systems*, 28, 2015.

[14] Pavlo Molchanov, Stephen Tyree, Tero Karras, Timo Aila, and Jan Kautz. Pruning convolutional neural networks for resource efficient inference. *arXiv preprint arXiv:1611.06440*, 2016.

[15] Mariia Seleznova, Dana Weitzner, Raja Giryes, Gitta Kutyniok, and Hung-Hsu Chou. Neural (tangent kernel) collapse. *Advances in Neural Information Processing Systems*, 36:16240–16270, 2023.

[16] Yanbing Liu, Jianwei Qin, Yan Liu, Xi Yue, Xun Liu, Guoqing Wang, Tianyu Li, Ye Ye, and Wei Li. Physics-constrained comprehensive optical neural networks. *Advances in Neural Information Processing Systems*, 37:92036–92054, 2024.

[17] Xinting Hu, Kaihua Tang, Chunyan Miao, Xian-Sheng Hua, and Hanwang Zhang. Distilling causal effect of data in class-incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3957–3966, 2021.

[18] Jinlin Xiang, Shane Colburn, Arka Majumdar, and Eli Shlizerman. Knowledge distillation circumvents nonlinearity for optical convolutional neural networks. *Applied Optics*, 61(9):2173–2183, 2022.

[19] Hrayr Harutyunyan, Ankit Singh Rawat, Aditya Krishna Menon, Seungyeon Kim, and Sanjiv Kumar. Supervision complexity and its role in knowledge distillation. *arXiv preprint arXiv:2301.12245*, 2023.

[20] 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*, 361(6406):1004–1008, 2018.

[21] Deniz Mengu, Yi Luo, Yair Rivenson, and Aydogan Ozcan. Analysis of diffractive optical neural networks and their integration with electronic neural networks. *IEEE Journal of Selected Topics in Quantum Electronics*, 26(1), 2020.

[22] Mingwei Yang, Elizabeth Robertson, Luisa Esguerra, Kurt Busch, and Janik Wolters. Optical convolutional neural network with atomic nonlinearity. *Optics Express*, 31(10):16451–16459, 2023.

[23] Albert Ryou, James Whitehead, Maksym Zhelyeznyakov, Paul Anderson, Cem Keskin, Michal Bajcsy, and Arka Majumdar. Free-space optical neural network based on thermal atomic nonlinearity. *Photonics Research*, 9(4):B128–B134, 2021.

[24] Tianyu Wang, Mandar M Sohoni, Logan G Wright, Martin M Stein, Shi-Yuan Ma, Tatsuhiro Onodera, Maxwell G Anderson, and Peter L McMahon. Image sensing with multilayer nonlinear optical neural networks. *Nature Photonics*, 17(5):408–415, 2023.

[25] Hanyu Zheng, Quan Liu, You Zhou, Ivan I Kravchenko, Yuankai Huo, and Jason Valentine. Meta-optic accelerators for object classifiers. *Science Advances*, 8(30):eabo6410, 2022.

[26] Tiankuang Zhou, Xing Lin, Jiamin Wu, Yitong Chen, Hao Xie, Yipeng Li, Jingtao Fan, Huaqiang Wu, Lu Fang, and Qionghai Dai. Large-scale neuromorphic optoelectronic computing with a reconfigurable diffractive processing unit. *Nature Photonics*, 15(5):367–373, 2021.

[27] Luocheng Huang, Quentin AA Tanguy, Johannes E Fröch, Saswata Mukherjee, Karl F Böhringer, and Arka Majumdar. Photonic advantage of optical encoders. *Nanophotonics*, 13(7):1191–1196, 2024.

[28] Zhiwei Xue, Tiankuang Zhou, Zhihao Xu, Shaoliang Yu, Qionghai Dai, and Lu Fang. Fully forward mode training for optical neural networks. *Nature*, 632(8024):280–286, 2024.

[29] Minho Choi, Jinlin Xiang, Anna Wirth-Singh, Seung-Hwan Baek, Eli Shlizerman, and Arka Majumdar. Transferable polychromatic optical encoder for neural networks. *Nature Communications*, 16(1):5623, 2025.

[30] Julie Chang, Vincent Sitzmann, Xiong Dun, Wolfgang Heidrich, and Gordon Wetzstein. Hybrid optical-electronic convolutional neural networks with optimized diffractive optics for image classification. *Scientific reports*, 8(1):1–10, 2018.

[31] Carlos Mauricio Villegas Burgos, Tianqi Yang, Yuhao Zhu, and A Nickolas Vamivakas. Design framework for metasurface optics-based convolutional neural networks. *Applied Optics*, 60(15):4356–4365, 2021.

[32] Zibo Hu, Shurui Li, Russell LT Schwartz, Maria Solyanik-Gorgone, Mario Miscuglio, Puneet Gupta, and Volker J Sorger. High-throughput multichannel parallelized diffraction convolutional neural network accelerator. *Laser & Photonics Reviews*, 16(12):2200213, 2022.

[33] Cong He, Dan Zhao, Fei Fan, Hongqiang Zhou, Xin Li adn Yao Li, Junjie Li, Fei Dong, Yin-Xiao Miao, Yongtian Wang, and Lingling Huang. Pluggable multitask diffractive neural networks based on cascaded metasurfaces. *Opto-Electronic Advances*, 7(23005), 2024.

[34] Kaixuan Wei, Xiao Li, Johannes Froech, Praneeth Chakravarthula, James Whitehead, Ethan Tseng, Arka Majumdar, and Felix Heide. Spatially varying nanophotonic neural networks. *Science Advances*, 10(45):eadp0391, 2024.

[35] Hanyu Zheng, Quan Liu, Ivan I Kravchenko, Xiaomeng Zhang, Yuankai Huo, and Jason G Valentine. Multichannel meta-imagers for accelerating machine vision. *Nature nanotechnology*, 19(4):471–478, 2024.

[36] Cong He, Dan Zhao, Fei Fan, Hongqiang Zhou, Xin Li, Yao Li, Junjie Li, Fei Dong, Yin-Xiao Miao, Yongtian Wang, et al. Pluggable multitask diffractive neural networks based on cascaded metasurfaces. *Opto-Electron. Adv*, 7(2):230005, 2024.

[37] Anna Wirth-Singh, Jinlin Xiang, Minho Choi, Johannes E Fröch, Luocheng Huang, Shane Colburn, Eli Shlizerman, and Arka Majumdar. Compressed meta-optical encoder for image classification. *Advanced Photonics Nexus*, 4(2):026009–026009, 2025.

[38] Anna Wirth-Singh, Jinlin Xiang, Minho Choi, Johannes Fröch, Luocheng Huang, Eli Shlizerman, and Arka Majumdar. Compressed meta-optical encoder for image classification. In *CLEO: Fundamental Science*, pages FF1J–1. Optica Publishing Group, 2024.

[39] Yuchi Huo, Hujun Bao, Yifan Peng, Chen Gao, Wei Hua, Qing Yang, Haifeng Li, Rui Wang, and Sung-Eui Yoon. Optical neural network via loose neuron array and functional learning. *Nature Communications*, 14(1):2535, 2023.

[40] Md Sadman Sakib Rahman and Aydogan Ozcan. Time-lapse image classification using a diffractive neural network. *Advanced Intelligent Systems*, 5(5):2200387, 2023.

[41] Chenyu You, Jinlin Xiang, Kun Su, Xiaoran Zhang, Siyuan Dong, John Onofrey, Lawrence Staib, and James S Duncan. Incremental learning meets transfer learning: Application to multi-site prostate mri segmentation. In *International Workshop on Distributed, Collaborative, and Federated Learning*, pages 3–16. Springer, 2022.

[42] Jinlin Xiang, Bozhao Qi, Marc Cerou, Wei Zhao, and Qi Tang. Dn-ode: Data-driven neural-ode modeling for breast cancer tumor dynamics and progression-free survivals. *Computers in Biology and Medicine*, 180:108876, 2024.

[43] Jaehoon Lee, Lechao Xiao, Samuel S Schoenholz, Yasaman Bahri, Roman Novak, Jascha Sohl-Dickstein, and Jeffrey Pennington. Finite versus infinite neural networks: An empirical study. In *Advances in Neural Information Processing Systems*, volume 33, pages 15156–15172, 2020.

[44] Guangda Ji and Zhanxing Zhu. Knowledge distillation in wide neural networks: Risk bound, data efficiency and imperfect teacher. *Advances in Neural Information Processing Systems*, 33:20823–20833, 2020.

[45] Yoann Morello, Emilie Grégoire, and Sam Verboven. Exploring task affinities through ntk alignment and early training dynamics in multi-task learning. In *NeurIPS 2024 Workshop on Mathematics of Modern Machine Learning*, 2024.

[46] Jinlin Xiang and Eli Shlizerman. Tkil: tangent kernel approach for class balanced incremental learning. *arXiv preprint arXiv:2206.08492*, 2022.

[47] Jinlin Xiang and Eli Shlizerman. Tkil: Tangent kernel optimization for class balanced incremental learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshop*, pages 3529–3539, 2023.

[48] Ziyang Zheng, Zhengyang Duan, Hang Chen, Rui Yang, Sheng Gao, Haiou Zhang, Hongkai Xiong, and Xing Lin. Dual adaptive training of photonic neural networks. *Nature Machine Intelligence*, 5(10):1119–1129, 2023.

[49] James Spall, Xianxin Guo, and Alexander I Lvovsky. Hybrid training of optical neural networks. *Optica*, 9(7):803–811, 2022.

[50] Hans-Christian Ruiz Euler, Marcus N Boon, Jochem T Wildeboer, Bram van de Ven, Tao Chen, Hajo Broersma, Peter A Bobbert, and Wilfred G van der Wiel. A deep-learning approach to realizing functionality in nanoelectronic devices. *Nature nanotechnology*, 15(12):992–998, 2020.

[51] Yubo Zhang, Rui Chen, Minho Choi, Johannes E Fröch, and Arka Majumdar. General image compression using random-psf metasurfaces and computational back-end. In *2024 Conference on Lasers and Electro-Optics (CLEO)*, pages 1–3. IEEE, 2024.

[52] Mengran Zhao, Shitao Zhu, Die Li, Thomas Fromenteze, Mohsen Khalily, Xiaoming Chen, Vincent Fusco, and Okan Yurduseven. Frequency-diverse bunching metasurface antenna for microwave computational imaging. *IEEE Transactions on Antennas and Propagation*, 2024.

[53] Anders Pors, Fei Ding, Yiting Chen, Ilya P Radko, and Sergey I Bozhevolnyi. Random-phase metasurfaces at optical wavelengths. *Scientific reports*, 6(1):28448, 2016.

[54] Matthieu Dupré, Liyi Hsu, and Boubacar Kanté. On the design of random metasurface based devices. *Scientific reports*, 8(1):7162, 2018.

[55] Parker R Wray and Harry A Atwater. Light–matter interactions in films of randomly distributed unidirectionally scattering dielectric nanoparticles. *ACS Photonics*, 7(8):2105–2114, 2020.

[56] Laj Cutrona, E Leith, C Palermo, and L Porcello. Optical data processing and filtering systems. *IRE Transactions on Information Theory*, 6(3):386–400, 2003.

[57] Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in neural information processing systems*, 31, 2018.

[58] Sanjeev Arora, Simon S Du, Wei Hu, Zhiyuan Li, Russ R Salakhutdinov, and Ruosong Wang. On exact computation with an infinitely wide neural net. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[59] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[60] Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009. Technical report.

[61] Yite Wang, Dawei Li, and Ruoyu Sun. Ntk-sap: Improving neural network pruning by aligning training dynamics. *arXiv preprint arXiv:2304.02840*, 2023.

[62] Ryan Vogt, Yang Zheng, and Eli Shlizerman. Lyapunov-guided representation of recurrent neural network performance. *Neural Computing and Applications*, 36(34):21211–21226, 2024.

[63] Caleb Zheng and Eli Shlizerman. Hyperpruning: Efficient search through pruned variants of recurrent neural networks leveraging lyapunov spectrum. *arXiv preprint arXiv:2506.07975*, 2025.

[64] Mingxing Tan and Quoc V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. *ArXiv*, abs/1905.11946, 2019.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly reflect the paper's scope and contributions, stating the proposed NTKD pipeline and supporting claims with both theory and experiments.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: This paper acknowledges limitations such as the shallow and linear nature of current ONNs, the difficulty of implementing nonlinearities and scaling up, and fabrication-induced issues in the Discussion section.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: The Supplementary provides the assumptions and a complete proof.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The Supplementary includes optical implementation and setup details.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The experiments use open-source benchmark datasets, and additional implementation details are provided in the Supplementary Material.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The implementation details are provided in the Supplementary Material.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Experiments are repeated with multiple random seeds, and results are reported with mean and standard deviation.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide estimates of MACs and system-level power consumption.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: This work complies with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: Our work does not have a significant societal impact, and it is a foundational research.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification:] We correctly credited all the assets used in the paper.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: This paper does not introduce any new datasets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: This research did not involve any crowdsourcing or human subject studies.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: This work does not involve human subjects or user studies and therefore does not require IRB approval.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.

- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: The core method development in this research does not involve LLMs.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.