NEURAL INFORMATION PROCESSING SYSTEMS

TEXAS The University of Texas at Austin | Electrical and Computer Engineering Cockrell School of Engineering

L^{2}ight: Enabling On-Chip Learning for Optical Neural Networks via Efficient in-situ Subspace Optimization

Jiaqi Gu, Hanqing Zhu, Chenghao Feng, Zixuan Jiang,
Ray T. Chen, David Z. Pan
ECE Department, University of Texas at Austin

December 17, 2021

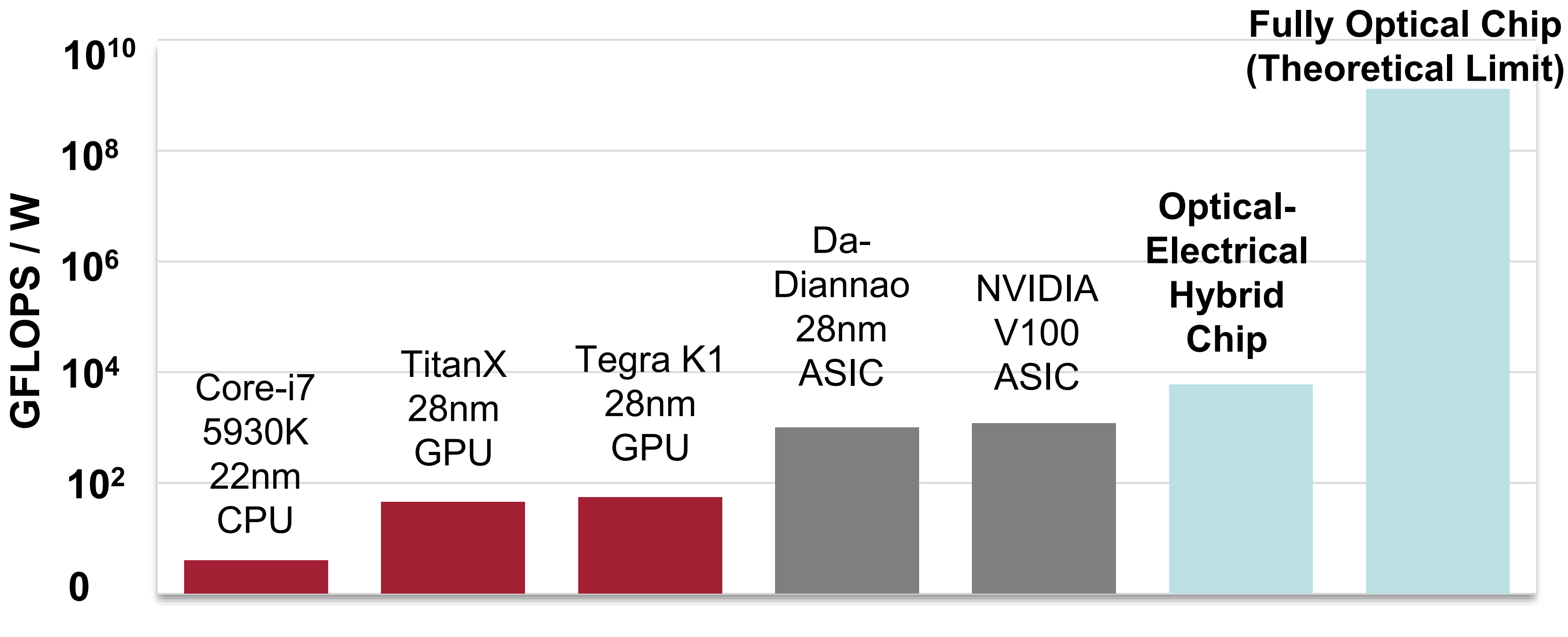
This work is supported in part by AFOSR MURI

jqgu@utexas.edu; https://jeremiemelo.github.io

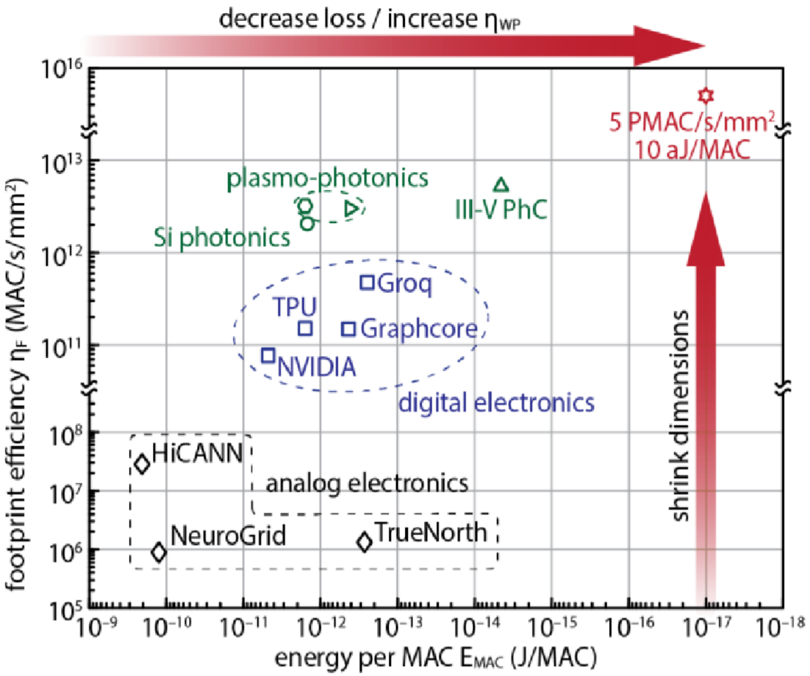
Optical Neurocomputing

- Moore's law is winding down
- Optics as next-generation AI solution

Ultra-high speed & Ultra-low energy cost



[Shen+, *Nature Photonics* 2017]



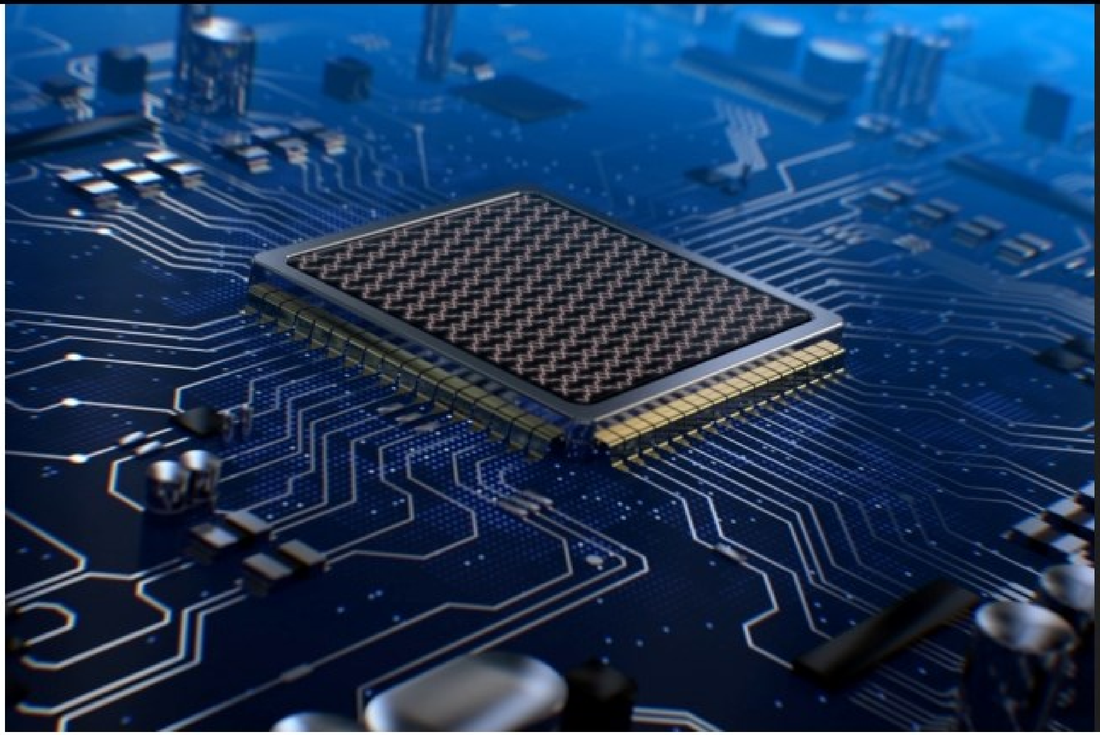
[Totovic+, *JSTQE* 2020]

Photonic AI Chips

Based on optics/photonics → photonic ICs

MIT News
ON CAMPUS AND AROUND THE WORLD

Browse or Search



FULL SCREEN

This futuristic drawing shows programmable nanophotonic processors integrated on a printed circuit board and carrying out deep learning computing.

Image: RedCube Inc., and courtesy of the researchers

New system allows optical “deep learning”

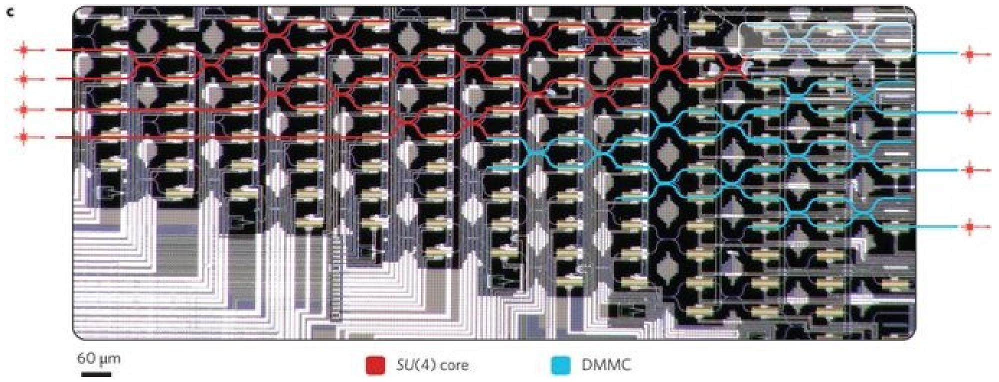
Neural networks could be implemented more quickly using new photonic technology.

LIGHTELLIGENCE

LIGHTMATTER

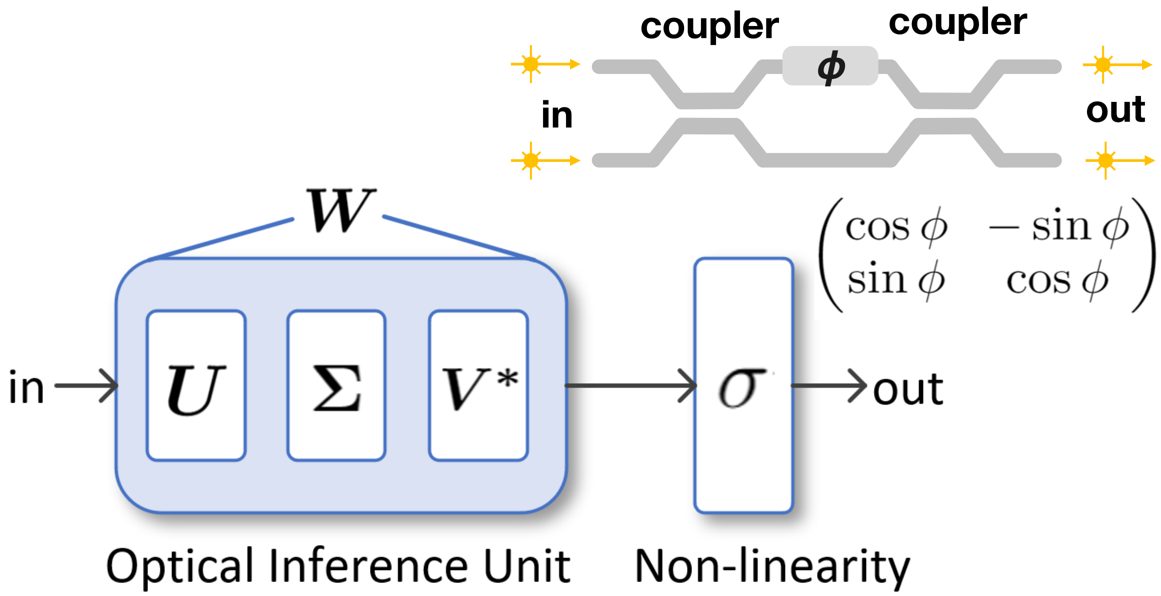
Optical Neural Networks (ONN)

- Emergence of photonic NNs
  - Ultra-fast speed (light in and light out)
  - >100 GHz photo-detection rate
  - Near-zero energy consumption if fixed



[Shen+, *Nature Photonics* 2017]

- Map weight matrix to MZI meshes
- Singular value decomposition (SVD)
  - $W = U\Sigma V^*$
- Unitary group parametrization (UP)
  - $U(n) = D \prod_{i=n}^{2} \prod_{j=1}^{i-1} R_{ij}(\phi_{ij})$

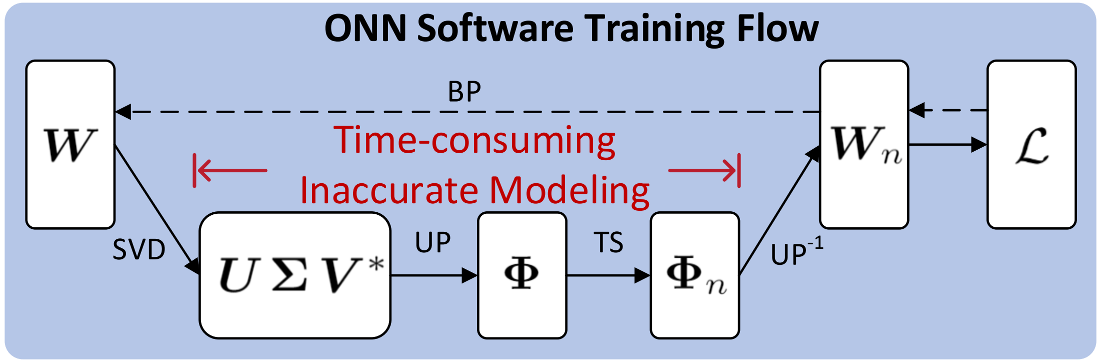


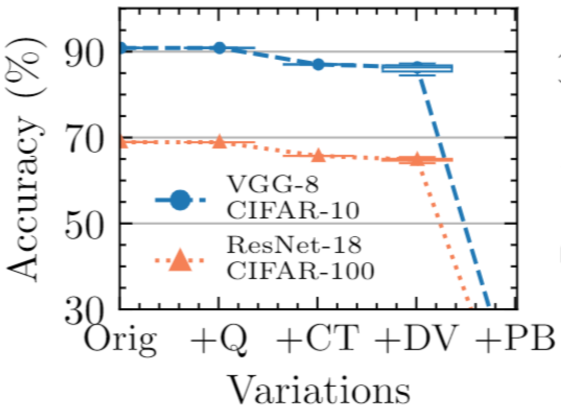
ONN On-Chip Training

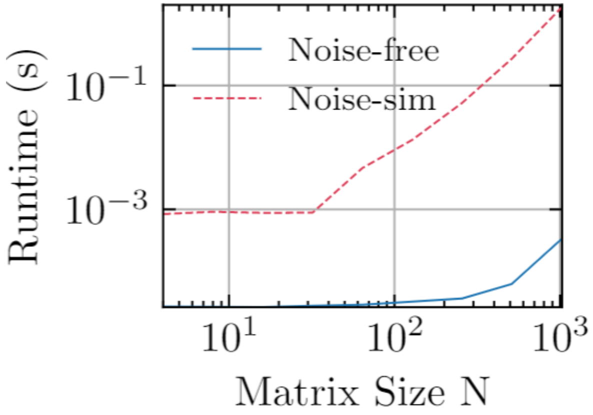
- What is ONN on-chip (on-device) training
  - *In-situ* **calibration and learning** on **non-ideal** photonic circuits

- Why on-chip training
  - Inaccurate sc
    - Severe perf
  - Inefficient and
    - Expensive c

Robust Deployment &
On-Chip Learnability

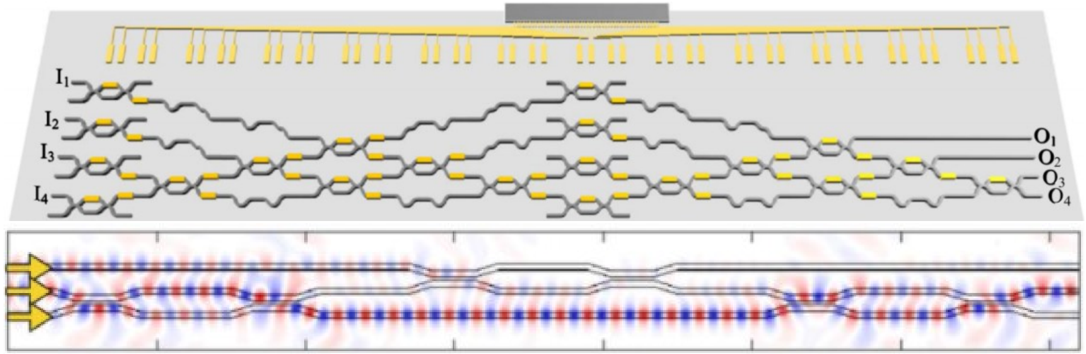




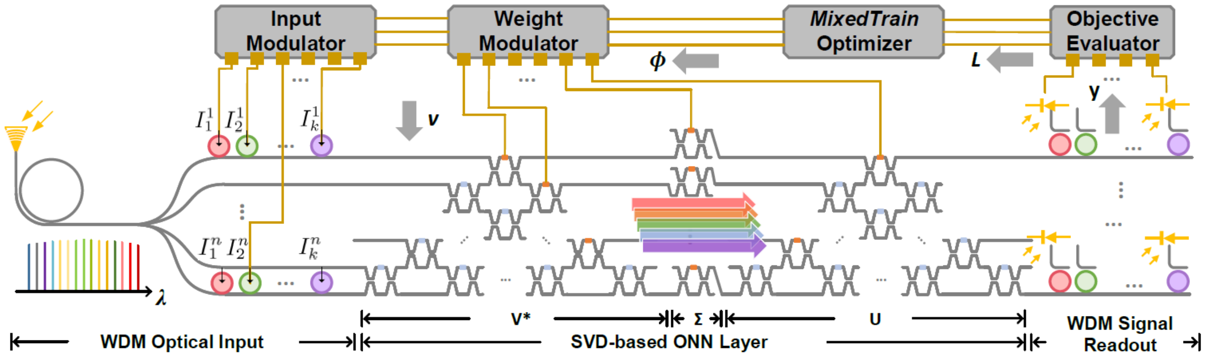


Prior On-Chip Training Protocols

- Unscalable: 100~1,000 MZIs
- Training instability/divergence
- Limited training efficiency



[Zhou+, *JSTQE*'19] [Hughes+, *Optica*'18]

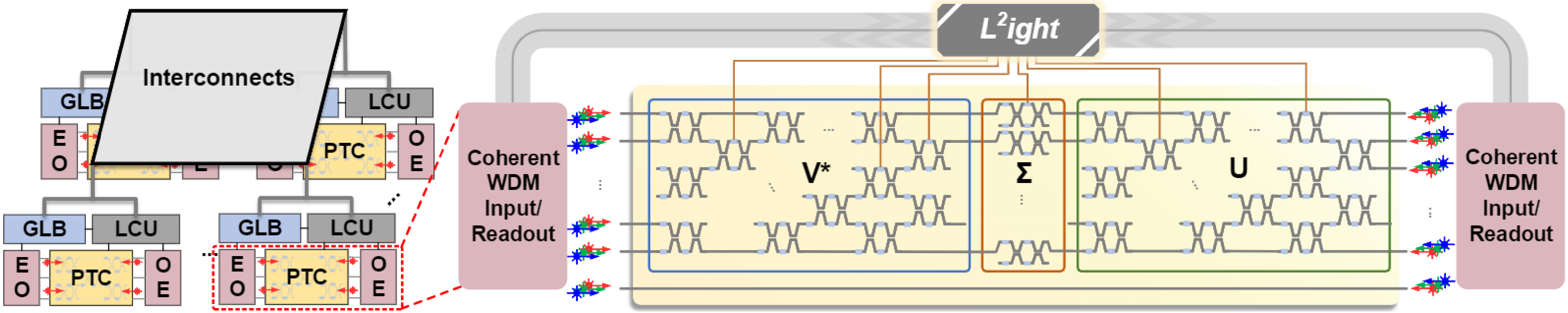


[Gu+, *DAC*'20] [Gu+, *AAAI*'21]

	BFT [NaturePhotonics'17]	PSO [OE'19]	AVM [Optica'18]	FLOPS [DAC'20]	MixedTrain [AAAI'21]	***Our L²ight***
#Params	~100	~100	~100	~1,000	~2,500	**~10 M**
Algorithm	ZO Search	Evolution (ZO)	Adjoint Method (FO)	ZO SGD	SZO-SCD	**ZO + FO**
Resolution Req.	Medium	High	Medium	High	Medium	**Medium**
Observability Req.	Coh. I/O	Coh. I/O	Coh. I/O+ Per device monitor	Coh. I/O	Coh. I/O	**Coh. I/O**

Our Contributions

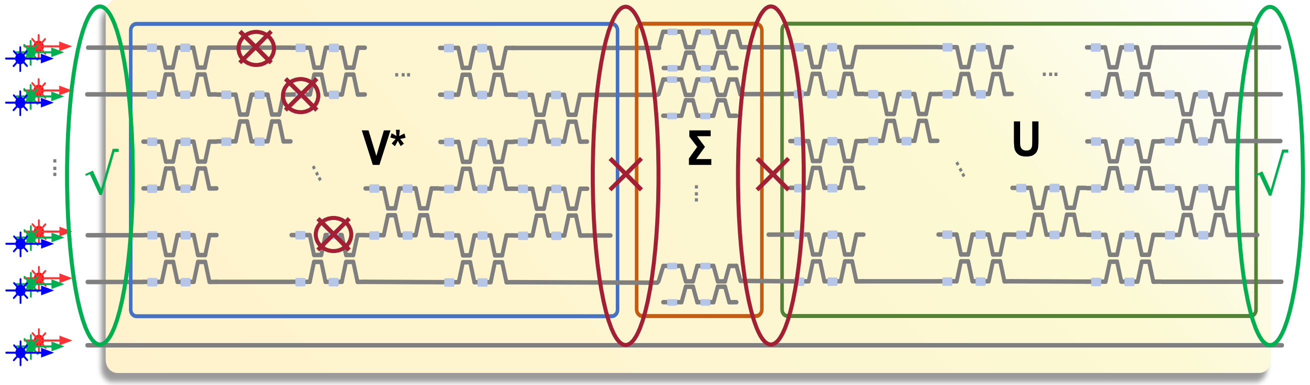
- Synergistic ONN On-Chip Learning Framework
  - Scalability: First framework that can handle *million-parameter* ONNs
  - Efficiency: Multi-level sparsity to boost efficiency by ***30***×
  - Learnability: *Subspace* optimization to enable on-device self-learnability
  - Robustness: In-situ noise consideration for *noise-resilient* ONNs



Problem Formulation and Challenges

- Optimize ***noisy*** MZI ***phases*** to minimize learning objective
  - Variables: $\Phi^U, \Phi^V, \Phi^\Sigma$
  - Non-ideality: cross-talk (Ω), Noise (Γ), Quantization ($\mathcal{Q}$), Phase bias (Φ_b)

- Challenges
  - Unobservable in-situ light fields
  - Limited input/output observability
  - Inaccessible gradients for Φ^U and Φ^V



$$\mathbf{\Phi}^* = \underset{\mathbf{\Phi}}{\operatorname{argmin}} \ \mathcal{L}\big(\boldsymbol{W}(\mathbf{\Omega\Gamma}\mathcal{Q}(\mathbf{\Phi}) + \mathbf{\Phi}_b); \mathcal{D}_{trn}\big),$$

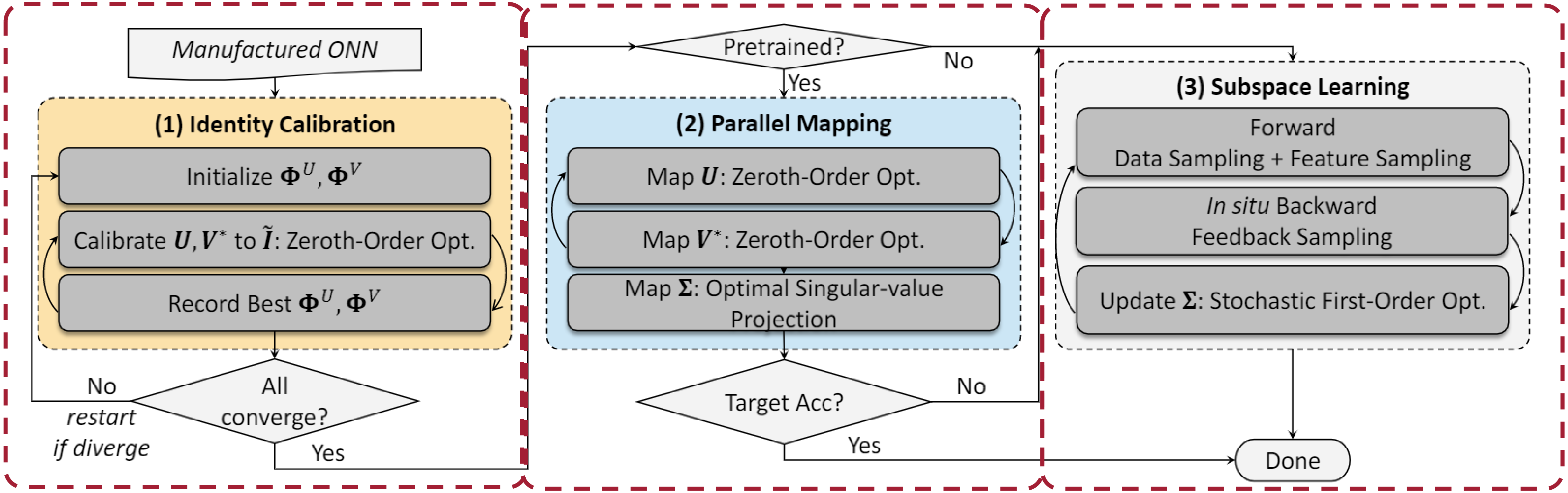
$$\text{s.t.}\ \boldsymbol{W}(\mathbf{\Phi}) = \big\{\boldsymbol{W}_{pq}(\mathbf{\Phi}_{pq})\big\}_{p=0,q=0}^{p=P-1,q=Q-1}, \quad \boldsymbol{W}_{pq}(\mathbf{\Phi}_{pq}) = \boldsymbol{U}_{pq}(\mathbf{\Phi}_{pq}^U)\mathbf{\Sigma}_{pq}(\mathbf{\Phi}_{pq}^S)\boldsymbol{V}_{pq}^*(\mathbf{\Phi}_{pq}^V)$$

$$\boldsymbol{U}_{pq}(\mathbf{\Phi}_{pq}^U) = \boldsymbol{D}_{pq}^U \prod_{i=k}^{2}\prod_{j=1}^{i-1} \boldsymbol{R}_{pqij}(\phi_{pqij}^U), \quad \boldsymbol{V}_{pq}^*(\mathbf{\Phi}_{pq}^V) = \boldsymbol{D}_{pq}^V \prod_{i=k}^{2}\prod_{j=1}^{i-1} \boldsymbol{R}_{pqij}(\phi_{pqij}^V),$$

$$\mathbf{\Sigma}_{pq}(\mathbf{\Phi}_{pq}^S) = \max(|\mathbf{\Sigma}_{pq}|)\texttt{diag}(\cdots, \cos\phi_{pq,i}^S, \cdots), \quad \mathbf{\Phi}_b \in \mathcal{U}(0, 2\pi),\ \mathbf{\Gamma} \in \mathcal{N}(\gamma, \sigma_\gamma^2).$$

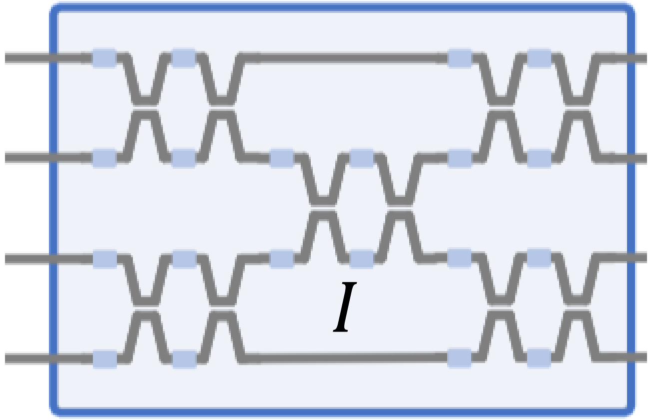
Proposed Framework: L²ight

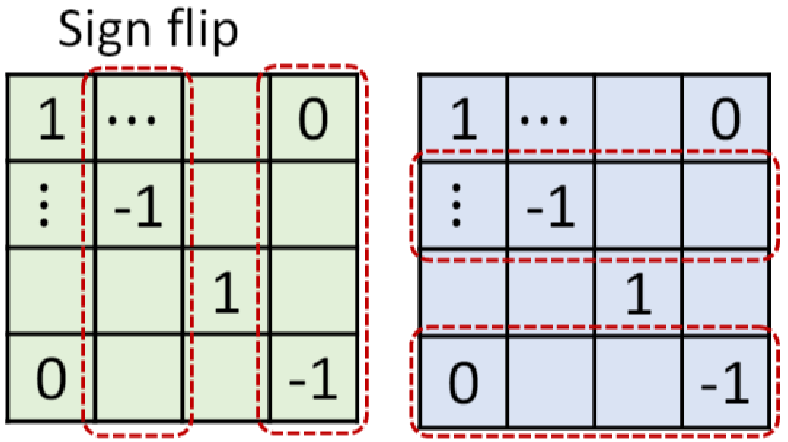
- Identity Calibration (*IC*): Variation-Agnostic Circuit State Preparation
- Parallel Mapping (*PM*): Alternate Projection-based Model Deployment
- Subspace Learning (*SL*): Hardware-Aware Multi-Level Sparse Training



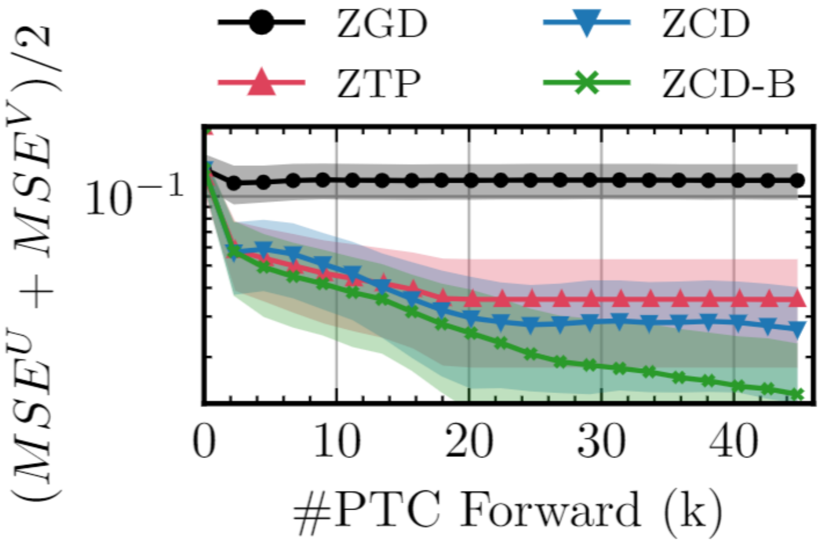
Step 1: Identity Calibration

- Prepare $\boldsymbol{U}$ and $\boldsymbol{V}^*$ to Identity projection
- $\min_{\Phi^U,\Phi^V} \sum_{p.q} \left|\left|\boldsymbol{U}_{pq}(\Phi^U_{pq}) - \boldsymbol{I}\right|\right|^2 + \left|\left|\boldsymbol{V}^*_{pq}(\Phi^V_{pq}) - \boldsymbol{I}\right|\right|^2$
- $\min_{\Phi} \sum_{p,q} ||\boldsymbol{U}_{pq}(\Phi^U_{pq})\boldsymbol{\Sigma}_{pq}\boldsymbol{V}^*_{pq}(\Phi^V_{pq})\boldsymbol{\Sigma}^{-1}_{pq} - \boldsymbol{I}||$
- Solve *batched* problem via zeroth-order optimization
- $\boldsymbol{U}$ converges to sign-flipping matrices $\tilde{\boldsymbol{I}}$



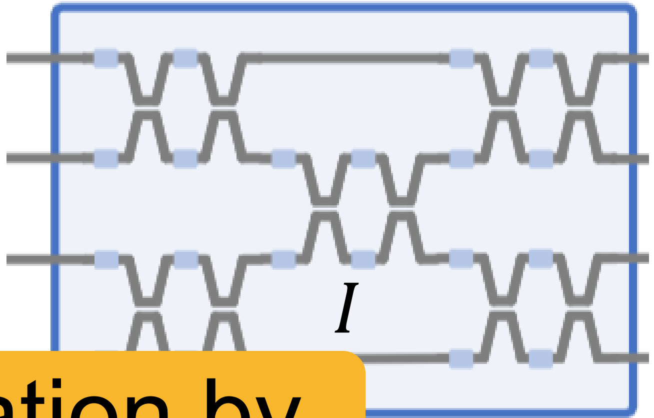


$\boldsymbol{U} = \tilde{\boldsymbol{I}}, |\boldsymbol{U}| = \boldsymbol{I}$ $\quad \boldsymbol{V}^* = \tilde{\boldsymbol{I}}, |\boldsymbol{V}^*| = \boldsymbol{I}$

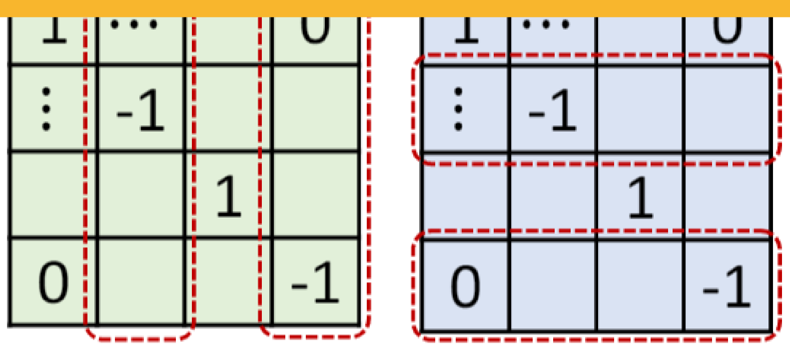


Step 1: Identity Calibration

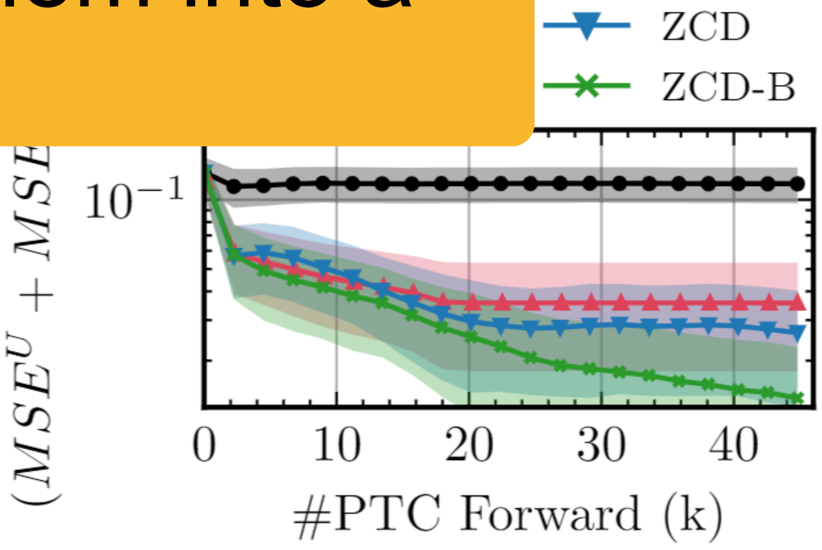
- Prepare $\boldsymbol{U}$ and $\boldsymbol{V}^*$ to Identity projection
- $\min_{\Phi^U,\Phi^V} \sum_{p.q} \left|\left|\boldsymbol{U}_{pq}(\Phi_{pq}^U) - \boldsymbol{I}\right|\right|^2 + \left|\left|\boldsymbol{V}_{pq}^*(\Phi_{pq}^V) - \boldsymbol{I}\right|\right|^2$
- $\min_{\Phi} \sum_{p,q} ||\boldsymbol{U}_{pq}(\Phi_{pq}^U)\boldsymbol{\Sigma}_{pq}\boldsymbol{V}_{pq}^*(\Phi_{pq}^V)\boldsymbol{\Sigma}_{pq}^{-1} - \boldsymbol{I}||$
- S
- $\boldsymbol{U}$



Efficient variation-agnostic calibration by partitioning large-scale problem into a batch of subtasks

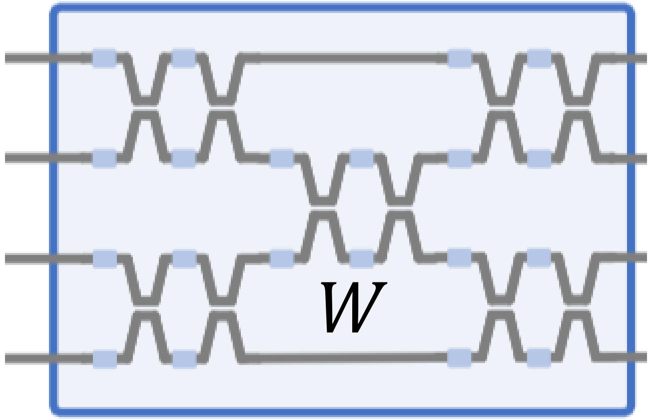


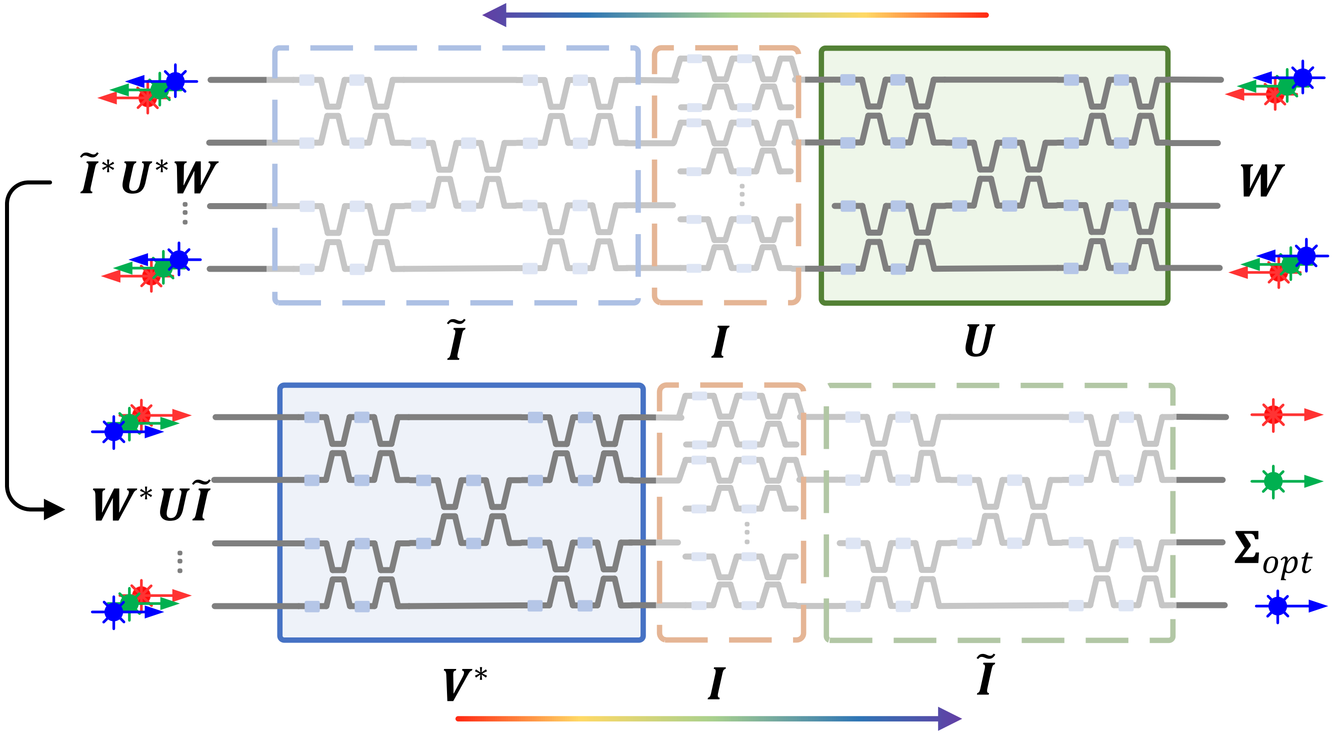
$\boldsymbol{U} = \tilde{\boldsymbol{I}}, |\boldsymbol{U}| = \boldsymbol{I}$ $\boldsymbol{V}^* = \tilde{\boldsymbol{I}}, |\boldsymbol{V}^*| = \boldsymbol{I}$



Step 2: Parallel Mapping

- Map pretrained matrix to optical mesh
- *Batched* regression: $\min_{\Phi} \sum_{p.q} \left|\left|\widetilde{\boldsymbol{W}}_{pq}(\Phi_{pq}) - \boldsymbol{W}_{pq}\right|\right|^2$
- Zeroth-order optimization on $\boldsymbol{U}$ and $\boldsymbol{V}^*$
- Analytical optimal projection (*OSP*) on $\boldsymbol{\Sigma}$
  - $\boldsymbol{\Sigma}_{\mathbf{opt}} = \mathbf{diag}\left(\left(\tilde{\boldsymbol{I}}^*\mathbf{V}^*\mathbf{W}^*\boldsymbol{U}\tilde{\boldsymbol{I}}\right)^*\right)$

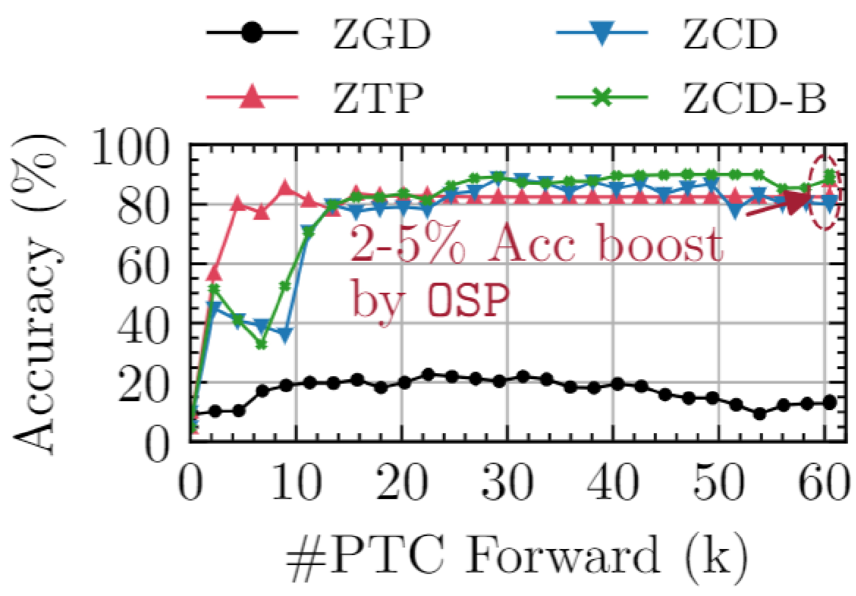


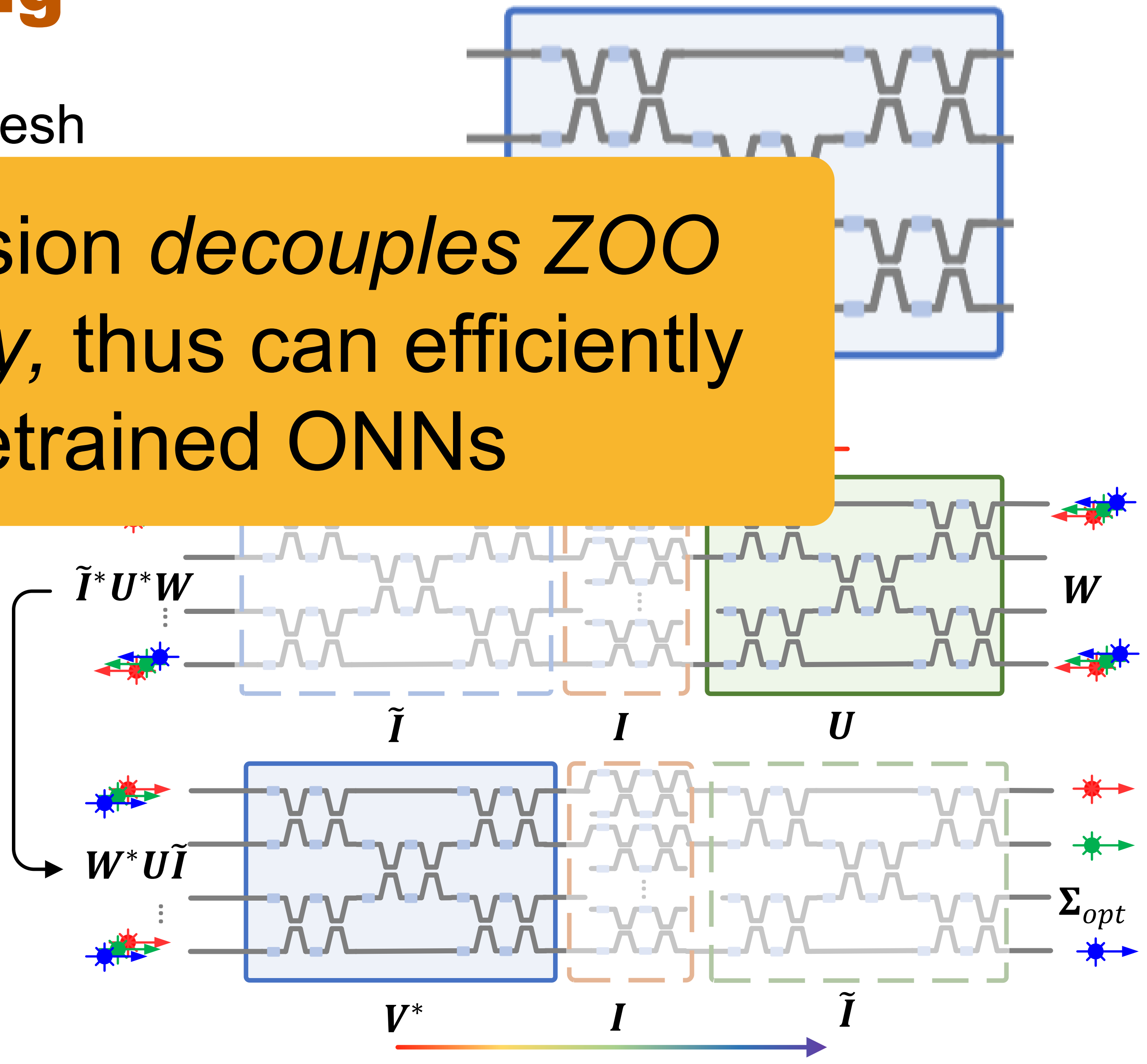


Step 2: Parallel Mapping

- Map pretrained matrix to optical mesh
- *Batch*
- Zerot
- Analy
  - Σ_{o}

Batched regression *decouples ZOO from stochasticity*, thus can efficiently deploy pretrained ONNs

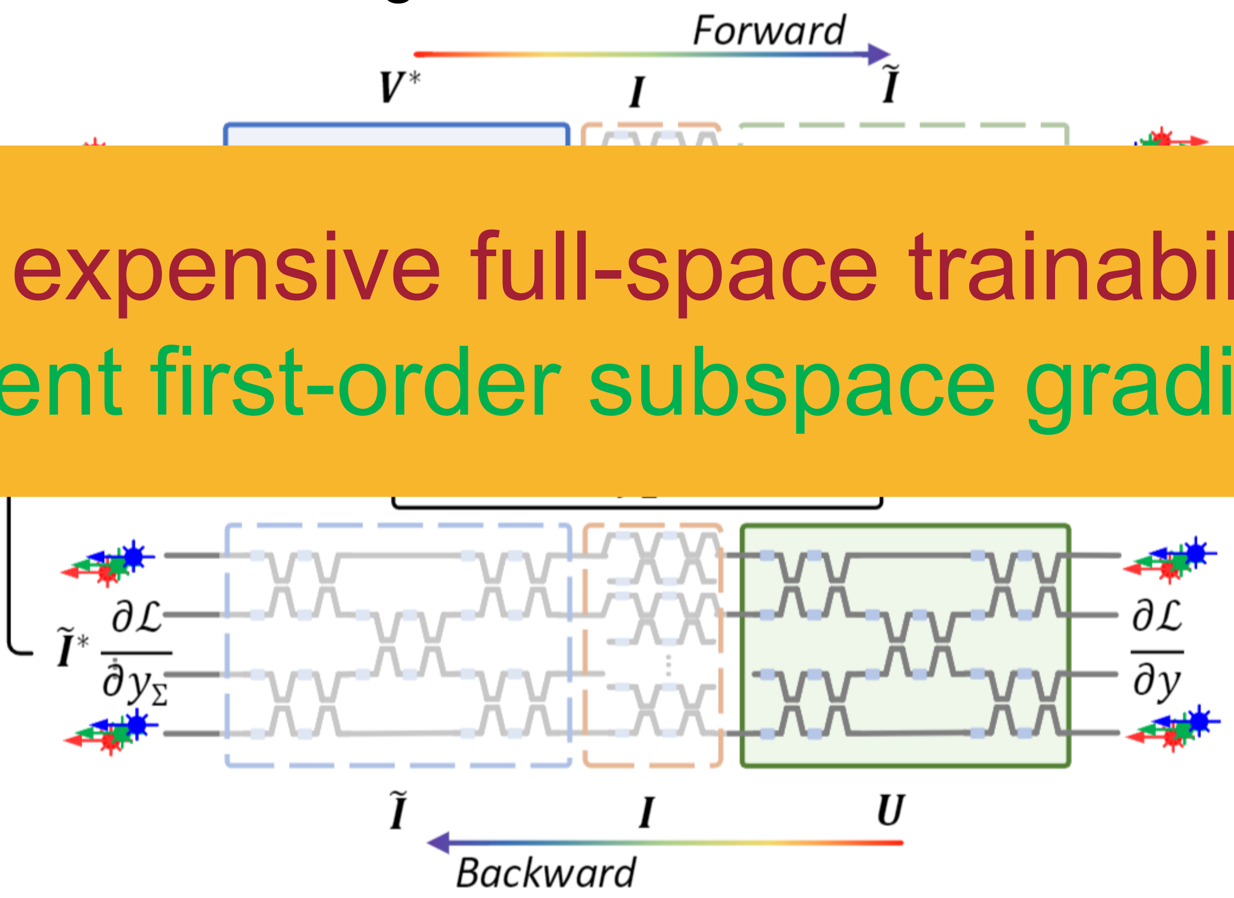




Step 3: Subspace Learning

- In-situ subspace gradient acquisition via *reciprocity*
- Shine light *forward/backward*
- Sign flips *cancel out* at diagonals

Only optimize Σ and freeze U and V^*

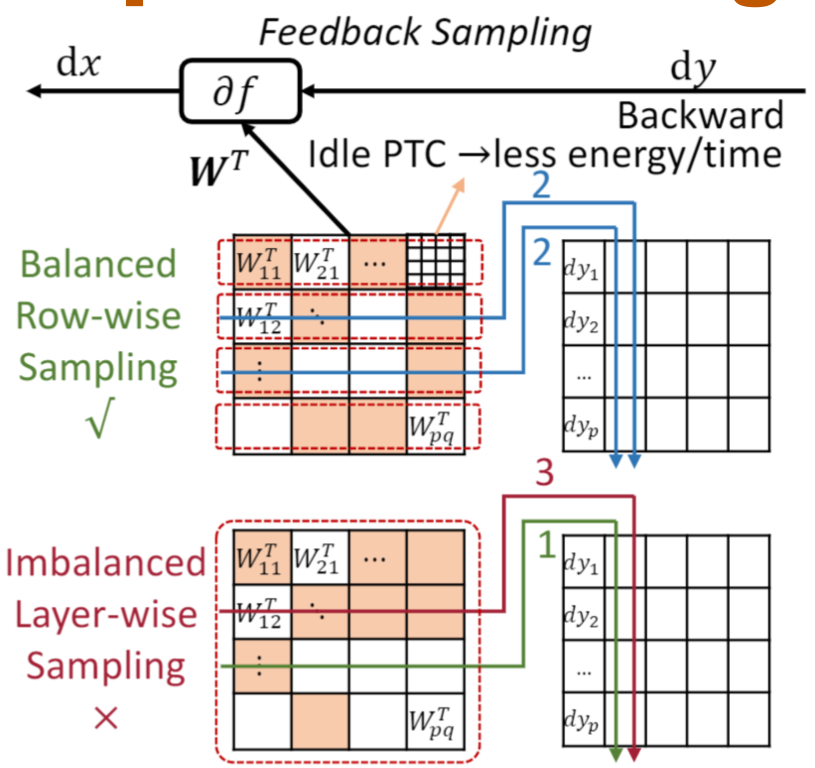


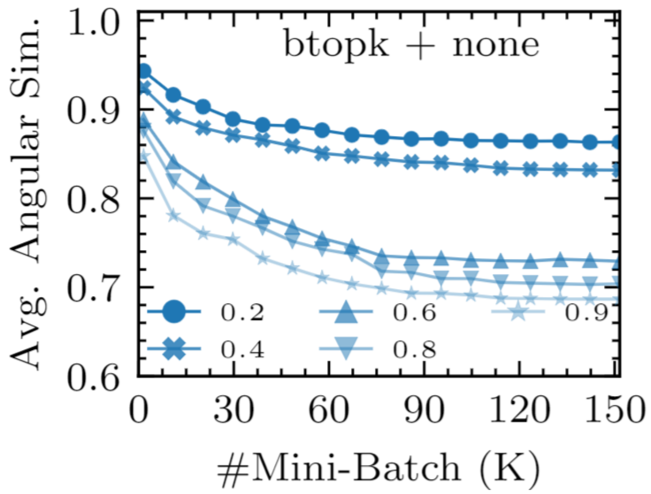
Trade expensive full-space trainability for efficient first-order subspace gradients

Efficiency: Multi-Level Sparse Subspace Learning

- Balanced feedback matrix sampling
  - Save cost on $\frac{\partial \mathcal{L}}{\partial x} = \boldsymbol{W}^T \frac{\partial \mathcal{L}}{\partial y}$
  - Sampling weight blocks for efficient error feedback (sparsity α_W)
  - Row-wise top-K sampling
    - *Lower variance* than uniform sampling
    - *Better load-balance* than naive top-K sampling
  - Gradients are well aligned with true grad.

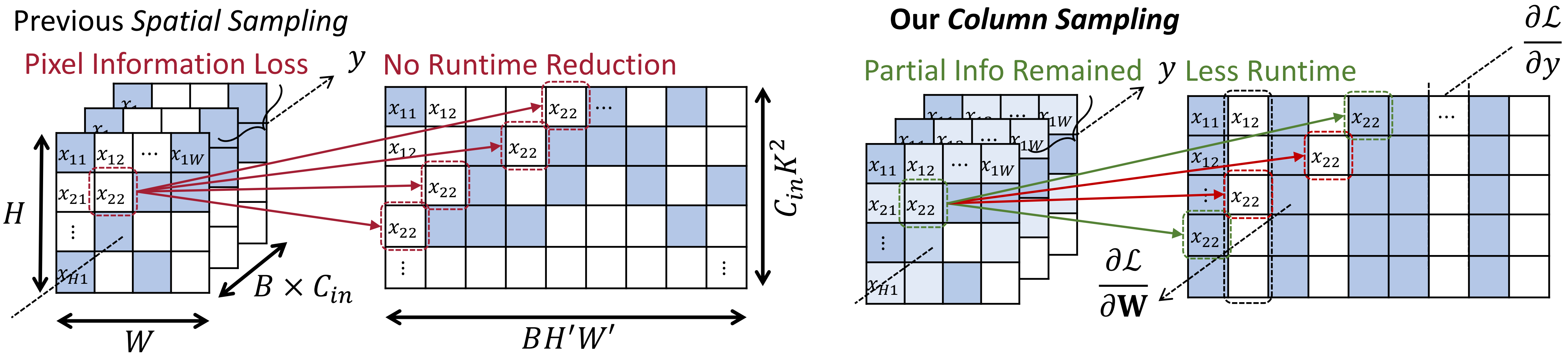
Feedback matrix $\boldsymbol{W}^T$ can be *approximated* for higher efficiency





Efficiency: Multi-Level Sparse Subspace Learning

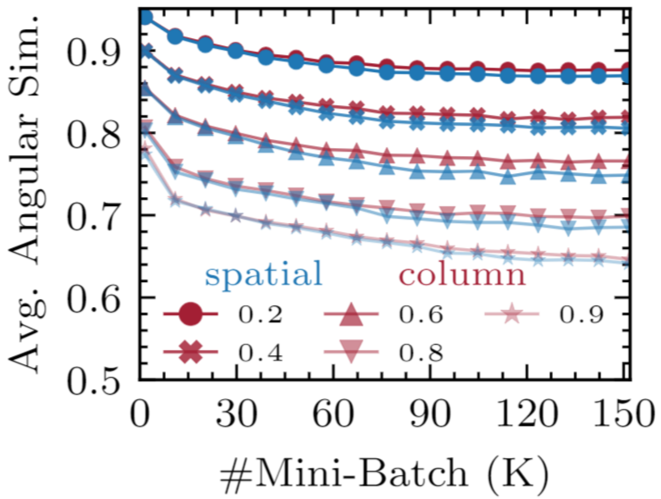
- Information-preserving column sampling
  - Save cost on $\frac{\partial \mathcal{L}}{\partial \boldsymbol{W}} = \frac{\partial \mathcal{L}}{\partial \boldsymbol{y}} * \boldsymbol{x}^T$
  - Sampling unrolled ***columns*** for efficient gradient computation (sparsity α_C)
  - Remains *partial pixel* information
  - *Structured sampling* can save runtime

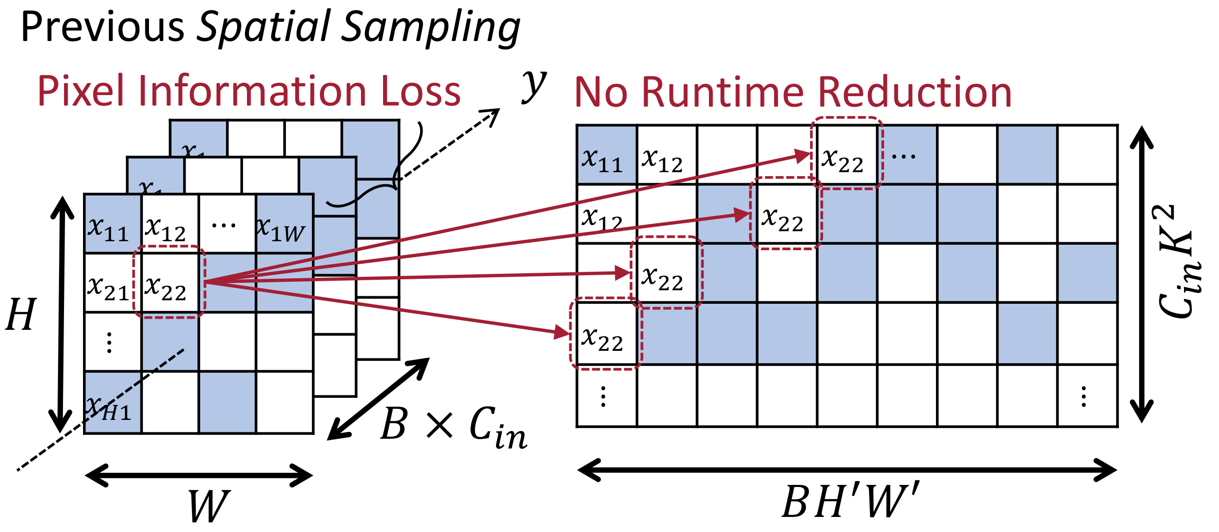


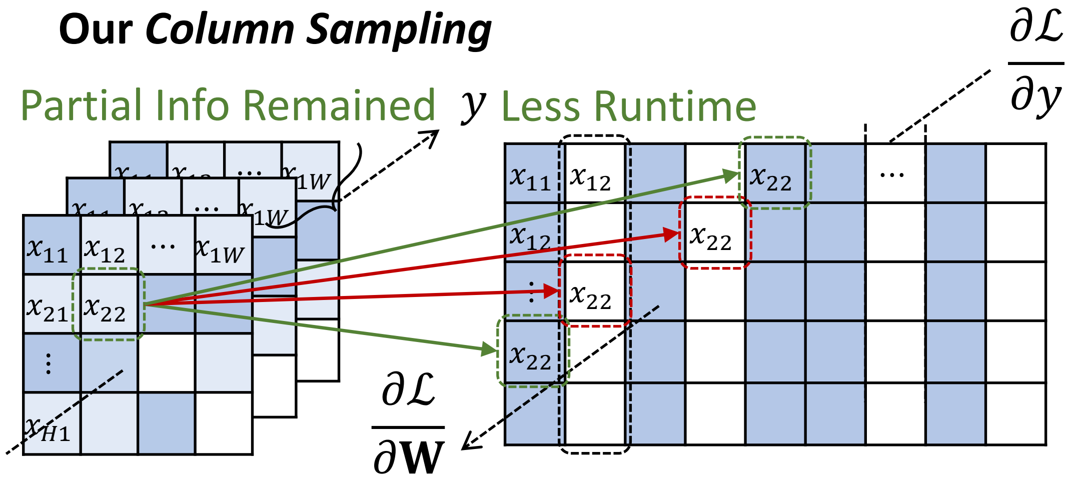
Efficiency: Multi-Level Sparse Subspace Learning

- Information-preserving column sampling
  - Save cost on $\frac{\partial \mathcal{L}}{\partial \mathbf{W}} = \frac{\partial \mathcal{L}}{\partial y} * x^T$
  - San
    (spa
  - Rer
  - *Structured sampling* can save runtime

Column sampling is more efficient & preserving more information

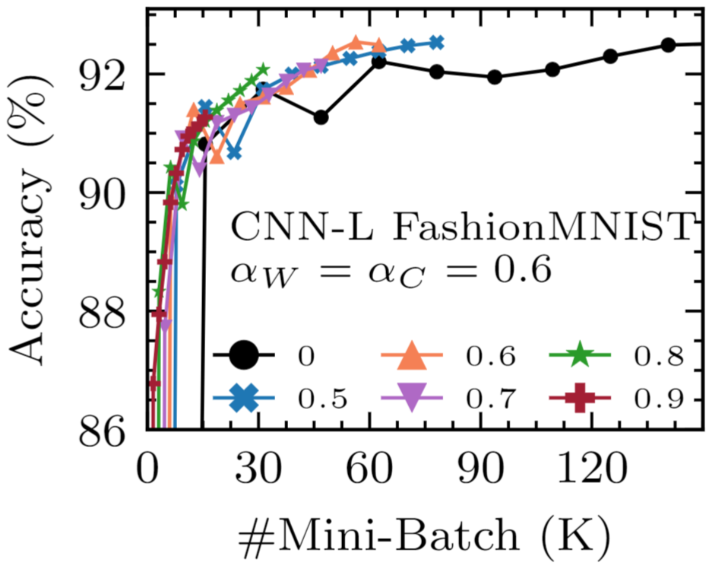






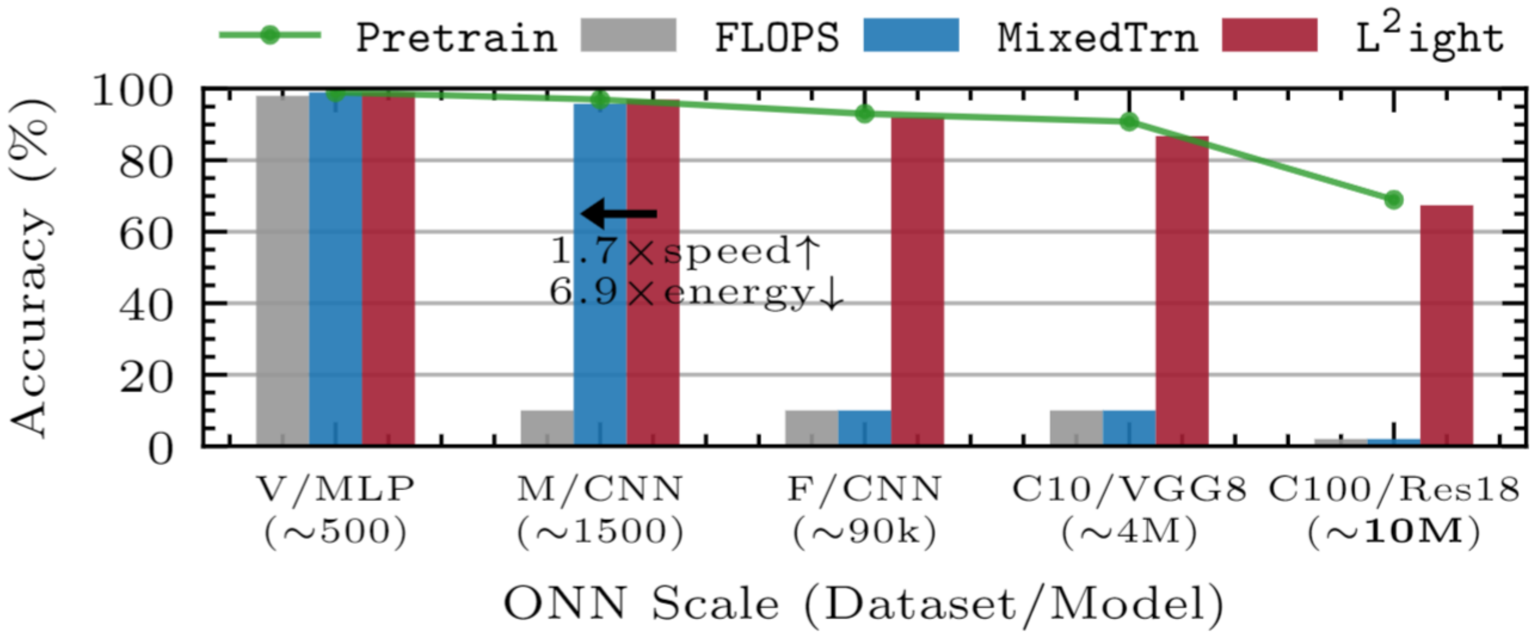
Efficiency: Multi-Level Sparse Subspace Learning

- Data sampling
  - Only train on a *subset of mini-batches* [E2-Train, NeurIPS'20]
  - Randomly skip interations with probability α_D
  - Direct speedup with marginal performance loss
  - Compatiable with feedback and column sampling



Experimental Results: Scalability

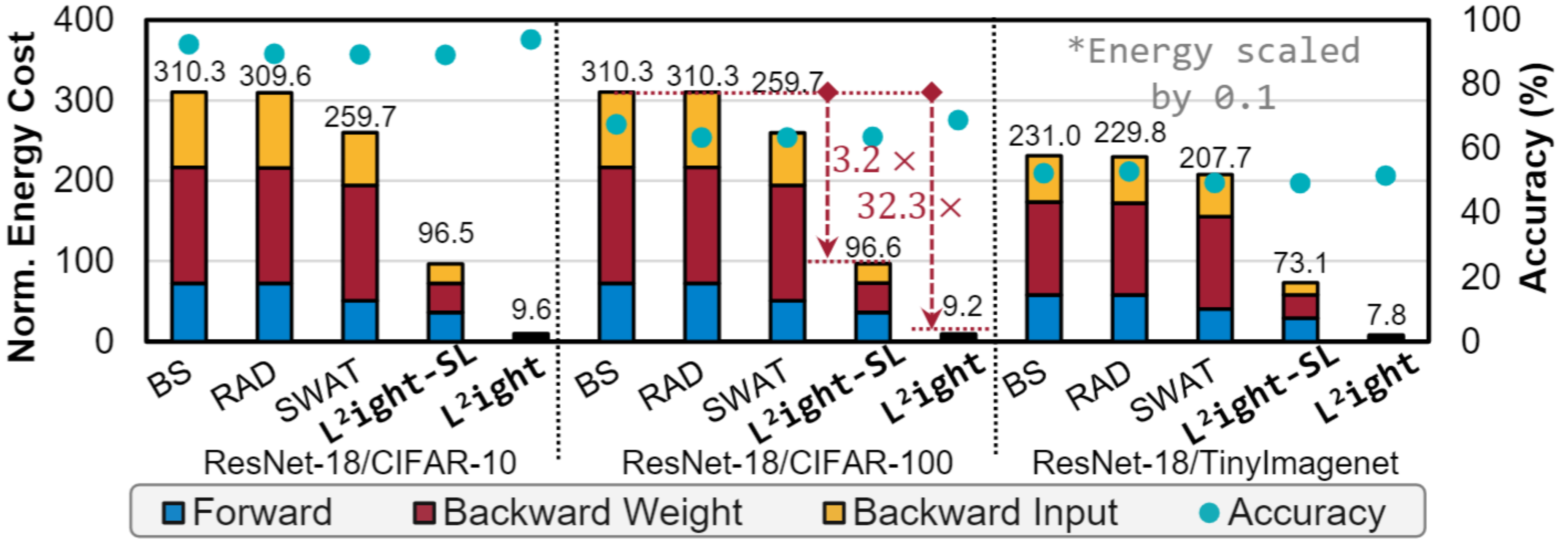
- **1,000×** more scalable than prior ONN on-chip training protocols
- High accuracy on million-parameter ONNs
- **1.7×** speedup and **6.9×** energy reduction on small ONNs than MixedTrain [Gu+, AAAI'21]



FLOPS: [Gu+, DAC'20]
MixedTrn: [Gu+, AAAI'21]

Experimental Results: Efficiency

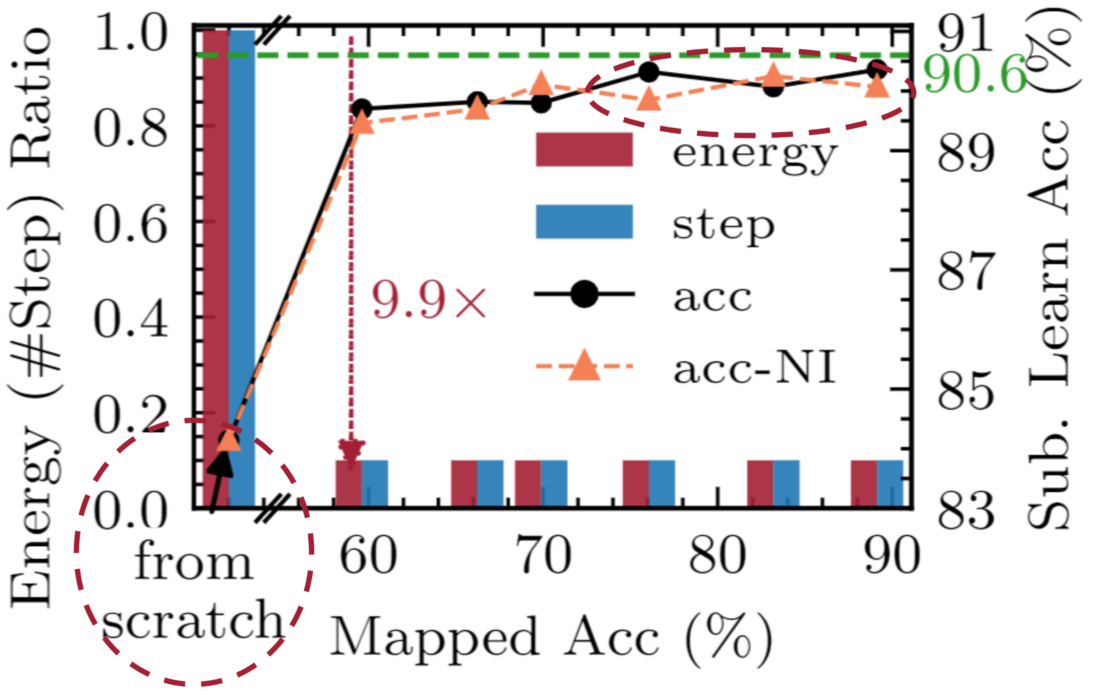
- Train from scratch: Multi-level sparse learning is **~3×** more efficiency than SoTA sparse training
- Train with mapping: Three-stage ***L²ight*** flow achieves **>30×** speed and energy efficiency improvement
- Nearly zero performance drop with heavy sparse sampling

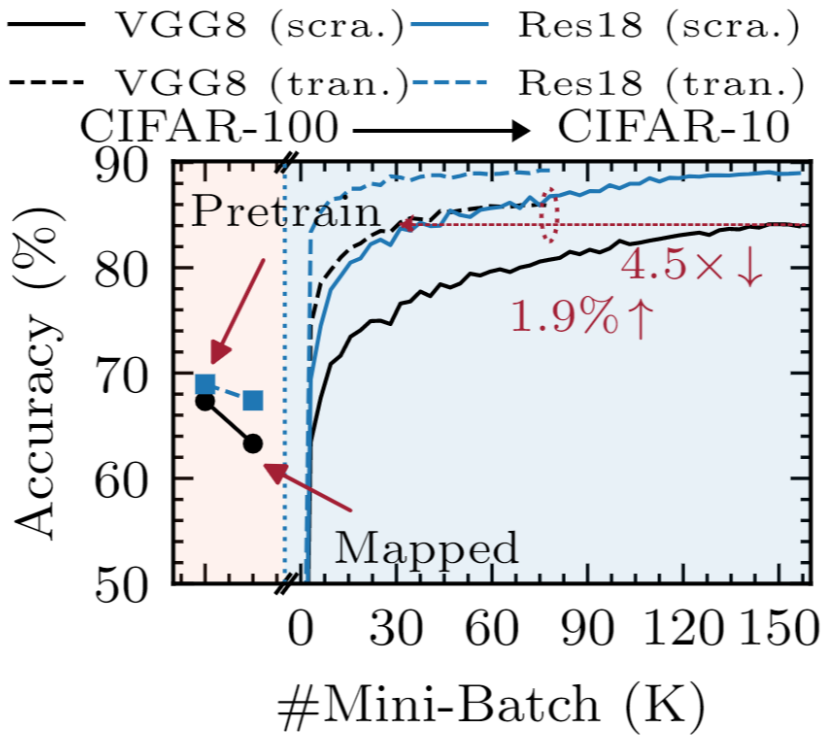


RAD: [Oktay+, ICLR'21]
SWAT: [Raihan+, NeurIPS'20]

Experimental Results: Self-learnability and Robustness

- Mapping can *improve solution quality* and save **~10**× hardware cost
- *Pure on-chip* learnability without mapping pretrained model
  - Enabled by in-situ subspace gradient acquisition
- High *noise tolerance* to non-ideal identity calibration $\tilde{I}$
- In-situ *transferability* in the restricted subspace

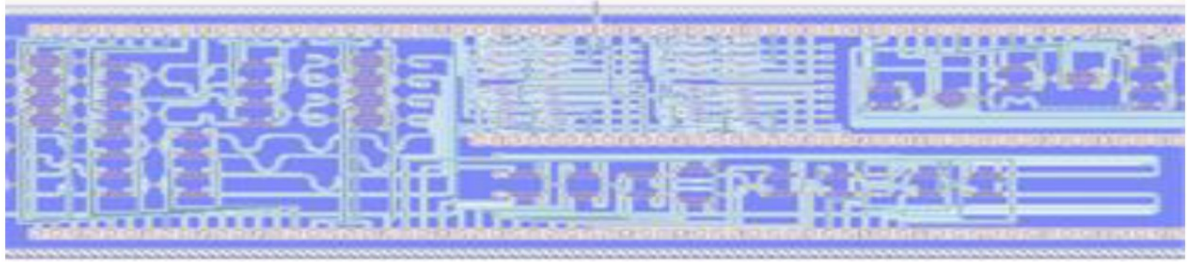




Conclusion

- ***L2ight:*** *First* scalable and efficient ONN on-chip training flow
- **Scalability**: **1,000×** more scalable than prior SoTA
- **Efficiency**: **30×** higher training efficiency via multi-level sparse subspace learning
- **Robustness**: hardware variation-agnostic flow with marginal accuracy loss

- Fure work
  - Explore new ONN architectures
  - Experimental demonstration on *real optical neural chip*



PyTorch

lumerical

SYNOPSYS®
PHOTONIC SOLUTIONS

amf ADVANCED MICRO FOUNDRY

Conclusion

- ***L2ight:*** *First* scalable and efficient ONN on-chip training flow
- **Scalability**: **1,000**× more scalable than prior SoTA
- **Efficiency**: **30**× higher training efficiency via multi-level sparse subspace learning
- **Robustness**: hardaware variation-agnostic flow with marginal accuracy loss

- Fure work
  - Explore new ONN architectures
  - Experimental demonstration on *real optical neural chip*