AI ACCELERATING SCIENCE: NEURAL OPERATORS FOR LITHOGRAPHY

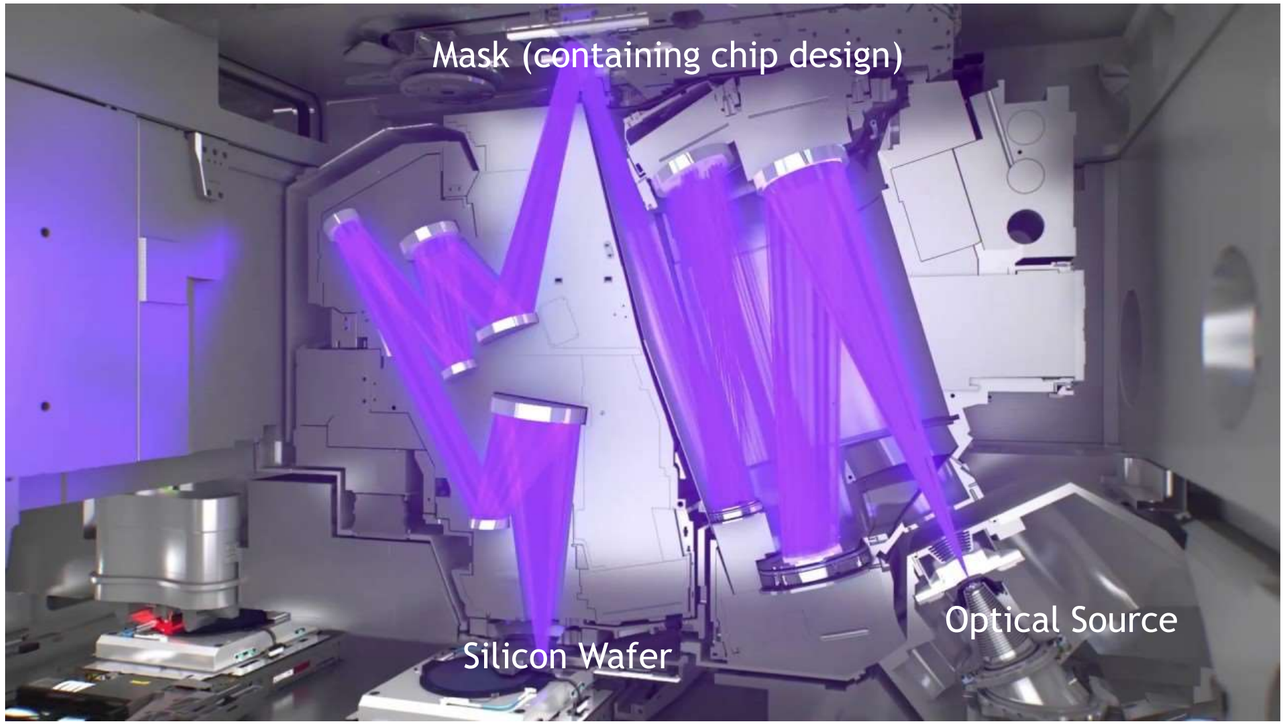
Anima Anandkumar

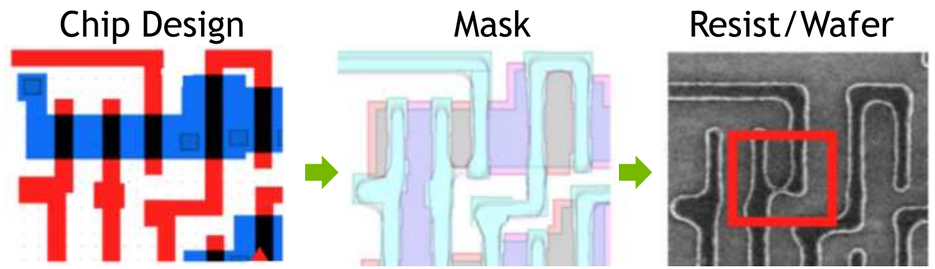
Senior Director of AI Research, NVIDIA

Bren Professor, Caltech CMS

COMPUTATIONAL LITHOGRAPHY

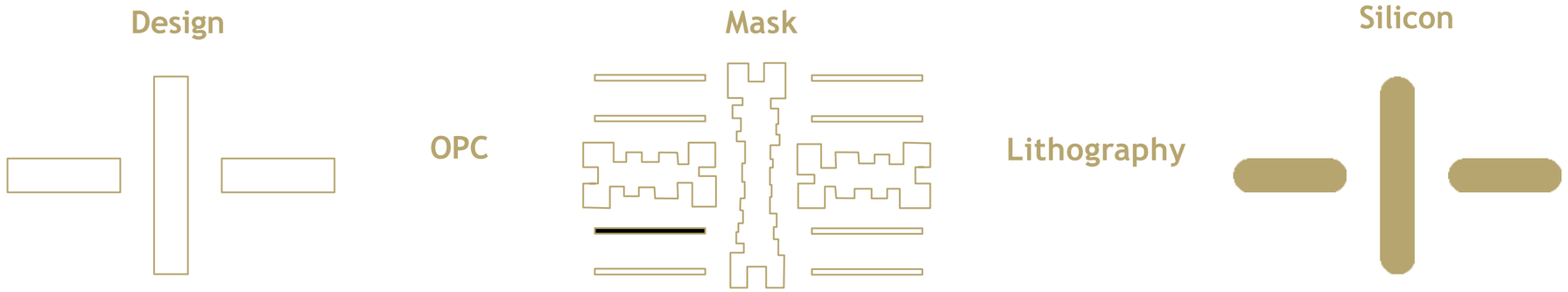
GPUs manufactured inside the lithography system

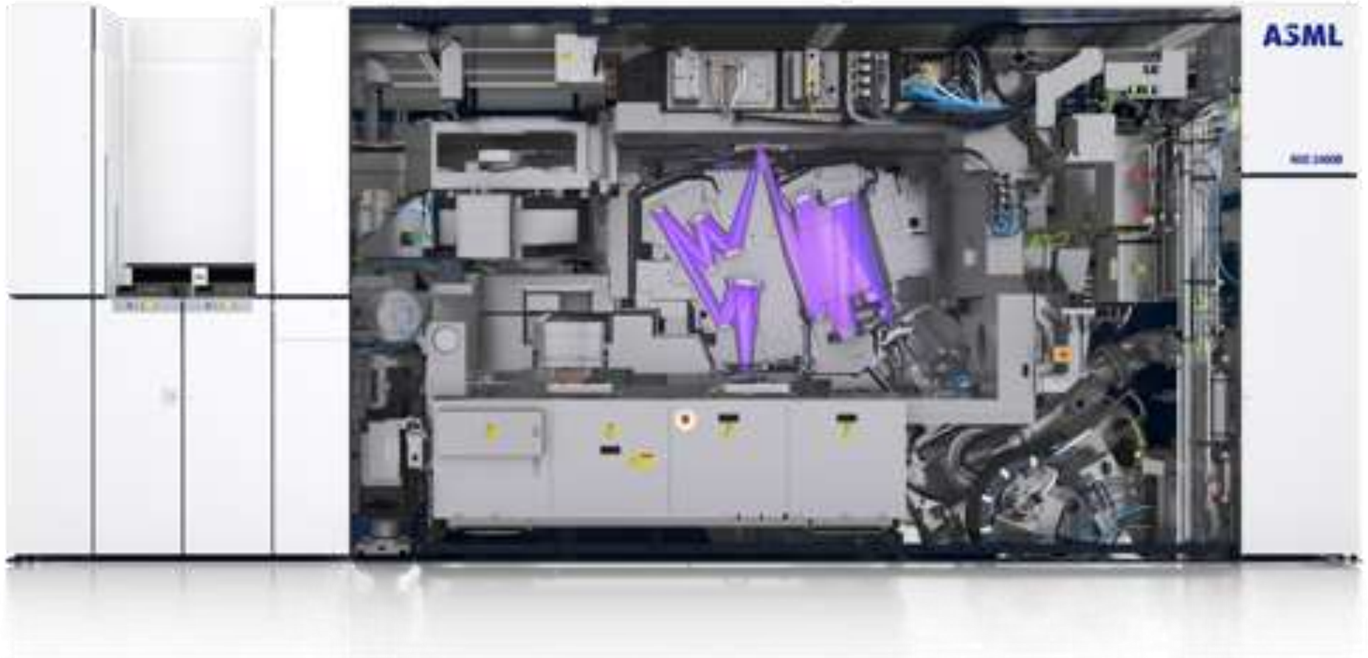




- Computational lithography is a critical research area that numerically models the behavior inside the lithography system.
- Traditional approaches take days to optimize and simulate a design on hundreds/thousands of CPU clusters, bottlenecking the turn-around-time.
- **Essential for chip yield improvement, and manufacturing cost reduction**.

OPTICAL PROXIMITY CORRECTION





OPC is the largest workload in semi design & manufacturing, tens of billions of CPU hours/yr

LITHOGRAPHY MODELING and MASK OPTIMIZATION

- Optical modeling maps a mask image to light intensity (aerial image) that is projected on a silicon wafer.
- Resist modeling deals with the interaction between light intensity and resist materials and determines the final shape formed on the silicon wafer.

Lithography Modeling

Computes the post-lithography shape on the silicon wafer given a mask design

Mask->Resist

Mask Optimization

Optimizes a mask such that the remaining pattern on the silicon wafer after the lithography process is as close as the desired shape (design)

An inverse process of lithography modeling

Design->Mask

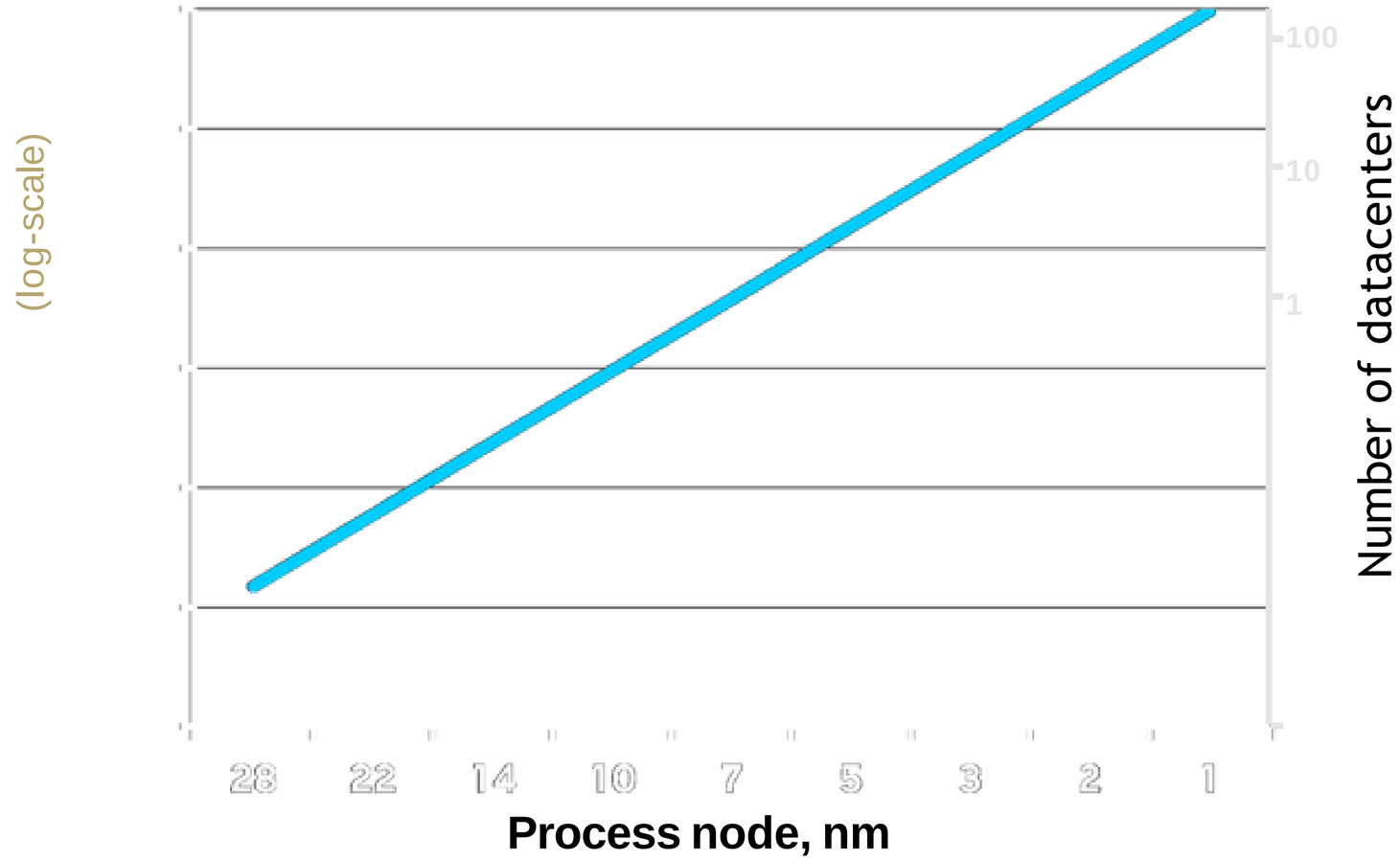
Computational Lithography Challenges

Lithography images 80 billion microscopic transistors in a single GPU onto silicon

Even a $200M camera doesn't have enough resolution for this without computational lithography

Must solve an inverse Physics problem, for trillions of polygons in a single GPU

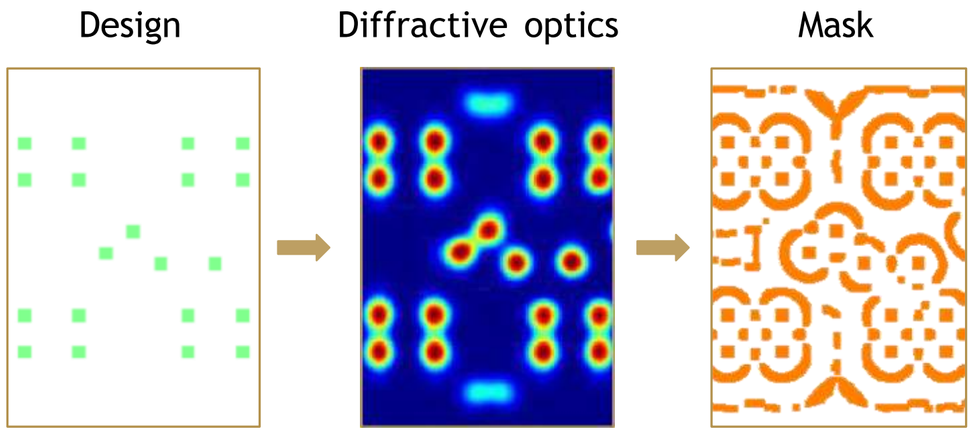
We cannot keep adding giant datacenters for this exponentially growing computation

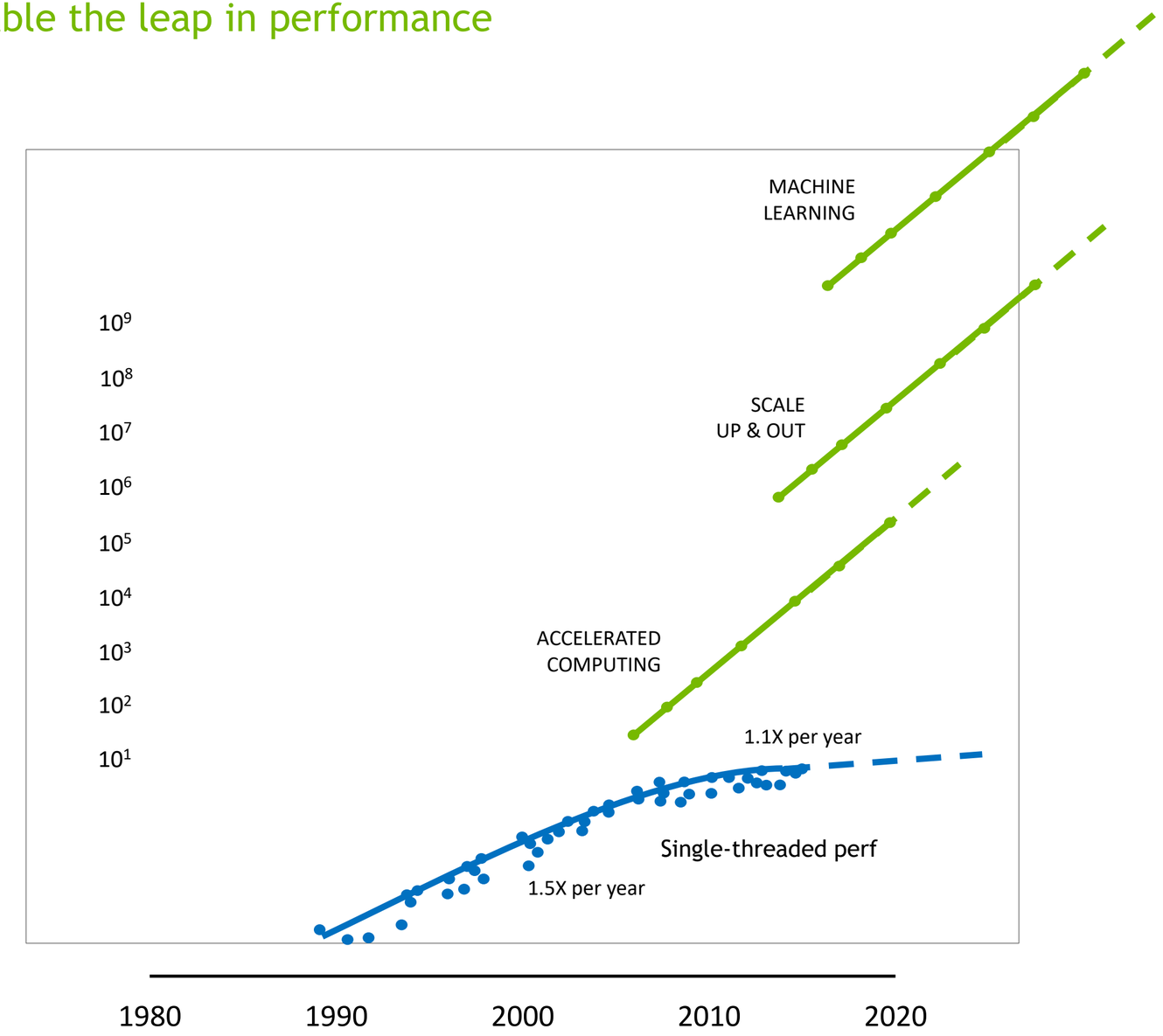


AI can allow foundries to deploy new lithography solutions, like ILT, necessary to continue semiconductor scaling

MILLION-X LEAP IN SCIENTIFIC COMPUTING

AI/ML to enable the leap in performance





AI COMPUTATIONAL LITHOGRAPHY

Motivations and Challenges

- AI is well-suited for image understanding tasks
  - Computational Lithography is capturing the relationship among Design, Mask, and Resist.

- AI is fast thanks to the Computing Power from GPU/CUDA
  - Traditional Lithography Simulation (10s) vs Single A100 ML Resist Prediction (5ms)
  - Traditional Mask Optimization Engine (100s) vs Single A100 ML Mask Optimization (5ms)

- Critical challenge of AI computational lithography: **Lacking Data**
  - ML models, simply speaking, are learning distributions, that are built upon well-distributed big data assumptions.
  - Chip data are hard to collect due to the long design cycle and IP protection.

Lithography Simulation with Conditional GAN Backbone

LithoGAN [Ye+,DAC'19]	TEMPO [Ye+,ISPD'20]	DAMO-DLS [Chen+,ICCAD'20]
▸ Thin mask model. ▸ Optical and resist modeling. ▸ Single via simulation on small clip only. ▸ Requires additional effort to predict via location. ▸ Max tile size: $1\mu m^2$. ▸ DCGAN	▸ Thick mask model. ▸ Optical modeling only. ▸ Requires thin mask aerial image as input. ▸ DCGAN	▸ Thin mask model. ▸ Optical and resist modeling. ▸ Multiple via simulation in a tile. ▸ Resolution: $4nm^2$/pixel. ▸ UNet++

- None of them considers the frequency domain characteristics of lithography simulation.

AI Computational Lithography

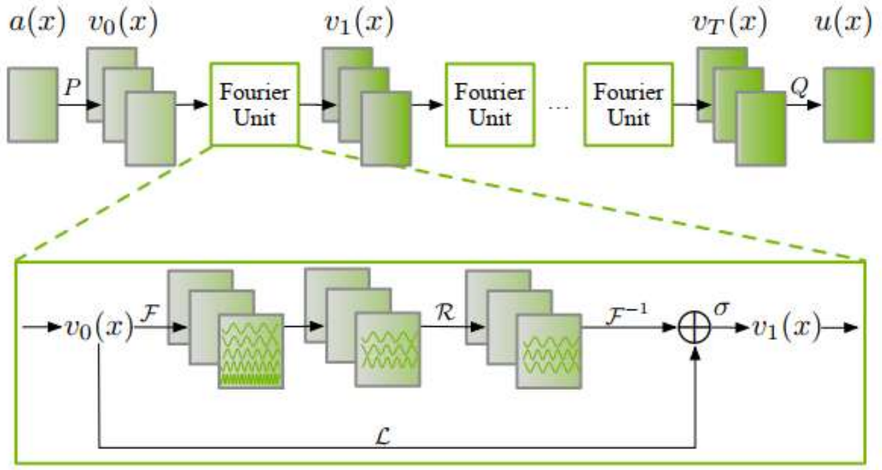
Inductive Bias: Fourier Neural Operator as Lithography Learner

- FNO: Learning Channel Mixing in Frequency Domain

$$\boldsymbol{V}_{t+1} = \sigma(\mathcal{F}^{-1}(\mathcal{F}(\boldsymbol{V}_t) \cdot \boldsymbol{W}_{\mathcal{R}})), \boldsymbol{W}_{\mathcal{R}} \in \mathbb{C}^{C \times C \times H \times W}$$

- Forward Lithography Process

$$\boldsymbol{I} = \sum_{k=1}^{l} \alpha_k |\mathcal{F}^{-1}(\mathcal{F}(\boldsymbol{h}_k) \odot \mathcal{F}(\boldsymbol{M}))|^2$$

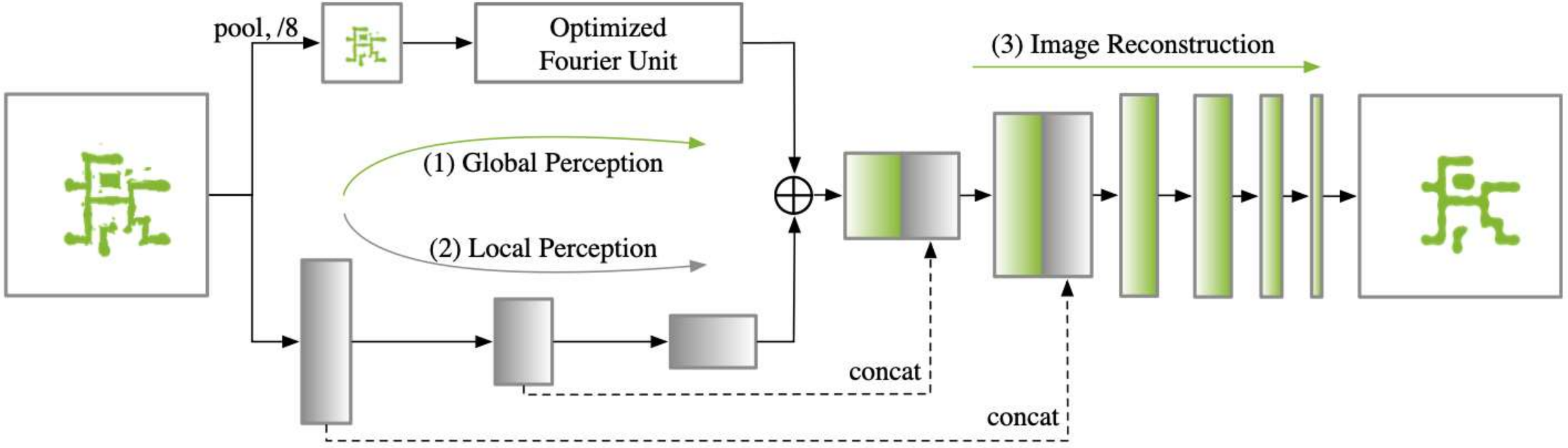


Analogy between lithography simulation and FNO

Step	Lithography Simulation	FNO
1	$\mathcal{F}(\boldsymbol{M})$: FFT on rasterized mask	$\mathcal{F}(\boldsymbol{V}_t)$: FFT on input space
2	$\mathcal{F}(\boldsymbol{h}_k)(\cdot)$: Linear transformation with lithography kernels	$\boldsymbol{W}_{\mathcal{R}}(\cdot)$: Linear channel mixing
3	$\mathcal{F}^{-1}(\cdot)$: Convert back to spatial domain	$\mathcal{F}^{-1}(\cdot)$: Convert back to spatial domain
4	$\alpha[\cdot]^2$: Weighted summation across intensity responses to all lithography kernels	σ: Some activation

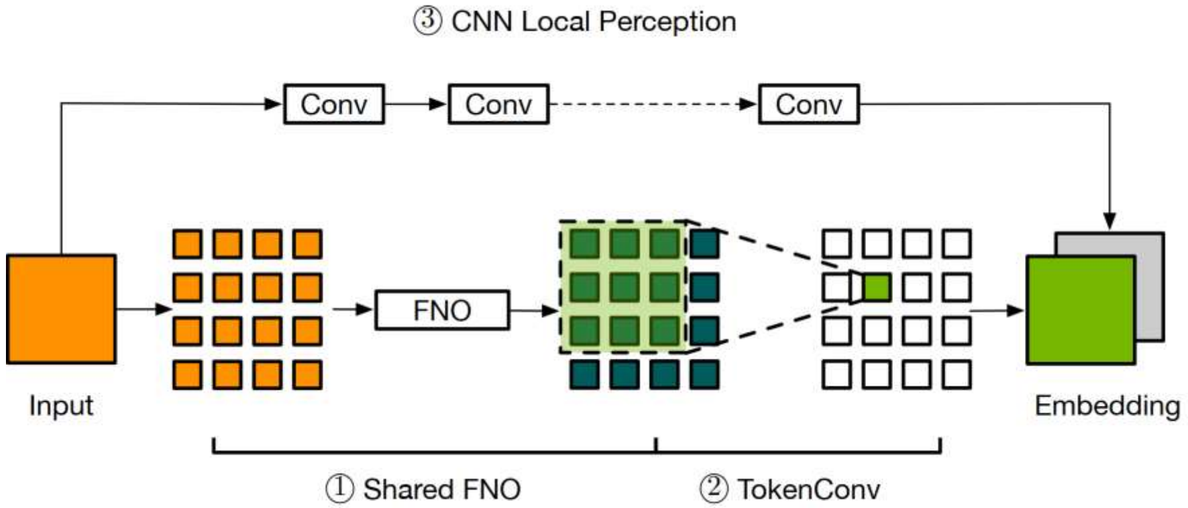
Li, Zongyi, et al. "Fourier neural operator for parametric partial differential equations." *arXiv preprint arXiv:2010.08895* (2020).

Global+Local Lithography Modeling



The overall contour prediction pipeline of the dual-band neural networks.

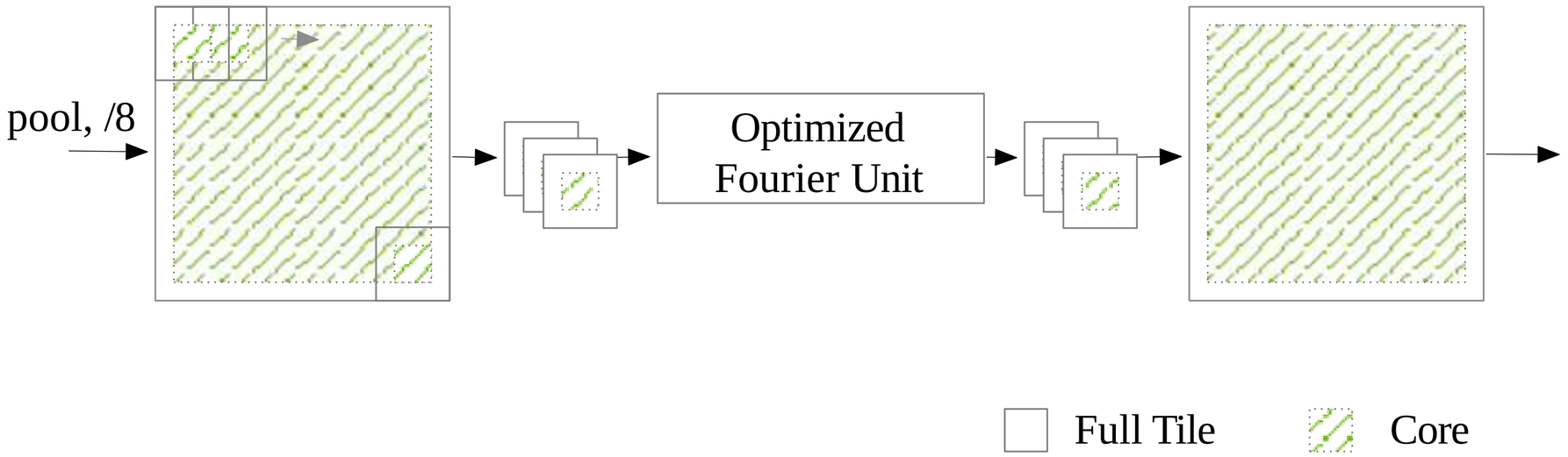
AI Computational Lithography



- Shared FNO:
  - Input images will be divided into non-overlapped patches, which will share the same FNO unit to learn global embedding.
- TokenConv:
  - Capture the spatial/long-range dependency among neighbor patches.
- Local Perception:
  - An auxiliary path with stacked convolution layers to capture local information.

- Design Note:
  - FNO introduces inductive bias of the lithography process
  - TokenConv contributes to long-range dependency
  - Patch size and tokenConv can be adjusted to accommodate different receptive field.

FULL CHIP GLOBAL PERCEPTION



- The light intensity at a location is determined by area surrounding location.
- Simulated contours near the tile boundary are not reliable.

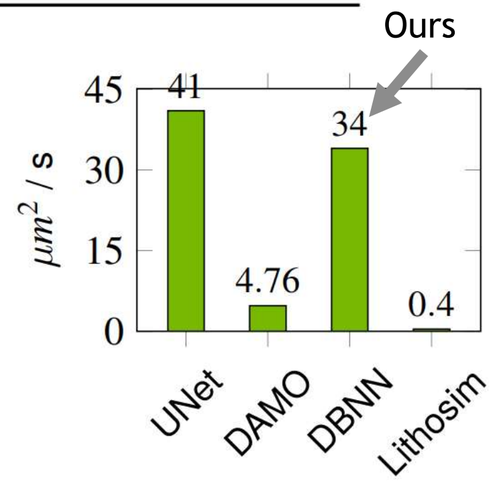
Our Contributions

AI for Lithography Modeling

Result Comparison with State-of-the-art

Benchmark	UNet		DAMO-DLS		Ours	
	mPA (%)	mIOU (%)	mPA (%)	mIOU (%)	mPA (%)	mIOU (%)
ISPD-2019 (L)	99.40	98.03	99.25	98.11	**99.43**	**98.27**
ISPD-2019 (H)	99.08	97.97	-	-	**99.21**	**98.45**
ICCAD-2013 (L)	97.30	95.38	98.94	96.97	**98.98**	**97.79**
ICCAD-2013 (H)	95.16	93.04	-	-	**99.12**	**97.77**

Benchmark	UNet		Ours	
	mPA (%)	mIOU (%)	mPA (%)	mIOU (%)
N14	94.39	91.64	**98.68**	**96.49**
N3	99.96	99.92	**99.97**	**99.94**



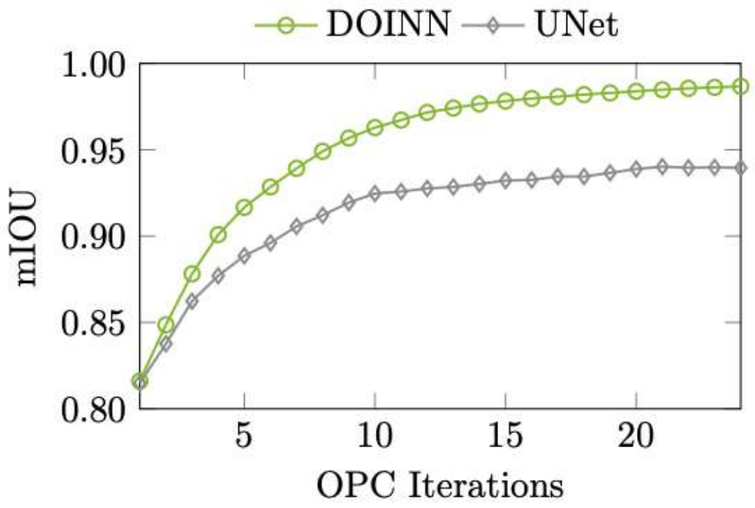
Predicted resist pattern vs simulated resist pattern

$$\text{mIOU}(P,G) = \frac{1}{k}\sum_{i=1}^{k}\frac{P_i \cap G_i}{P_i \cup G_i}.$$

$$\text{mPA}(P,G) = \frac{1}{k}\sum_{i=1}^{k}\frac{P_i \cap G_i}{G_i}.$$

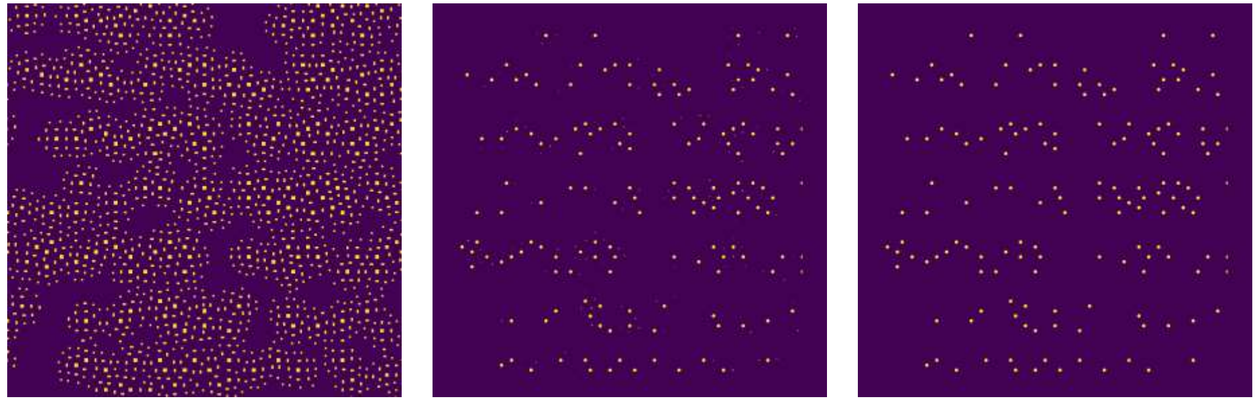
- Compared to the state-of-the-art ML-based lithography simulator, we have: a **20X** smaller model size (1.3M vs 20M), 2% higher simulation accuracy, and 10X faster training convergence (10 epoch vs 100 epoch), 7X faster simulation speed.
- Compared to an open-source physical lithography simulator (Lithosim) or Calibre, we have: a <1% accuracy loss with **85X** speedup.

Full Chip Simulation

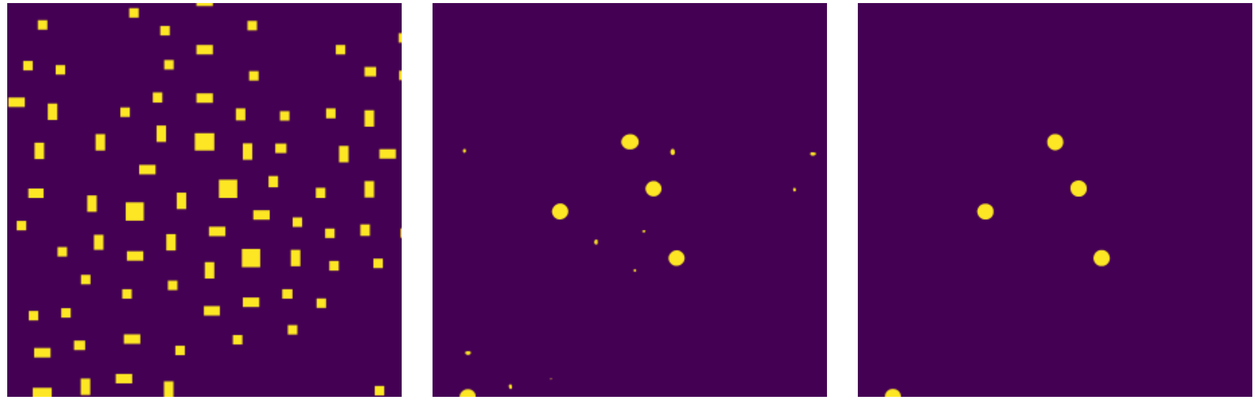


Lithography modeling performance on subtle perturbations.

ISPD-2019-Full	mPA (%)	mIOU (%)
DBNN	96.30	92.03
DBNN-FS	**99.25**	**98.23**



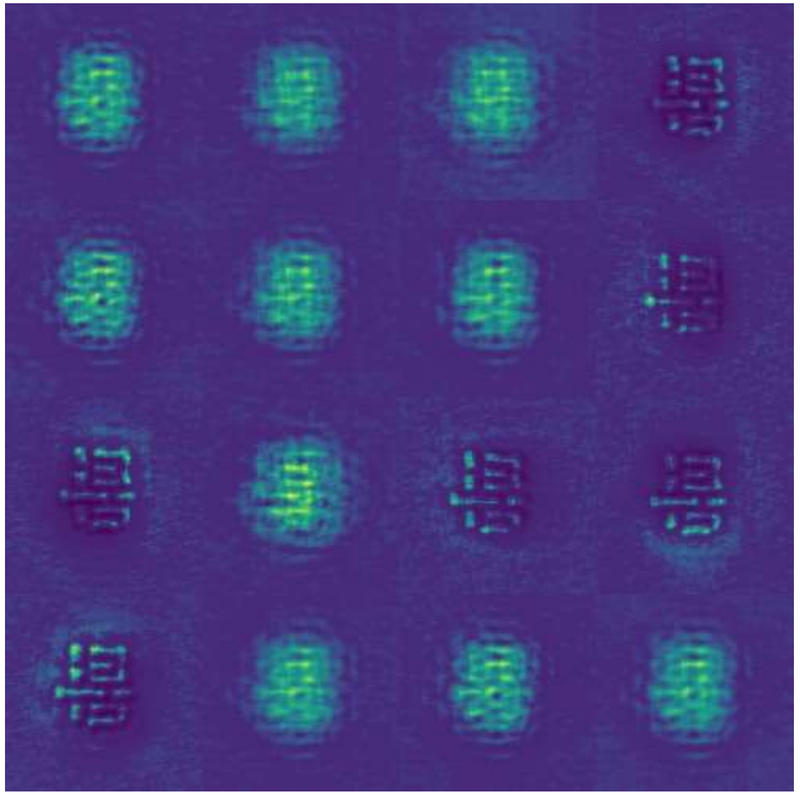
(a) Mask (b) DOINN (c) DOINN-FS



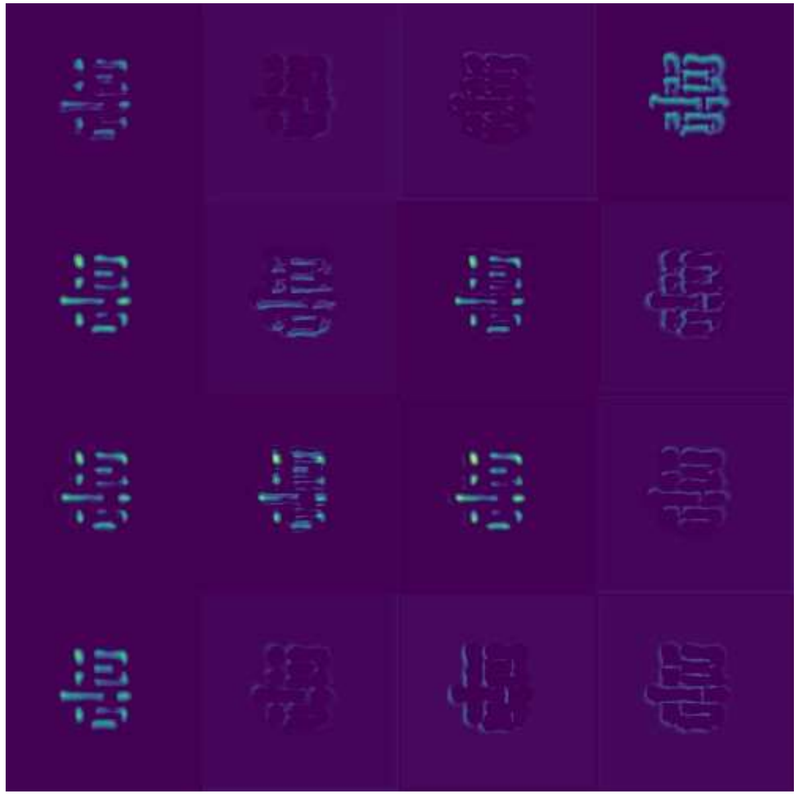
(d) Mask (e) DOINN (f) DOINN-FS

NVIDIA

What Does DOINN Learn?



(a) Global Perception

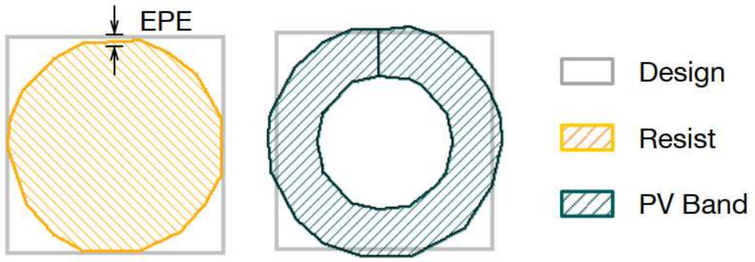


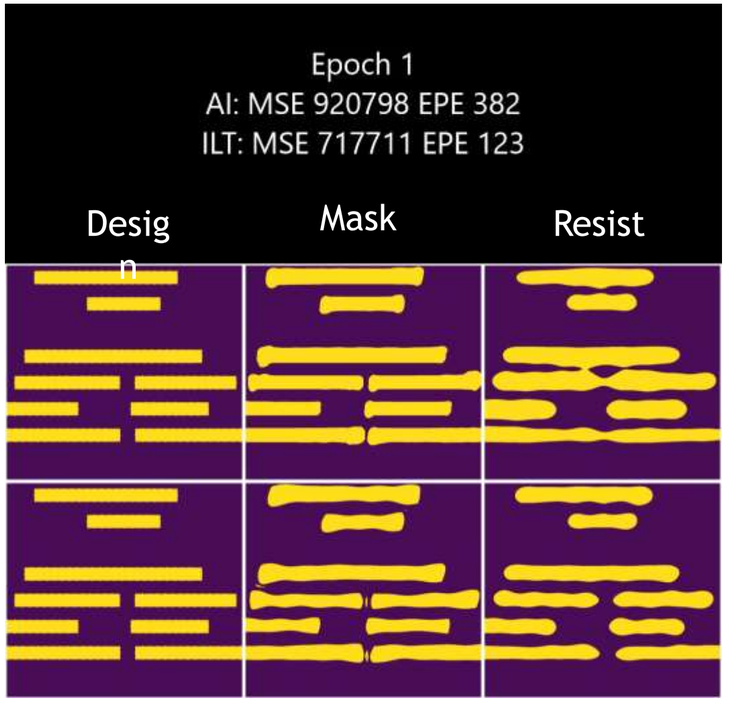
(b) Local Perception

Our Contributions

AI for Inverse Lithography (Mask Optimization)

Evaluation of Mask Quality





Compared to the state-of-the-art academic mask optimization engine, we can solve the mask optimization in single inference.

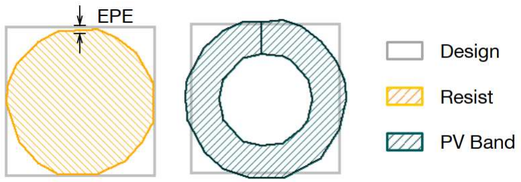
600x speedup and better mask quality than numerical solvers provide better mask quality: **57X smaller EPE violation**

progressive self-training of FNO with better design samples, which is not feasible with a traditional solver.

Our Contributions

AI for Inverse Lithography (Mask Optimization)

Evaluation of Mask Quality

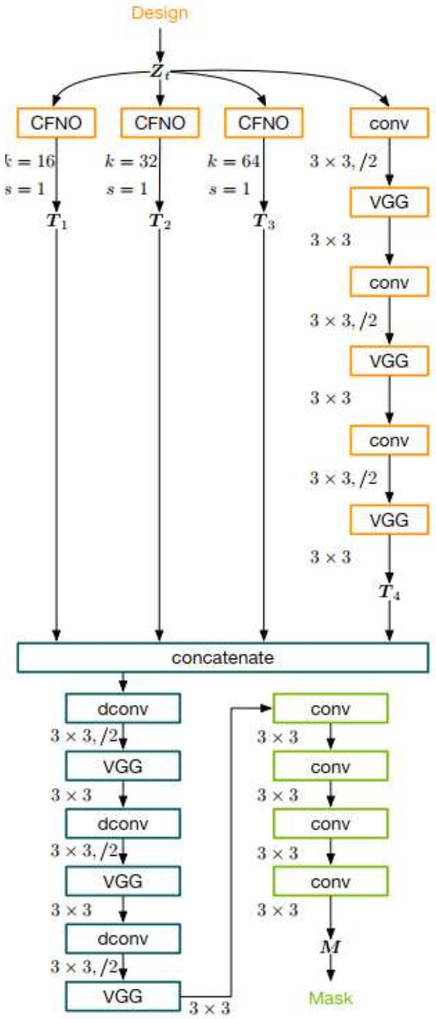


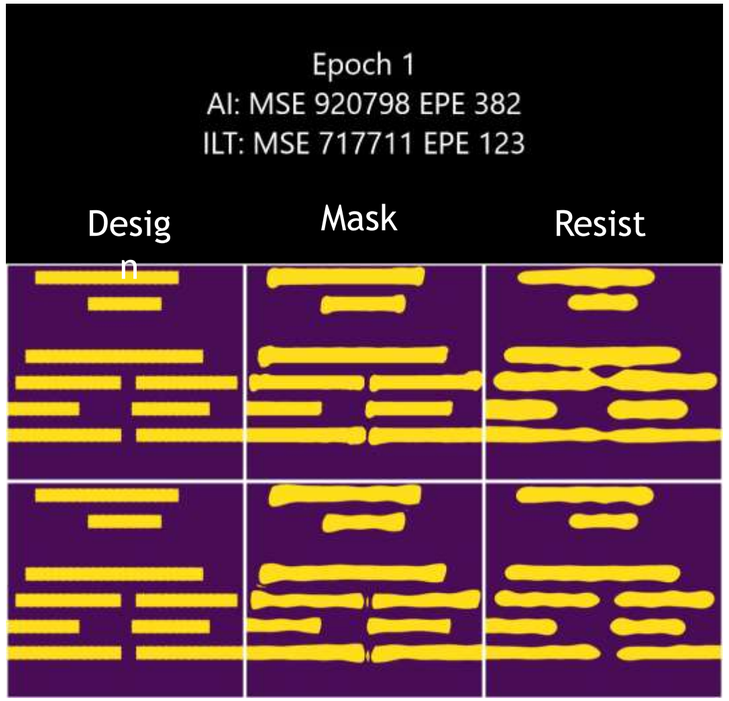
Definition 1 (EPE Violation[20]). EPE is measured as the geometric distance between the target edge and the lithographic contour printed at the nominal condition. If the EPE measured at a point is greater than certain tolerance value, we call it an EPE violation.

Definition 2 (MSE). MSE measures the pixel-wise difference between the design and the resist image as in:

$$\text{MSE} = ||Z - Z_t||_F^2. \quad (3)$$

Definition 3 (PVB Area[20]). This is evaluated by running lithography simulation at different corners on the final mask solution. Once run, a process variation band metric will be defined as the XOR of all the contours. The total area of the process variation band is defined as PVB Area.





- Compared to the state-of-the-art academic mask optimization engine: We can solve the mask optimization in single inference, achieve a speedup of an order of magnitude (<1s on A100 vs a couple of hours per tile), and provide better mask quality in terms of **57X smaller EPE violation** (2.7 vs 165.2).

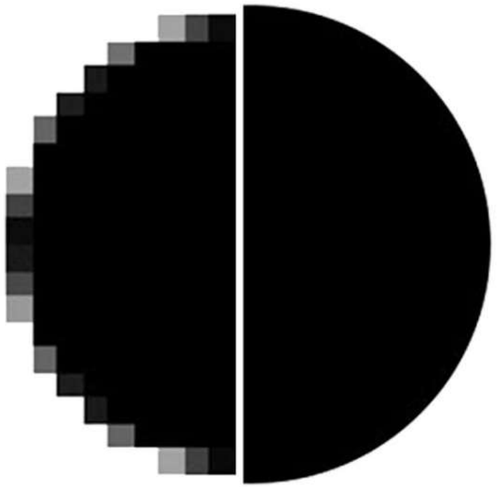
FOURIER NEURAL OPERATOR

DISCRETIZATION-INVARIANT LEARNING

One AI model for any discretization: no re-training

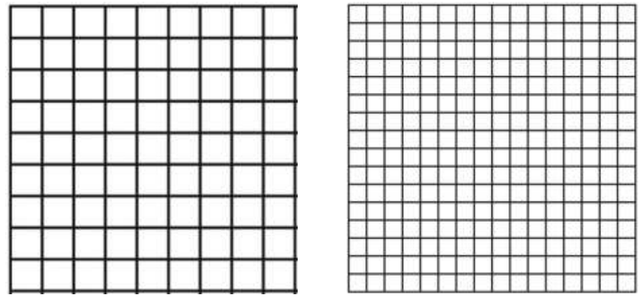
Neural Network

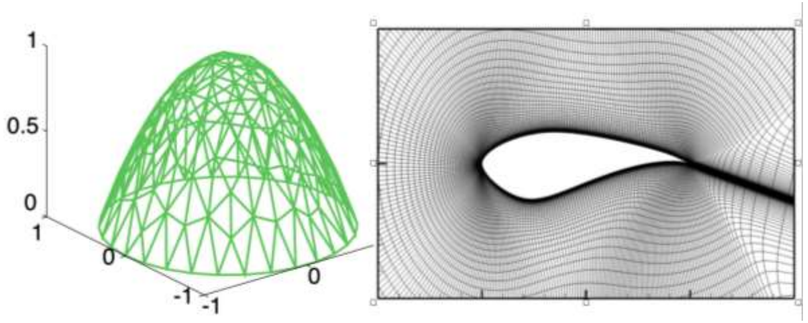
Input and output at fixed resolution



Neural Operator

Input and output at any points in domain

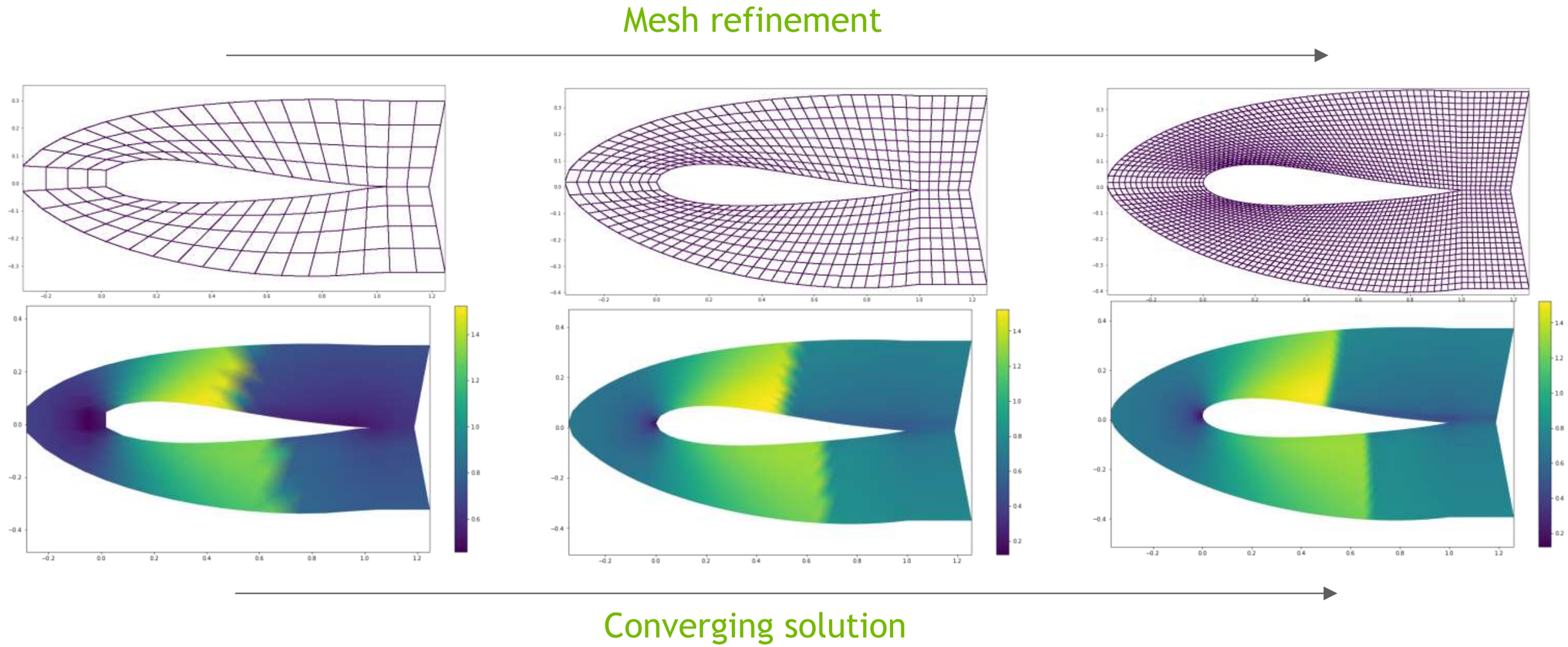




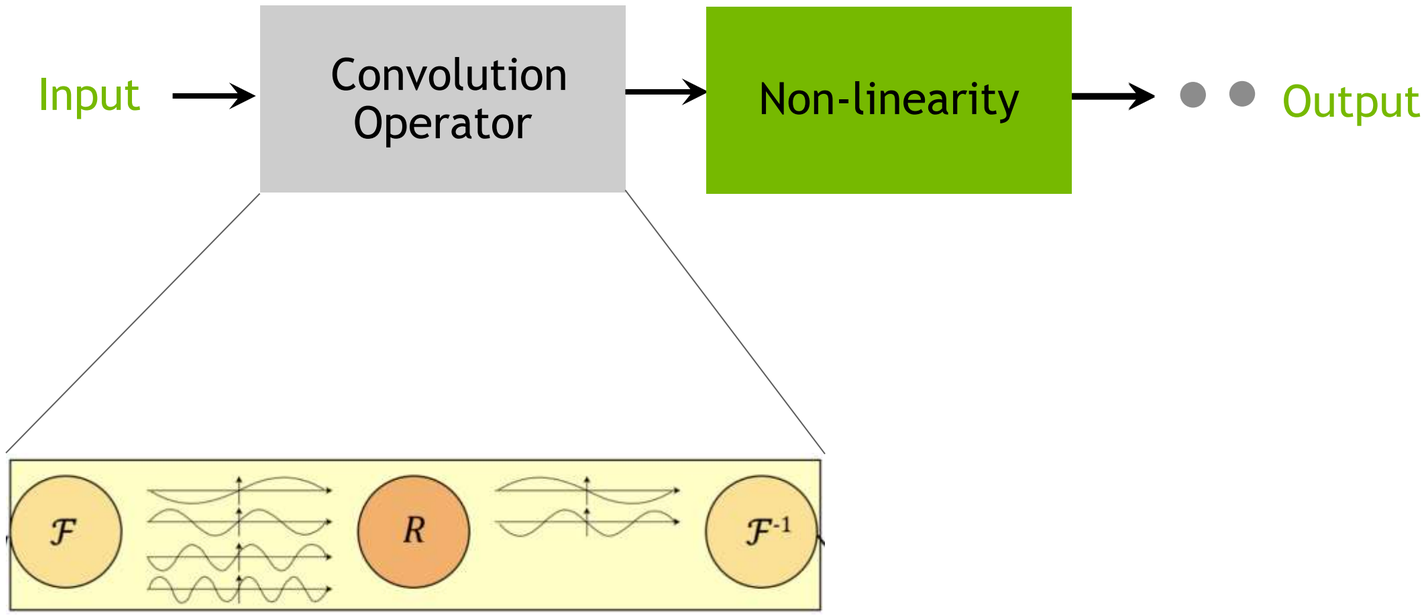
DISCRETIZATION-INVARIANCE OF NEURAL OPERATOR

Definition: a trained AI model is discretization-invariant if

- We can query at any point.
- Converges upon mesh refinement to a limit.

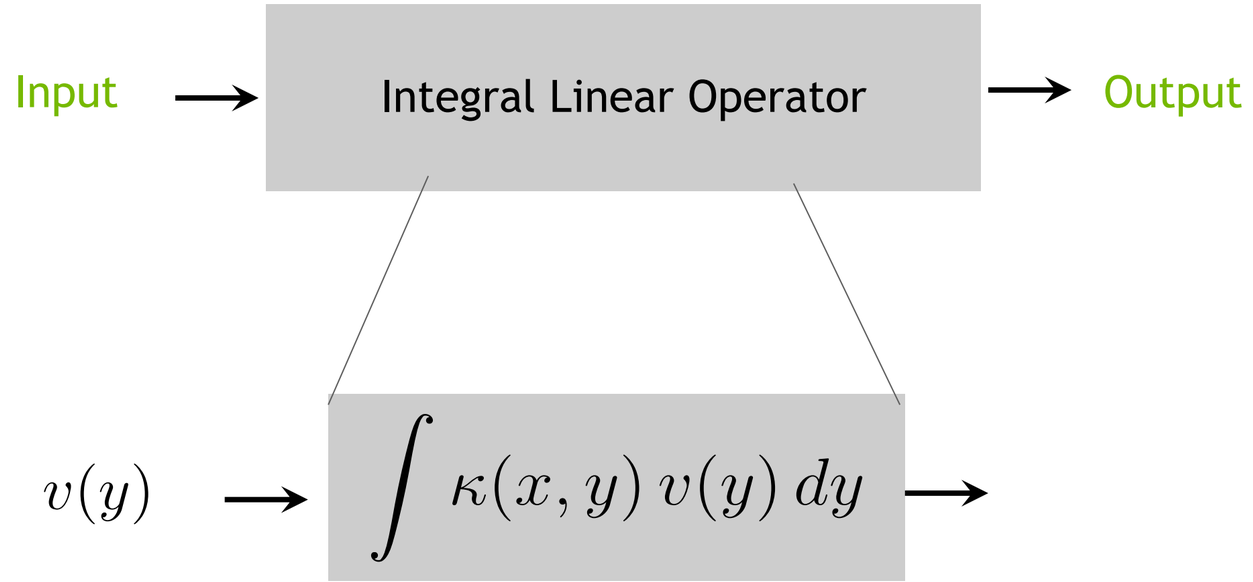


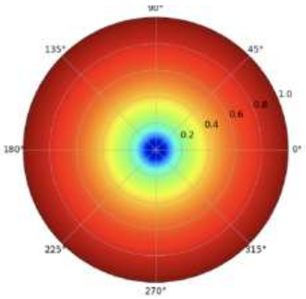
FNO: FOURIER NEURAL OPERATOR



- Convolution = multiplication in frequency domain.
- Learning weights in frequency domain.
- Fourier Transform implements convolution and also discretization invariant.

INTEGRAL OPERATOR FOR SOLVING LINEAR PDE



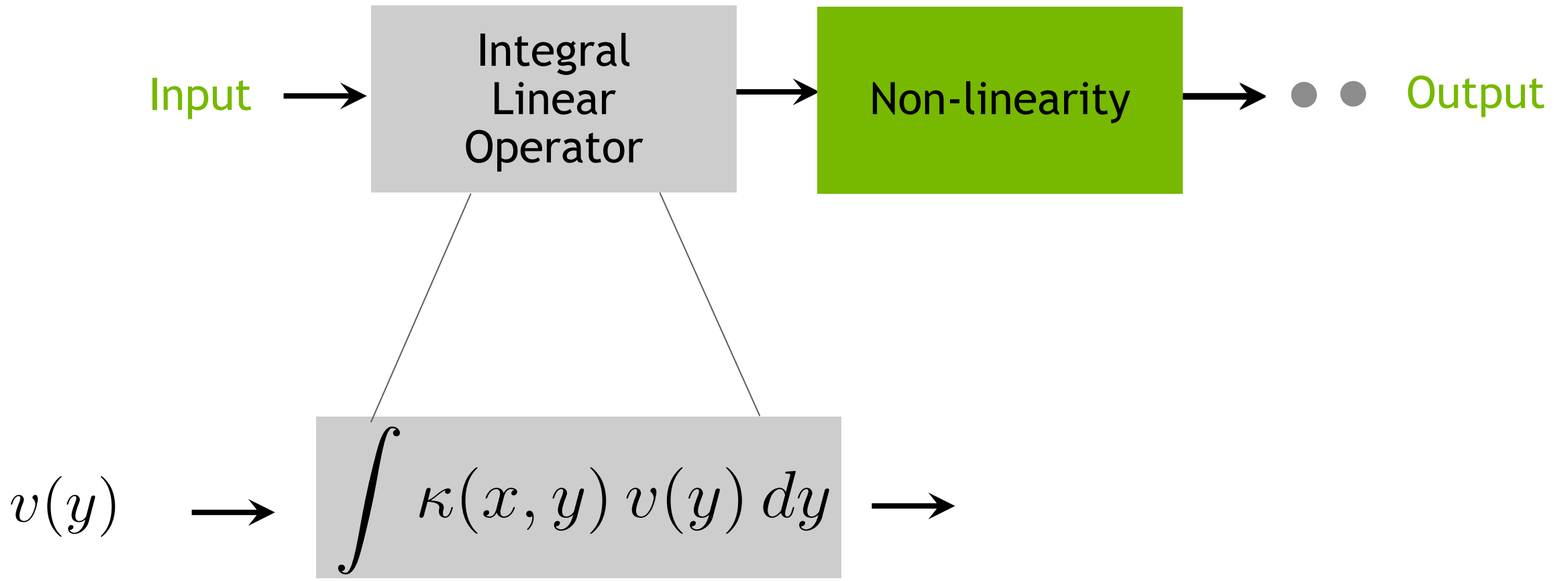


$K(x, y)$

Kernel of integral operator
For heat diffusion

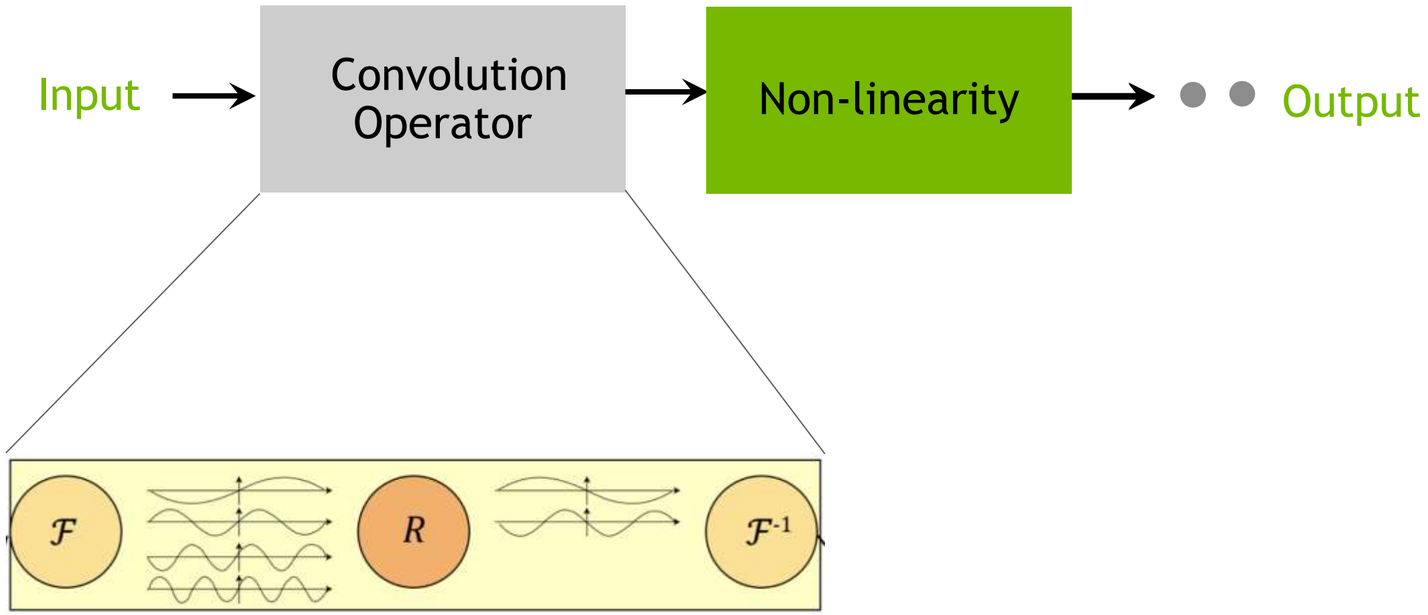
- Integral operator outputs functions (not just finite-dimensional vectors).
- Integral operator is discretization invariant.

NEURAL OPERATOR: A GENERAL FRAMEWORK



- Integral operator outputs functions (not just finite-dimensional vectors).
- Integral operator is discretization invariant.

FNO: FOURIER NEURAL OPERATOR



- Convolution = multiplication in frequency domain.
- Learning weights in frequency domain.
- Fourier Transform implements convolution and also discretization invariant.

DISCRETIZATION-INVARIANCE

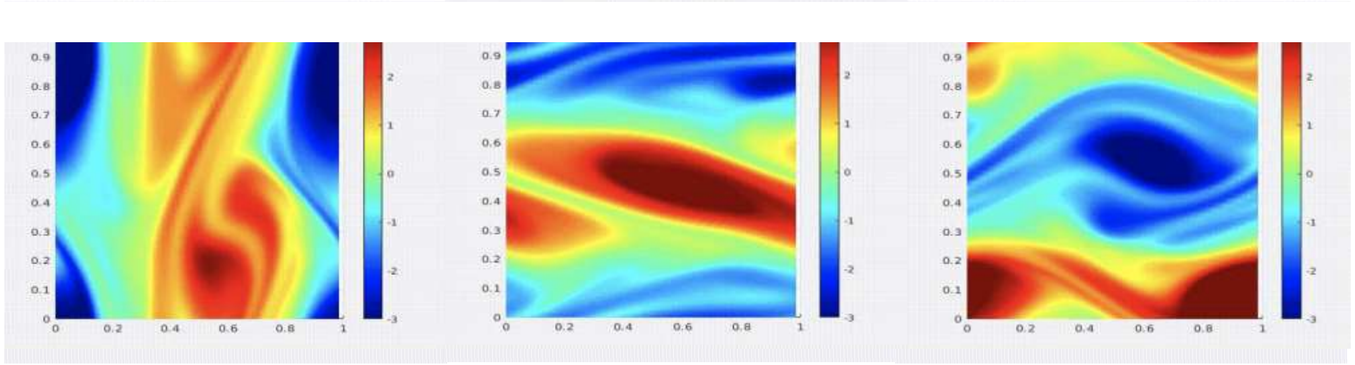
Property \ Model	NNs	DeepONets	Interpolation	Neural Operators
Discretization Invariance	✗	✗	✓	✓
Is the output a function?	✗	✓	✓	✓
Can query the output at any point?	✗	✓	✓	✓
Can take the input at any point?	✗	✗	✓	✓
Universal Approximation	✗	✓	✗	✓

- Neural operators are discretization-invariant.
- Neural operators are universal approximators in function spaces.

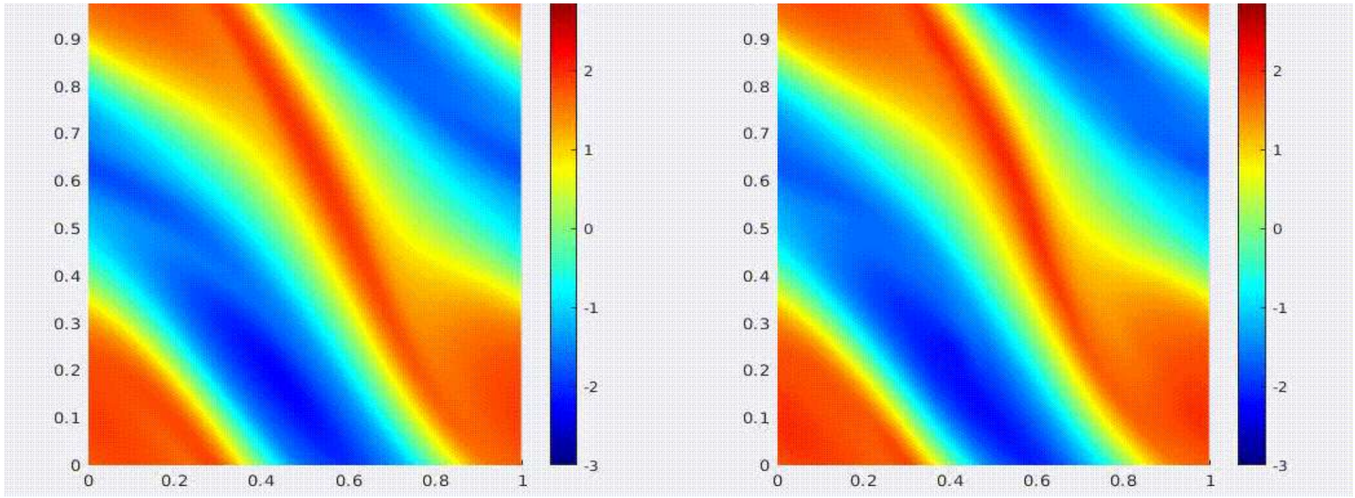
DEMONSTRATING DISCRETIZATION INVARIANCE OF FNO

Zero-shot super-resolution

Train using coarse resolution data

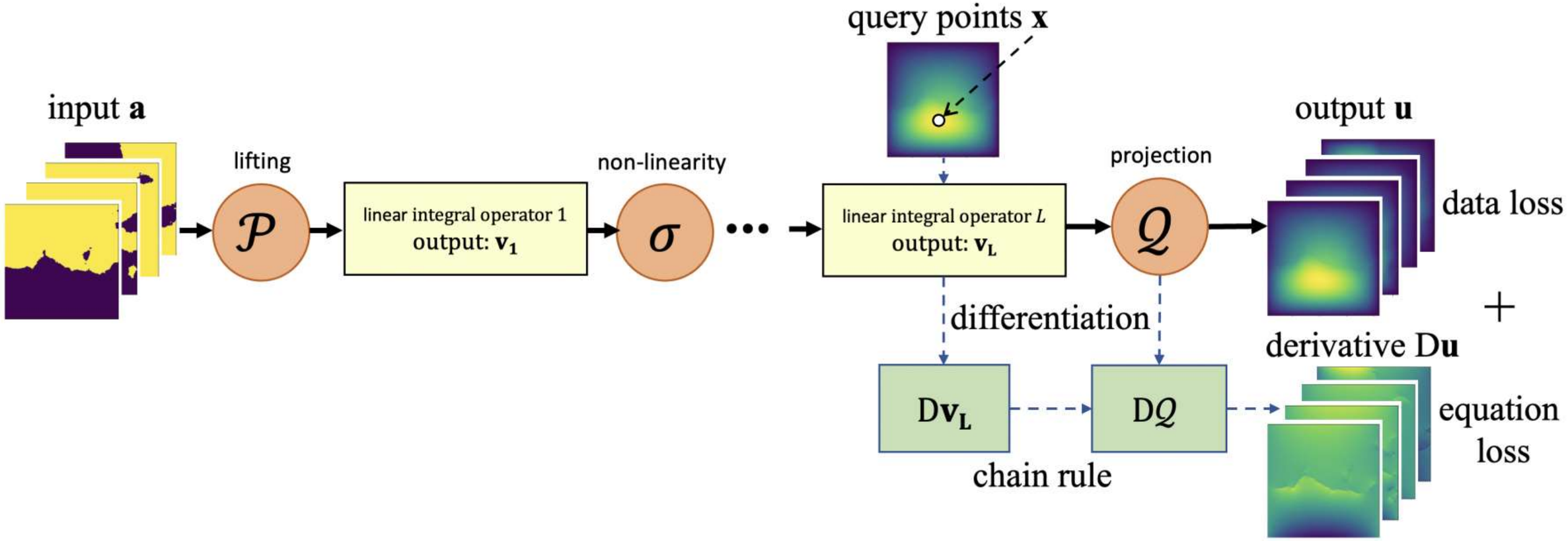


Directly evaluate on higher resolution (no re-training)



PHYSICS INFORMED NEURAL OPERATORS

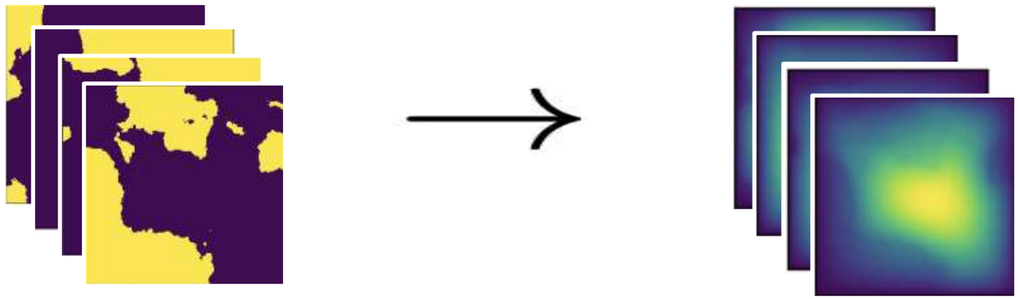
PINO: PHYSICS-INFORMED NEURAL OPERATOR



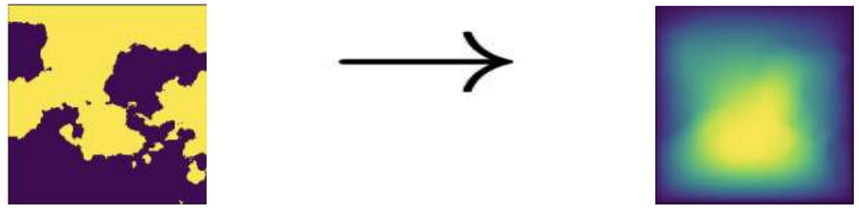
PINO: PHYSICS-INFORMED NEURAL OPERATOR

PINO can learn solution operator for a family of equations and fine-tune on an instance

Operator learning

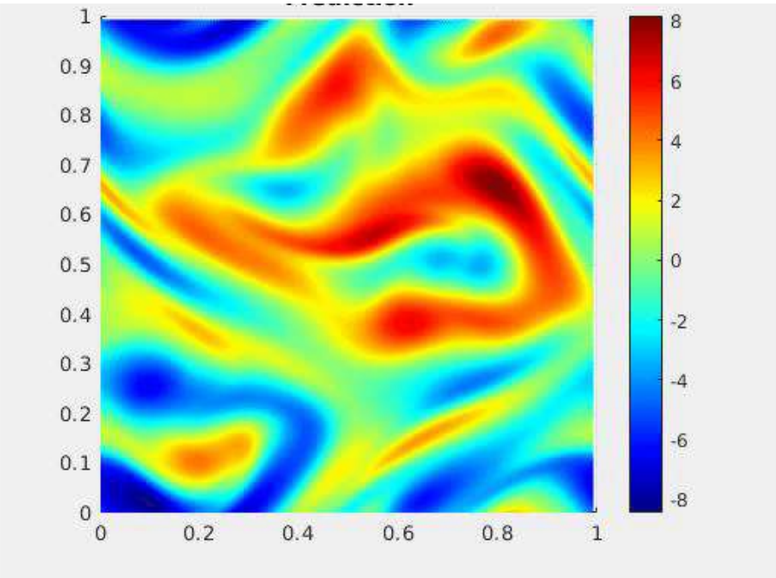


Instance-wise finetuning

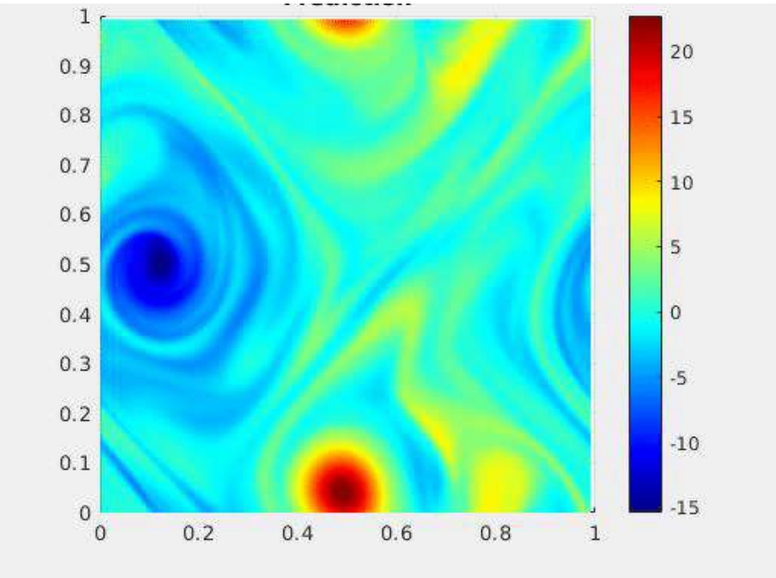


TRANSFER LEARNING WITH PINO

Operator learned on Re100, fine-tune to Re500. Converges 3x faster.



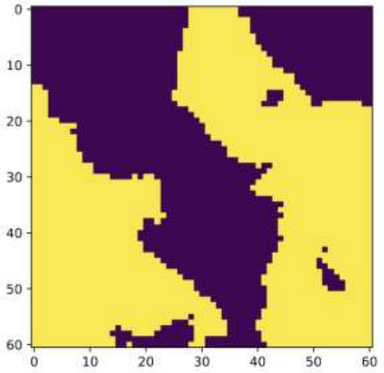
Re100



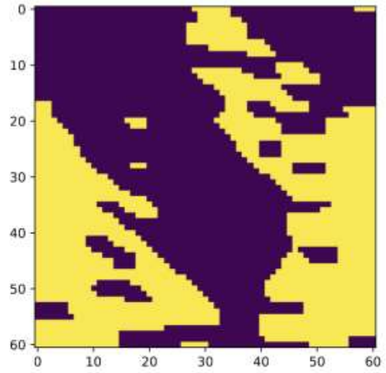
Re500

INVERSE PROBLEMS WITH PINO

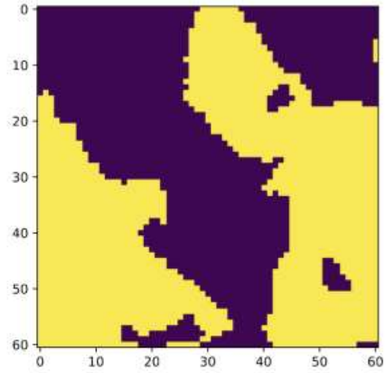
Inverse problem: given solution of forward simulation, recover input.
PINO makes the inverse prediction more physically valid.



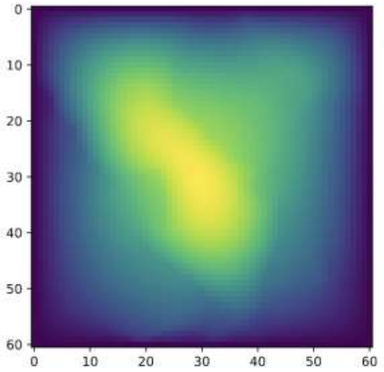
(a) Ground truth input a



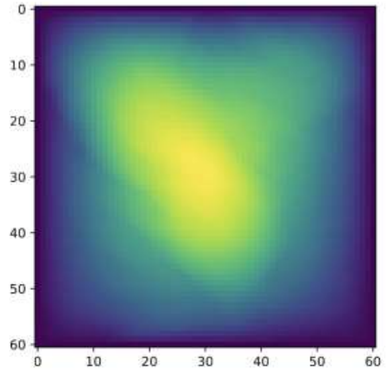
(b) Inversion using only data constraint



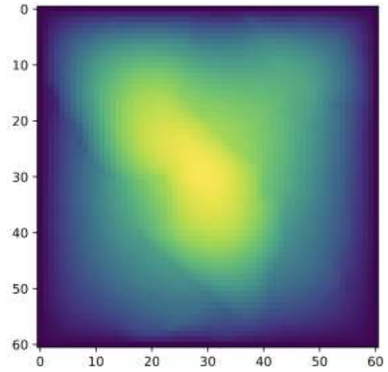
(c) Inversion using data and PDE constraints



(d) Observed output function



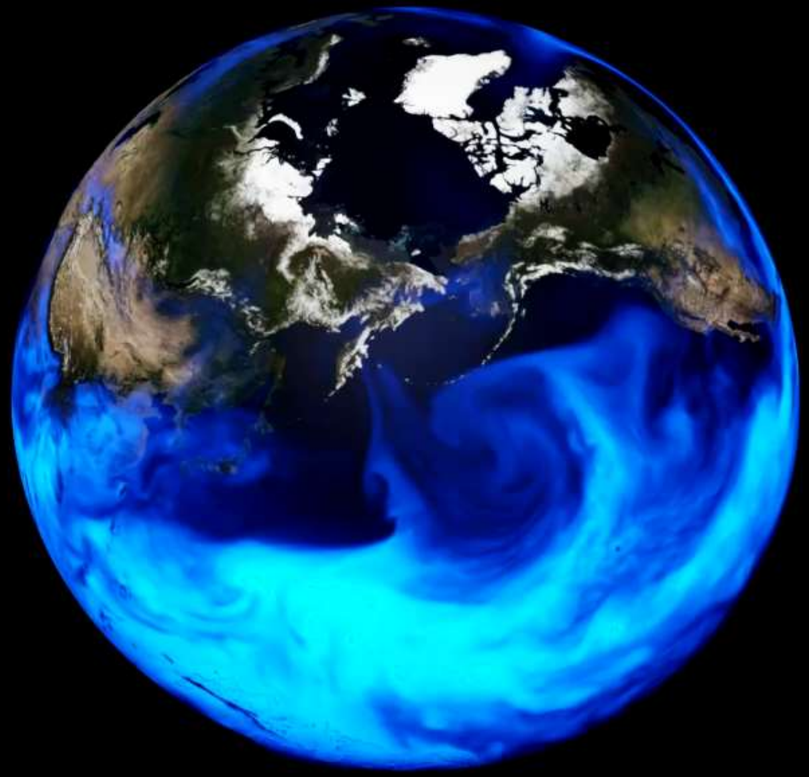
(e) Output function of inversion using only data constraint



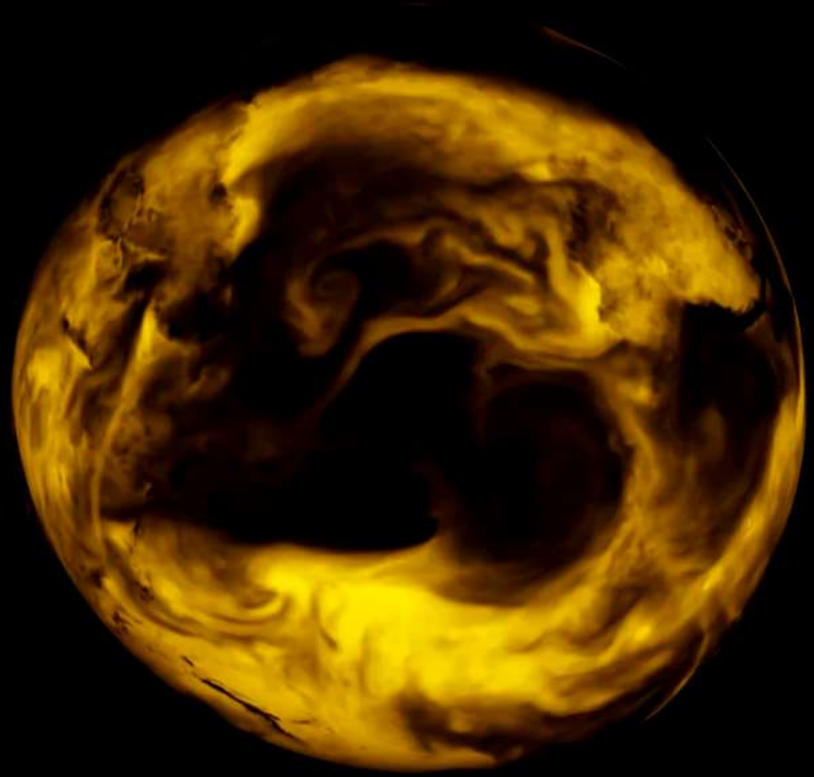
(f) Output function of inversion using data and PDE constraints


AI FOR WEATHER AND CLIMATE

Ground Truth



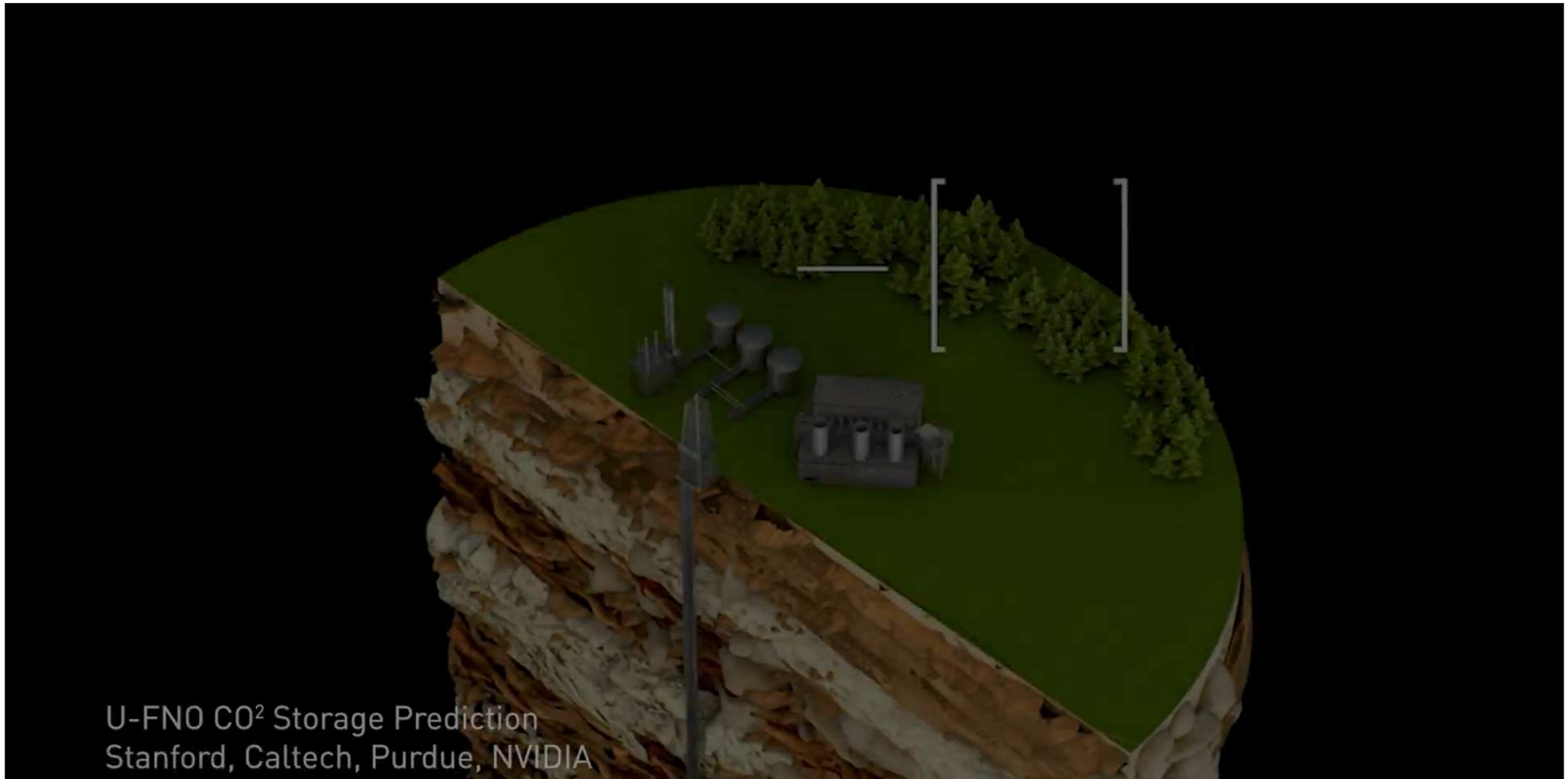
FourCastNet



Our AI (FourCastNet) is **45,000 times** faster than current weather models

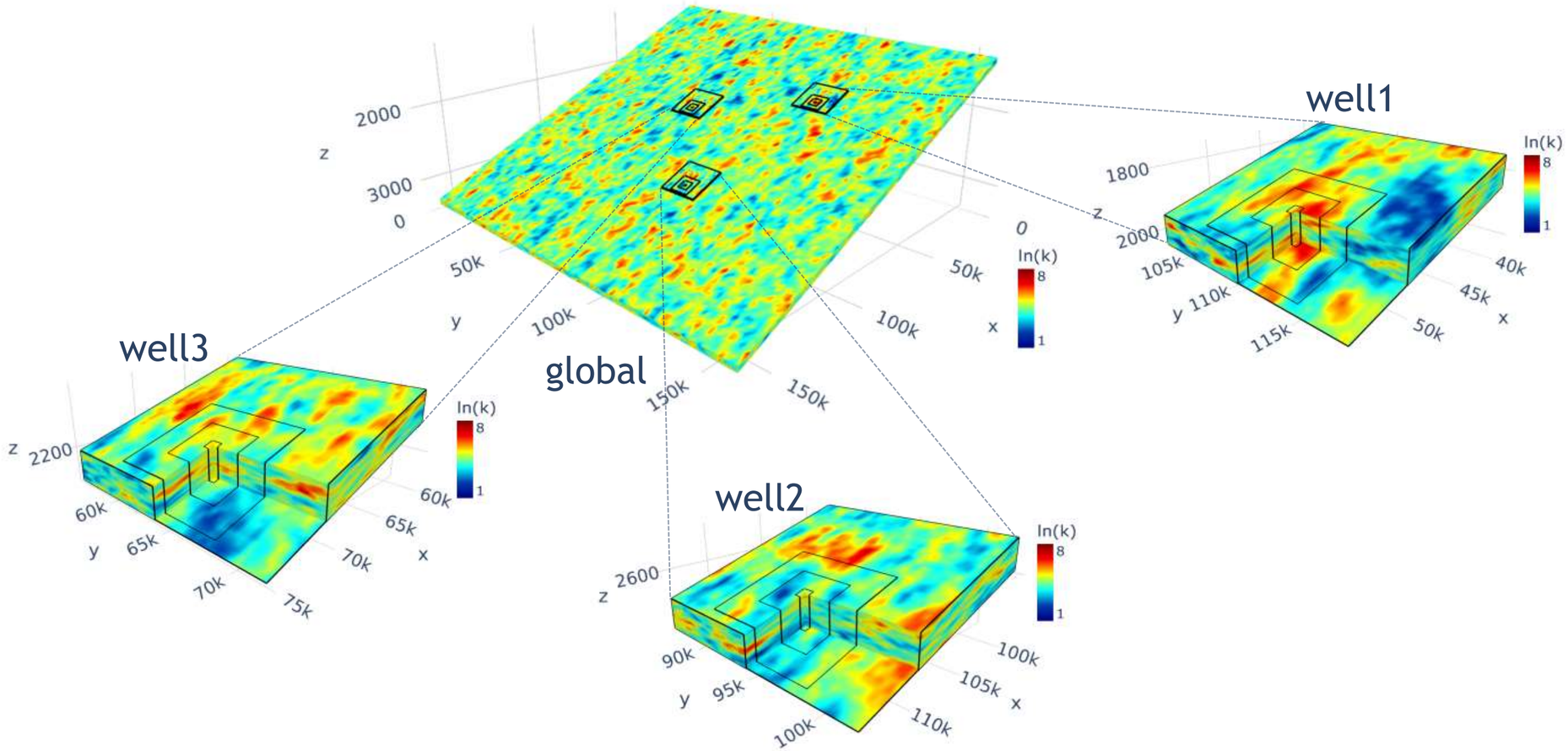
CLIMATE CHANGE MITIGATION: MODELING CO_2 STORAGE

Our AI Method accelerates by 700,000 times



FOUR-DIMENSIONAL CCS MODELING WITH AI (FNO)

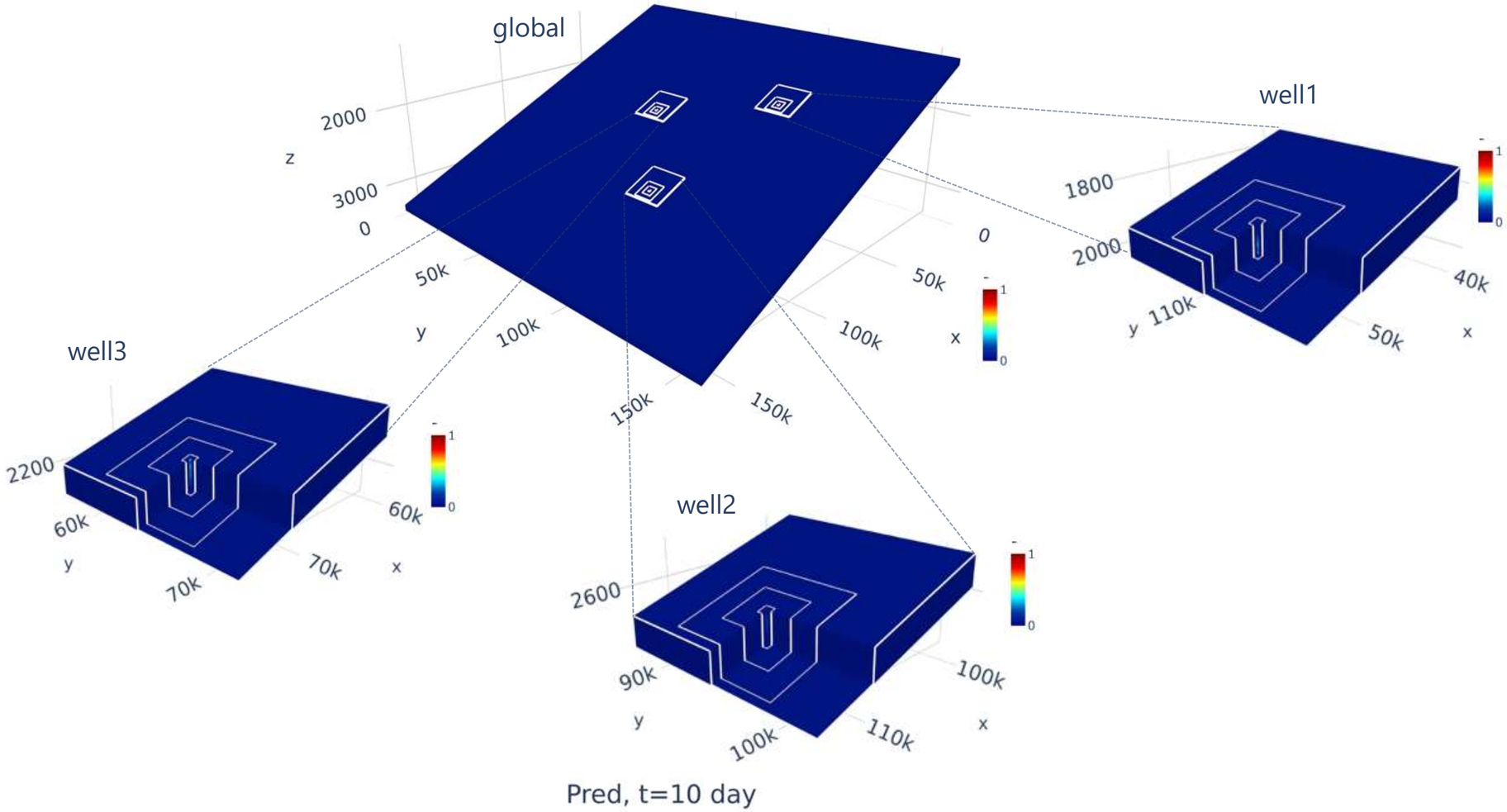
Our AI Method accelerates by 700,000 times



Permeability Heat Map

FOUR-DIMENSIONAL CCS MODELING WITH AI (FNO)

Our AI Method accelerates by 700,000 times



Gas Saturation

Summary

- Computational Lithography Challenges
  - Slow, costly, …
- AI for Computational Lithography
  - Promising
  - Lacking Data
- FNO and CFNO Backbone
  - Carrying good inductive biases to be plugged into machine learning models
- Our Contributions
  - AI lithography modeling
  - AI mask optimization