GTC 2018:
Physics-Based AI for Semiconductor Inspection Using a GPU Based Optical Neural Network

Jing Zhang, Principle Software Engineer

KLA-Tencor, AI group
March 29 2018

KLA-Tencor

KLA-Tencor Background

Process Control, Portfolio, Challenges

KLA-Tencor Overview

>40 years

Global Leader in Process Control since 1976

~23,000
tools installed worldwide

~6,400
global employees

17
countries

$3.5B
FY17 revenue

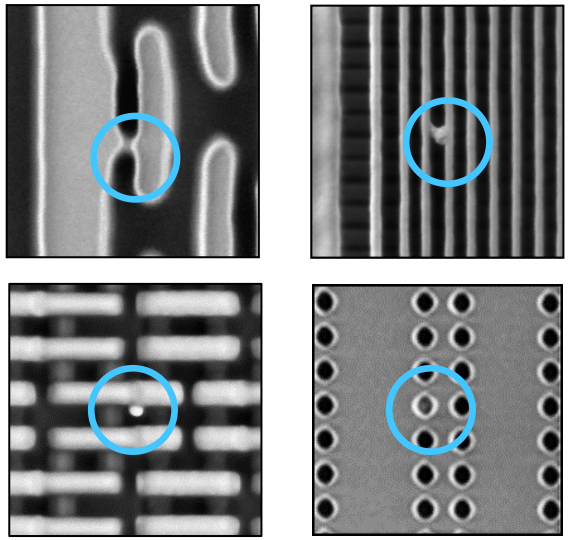
$2.1B
R&D investment over last 4 fiscal years

Our Focus: *Process Control = Foundation of Semiconductor Quality*

"We're blind without you guys…"

Inspection

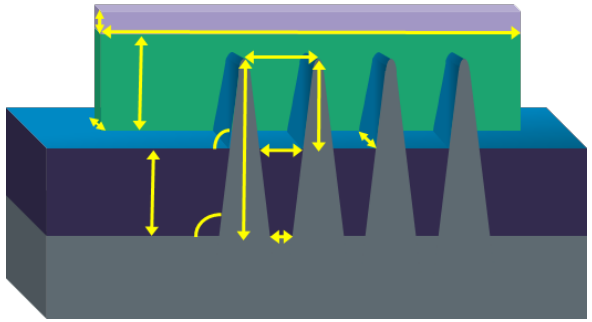
Find Critical Defects

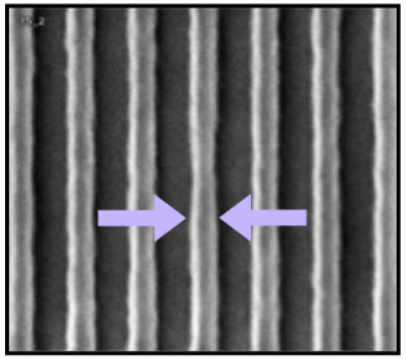


You can't fix
what you can't find

Metrology

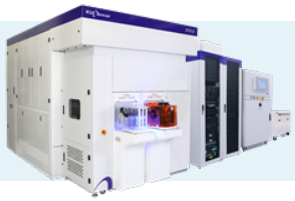
Measure Critical Parameters



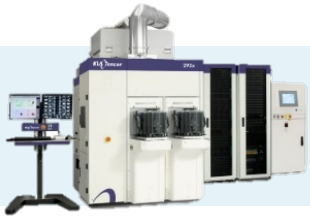


You can't control
what you can't measure

KLA-Tencor’s Inspection Portfolio



3900 Series
Broadband Plasma Wafer



2930 Series
Broadband Plasma Wafer

Puma™ 9980
Laser Scanning Wafer

8920
High Sampling Wafer

CIRCL™
All-Surface Wafer

Surfscan® SP5XP
DUV Unpatterned Wafer

eDR7280™
Wafer SEM Review

eS805™
e-beam Wafer

CIRCL-AP™
Wafer-Level Packaging

ICOS® T3 & T7
Component

Klarity® Defect & ACE
Defect Data Management

RDC
Reticle Data Analysis

Teron™ 640e
Reticle (Mask Shop)

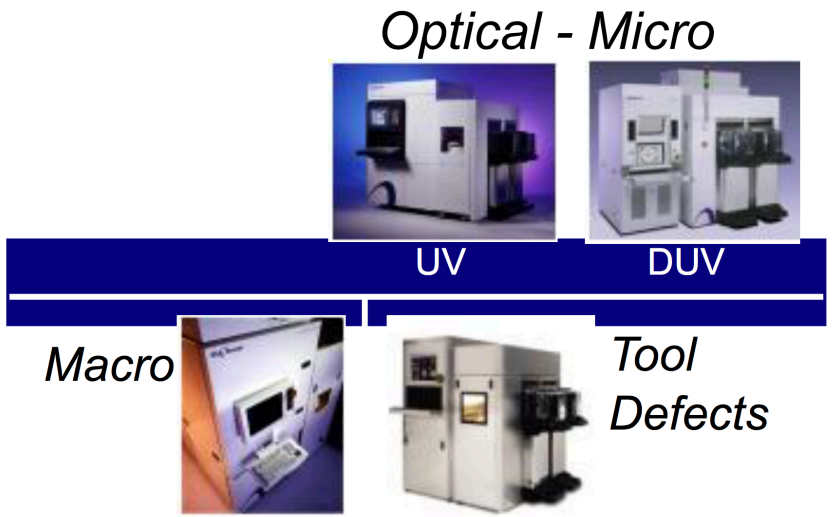
FlashScan™ 200 Series
Reticle Blank

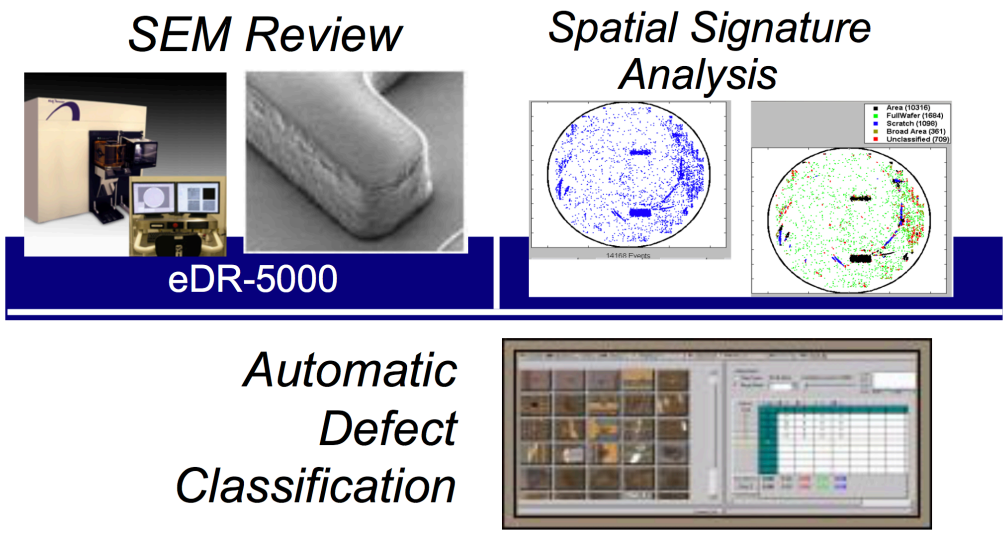
Teron™ SL655
Reticle (IC Fab)

Comprehensive wafer, reticle and component inspection with advanced data analysis supports defect discovery, process optimization and production monitoring

KLA-Tencor

KLA-Tencor Typical Inspection Paradigm





Find Defects

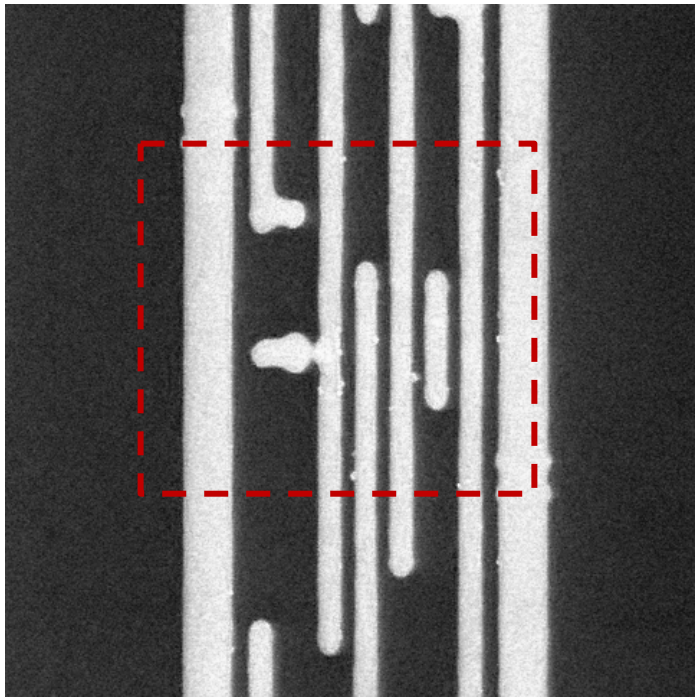
Identify/Verify Defects

KLA-Tencor

Detecting defects without Fully Resolving in Optical Images

Optical Image

(Fast/Low Resolution)

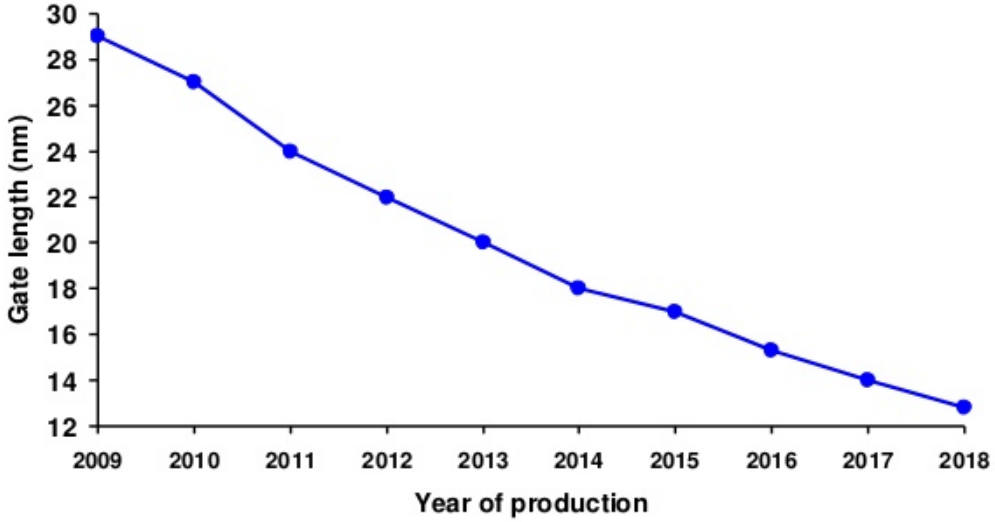


SEM Image

(Slow/High Resolution)

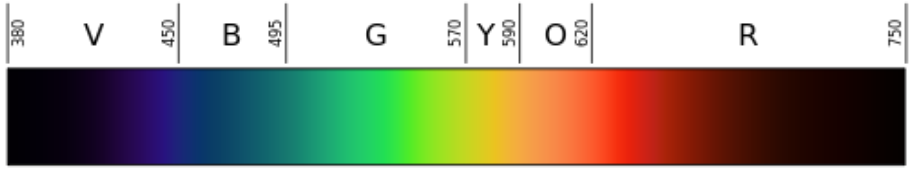
Moore's Law leads to the Challenges for Inspection

The *Actual* Moore's Law
(About *transistor size*.)



[International Technology Roadmap for Semiconductors, 2011]

Approximated Optical Resolution:



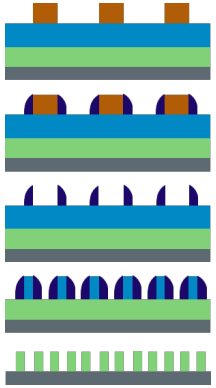
	Red	Green	Violet	UV
Wavelength	700 nm	500 nm	400 nm	200 nm
Min Pitch	390 nm	280 nm	220 nm	110 nm

[Computed based on: X. Ma, G. R. Arce, Computational Lithography.]

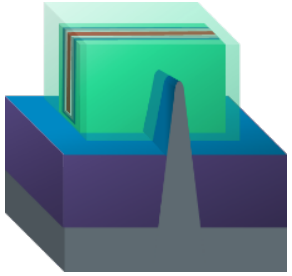
KLA Tencor

Significant Technical Challenges

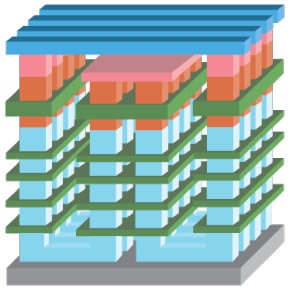
Multi-Patterning



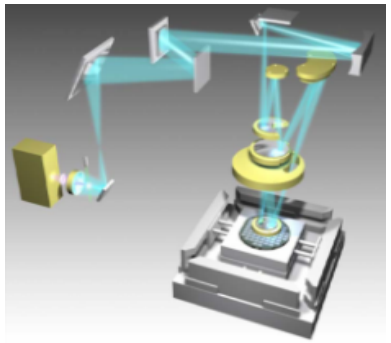
3D Transistors

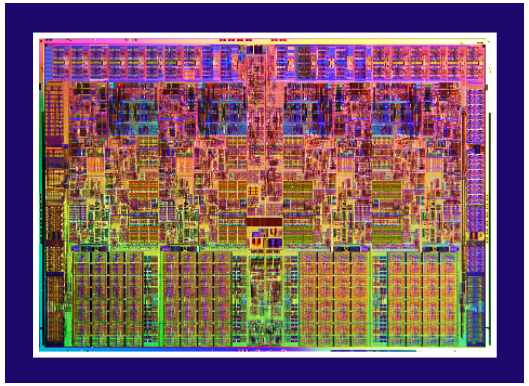


3D Memory

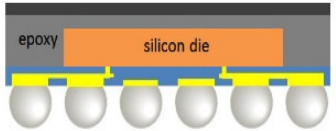


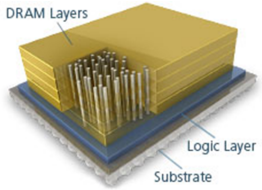
Advanced Lithography



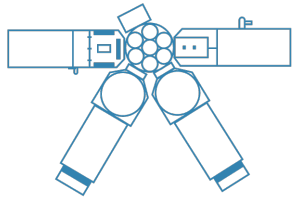


Advanced Packaging

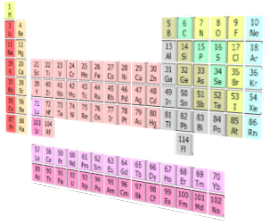




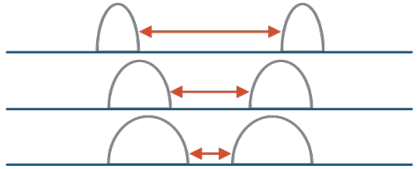
New Processes



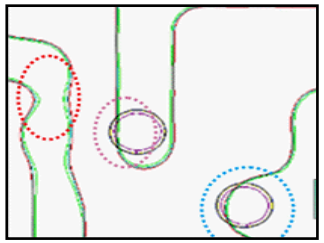
Novel Materials



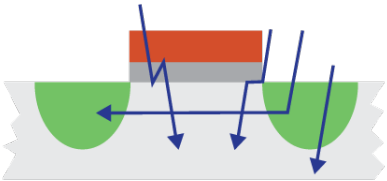
Scaling Challenges



Eroding Process Margins



Increased Variability



KLA Tencor

Motivation

Physical Modeling	Machine Learning
Knowledge: explicitly abstracted into mathematical formula.	**Knowledge**: implicitly represented as learned parameters or dataset.
Data: Limited	**Data**: Plentiful
PROGRAM = ALGORITHMS + data	**MODEL = DATA + network**
*Use when the physical theory is **well-understood** and the solution is **efficient**.*	*Use when the first principle theory is **missing** but data is **plentiful.***

Motivation

Physical Modeling	Machine Learning
Knowledge: explicitly abstracted into mathematical formula.	**Knowledge**: implicitly represented as learned parameters or dataset.
Data: Limited	**Data**: Plentiful
PROGRAM = ALGORITHMS + data	**MODEL = DATA + network**
*Use when the physical theory is **well-understood** and the solution is **efficient**.*	*Use when the first principle theory is **missing** but data is **plentiful.***
It is often too complicated to fully understand every physical process for semiconductor manufacturing.	*Acquisition of meaningful data can be very time-consuming.*

Can we combine these two to address our challenges?

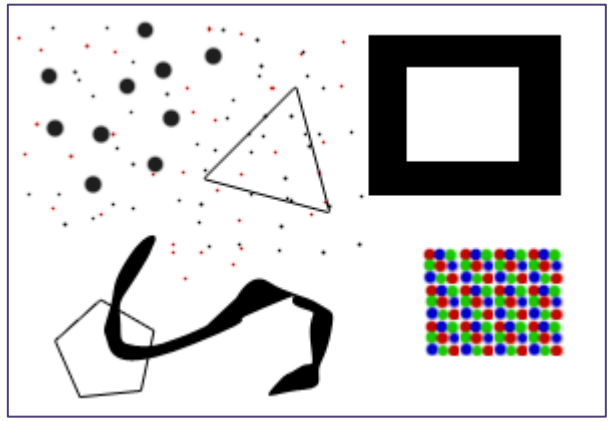
Key Technique Questions

- Is it possible to combine a physical model with Machine Learning?
- How does the combined model meet robustness requirements?
- How does the combined model deal with limited data?
- How to train the combined model in unsupervised manner when we don't have the ground truth?
- Is computation cost of the combined model economic?

Physics-based Deep Learning to Solve Inverse Problem

Inverse Problem: Deconvolution as an example

Ignoring geometry transformation/distortion and sampling errors, the **blur process** is simply modeled by,



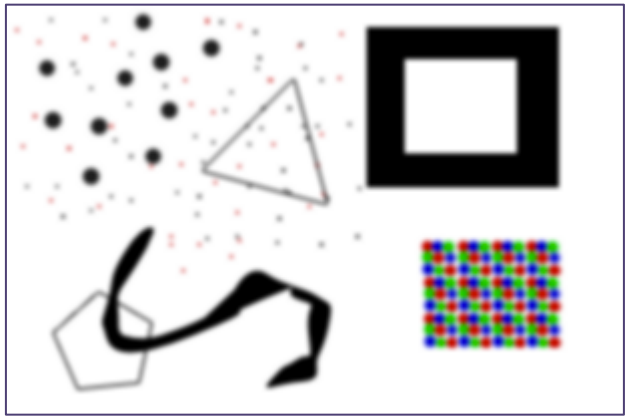
Objective Image

*

0.01	0.02	0.03	0.02	0.01
0.02	0.06	0.08	0.06	0.02
0.03	0.08	0.11	0.08	0.03
0.02	0.06	0.08	0.06	0.02
0.01	0.02	0.03	0.02	0.01

Kernel (PSF)

=



Observed Image

Deconvolution is to restore the "real" image from the possibly **band-limited, noisy and distorted** observation, provided some prior information.

Conventional Methods to Solve Deconvolution

- Bayesian Inference Methods, e.g., MAP:

$$\min_{\{\mathbf{x}\}} \quad \underbrace{\left(\mathbf{y} \otimes k - \mathbf{x} - \mathbf{x} \cdot \log \frac{\mathbf{y} \otimes k}{\mathbf{x}}\right)}_{\text{Poisson Distribution}} + \underbrace{\lambda \cdot \Psi(\bigtriangledown \mathbf{x})}_{\text{Prior/Regularization}}$$

- Optimization Methods, e.g., Alternating Direction Method of Multipliers (ADMM):

$$\min_{\{\mathbf{x}\}} \quad \underbrace{\frac{1}{2}\|\mathbf{x} \otimes k - \mathbf{y}\|_2^2}_{\mathrm{f(x)}} + \underbrace{\lambda\|\mathbf{z}\|_1}_{\mathrm{g(z)}}$$

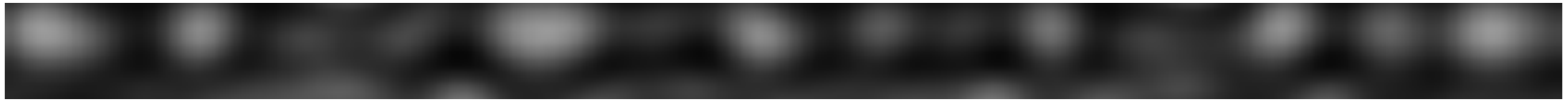
$$s.t. \quad \bigtriangledown \mathbf{x} = \mathbf{z}$$

$$\min_{\{\mathbf{x},\mathbf{z},\beta\}} L_\rho(\mathbf{x}, \mathbf{z}, \beta) = f(\mathbf{x}) + g(\mathbf{z}) + \frac{\rho}{2}\|\mathbf{x} - \mathbf{z}\|_2^2 - \beta^T(\mathbf{x} - \mathbf{z})$$

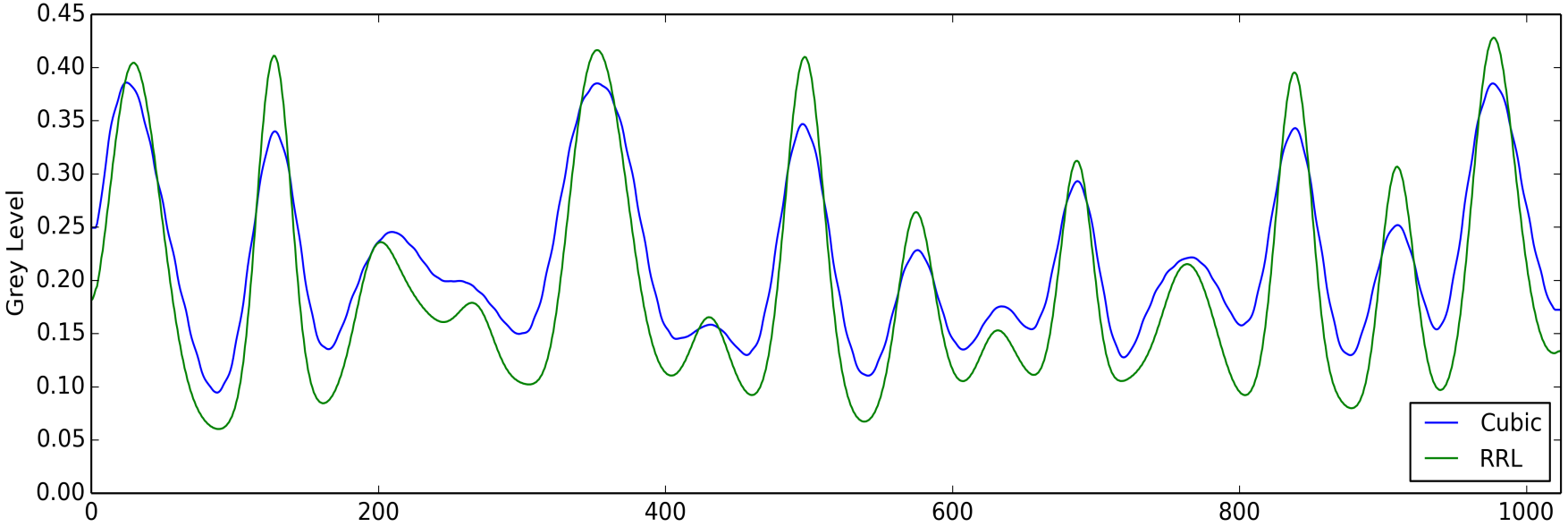
[R. Wang and D. Tao, Recent Progress in Image Deblurring. arXiv:1408.6838, 2014.]

Deconvolution – Beyond Resolution Limit

- Original:



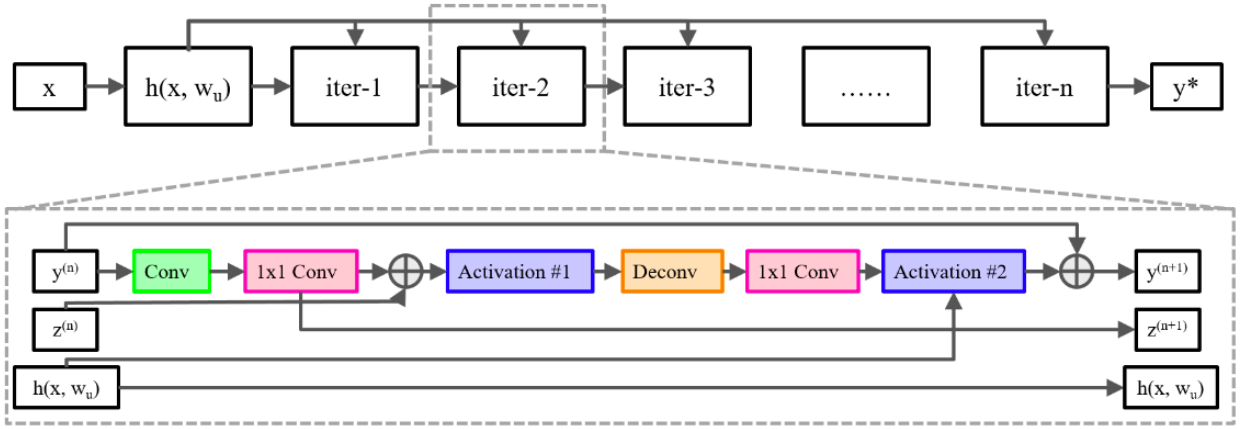
- Deconvolution:



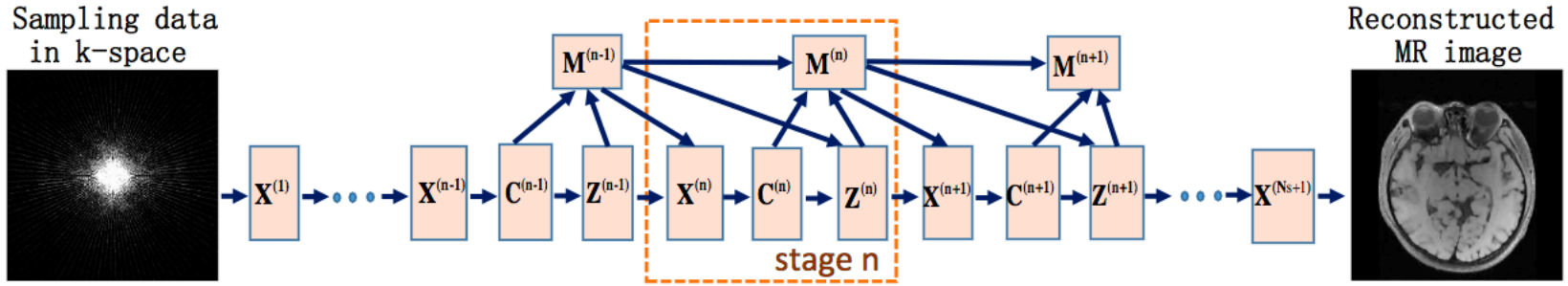
KLA Tencor

Can we mapping the iterative algorithm to Neural Network?

- S. Wang, S. Fidler, R. Urtasun, Proximal Deep Structured Models, ***NIPS*** 2016.



- J. Sun, H. Li, Z. Xu, Deep ADMM-Net for compressive sensing MRI, ***NIPS*** 2016.



ADMM Network = Layer mapping + Loop unrolling

2016

$$\mathbf{x}^{(n+1)} = \mathbb{F}^{-1}\left\{\frac{\mathbb{F}\left\{\mathbf{k}^T \otimes \mathbf{y} + \frac{\mathbf{x}^{(n)}}{2\mu} - \frac{\rho}{2}\left(\sum_i W_i^T \otimes \left(W_i \otimes \mathbf{x}^{(n)} - (\mathbf{z}_i^{(n)} - \mathbf{u}_i^{(n)})\right)\right)\right\}}{\mathbb{F}\left\{\mathbf{k}^T \otimes \mathbf{k} + \frac{1}{2\mu}\right\}}\right\}$$

$$\mathbf{z}_i^{(n+1)} = S_{\lambda/\rho}\left(W_i \otimes \mathbf{x}^{(n+1)} + \mathbf{u}_i^{(n)}\right)$$

$$\mathbf{u}_i^{(n+1)} = \mathbf{u}_i^{(n)} + \left(W_i \otimes \mathbf{x}^{(n+1)} - \mathbf{z}_i^{(n+1)}\right)$$

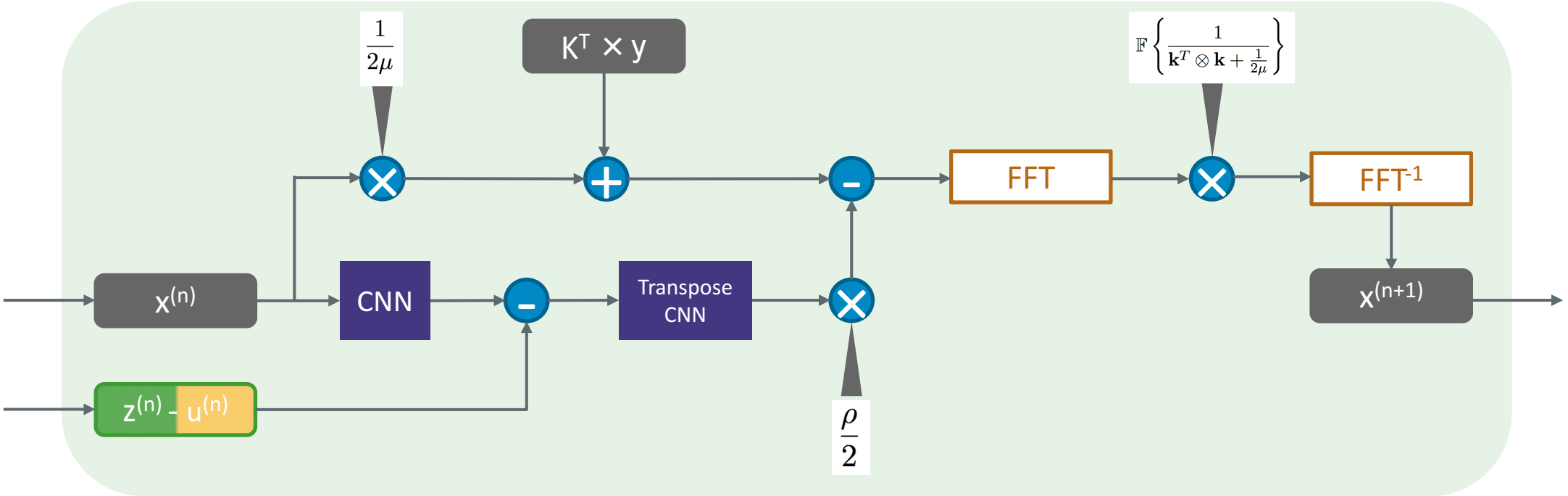
ADMM Network = Layer mapping + Loop unrolling

2016

$$\mathbf{x}^{(n+1)} = \mathbb{F}^{-1}\left\{\frac{\mathbb{F}\left\{\mathbf{k}^T \otimes \mathbf{y} + \frac{\mathbf{x}^{(n)}}{2\mu} - \frac{\rho}{2}\left(\sum_i W_i^T \otimes \left(W_i \otimes \mathbf{x}^{(n)} - (\mathbf{z}_i^{(n)} - \mathbf{u}_i^{(n)})\right)\right)\right\}}{\mathbb{F}\left\{\mathbf{k}^T \otimes \mathbf{k} + \frac{1}{2\mu}\right\}}\right\}$$

$$\mathbf{z}_i^{(n+1)} = S_{\lambda/\rho}\left(W_i \otimes \mathbf{x}^{(n+1)} + \mathbf{u}_i^{(n)}\right)$$

$$\mathbf{u}_i^{(n+1)} = \mathbf{u}_i^{(n)} + \left(W_i \otimes \mathbf{x}^{(n+1)} - \mathbf{z}_i^{(n+1)}\right)$$



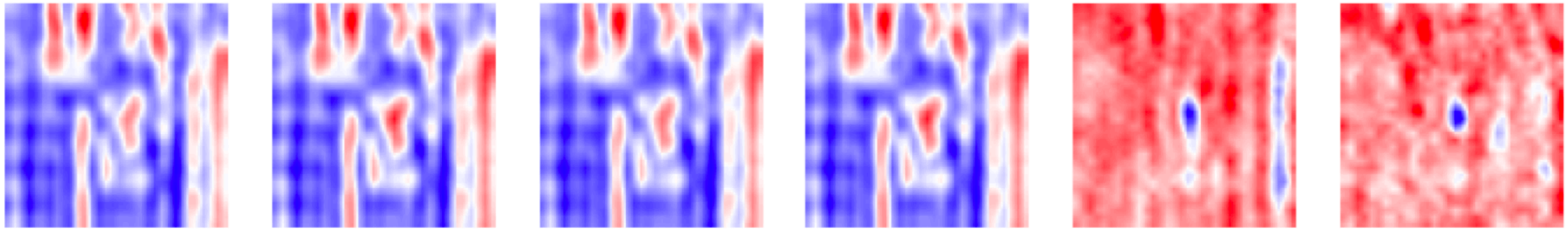
KLA Tencor

ADMM Network Performance

GUID=22608

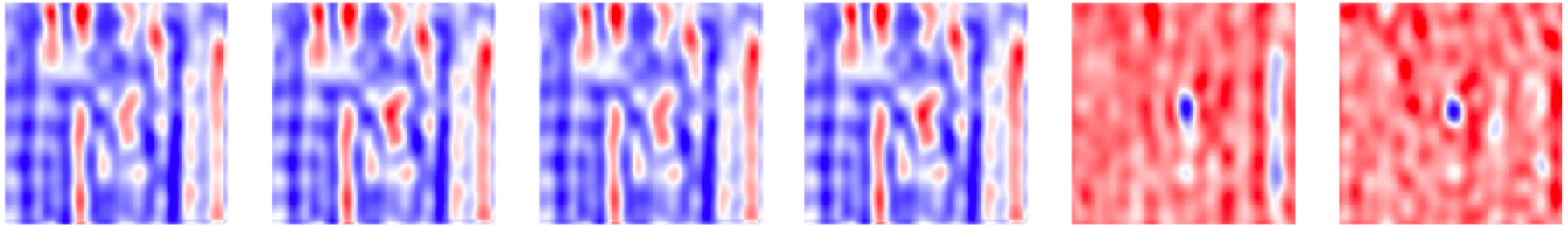
Ref 1 | Def 1 | Ref 2 | Def 2 | Difference 1 SNR = 15.9 db | Difference 2 SNR = 9.7 db

Original:



Difference 1 SNR = 16.6 db | Difference 2 SNR = 15.4 db

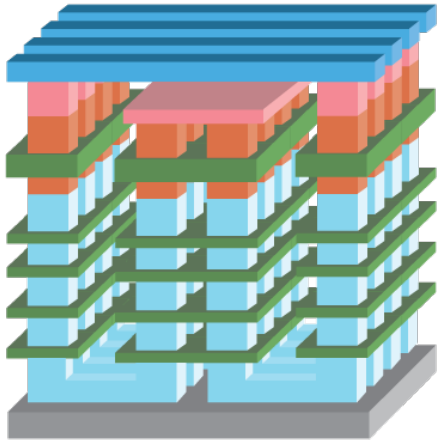
ADMM Network:



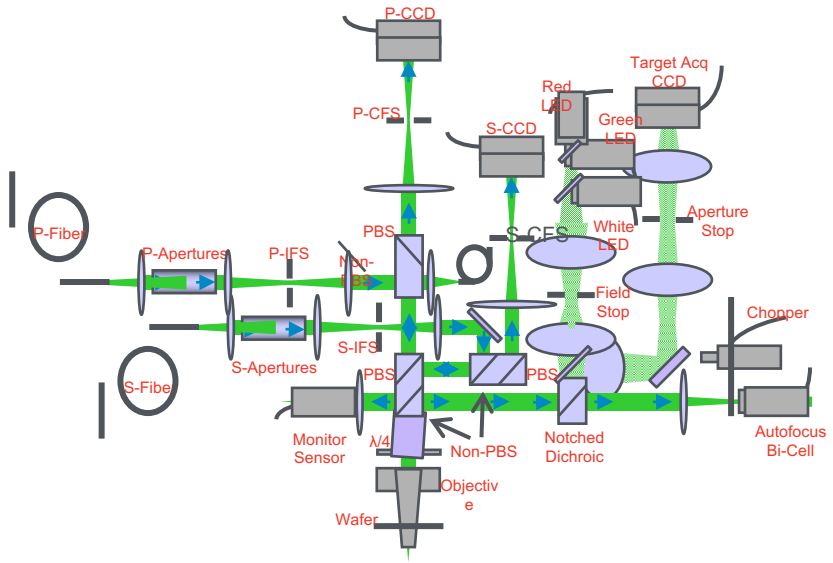
Practical and Theoretical Limitation

- If we don't have a explicit solution from first principle, it very difficult to do "direct mapping".
  - We can not solve it, if we ***don't fully understand*** the physical process…
  - Sometimes, an approximated solution can be ***complex*** also.

3D Memory



Optics



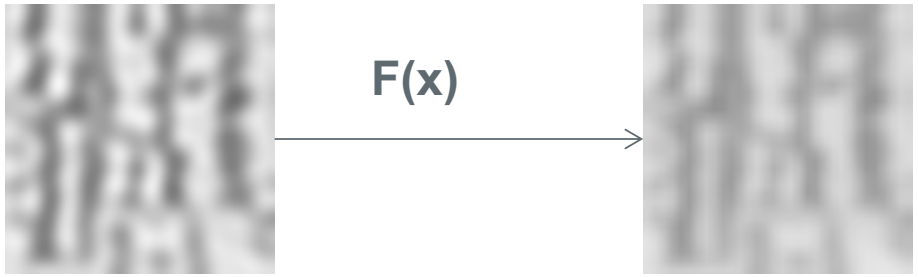
KLA Tencor

Optical Neural Network

Optical Neural Network

2016

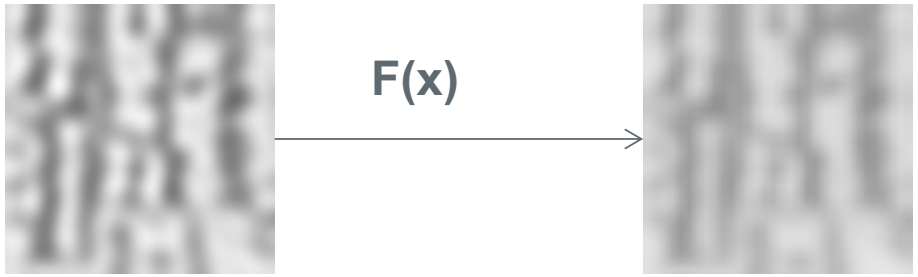
- Physical process can be modeled from first principle theories, e.g., Fourier Optics, Maxwell Equation, etc.



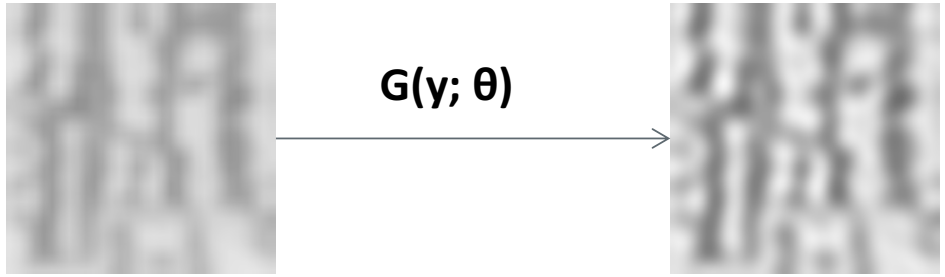
Optical Neural Network

2016

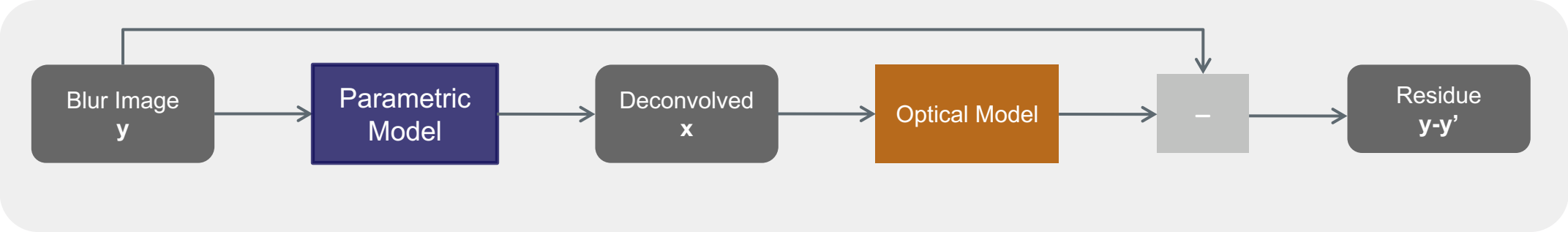
- Physical process can be modeled from first principle theories, e.g., Fourier Optics, Maxwell Equation, etc.



- The inverse process is approximated via a parametric model, e.g., Neural Network.



Optical Neural Network

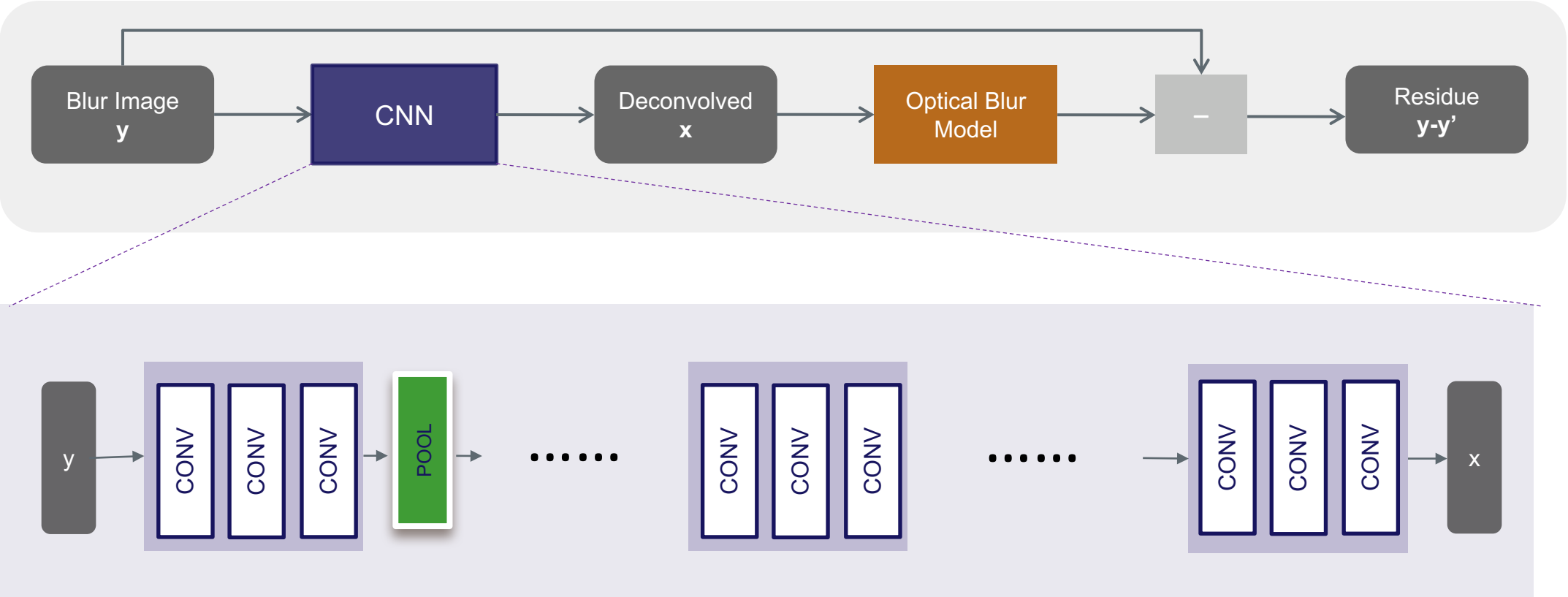


- The learning of such inverse model can be performed via minimizing the loss function:

$$\arg\min_{\theta} \quad \frac{1}{2}\|\mathbb{F}(\mathbb{G}(\mathbf{y};\theta)) - \mathbf{y}\|_2^2$$

[Patent Applied for in the US and abroad.]

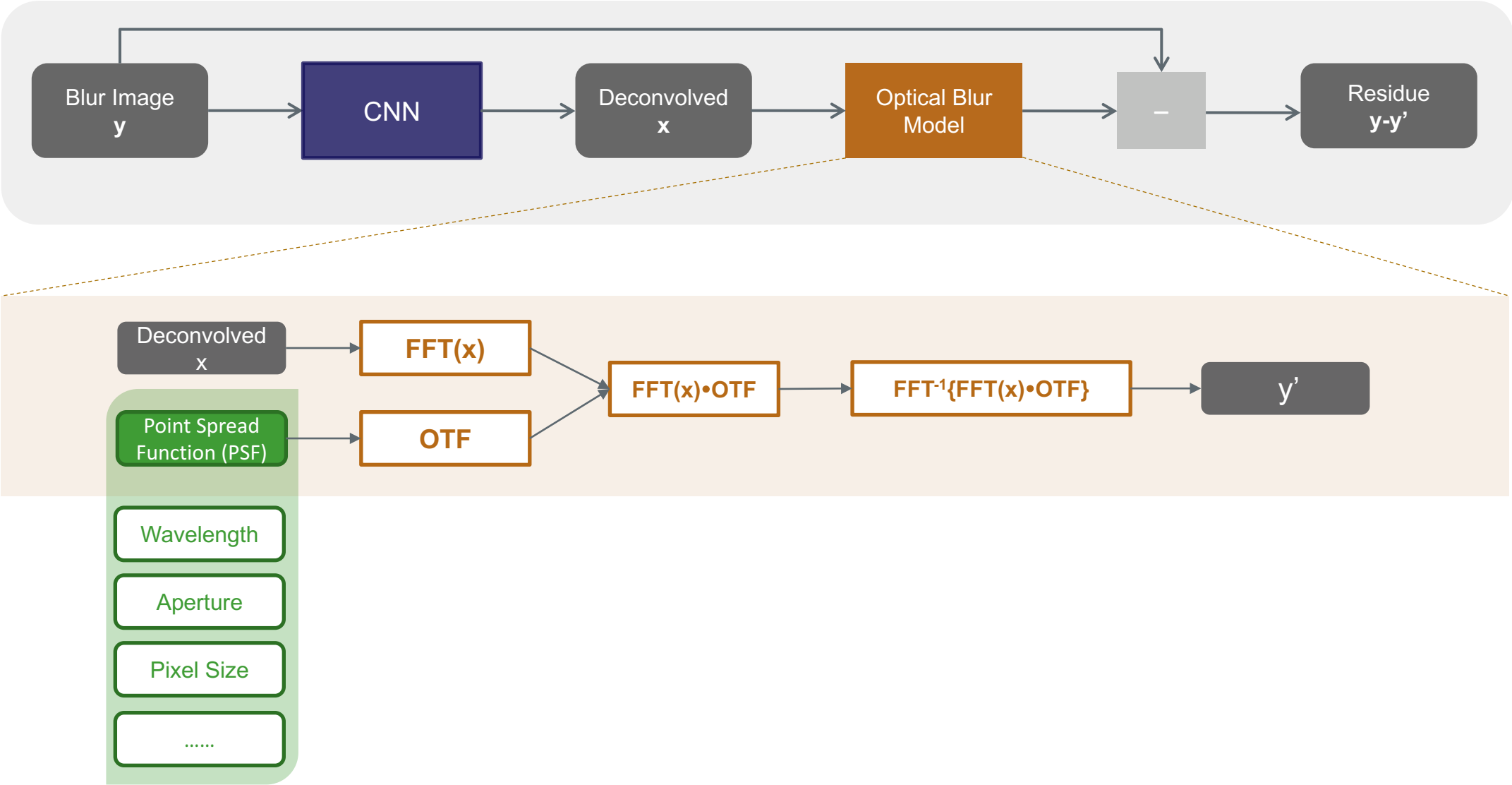
Parametric Model as the Inverse Method



- Parametric Inverse Model G(y; θ)
  - Convolution Neural Network
  - Domain Knowledge

KLA Tencor

Differentiable Optical Model as the Forward Method



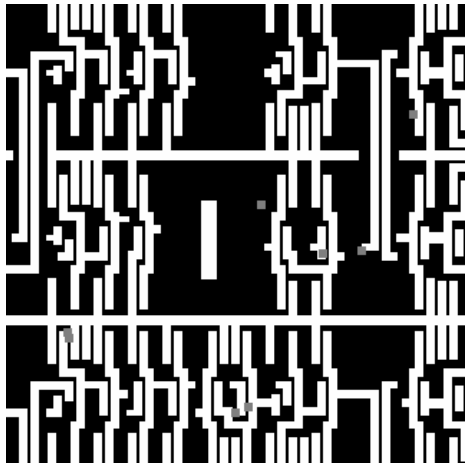
KLA Tencor

ONN – Synthetic Experiment

Simulated Image

ONN Deconvolved

CAD Design

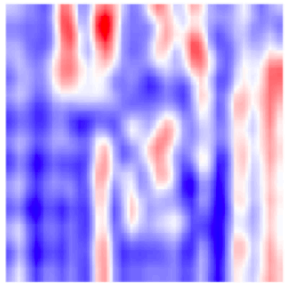
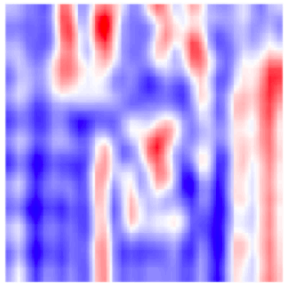
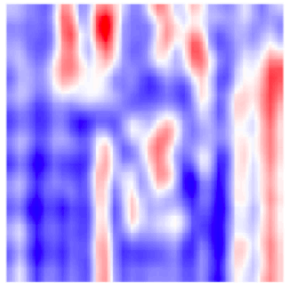
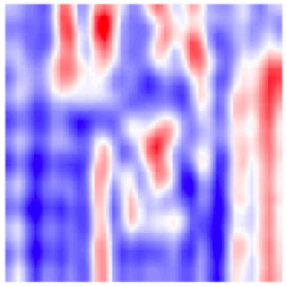
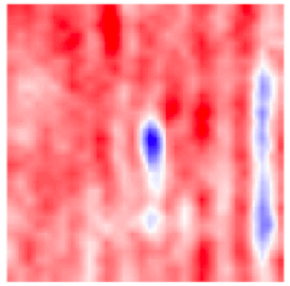
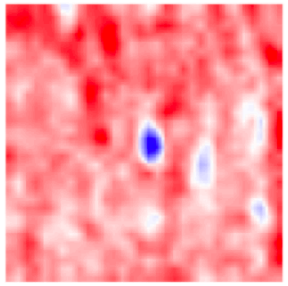
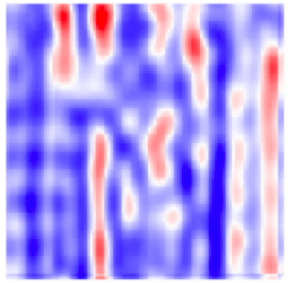
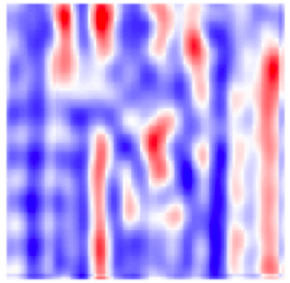
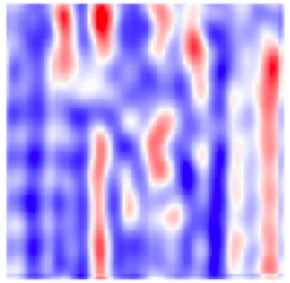
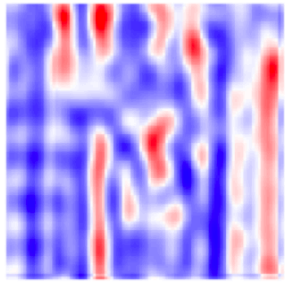
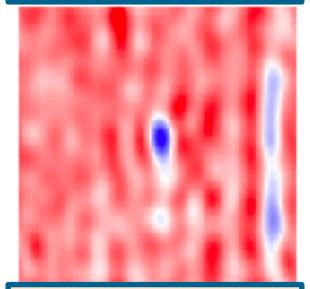
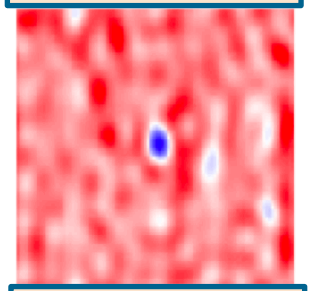
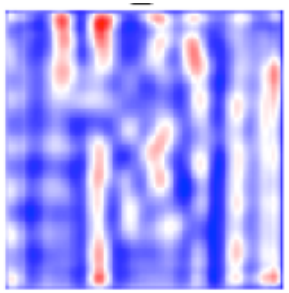
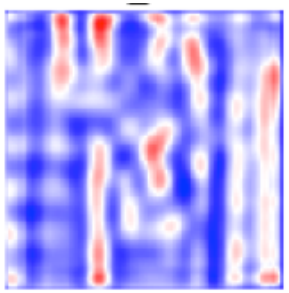
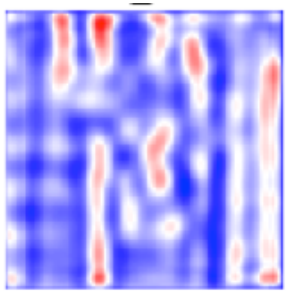
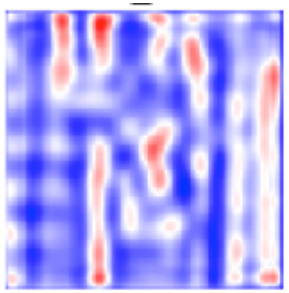
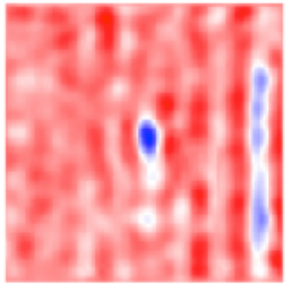
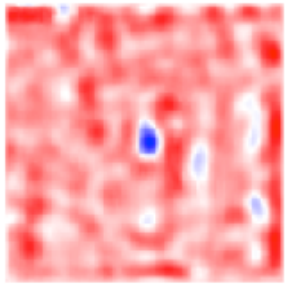


Lucy-Richardson

[Synthetic image is generated from incoherent simulation based on known PSF and shot noise to added to simulate the sensor noise.]

ONN – Real Experiment

GUID=22608

	Ref 1	Def 1	Ref 2	Def 2	Difference 1 SNR = 15.9 db	Difference 2 SNR = 9.7 db
Original:						
ADMM Network:					SNR = 16.6 db 	SNR = 15.4 db 
ONN :					SNR = 16.2 db 	SNR = 14.9 db 

Performance

	PSNR 1 [db]	PSNR 2 [db]	PSNR Difference	Throughput [ms]
Original	14.1 db	14.0 db	2.8 db	-
ADMM Network	15.6 db **(+1.5 db)**	15.6 db **(+1.6 db)**	**1.7 db**	0.77 ms
Optical Neural Network	15.4 db **(+1.3 db)**	15.4 db **(+1.4 db)**	**1.7 db**	0.78 ms

- For both ADMM and ONN, PSNR is improved by +1.4 db on average, and up to +6 db; equivalently the peak/std is improved by +20%.
- PSNR difference is reduced by 1.1 db, which suggests the noise discrepancy is improved.

Benchmark is performed as follow: 32x32 image patch as input, 128x128 image patch as output; Pascal Titan X, Tensorflow, CUDNN 5.1.5 are used on experiment machine. ADMM network is unrolled to have fixed 15 iterations.

Summary

- Is it possible to combine a physical modeling with Machine Learning?
  - Yes, e.g., Optical Neural Network.
- How does the combined model meet robustness requirements?
  - Yes, with enough constraints.
- How does the combined model deal with limited data?
  - Yes, the forward physical model helps.
- How to train the combined model in unsupervised manner when we don't have the ground truth?
  - Yes, ONN is unsupervised.
- Is computation cost of the combined model economic?
  - Better.

Summary

- Is it possible to combine a physical modeling with Machine Learning?
  - Yes, e.g., Optical Neural Network.
- How does the combined model meet robustness requirements?
  - Yes, with enough constraints.
- How does the combined model deal with limited data?
  - Yes, the forward physical model helps.
- How to train the combined model in unsupervised manner when we don't have the ground truth?
  - Yes, ONN is unsupervised.
- Is computation cost of the combined model economic?
  - Better.

Thanks to Nvidia for enabling Deep Learning!!!

Thank you!

KLA Tencor

Backup Slides