Synergy of physics and learning-based models in computational imaging and display

Zihao Wang

PhD defense, July 23, 2020

Committee:

Prof. Oliver Cossairt (advisor),

Prof. Aggelos Katsaggelos,

Prof. Jack Tumblin,

Dr. Nathan Matsuda

Computational imaging

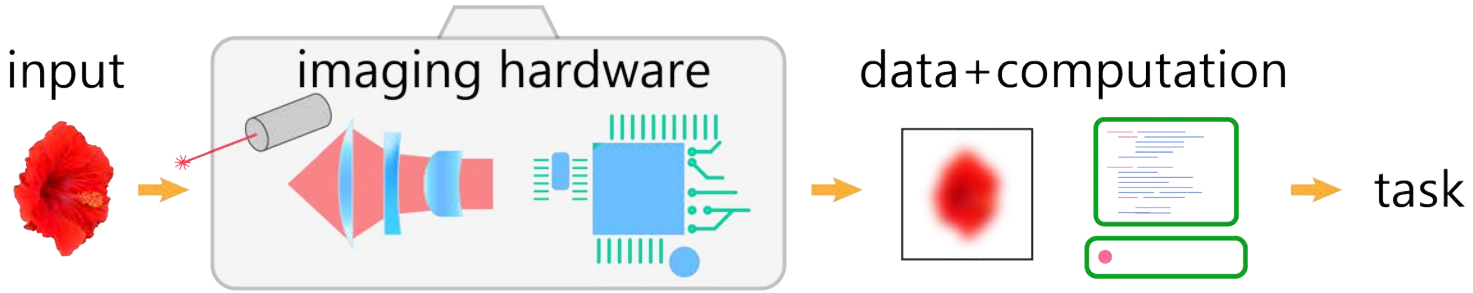
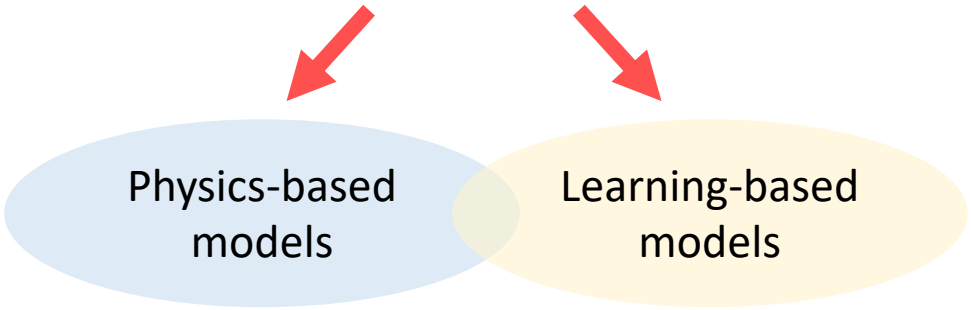


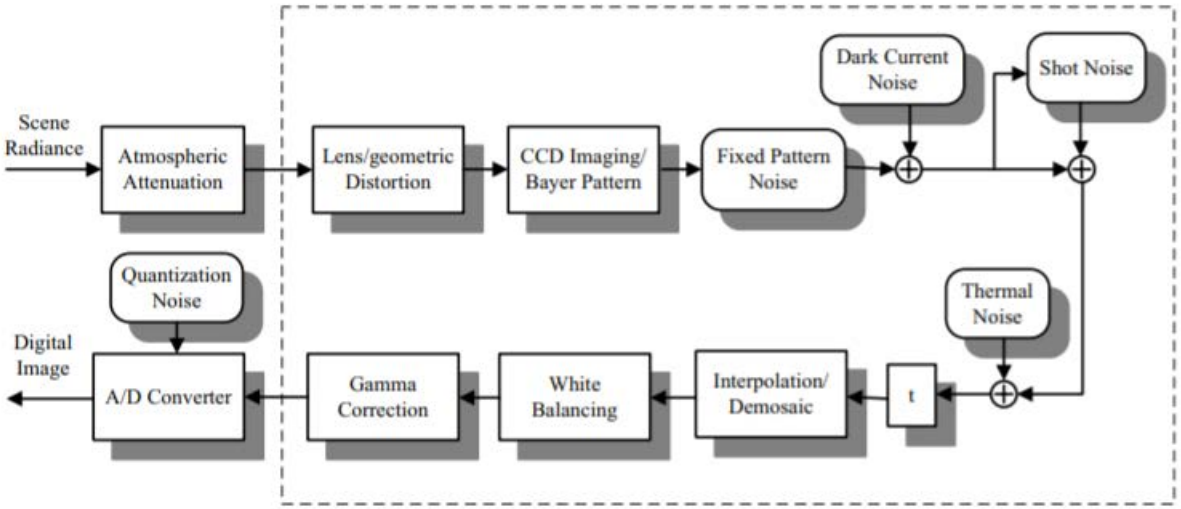
Image formation models



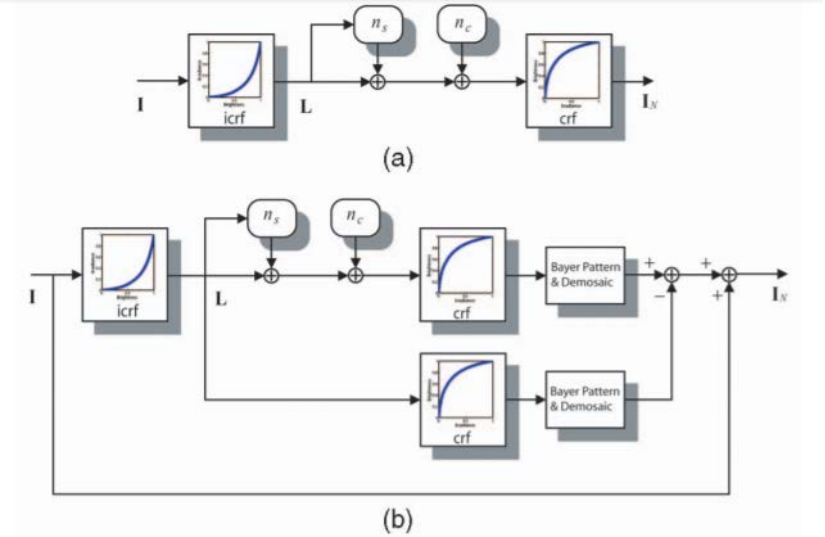
Example: image denoising

Liu et al. CVPR'06, PAMI'07

In-camera ISP pipeline



Noise modeling



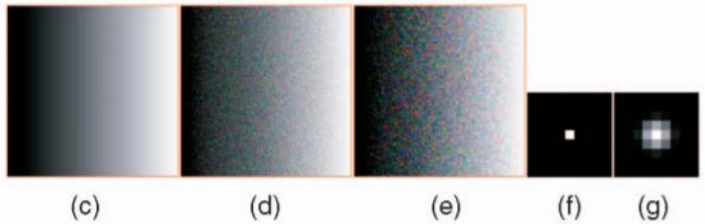


Fig. 4. Block diagrams showing noise simulations for color camera images. (a) shows independent white noise synthesis and (b) adds CCD color filter pattern sensing and demosaicing to model spatial correlations in the camera noise [28]. (c) Test pattern. (d) and (e) The synthesized images of (a) and (b). (f) and (g) The corresponding autocorrelation.

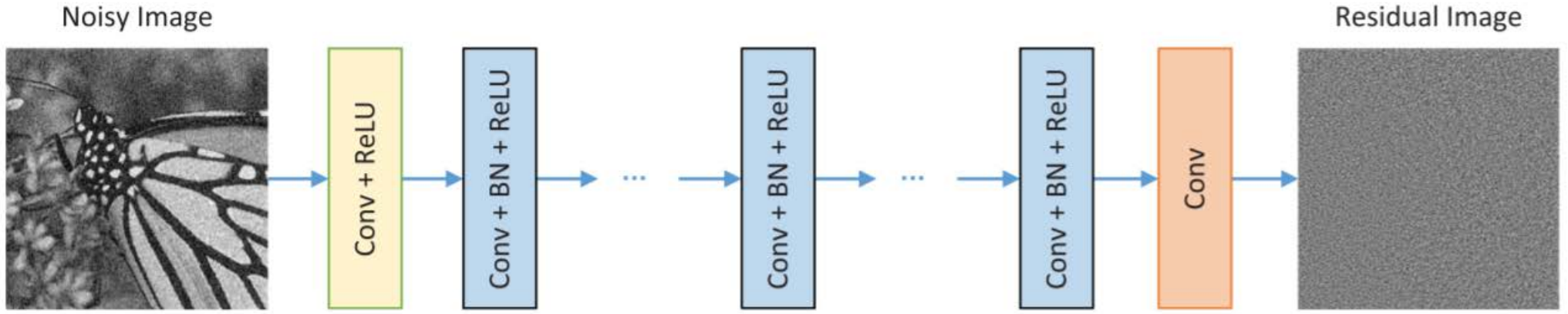
noisy image

denoised image

Example: image denoising (10 years later)

Zhang et al. TIP'17

DnCNN



noisy image | denoised image | noisy image | denoised image

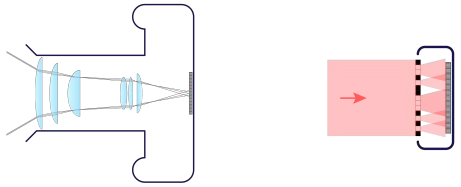
Non-universal definition

Physics-based models

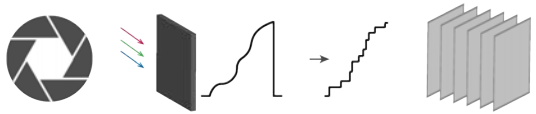
Analytical models from the image formation (hardware)
few parameters

Models used in this talk

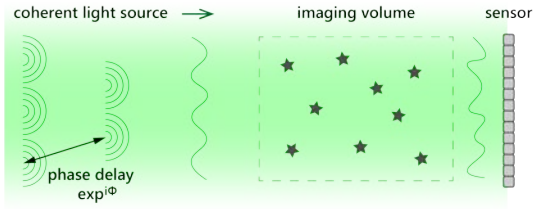
Optical elements



Sensing pipeline



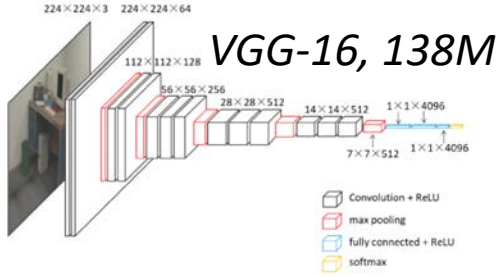
Light transport



Learning-based models

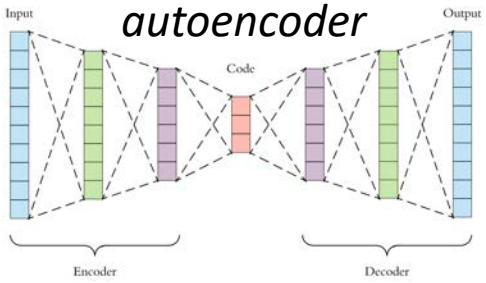
Statistical priors from lots of data
millions of parameters

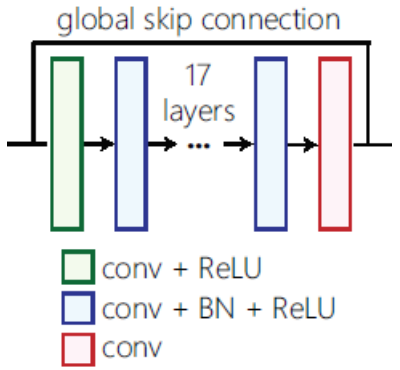
Models used in this talk (conv. nets)



VGG-16, 138M

autoencoder

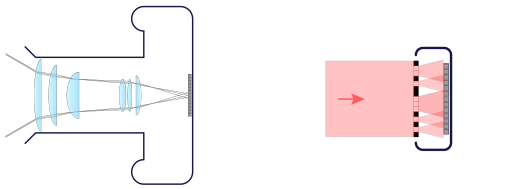


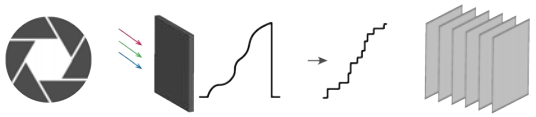


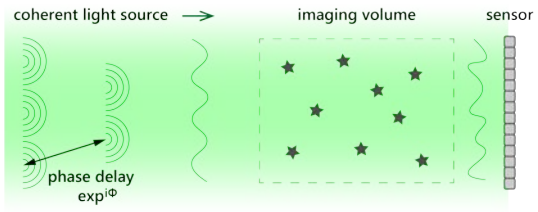
ResNet block, 0.6M

Overview

Physics-based models







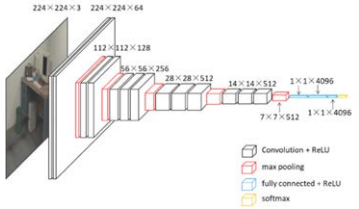
Synergistic models

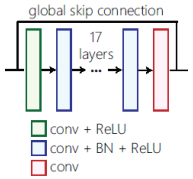
1. Privacy preserving action recognition, *CVPRW'19*

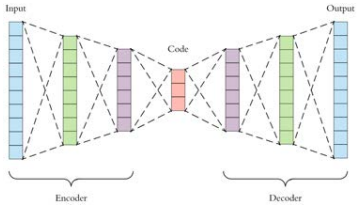
2. High frame-rate video frame synthesis, *ICCVW'19*

3. 3D holographic display design, image compression

Learning-based models







Why are learning-based models not enough?

Imaging hardware

Datasets + neural networks

task

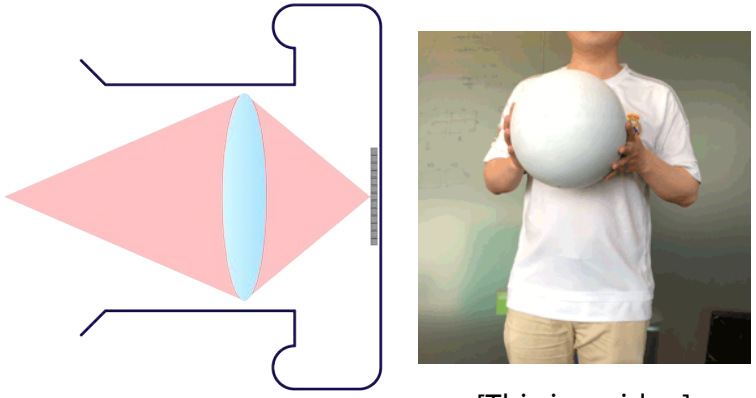
Synergistic models

Image formation

Neural networks

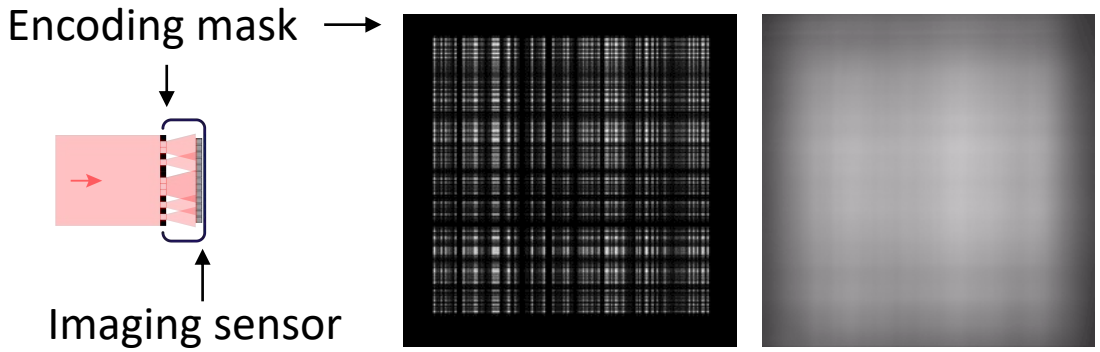
Pre-capture privacy preservation

Lens-based cameras



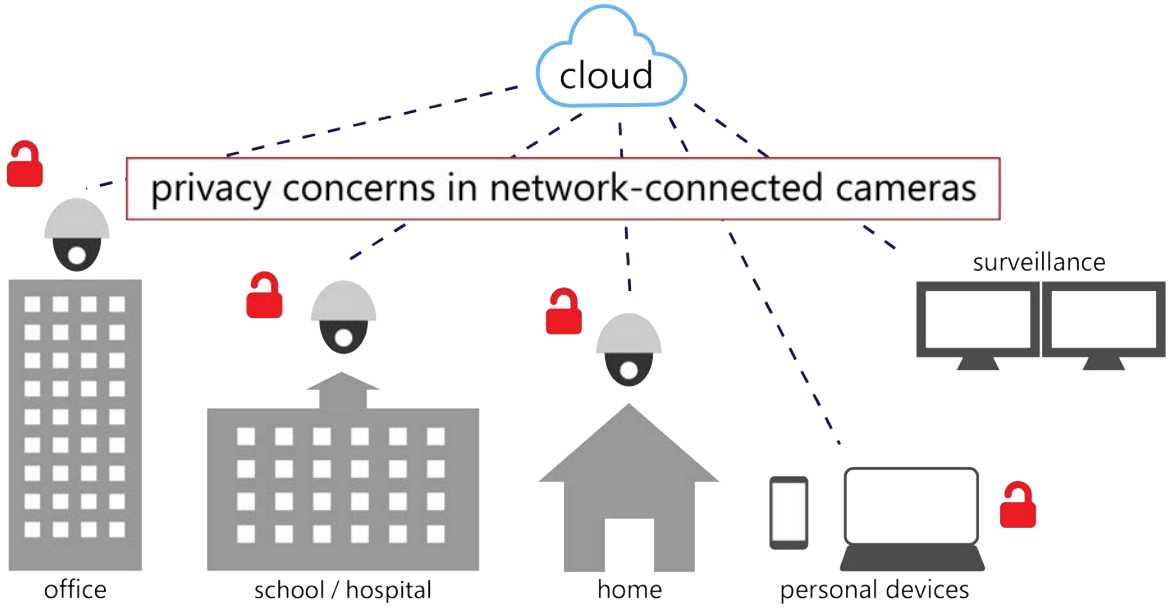
[This is a video]

Lens-free coded aperture cameras



[This is a video]

Restoration can be done with mask info., not without masks.



Privacy preserving action recognition:
Mask-free & restoration-free

Privacy preserving action recognition

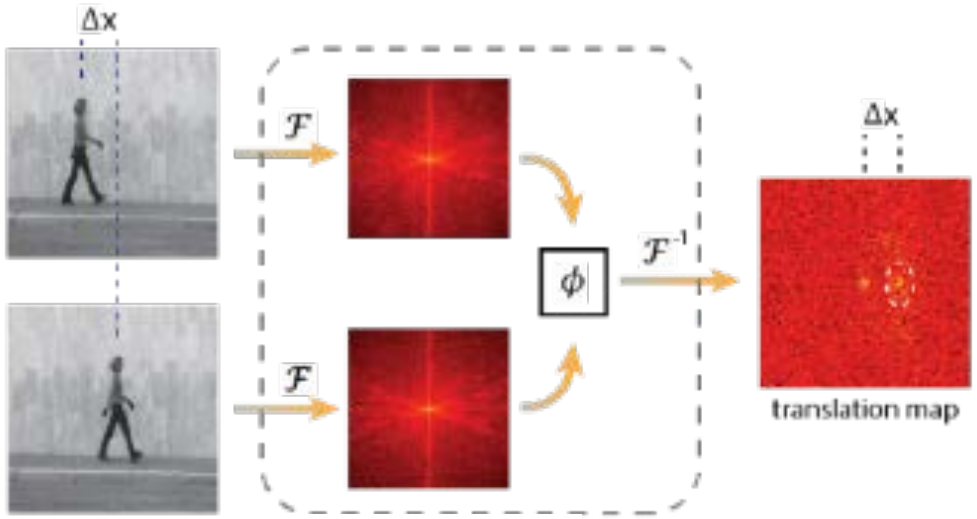
	(%)	Training	Validation
Overfitting!	CA	50th: 79.06	50th: 63.21
Info. is still there!	grayscale	50th: 99.56	50th: 94.39

Need to remove the mask effect
Design mask-independent features

Task: 5-class classification. (RGB2gray *vs.* synthetic CA images)
"writing on board", "Wall pushups", "blowing candles", "pushups", "mopping floor" from UCF-101
3 frames for each sample
Classifier: VGG-16

Mask-independent motion features

Translation feature

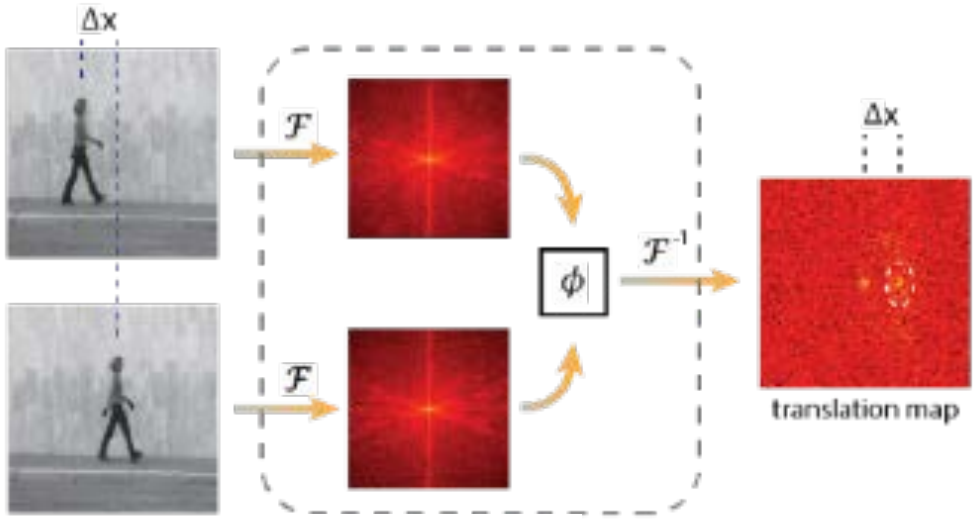


phase correlation

- $o_2(\mathbf{p}) = o_1(\mathbf{p} + \Delta\mathbf{p})$
- $\mathrm{O}_2(\boldsymbol{\nu}) = \phi(\Delta\mathbf{p})\mathrm{O}_1(\boldsymbol{\nu})$ Fourier transform
- $\mathrm{C}(\boldsymbol{\nu}) = \frac{\mathrm{O}_1 \cdot \mathrm{O}_2^*}{|\mathrm{O}_1 \cdot \mathrm{O}_2^*|} = \phi^* \frac{\mathrm{O}_1 \cdot \mathrm{O}_1^*}{|\mathrm{O}_1 \cdot \mathrm{O}_1^*|} = \phi(-\Delta\mathbf{p})$ Cross power spectrum
- $c(\mathbf{p}) = \delta(\mathbf{p} + \Delta\mathbf{p})$ Inverse Fourier transform

Mask-independent motion features

Translation feature



phase correlation

- $d_2(\mathbf{p}) = o_2(\mathbf{p}) * a = d_1(\mathbf{p} + \Delta\mathbf{p})$ a: coded aperture
- $\mathrm{D}_2(\boldsymbol{\nu}) = \mathrm{O}_2 \cdot \mathrm{A} = \phi \cdot \mathrm{O}_1 \cdot \mathrm{A}$ Fourier transform
- $\mathrm{C}_d(\boldsymbol{\nu}) = \frac{\mathrm{D}_1 \cdot \mathrm{D}_2^*}{|\mathrm{D}_1 \cdot \mathrm{D}_2^*|} = \phi^* \frac{\mathrm{O}_1 \cdot \mathrm{A} \cdot \mathrm{A}^* \cdot \mathrm{O}_1^*}{|\mathrm{O}_1 \cdot \mathrm{A} \cdot \mathrm{A}^* \cdot \mathrm{O}_1^*|} \approx \phi^*$ Cross power spectrum
- $c_d(\mathbf{p}) = \delta(\mathbf{p} + \Delta\mathbf{p})$ Inverse Fourier transform

Mask-independent

Mask-independent motion features

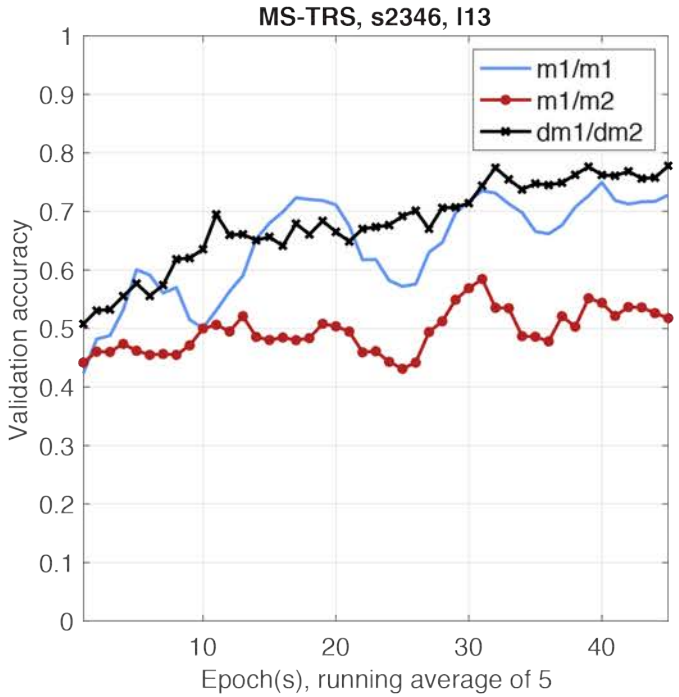
Translation features
phase correlation

Mask-independent

Rotation & Scale features
Fourier-Merlin transform (Cartesian to Polar)
+ phase correlation

Not mask-independent

Augmentation: change masks during training



Training w/ randomly varying masks
vs.
w/ single fixed mask, test w/ a different mask

More results are in the paper

Why synergistic models?

Imaging hardware

Synergistic models

task

Physics-based model

Learning-based model

#1 coded aperture camera

Deep classifier

Privacy preserving action recognition

Learning-based models alone do not perform well

Motion features

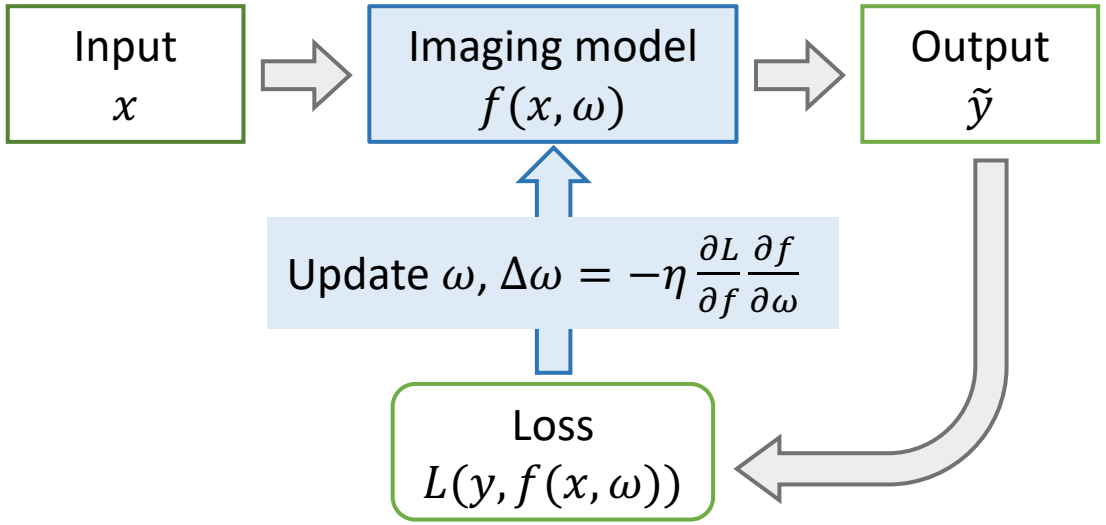
Physics-based models have limitation, e.g. mask-dependence

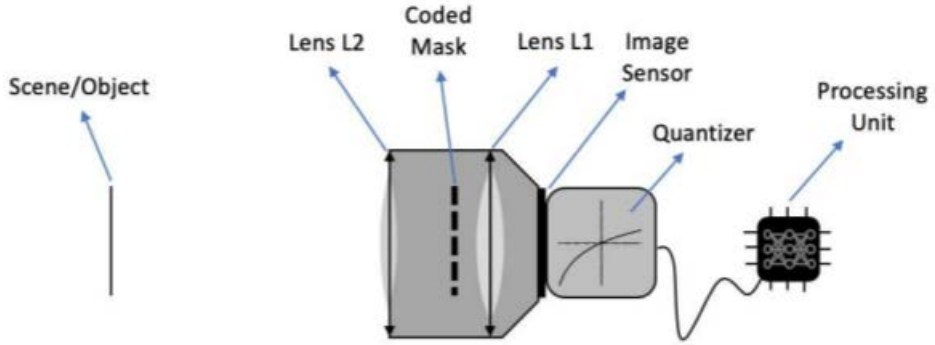
Motion features

Mask augmentation

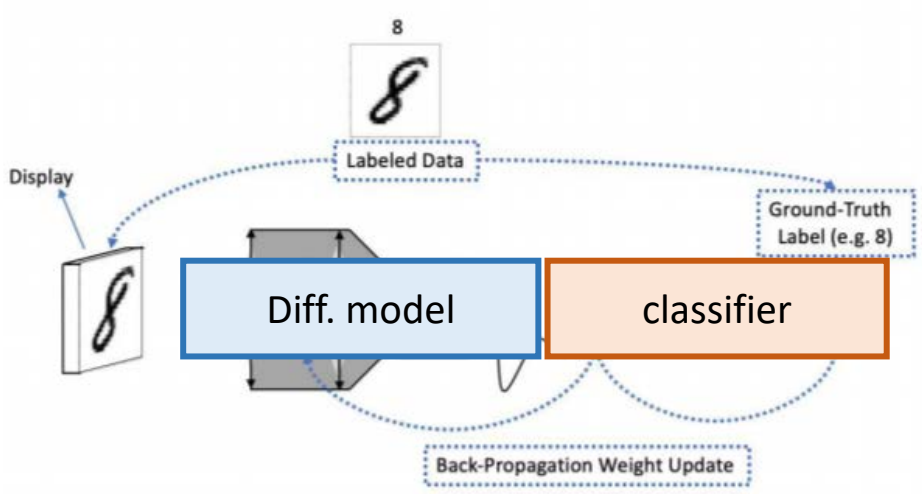
Deep classifier

Differentiable imaging model





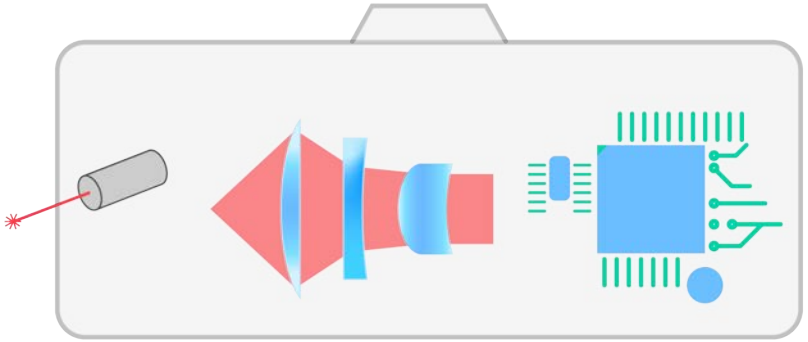
System setup



The coded mask is the first layer of the ConvNet

Task-specific differentiable imaging

Part of the learnable parameters ω has physical counterparts



LED sequence

Microscopy

Kellman et al. ICCP'19

ISP parameters

Object detection/segmentation

Mosleh et al. CVPR'20

SR &
extended DoF

Sitzmann et al. TOG'18

HDR

Metzler et al. CVPR'20

Sun et al. CVPR'20

Bayer pattern

Color imaging

Chakrabarti et al. NIPS'16

HDR &
compressive video

Martel et al. ICCP'20

phase plate

Depth estimation

Haim et al. TCI'18 *He et al. TIP'18*

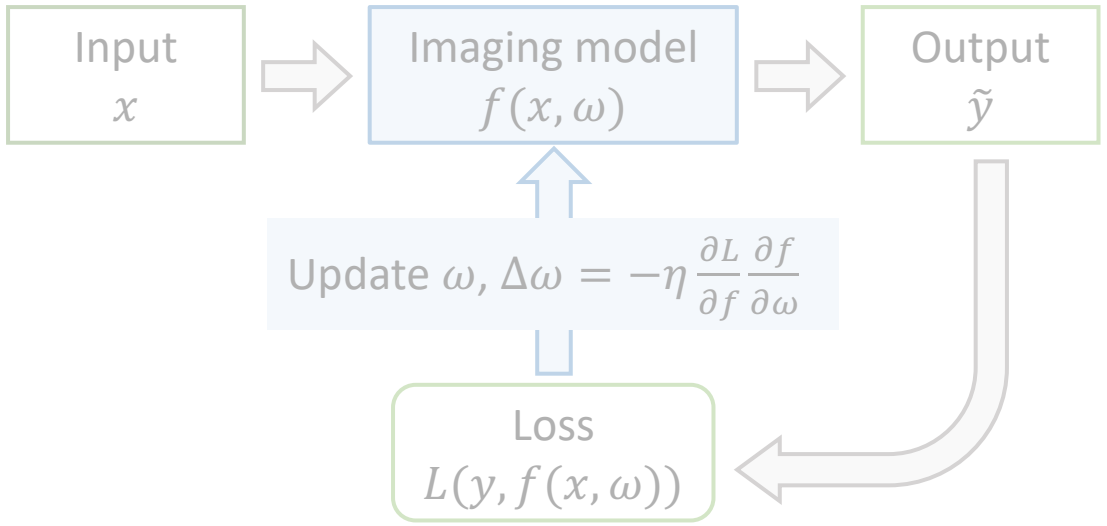
Chang et al. ICCV'19 *Wu et al. ICCP'19*

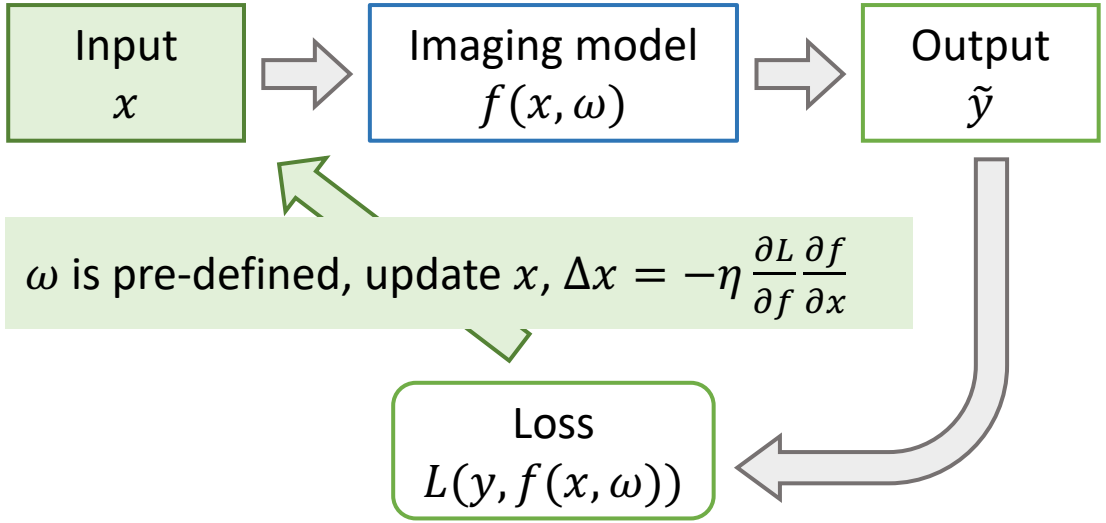
Exposure coding

Action recognition

Okawara et al. ICCP'20

Task-specific differentiable imaging





Differentiable inverse solver

Optical tomography

Kamilov et al. Optica'15

Neural 3D rendering

Kato et al. CVPR'18 …

High frame-rate video frame synthesis

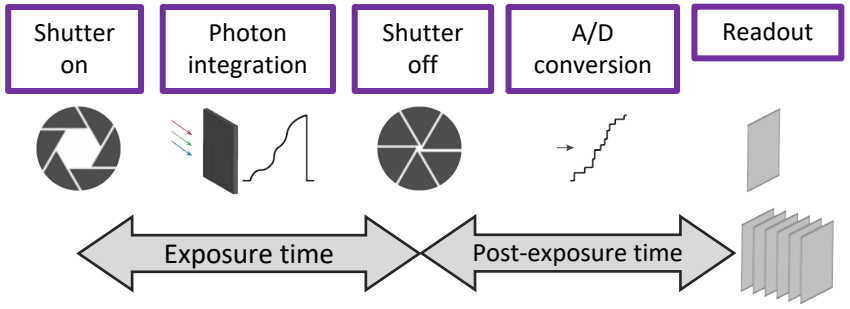
ICCVW'19

Holographic display

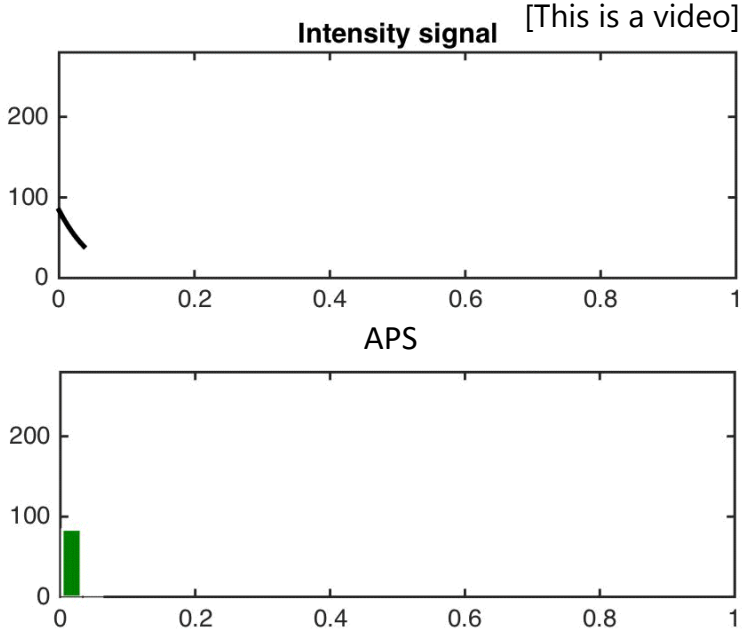
Chakravarthula et al. TOG'19

Background: what's making a camera slow?

Active pixel sensors (APS):
agnostic to scene motion



[This is a video]

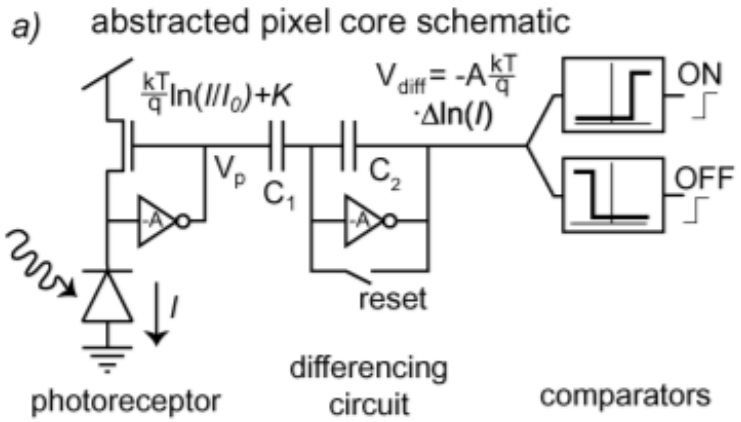


[This is a video]

Background: what's making a camera slow?

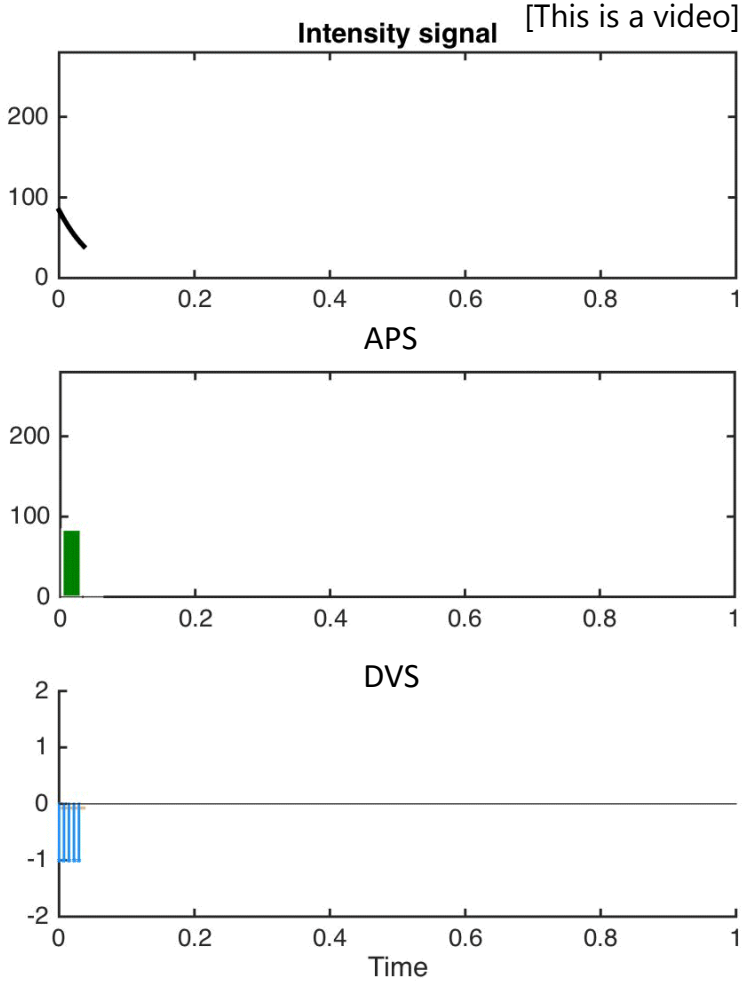
Active pixel sensors (APS):
agnostic to scene motion

Dynamic vision sensors (DVS):



Lichtsteiner et al. IEEE SSC 2008

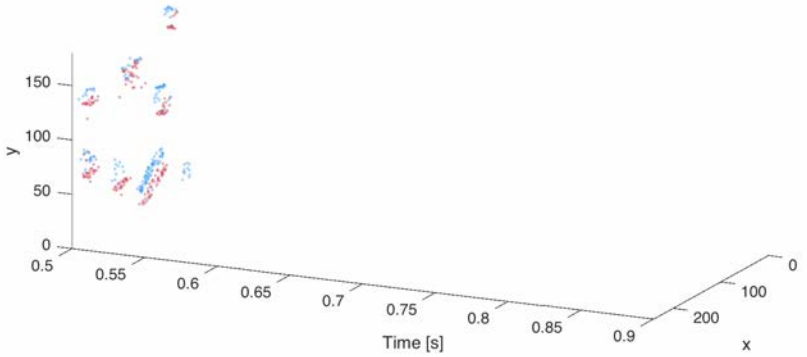
[This is a video]



Fast (~1us latency)

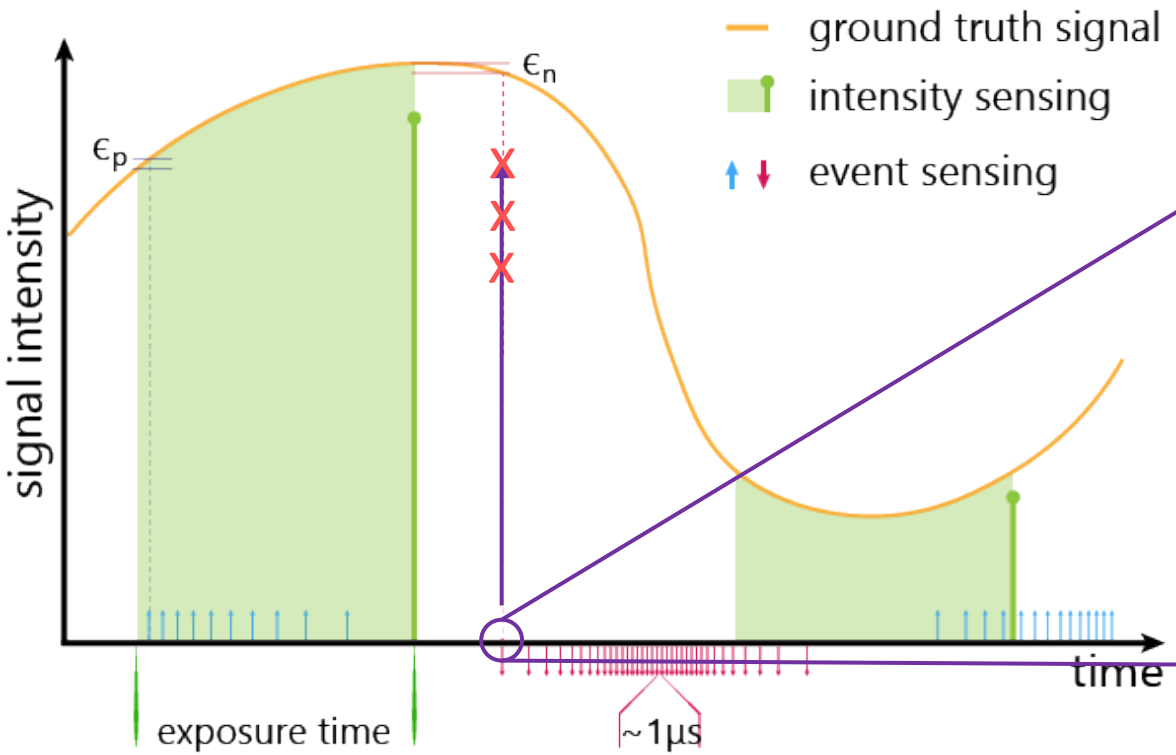
But only binary data; also noisy

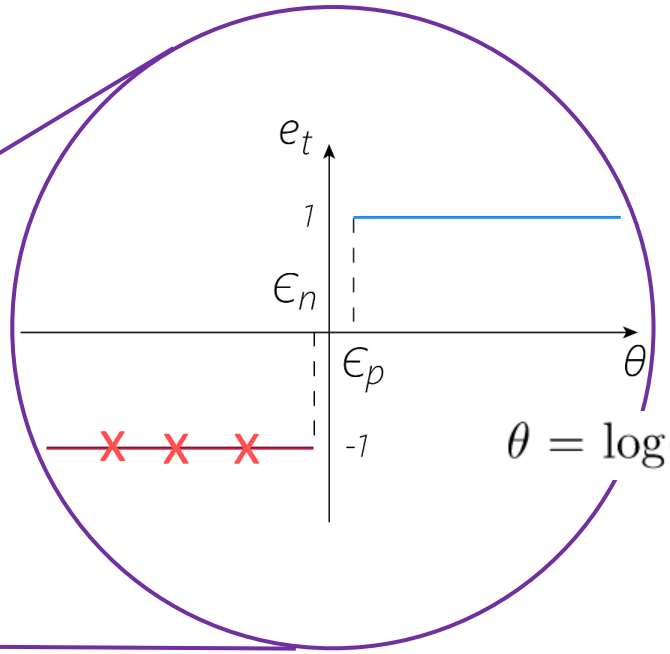
[This is a video]



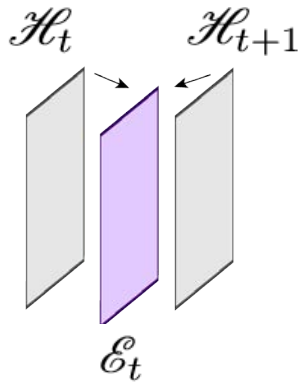
[This is a video]

Synthesizing dense frames from intensity (APS) and events (DVS)

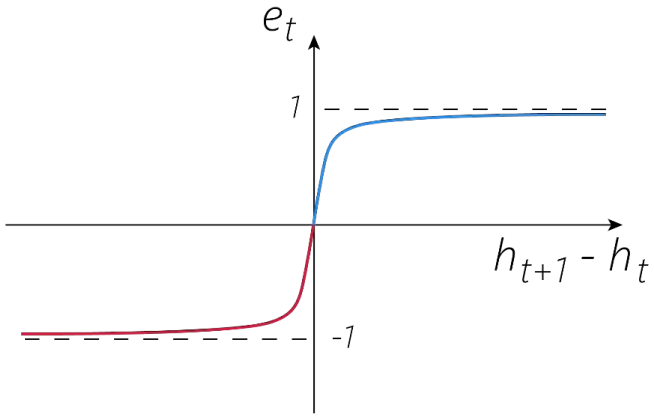




$$\theta = \log(I_t + b) - \log(I_0 + b)$$



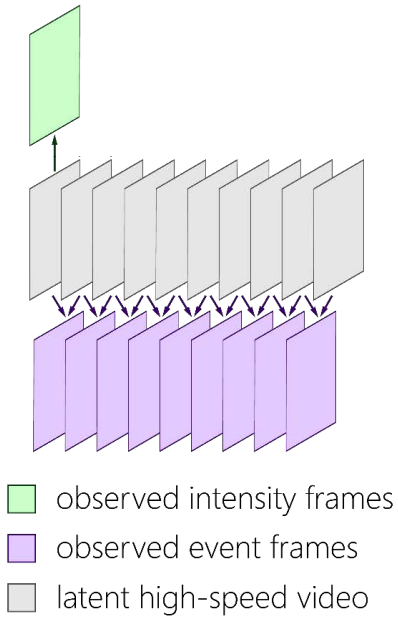
$$\mathscr{E}_t = \tanh\left\{\alpha[\mathscr{H}_{t+1} - \mathscr{H}_t]\right\}$$

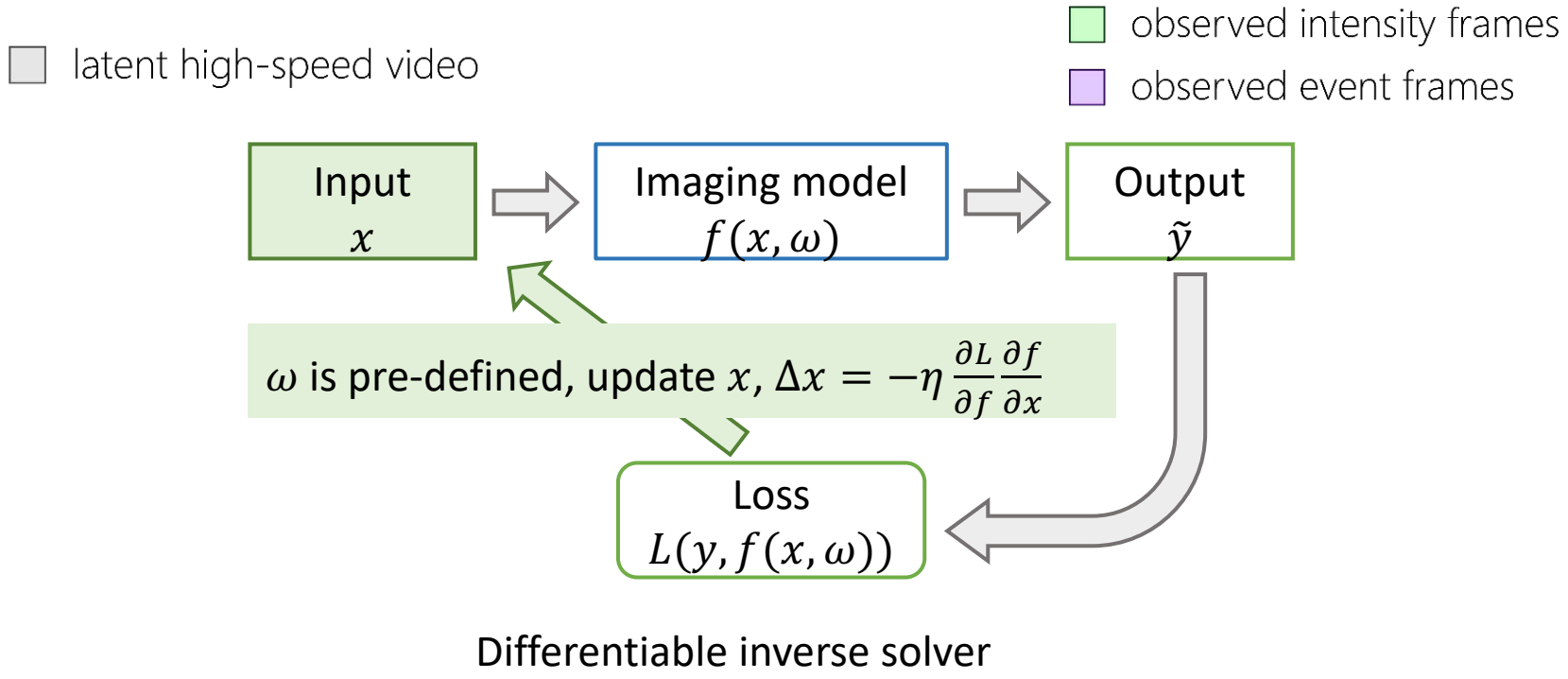


Approx. differentiable model

Differentiable inverse solver for video frame synthesis

Fusion setting:
Known first intensity frame
Known subsequent event frames

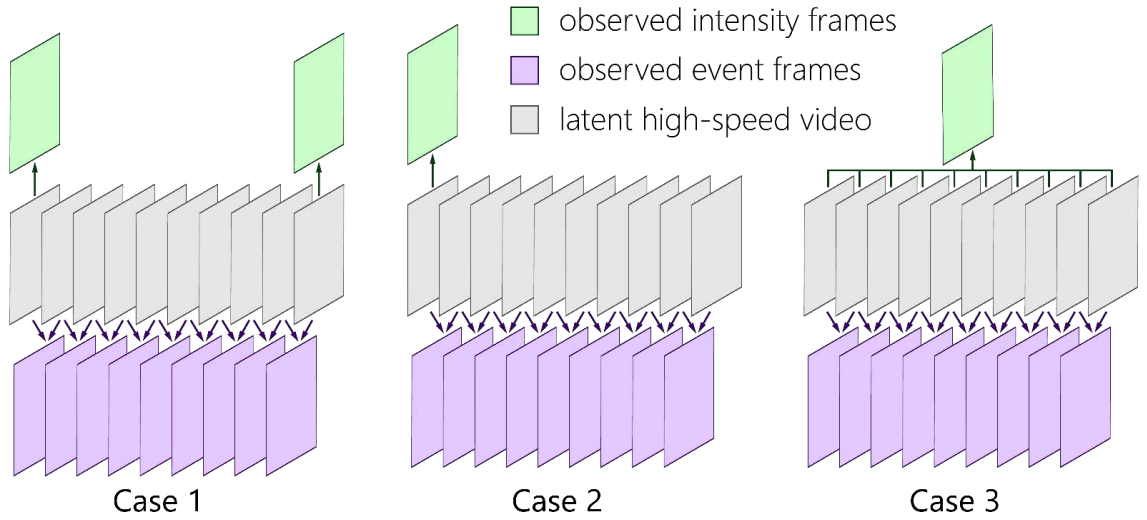




Differentiable inverse solver

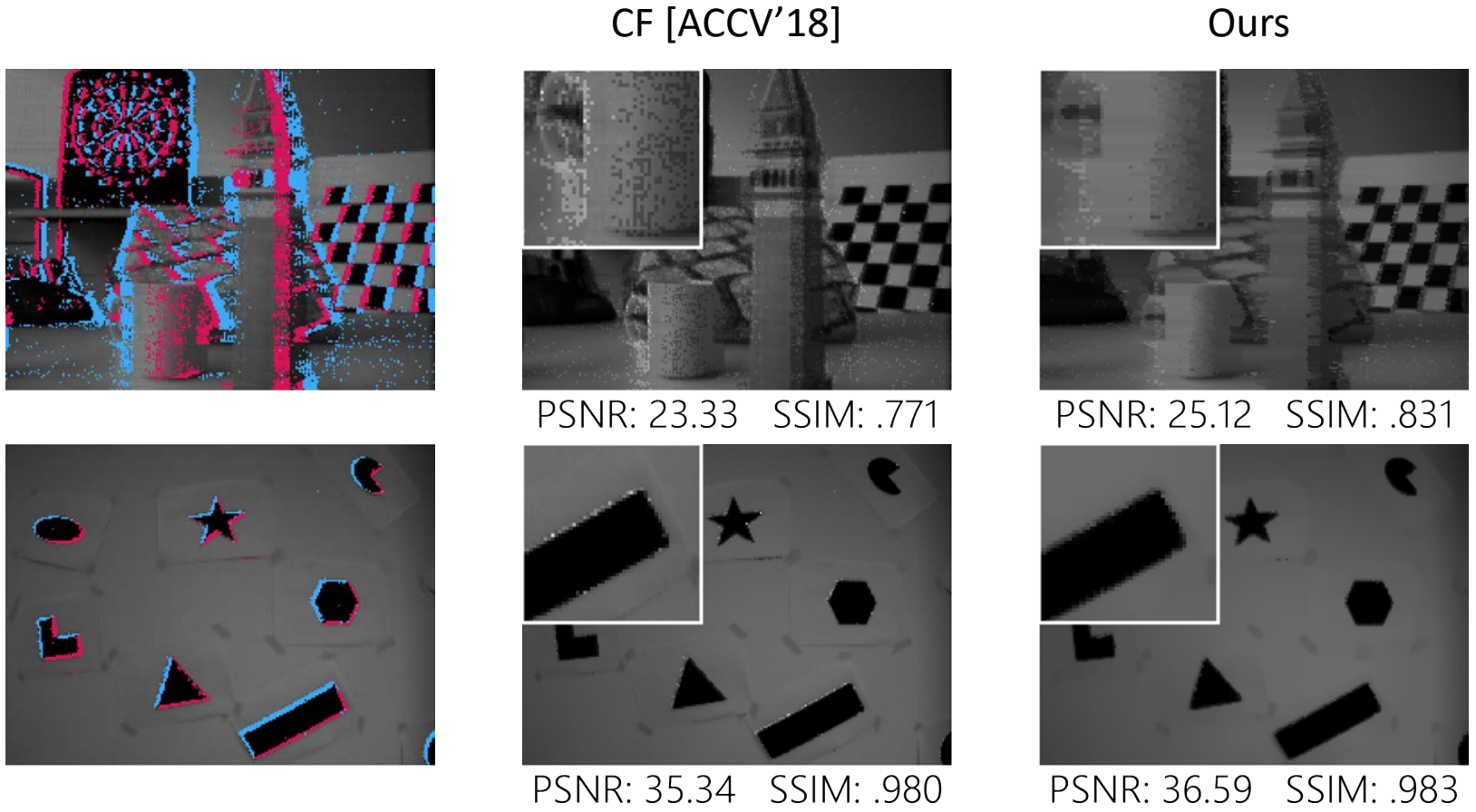
Variants of diff. inverse solvers

Differentiable inverse solvers are flexible to solve different fusion settings



Results

- Prediction case
  - Given start frame and future events, recover future frames



Results

- Interpolation case
  - Given start & end frames + events in-between, recover intermediate frames

Low-speed intensity frames
(2 frames)

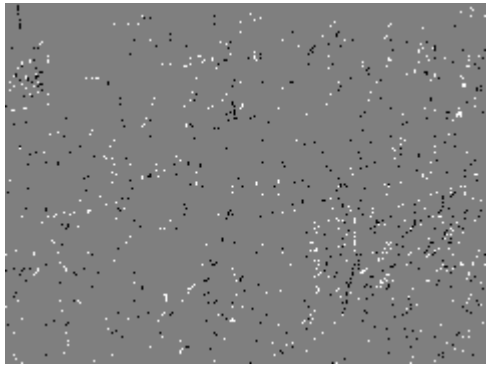
[This is a video]

The middle frame is withheld for evaluation

High-speed video
(21 frames)

[This is a video]

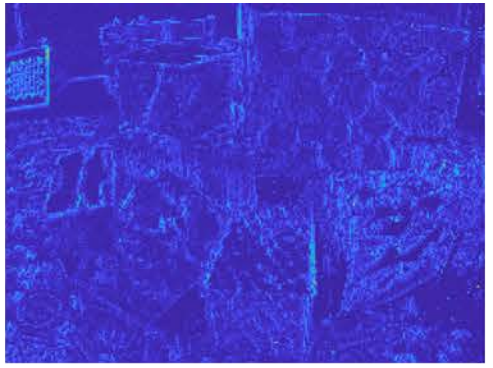
[This is a video]



Event frames (20 frames)

PSNR: 33.41 SSIM: .955

Frame #10



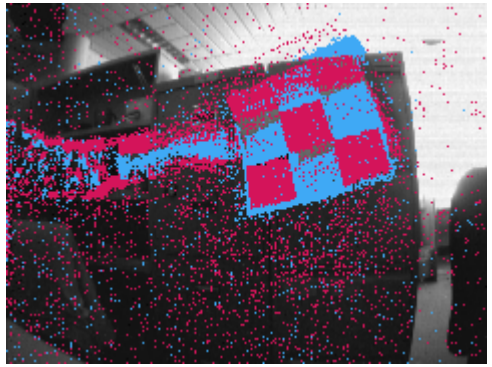
Error map of Frame #10

Results

- Motion deblur case
  - Given a blurry image + events in-exposure, recover intermediate sharp frames.

Blurry images

Events during exposure

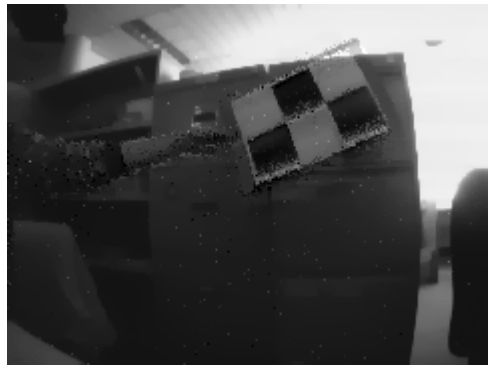


EDI [CVPR'19]

Ours

Reconstruction

[This is a video]



[This is a video]

Synergy: diff. inverse solver + residual learning

Imaging hardware

Synergistic models

task

Physics-based model

Learning-based model

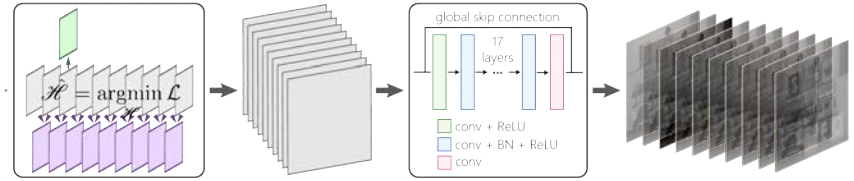
#1 APS + DVS camera

Differentiable inverse solver

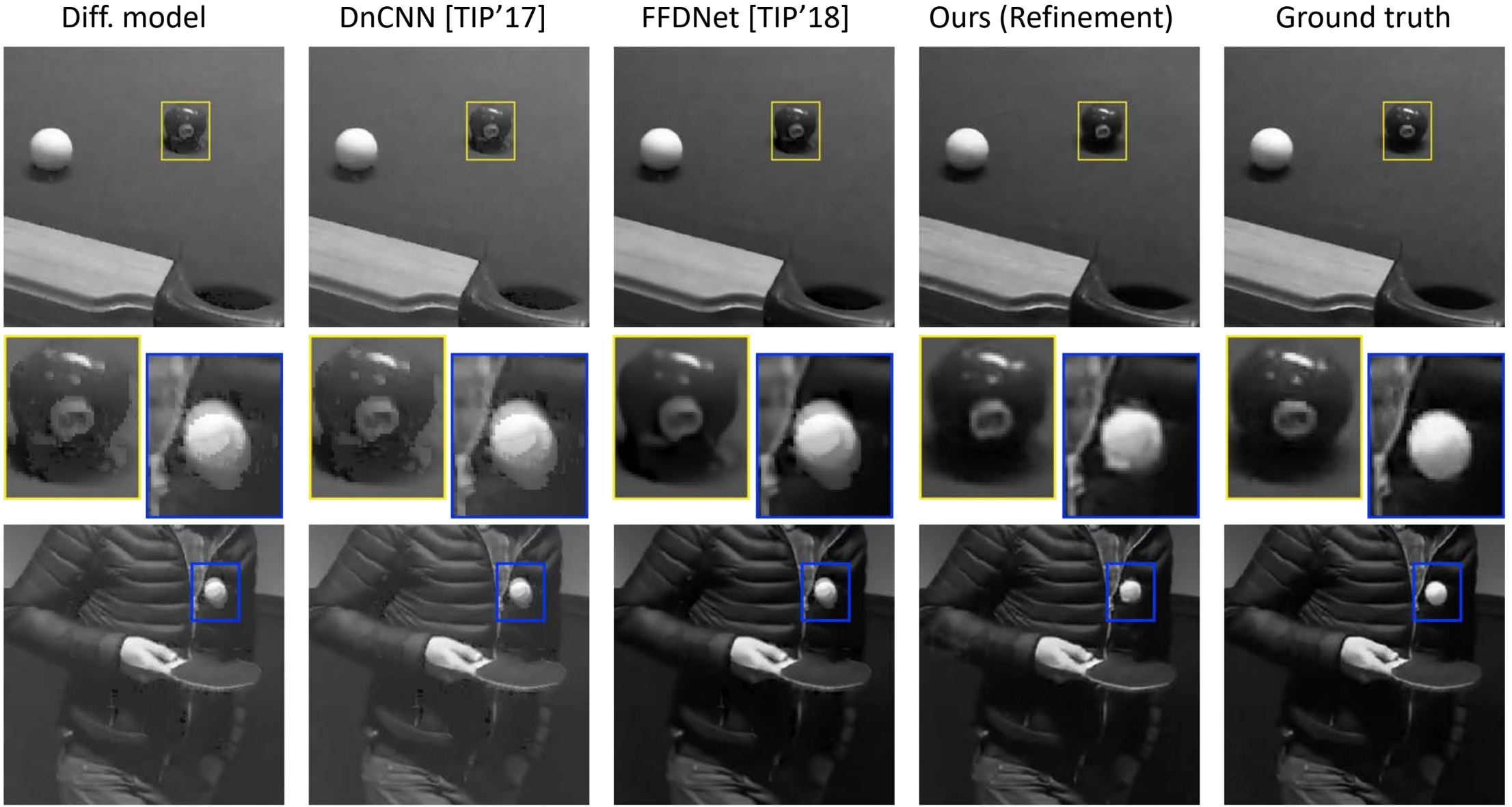
Video frame synthesis: Interpolation, extrapolation, motion deblur

Augmentation

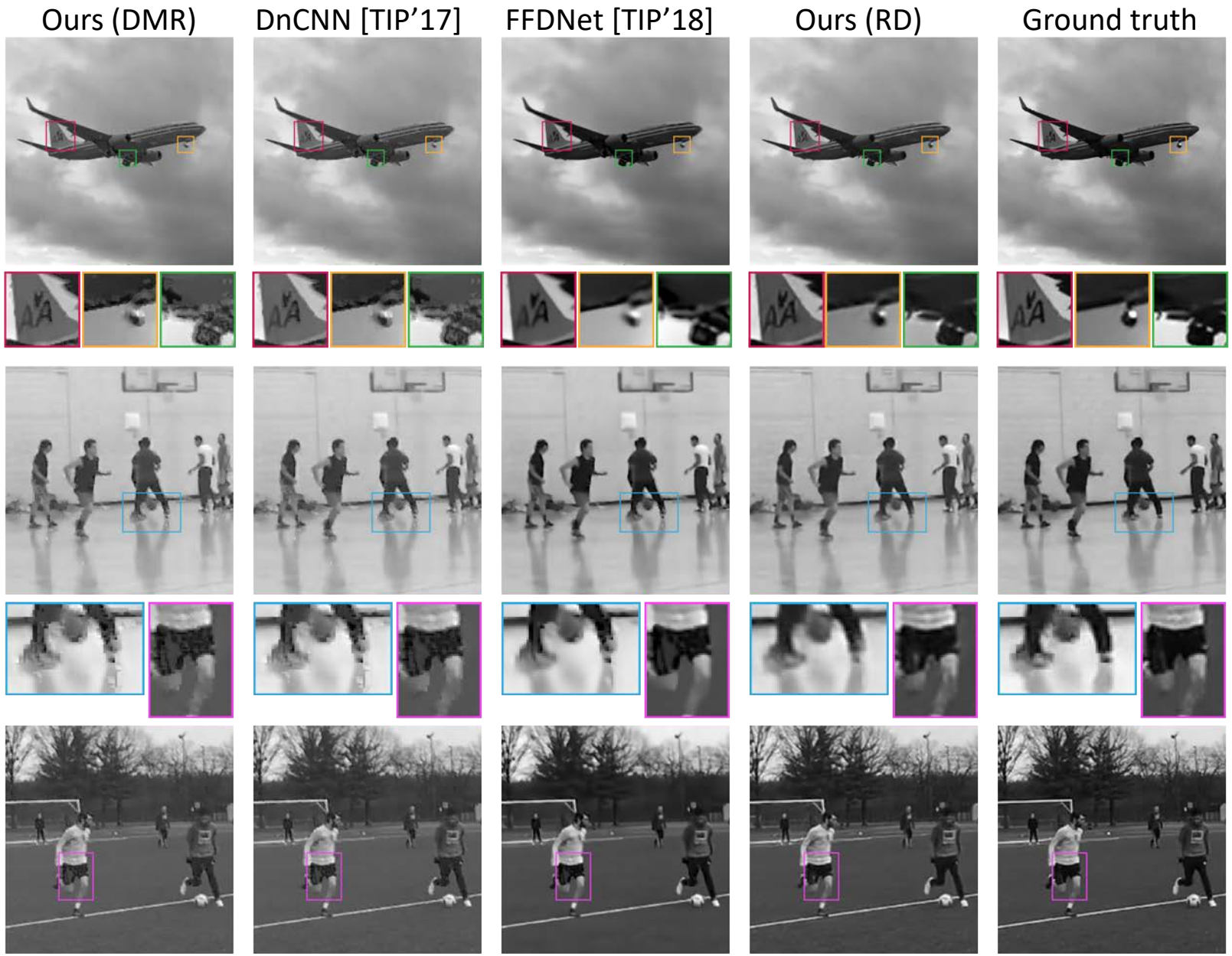
Residual refinement



Results (residual refinement)



More results (residual refinement)



clip name	metric	DMR	DnCNN	FFDNet	Ours
airplane	PSNR	30.91	31.10	30.92	**31.38**
	SSIM	0.975	0.982	0.976	**0.982**
basketball	PSNR	23.55	24.05	23.47	**24.06**
	SSIM	0.963	0.971	0.964	**0.972**
soccer	PSNR	29.96	31.08	30.13	**31.29**
	SSIM	0.961	0.974	0.962	**0.975**
billiard	PSNR	36.46	35.42	**36.48**	36.46
	SSIM	0.982	0.986	0.983	**0.987**
ping pong	PSNR	32.46	32.26	**32.50**	32.24
	SSIM	0.974	0.978	0.975	**0.979**

Results

Comparison with non-event-based frame interpolation approach.

Interpolation using APS-only

SepConv [CVPR'17]

[This is a video]

Ground truth

[This is a video]

Interpolation using APS + DVS

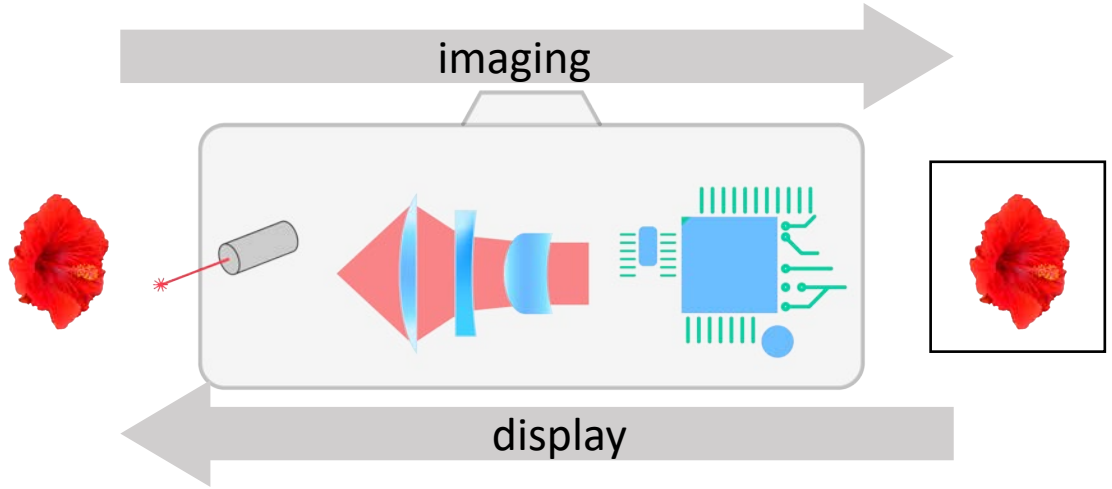
Ours (DMR + Refinement)

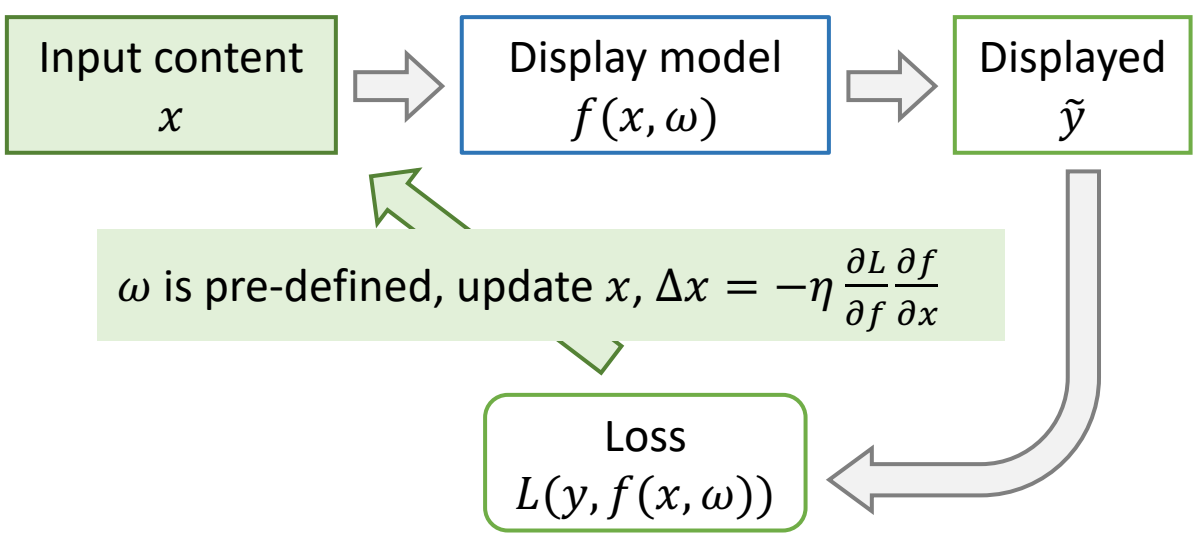
[This is a video]

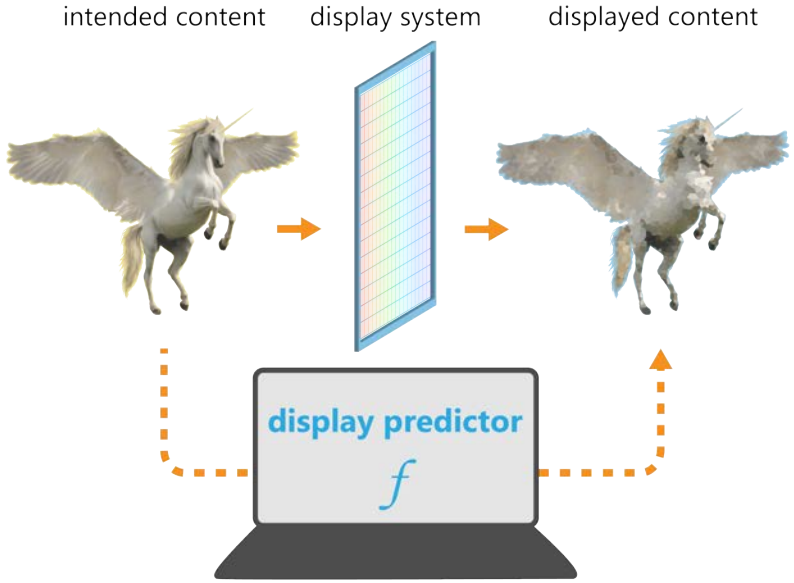
Physics-based models as means of data augmentation

Imaging hardware	Synergistic models			task
	Physics-based model		Learning-based model	
#1 coded aperture camera	Motion features	Augmentation	Deep classifier	*Privacy preserving action recognition*
	# of input images x *# of masks*			
#2 APS+DVS camera	Diff. fusion model	Augmentation	Residual learning	*Video frame synthesis*
	# of input videos x *# of reconstruction parameters*			

Computational display: a reverse process of imaging

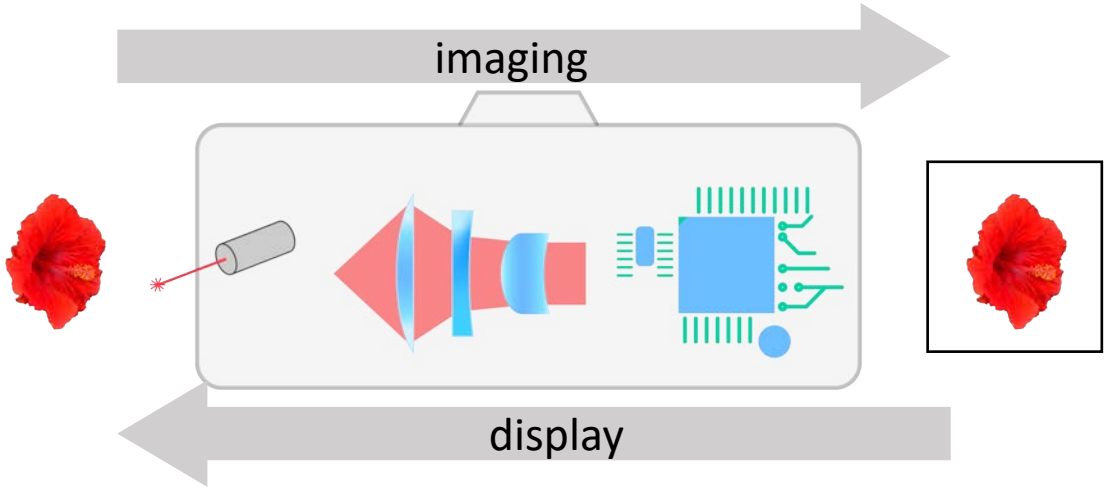


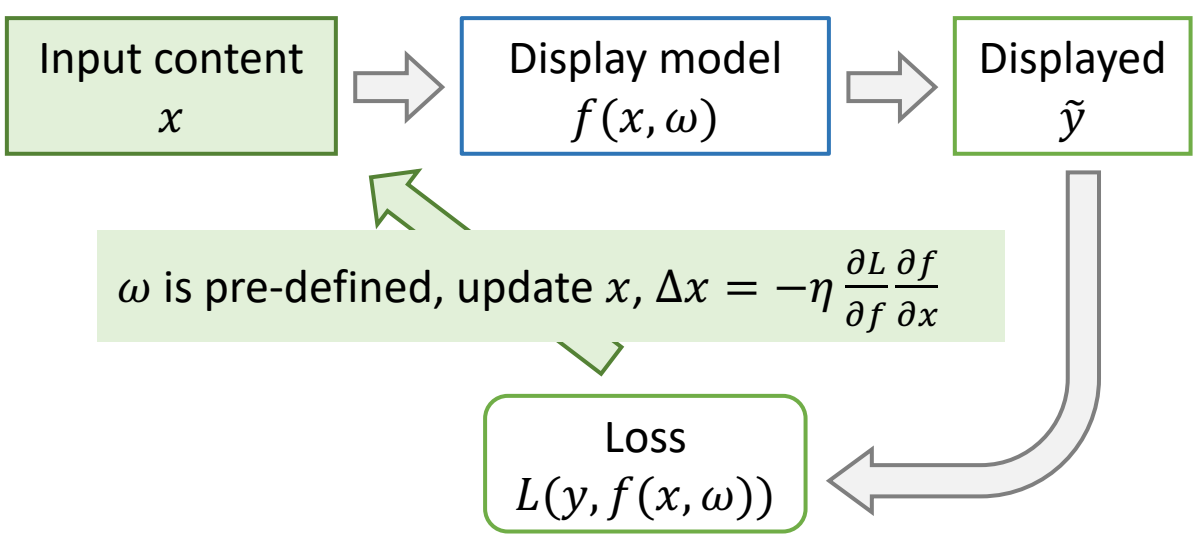


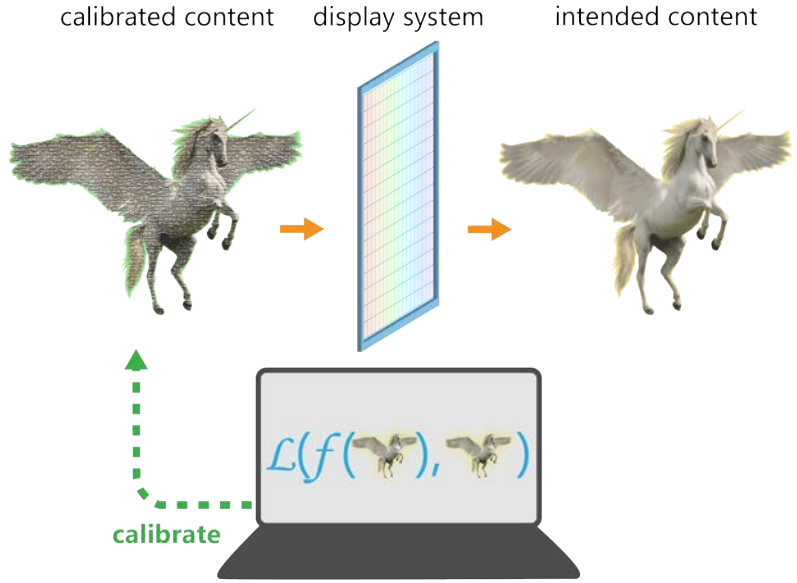


Differentiable display model: optimize content to display

Computational display: a reverse process of imaging





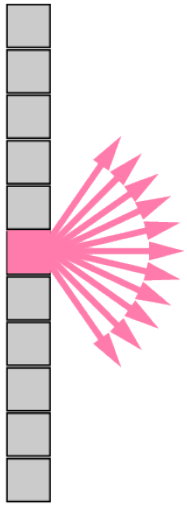


Differentiable display model: optimize content to display

What are we talking about when we talk about 3D display? – parallax

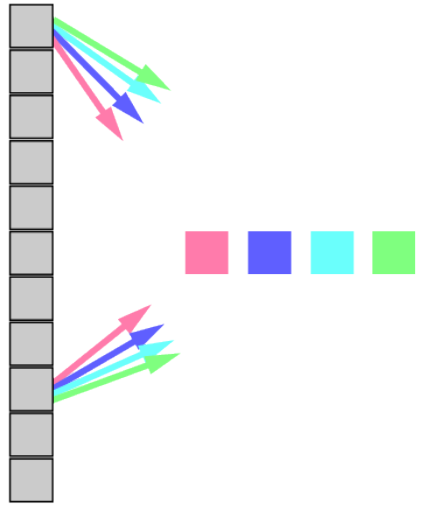
[video of a recorded volume hologram]

2D display



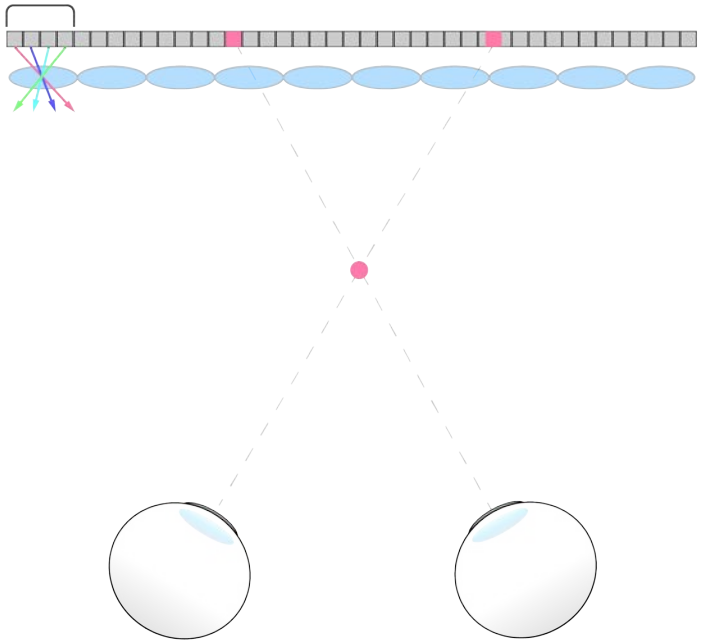
Program 2D pixels

3D display



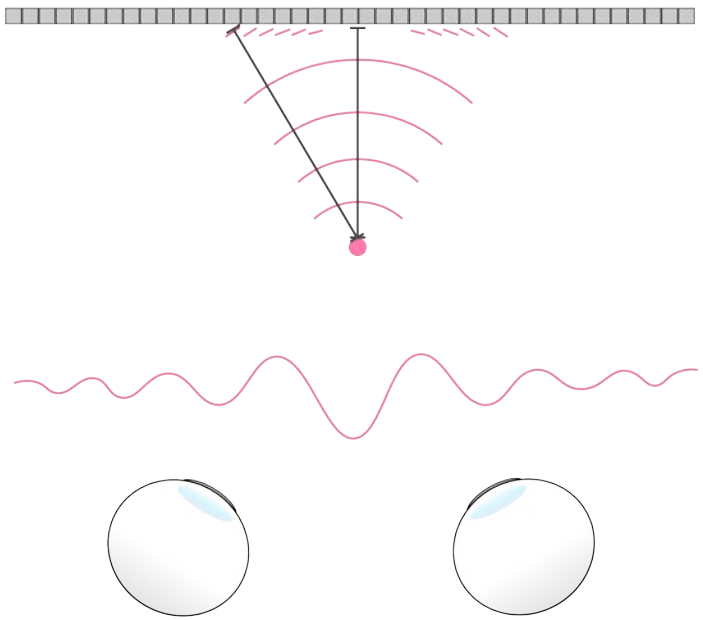
Program 4D rays
- 2D in spatial
- 2D in angular

Ray programming: light field vs. holographic display



Light field display:

- Use microlens/pinhole array to bend light



Holographic display:

- Use coherent light and phase-based Spatial Light Modulators

The total number of pixels (spatial-bandwidth product) remains unchanged:

- 2D display: N x N pixels
- 3D display: (N/M x N/M) x (M x M)

M: angular resolution

Challenges:

- computation: $10^{12} - 10^{14}$ pixel/s for 3D iPad or 3D iMac, (not for HMD)
  - Currently can achieve $10^{9} - 10^{11}$
- hardware: pixel pitch, form factor etc.

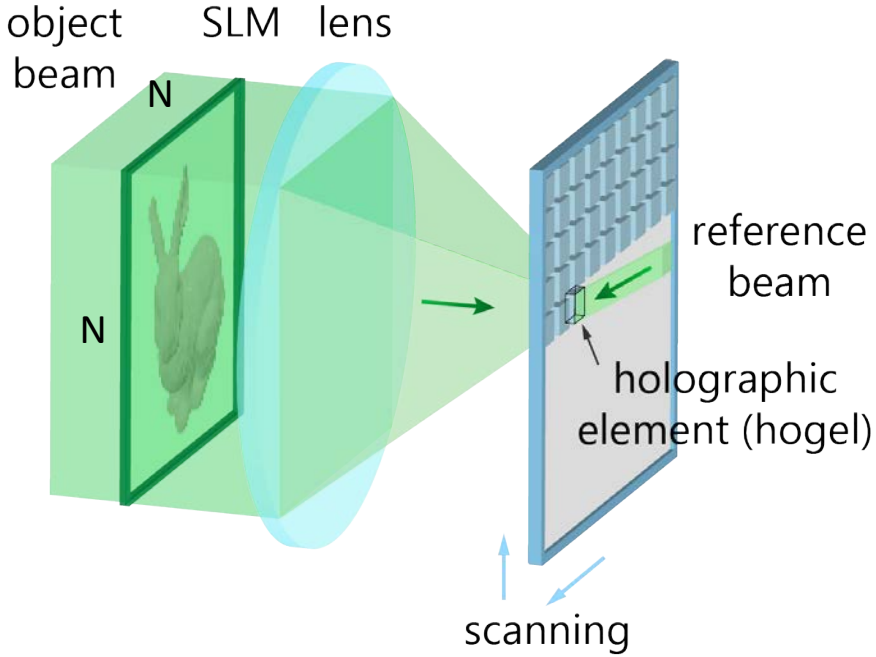
[Calculation adapted from *Yamaguchi, M. JOSAA 2016*]

Can we program N x N x M x M rays using N x N pixels?

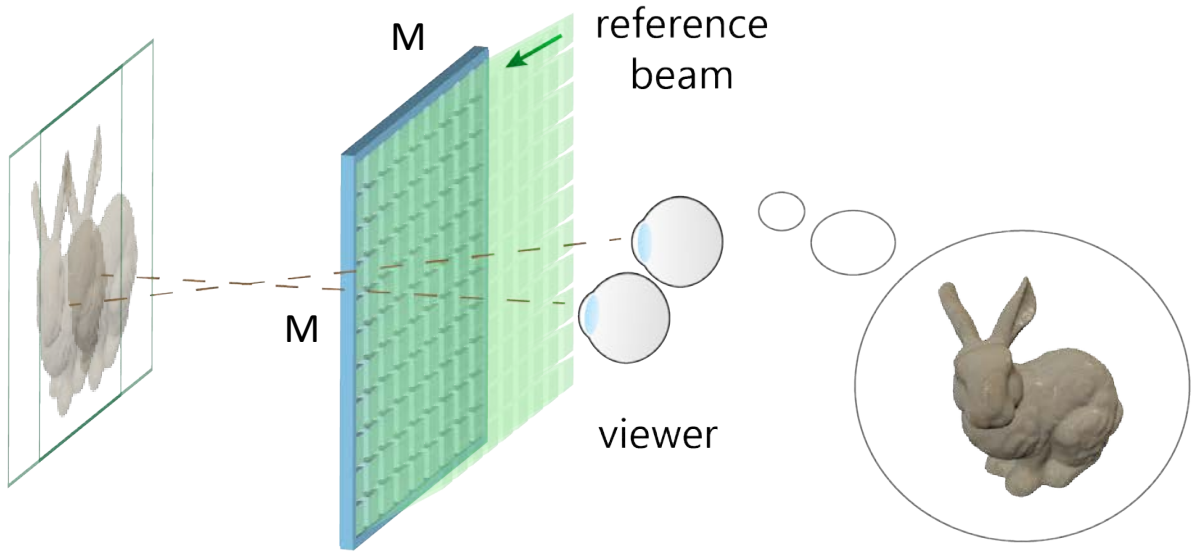
Holographic stereogram

Holographic stereograms

Recording (preparing the HS)



Playback (viewing the HS)

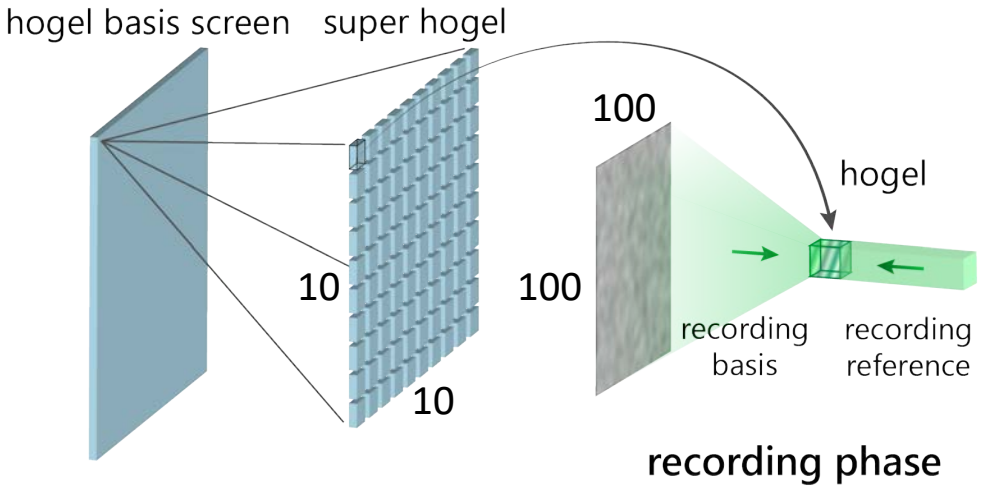


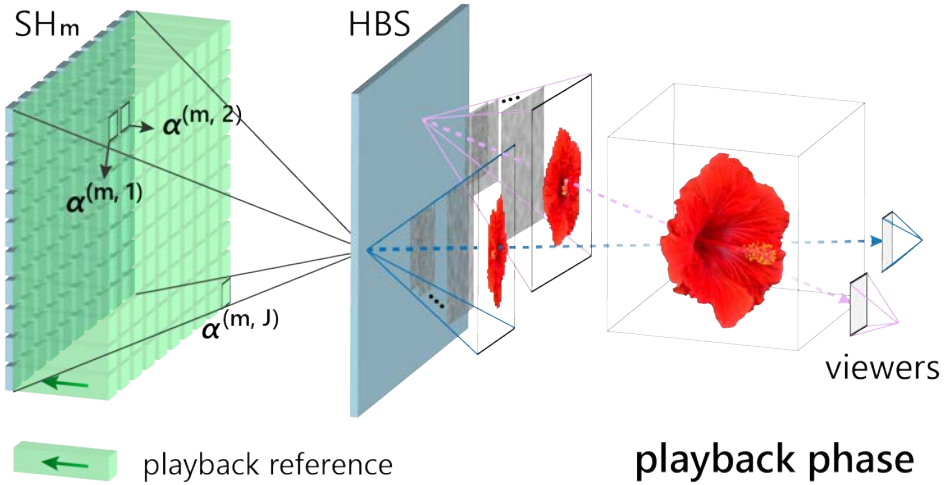
See-through: viewers are viewing the reflected light.

SLM has N x N pixels, and the stereogram has M x M hogels.
Total rays: N x N x M x M
But this is static 3D image.

Add programmability to the ref. beam

Hogel basis display





HBS: 1000 x 1000 hogels = 100 x 100 super hogels (SH)
Each hogel records a 100 x 100 basis pattern

One SH produce one viewpoint image.

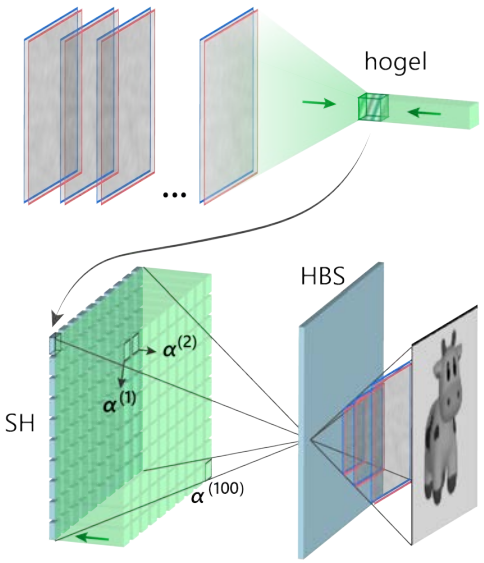
$$I_{SH_m} = \left| \sum_{j=1}^{J} \alpha^{(m,j)} B^{(j)} \right|^2$$

Spatial averaging for SH:
viewing distance should satisfy SH size within visual acuity.

HBS: 1000x1000 ref. pixels, i.e. $\alpha^{(m,j)}$
can produce 100x100 views;
100x100x100x100 rays.

One super hogel (SH)

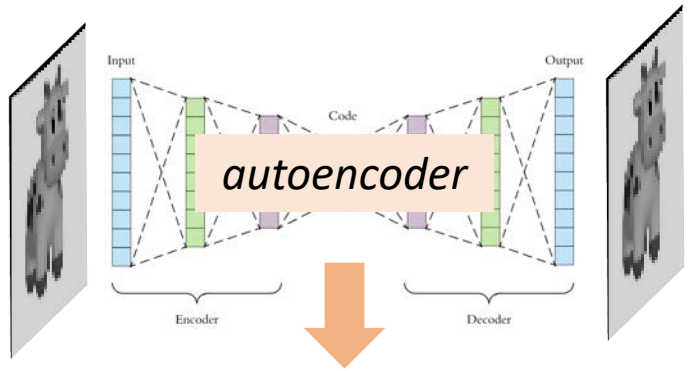
Display system hardware

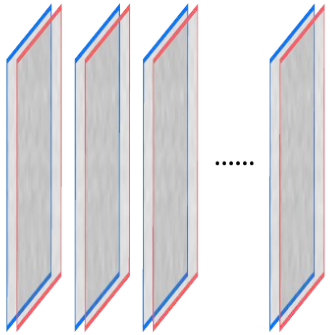


Hardware unavailable due to COVID-19

Construct a super hogel: record each volume hologram & playback

Neural networks software



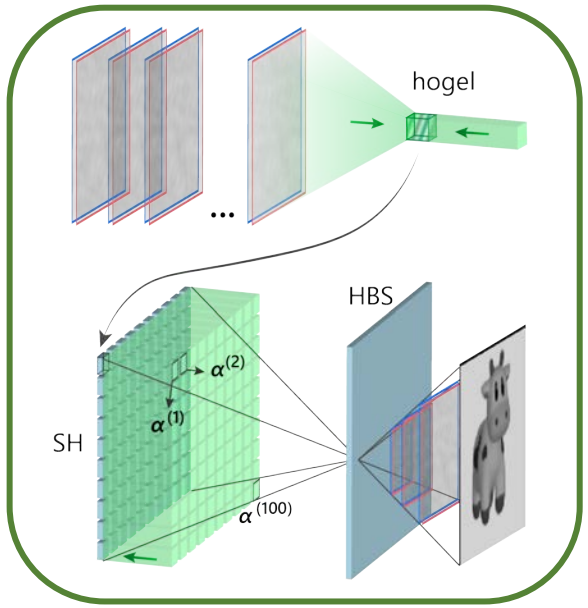


Optimize the hogel bases + coeff.

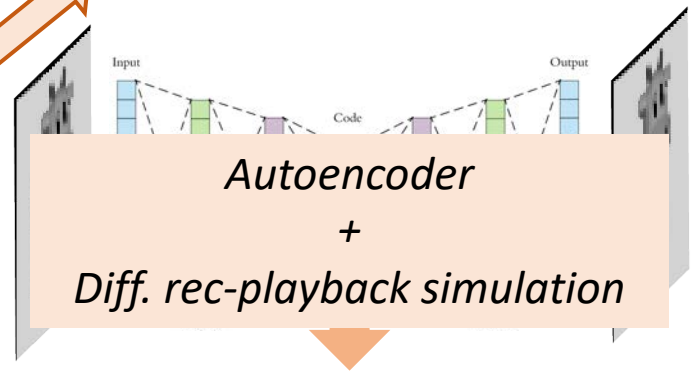
One super hogel (SH)

Display system hardware

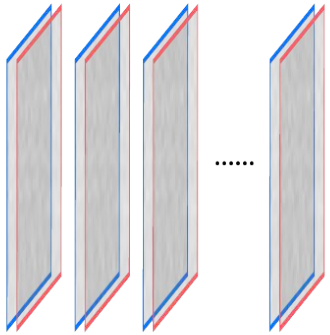
Differentiable modeling of recording-playback



Neural networks software



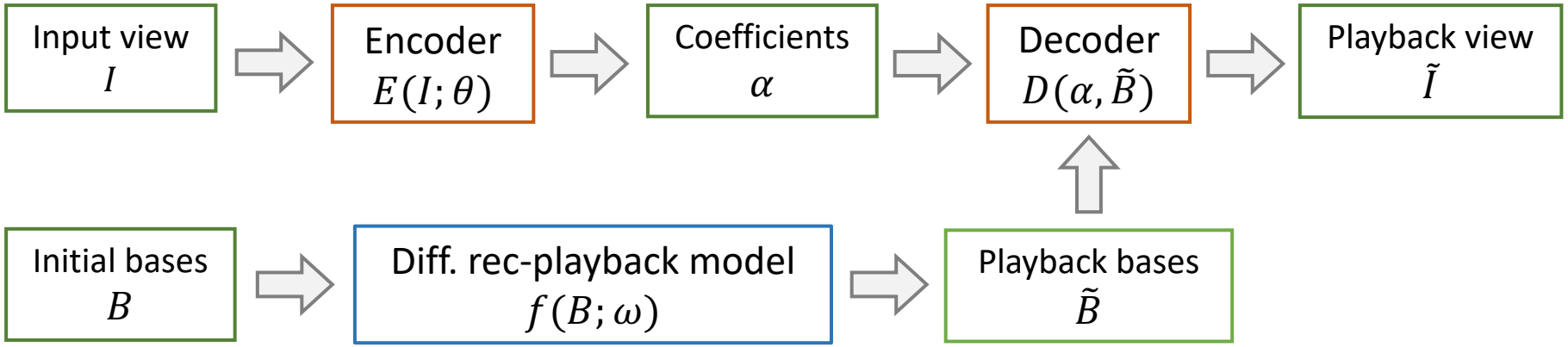
Autoencoder
+
Diff. rec-playback simulation



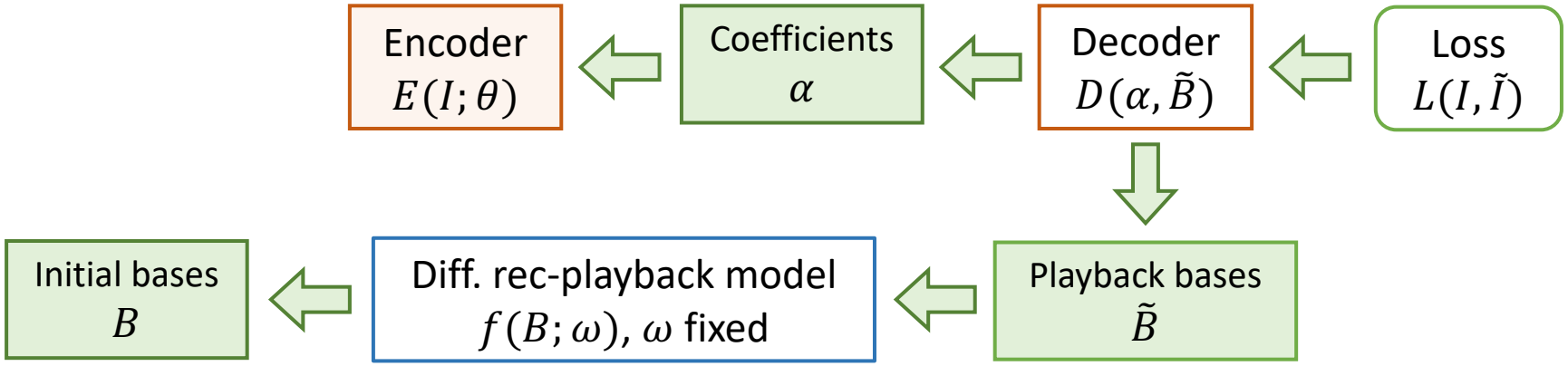
Optimize the hogel bases + coeff.

Synergistic model for (super) hogel basis learning

Forward pass



Backprop.



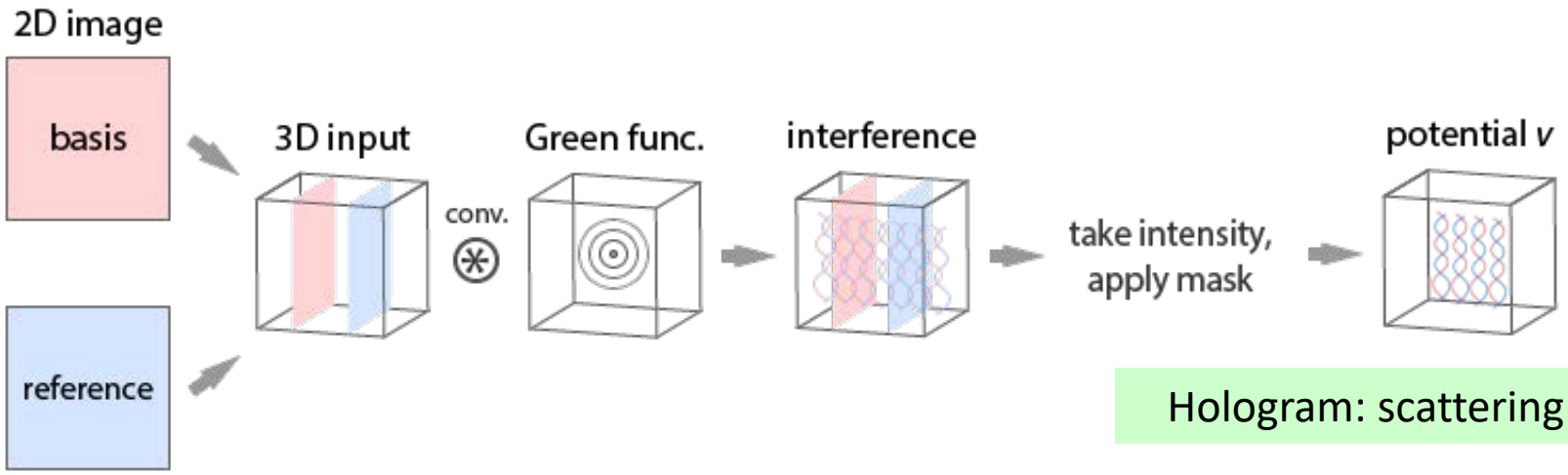
(Physics-in-the-loop autoencoder)

Physics-based (recording-playback) model

Solve the inhomogeneous Helmholtz equation, based on Born approx.

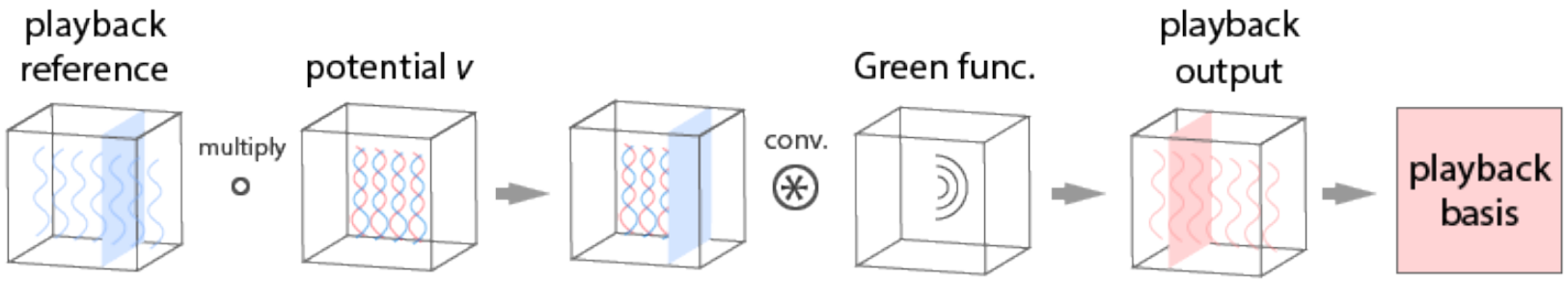
3D convolution with the Green's function (PSF): $G(x) = \frac{\exp(ik_0|x|)}{|x|}$

Record



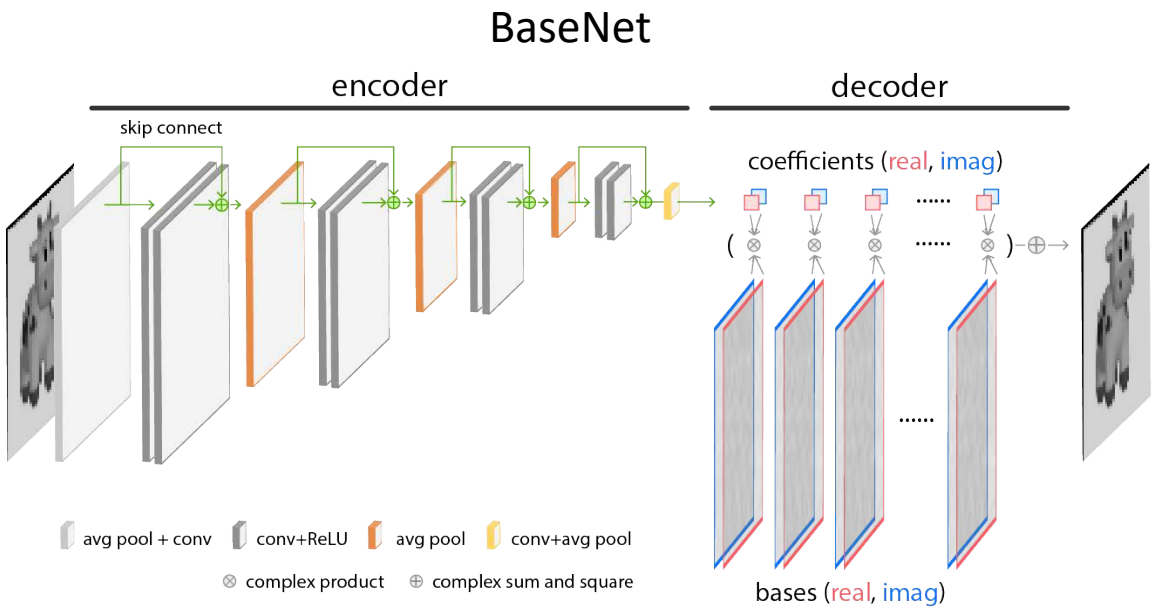
Hologram: scattering potential

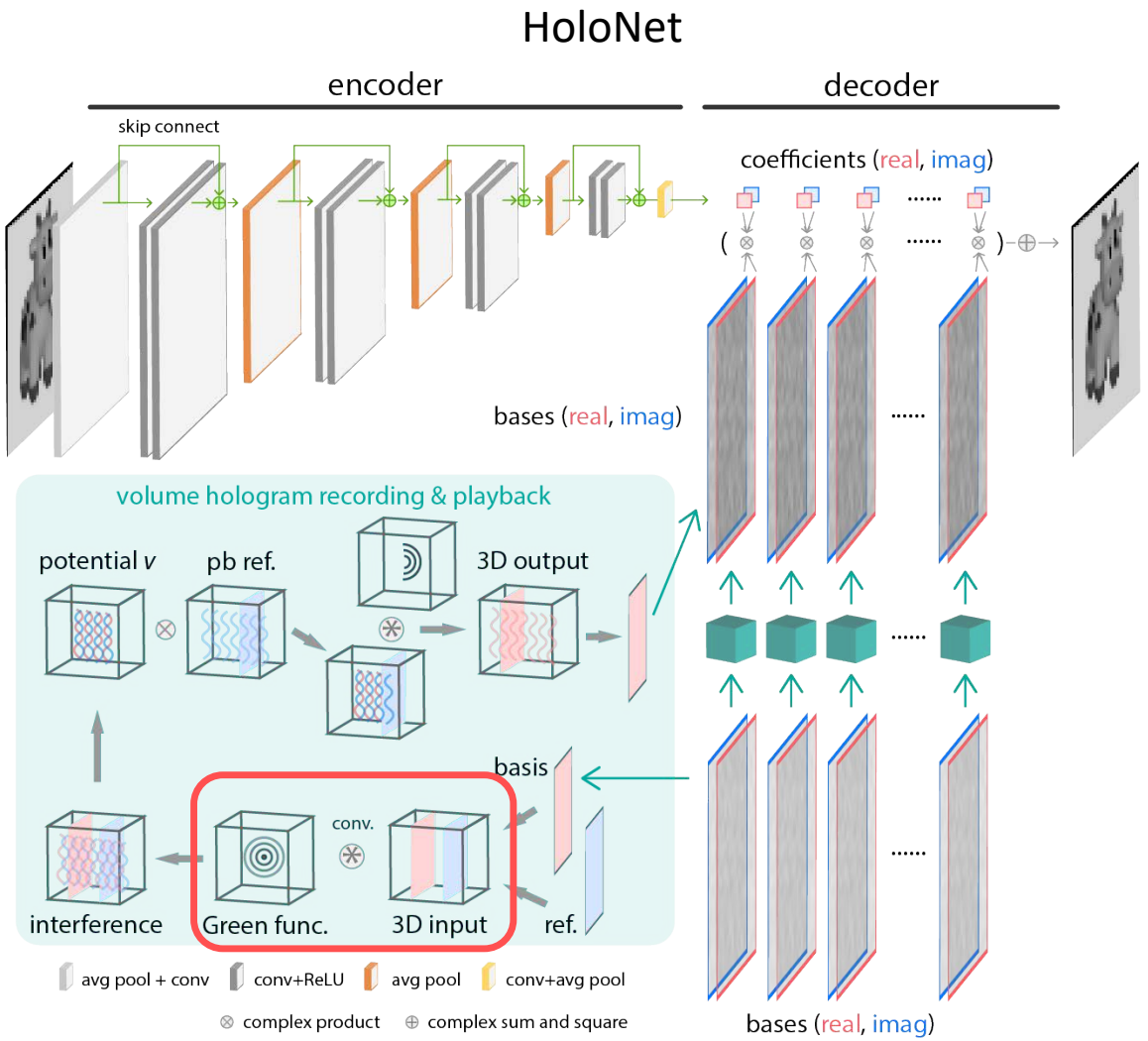
Playback



Hologram

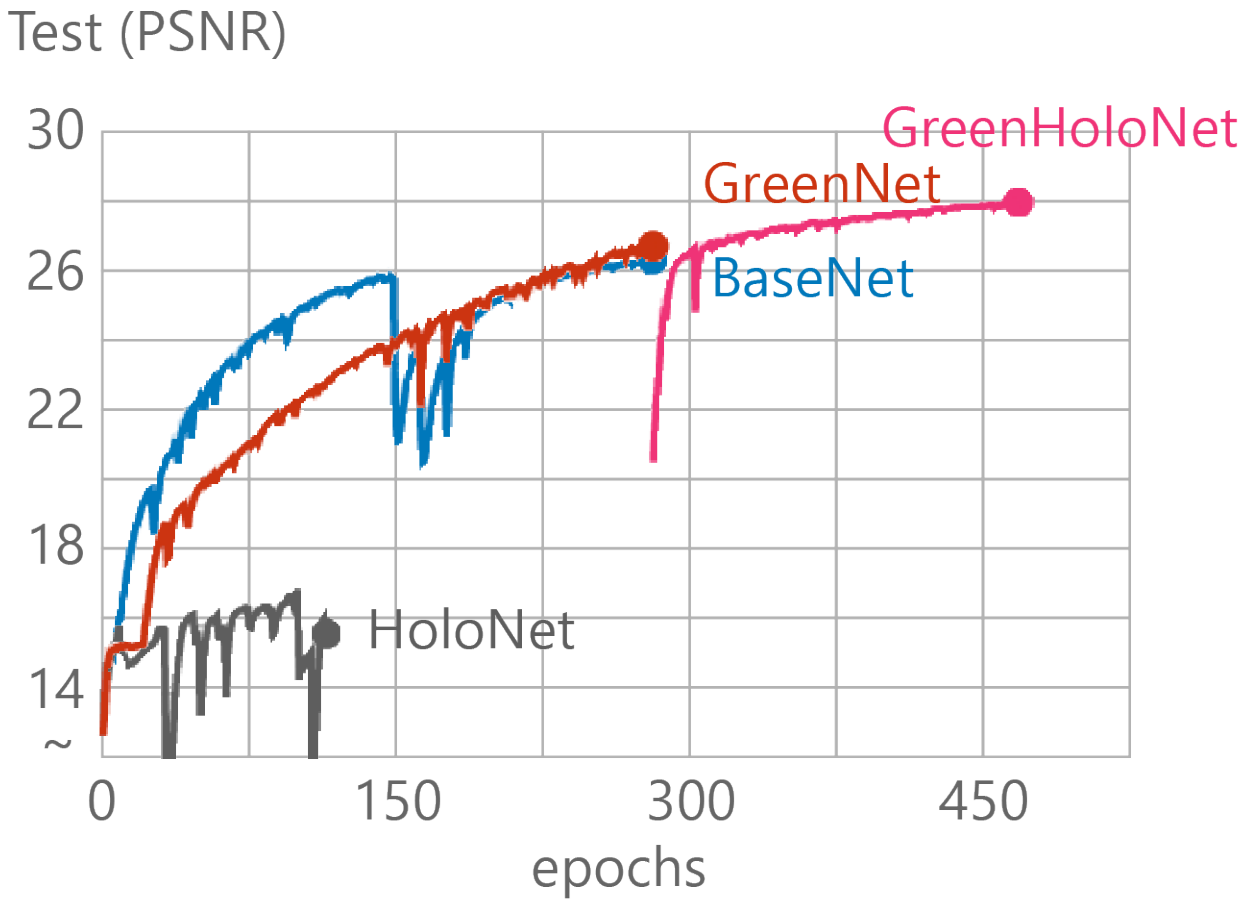
Learning-based model vs synergistic model



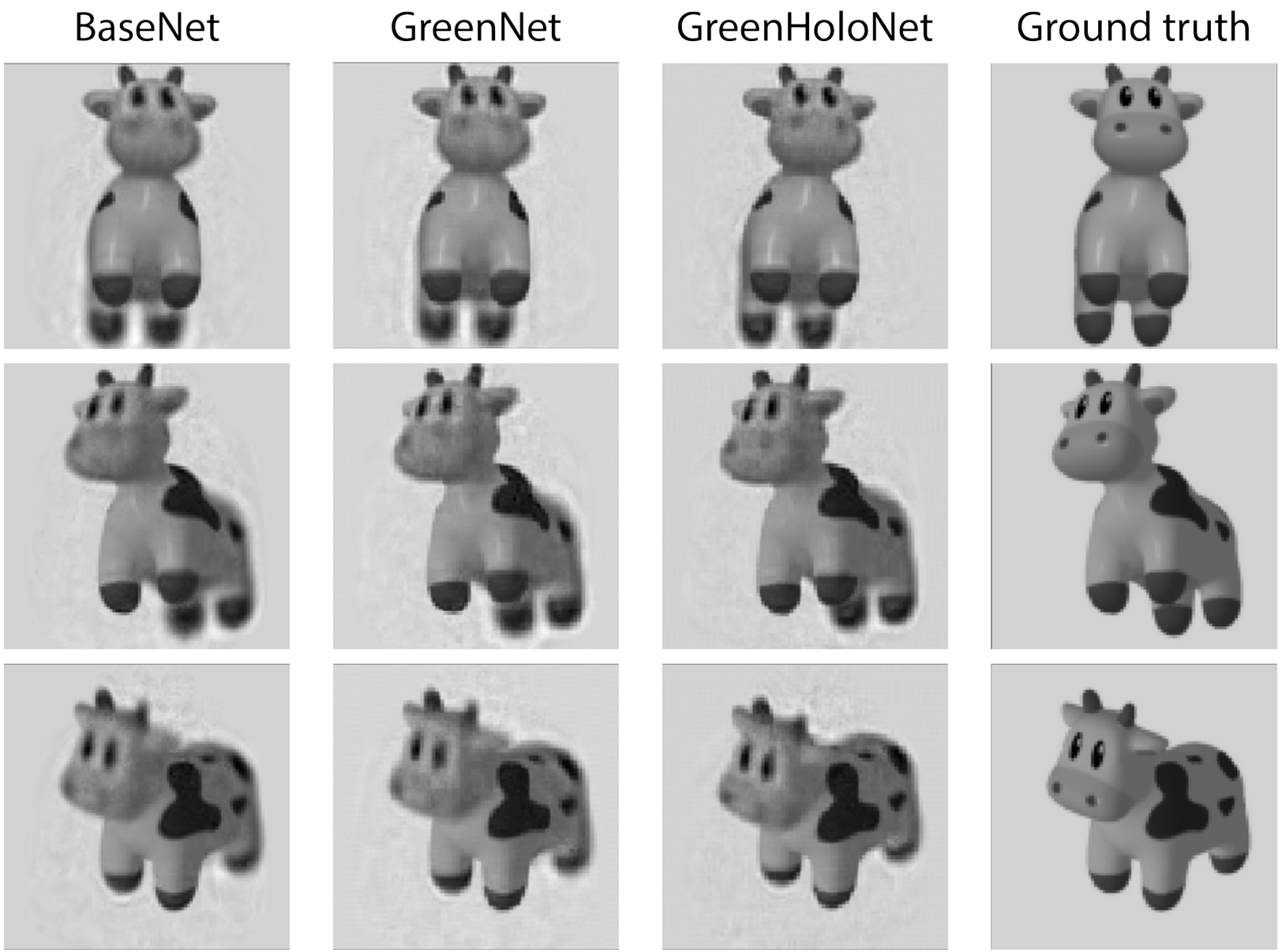


GreenNet: each basis will perform 3D conv. w/ the Green's function.

Results

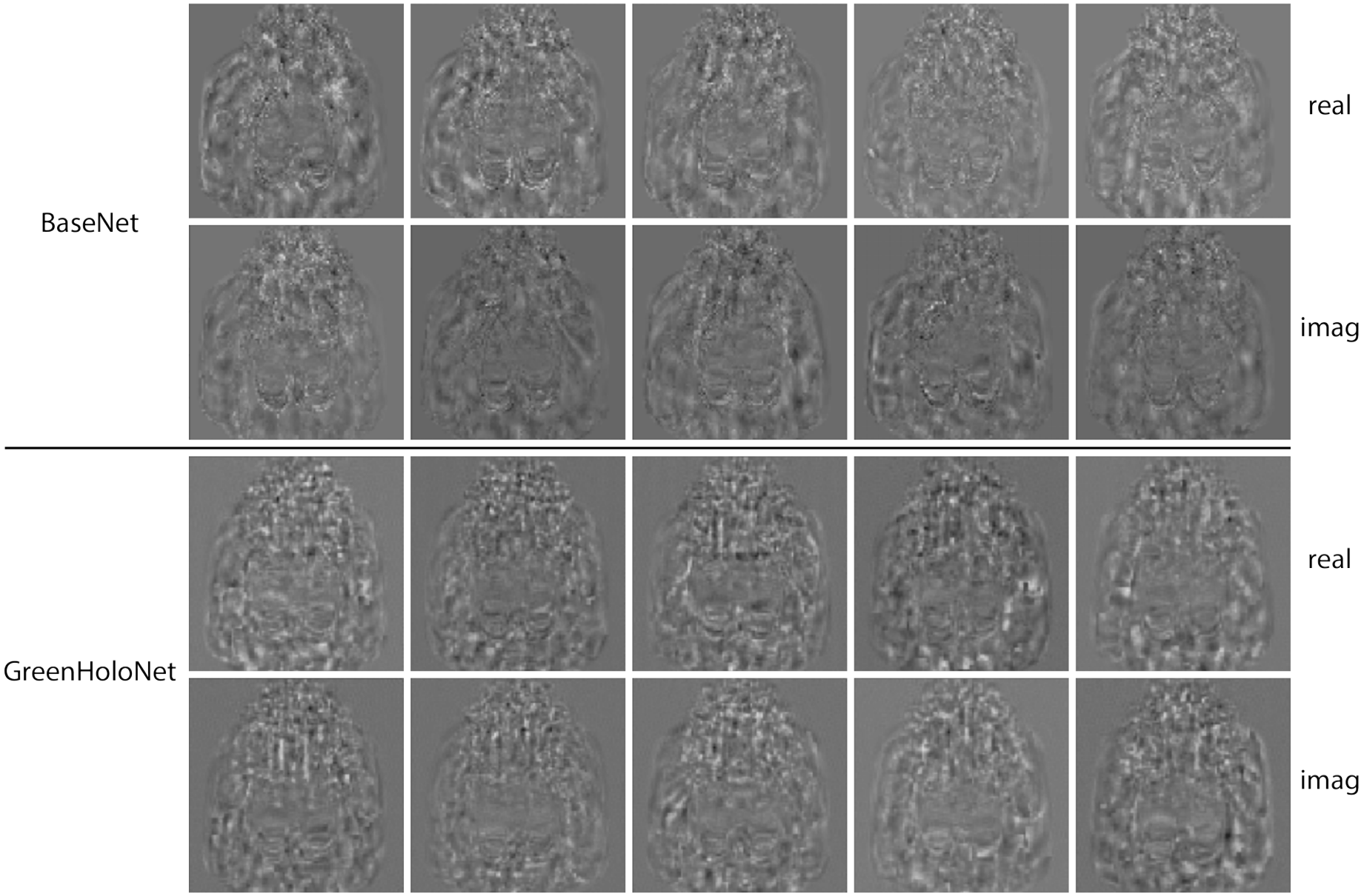

Test (PSNR)
30
26
22
18
14
~
0
150
300
450
epochs
GreenHoloNet
GreenNet
BaseNet
HoloNet

Results



Network	PSNR (dB)
BaseNet	26.1
GreenNet	26.5
GreenHoloNet	27.9

Visualization of example learned bases

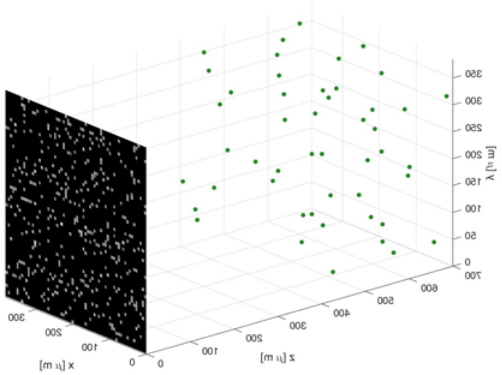


Summary

Imaging/display hardware	Synergistic models			task
	Physics-based model	Learning-based model		
#3 Volume hologram	Diff. rec-playback model	Partial pre-training	Autoencoder	*Basis learning / image compression*
	[differentiable modeling]			
#2 APS+DVS camera	Diff. fusion model	Augmentation	Residual learning	*Video frame synthesis*
	[physics-based model as data generator]			
#1 coded aperture camera	Motion features	Augmentation	Deep classifier	*Privacy preserving action recognition*

Holographic compressive video

OE'17



Recover 3D videos from single images

Coded exposure, compressive video

Holographic imaging

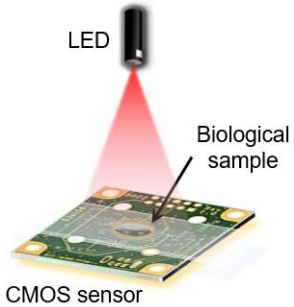
Synergistic models

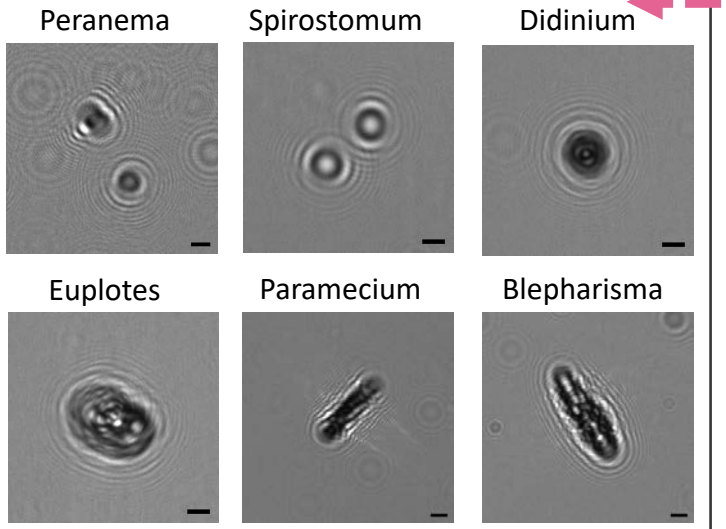
Lens-free coded aperture cameras, *CVPRW'19*

High speed RGB-event cameras, *ICCVW'19*

3D holographic display

On-chip holographic microscopy, *BOE'17, DH'17, COSI'17*





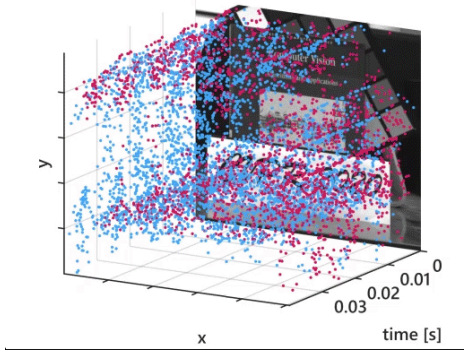
Dictionary learning super resolution +

phase retrieval

Guided event filtering

CVPR'20 & journal extension

image+events



Event denoising and super resolution w/

HR image guidance

Lossy event compression

Srutarshi et al. in submission '20

Acknowledgement

Advisor: Ollie Cossairt,

Committee: Aggelos Katsaggelos, Jack Tumblin, Nathan Matsuda

Internship mentors: Dikpal Reddy (Light co.), Sing Bing Kang (Microsoft Research), Xiaokai Li (Apple Inc.)

Coauthors: Lei Tian (Boston U.), Roarke Horstmeyer (Duke U.), Boxin Shi (PKU), Leo Spinoulas, Kuan He, Donghun Ryu, George Chen, Qiqin Dai, Weixin Jiang, Francesco Pittaluga, Vibhav Vineet, Sudipta Sinha, Srutarshi Banerjee

Coauthors-to-be: Prasan Shedligeri, Hamid Hasani, Florian Willomitzer, Florian Schieffers

Comp-photo-lab members

My parents!