

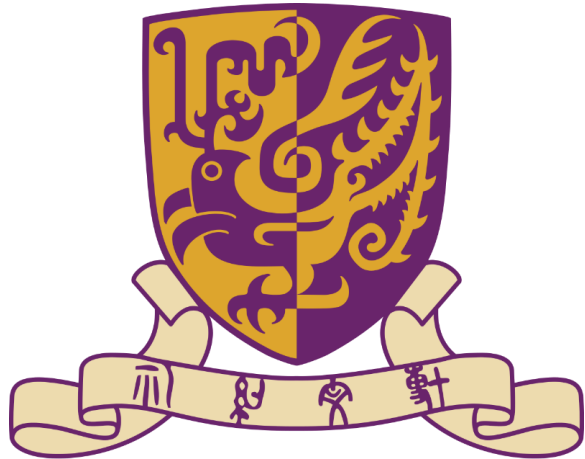
How to make a smart camera pipeline

Tianfan Xue

The Chinese University of
Hong Kong



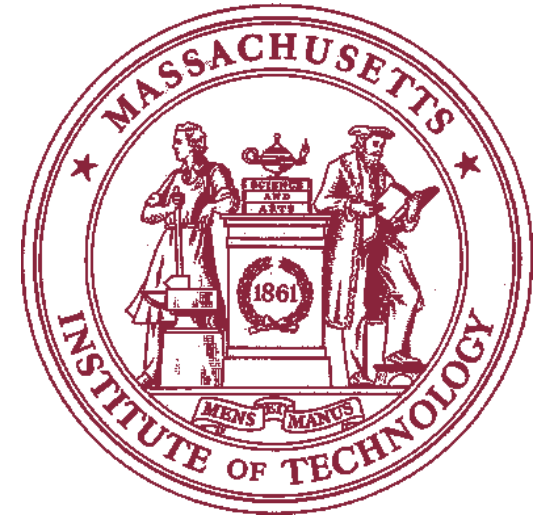
About myself



The Chinese Univ. of HK
Assist Prof.



Google Research, Gcam
Staff Eng.
Manager: David Saleson
2017-2022



MIT
Ph.D.
Advisor: William T. Freeman
2012-2017

Smart camera = ML algorithms applied images?



Object detection



Face beautification
[Leyvand et al., 2006]

Camera may not capture visual signal for ML system



One frame

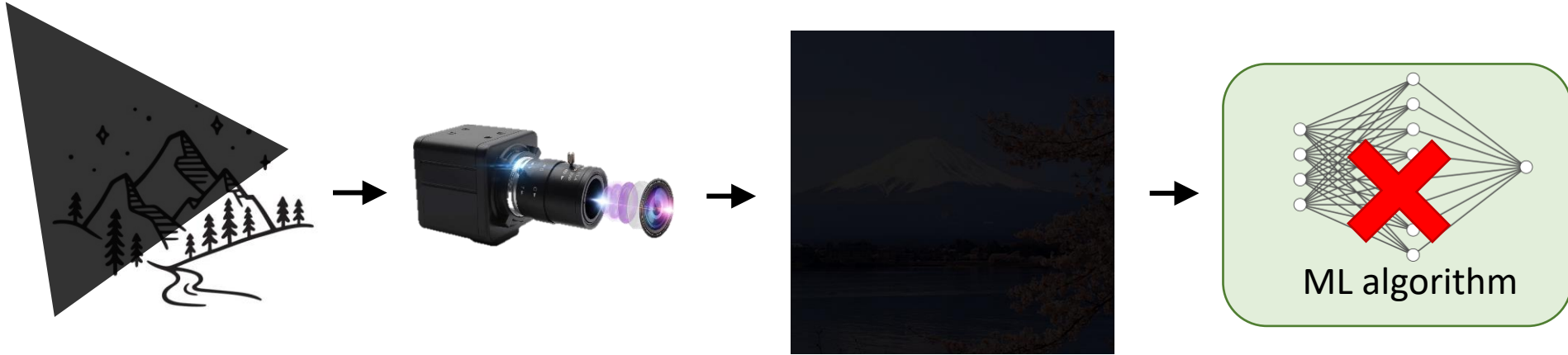


A frame 2s later

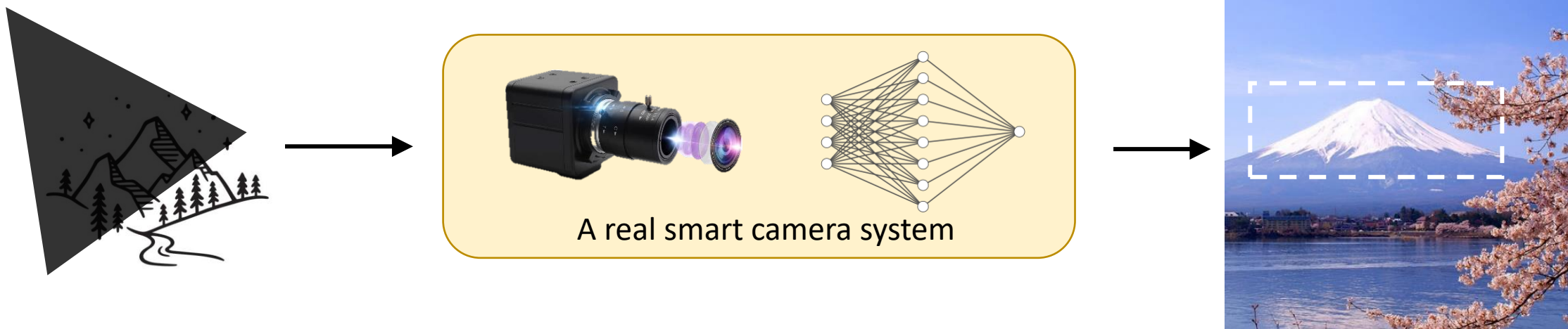
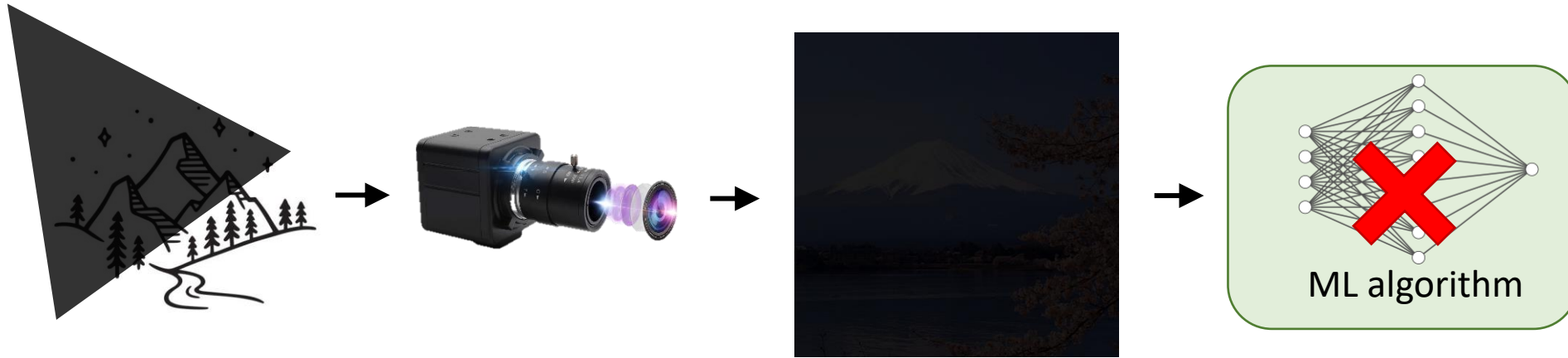
Images are too dark for ML to detect this biking person

Video from a fatal car crash

Separate design may fail

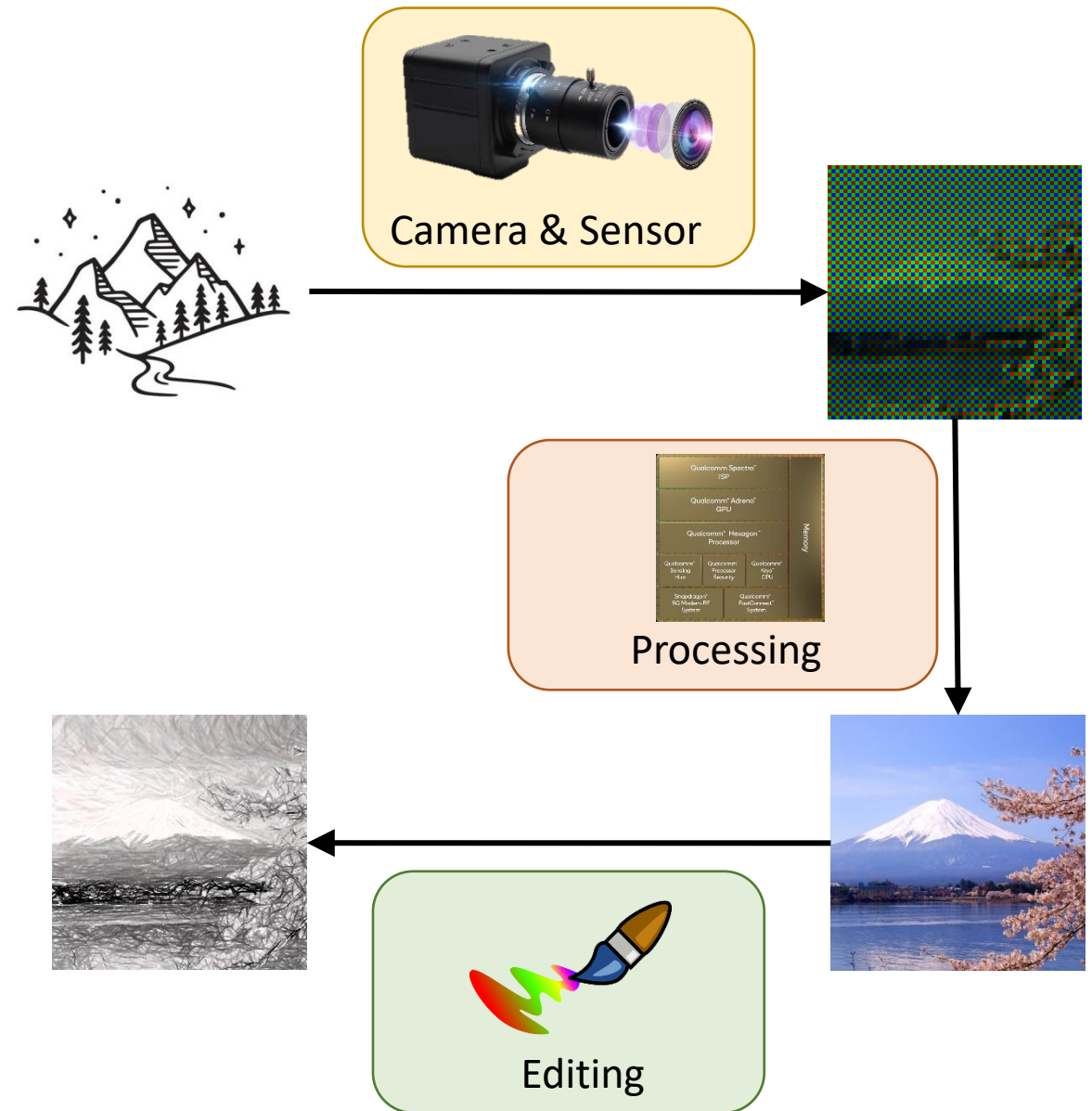


Machine learning embedded in the camera



Overview

- **Capturing:** multiple source fusion
- **Processing & editing**
 - **Training data:** synthetic data
 - **Network:** combine classic image processing algorithm and machine learning



Overview

- **Capturing:** multiple source fusion
- **Processing & editing**
 - **Training data:** synthetic data
 - **Network:** combine classic image processing algorithm and machine learning

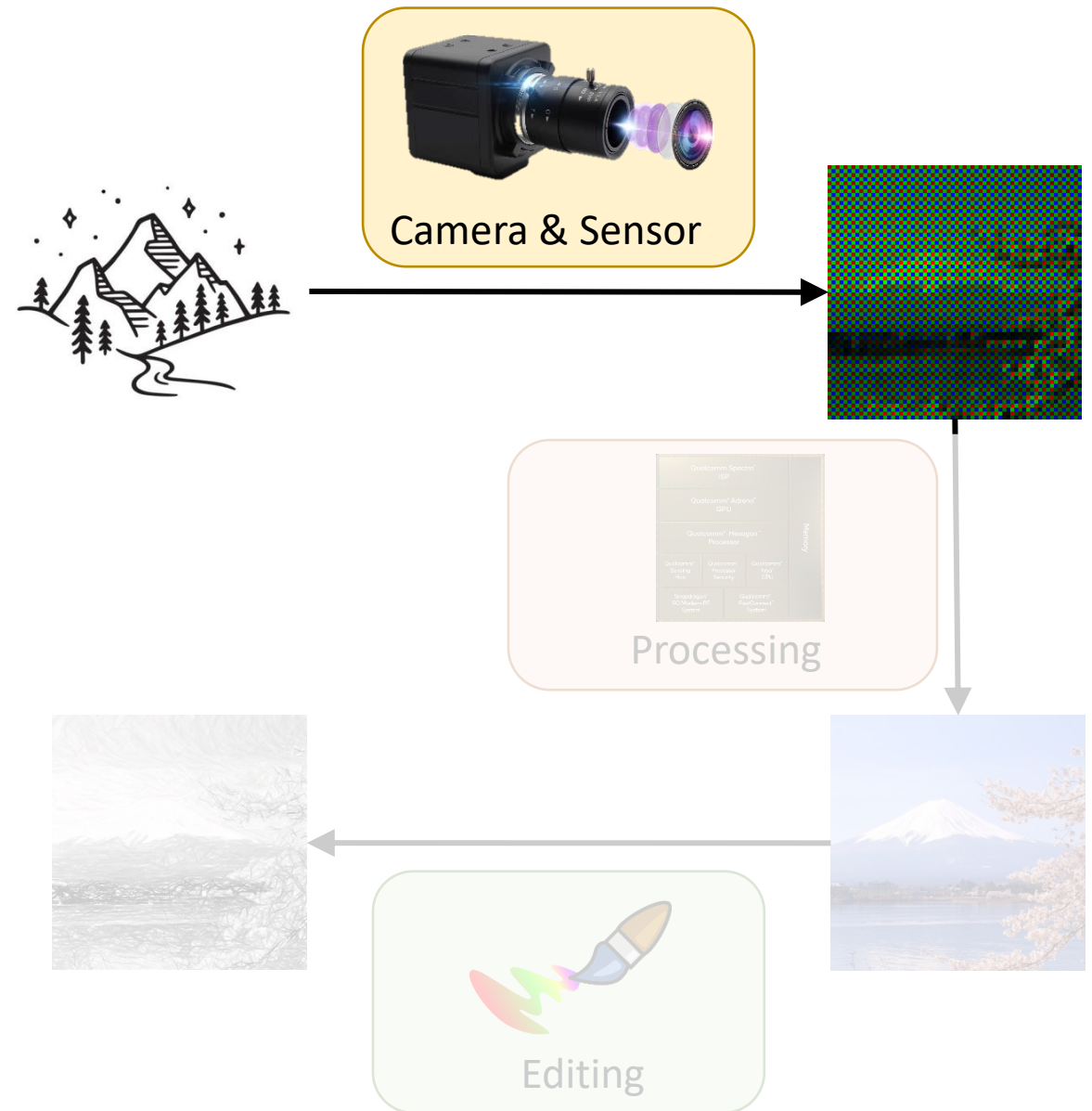


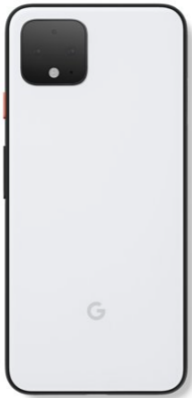


Image in lowlight



What is this object?

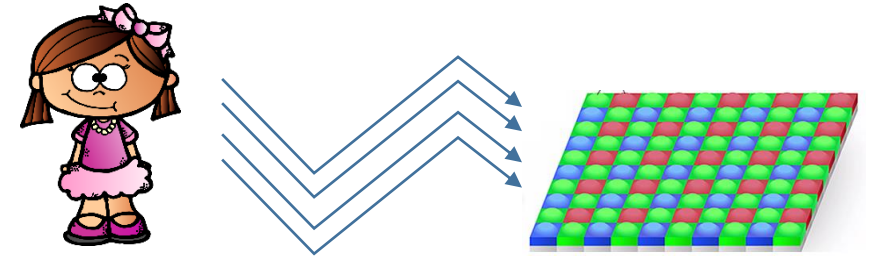
"Handheld Mobile Photography in Very Low Light", SIGGRAPH Asia 2019



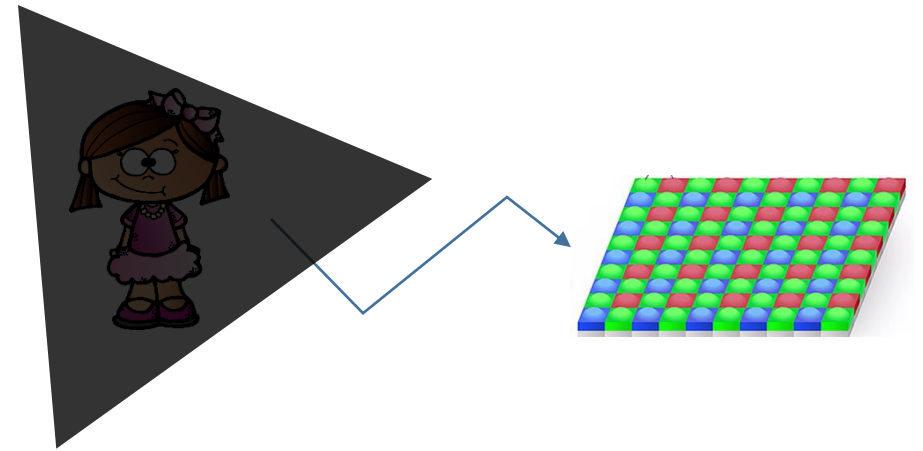
by our night sight algorithm



Not enough photons in lowlight

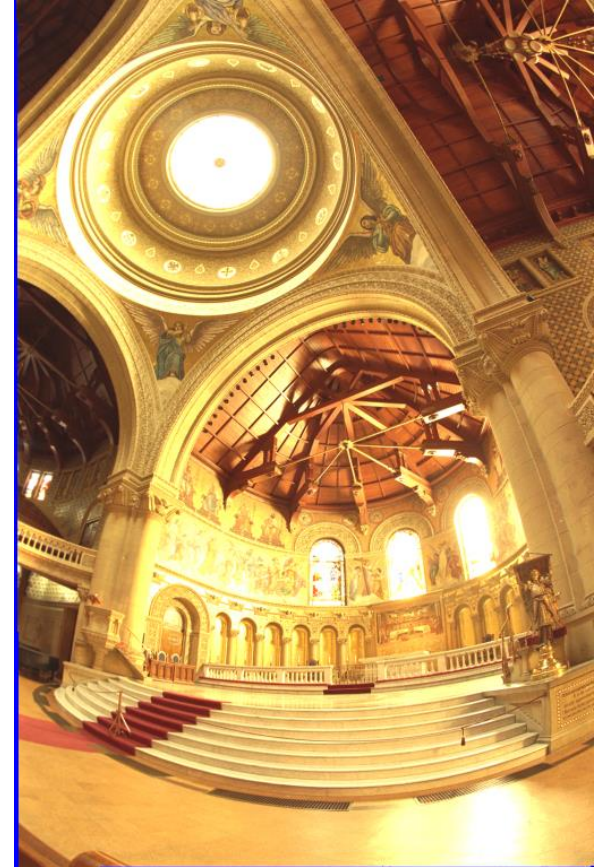
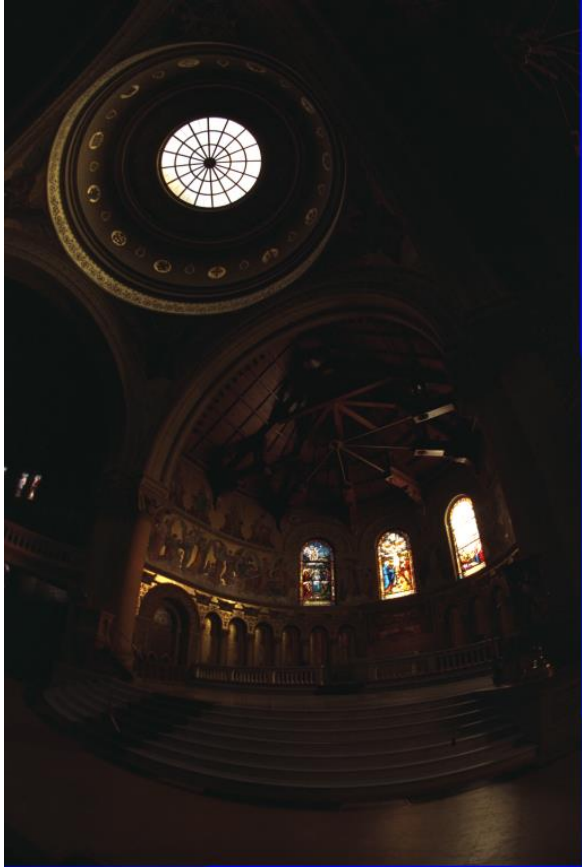


Good lighting



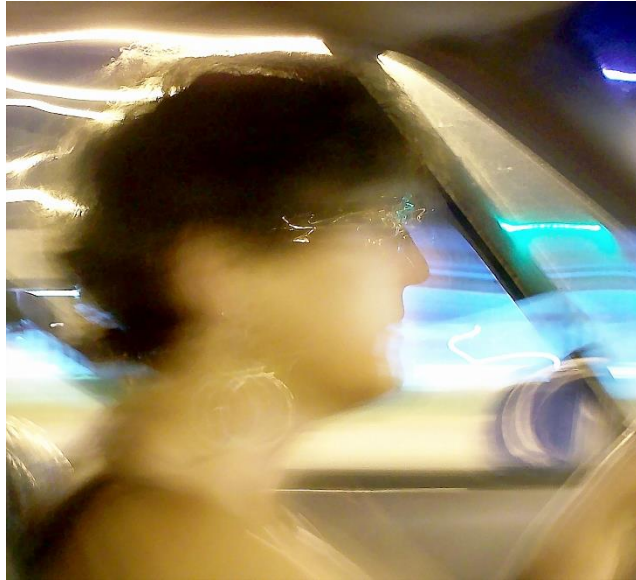
Lowlight

Exposure bracketing



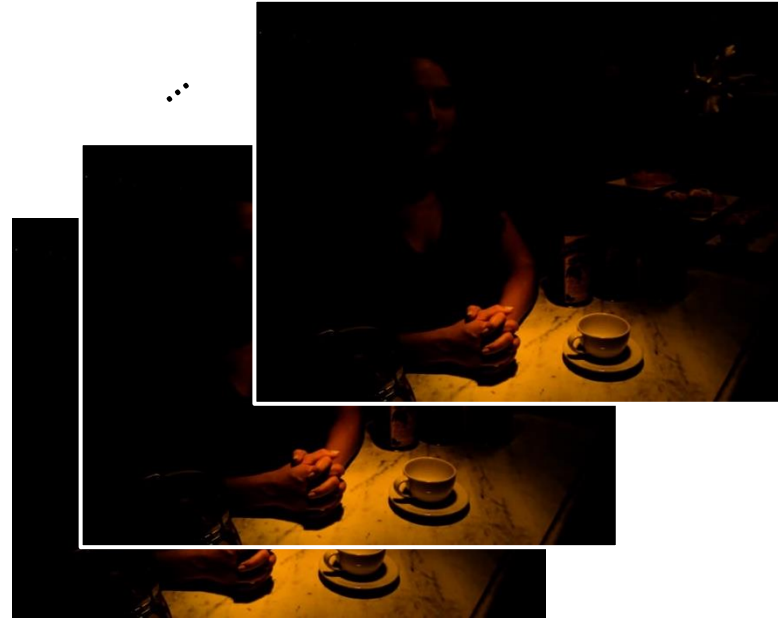
[Debevec et al., 2007]
[Gallo and Sen 2016]

From long exposure to burst photography



Long-exposure

Motion blur



Burst photography

burst of short exposed frames

HDR+



S. Hasinoff et al, "Burst photography for high dynamic range and low-light imaging on mobile cameras ", SIGGRAPH Asia 2016.

Night sight



O. Liba, et al., "Handheld Mobile Photography in Very Low Light", SIGGRAPH Asia 2019.

Multiple captures also helps to remove reflection



Images with reflection



Reflection-free image

Reflection removal using stereo input



2 frames from stereo camera

Output

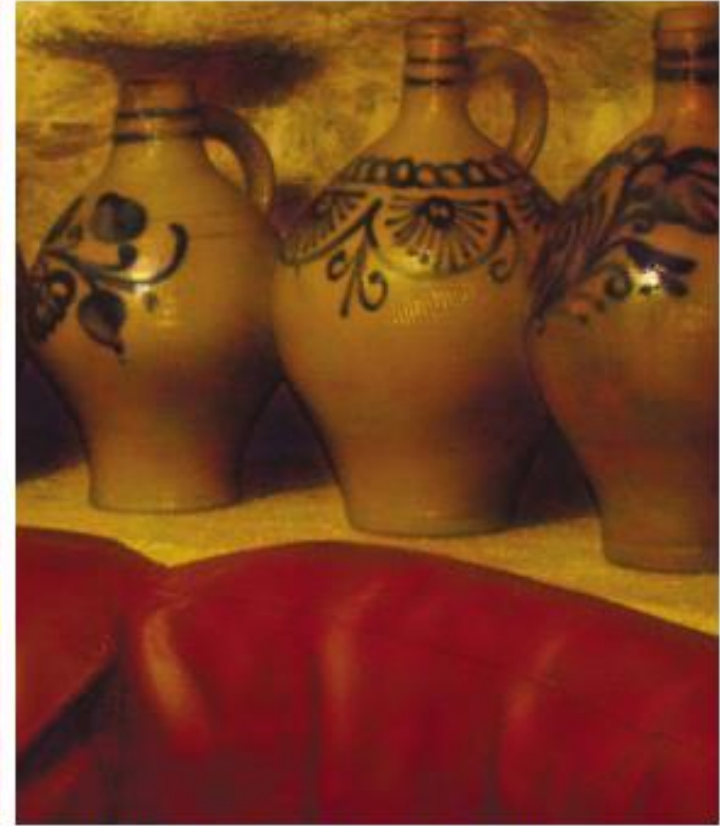
Flash / Non-flash Photography



Flash



No-Flash



Detail Transfer with Denoising

G. Petschnigg et al., "Digital Photography with Flash and No-Flash Image Pairs", SIGGRAPH 2004.

Stereoscopic Dark Flash for Low-light Photography



RGB

IR (infrared)

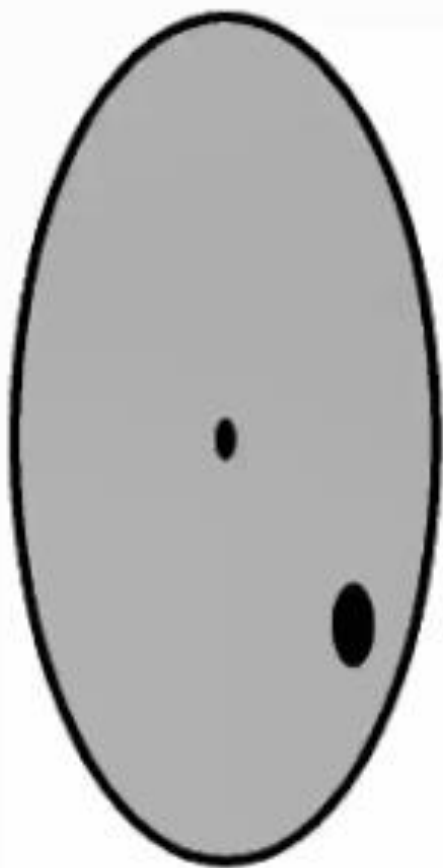


Merged result

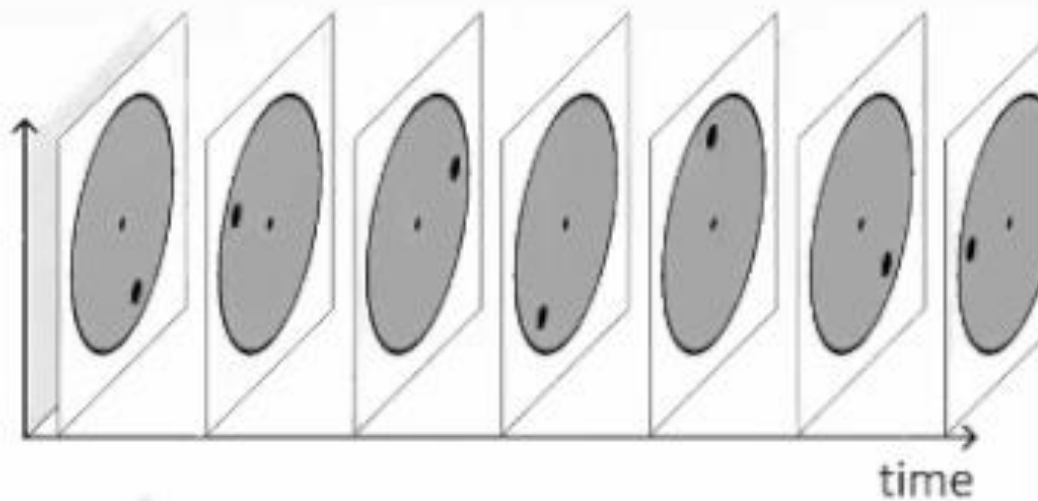
Jian Wang, Tianfan Xue, Jonathan T. Barron, Jiawen Chen

ICCP 2019

Event camera



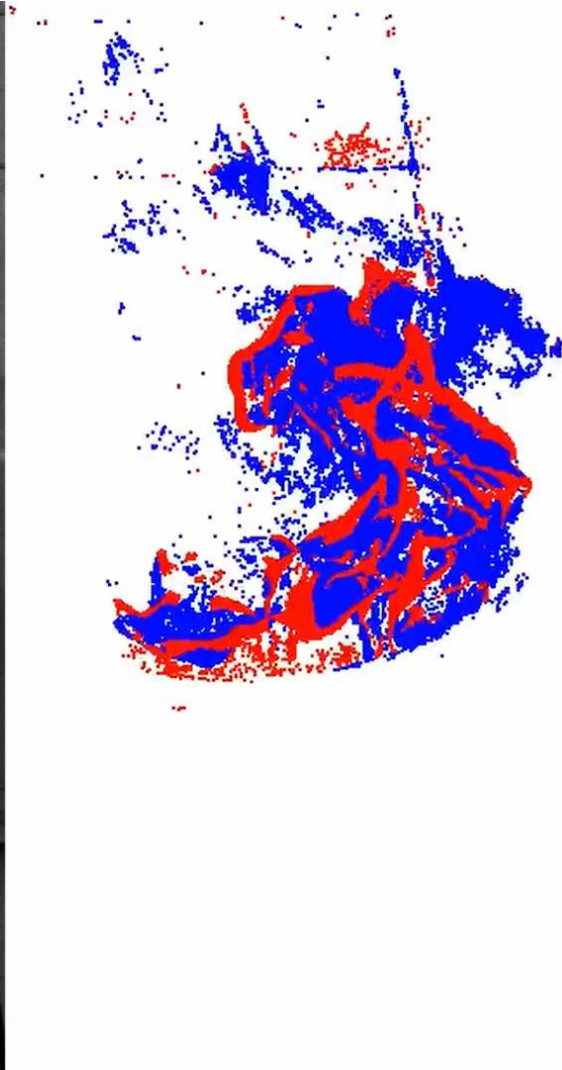
**standard
camera
output:**



**DVS
output:**

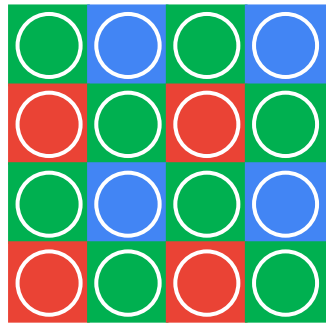


Use event camera to recover high-speed motion

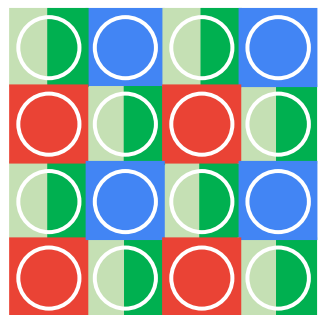


S. Tulyakov et al., "Time Lens: Event-based Video Frame Interpolation", CVPR 2021.

Depth and deblurring from DP images

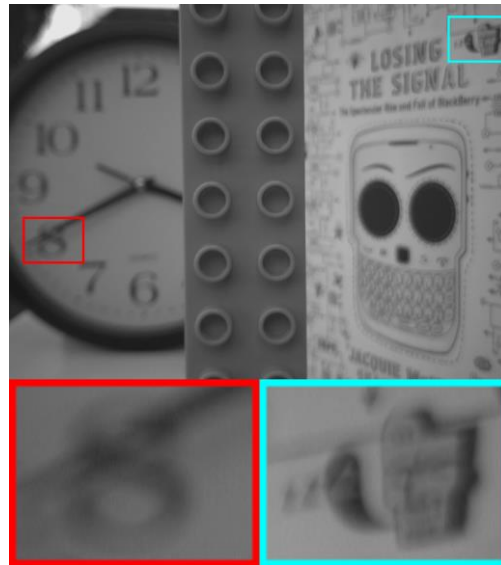


Regular sensor

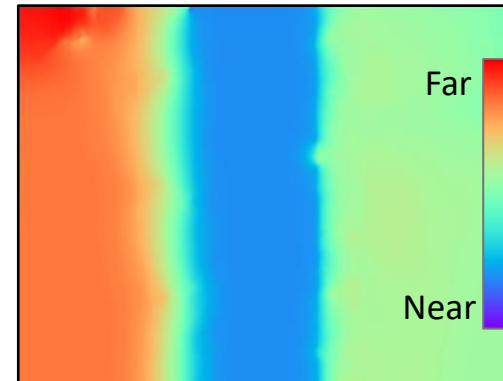


DP sensor

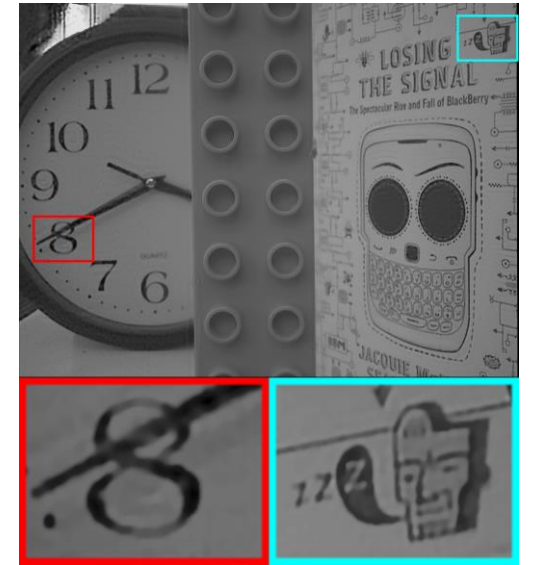
Input DP image



Depth map



All-in-focus image



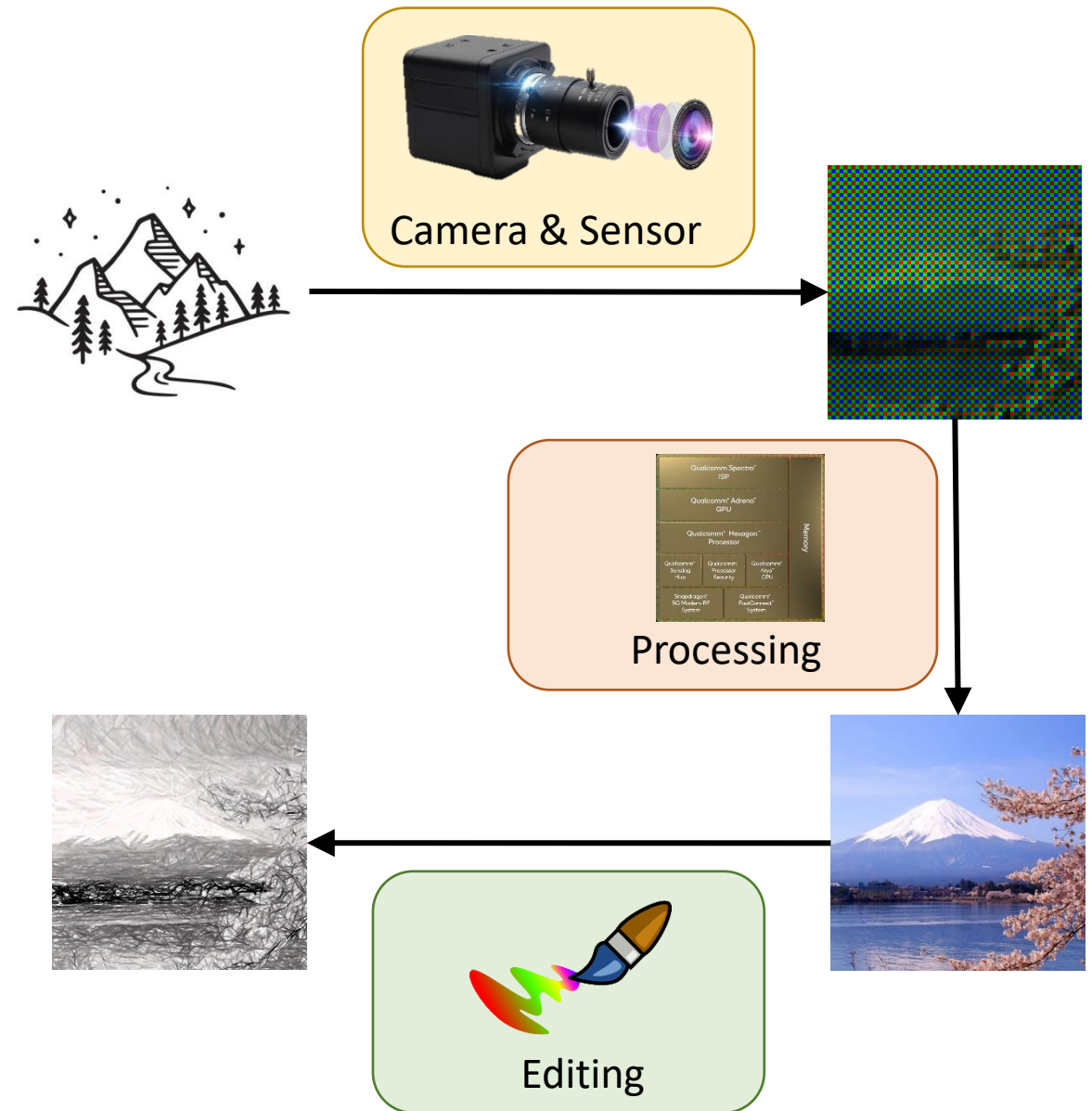
Defocus Map Estimation and Deblurring from a Single Dual-Pixel Image

S. Xin, N. Wadhwa, T. Xue, J. T. Barron, P. P. Srinivasan, J. Chen, I. Gkioulekas, R. Garg

ICCV 2021

Overview

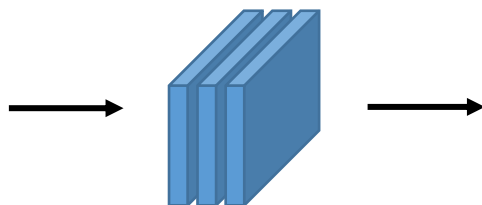
- **Capturing:** multiple source fusion
- **Processing & editing**
 - **Training data:** synthetic data
 - **Network:** combine classic image processing algorithm and machine learning



An input/output pair is needed for ML training



Noisy input

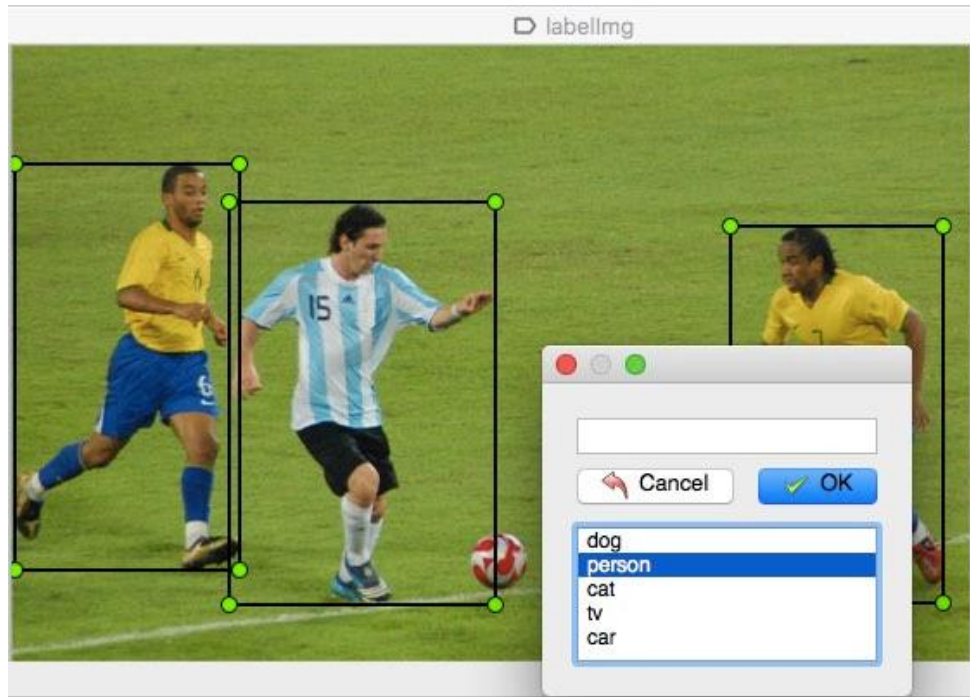


Neural network



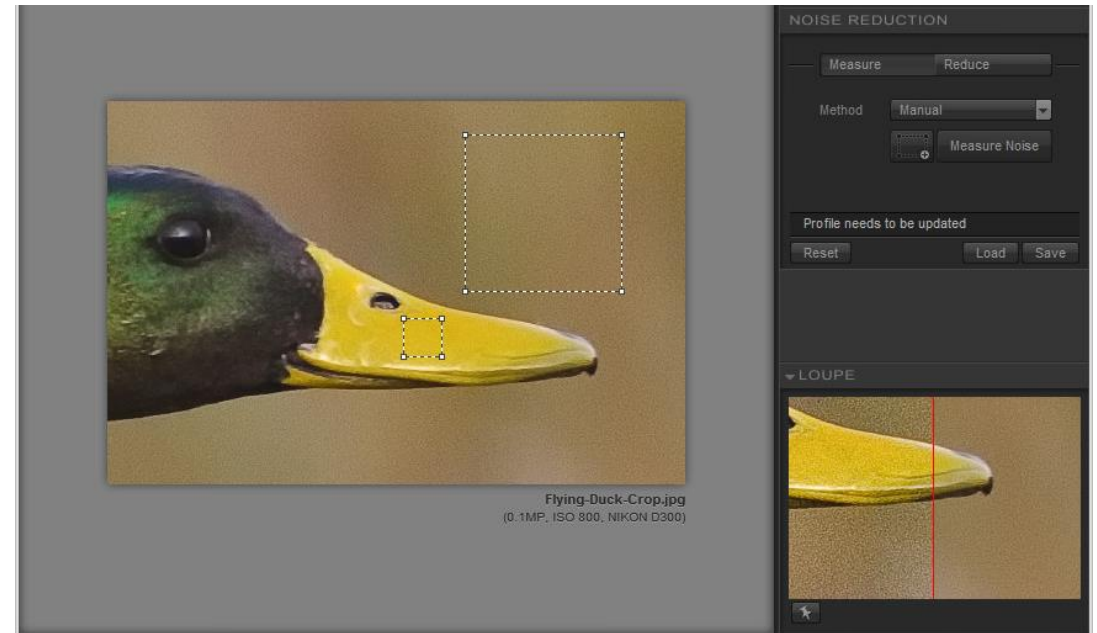
Denoised output

Ground truth output are hard to label



Labeling detection is **easy**.
few seconds / image

Image Credit: <https://github.com/tzutalin/labelImg>



Label denoising is **hard**:
few hours / image

Image Credit: Nik Collection

Capturing ground truth requires a lot of manual efforts



Collecting ground truth for denoising (<100/day)

Can we use images on the web



<100 images / day



No. of images uploaded to internet:
3,000,000,000,000 images / day

by Leon Seibert, Unsplash

Apply degeneration to images on the web

Input/output pairs



Clean images from web

Synthesize
degeneration



Degenerated images

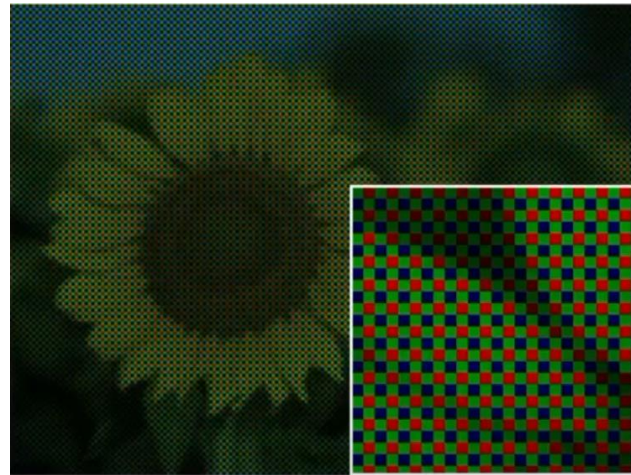
Processing
network



Recovered clean image

How to generate realistic degeneration?

Real noise does not directly apply to sRGB

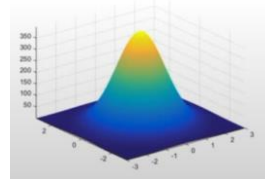


Camera pipeline



Raw image

sRGB image



Noise

$$y \sim \mathcal{N}(\mu = x, \sigma^2 = \lambda_{read} + \lambda_{shot}x)$$

Noise model



Synthesize raw from sRGB

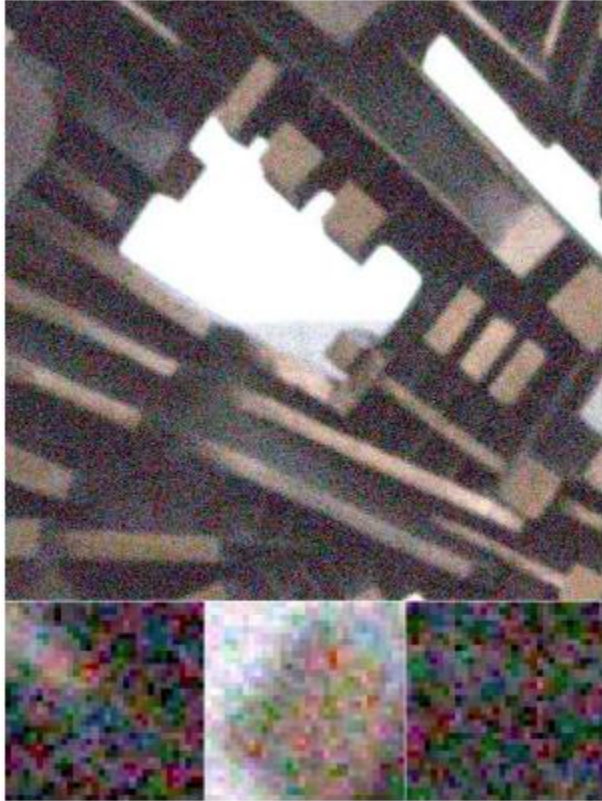
“Reprocessing”



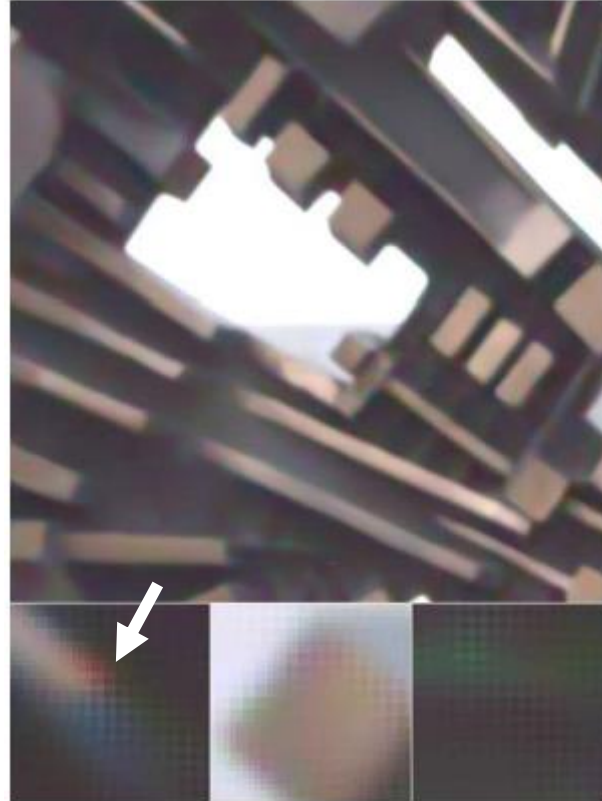
“Unprocessing”



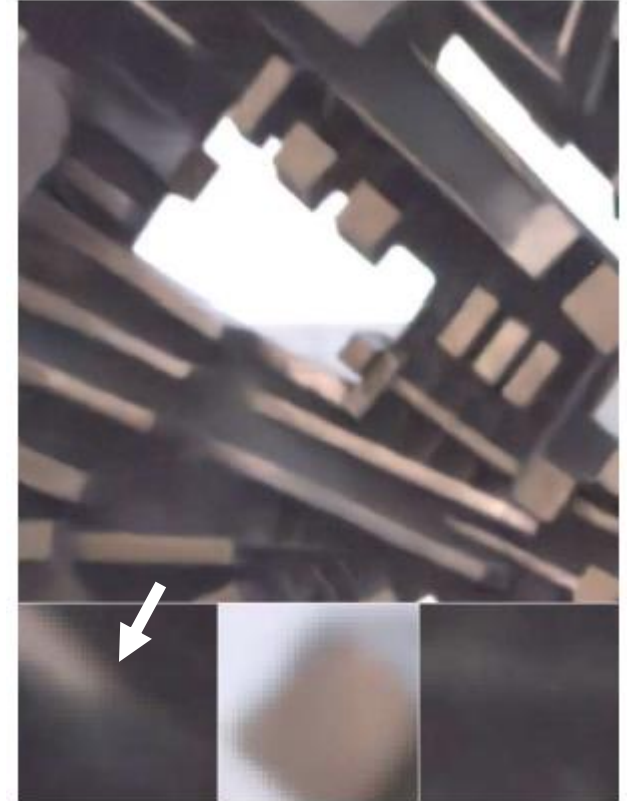
Unprocess improves the image quality



Noisy input



N3Net



Ours

Simulate realistic rain drops

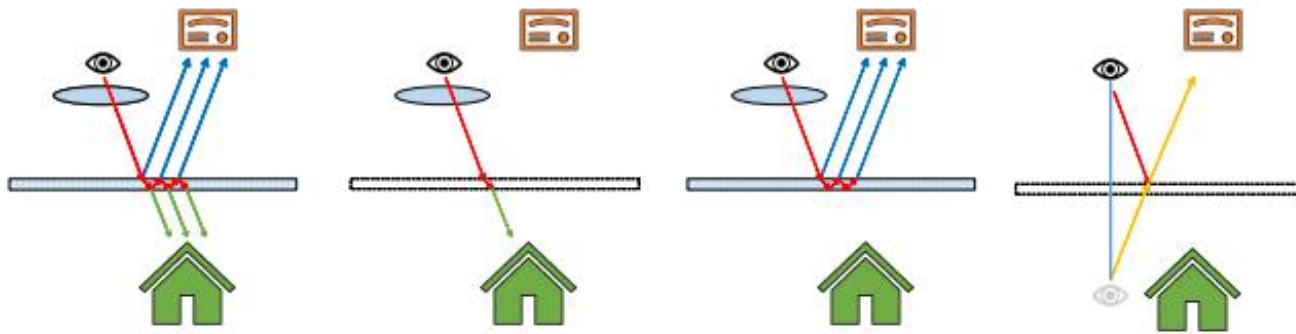


(a) Heavy rain.



(b) Rain accumulation.

Sometimes, it is important to understand 3D geometry in the simulation



RGB image



Syn. Depth



Syn. Mesh



(1) I



(2) T



(3) \tilde{R}



(4) R

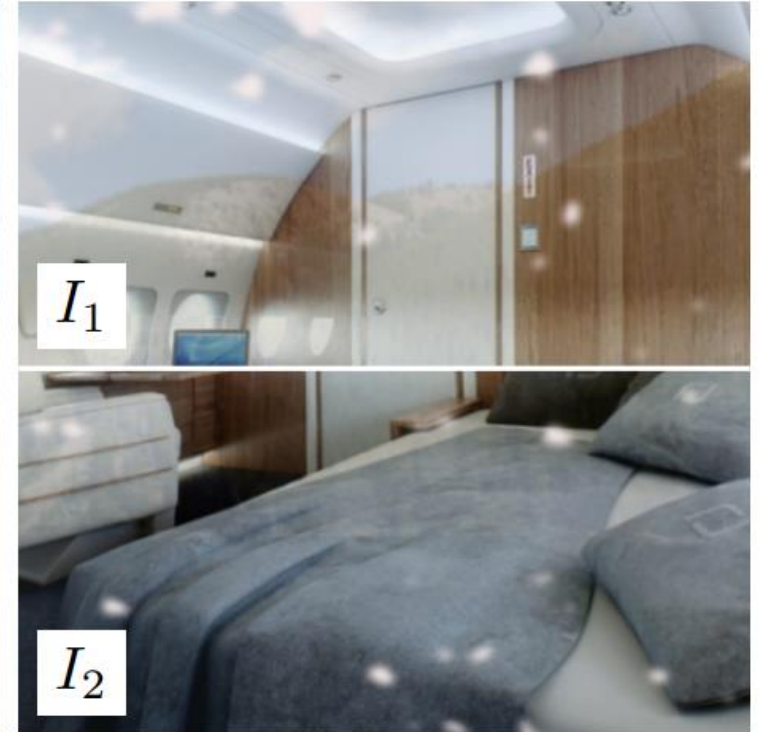
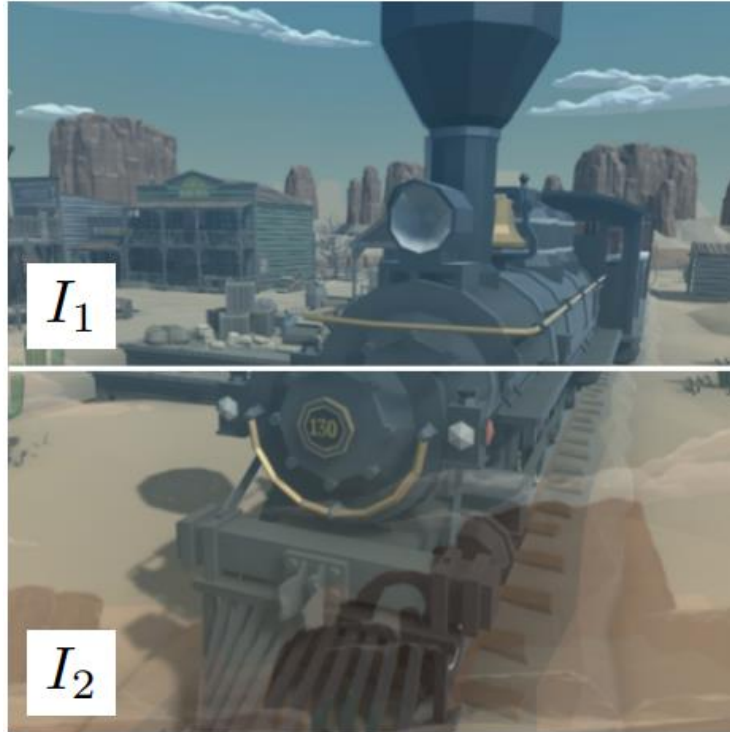
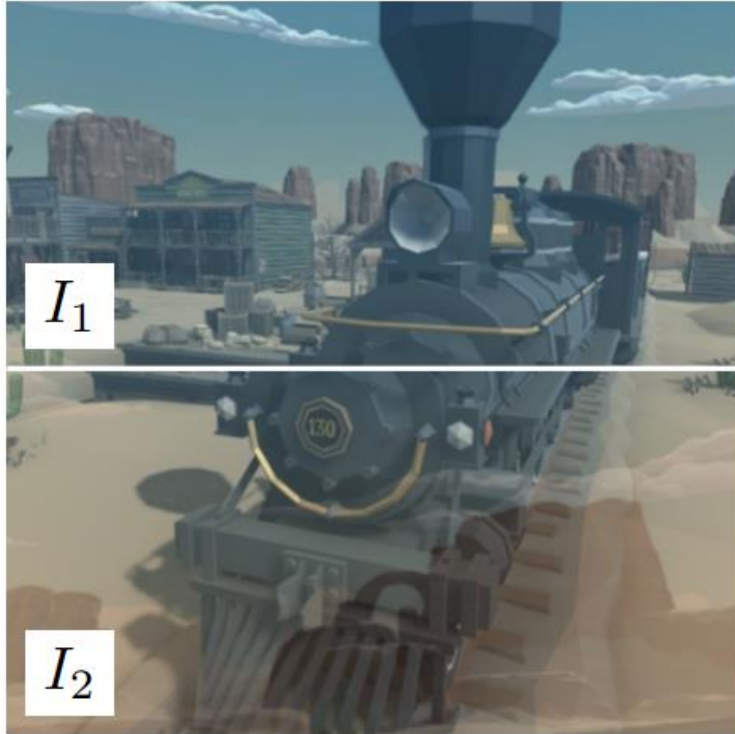
Symbol	Definition
T	Front scene image
R	Back scene image
\tilde{R}	Back scene image reflected by a glass
X^*	Predicted image of X
T/R	T or R

We can even resort to rendering engine



S. Niklaus et al., "Learned dual-view reflection removal", WACV, 2021.

We can even resort to rendering engine

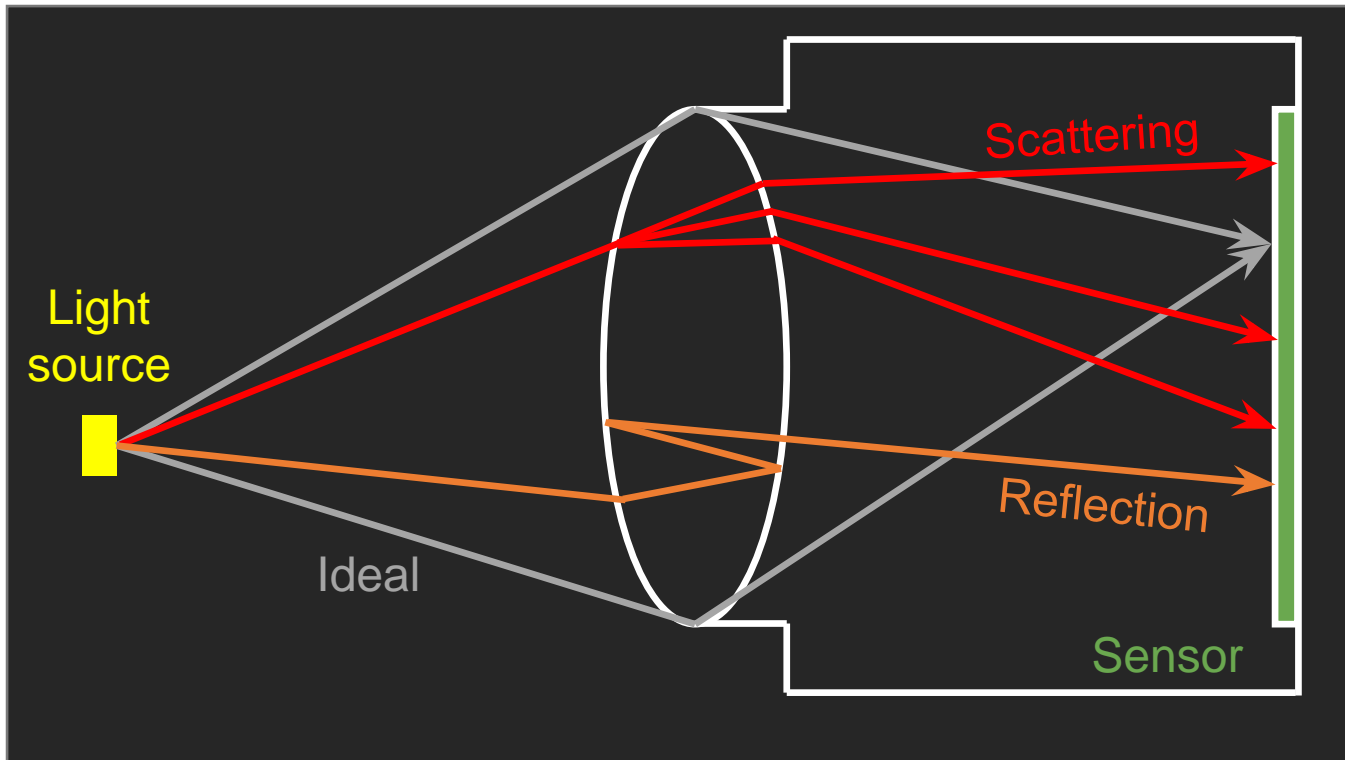


S. Niklaus et al., "Learned dual-view reflection removal", WACV, 2021.

Lens flare

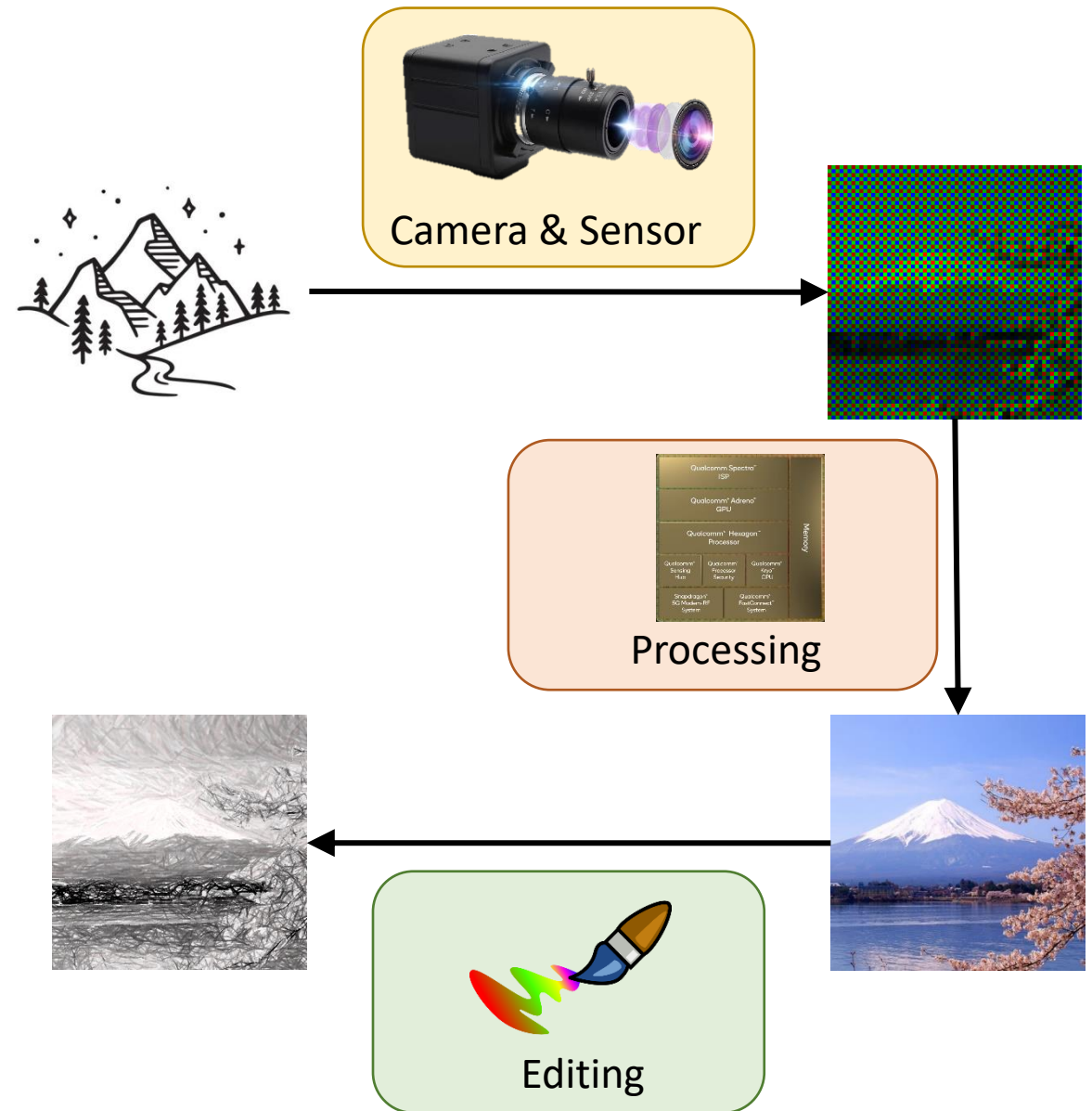


Flare formation

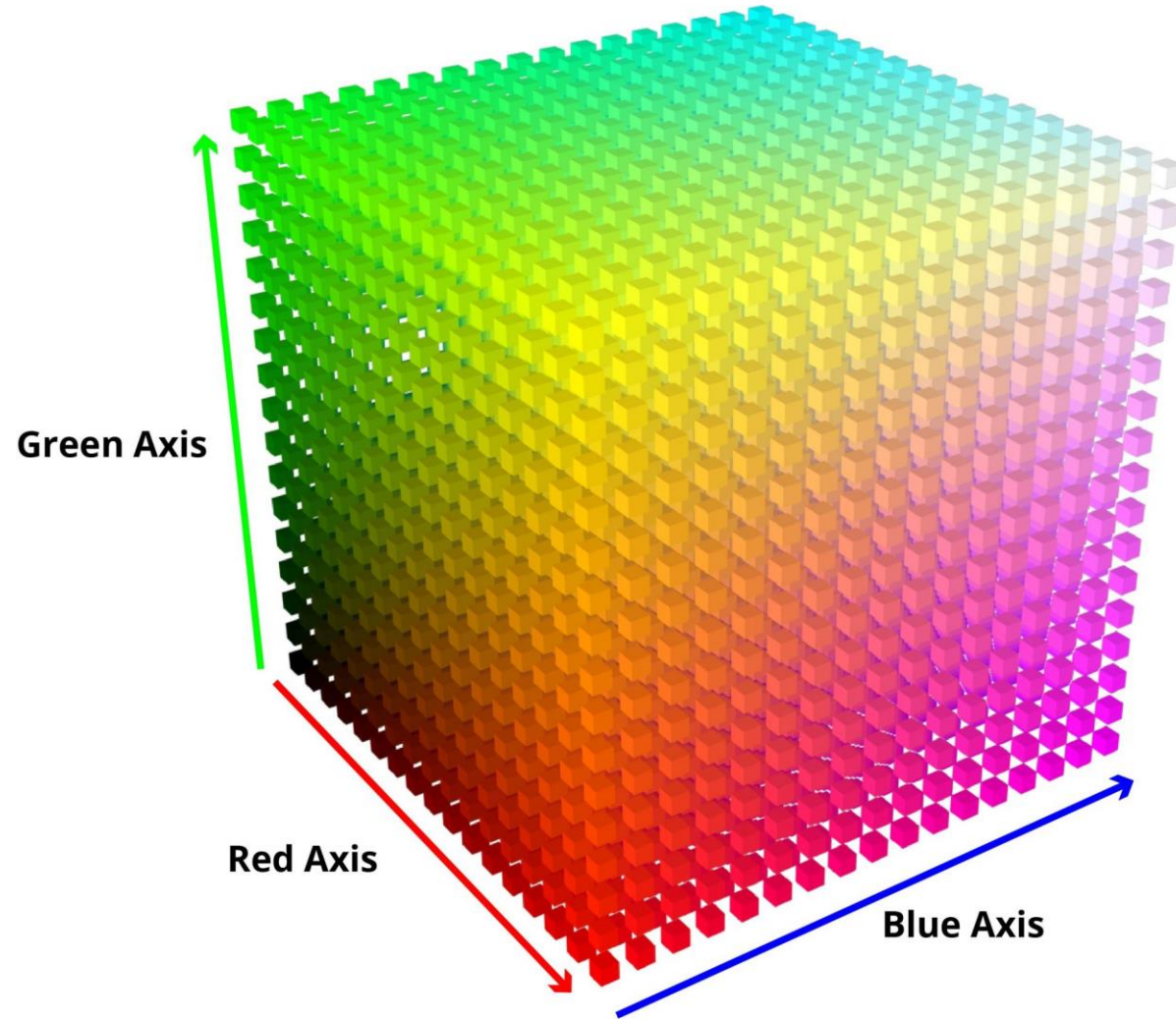


Overview

- **Capturing:** multiple source fusion
- **Processing & editing**
 - **Training data:** synthetic data
 - **Network:** combine classic image processing algorithm and machine learning

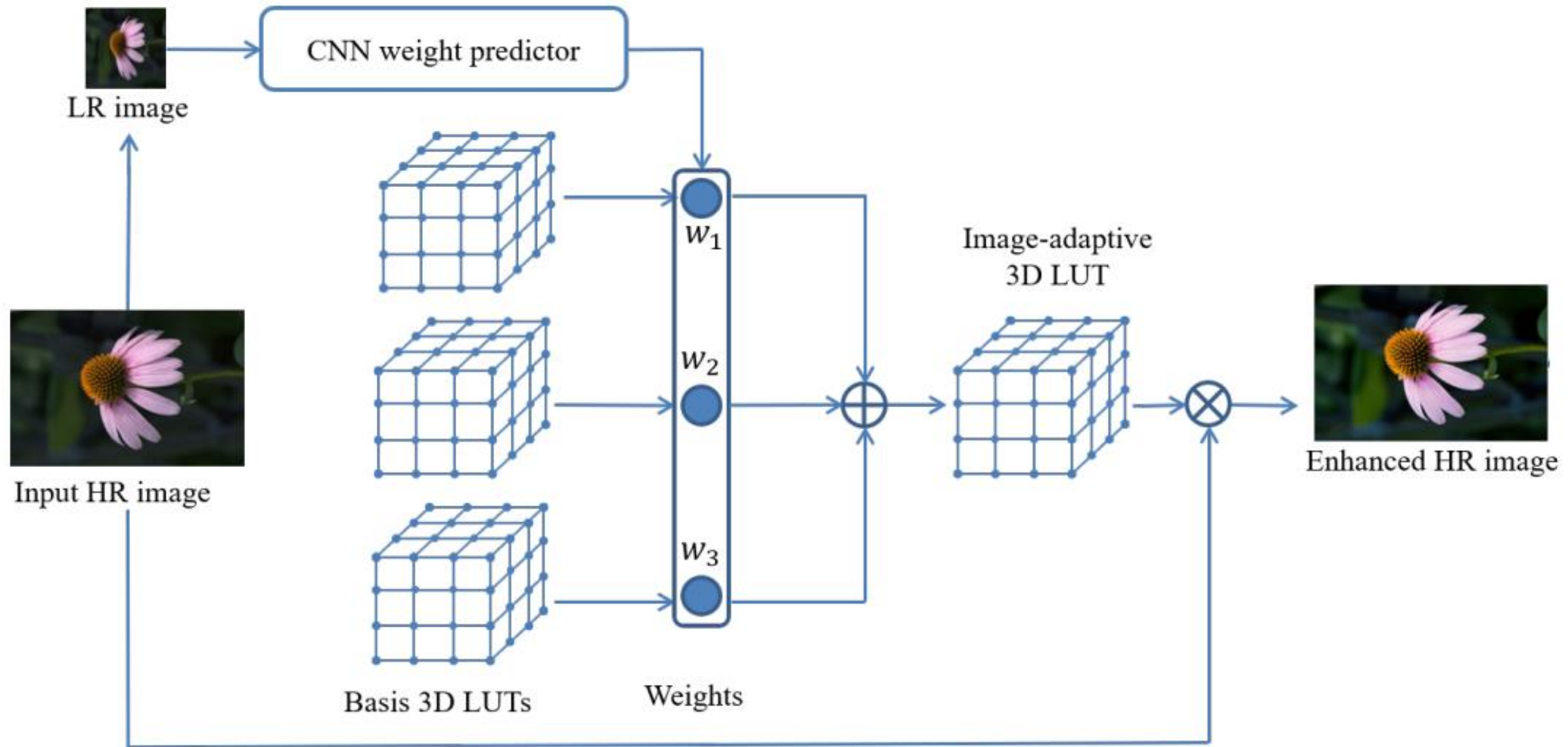


Representing point operation: 3D LUT



<https://www.bromptontech.com/what-is-a-3d-lut/>

Learning to enhance -> Learning 3D LUT

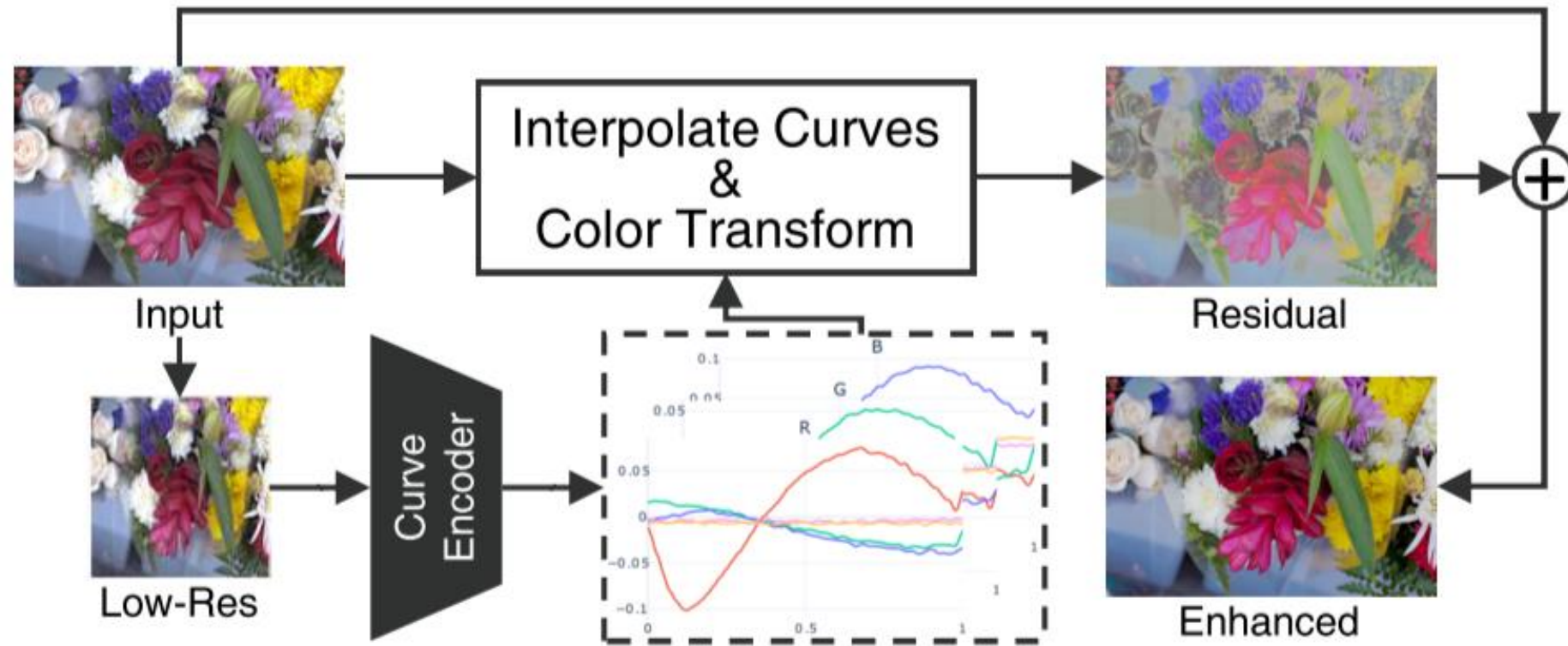


H. Zeng et al., "Learning Image-adaptive 3D Lookup Tables for High Performance Photo Enhancement in Real-time", T-PAMI, 2020

Learning 3D LUT significantly reduces the time cost

Resolution	1920×1080	3840×2160	6000×4000
Pix2Pix [49]	1.2e2	N.A.	N.A.
CycGAN [50]	5.6e2	N.A.	N.A.
DPE [7]	8.6e1	N.A.	N.A.
White-Box [9]	5.0e3	9.1e3	2.0e4
Dis-Rec [8]	2.5e4	1.1e5	3.3e5
UIE [11]	1.0e4	2.0e4	3.3e4
HDRNet [2]	4.5e1	2.1e2	5.9e2
UPE [10]	4.5e1	2.1e2	5.9e2
Ours	0.64	1.66	3.76

Image Enhancement: Using color tran. and global curve



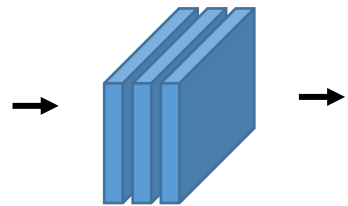
Direct prediction is expensive



Style image



Input



Neural network



Stylized output

[PhotoWCT: Li et al., ECCV 2018]

[WCT2: Yoo et al., ICCV 2019]

[LST: Li et al., CVPR 2019]

All of them are OOM when applied to 4MP image

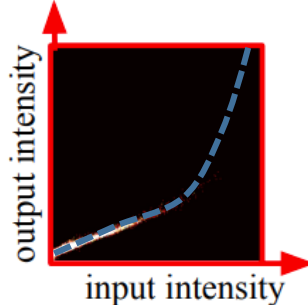
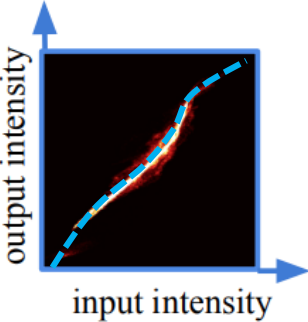
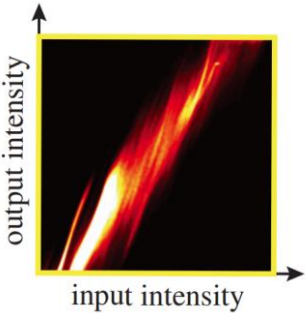
Can we approximate it using tone curves?



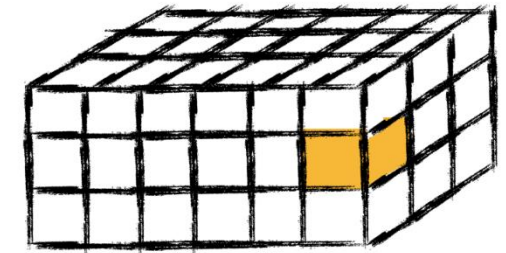
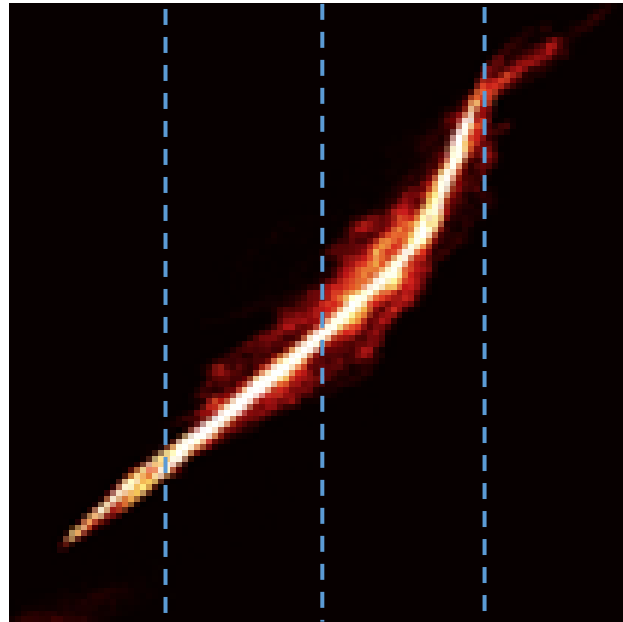
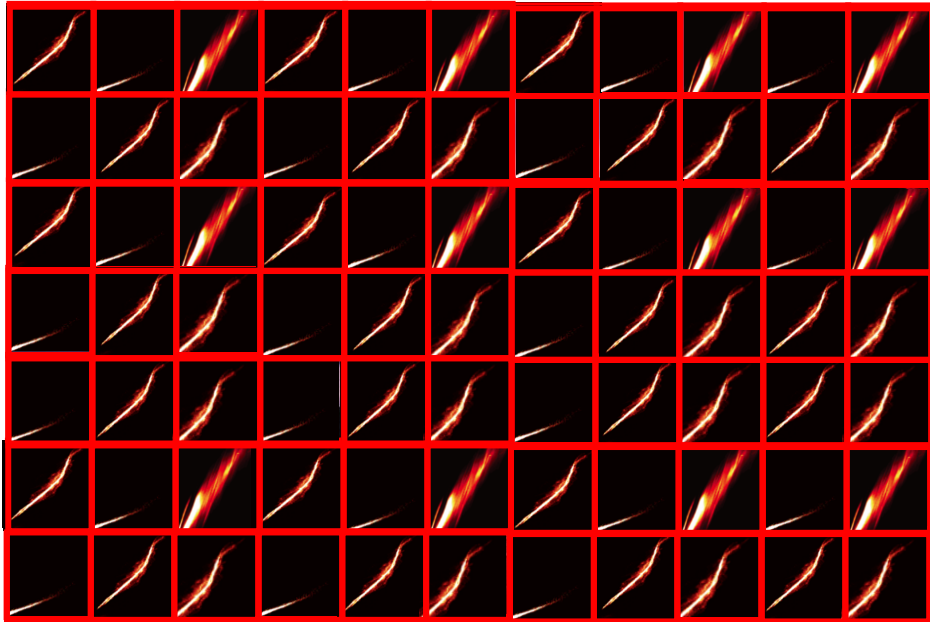
Input



Stylized output



Model a set of tone curves as bilateral grid



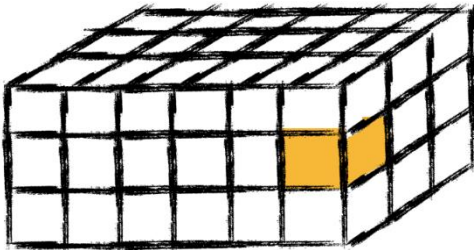
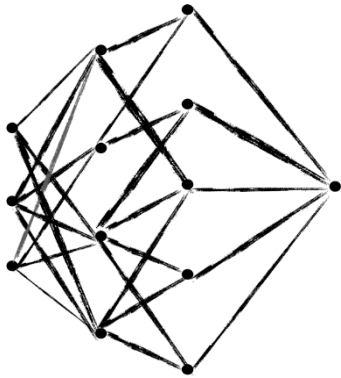
$$\begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \end{bmatrix} \begin{bmatrix} r \\ g \\ b \\ 1 \end{bmatrix}$$

Bilateral grid

J. Chen, A. Adams, N. Wadhwa, S. Hasinoff, "Bilateral guided upsampling", 2017

M. Gharbi, J. Chen, J. Barron, S. Hasinoff, F. Durand, "Deep Bilateral Learning for Real-Time Image Enhancement", SIGGRAPH 2017

Style transfer using a set of tone curves



$$\begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \end{bmatrix} \begin{bmatrix} r \\ g \\ b \\ 1 \end{bmatrix}$$

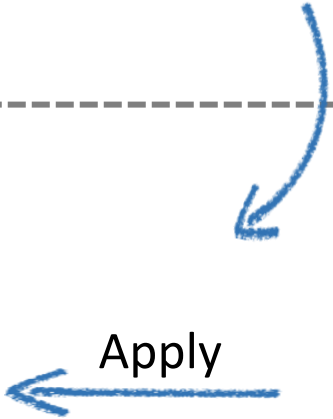
Tone curves baked in bilateral grid

Low resolution: 256x256

Full resolution: 4Kx3K



Output



Input

Performance

Image Size	PhotoWCT	LST	WCT ²	Ours
512 × 512	0.68s	0.25s	3.85s	< 5 ms
1024 × 1024	1.51s	0.84s	6.13s	< 5 ms
1000 × 2000	2.75s	OOM	10.94s	< 5 ms
2000 × 2000	OOM	OOM	OOM	< 5 ms
3000 × 4000	OOM	OOM	OOM	< 5 ms

Latency

Mean Score	PhotoWCT	LST	WCT ²	Ours
Photorealism	2.02	2.89	4.21	4.14
Stylization	3.10	3.19	3.24	3.49
Overall quality	2.23	2.84	3.60	3.79

User study of visual quality





Results on 12MP image

HDRnet tonemapping



12 megapixel 16-bit linear input
(tone-mapped for visualization)



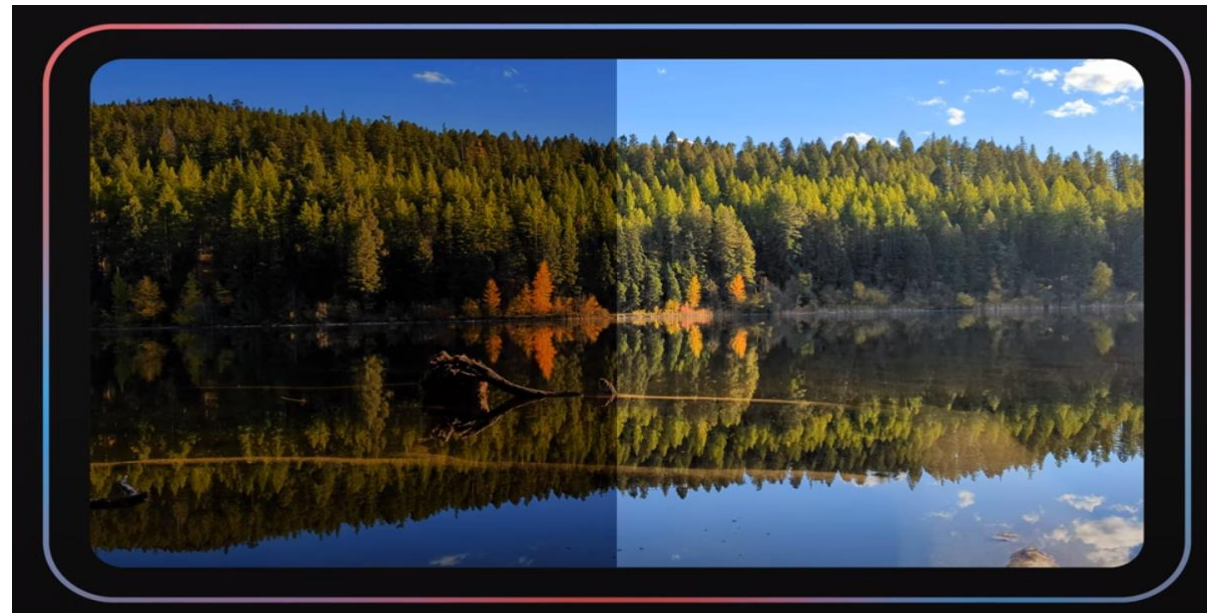
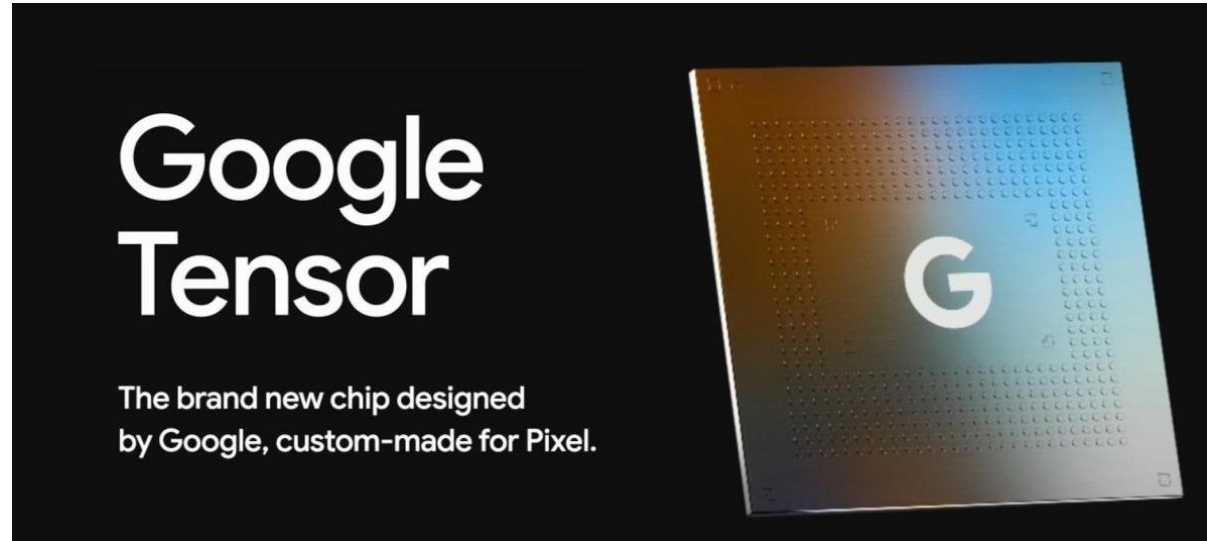
tone-mapped with HDR+
400 – 600 ms



processed with our algorithm
61 ms, PSNR = 28.4 dB

M. Gharbi, J. Chen, J. Barron, S. Hasinoff, F. Durand, “ Deep Bilateral Learning for Real-Time Image Enhancement”, SIGGRAPH 2017

Used by Google Tensor Chip



Denoising using spatially varying kernels



Input



MalleConv

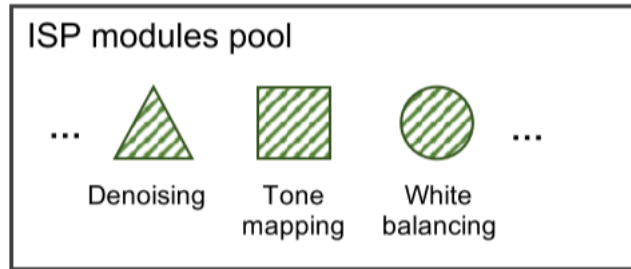


Selected kernel 1

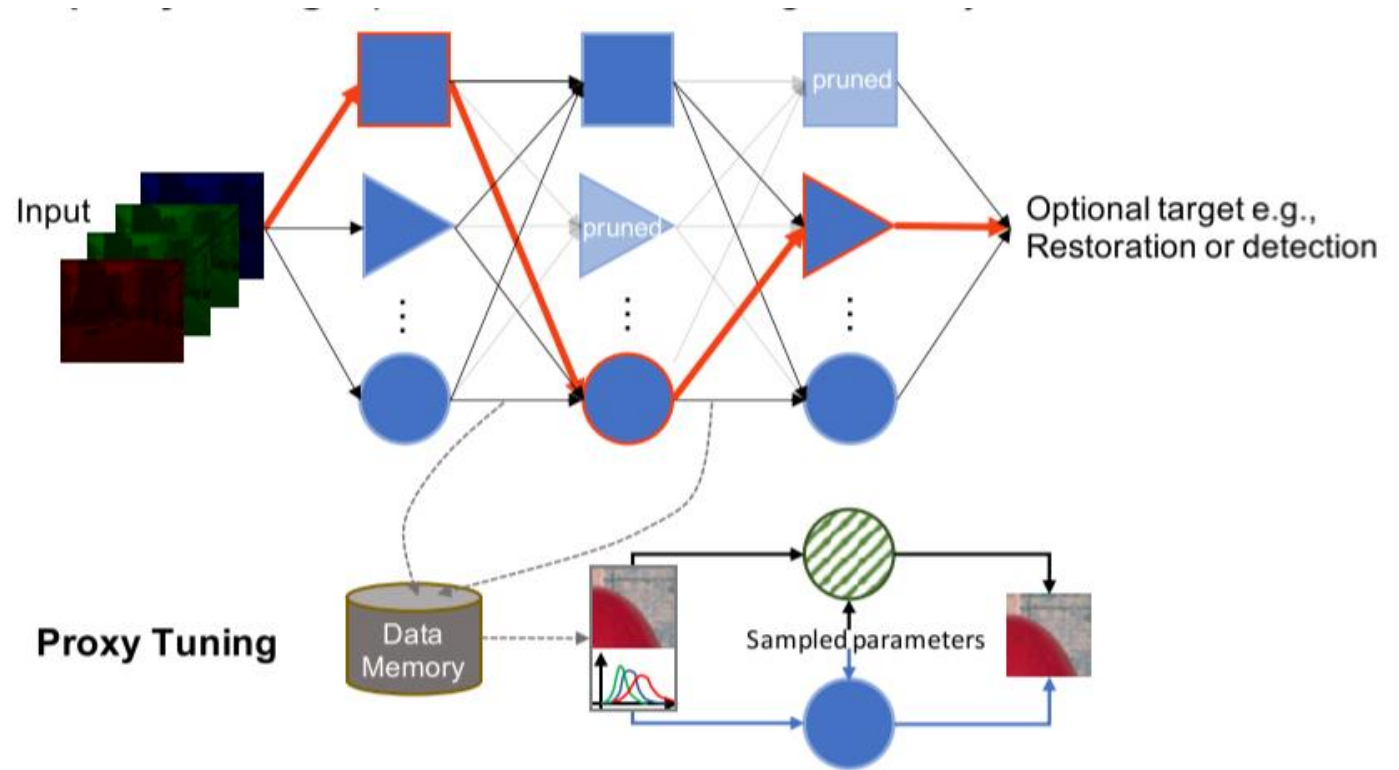


Selected kernel 2

We can even learn to reorder different modules

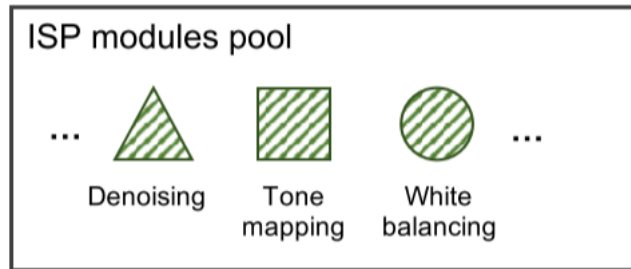


Basic camera modules

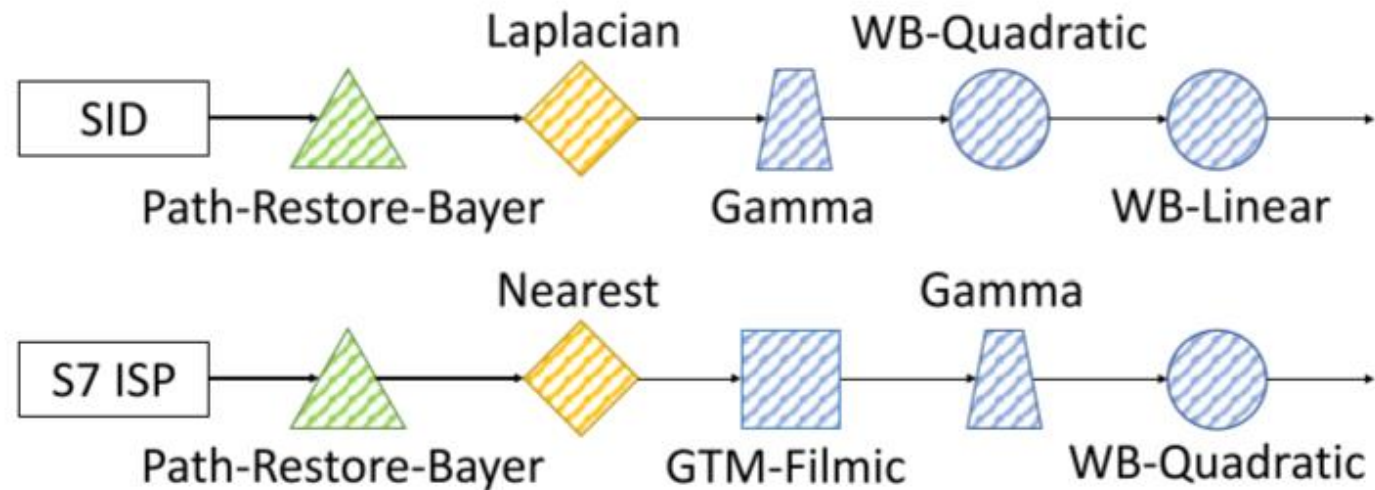


Learn a task-specific pipeline

We can even learn to reorder different modules



Basic camera modules



Different task may need different pipelines

Future smart cameras research



Simulation is important to collect training data

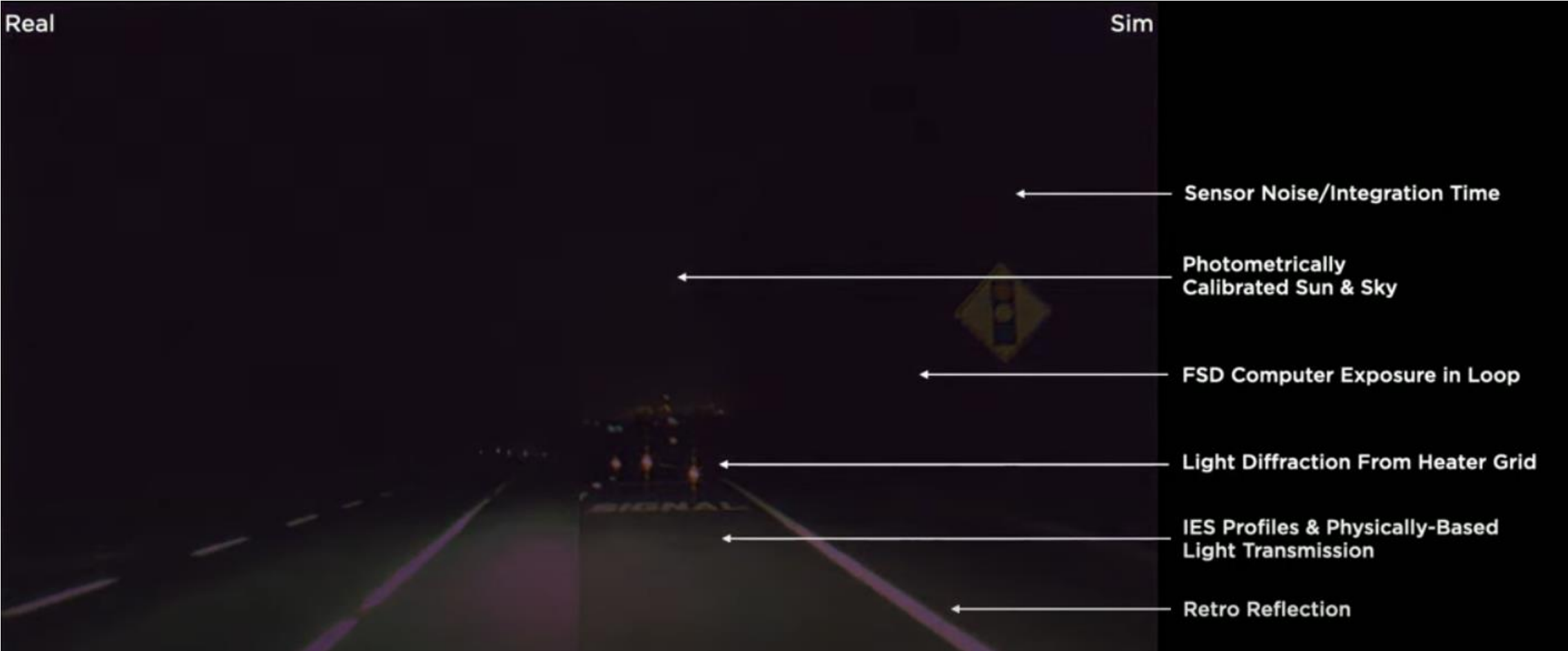
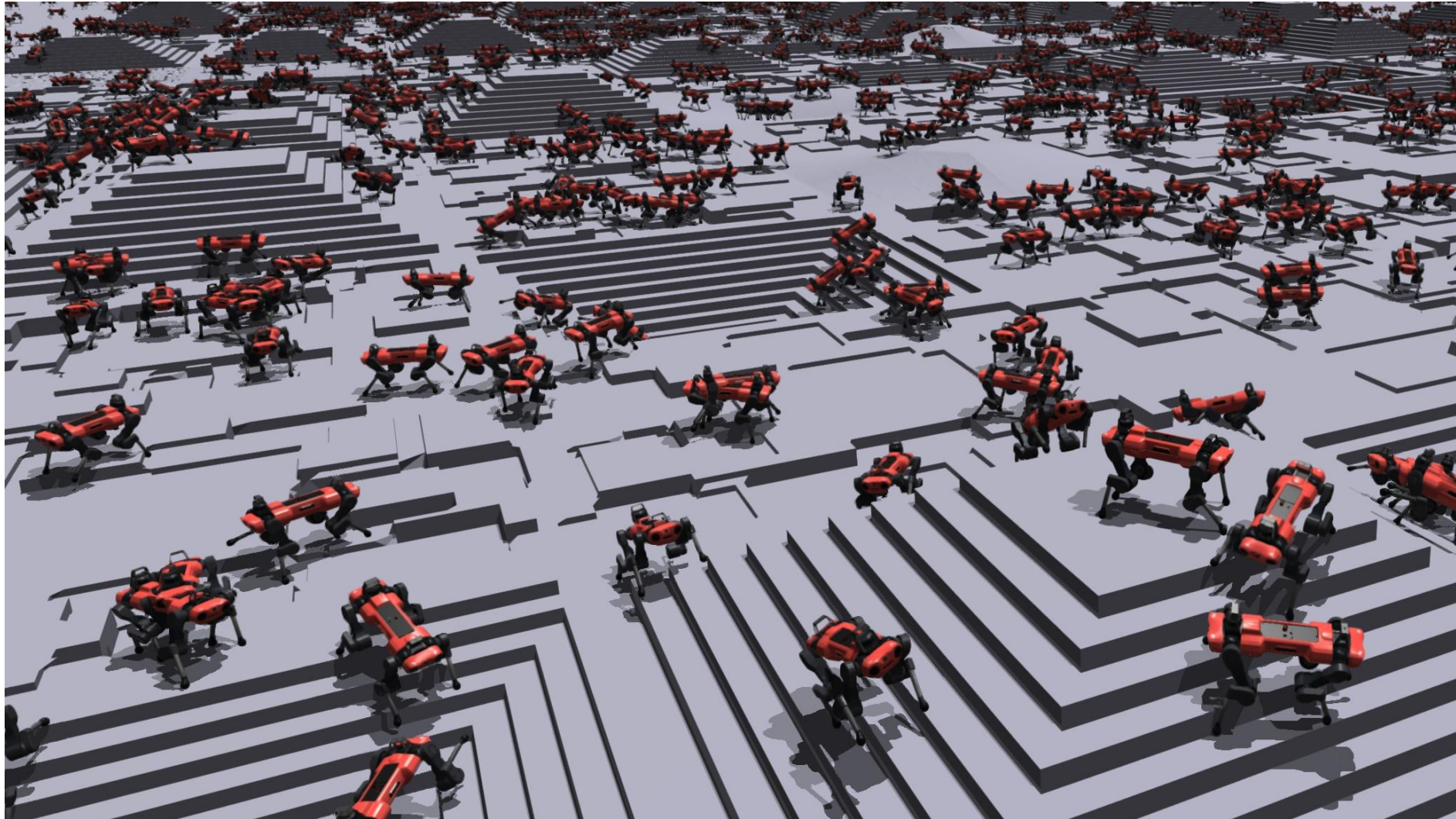


Image credit: Tesla AI Day

Is there Isaac Gym for computational photography?



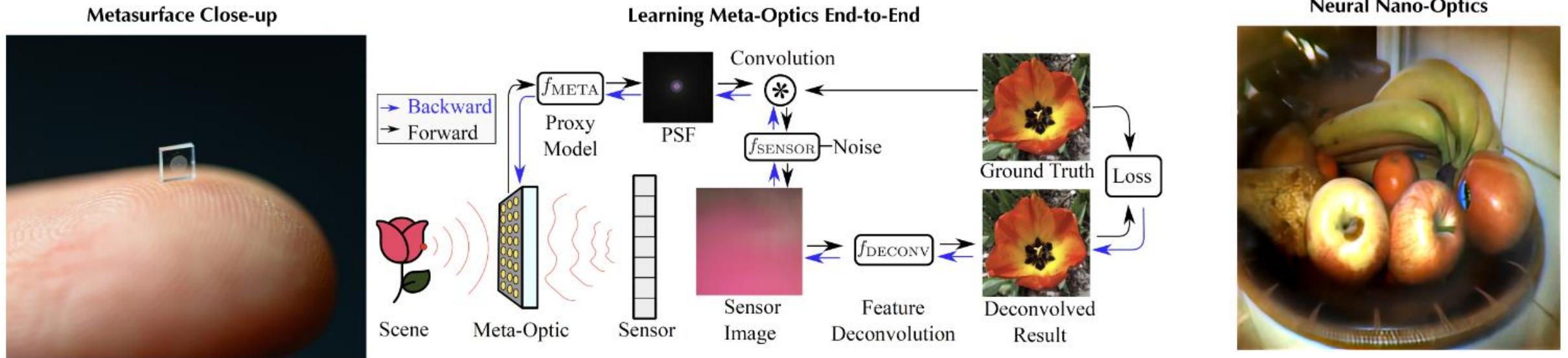
Issac Gym by NVIDIA, for robotic algorithm design

Computational Photography and Hardware



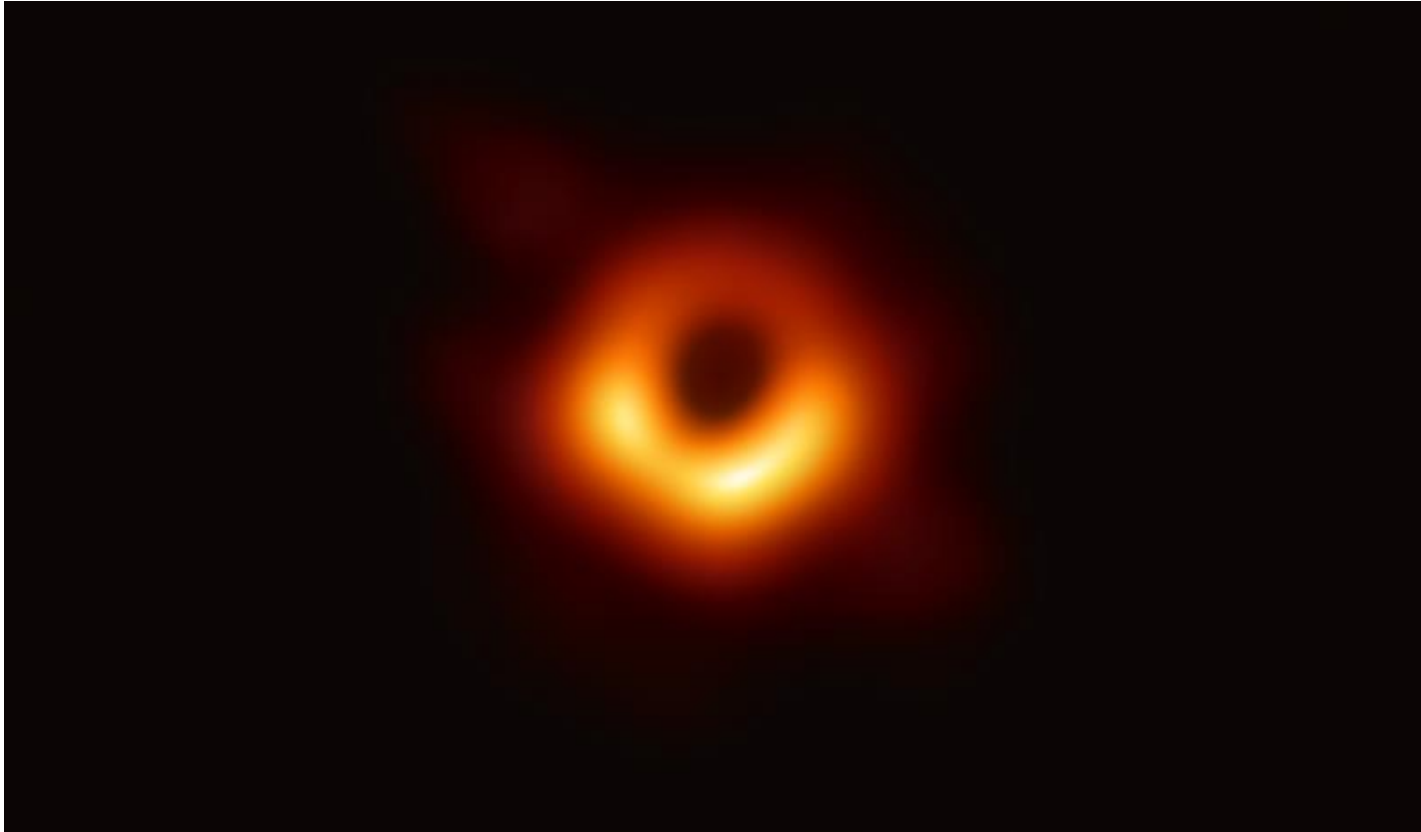
Image credit: <https://bit.ly/2mmFtKP>

Joint lens and algorithm design

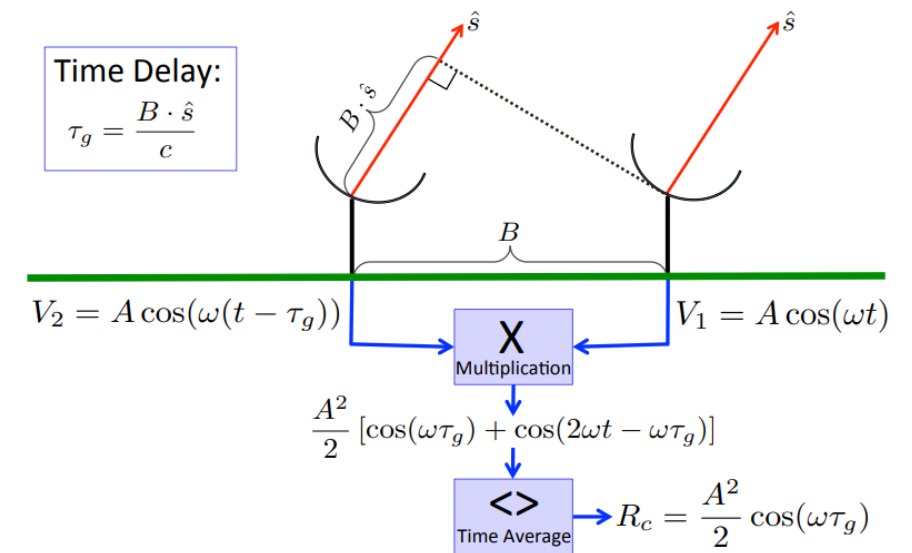


Tseng et al., “Neural nano-optics for high-quality thin lens imaging”, 2021

Camera is not only for better selfies



First black hole image
[image credit: NASA]



VLBI image formation

How VR may impact us

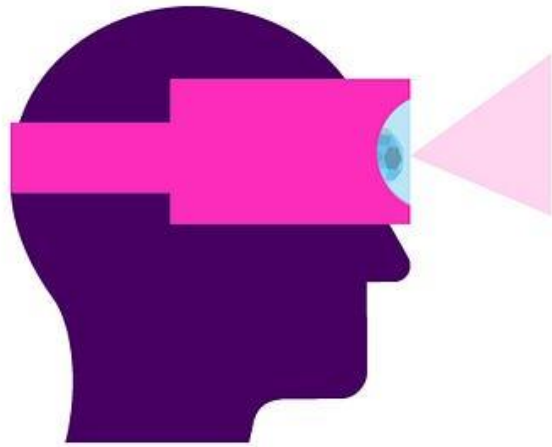


Can we capture more 3D content using our cameras?

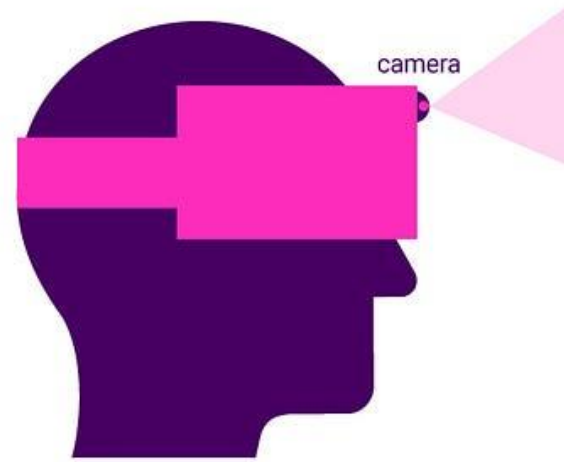


B. Mildenhall et al., "NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis", ECCV, 2020

Can cameras be as good as our eyes?



Optical see through displays



Video see through displays

Can our algorithms be as fast as to deal with 8K60fps?





AIGC and Computational Photography



Where is the boundary between **editing** and **synthesis**



Which one is real, which one is fake?

<https://www.bilibili.com/video/BV16M4y1q7B5>

Where is the boundary between **editing** and **synthesis**



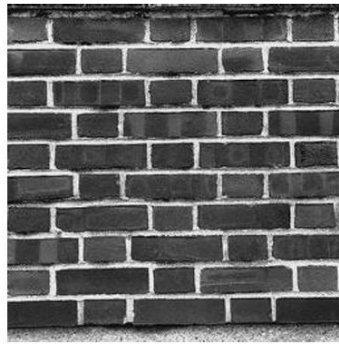
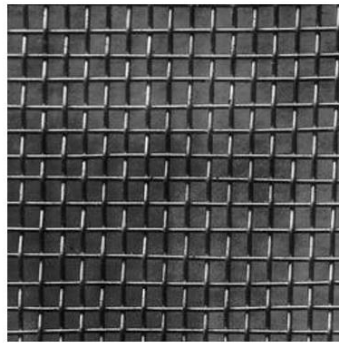
Robin et al., “High-Resolution Image Synthesis with Latent Diffusion Models”, CVPR 22.

Where is the boundary between **editing** and **synthesis**



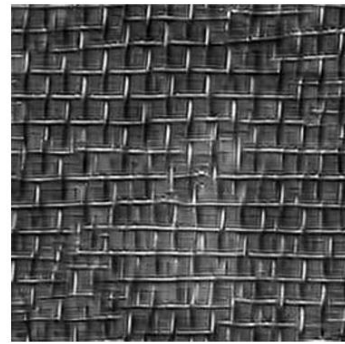
Generated by Midjourney v5: <https://petapixel.com/2023/03/17/midjourney-v5-creates-photorealistic-images-and-even-does-hands-correctly/>

Where is the boundary between editing and synthesis



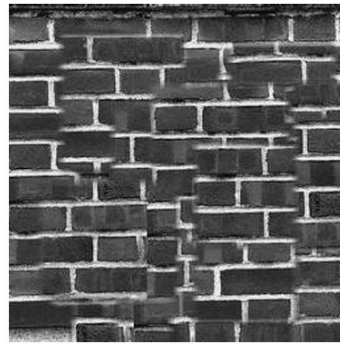
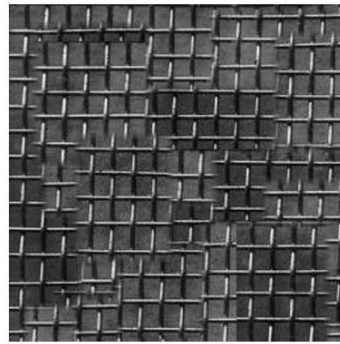
... of a visual cortical neuron—their
... describing the response of that neuro
... ht as a function of position—is perhap
... functional description of that neuron.
... seek a single conceptual and mathem
... describe the wealth of simple-cell recep
... id neurophysiologically¹⁻³ and inferred
... especially if such a framework has the
... it helps us to understand the functio
... leeper way. Whereas no generic mos
... ussians (DOG), difference of offset C
... rivative of a Gaussian, higher derivati
... function, and so on—can be expect
... mple-cell receptive field, we noneth

input texture



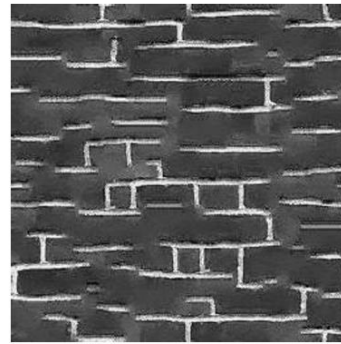
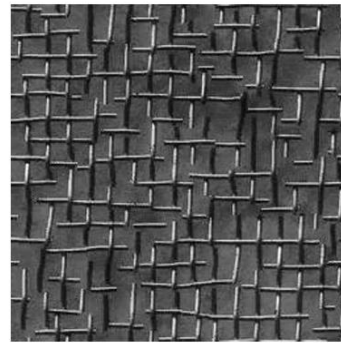
... hal m...
... and mathem...
... ht ap...
... funs and inferred...
... sd neurophysiol...
... especially if suc...
... id helps us to...
... eeper way. We...
... rissians (DOG)...
... leeper...
... ussi...
... rivat...
... fun...
... mplogically¹⁻³ an

Portilla & Simoncelli [17]



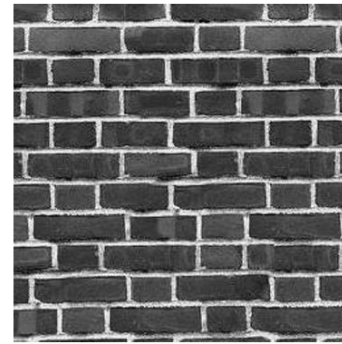
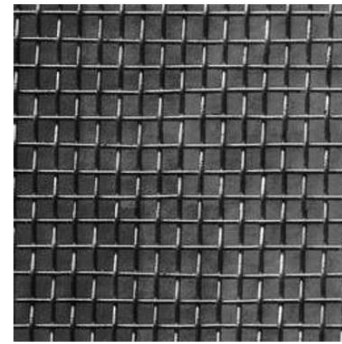
... de and so on...
... ht ap...
... funs and inferred...
... sd neurophysiol...
... especially if suc...
... id helps us to...
... eeper way. We...
... rissians (DOG)...
... leeper...
... ussi...
... rivat...
... fun...
... mplogically¹⁻³ an

Xu et al. [21]



... cols...
... o car...
... esoeao so...
... euogrs e...
... fy a...
... neisntu...
... nse...
... as hal...
... n...
... n si...
... tunnting...
... int...

Wei & Levoy [20]



... sition—is per...
... of that neur...
... ual and matheurophysiologically¹⁻³ and
... simple-cell pecially if such a framewor
... y¹⁻³ and inferlps us to understand th
... amework has perhay. Whereas no ge
... and the fumeuroi(DOG), difference o
... s no generic a single conceptual and m
... rence of offse the wealth of simple-ce
... , higher deriescribing the response of t
... —can be expes a function of position—
... helps us to understand thription of th
... per way. Whereas no gonceptual an
... sians (DOG), differencealth of simple

Image Quilting

A. Efros and W. T. Freeman, “Image quilting for texture synthesis and transfer”, CG 2001

Do we even need a powerful lens?

Only less powerful camera hardware is needed in the future?



iPhone SE

iPhone 12 mini

iPhone 7, 8, SE 2

iPhone X, XS, 11 Pro

iPhone 12, 12 Pro

iPhone 11, XR

iPhone 11 Pro Max, XS Max, 7 Plus, 8 Plus

iPhone 12 Pro Max

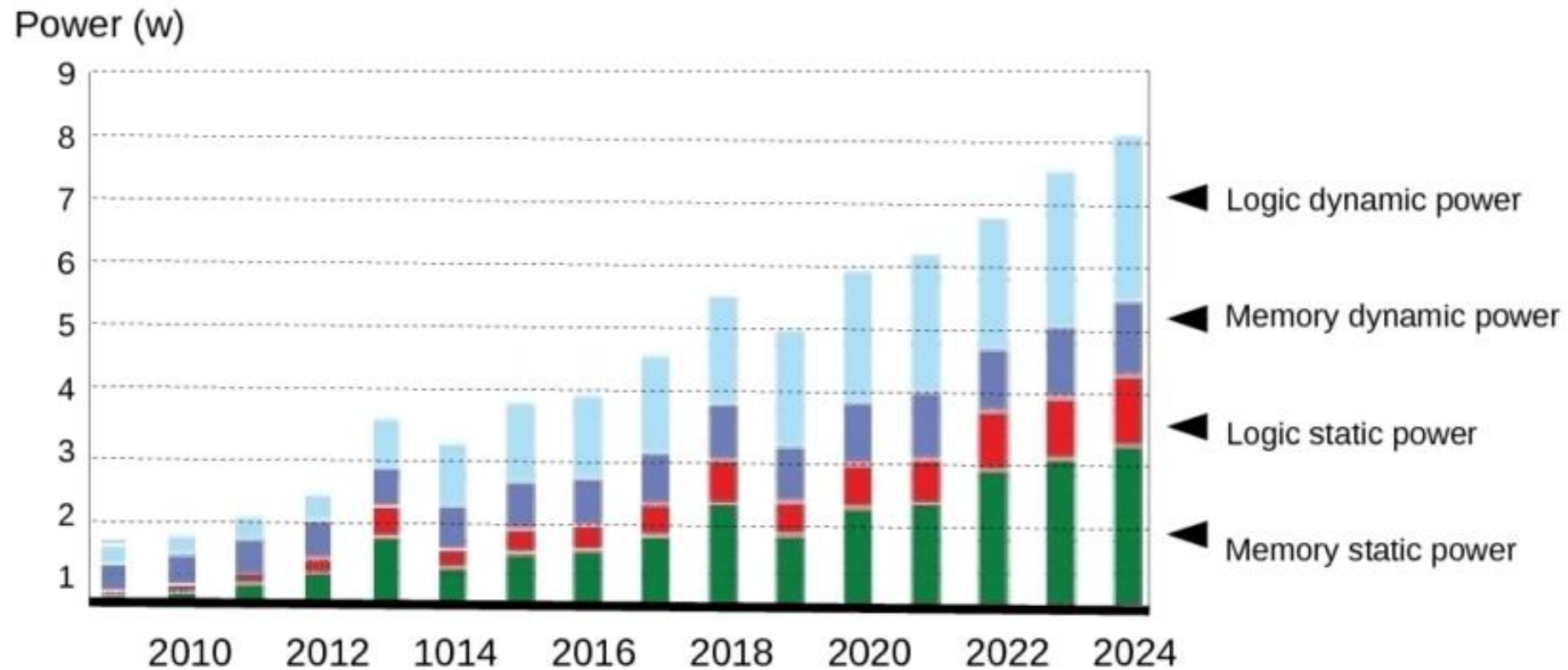


Cameras become **more powerful** in past 10y

Gap between
academic
research and
industrial design



Industrial not only care quality, but also speed & power



SoC consumer portable power consumption

Yahia Benmoussa, "Performance and Energy Consumption Characterization and Modeling of Video Decoding on Multi-core Heterogenous Mobile SoC and their Applications"

High PSNR != Better Image

bicubic
(21.59dB/0.6423)



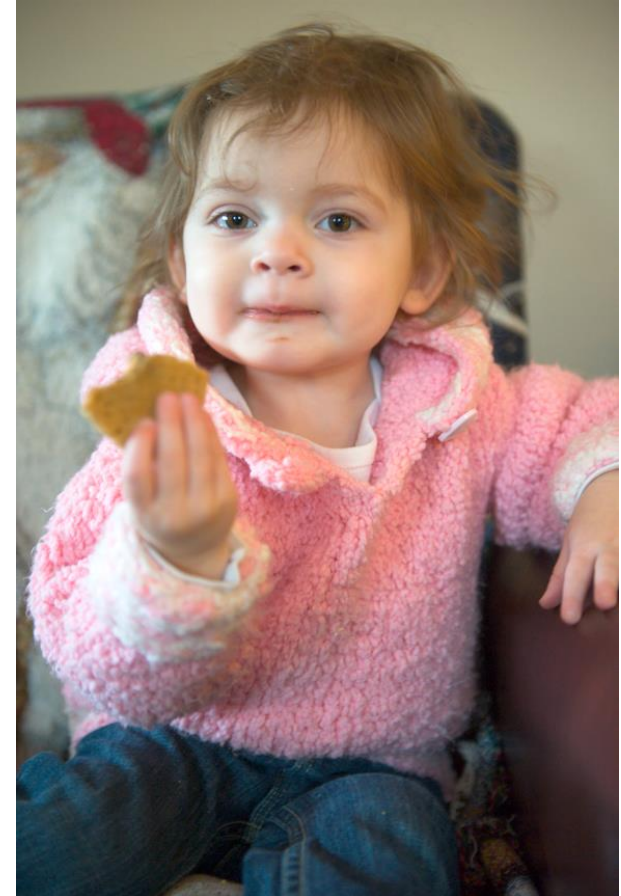
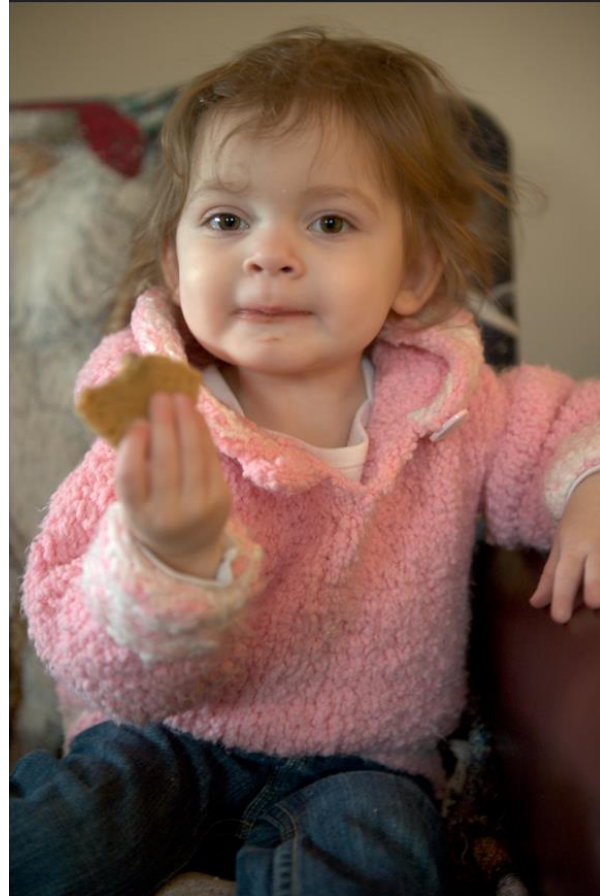
SRResNet
(23.53dB/0.7832)



SRGAN
(21.15dB/0.6868)



Diff users may even have diff preferences



3 experts gave very different tunings

It is even hard to describe what is best



Some photographers don't like over-smoothed image, and call it “**like oil painting**”



Some photographers don't like HDR image, and call it “**Cartoon-look**”

Collaborators and acknowledgement



Bill Freeman



Fredo Durand



Antonio Torralba



Josh Tenenbaum



Michael Geosele



Rick Szeliski



Daniel



Ce Liu



Orly Liba



Miki Rubinstein



Joseph Lim



Yuandong Tian



Hossein Mobahi



Jiajun Wu



Katie Bouman



Neal Wadhwa



Chengkai Zhang



Yun-Ta Tsai



Jon Barron



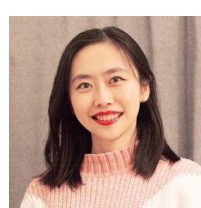
Sam Hasinoff



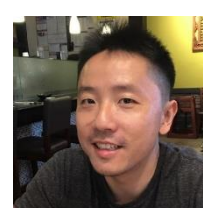
Dillon Sharlet



Jiawen Chen



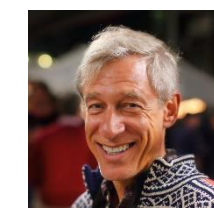
Xide Xia



Zheng Sun



Kiran Murthy



Marc Levoy



Shumian Xin



Pratul Srinivasan



Ioannis Gkioulekas



Rahul Garg



Jian Wang



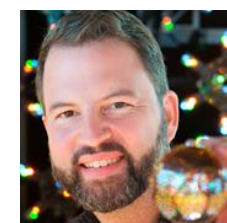
Qiurui He



Brian Kulis



Simon Niklaus



Paul Debevec



Xiuming Zhang



Q&A

