D3 CENTER
DIGITAL DESIGN, DATABILITY, AND DECISION INTELLIGENCE

Deep Sensing

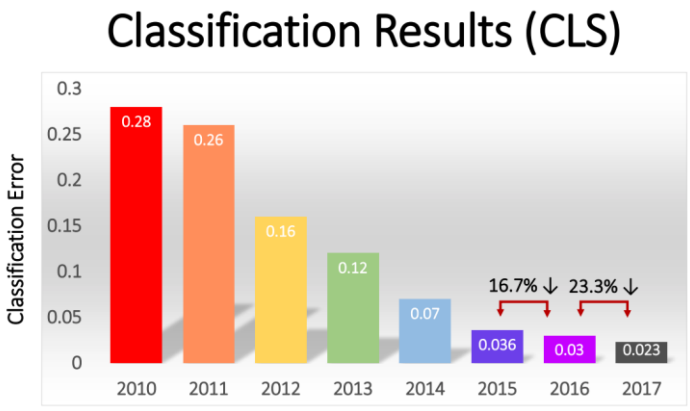
-Jointly optimizing a sensing and processing-

Hajime Nagahara

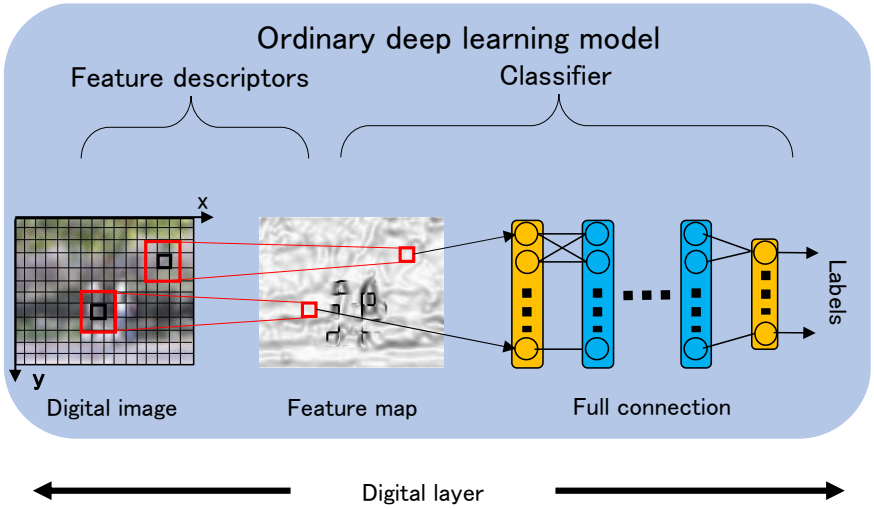
D3 Center, The University of Osaka

Limitation of an appearance-based recognition

- Optimization for feature descriptors and classifier by learning.
- The recognition accuracy outperformed from Human (ILSVRC2015).
- The optimizations are only for the pipeline after the digitization as RGB image.

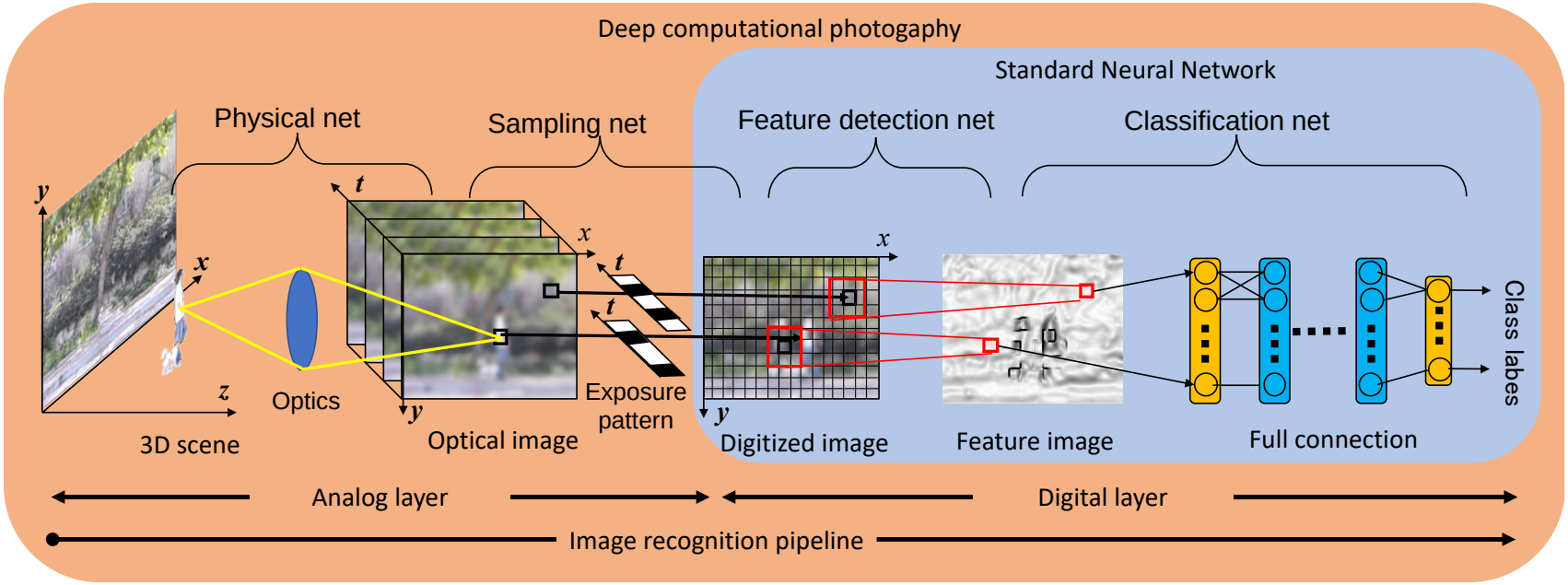


ILSVRC (Object recognition with 1000 classes)



Regular recognition model classifies appearance differences

Concept of Deep sensing

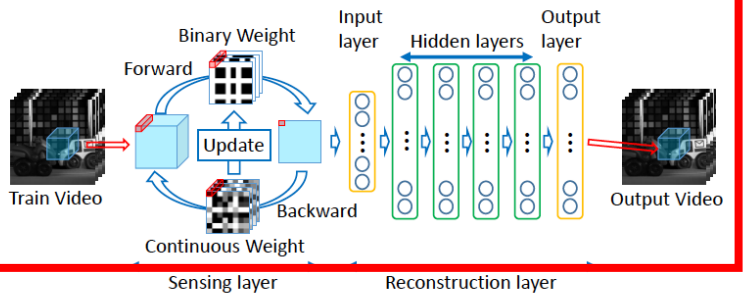


H30-R1挑戦的研究(萌芽)
R2-5 挑戦的研究(開拓)
R5-9 基盤研究S

Proposing a general framework to design a camera hardware by DNN optimization

Our research examples and history

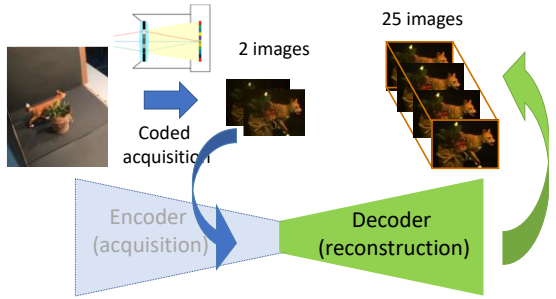
Joint optimization for Compressive video sensing [Yoshida+ ECCV2018]
Joint optimization for Compressive color video [Yoshida+ Sensors2023]



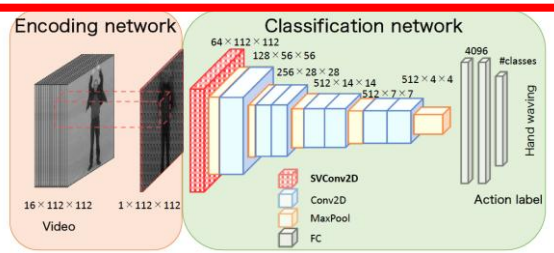
Learning to Capture Light Fields through A Coded Aperture Camera [Inagaki+ ECCV2018]

Acquiring a Dynamic Light Field through a Single-Shot Coded Image [Mizuno+ CVPR2022]

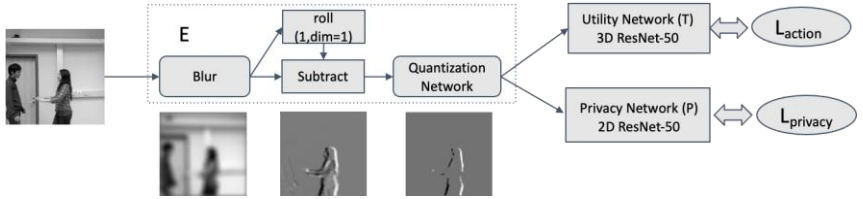
Time-Efficient Light-Field Acquisition Using Coded Aperture and Events [Habuchi+ CVPR2024]



Action Recognition from a Single Coded Image [Okawara+ ICCP2020]



Privacy Preserving Action Recognition via Motion Difference Quantization [Kuwamat+ ECCV2022]

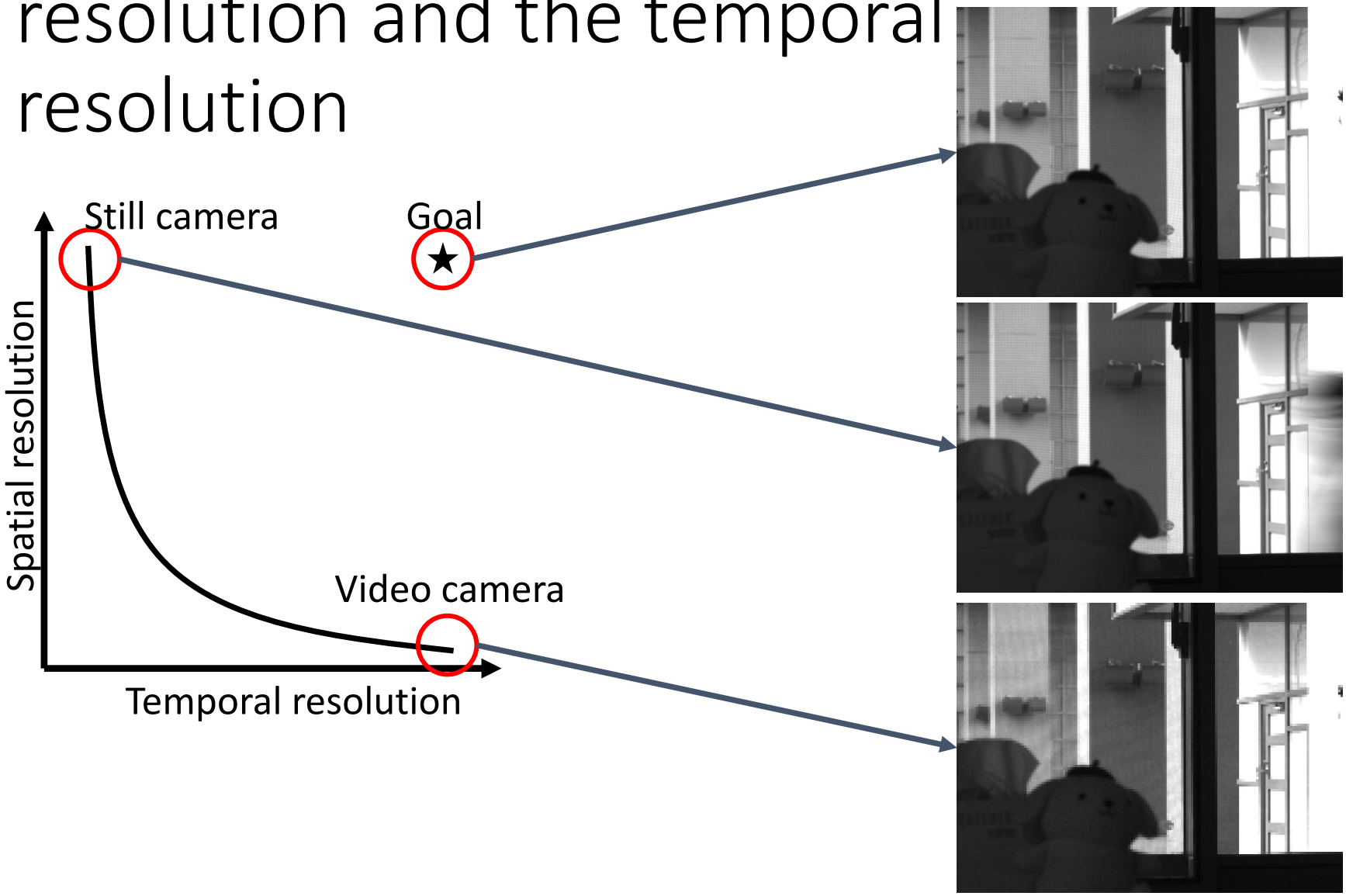


Joint optimization for Compressive video sensing

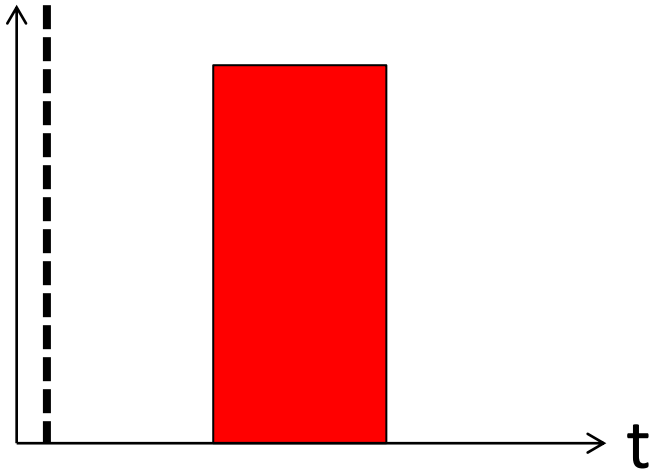
Michitaka Yoshida, Akihiko Torii, Masatoshi Okutomi,
Kenta Endo, Yukinobu Sugiyama, Rin-ichiro Taniguchi,
Hajime Nagahara

ECCV2018

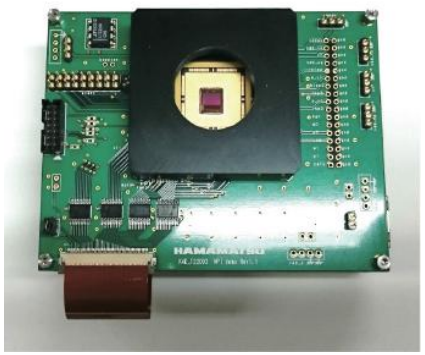
A trade-off between the spatial resolution and the temporal resolution

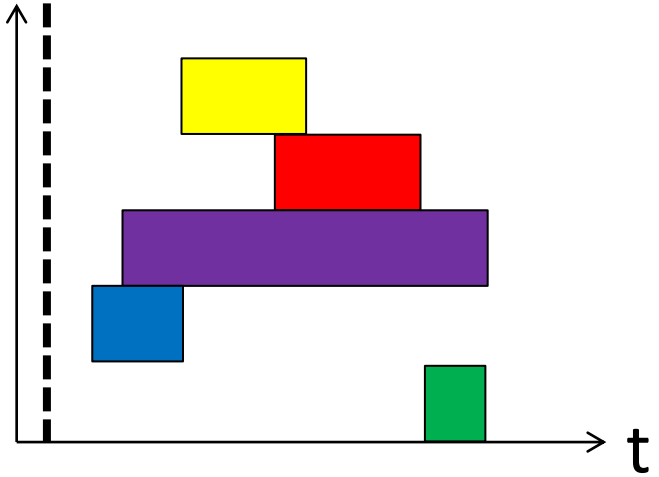


Random shutter



Global shutter





Random shutter

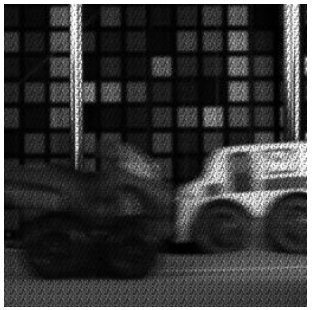
Compressive sensing is auto-encoder?

- Modeling as an auto-encoder
- The encoder is physical encoder like camera

16 frames

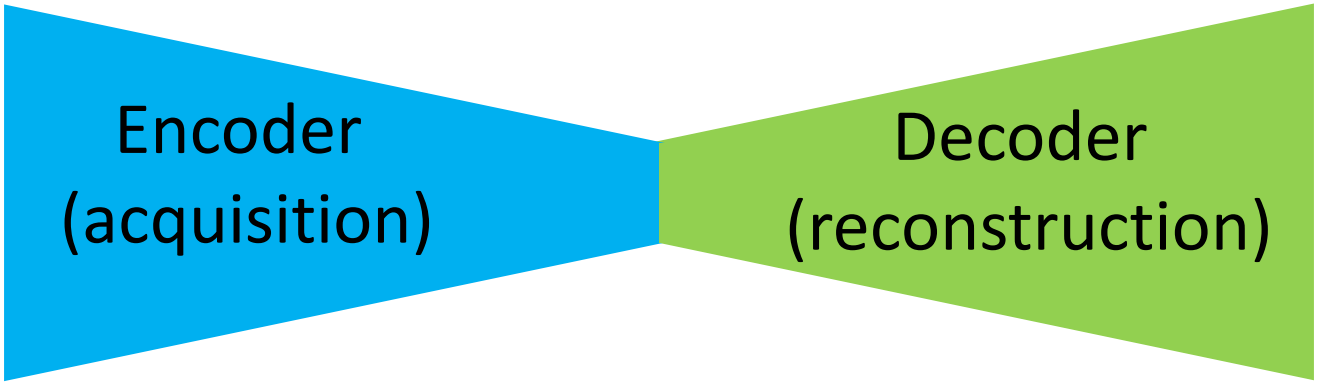
Coded acquisition

Single image

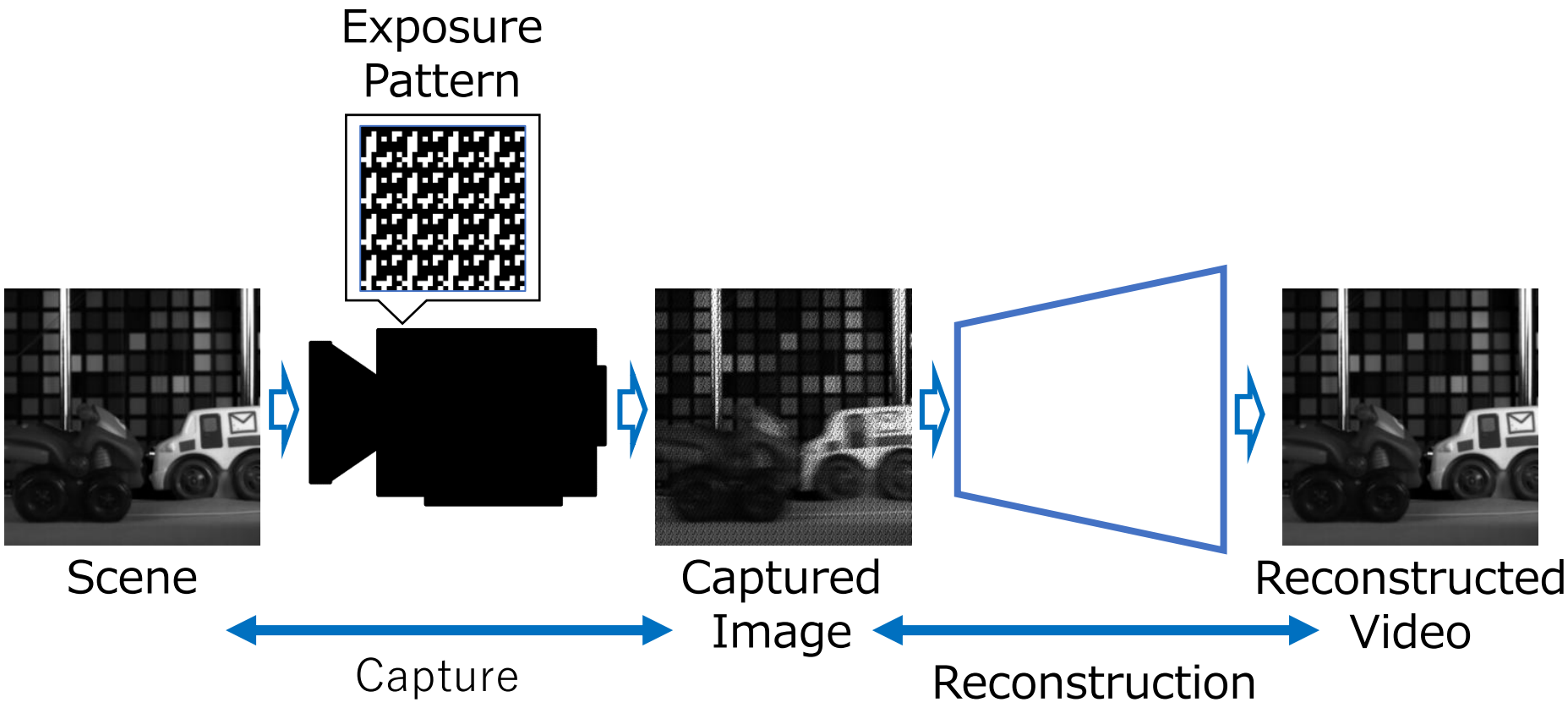


Computational reconstruction

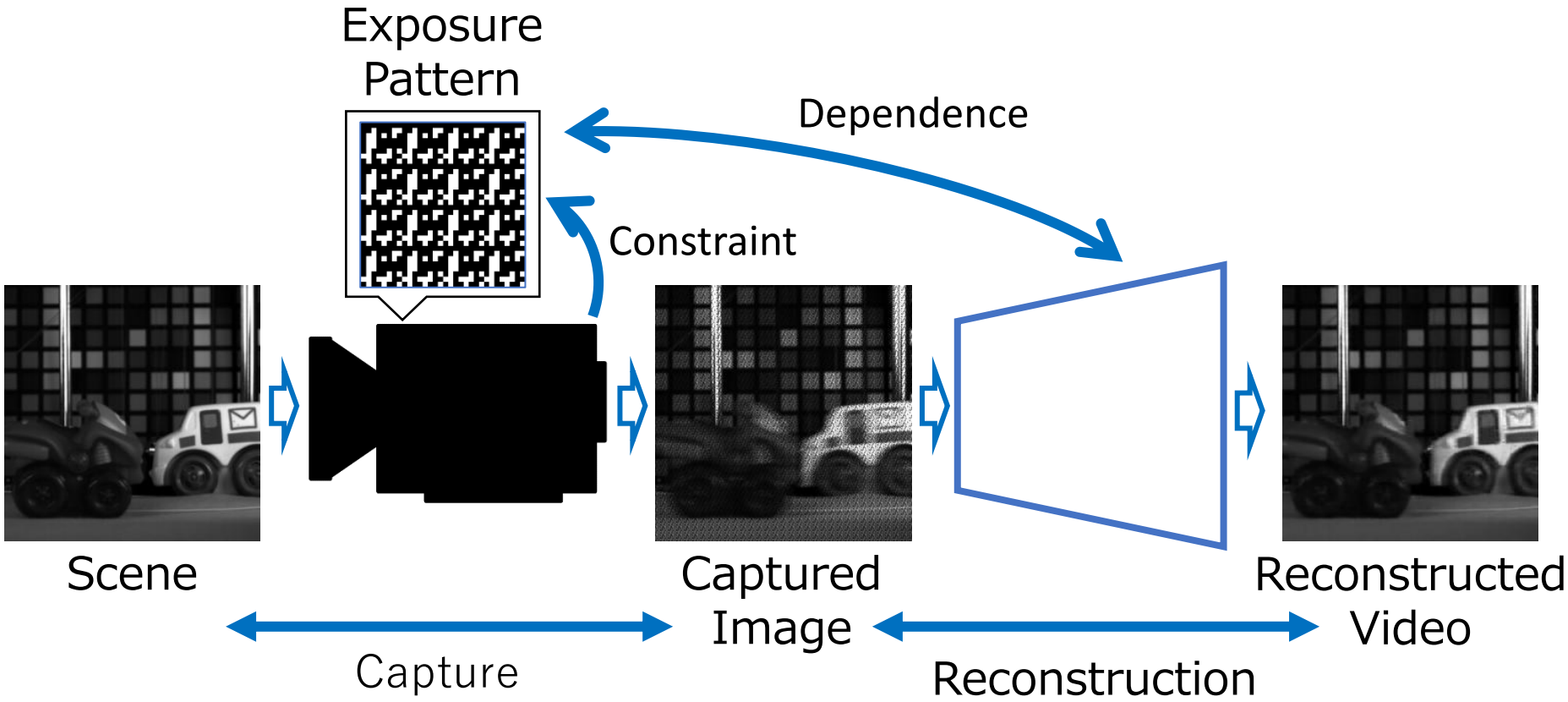
16 frames



Compressive video sensing



Compressive video sensing

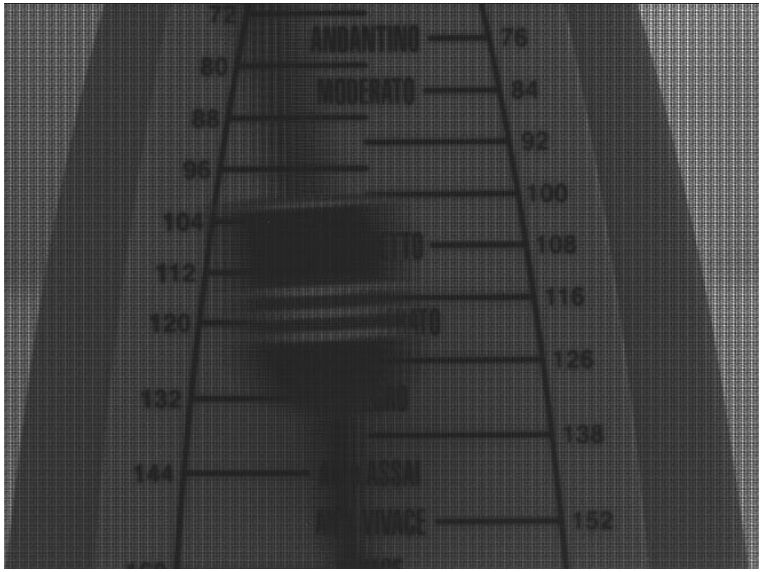


Necessary to jointly optimize

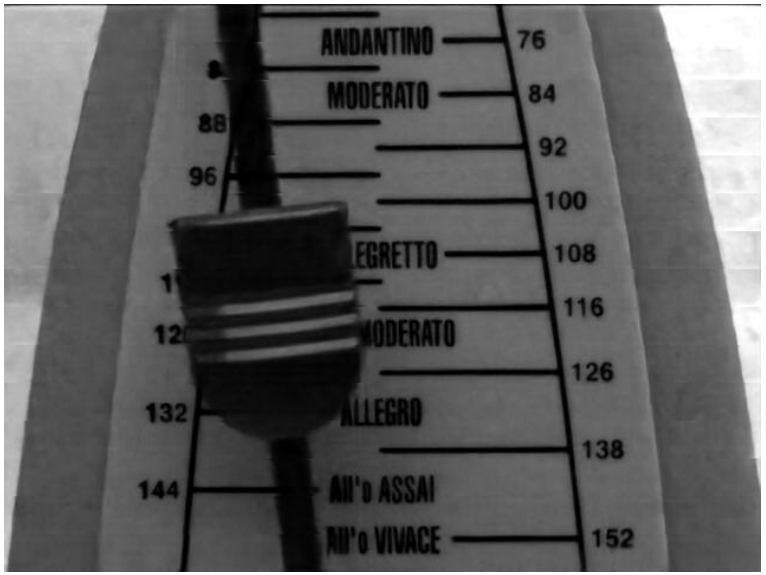
Real experiments

- Spatial constraint on exposure pattern
- Capturing coded images with 71 FPS
- Set 16 exposure patterns per frame

Real experiments

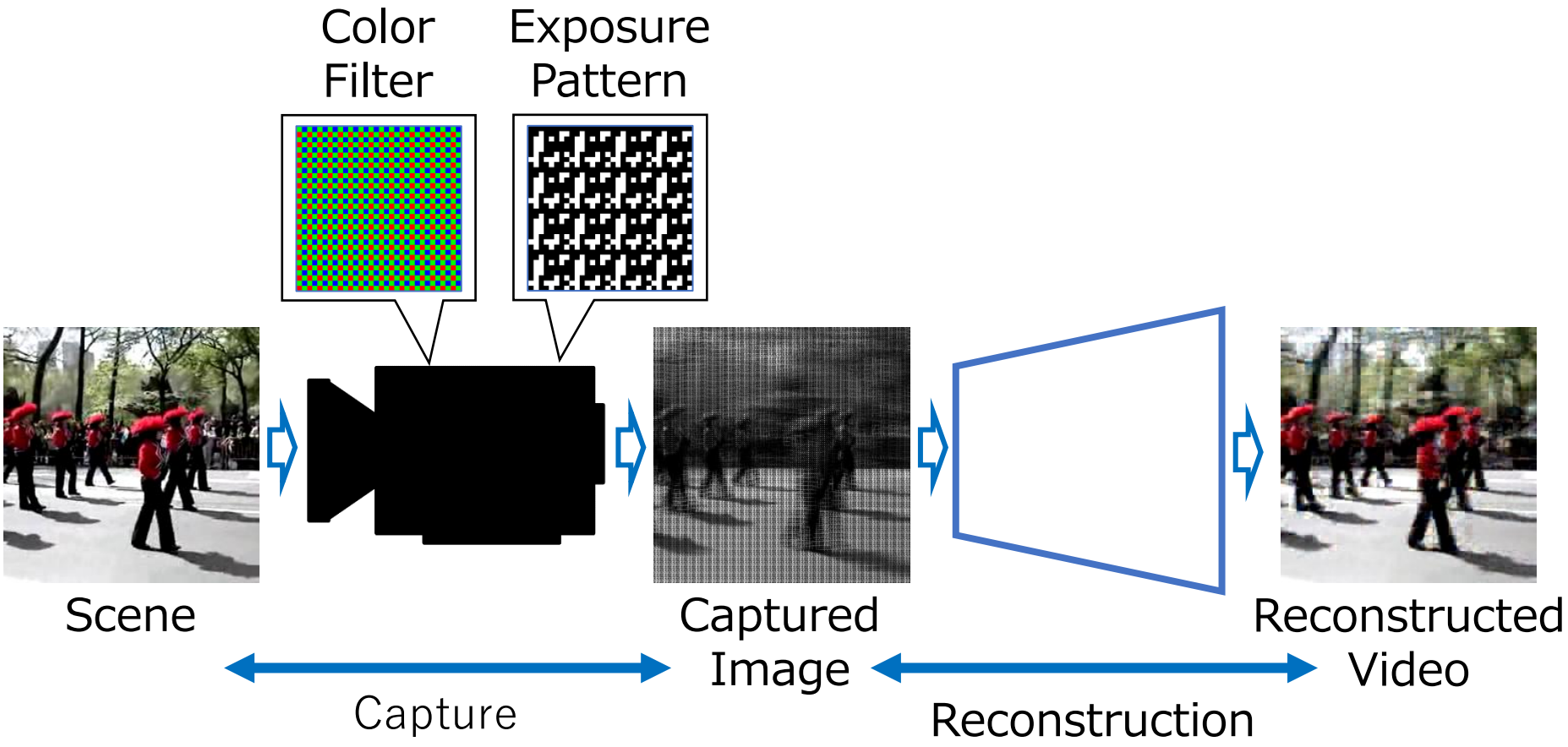


Captured Images
(71fps)



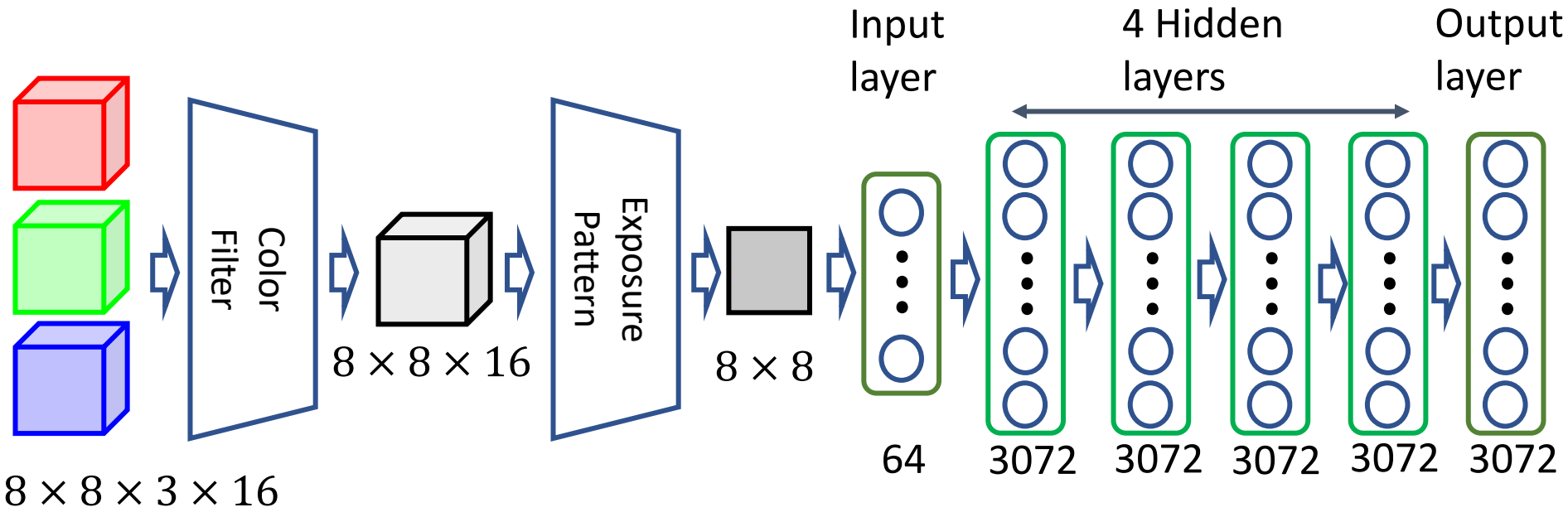
Reconstructed Video
(x16, 1136fps)

Compressive color video sensing

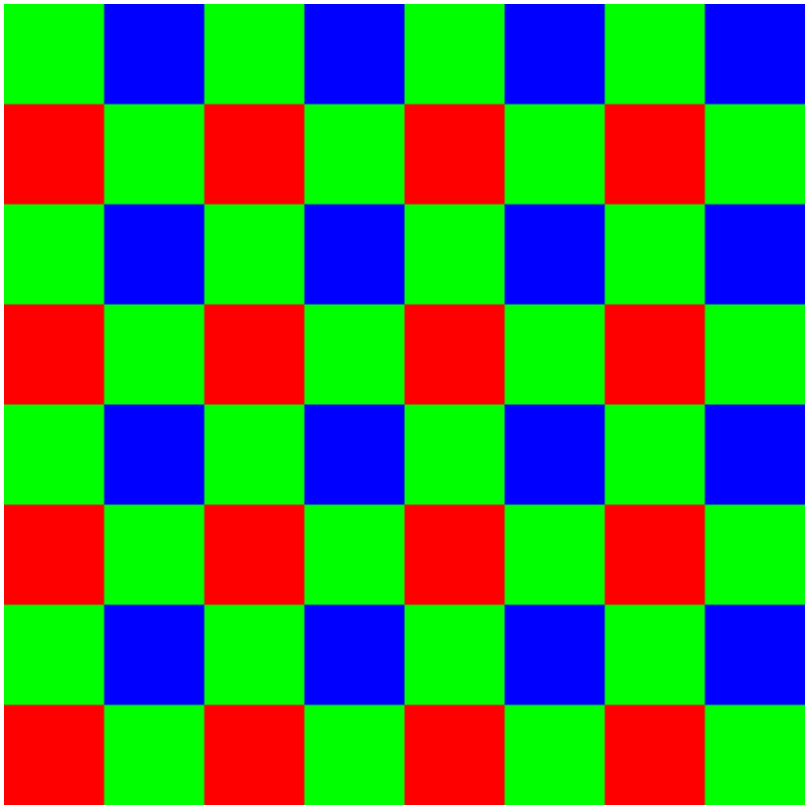


Network structure to optimize the color filter, exposure pattern and reconstruction

- Input : RGB Video ($8 \times 8 \times 3 \times 16$)
- Output : RGB Video ($8 \times 8 \times 3 \times 16$)

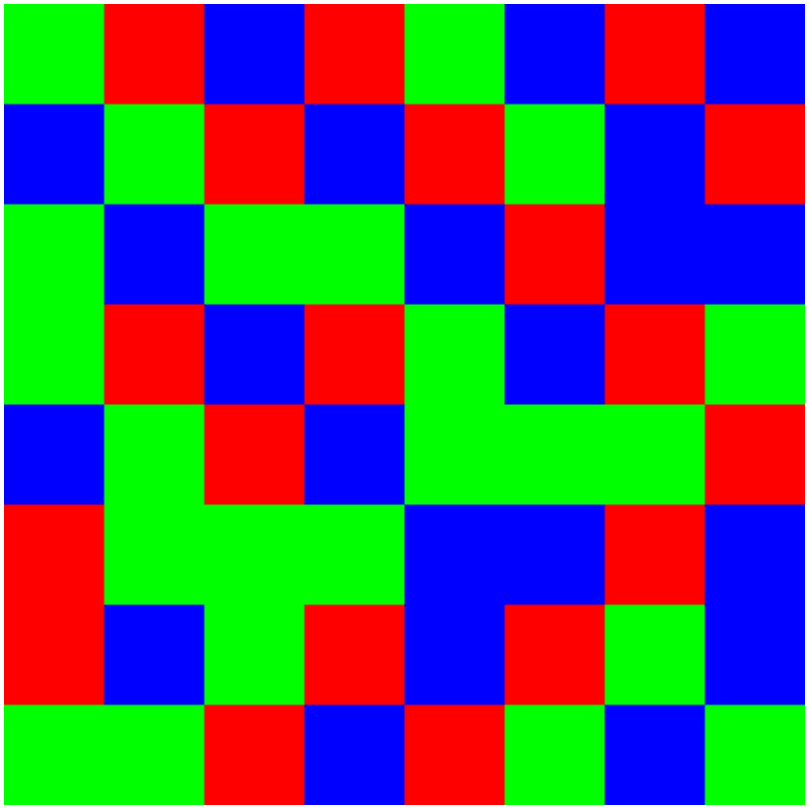


Color filter (8 × 8)



Bayer Filter

R:16 G:32 B:16



Optimized Filter
(100 epoch)

R:19 G:23 B:22

Reconstruction results

GT

w/o optimization

w/ optimization

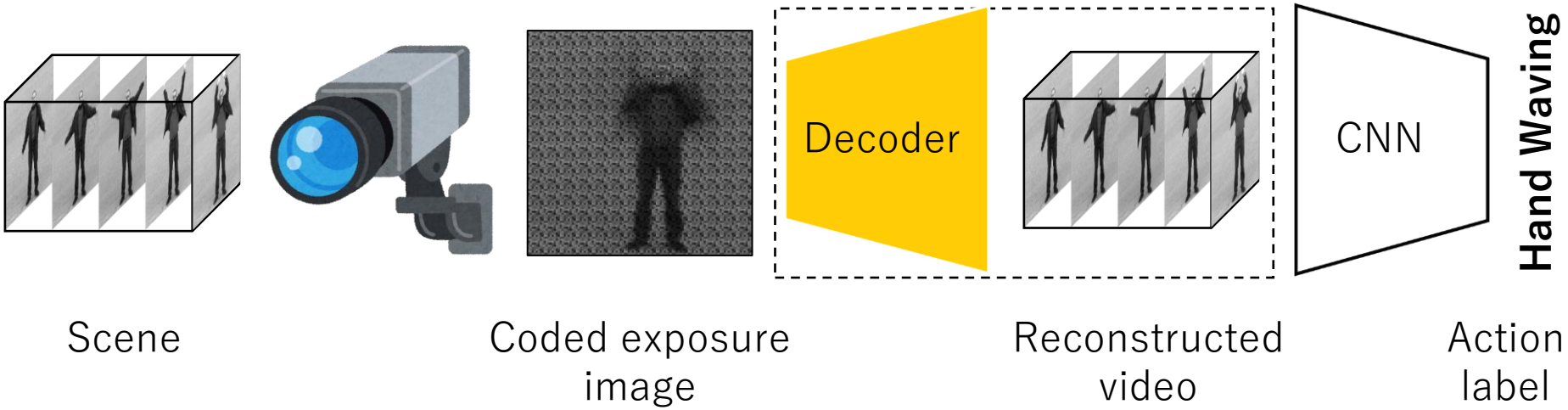
Action recognition from a single coded image

Tasashi Okawara, Michitaka Yoshida,

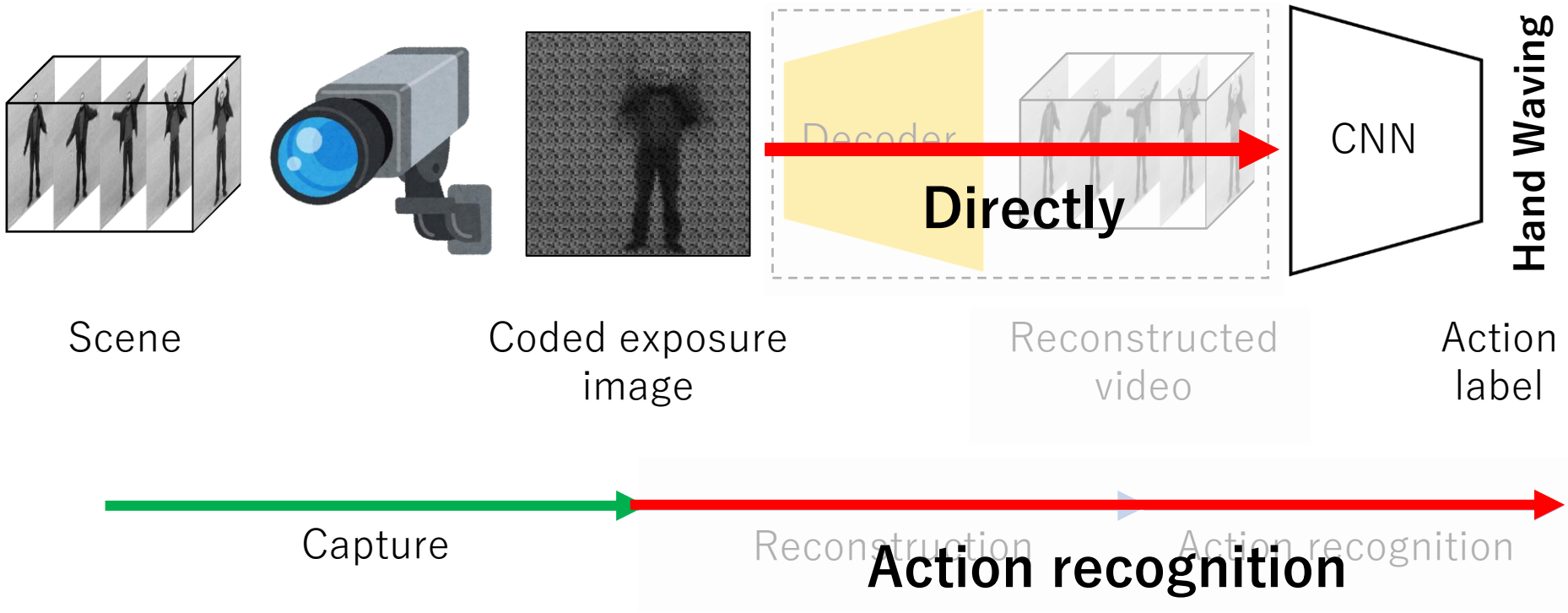
Hajime Nagahara, Yasushi Yagi

ICCP2020

Apply compressive video sensing to video analysis

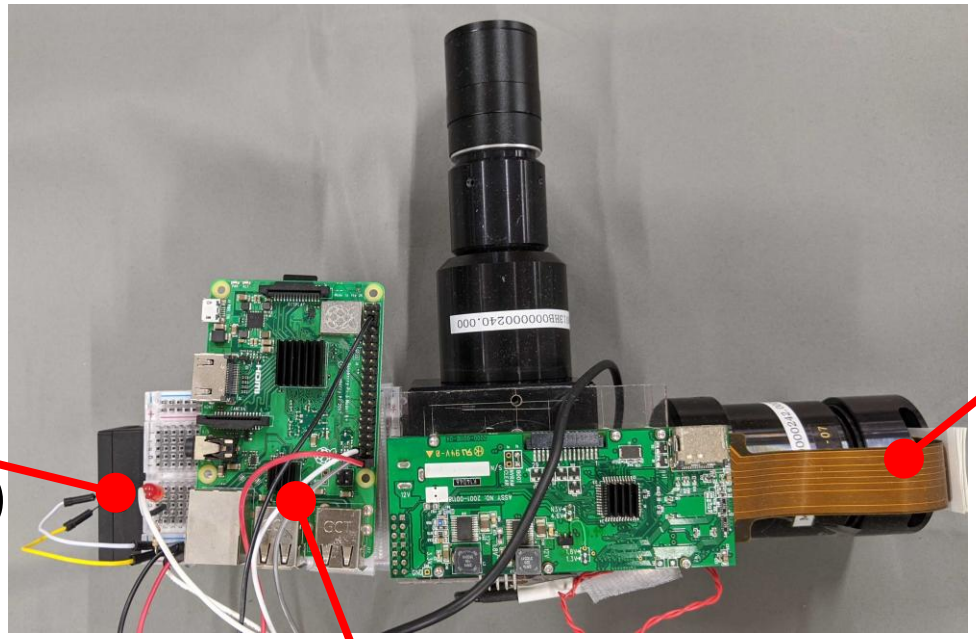


[Proposed] Action recognition from a single coded image



Prototype camera for real experiments

Coded exposure camera



Camera
(Global Shutter)

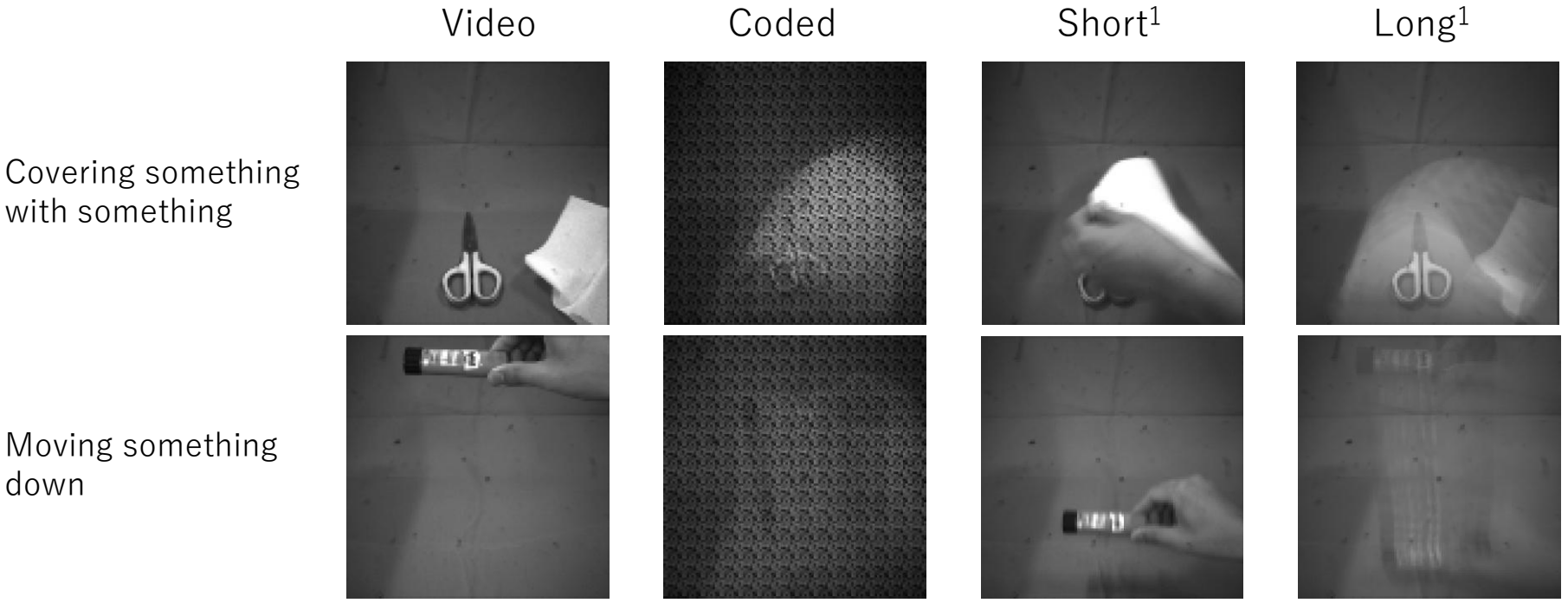
Liquid Crystal
on Silicon
(LCoS)

Raspberry Pi

Dataset in real environment

- Chose 25 actions from the Something-Something V2 dataset
- Captured 100 coded images and 100 videos with 4 objects per action

Captured image and videos



1: generate from a video

Results

Input		Model	Simulation			Real		
			Top1	Top3	Top5	Top1	Top3	Top5
Video (upper bound)		C3D	47.1	69.4	76.9	71.0	88.0	88.0
Single image	**Coded exposure (Proposed)**	SVC2D	41.6	58.9	67.2	72.0	84.0	88.0
	Long exposure	C2D	13.8	30.4	39.4	20.0	40.0	52.0
	Short exposure	C2D	14.6	32.5	40.5	21.0	47.0	60.0

Privacy Preserving Action Recognition via Motion Difference Quantization

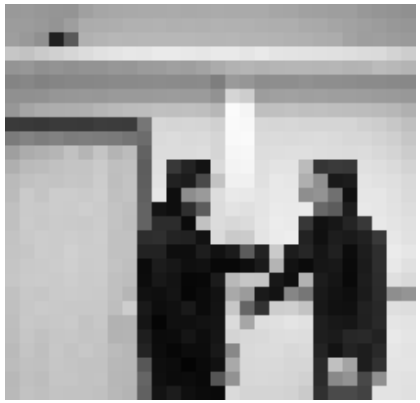
Sudhakar Kumawat and Hajime Nagahara

ECCV2022

Privacy-preserving action recognition

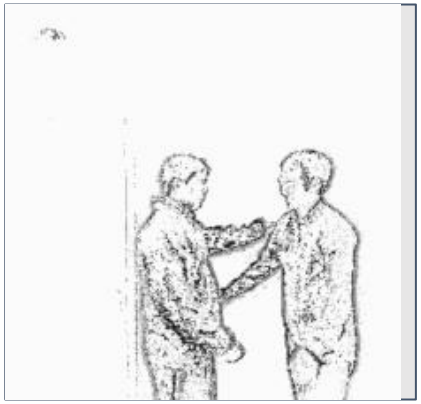
Some methods for privacy-preserving action recognition.

Original Scene



Downsampled

Blurred

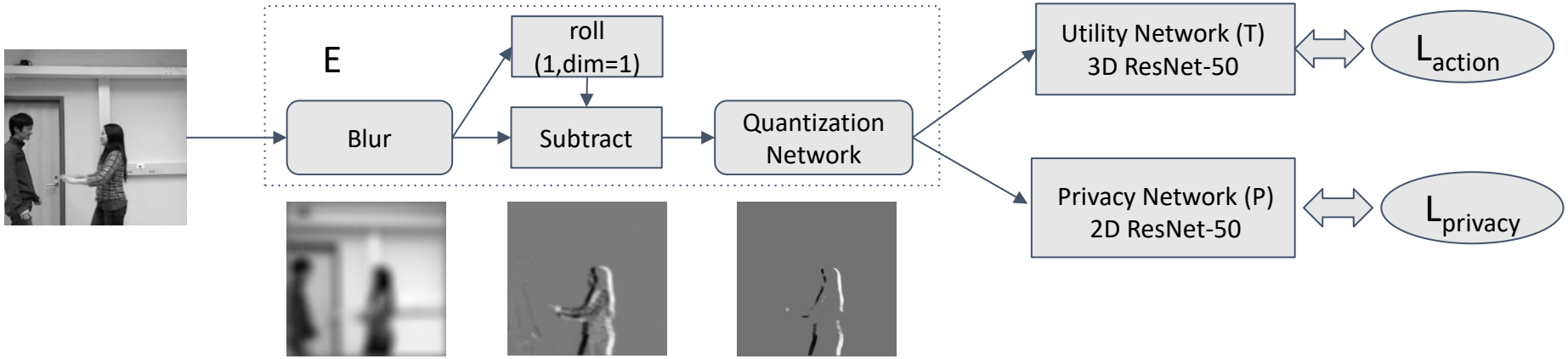


DVS Sensor

Goal: To develop an efficient encoder for the camera system that allows important features for action recognition while protecting actor(s) visual privacy.

Blur Difference Quantization

- Finding good balance of action recognition and image quality by adversarial learning



Given a frame v_i, we define a video as $V = \{v_i | i = 1, 2, .., t\}$ where t is the number of frames.

Blur module: $B_{v_i} = G_\sigma v_i$, where $G_\sigma = \frac{1}{2\pi\sigma^2} exp(-\frac{x^2+y^2}{2\sigma^2})$

Difference module: $D(B_{v_i}, B_{v_j}) = B_{v_i} - B_{v_j}$

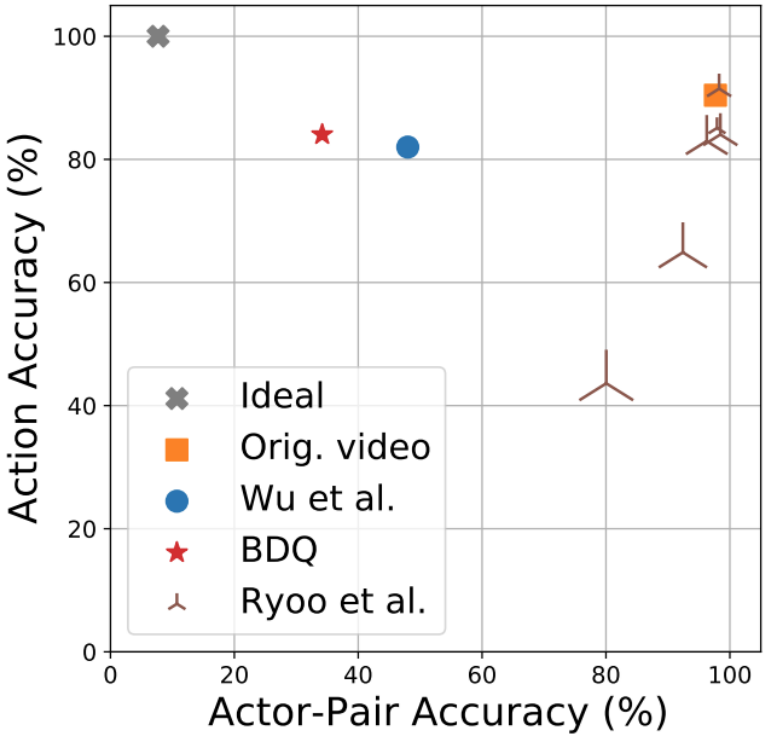
Quantization module: $\mathcal{Q}(D(B_{v_i}, B_{v_j})) = \sum_{n=1}^{N-1} \sigma(H(D(B_{v_i}, B_{v_j}) - b_i))$, where $N = 16$, t- Temperature, $b_i = \{0.5, 1.5, \ldots, N - 1.5\}$

BDQ Training: Repeat following two steps iteratively until convergence.

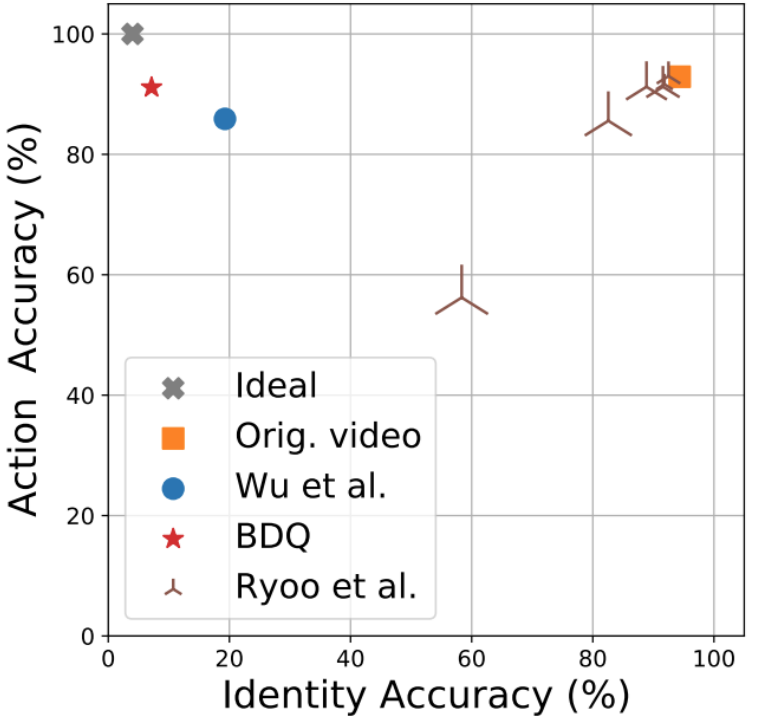
1. Fix P, train E and T using loss function $\mathcal{L}(V, \theta_E, \theta_T) = \mathcal{XE}(T(E(V)), L_{action}) - \alpha\mathcal{E}(P(E(V)))$
2. Fix E and T, train P using loss function $\mathcal{L}(V, \theta_P) = \mathcal{XE}(P(E(V)), L_{privacy})$

Evaluation on

SBU dataset (Actions-8, Privacy-13)



KTH dataset (Actions-6, Privacy-25)



Wu et al. PAMI2020 used a UNet like encoder-decoder for video degradation.

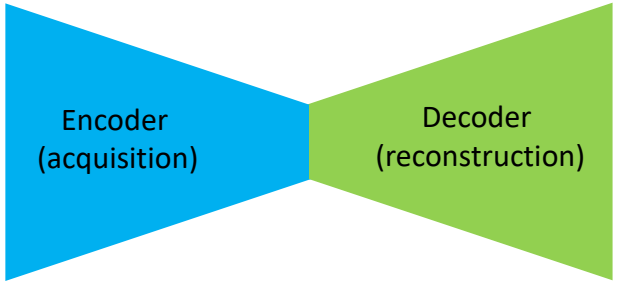
Ryoo et al. AAAI2017 used downsampling for video degradation.

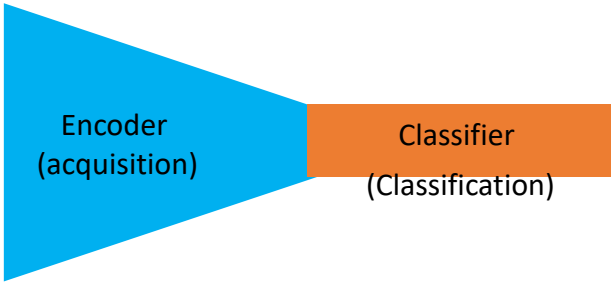
Method	Params.	Size	FLOPs
Wu *et al*	1.3M	3.8Mb	166.4G
BDQ	16	3.4Kb	120.4M

Conclusions

- Optimize the optics and image sensor as an encoder.

- Reconstruction
  - Important cues for reconstruction.
  - Omitting redundant information which decoder can be interpolated.
- Classification
  - Distinguishable features btw classes.
  - Filter out the redundant information.
  - Rich information is not always better: Cost, data rate, and privacy.





- We are expecting to use the controllable/programmable image sensor which encode the image on pixel on chip.