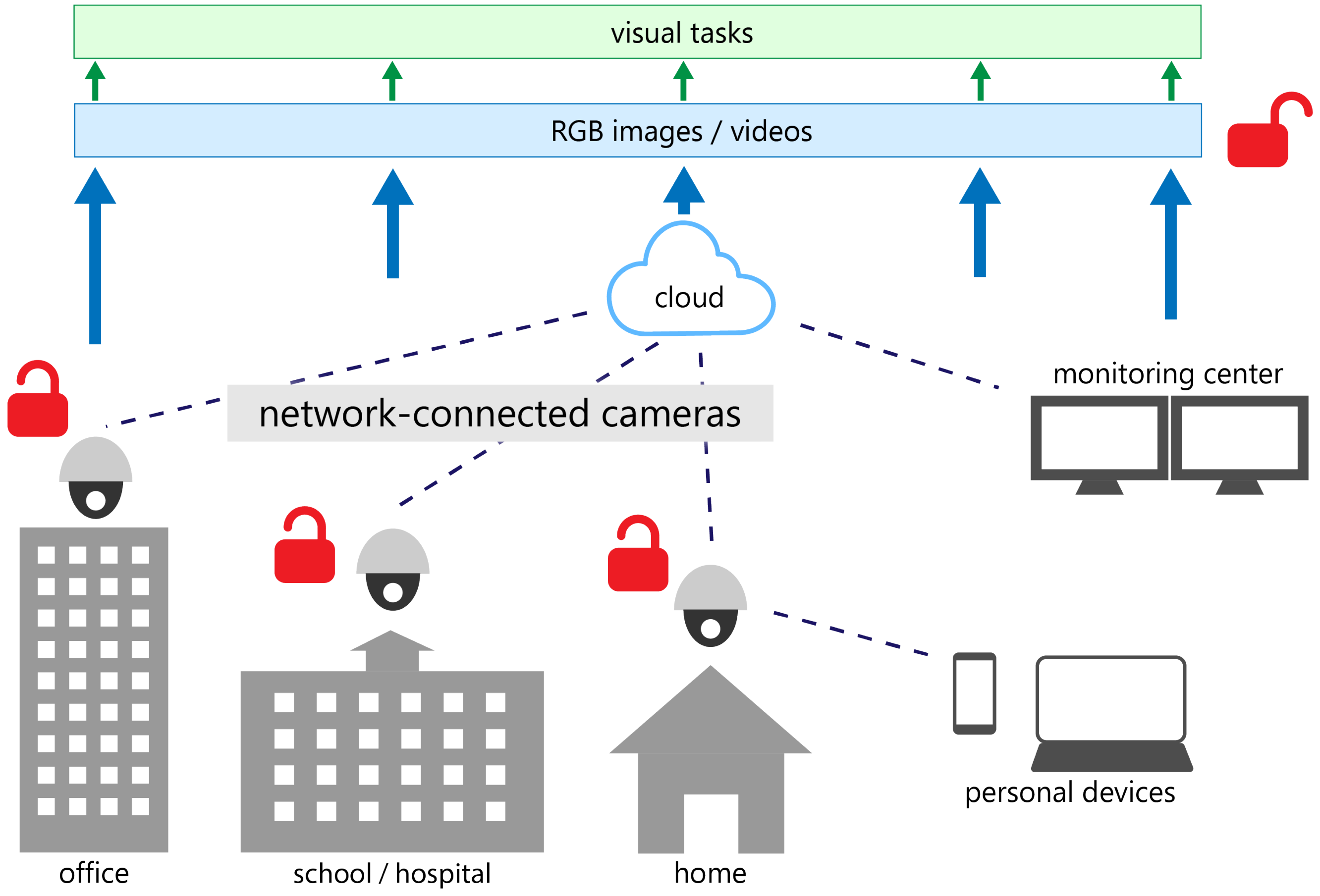
Privacy-Preserving Action Recognition using Coded Aperture Videos

Zihao W. Wang[1], Vibhav Vineet[2], Francesco Pittaluga[3],
Sudipta N. Sinha[2], Oliver Cossairt[1], Sing Bing Kang[4]

[1]Northwestern University [2]Microsoft Research [3]University of Florida [4]Zillow Group

Motivation

- Monitoring cameras are widely deployed in public and/or private places.
- Yet the network-connected cameras are subject to attacks!
- Moreover, mainstream visual algorithms always interface with RGB images / videos, which are privacy revealing.



Goal:

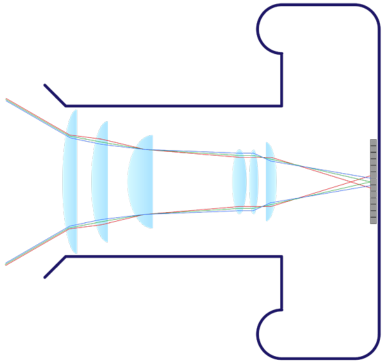
To preserve privacy before capturing the images.

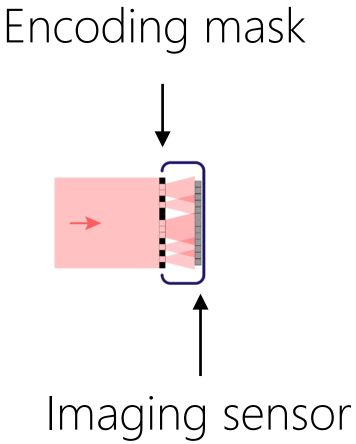
To execute visual task(s) without looking at the original visual data.

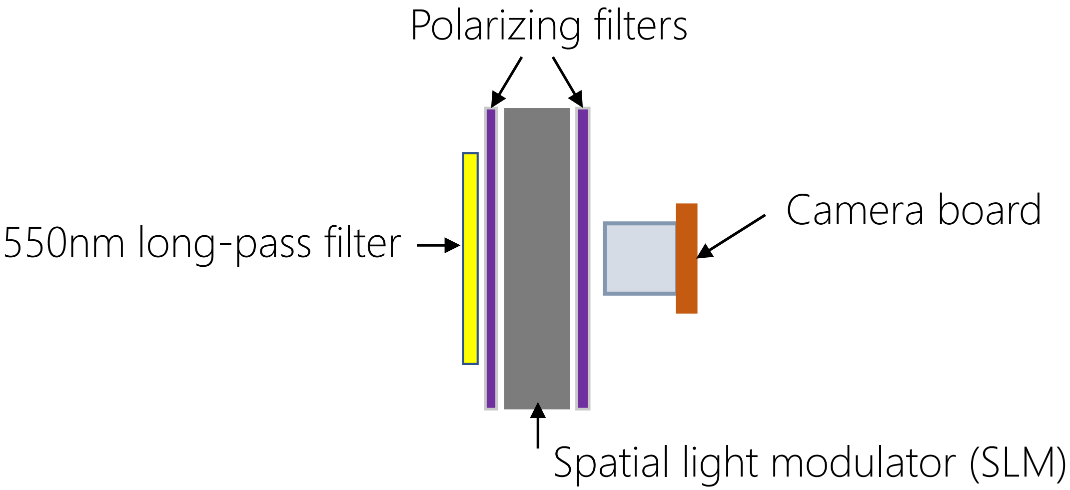
Previous work on pre-capture privacy cameras

- *Dai et. al., 2015*
  - Multiple extremely low-res cameras
    - 100x100 to 1x1
  - Multi-camera issues (calibration, effective FOV, …)
- *Pittaluga and Koppal, 2017*
  - Optical defocus blur
  - Estimating blur kernels (Gaussian) is not hard
- *Kulkarni and Turaga, 2016*
  - Compressive (single pixel) imaging
  - Costly to implement (DMD is required)
- Ours:
  - Single camera
  - Severely blurred, kernels are very hard to retrieve
  - Easy to build

Lens-free coded aperture imaging







	Lens-based cameras	Lens-free coded aperture cameras
Strength	High visual quality; CV algorithms friendly	No lens issues; Easy to build
Weakne ss	Lens issues (distortion; aberration; heavy)	Noise due to low light; Requires heavy computation
Privacy	revealing	preserving

Image formation (in matrix form)

$$\mathbf{d} = \mathbf{A}\mathbf{o}$$

$\mathbf{o} \in \mathbb{R}^{mn\times 1}$ object image

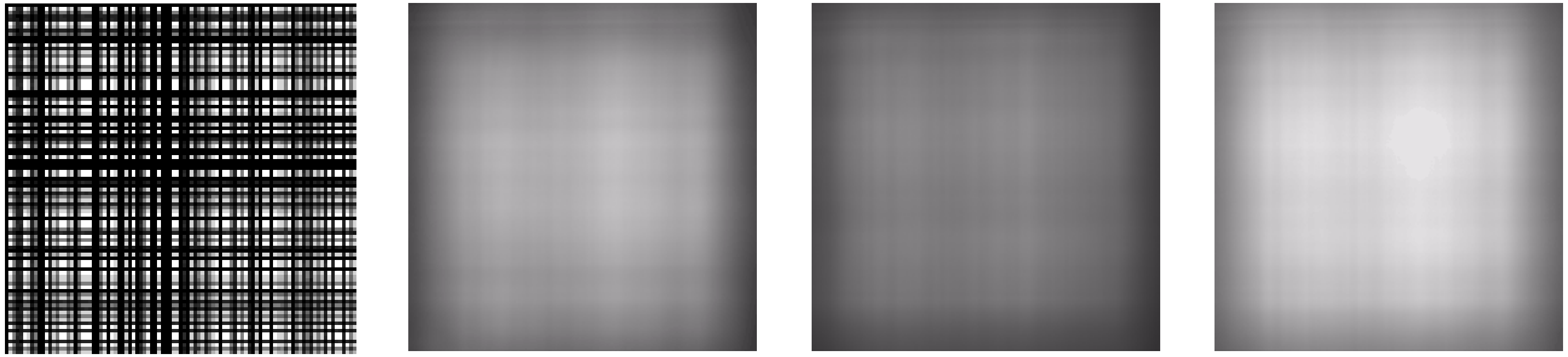
$\mathbf{A} \in \mathbb{R}^{\boxed{mn\times mn}}$ coded aperture PSF

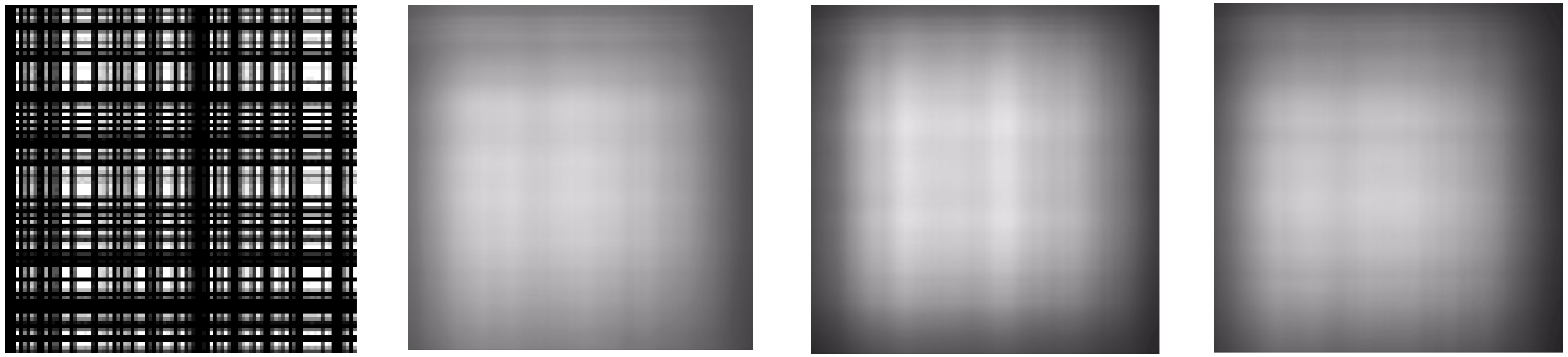
$\mathbf{d} \in \mathbb{R}^{mn\times 1}$ detected image

If $m = n = 10^3$,
A will be $10^6 \times 10^6$

What is this?

Hint: encoding mask





Answer

Coded aperture images are difficult to understand by human beings!

Can machines do a better job?

A pilot study

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

Same training settings...

Results:

5-class classification	Training accuracy (%)	Validation accuracy (%)
RGB2gray images	50^{th}: 99.56 (Max: 99.86)	50^{th}: 94.39 (Max: 95.91)
Synthetic CA images	50^{th}: 79.06 (Max: 92.65)	50^{th}: 63.21 (Max: 83.96)

Directly classifying CA images will quickly result in overfitting!

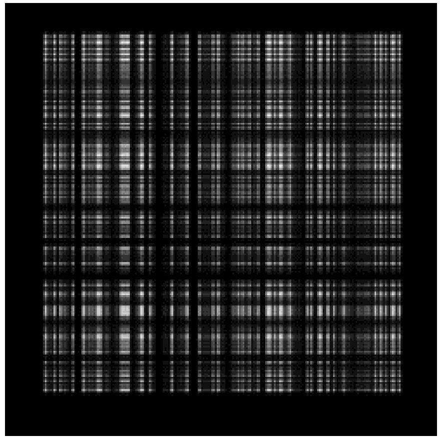
What about reconstruction? Possible but non-trivial.

Scene reconstruction from coded aperture image

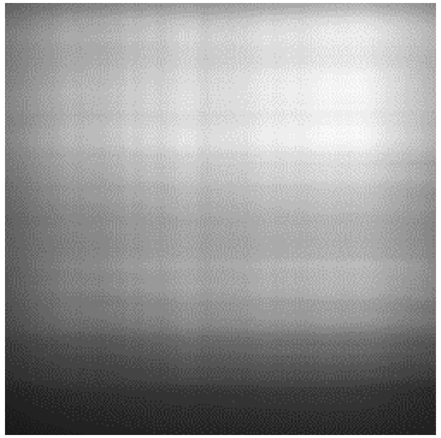
We follow the deconvolution method from *DeWeert & Farm 2015*

- The mask used is separable on x and y axis, so as to reduce complexity.
- The mask code on one dimension is called “Maximum Length Sequence” (MLS), a family of random binary code.

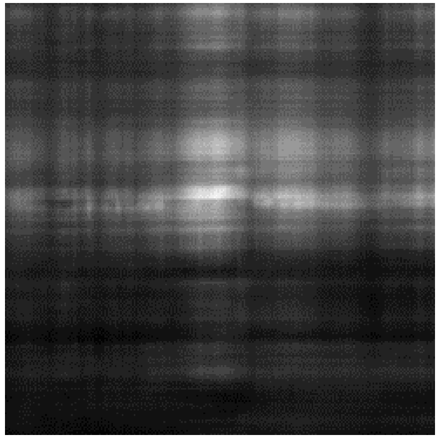
Coded mask on SLM



Captured image



Reconstructed image



Reference

For high quality:

- preprocessing (hot pixel removal, denoising)
- PSF calibration
- many iterations

Expensive for executing visual tasks!

Challenges and opportunities

CA image | Mask PSF | Scene reconstruction | Visual task

All available | √ | ? | DeWeert & Farm 2015, Asif et. al. 2017

Available | 1. Estimate PSF; 2 Reconstruction | ? | Challenging for large kernels

Available | PSF-free reconstruction | ? | Strong constraints on objects

Available | Reconstruction-free vision | proposed

CA images | Image processing | DNN

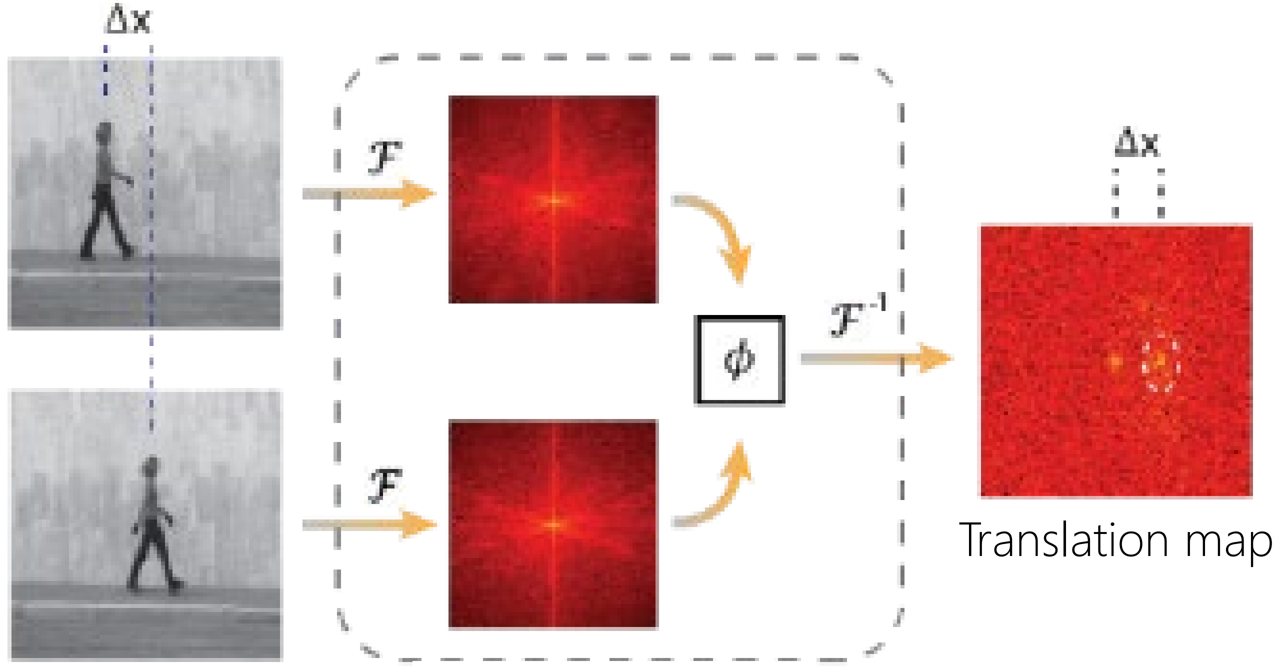
Action recognition in this work.

Second layer of privacy, serves as interface to DNNs.

First layer of privacy, requires optics knowledge for recovery.

Motion characterization (Translation)

- $o_2(\mathbf{p}) = o_1(\mathbf{p} + \Delta\mathbf{p})$ — $\mathbf{p} = [x, y]^T; \Delta\mathbf{p} = [\Delta x, \Delta y]^T$
- $\mathrm{O}_2(\mathbf{v}) = \phi(\Delta\mathbf{p})\mathrm{O}_1(\mathbf{v})$ — Fourier transform, $\mathbf{v} = [\xi, \eta]^T$ $\quad \phi(\Delta\mathbf{p}) = e^{i2\pi(\xi\Delta x + \eta\Delta y)}$
- $\mathrm{C}(\mathbf{v}) = \frac{\mathrm{O}_1 \cdot \mathrm{O}_2^*}{|\mathrm{O}_1 \cdot \mathrm{O}_2^*|} = \phi^* \frac{\mathrm{O}_1 \cdot \mathrm{O}_1^*}{|\mathrm{O}_1 \cdot \mathrm{O}_1^*|} = \phi(-\Delta\mathbf{p})$ — Cross power spectrum
- $c(\mathbf{p}) = \delta(\mathbf{p} + \Delta\mathbf{p})$ — Inverse Fourier transform



phase correlation

Translation features for coded aperture images

- $d_2(\mathbf{p}) = o_2(\mathbf{p}) * a = d_1(\mathbf{p}')$

$\mathbf{p} = [x, y]^T; \mathbf{p}' = \mathbf{p} + \Delta\mathbf{p}$

- $\mathrm{D}_2(\boldsymbol{\nu}) = \mathrm{O}_2 \cdot \mathrm{A} = \phi \cdot \mathrm{O}_1 \cdot \mathrm{A}$

Fourier transform

- $\mathrm{C}_d(\boldsymbol{\nu}) = \frac{\mathrm{D}_1 \cdot \mathrm{D}_2^*}{|\mathrm{D}_1 \cdot \mathrm{D}_2^*|} = \phi^* \frac{\mathrm{O}_1 \cdot \mathrm{A} \cdot \mathrm{A}^* \cdot \mathrm{O}_1^*}{|\mathrm{O}_1 \cdot \mathrm{A} \cdot \mathrm{A}^* \cdot \mathrm{O}_1^*|} \approx \phi^*$

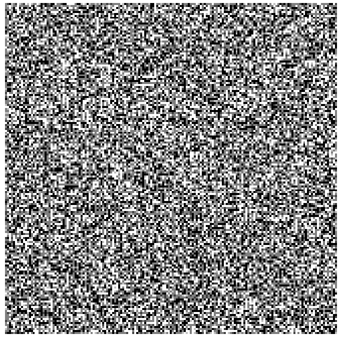
Cross power spectrum

- $c_d(\mathbf{p}) = \delta(\mathbf{p}')$

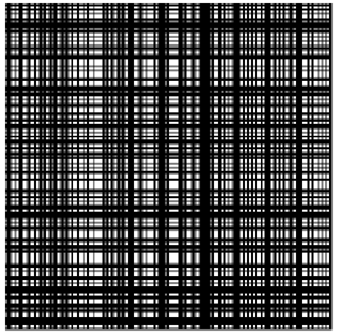
Inverse Fourier transform

The mask spectrum in Fourier space should contain as many non-zeros as possible, so that Translation features are invariant to mask patterns.

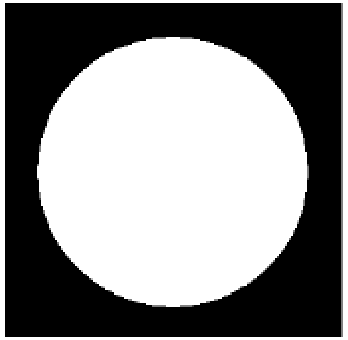
To evaluate, we qualitatively compare three masks (50%)



Mask 1



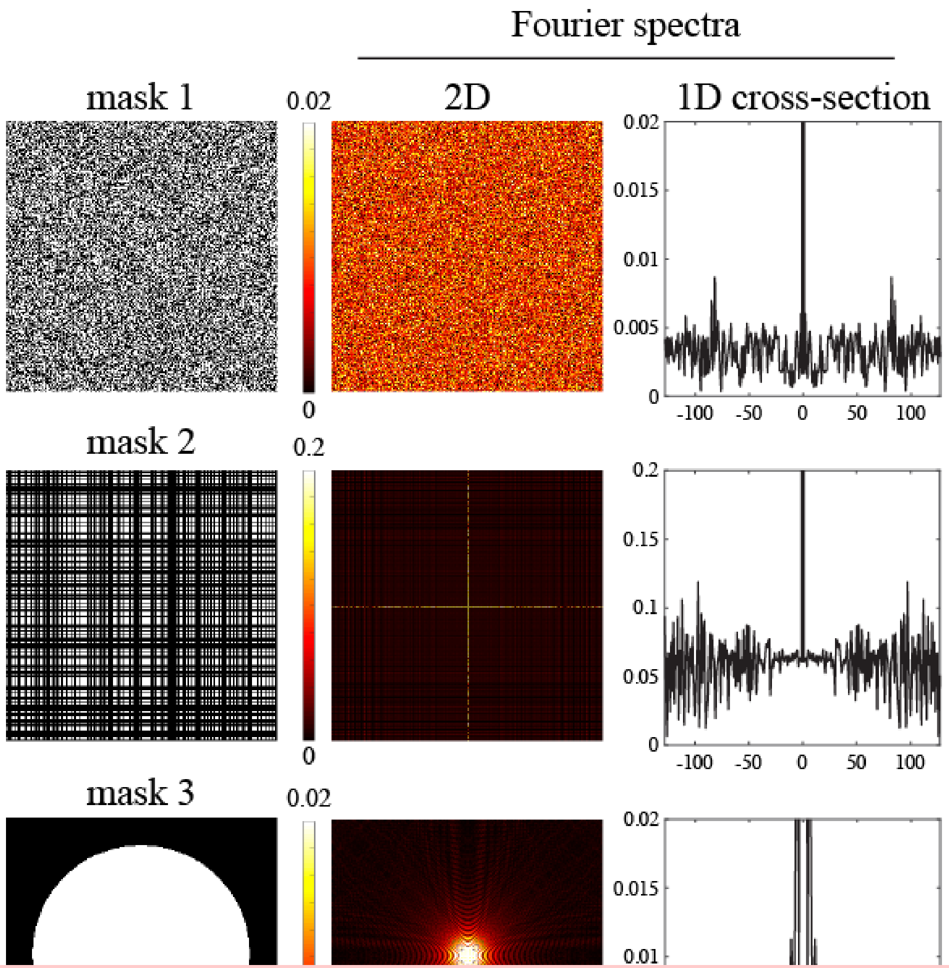
Mask 2

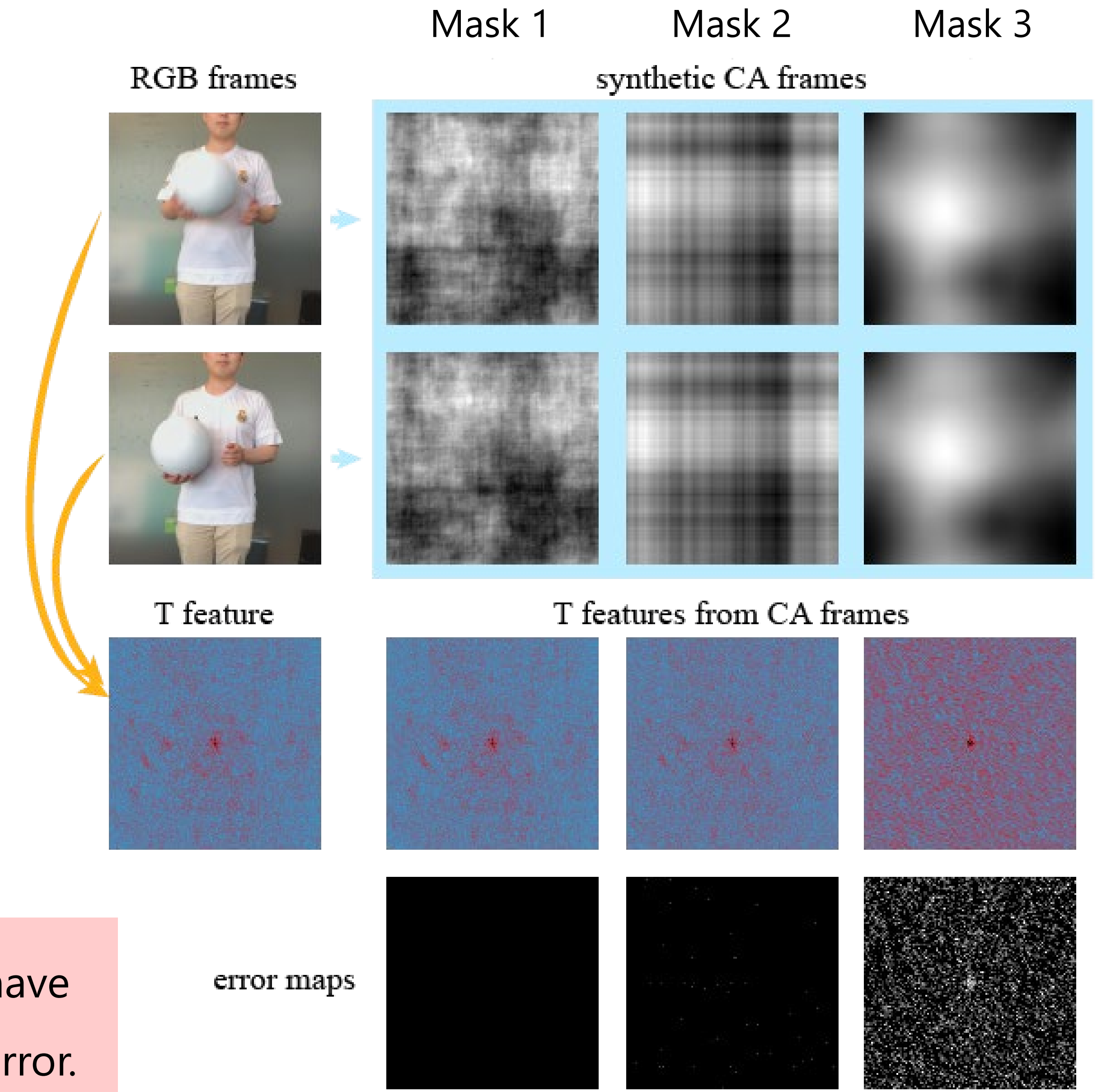


Mask 3

Mask 1: complete random;
Mask 2: x-y separated MLS;
Mask 3: a circular shape aperture

Comparison of mask patterns





Compared to T feature without mask, Mask 1 and 2 have little effect on T features, but Mask 3 has noticeable error.

Classifying T features

We repeat the 5-class classification task, but first converting RGB images to T features:

RGB->gray->CA->T

We show the progress of validation accuracy for 50 epochs.

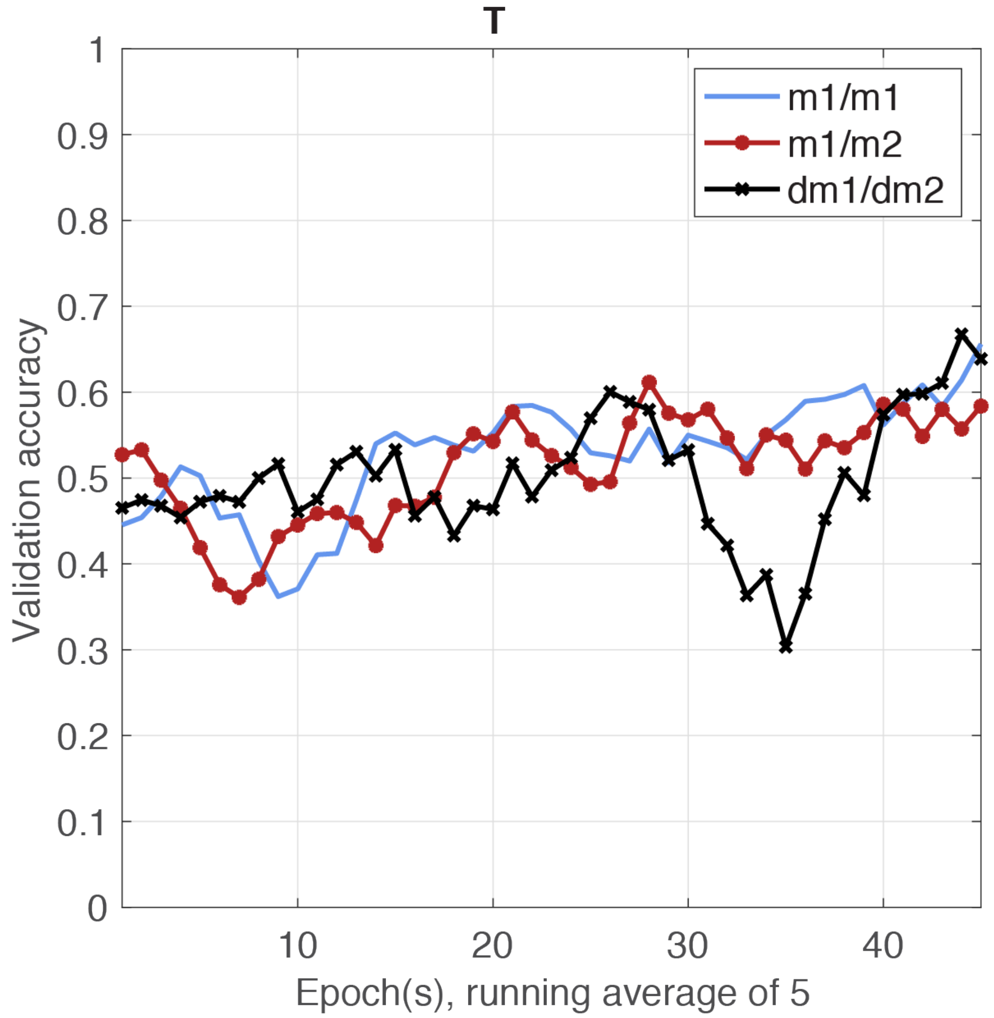
We compare 3 strategies:

"m1/m1": train and validate using the same mask;

"m1/m2": train and validate using two different masks;

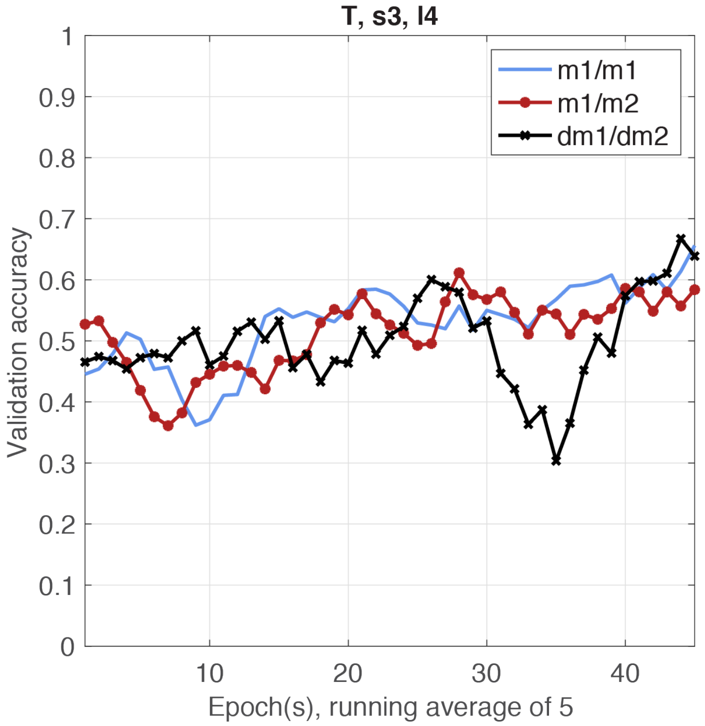
"dm1/dm2": train and validate using randomly generated masks changes for each epoch.

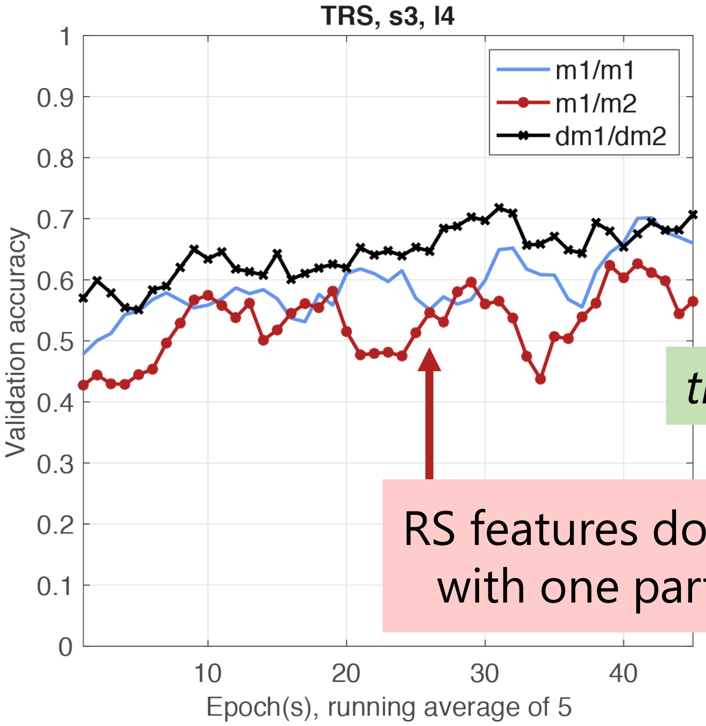
The results validates the "mask-invariant" property for T features.



Further improvements (Rotation & Scale features)

- $o_2(\mathbf{p}) = o_1(s\mathbf{R}\mathbf{p})$ — s: scaling factor; $\mathbf{R}$: rotation matrix
- $\mathrm{O}_2(\mathbf{v}) = \mathrm{O}_1(s\mathbf{R}\mathbf{v})$ — Rotation and scale preserves in Fourier domain
- $|\mathrm{O}_2(\mathbf{q})| = |\mathrm{O}_1(\mathbf{q} + \Delta\mathbf{q})|$ — Transform to log-polar space. $\mathbf{p} = [x, y]^T \Rightarrow \mathbf{q} = [\log(\rho), \theta]^T$
- Use phase correlation again to obtain RS feature map.
- Together with Translation features, to form TRS feature maps.



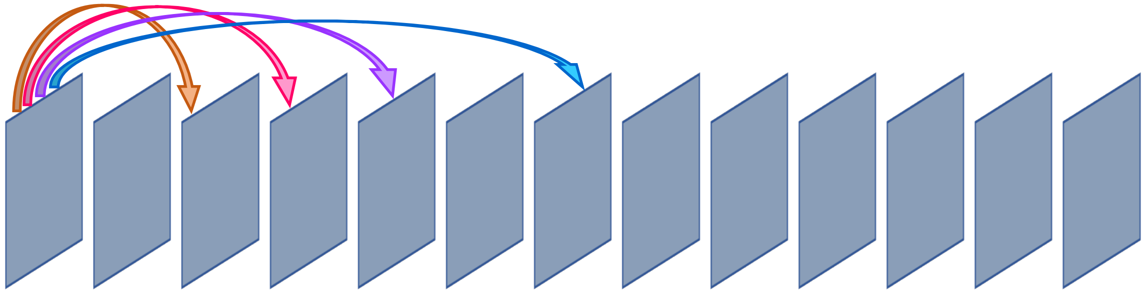


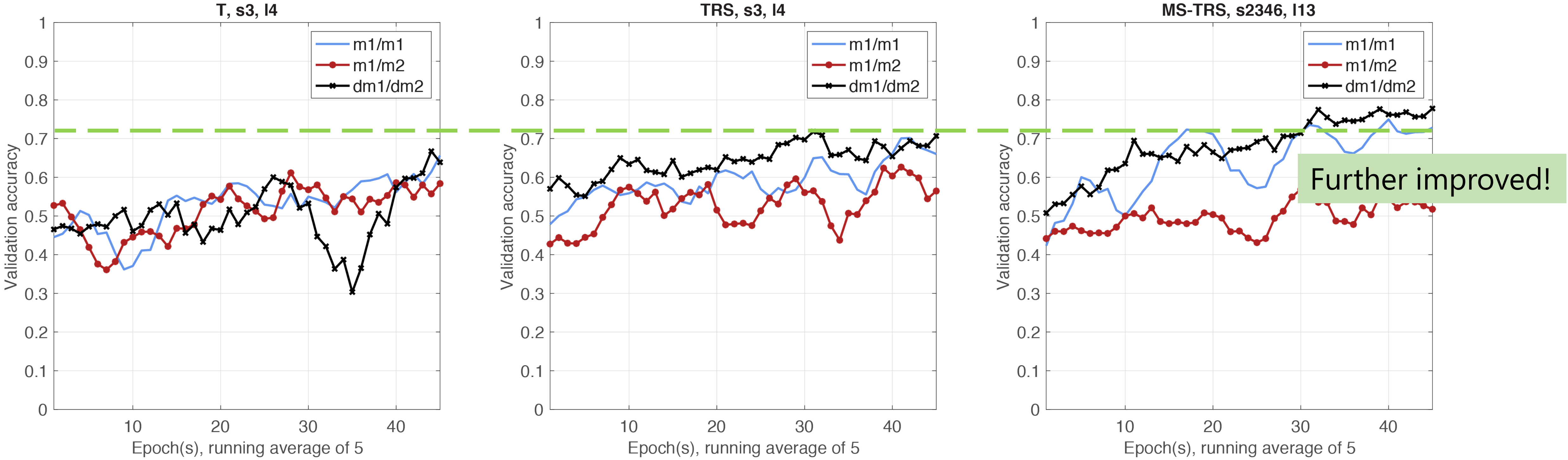
training with varying masks improves accuracy!

RS features does not have "mask invariant" property. Training with one particular mask does not apply to another mask.

Further improvements (TRS features at multiple time strides)

- Compute TRS at multiple time strides.
  - e.g. for a 13-frame video (*l13*), strides of *{s2, s3, s4, s6}* results in *(6+4+3+2)x2=30* TRS images.
  - We name it MS-TRS.





More testing results

We focus on indoor actions with stationary cameras.

Datasets: 22 classes selected from UCF-101; (for searching for best MS-TRS combinations)

9-class body motion: Hula hoop, mopping floor, body weight squat, …

13-class subtle motion: Apply eye makeup, apply lipsticks, brushing teeth, …

Results:

Training / validation (%)	9-class body	13-class subtle	22-class indoor
s346, l19	90.5 / 83.4	86.1 / 76.4	88.6 / 72.8

s346, l19 is best for cost & performance.

Salient / body motion > subtle / local motion.

More testing results

We focus on indoor actions with stationary cameras.

Datasets: 1) 22 classes selected from UCF-101; (for searching for best combinations)

2) 2 UCF classes + 8 classes from NTU RGB-D dataset (we only use RGB).

"jumping jack" & "body weight squat" are from UCF-101

Testing protocol: video-wise. 3 spatial_scales x 5 time_intervals = 15 clips are sampled for each testing video.

Features: *s346, l19* (19-frame clip, compute TRS at strides of {3, 4, 6})

Results:

predicted class (columns) / true class (rows)

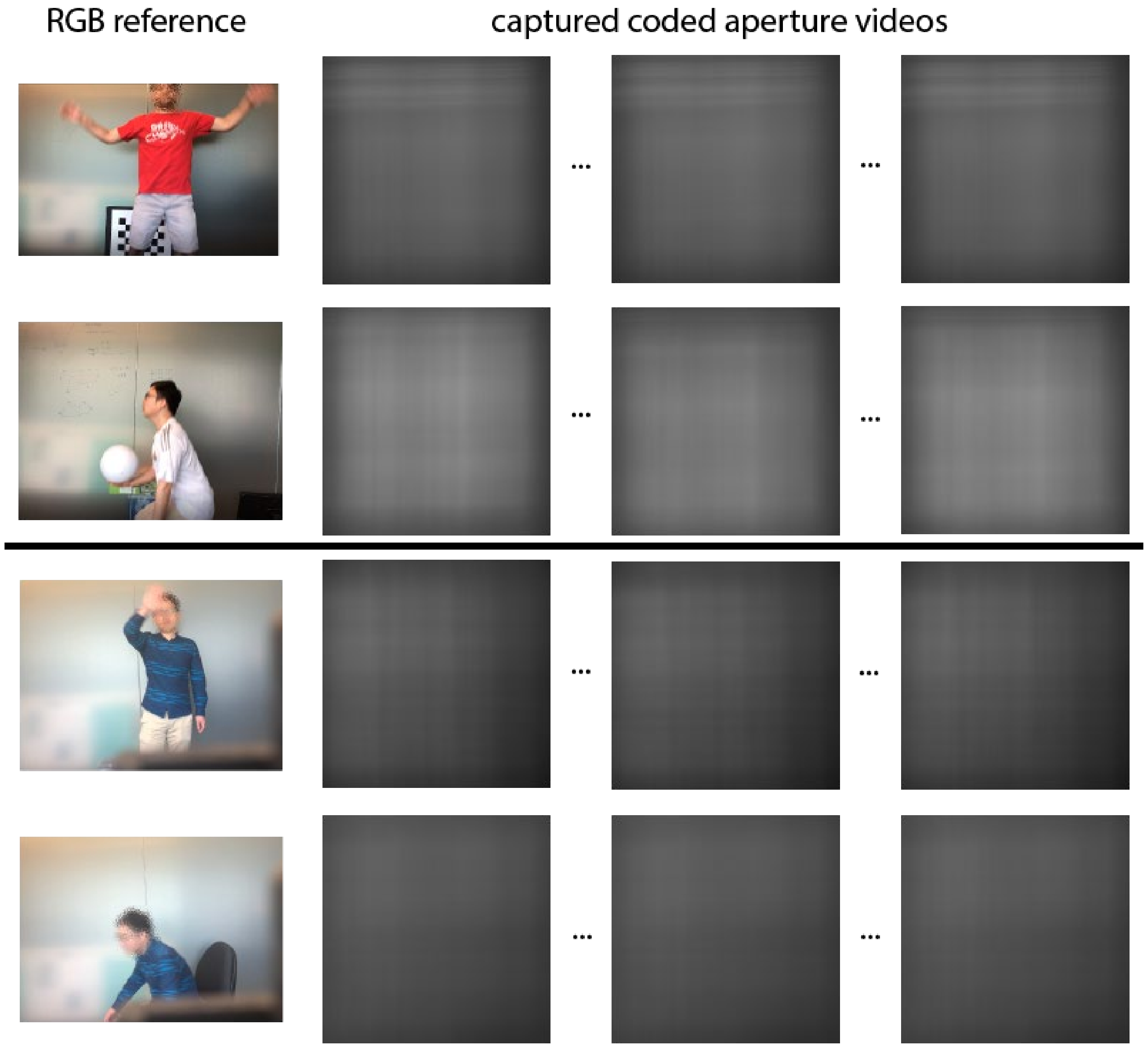
	1	2	3	4	5	6	7	8	9	10
1	97.1	0.0	2.9	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	0.0	94.3	0.0	0.0	0.0	0.0	2.9	2.9	0.0	0.0
3	0.0	8.6	91.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0
4	0.0	10.8	2.7	81.1	5.4	0.0	0.0	0.0	0.0	0.0
5	0.0	0.0	3.3	20.0	76.7	0.0	0.0	0.0	0.0	0.0
6	0.0	28.6	8.6	0.0	0.0	57.1	5.7	0.0	0.0	0.0
7	0.0	37.1	5.7	0.0	0.0	5.7	51.4	0.0	0.0	0.0
8	2.9	51.4	2.9	0.0	0.0	0.0	11.4	31.4	0.0	0.0
9	0.0	65.6	0.0	0.0	0.0	0.0	15.6	6.3	12.5	0.0
10	0.0	31.4	2.9	0.0	0.0	0.0	42.9	8.6	8.6	5.7

1	Hopping
2	Staggering
3	Jumping up
4	Jumping jack
5	Body weight squat
6	Standing up
7	Sitting down
8	Throw
9	Clapping
10	Handwaving

Salient body motion

Subtle local motion

Testing real CA videos



Successful classes:

- body weight squat (3 videos) 100% top-1
- jumping jack (5 videos) 100% top-2
- standing up (1 video) 100% top-3

Others (e.g. 8 “handwaving” and 2 “sitting down”) are unsuccessful

Limitation & Future work:

Domain gap between training (synthetic CA) and testing (real CA); fancy forward simulation (diffraction effects)

No existing real CA dataset available, collecting one might be worthy.

Concluding remarks

- Lens-free coded aperture cameras are useful for privacy preserving vision.
  - Captured data is visually incomprehensible.
  - The encoding mask / PSF is required to perform reconstruction.
  - The masks can be randomly generated for each camera.
    - For hackers to obtain the PSF, they need to break into the room and light a point source.
- MS-TRS for privacy-preserving motion features.
  - Non-invertible (phase correlation).
  - Mask-invariant (Different masks result in the same features).
    - True for T features.
    - RS features can be achieved by training with varying masks.

Thank you!