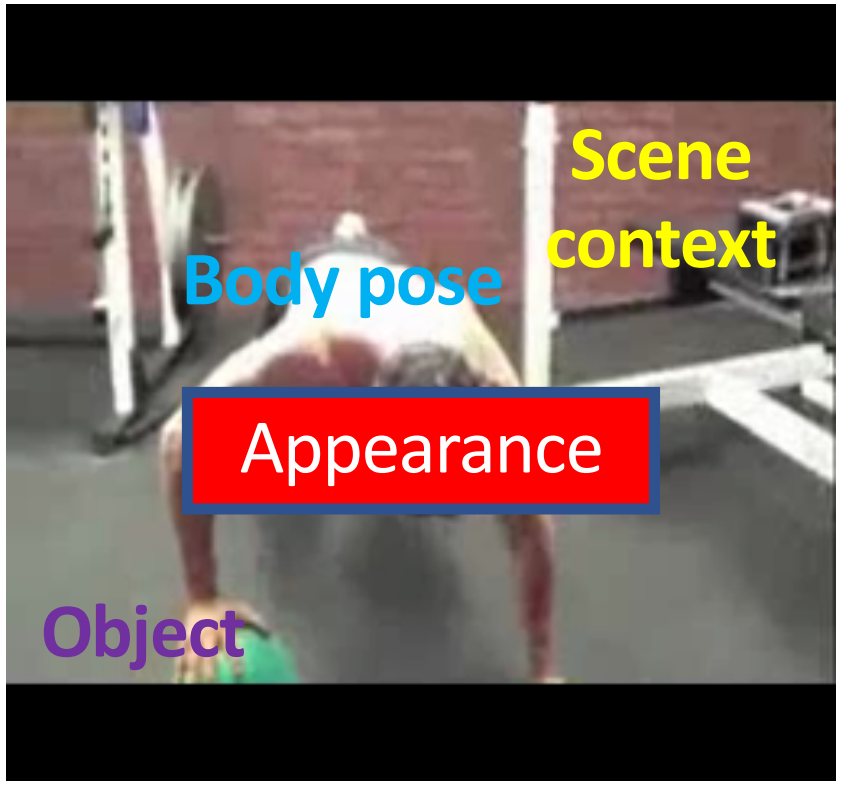
CVPR 2018, June 2018

# Im2Flow: Motion Hallucination from Static Images for Action Recognition

Ruohan Gao, Bo Xiong, Kristen Grauman
The University of Texas at Austin

# Video-level Action Recognition

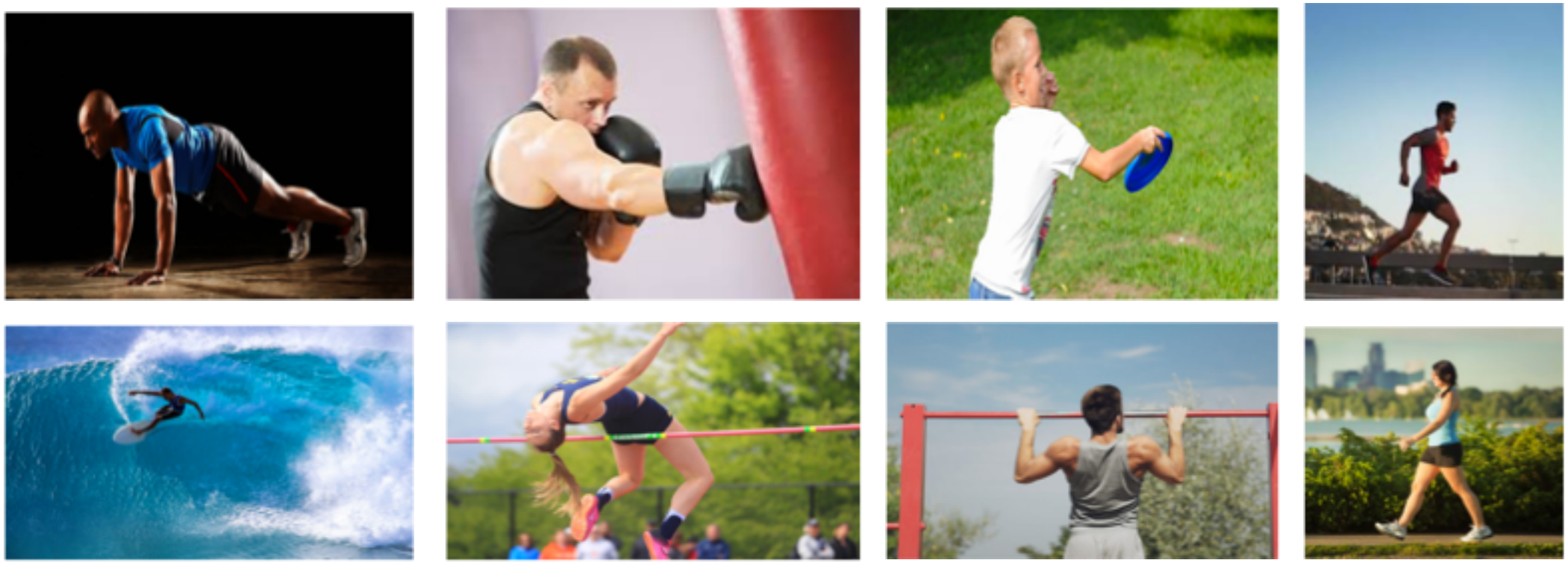




Video-level action recognition methods exploit both appearance and motion:

[Simonyan & Zisserman, NIPS 2014 ; Ji *et al.* TPAMI 2013; Karpathy *et al.* CVPR 2014; Ng *et al.* CVPR 2015; Donahue *et al.* CVPR2015; Fernando *et al.* CVPR 2015; Feichtenhofer *et al.* CVPR 2016; Wang *et al.*, ECCV 2016; Varol *et al.* TPAMI 2017; Girdhar *et al.*, CVPR 2017; Tran *et al.* CVPR 2018; Feichtenhofer *et al.*, CVPR 2018]

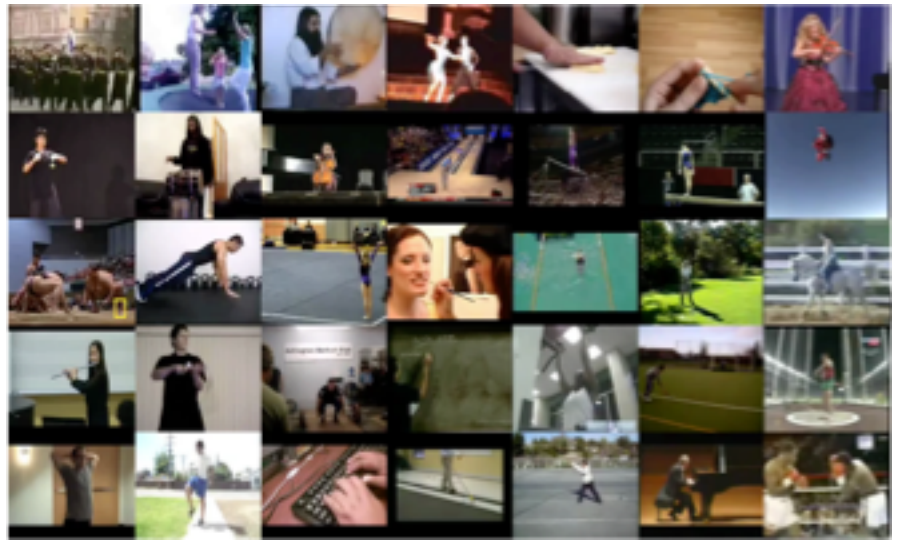
# Static-Image Action Recognition



Without motion, static-image action recognition methods exploit various high-level cues, such as body pose, objects and scene context.

[Thurau & Hlavác, CVPR 2008; Gupta *et al.* TPAMI 2009; Delaitre *et al.* BMVC 2010; Maji *et al.* CVPR 2011; Yao *et al.* ICCV 2011; Delaitre *et al.*, NIPS 2011; Prest *et al.* TPAMI 2012; Sener *et al.*, ECCV 2012; Chen & Grauman CVPR 2013; Sharma *et al.* CVPR 2013; Chao *et al.* ICCV 2015; Gkioxari *et al.*, CVPR 2015]

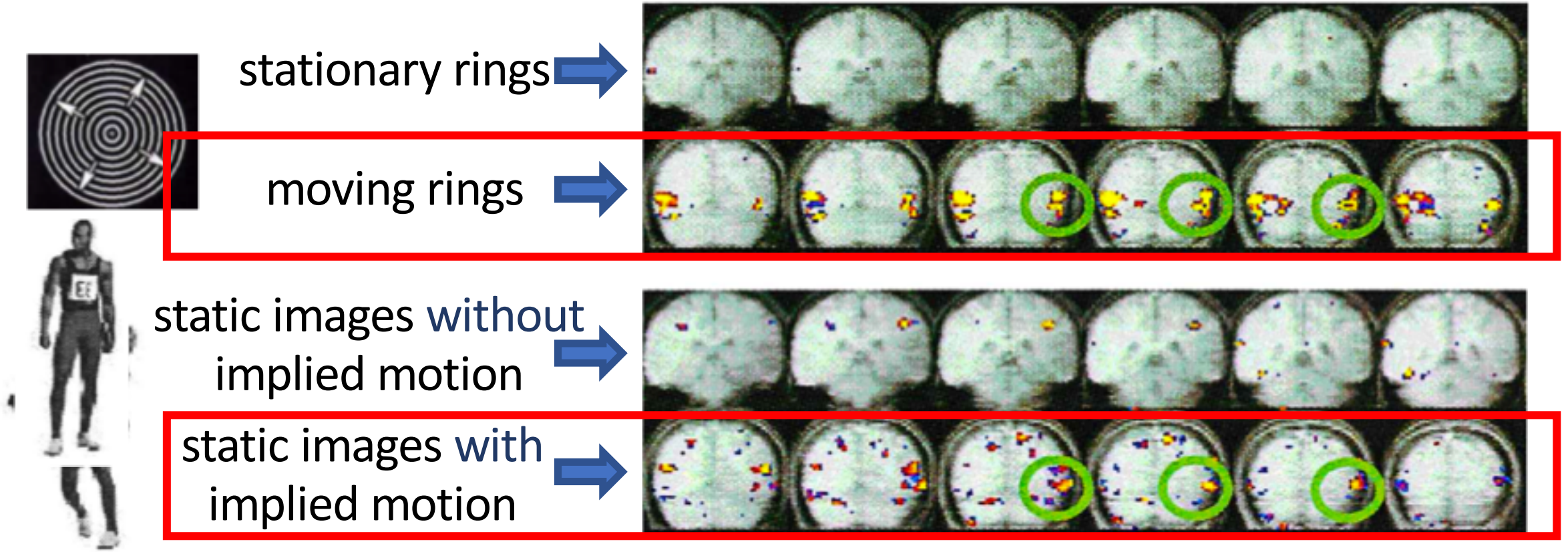
# But is motion really absent in static images?

# Implied Motion Perception in the Brain

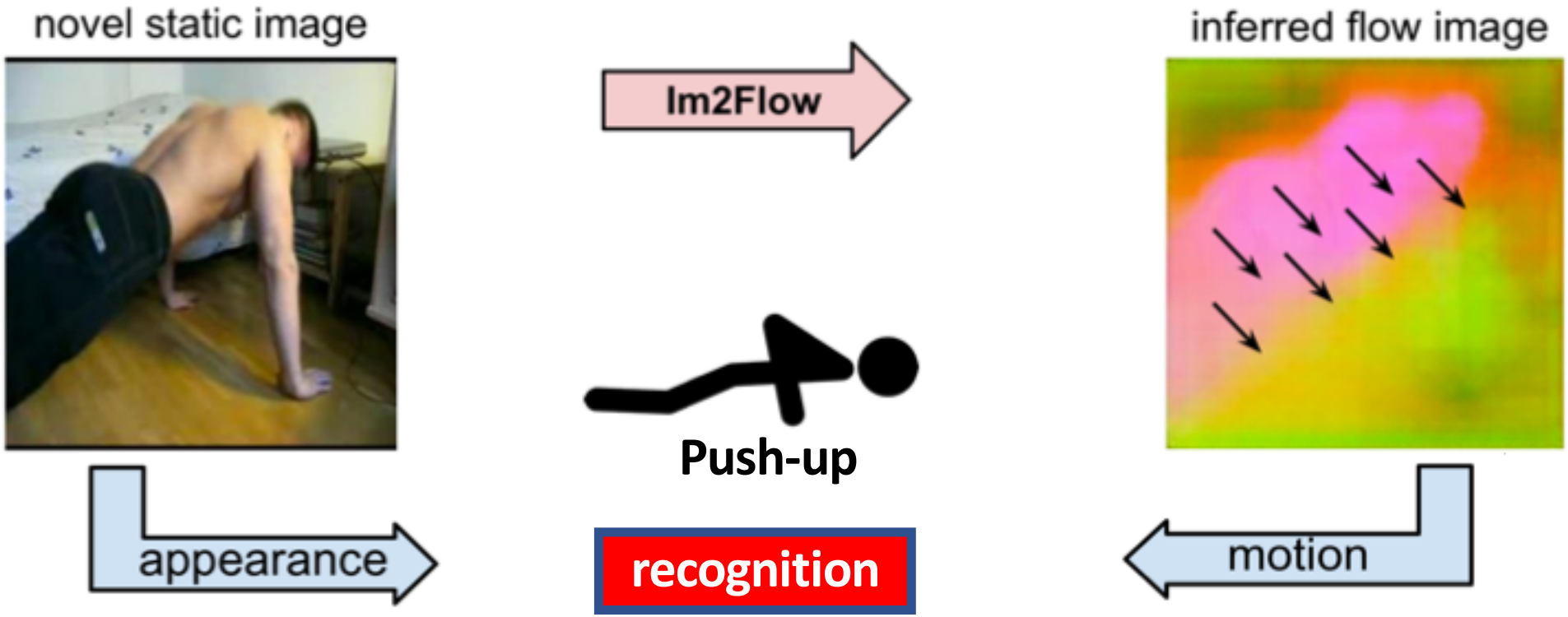
[Kourtzi & Kanwisher, 2000]

Activation in human's MT/MST cortex by static images with implied motion



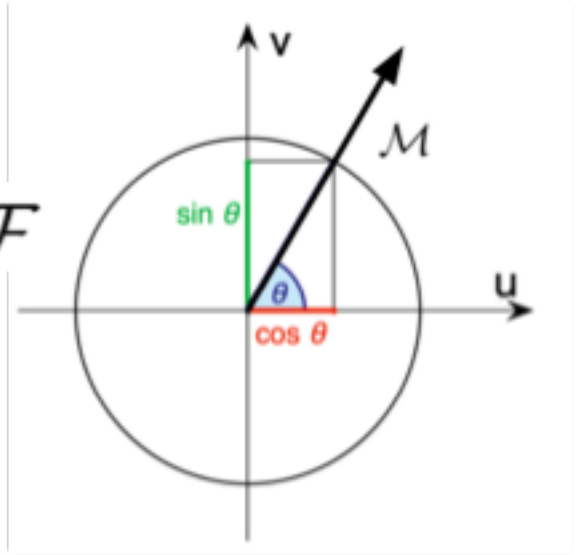
# Our Idea

learn a motion prior from unlabeled videos

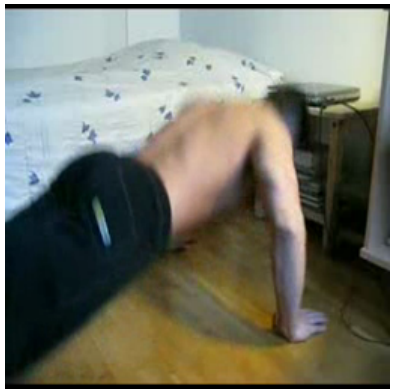


# Problem Formulation

- Encode optical flow as a single 3-channel flow image $\mathcal{F}$

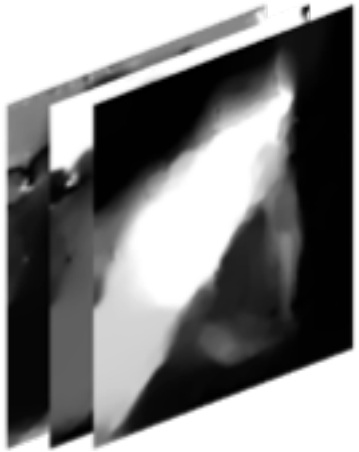



$$\mathcal{F}_1 = \sin(\theta) = \frac{v}{\mathcal{M}}; \quad \mathcal{F}_2 = \cos(\theta) = \frac{u}{\mathcal{M}}; \quad \mathcal{F}_3 = \mathcal{M}$$



Im2Flow

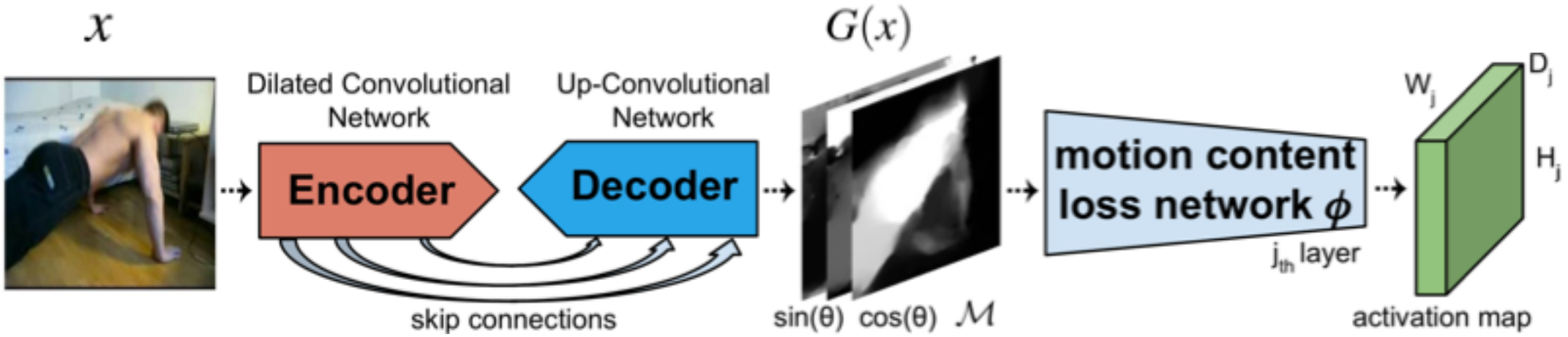
G



sin(θ) cos(θ) $\mathcal{M}$

- $\mathrm{X}$: static images containing an action, $\mathrm{Y}$: flow images
- Goal: mapping G from $\mathrm{X} \rightarrow \mathrm{Y}$ to infer flow from an individual image

# Im2Flow Network Architecture




- Total loss: $L = L_{pixel} + \lambda L_{content}^{\phi,j}$
- Pixel error loss: $L_{pixel} = \mathbb{E}_{p,q \in \{x_i, y_i\}_{i=1}^{N}} [||y_i - G(x_i)||_2]$
- Motion content loss: $L_{content}^{\phi,j} = \frac{1}{D_j \times H_j \times W_j} \mathbb{E}_{p,q \in \{x_i, y_i\}_{i=1}^{N}} [||\phi_j(y_i) - \phi_j(G(x_i))||_2]$

# Flow Prediction Results

Evaluate on test set from UCF-101, HMDB-51, and Weizmann datasets

- all pixels in the whole image
- masks on canny edges
- masks on foreground (FG) regions

| UCF-101 | EPE ↓ | EPE-Canny | EPE-FG | DS ↑ | DS-Canny | DS-FG | OS ↑ | OS-Canny | OS-FG |
|---|---|---|---|---|---|---|---|---|---|
| Pintea *et al.* 2014 | 2.401 | 2.699 | 3.233 | -0.001 | -0.002 | -0.005 | 0.513 | 0.544 | 0.555 |
| Walker *et al.* 2015 | 2.391 | 2.696 | 3.139 | 0.003 | 0.001 | 0.014 | 0.661 | 0.673 | 0.662 |
| Nearest Neighbor | 3.123 | 3.234 | 3.998 | -0.002 | -0.001 | -0.023 | 0.652 | 0.651 | 0.659 |
| Ours | **2.210** | **2.533** | **2.936** | **0.143** | **0.135** | **0.137** | **0.699** | **0.692** | **0.696** |

End-Point-Error
(lower better)

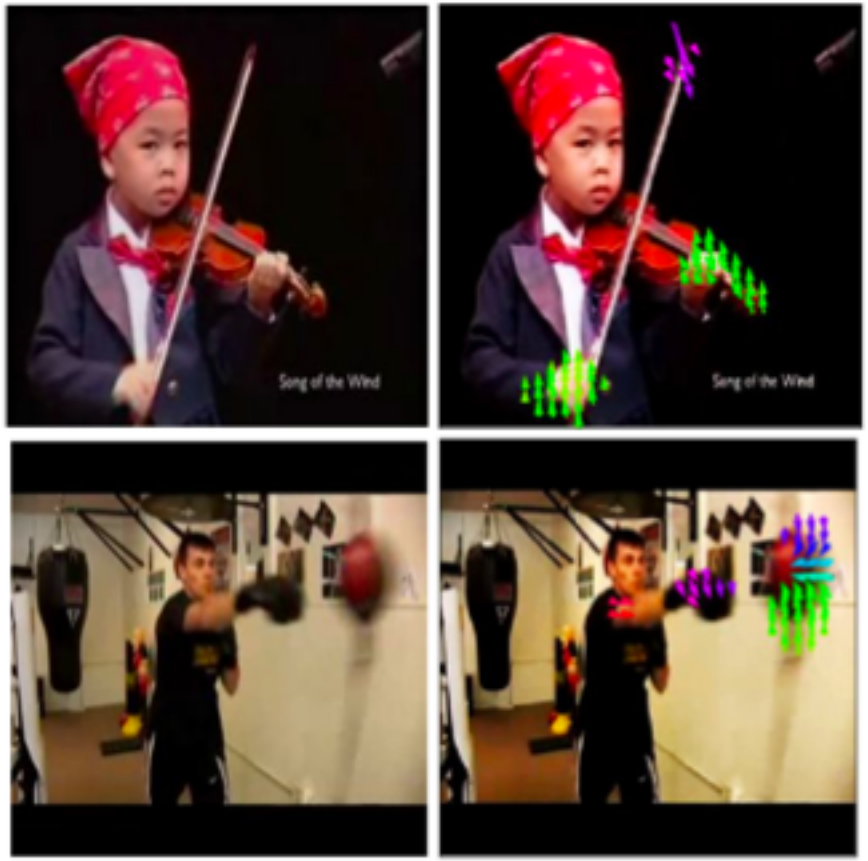
Direction Similarity
(higher better)

Orientation Similarity
(higher better)

# Qualitative Results

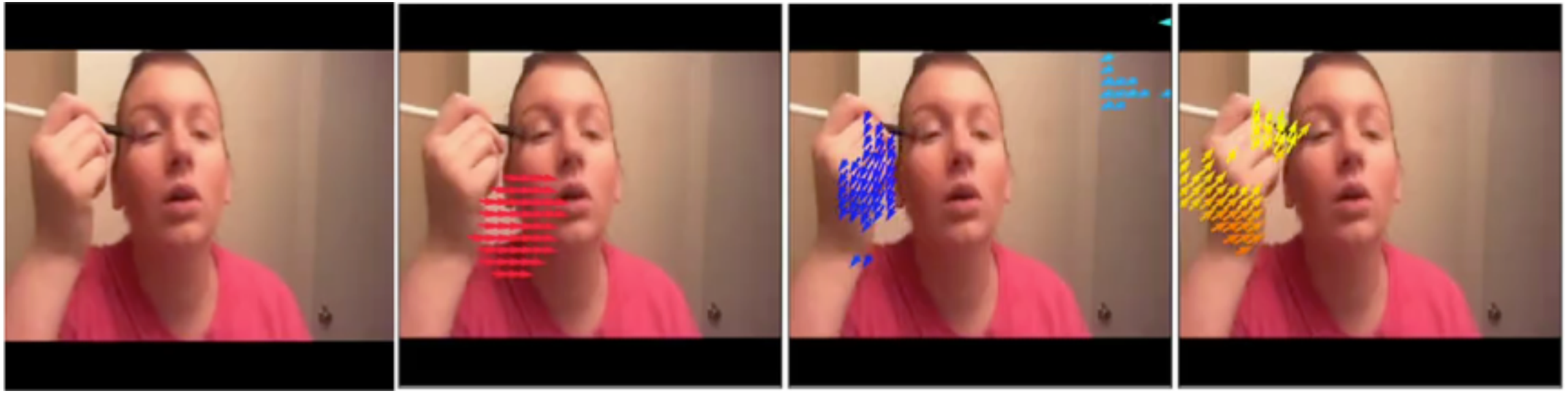
Input image Prediction

Input image Prediction



# Qualitative Results

Input image | Walker et al. 2015 | Im2Flow (Ours) | Ground-truth Flow



***Per-image*** flow prediction results without temporal smoothing.

# Failure Case

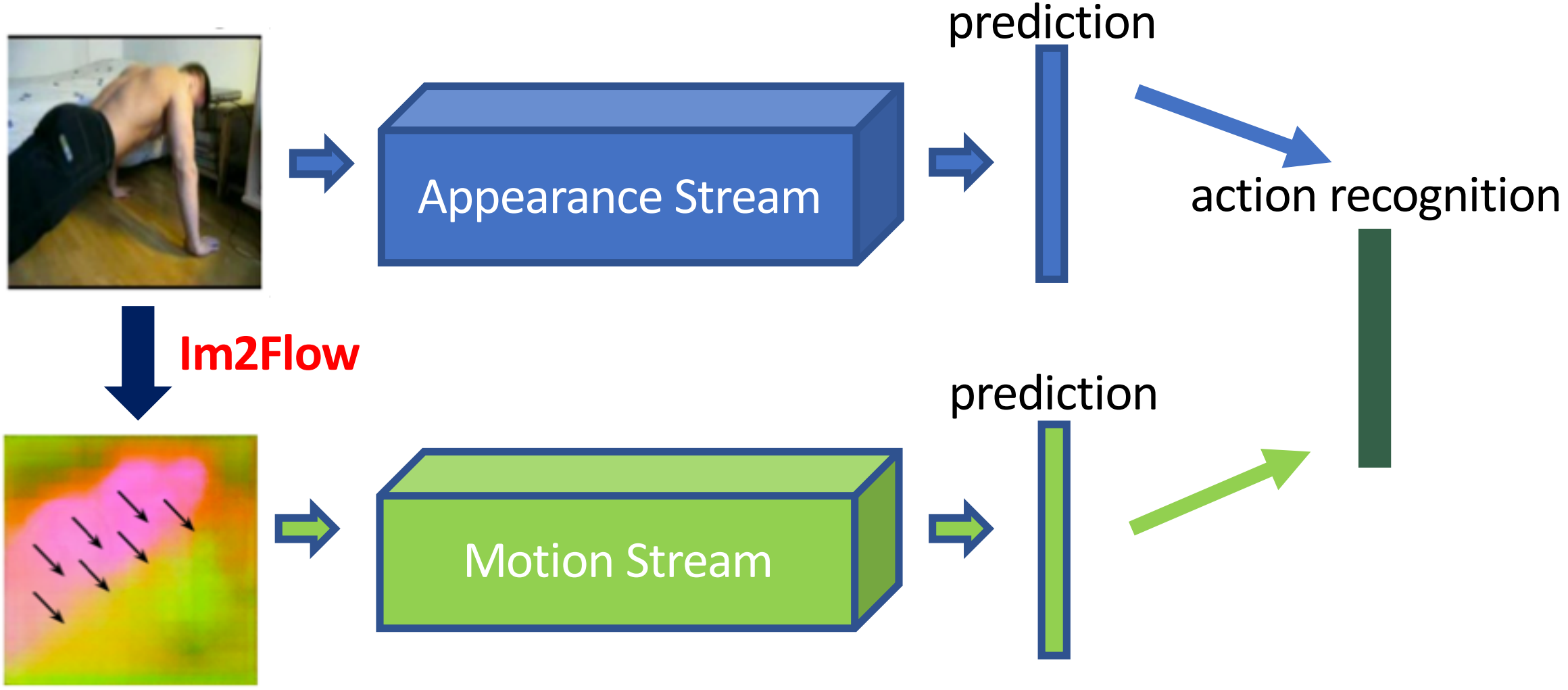
Input image | Walker et al. 2015 | Im2Flow (Ours) | Ground-truth Flow

- Background is too diverse
- Motion present in static images is too subtle

# Motion Potential

Which images are the most suggestive of coming events?

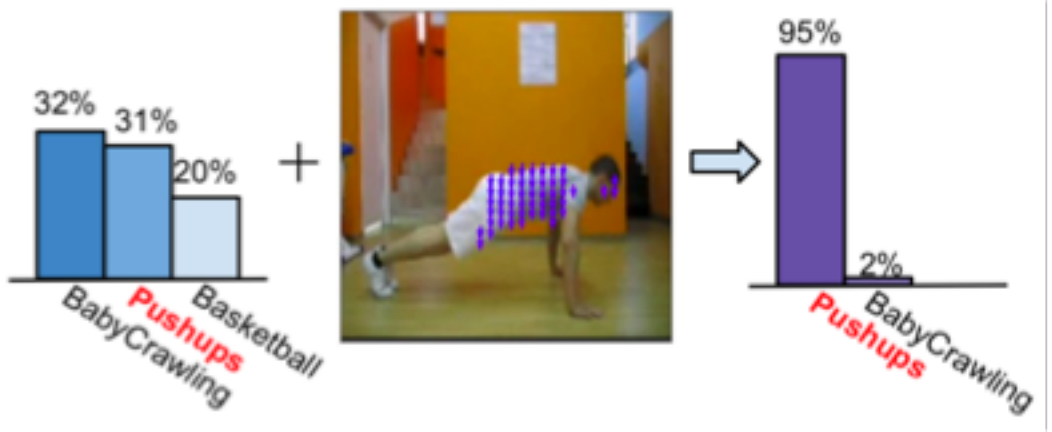
# Static-Image Action Recognition

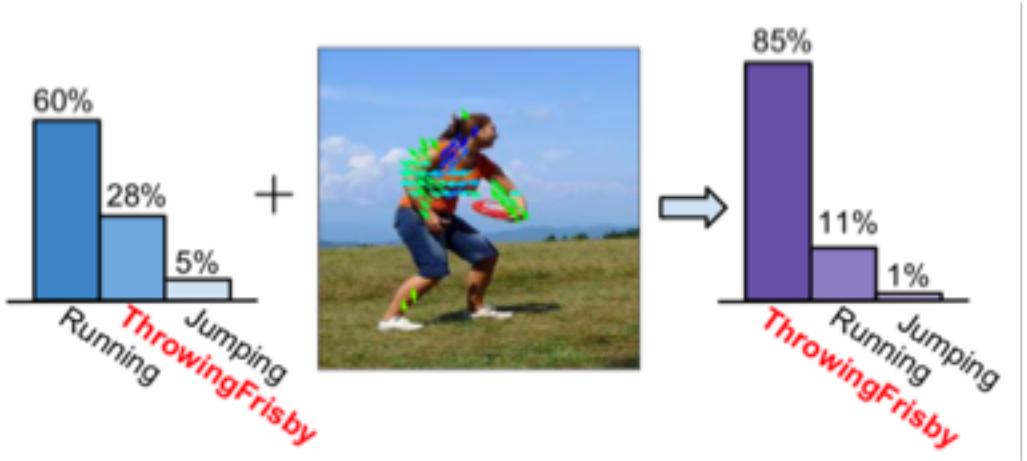


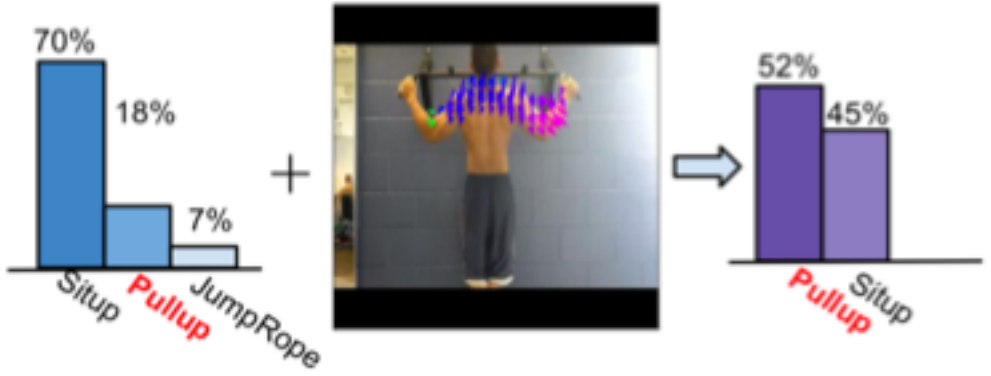

[Simonyan & Zisserman, NIPS 2014]

# How does the inferred motion help recognition?

# How does the inferred motion help recognition?

# Static-Image Datasets

**UCF-static**

13,320 images of 101 classes
[Soomro *et al.* 2012]

**HMDB-static**

5,100 images of 51 classes
[Kuehne *et al.* ICCV 2011]

**PennAction**

2,326 images of 15 classes
[Zhang *et al.* ICCV 2013]

**Willow**


911 images of 7 classes
[Delaitre *et al.* BMVC 2010]

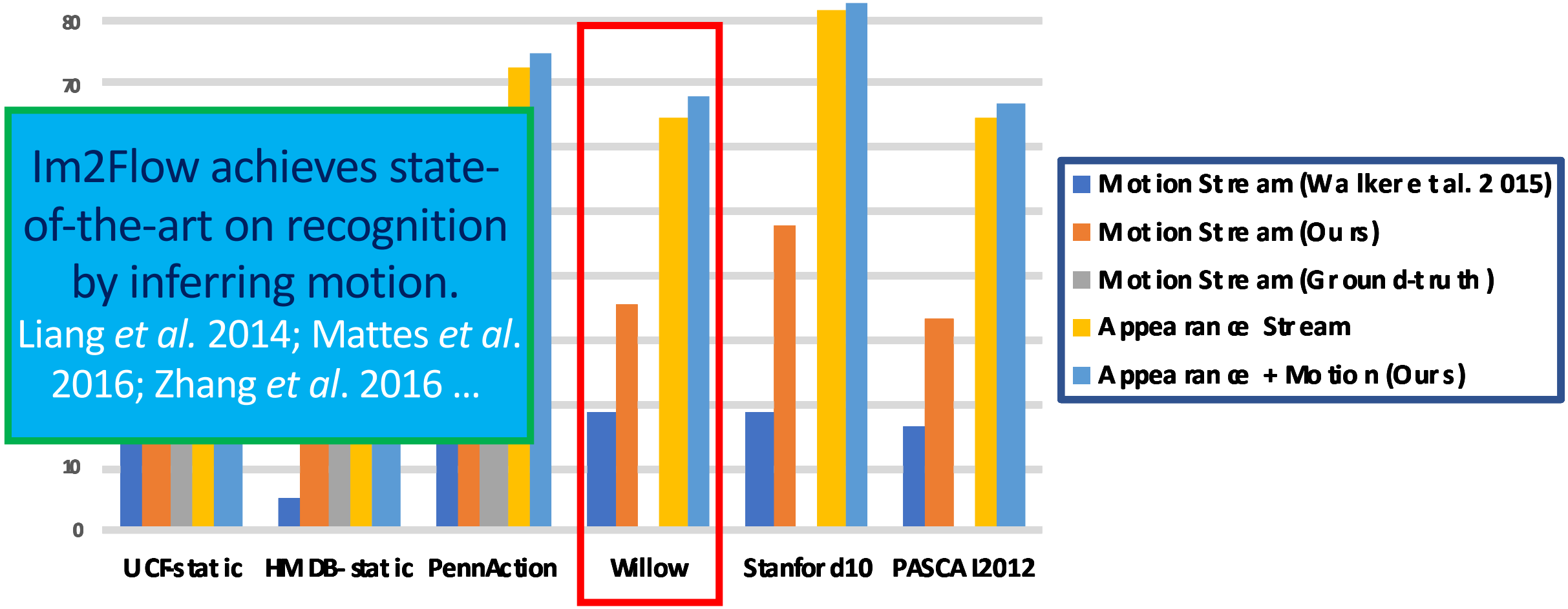
**Stanford 10**


2,672 images of 10 classes
[Yao *et al.* ICCV 2011]

**PASCAL 2012 Actions**


5,303 images of 10 classes
[Everingham *et al.* 2012]

# Action Recognition Results




Im2Flow achieves state-of-the-art on recognition by inferring motion.

Liang *et al.* 2014; Mattes *et al.* 2016; Zhang *et al.* 2016 …

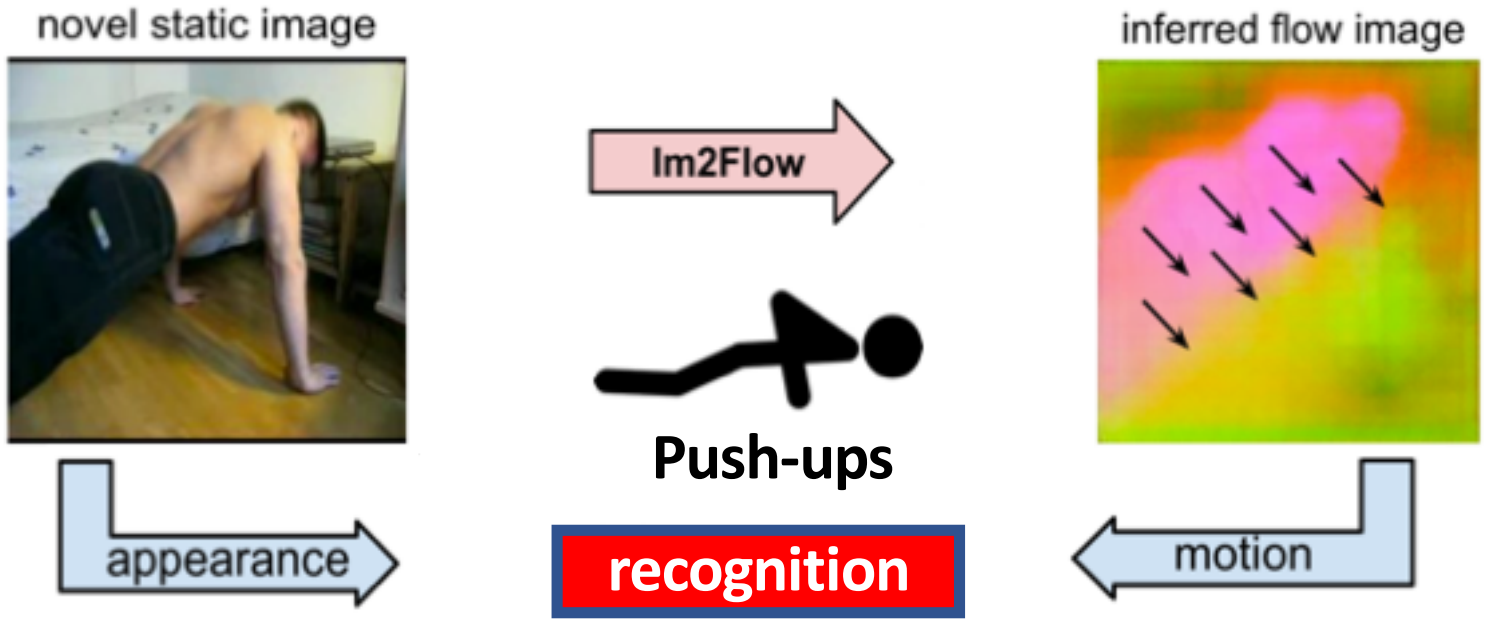
- Inferred motion from our Im2Flow framework works well for recognition
- 1-6% relative gain for ours vs. appearance stream

# Static-Image Dynamic Scene Recognition

- YUP++ Dynamic Scenes Dataset [Feichtenhofer *et al.* CVPR 2017]
  — waterfall, falling trees, rushing river, railway ……

# Conclusion

- Hallucinate the motion implied by a single snapshot
- Im2Flow network for dense flow prediction
- Enhance recognition of actions and dynamic scenes




**Poster: [A3]**

Project page: http://vision.cs.utexas.edu/projects/im2flow/