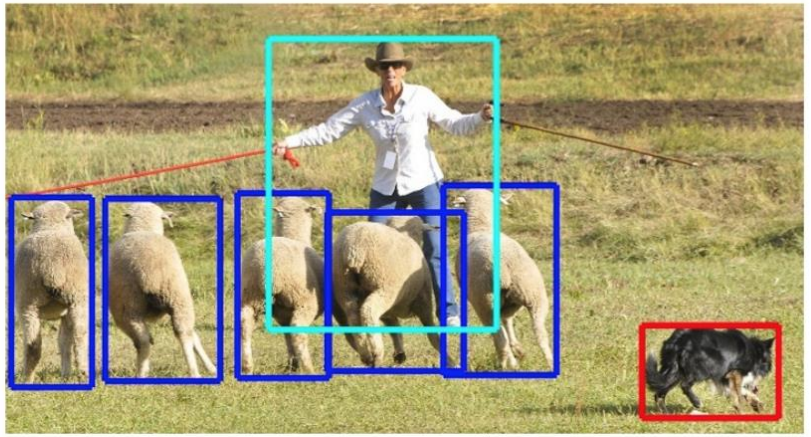
Im2Flow: Motion Hallucination from Static Images for Action Recognition

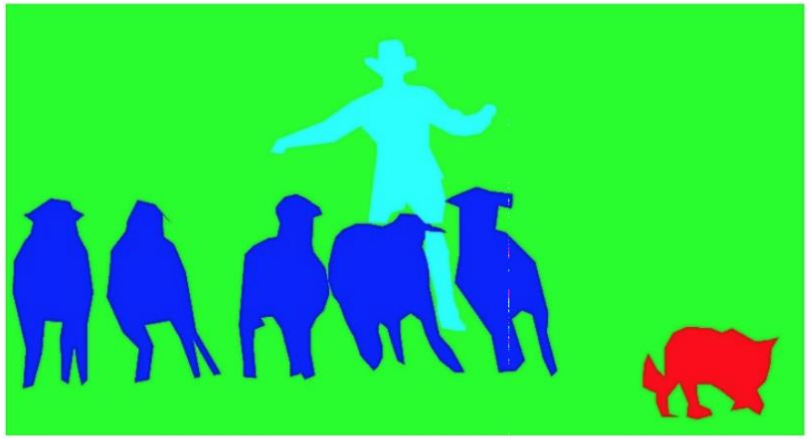
RUOHAN GAO BO XIONG KRISTEN GRAUMAN

Image Classification

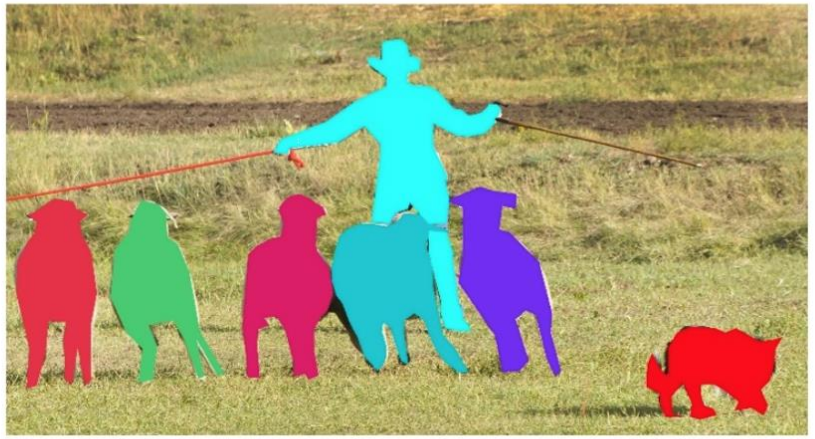


Object Detection/Localization

Action Recognition?
What is an action?



Semantic Segmentation



Instance Segmentation

Problem: Action Recognition

- Action is the most elementary human-surrounding interaction with a meaning.
- Multi-classification Problem
  - Input: Video or Image
  - Output: Labels (categories of actions)
  - Human Action Recognition

Input:

Output: Action 1, Action 2, ... Action N

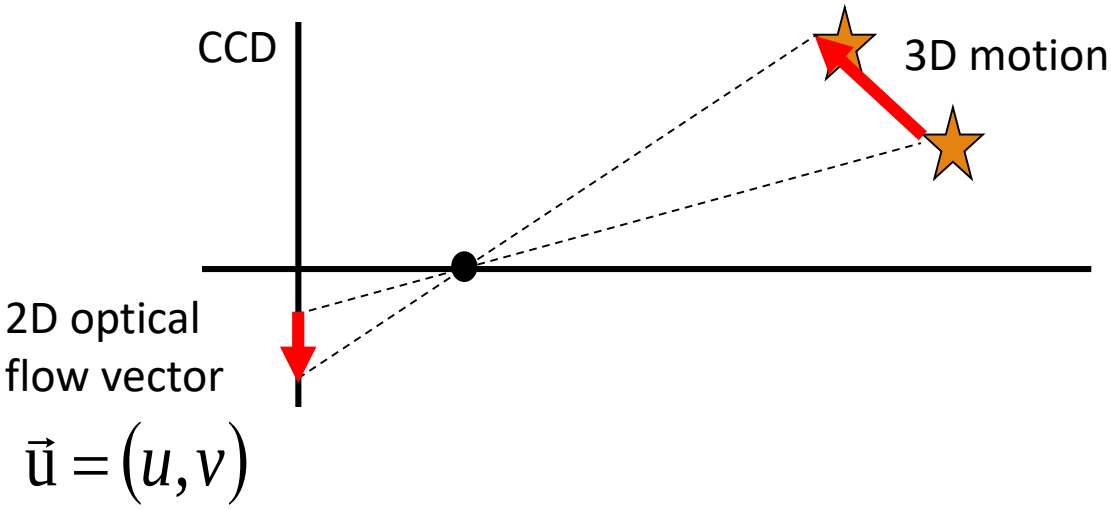
Running, Kicking and Jumping

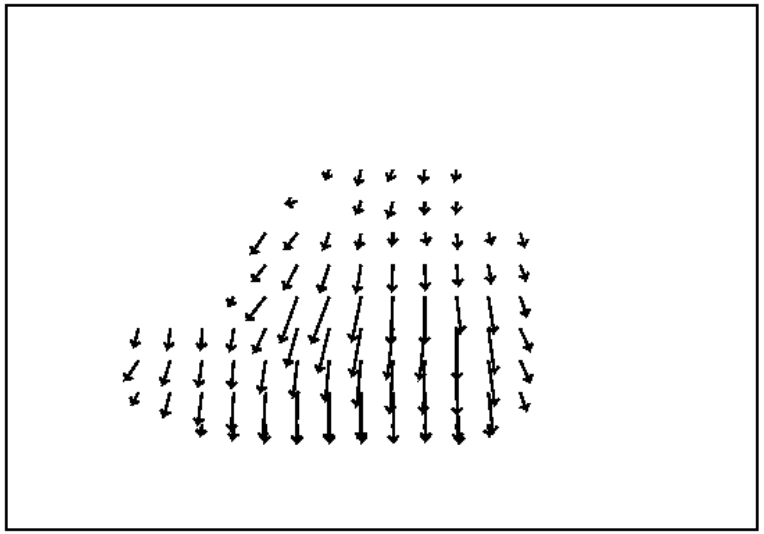
Video-based Action Recognition

- Classify human actions in video clips
- Simplification: **Trimmed Video with action lables**
  - Datasets: UCF101; HMDB51; MSR Action 3D;
- Temporal Action Detection/Localization: Untrimmed Video

Video-based Action Recognition

- Rich Temporal Information + Motion Information (**Optical Flow**)
- Motion field = real 3D scene motion
- **Optical flow** = projection of motion field, the apparent motion of brightness patterns
  - 2D vector represents Instantaneous velocity



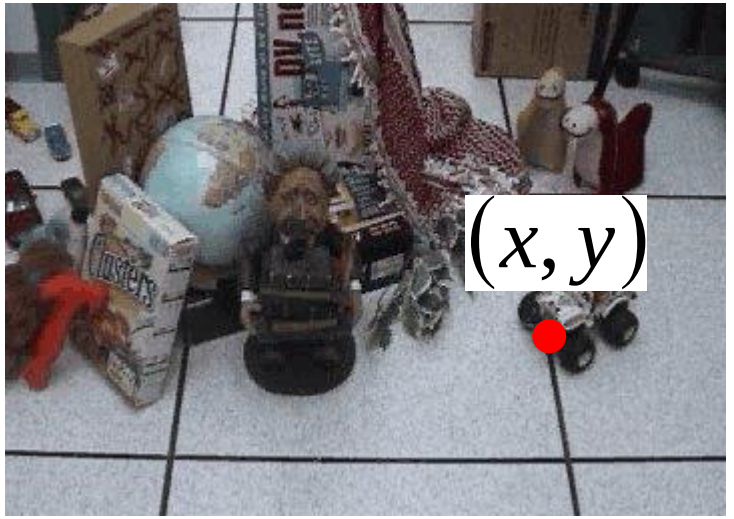


Pierre Kornprobst's Demo

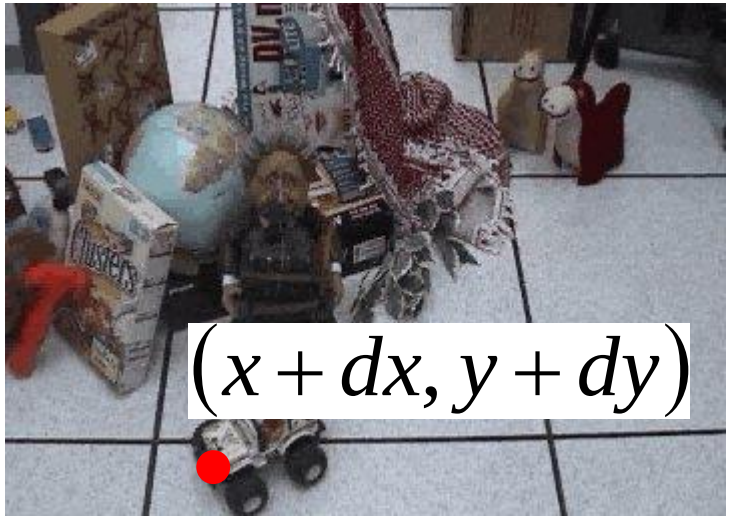
Optical Flow Estimation

- Brightness constant
- Motion is tiny
- Spatial consistency

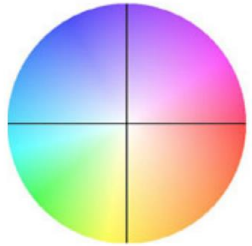
Time = *t*



Time = *t+dt*

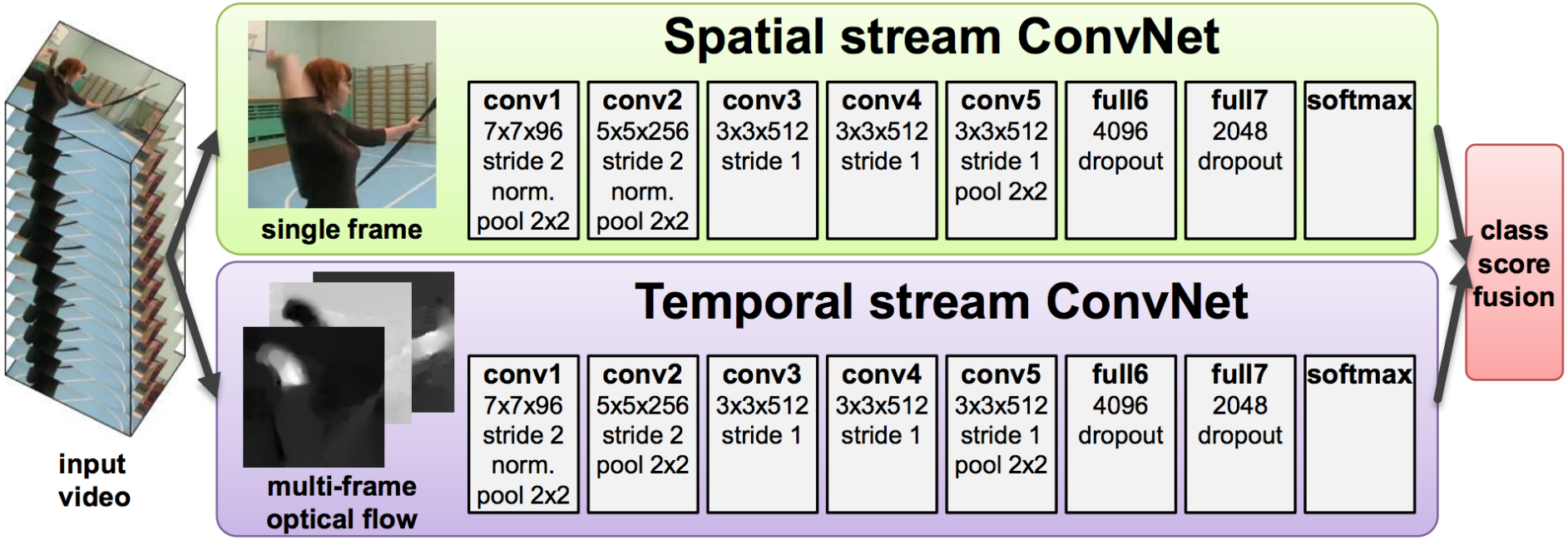


$$I(x, y, t) = I(x+dx, y+dy, t+dt)$$



Optical Flow and Action Recognition

- iDT（improved dense trajectories)
  - DT: OF > trajectories (HOF, HOG, MBH, trajectory) > FV（Fisher Vector） > SVM
  - iDT: matching using optical flow and SURF
- Two Stream Network (UCF101-88.0%，HMDB51-59.4%)

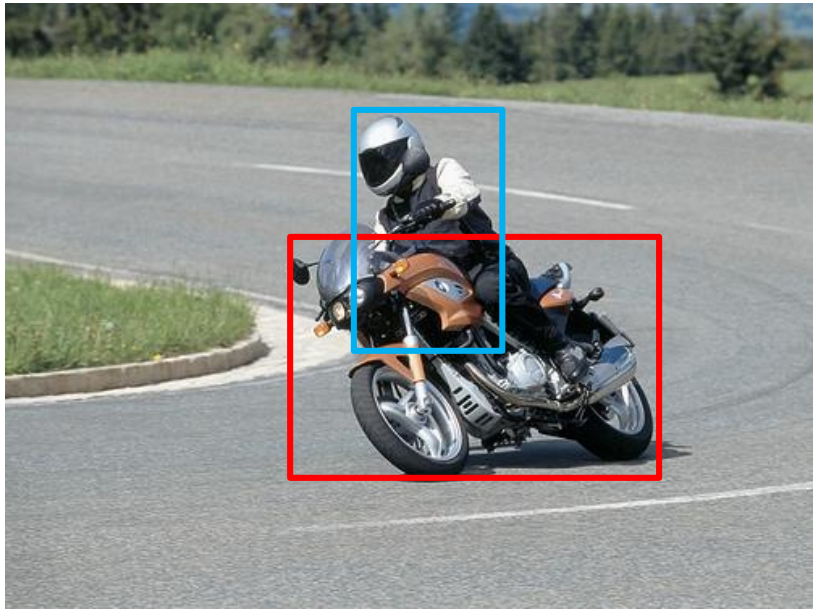


Why Optical Flow needed for Action Recognition?

- On the Integration of Optical Flow and Action Recognition
  - Invariant to appearance, even when the flow vectors are inaccurate.

Static Image Action Recognition

- Representation based solution
  - high-level cues: human body or body parts, objects , human-object interactions, and scene context
- **Big Issue!**
  - **No Temporal information? No Motion information?**



Solution: Motion Hallucination

- Train a U-Net (adapted) on Youtube data to learn motion (static frame > 5 predicted OFs)
- Losses: a pixel error loss and a motion content loss $L = L_{pixel} + \lambda L_{content}^{\phi,j}$
- two-stream CNN architecture

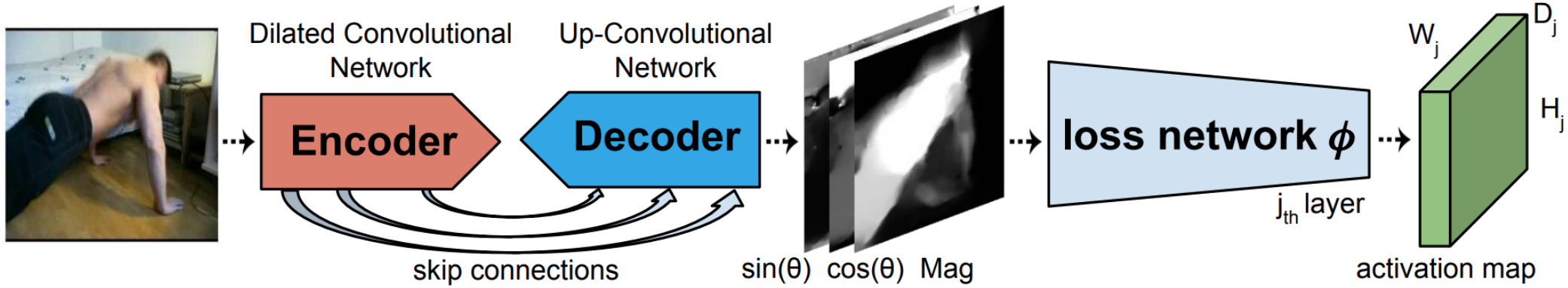


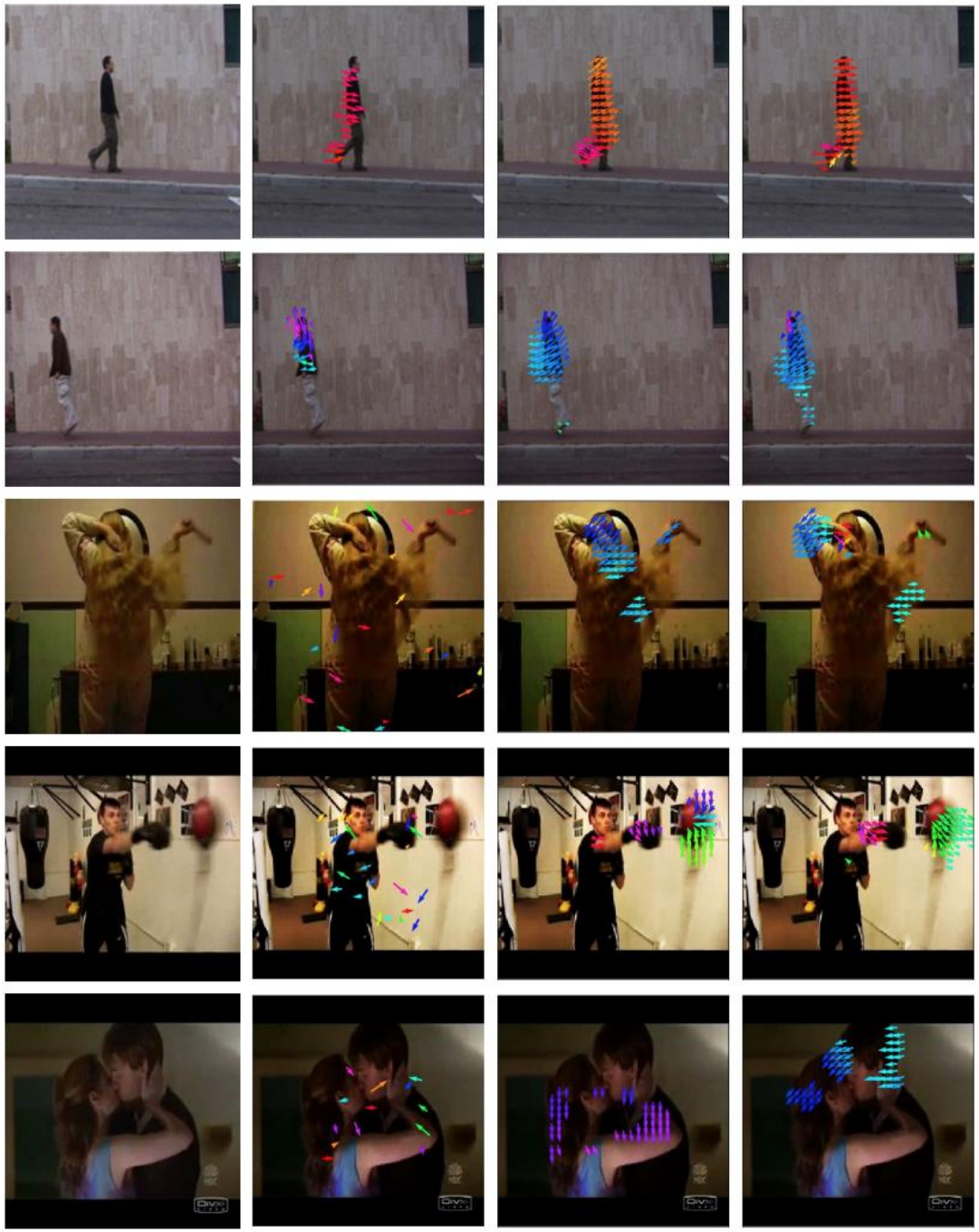
Figure 2. The network architecture of our Im2Flow framework. The network is an encoder-decoder that takes a static image as input and generates the corresponding 3-channel flow map $\mathcal{F}$ as output. Our training objective is a combination of the L_2 loss in the pixel space and in the deep feature space. A motion content loss network encourages the predicted flow image to preserve high-level motion features.

Flow Prediction

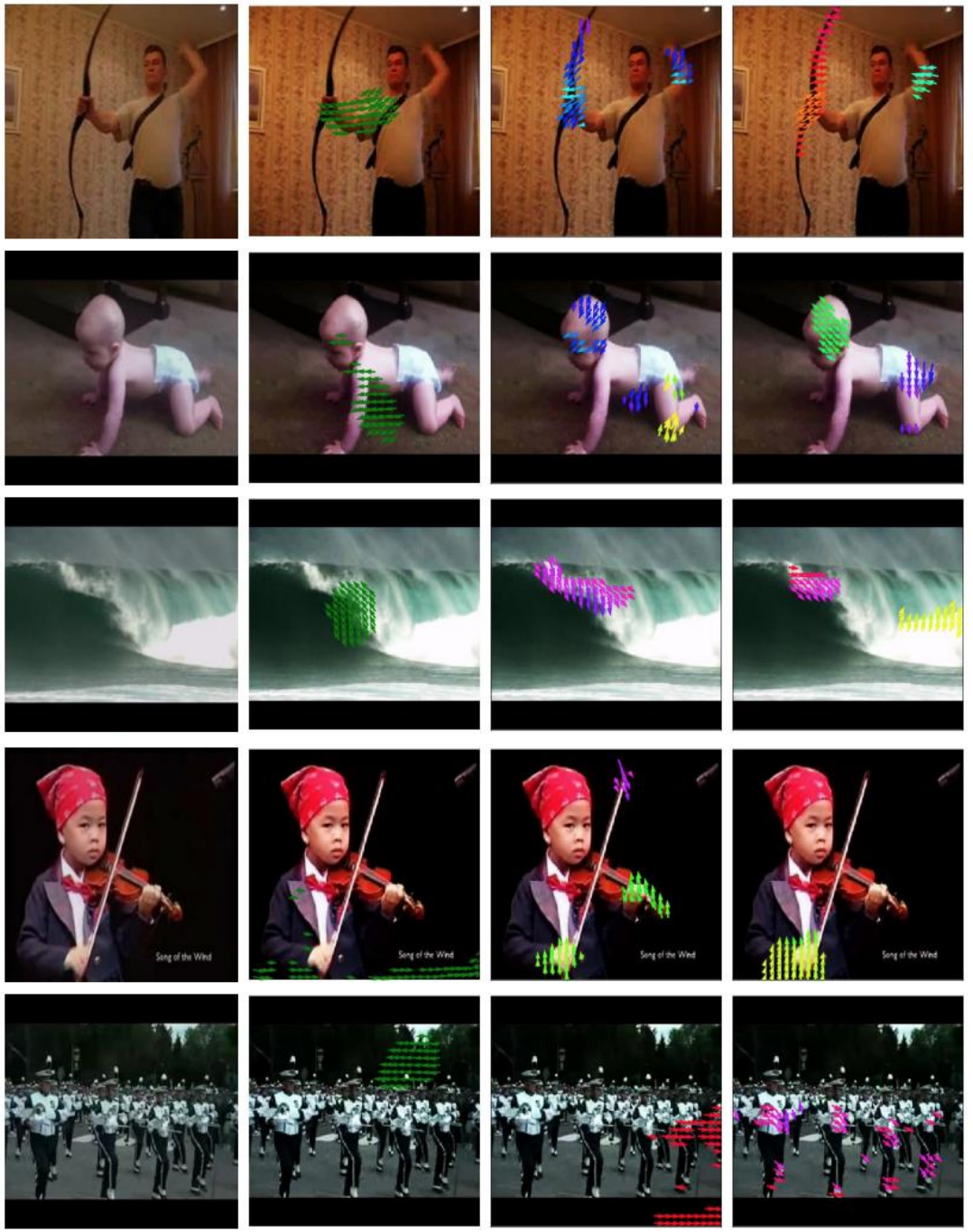
- 3 datasets: UCF-101, HMDB-51, and Weizmann.
- Evaluation metrics:
  - End-Point-Error (EPE) $\sqrt{(u_0 - u_1)^2 + (v_0 - v_1)^2}$
  - Direction Similarity (DS)
  - Orientation Similarity (OS)

UCF-101	EPE ↓	EPE-Canny	EPE-FG	DS ↑	DS-Canny	DS-FG	OS ↑	OS-Canny	OS-FG
Pintea *et al.* [58]	2.401	2.699	3.233	-0.001	-0.002	-0.005	0.513	0.544	0.555
Walker *et al.* [78]	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**

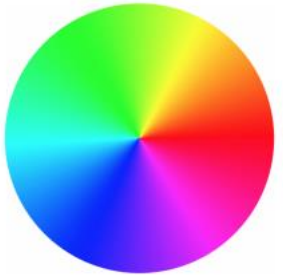
Quantitative results



(a) Input Image (b) Pintea *et al.* [58] (c) Ours (d) Ground-truth



(a) Input Image (b) Walker *et al.* [78] (c) Ours (d) Ground-truth

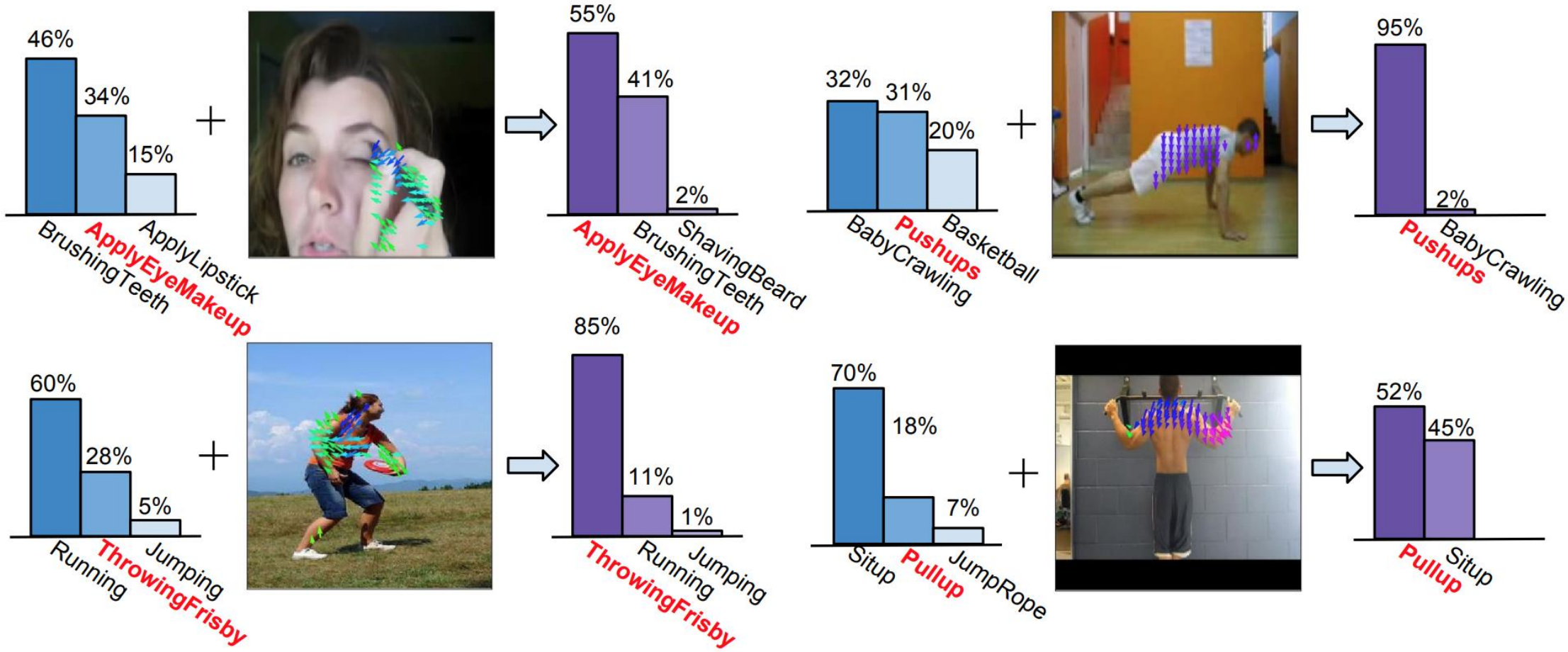


Action Recognition

- 3 static-image datasets (video datasets): **UCF-101, HMDB-51, Penn Action**
- 3 static-image action benchmarks: **Willow, Stanford10, PASCAL2012 Actions**
- **YUP++ Dynamic Scenes**

- 4 Baselines
  - Appearance Stream
  - Motion Stream (Ground-truth)
  - Motion Stream (Walker)
  - Appearance + Appearance

		UCF-static	HMDB-static	PennAction	Willow	Stanford10	PASCAL2012
	Appearance Stream	63.6	35.1	73.1	65.1	81.3	65.0
Motion Stream	Motion Stream (Walker *et al.* [78])	*14.3	4.96	21.2	18.8	19.0	15.9
Motion Stream	Motion Stream (Ours-UCF)	-	**13.9**	51.0	35.7	46.4	32.5
Motion Stream	Motion Stream (Ours-HMDB)	**24.1**	-	42.4	30.6	42.2	30.1
Motion Stream	Motion Stream (Ours-UCF+HMDB)	-	-	**51.1**	**35.9**	**48.4**	**32.7**
Motion Stream	→ Motion Stream (Ground-truth Motion)	38.7	20.0	52.4	-	-	-
Two-Stream	Appearance + Appearance	64.0	35.5	73.4	65.8	81.3	65.1
Two-Stream	Appearance + Motion (Walker *et al.* [78])	*64.5	35.9	73.1	65.9	81.5	65.0
Two-Stream	Appearance + Motion (Ours-UCF)	-	**37.1**	74.5	67.4	82.1	66.0
Two-Stream	Appearance + Motion (Ours-HMDB)	**65.5**	-	74.3	67.1	81.9	65.6
Two-Stream	Appearance + Motion (Ours-UCF+HMDB)	-	-	**74.5**	**67.5**	**82.3**	**66.1**
Two-Stream	→ Appearance + Motion (Ground-truth Motion)	68.1	39.5	77.4	-	-	-



inferred motion can help static image action recognition

	Accuracy	mAP
Appearance	74.3	79.3
Ground-truth Motion	55.5	62.0
Inferred Motion	30.0	37.0
Appearance + Appearance	75.2	79.8
Appearance + Inferred Motion	**78.2**	**82.3**
Appearance + Ground-truth Motion	79.6	83.6

Static-image action recognition results (in %) on the static-YUP++ dataset

	mAP (%)
Delaitre *et al.* [6]	59.6
Sharma *et al.* [65]	67.6
Khan *et al.* [41]	68.0
Zhang *et al.* [90]	77.0
Liang *et al.* [51]	80.4
Mettes *et al.* [57]	81.7
Ours (AlexNet as base network)	74.0
Ours (VGG-16 as base network)	87.2
Ours (ResNet-50 as base network)	**90.5**

Comparison to other recognition models on Willow

Conclusion

- Approach: hallucinate the motion from static image and use it as an auxiliary cue for action recognition
- state-of-the-art performance on optical flow prediction from an individual image
- Standard two-stream network to enhance recognition of actions and dynamic scenes by a good margin