Deep neural nets for human pose estimation in videos

Tomas Pfister, James Charles, Andrew Zisserman
Department of Engineering Science
University of Oxford
http://www.robots.ox.ac.uk/~vgg

Aim:

Estimate 2D upper body joint positions (wrist, elbow, shoulder, head) with high accuracy in real-time

Outline

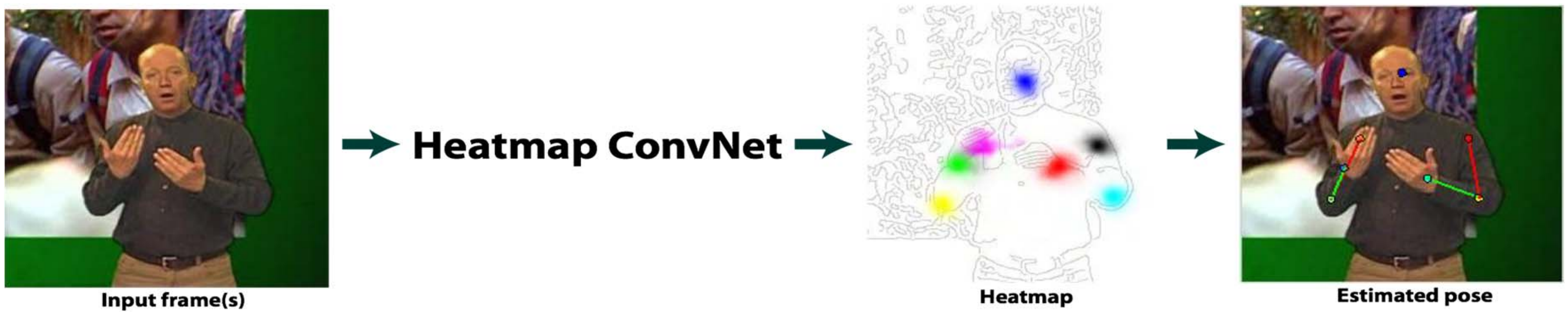
- Two types of loss functions for pose estimation
  - Coordinate net
  - Heatmap net
- Optical flow for pose estimation in videos
- Results (cf state of the art)

Method overview: single frame learning

<u>1. Coordinate Net</u>

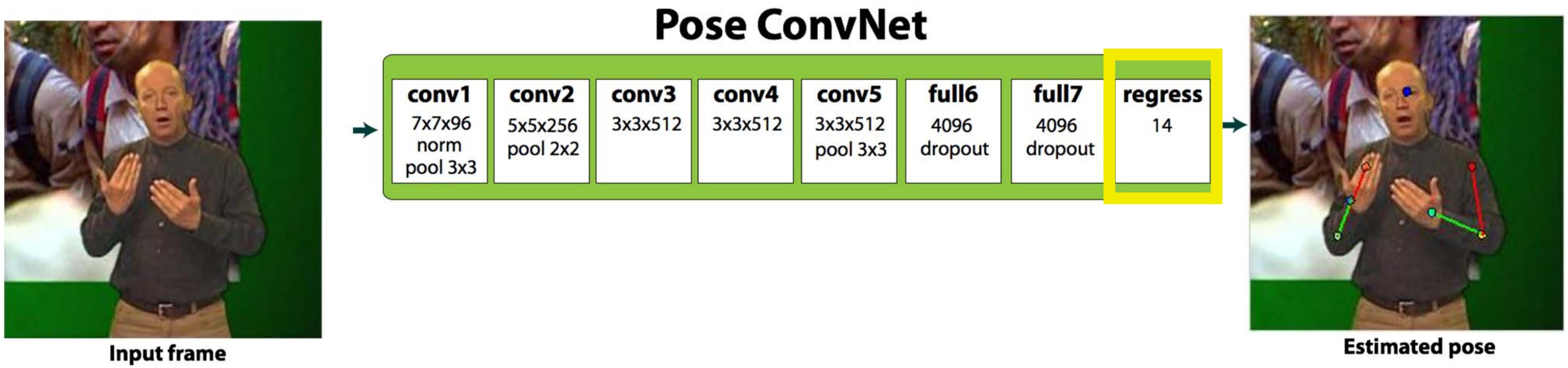
e.g. DeepPose CVPR14, Pfister et al ACCV14

<u>2. Heatmap Net</u>



e.g. Jain et al ICLR14, Tompson et al CVPR15

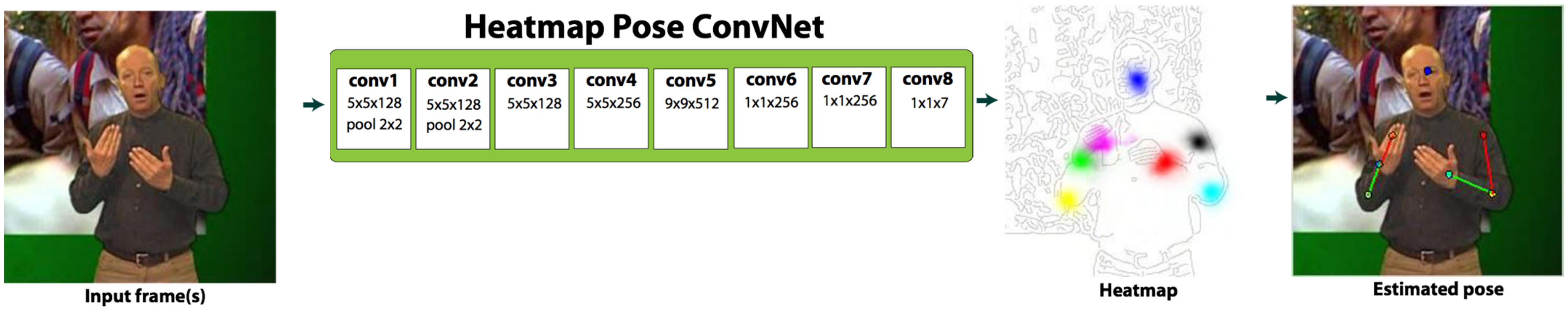
Coordinate Net: regress joint positions



Training loss: L2 on joint positions

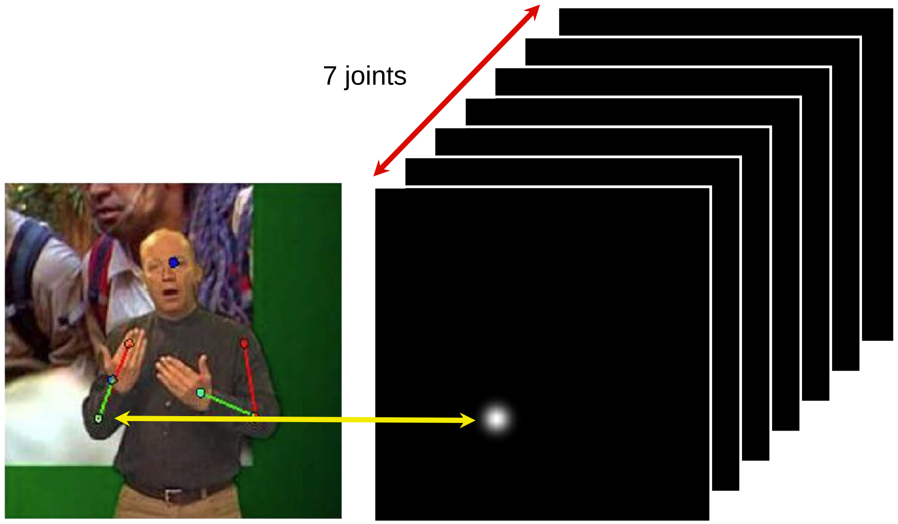
OverFeat like architecture

Heatmap Net: regress heatmap for each joint



256 x 256

64 x 64



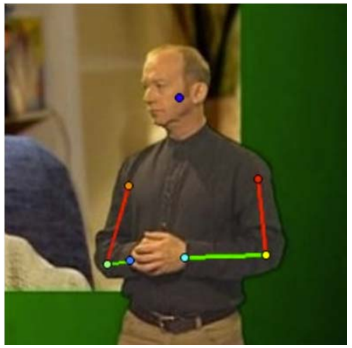
Represent joint position by Gaussian

Training loss: L2 on pixels

Comparison

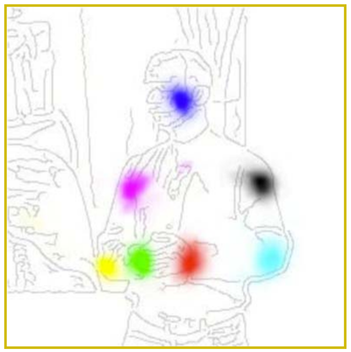
Regression target

Coordinate Net



Coordinates

Heatmap Net



Heatmap

BBC sign language videos data set

Training:

15 videos each 0.5-1hr long, all frames annotated

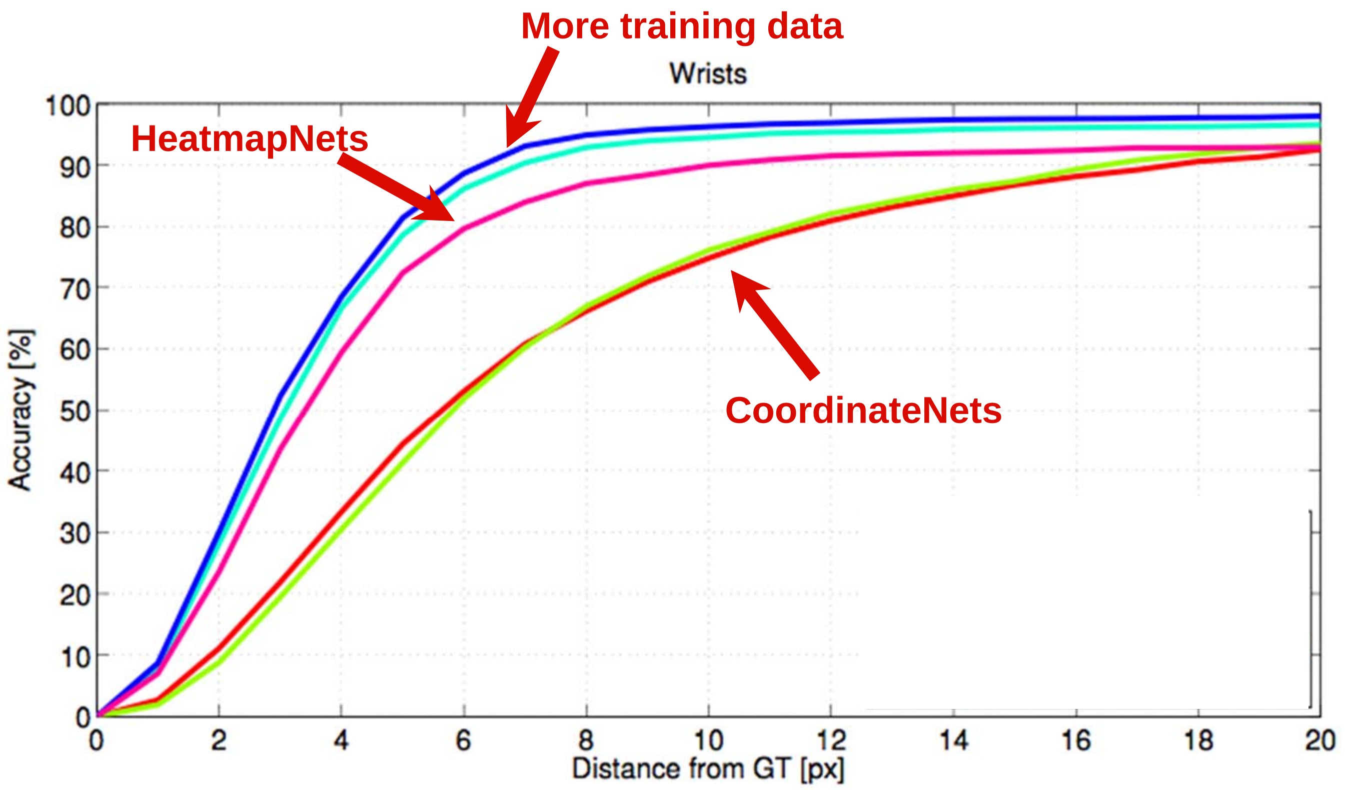
Testing:

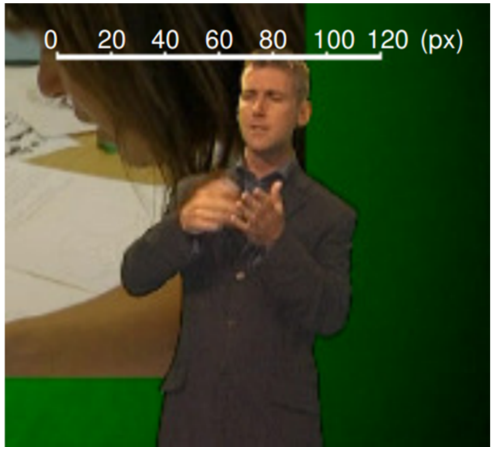
5 videos, 200 annotated frames per video

Extended Training:

72 videos with noisy automated annotations

Results on architecture comparison





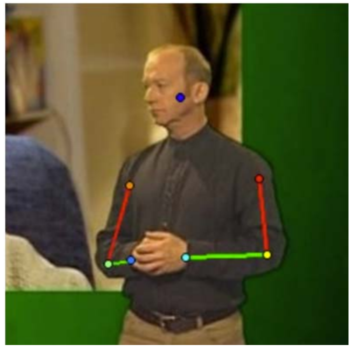
- Heatmap net superior to coordinate net
- Performance of coordinate net saturates with more training data

Evaluated on BBC Pose

Why is the heatmap network superior?

Regression target

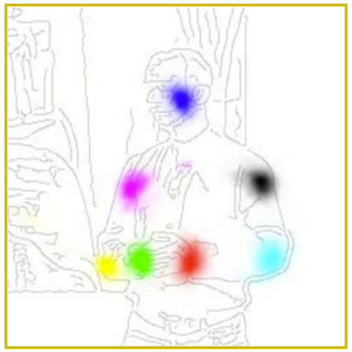
1. Can represent multimodal estimates, so can model uncertainty/confidence
2. In training there is an error signal from every pixel, so better smoothing for back propagation



Coordinates

Coordinate Net

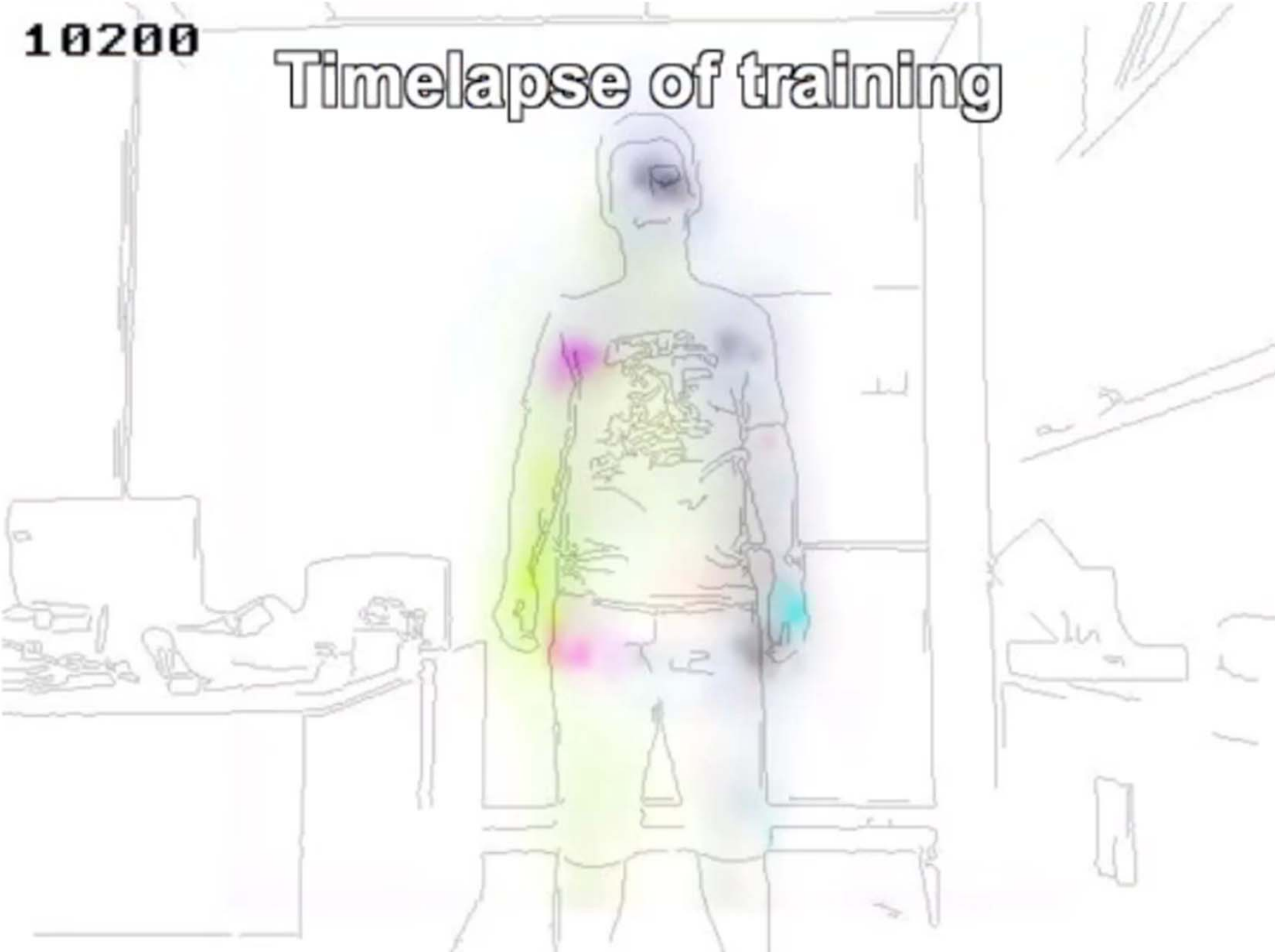
Also, it is easier to visualize (and understand) what is being learnt



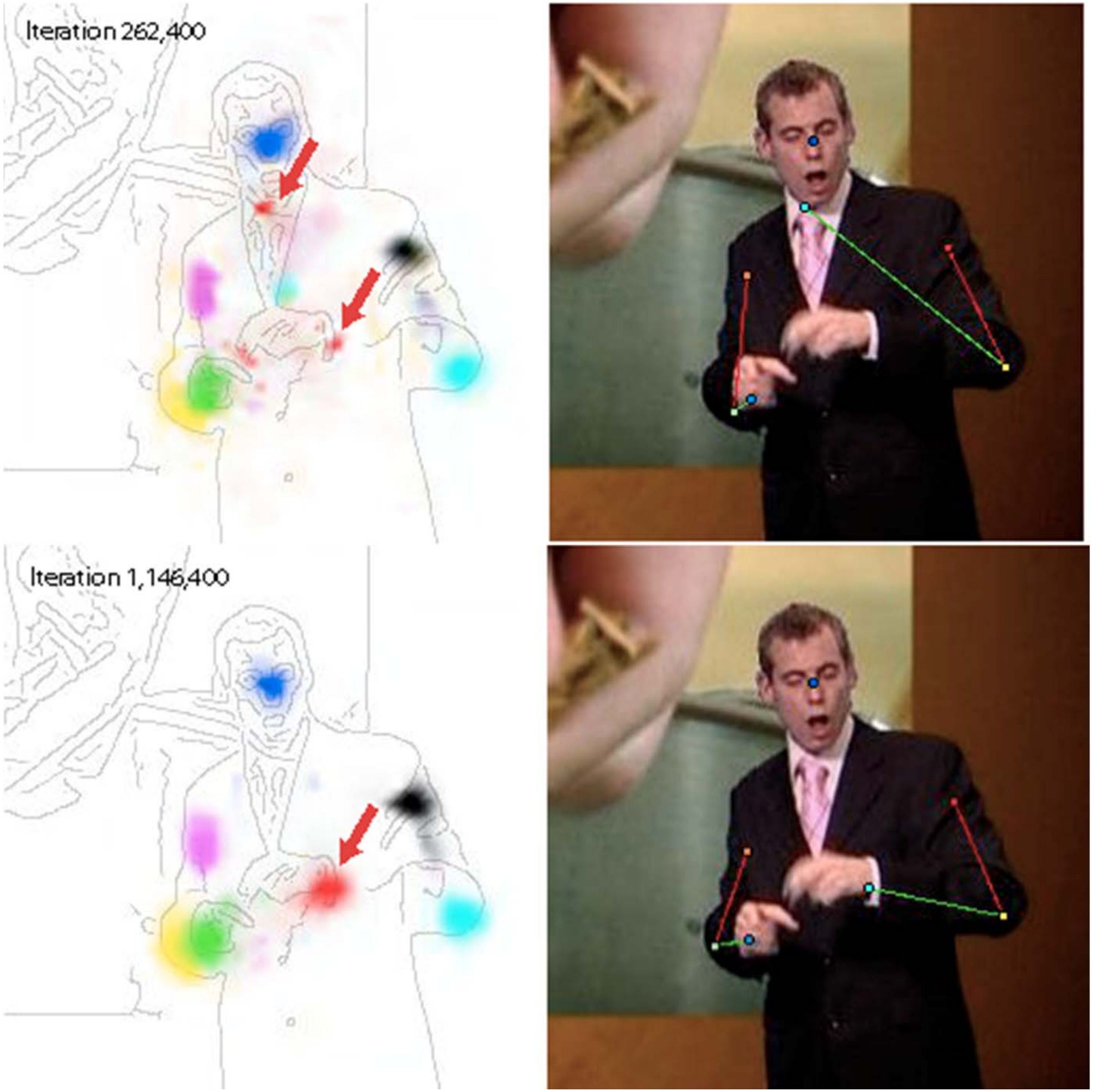
Heatmap

Heatmap Net

Timelapse of training



Multiple modes example

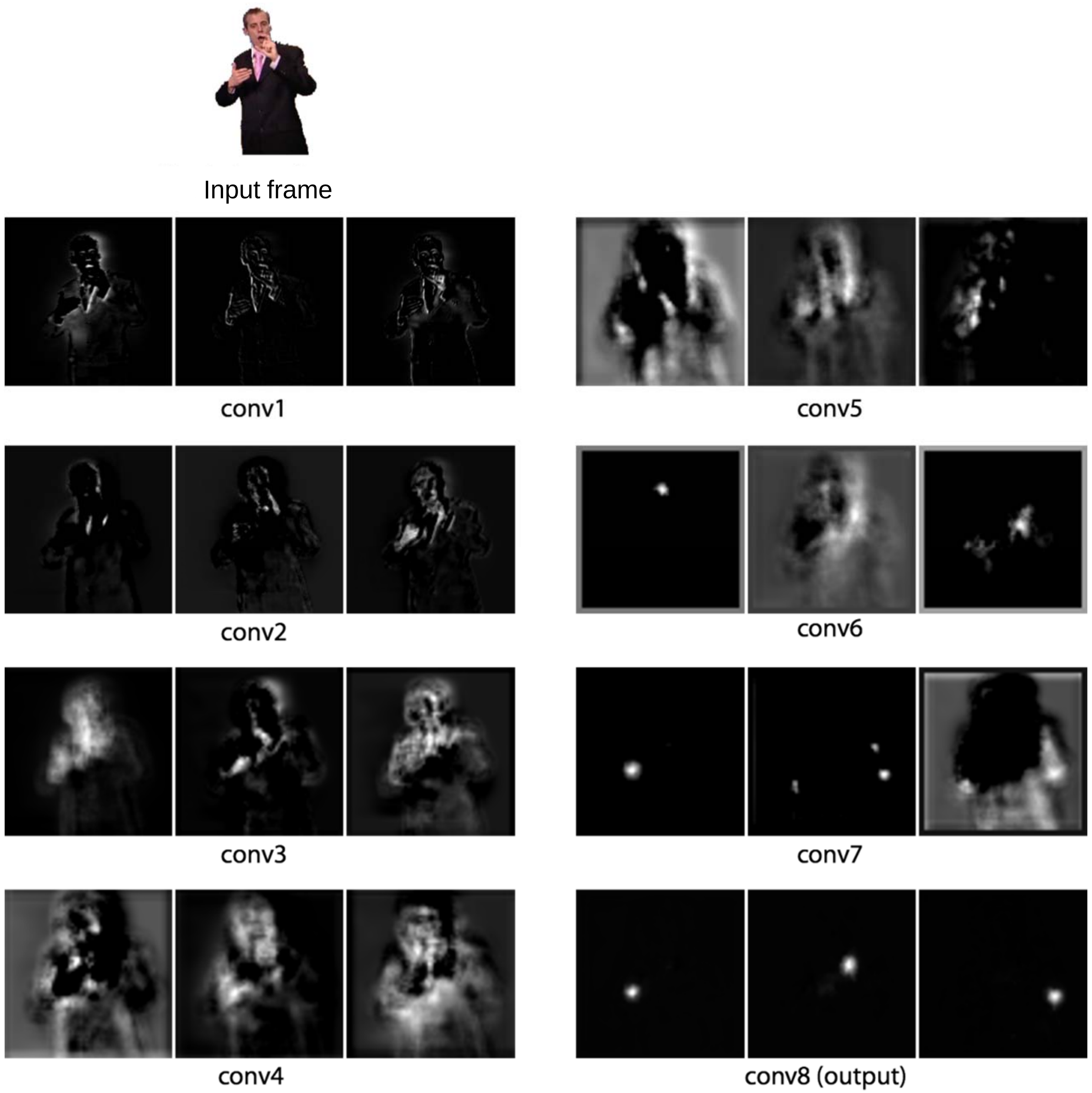


early in training

late in training

What do the layers learn?

Three randomly selected activations from each layer



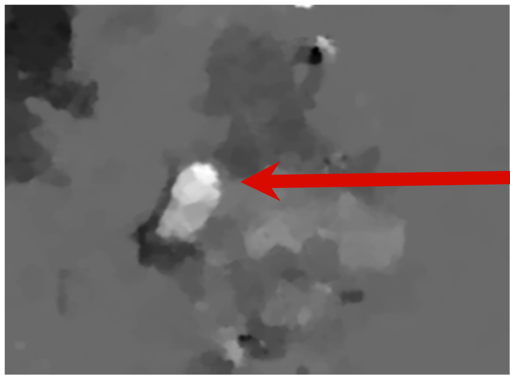
Edges

Body parts
(some)

Learning from videos

- Temporal information

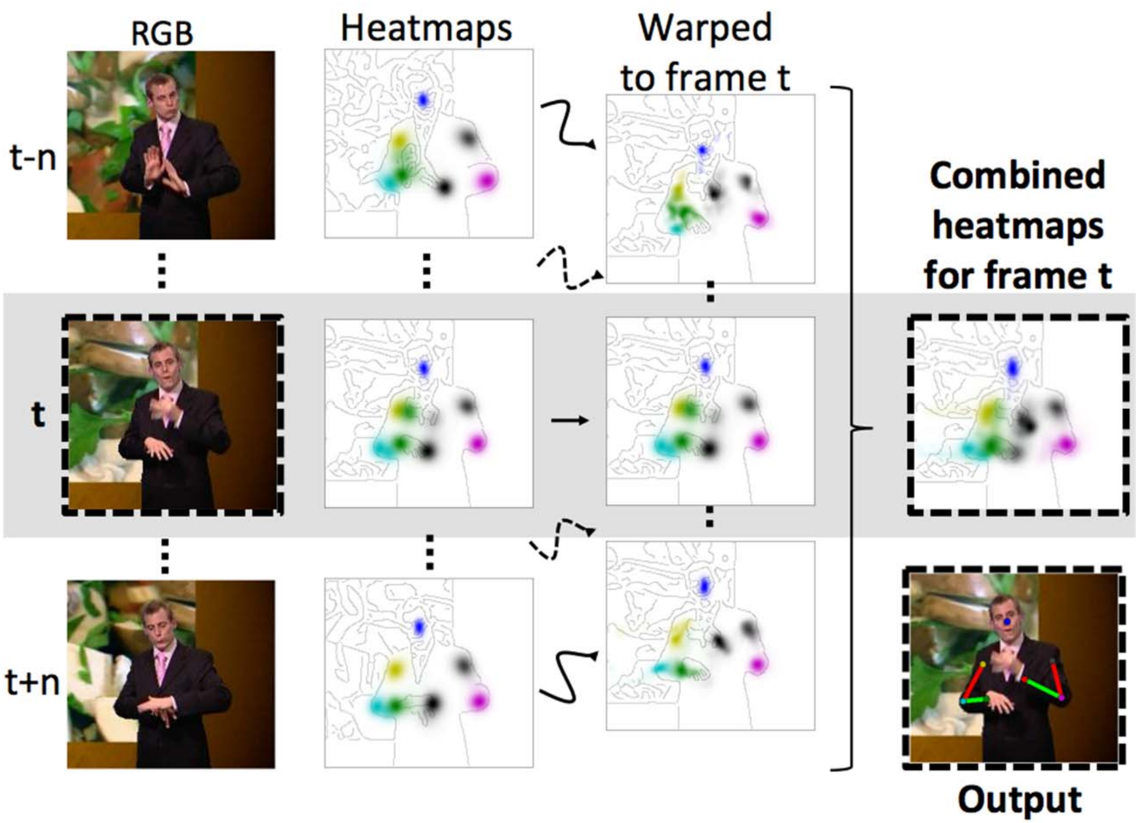
  - How do we learn from temporal information with a ConvNet?



Hand moving
in x direction

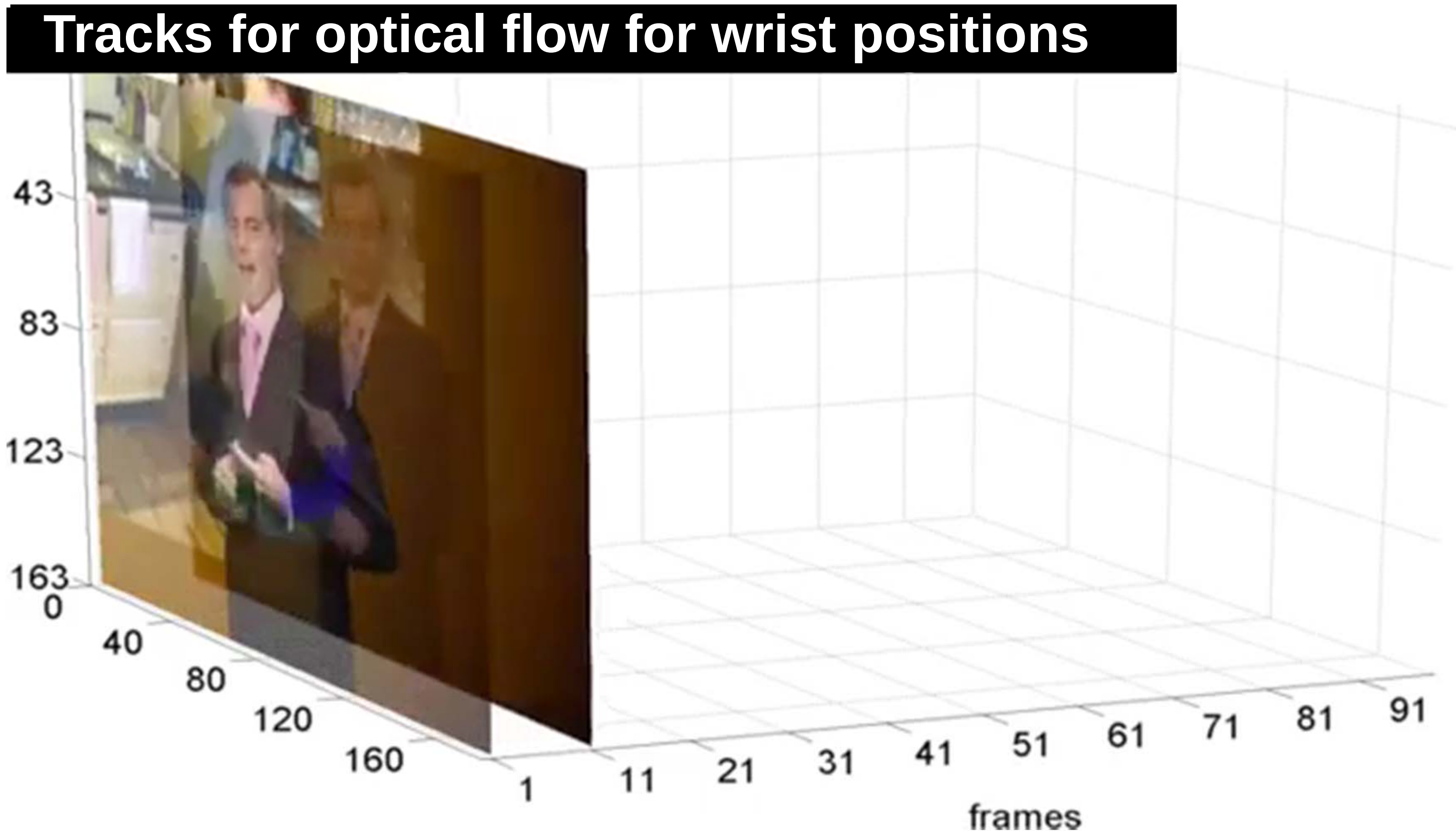
Late fusion using flow

Warp the heatmaps from previous/next frames & **combine**



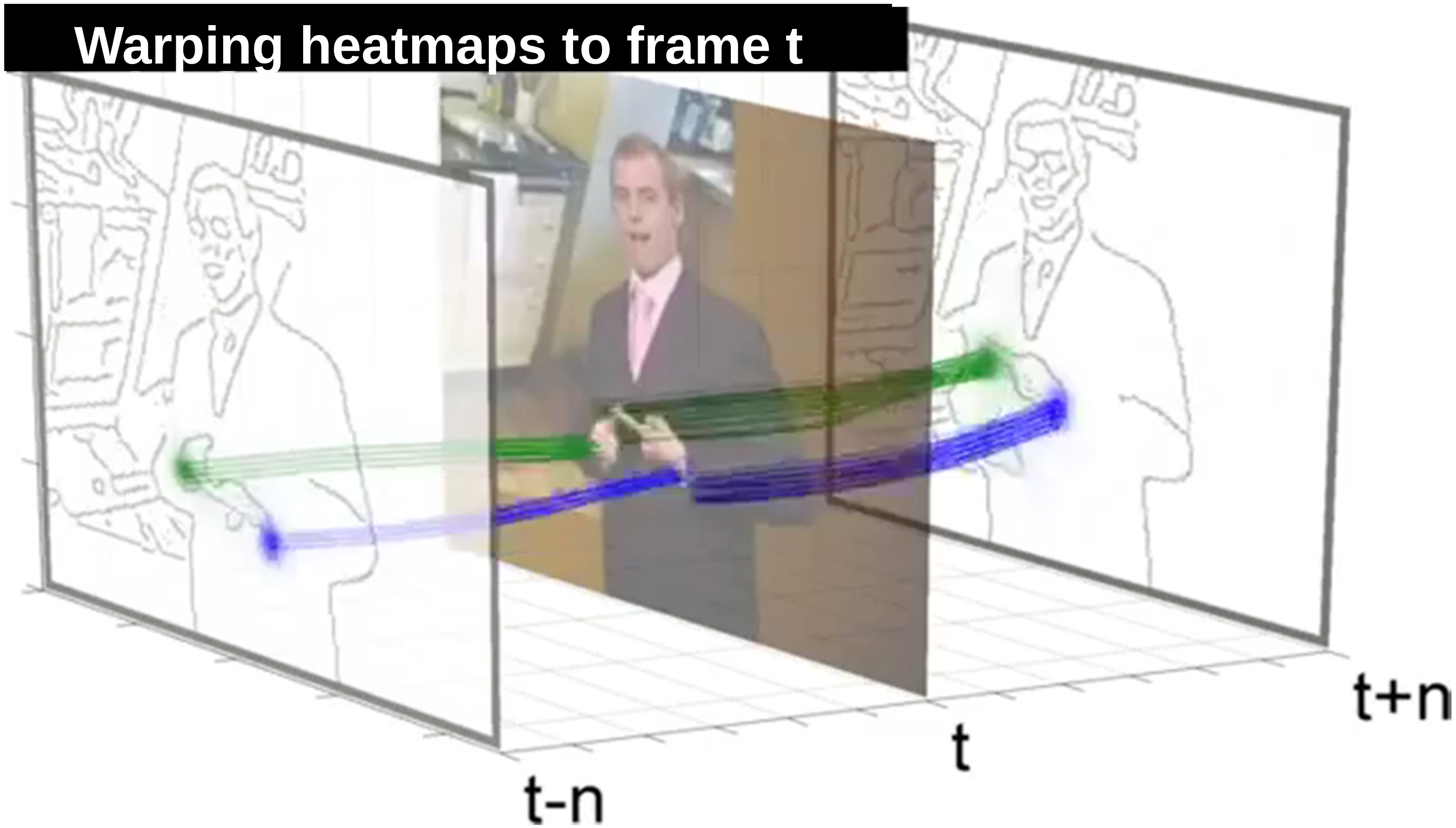
Cf S. Zuffi et al., Estimating human pose with flowing puppets. Proc. ICCV, 2013
Charles et al., Upper Body Pose Estimation with Temporal Sequential Forests, BMVC 2014

Optical flow
Example: Heatmap Net & Optical flow



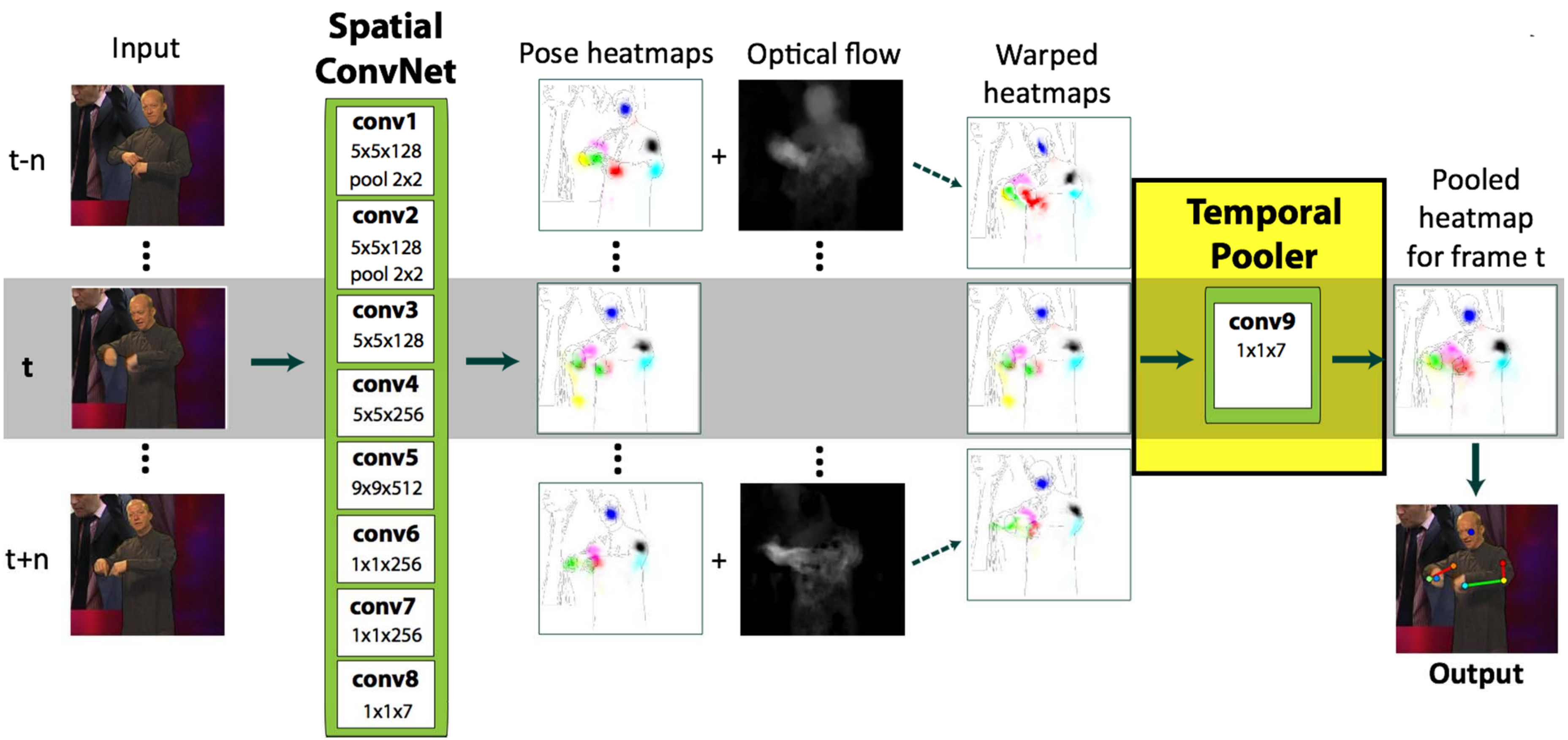
Flow: Brox et al GPU flow from OpenCV, or FastDeepFlow

Optical flow
Example: Heatmap Net & Optical flow

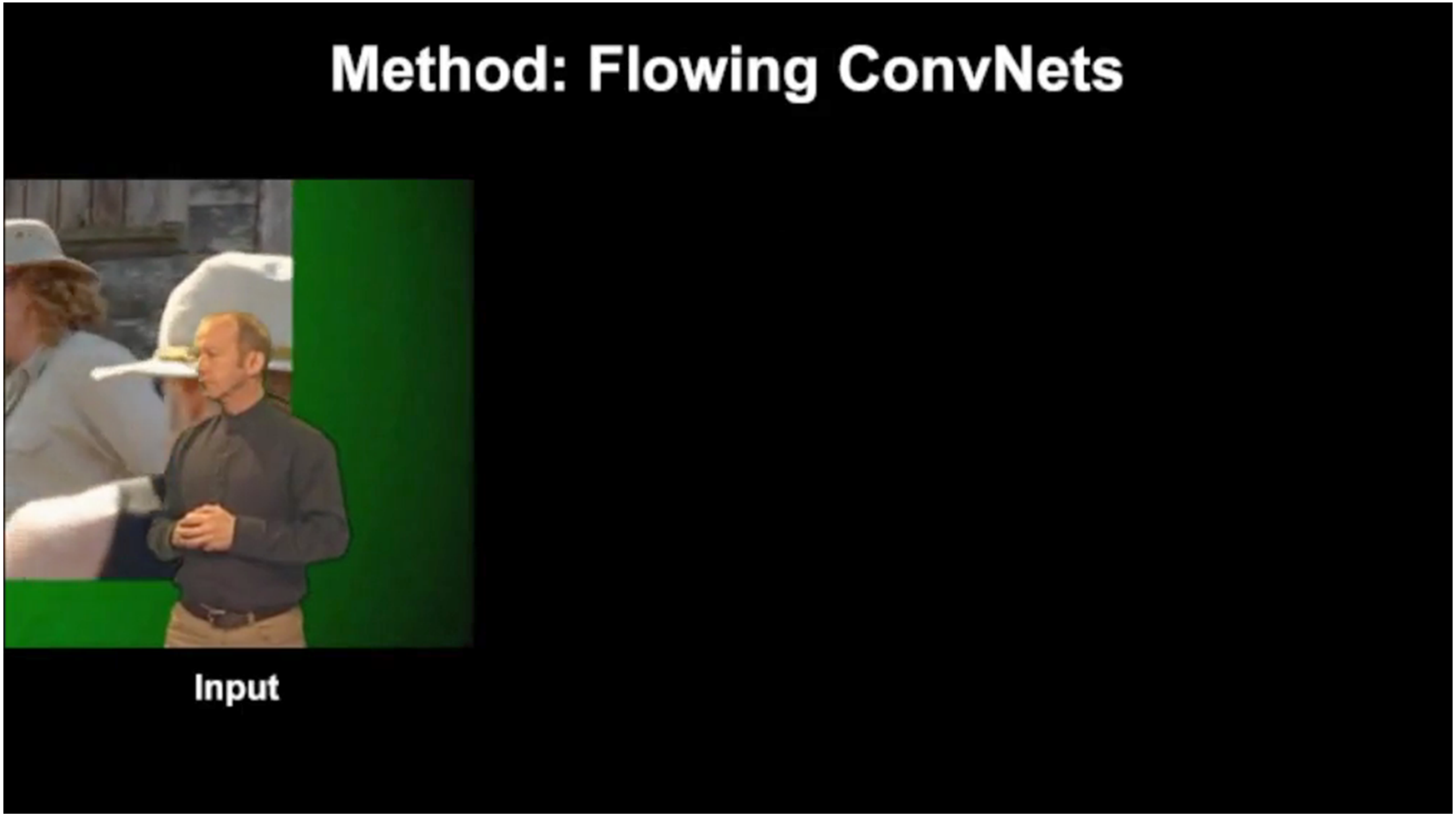


Flowing ConvNets

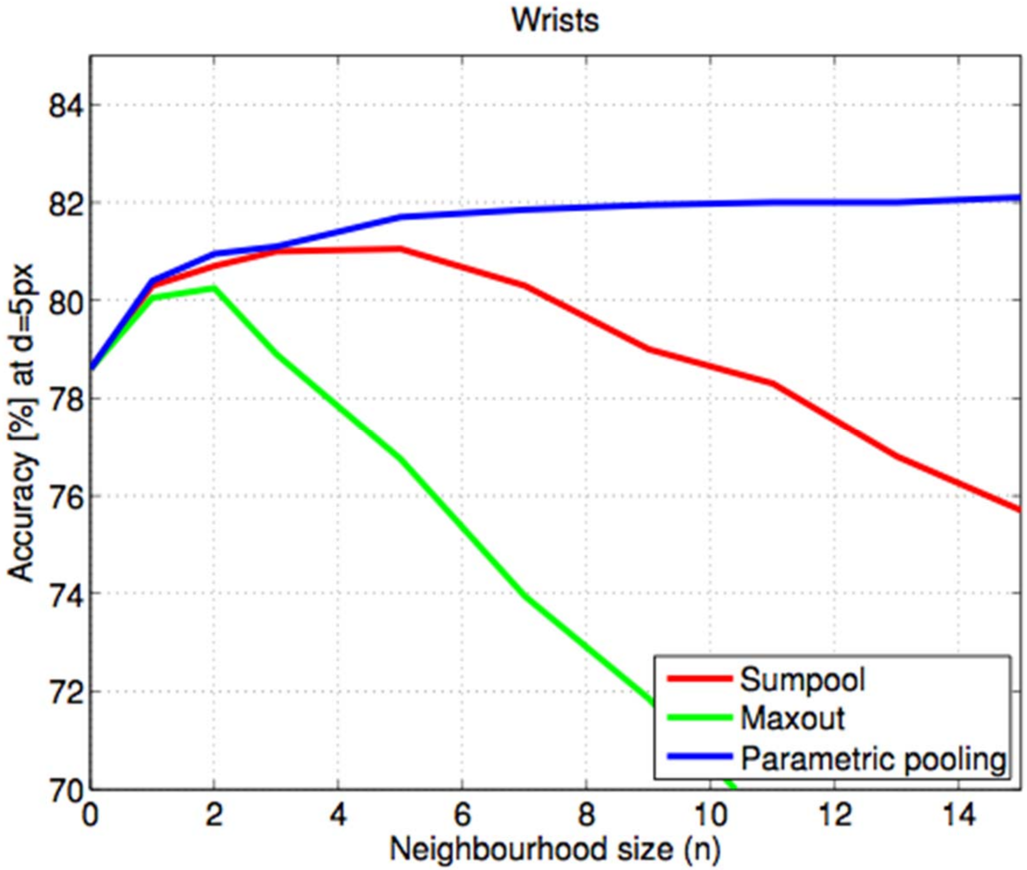
- Learn the **pooling** of the warped heatmaps



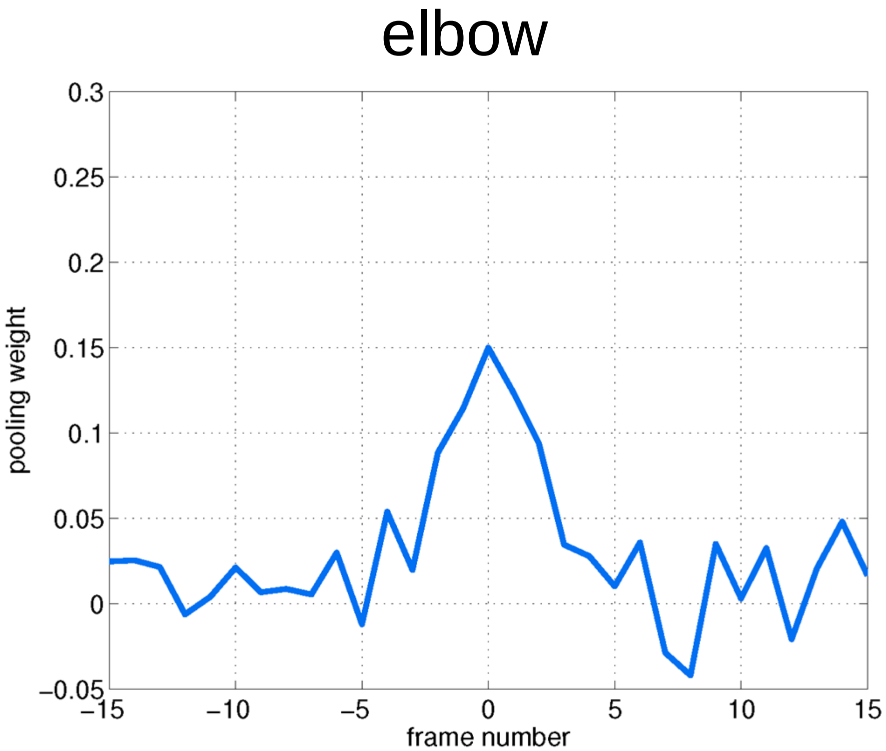
Results: with/without optical flow

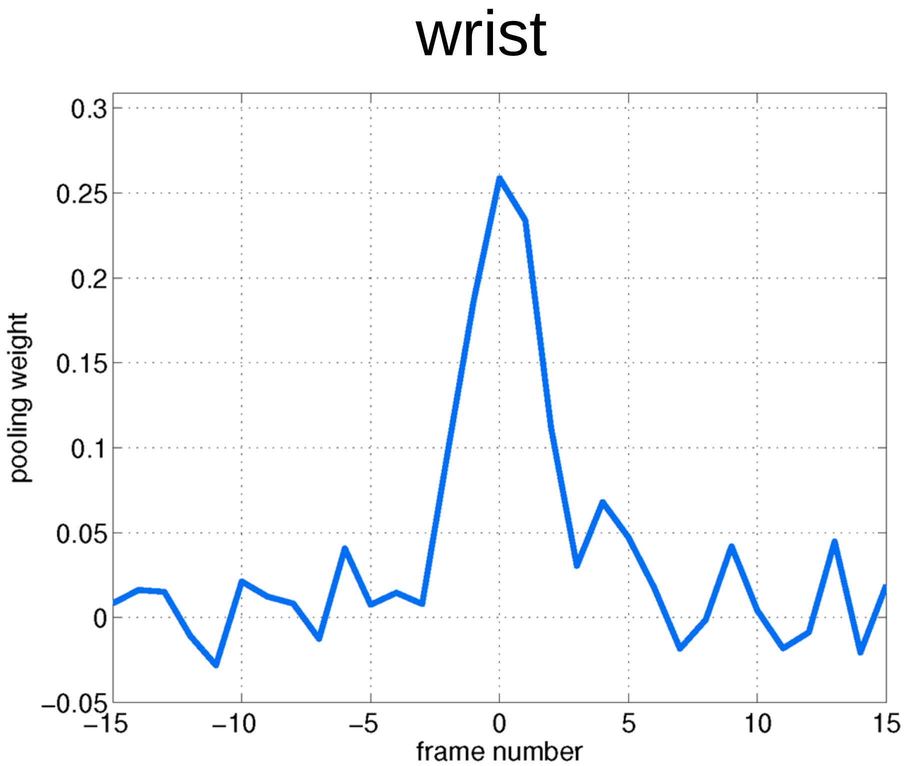


Results
Comparison of pooling types



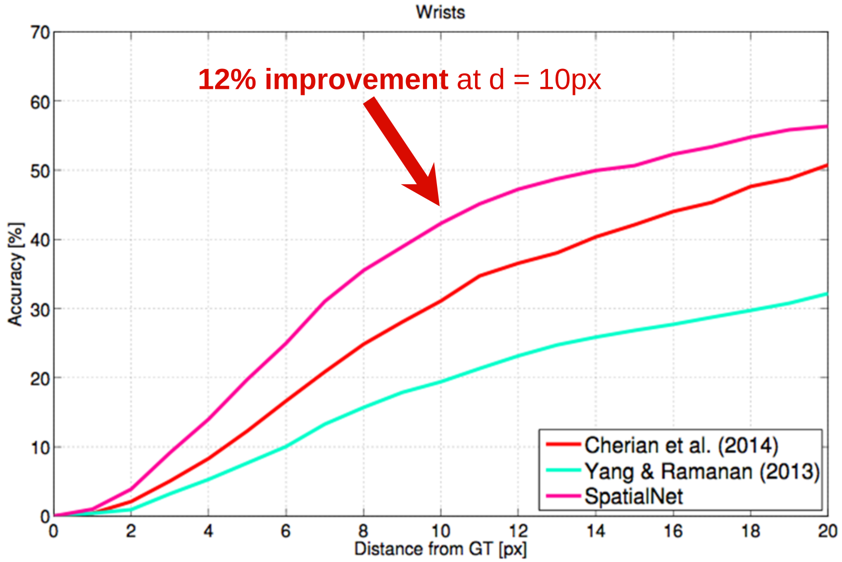
Results
Learnt optical flow pooling weights



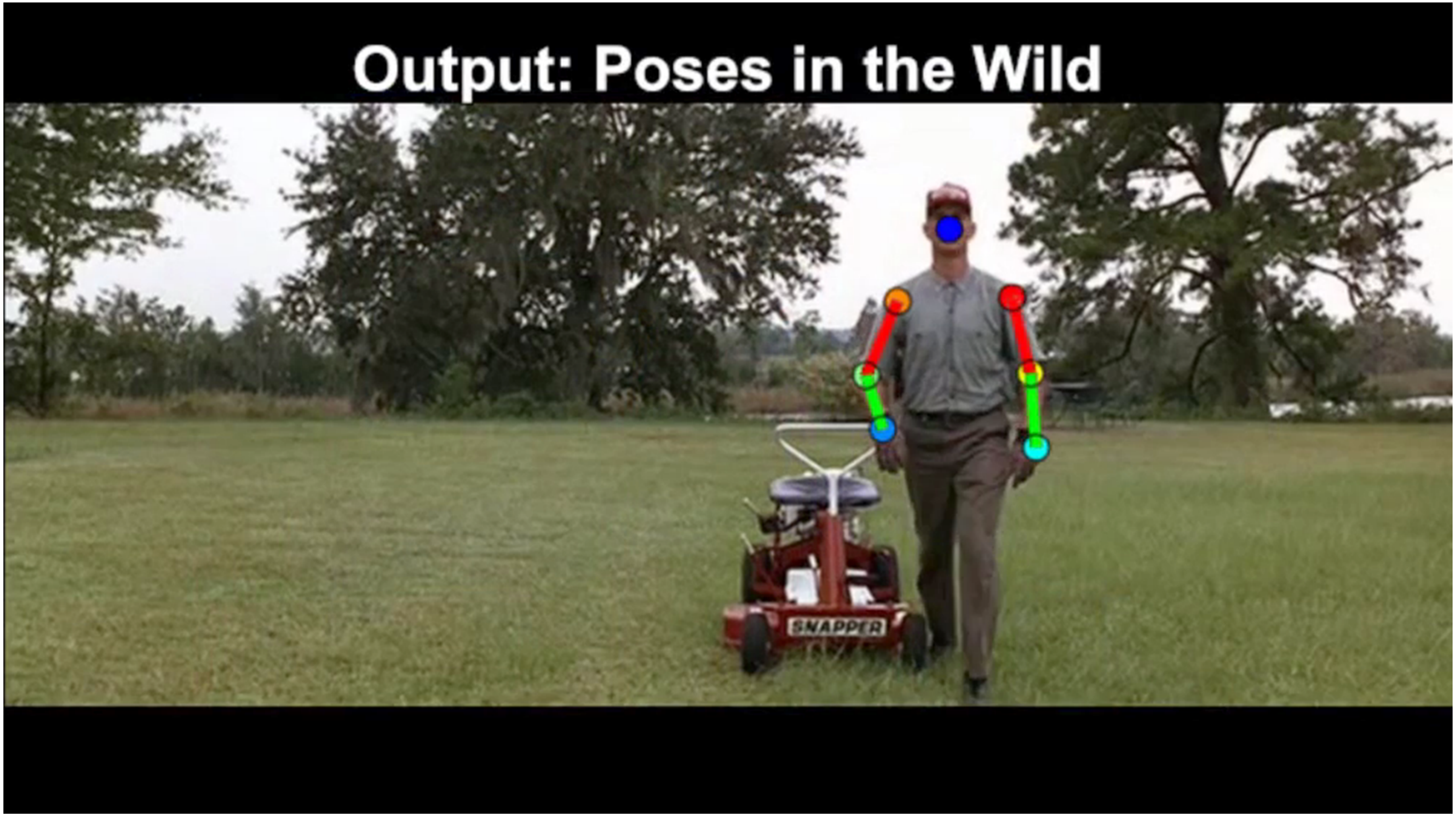


Results
Comparison to the state of the art

Poses in the Wild



Results: Example pose estimation

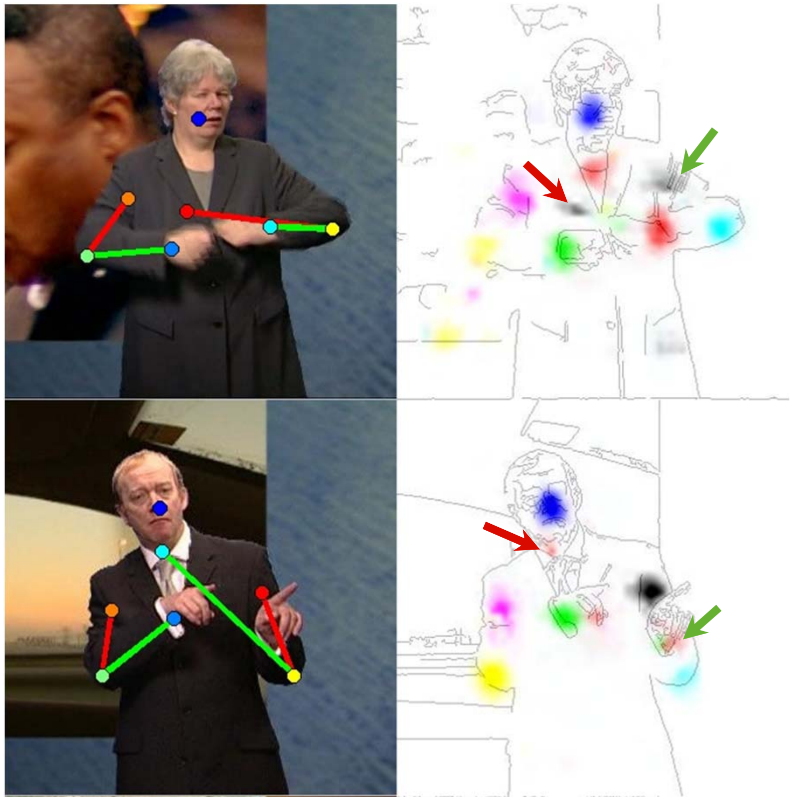


50fps on 1 GPU without optical flow, 5fps with optical flow

Results
Failure cases

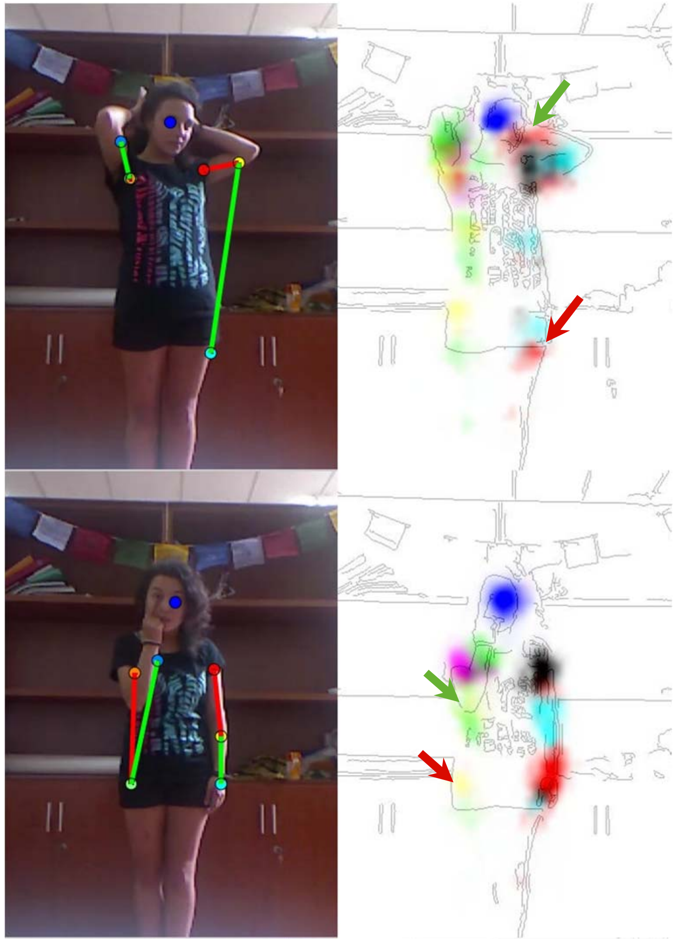
Main failure case: Picking the wrong mode

BBC Pose

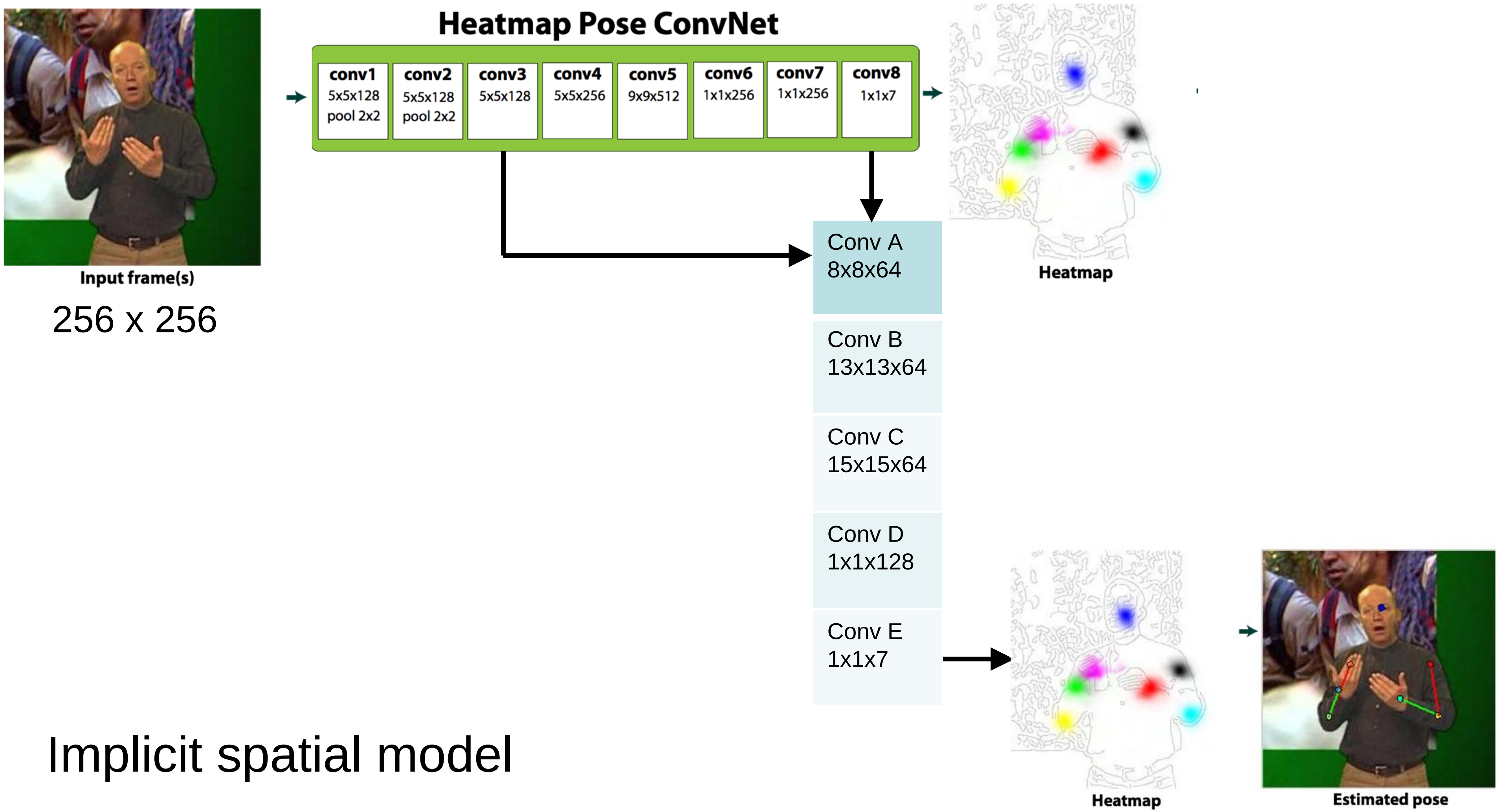


Correctable with a **spatial model**

ChaLearn

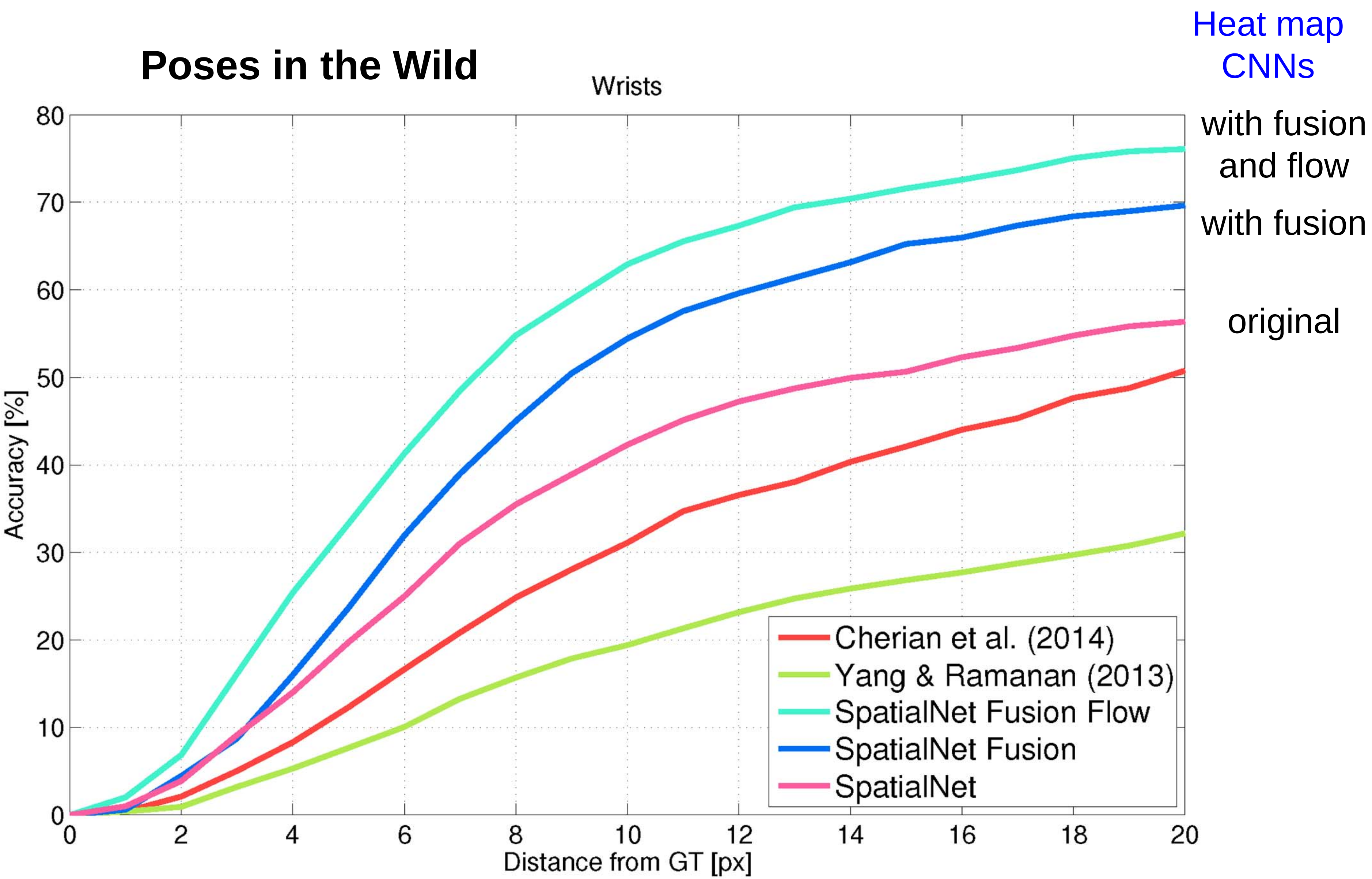


Additional Pooling Fusion Layers



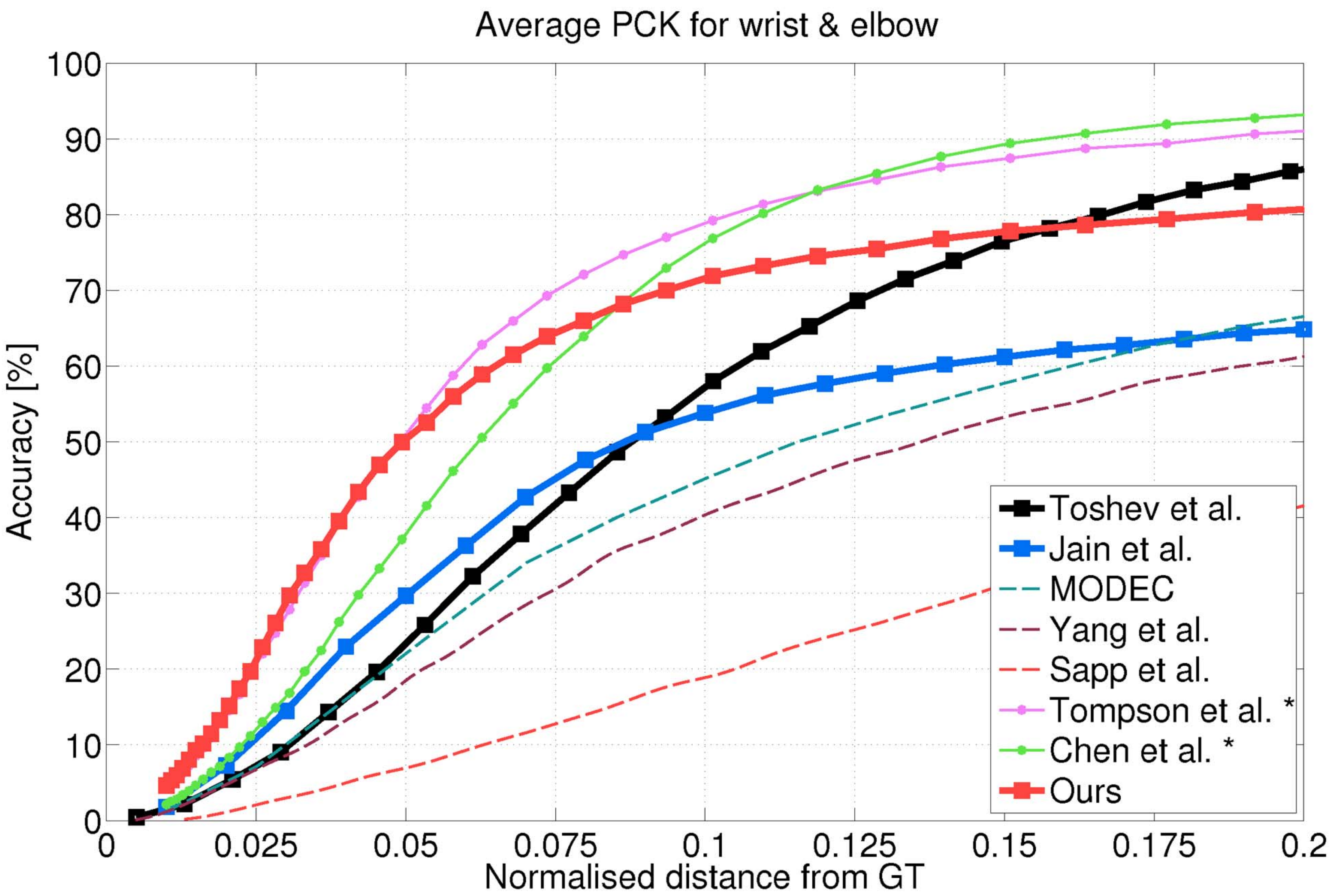
Implicit spatial model

Results: Additional Pooling Fusion Layers



Results: Additional Pooling Fusion Layers

FLIC: single image predictions



Summary

- Deep Heatmap ConvNet achieves state of the art with implicit spatial models
- Performance improved by optical flow pooling
- Futures:
  - Robust regression
  - Data dependent flow channel pooling
  - More training data