Hands Deep in Deep Learning for Hand Pose Estimation

Markus Oberweger, Paul Wohlhart, Vincent Lepetit
Institute for Computer Graphics and Vision
Graz University of Technology, Austria

10.2.2015

Motivation

- Important for Human Computer Interaction, Augmented Reality
- Leap Motion, Nimble VR, Oculus Rift, Microsoft Hololens, …
- Estimate hand pose from single depth image

Motivation

- Difficult: self-occlusions, self-similarity, many DoF
- Convolutional Neural Network regression
  → high speed, high accuracy

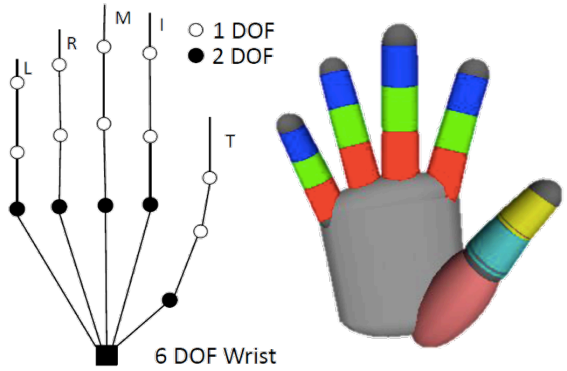
Motivation

Motivation

Tang et al.

Ours

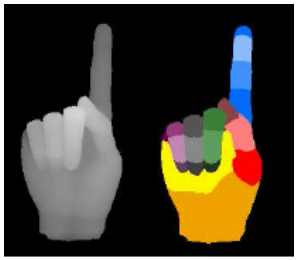
Related Work

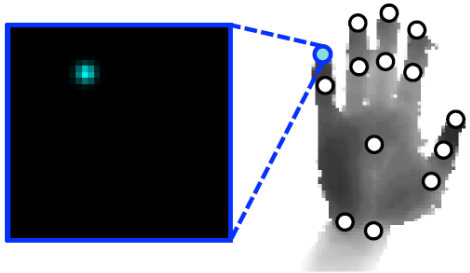


- Generative models [Oikonomidis CVPR'11], [Qian CVPR'14], [Ballan ECCV'12], [de La Gorce TPAMI'11], [Xu ICCV'13], [Melax GIC'13]

  → Problem: initialization, local minima

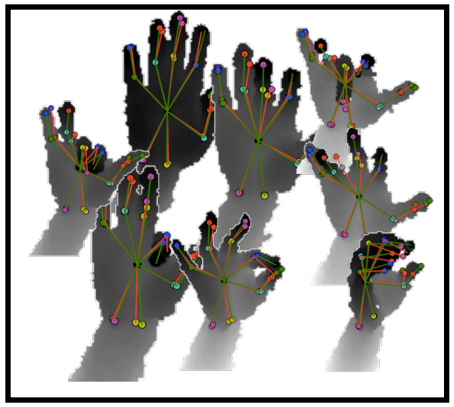
- Discriminative models [Keskin ICCV'11], [Keskin ECCV'12], [Kuznetsova ICCV'13], [Tang ICCV'13], [Tang CVPR'14], [Tompson ToG'14]



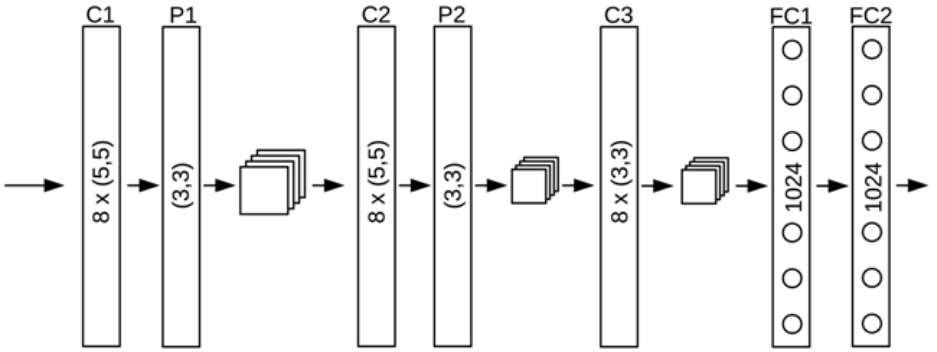


Method

Training Set

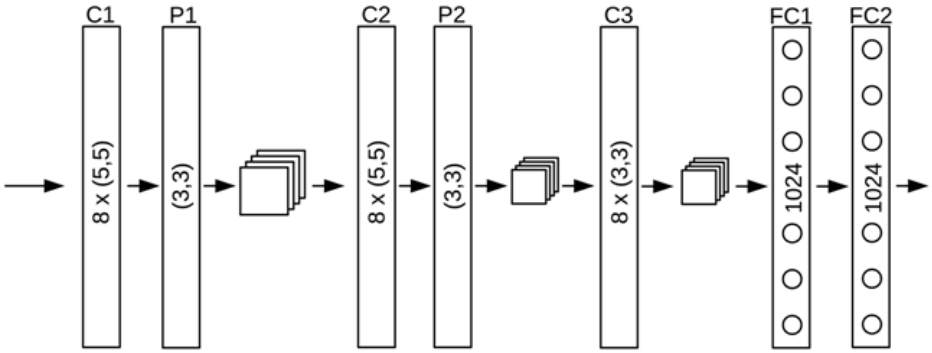


CNN

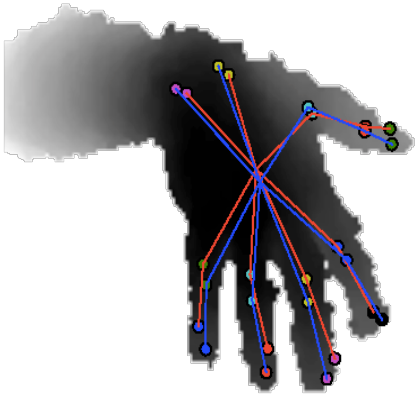


Input depth image

Trained CNN



Predicted pose

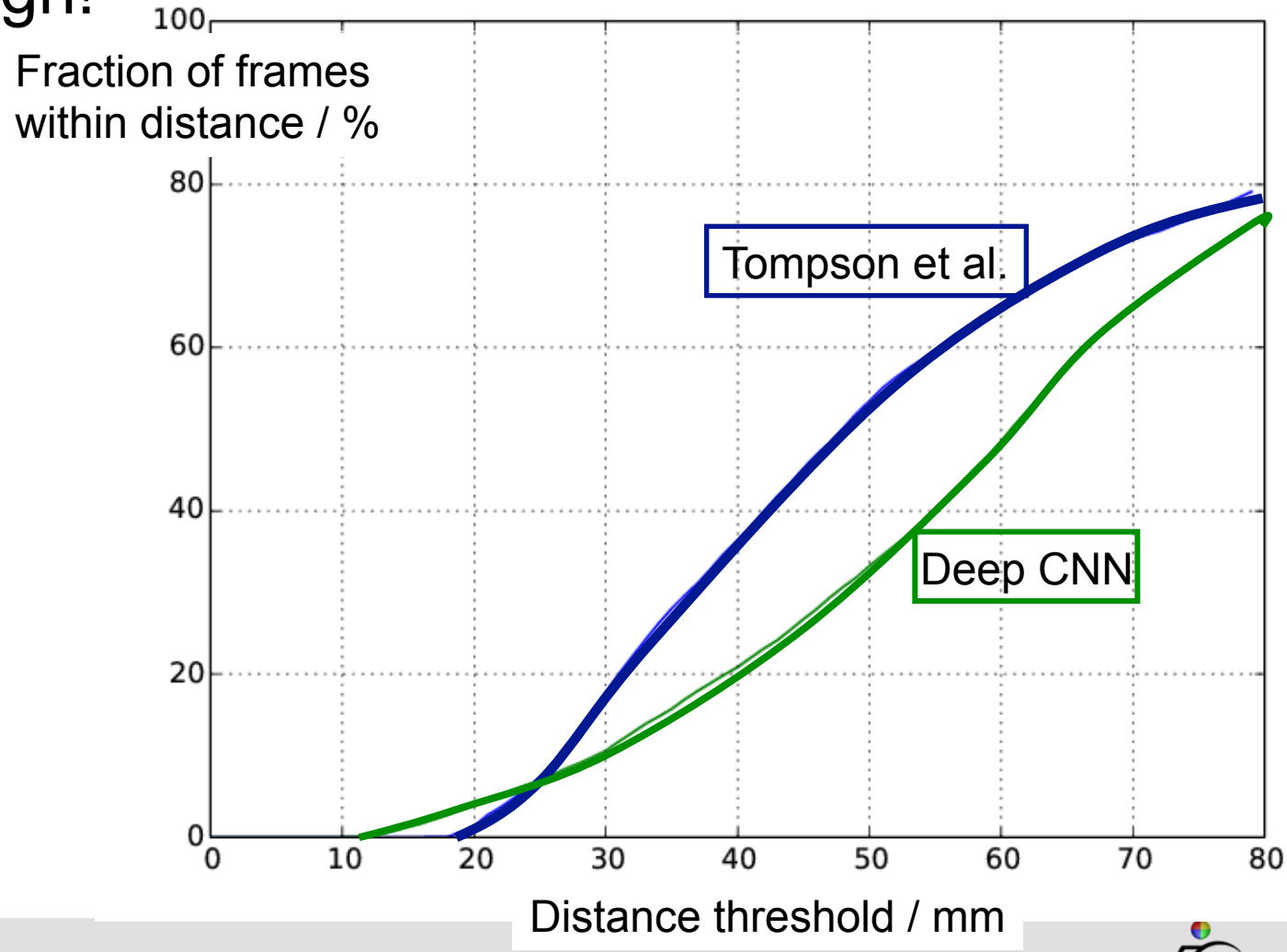


CNN

Straight forward: Deep CNN

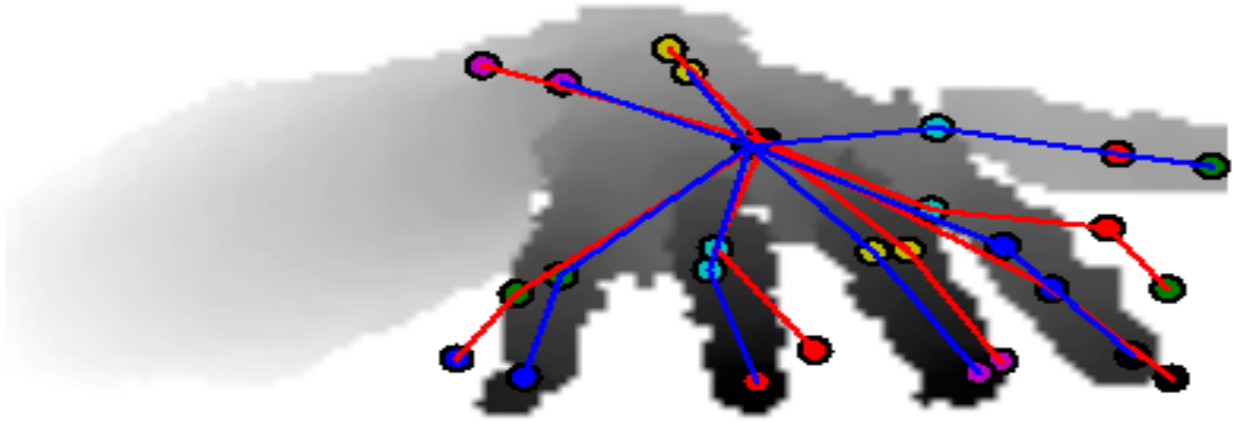
CNN

Not good enough!



Pose Prior

Deep CNN: works, somehow…



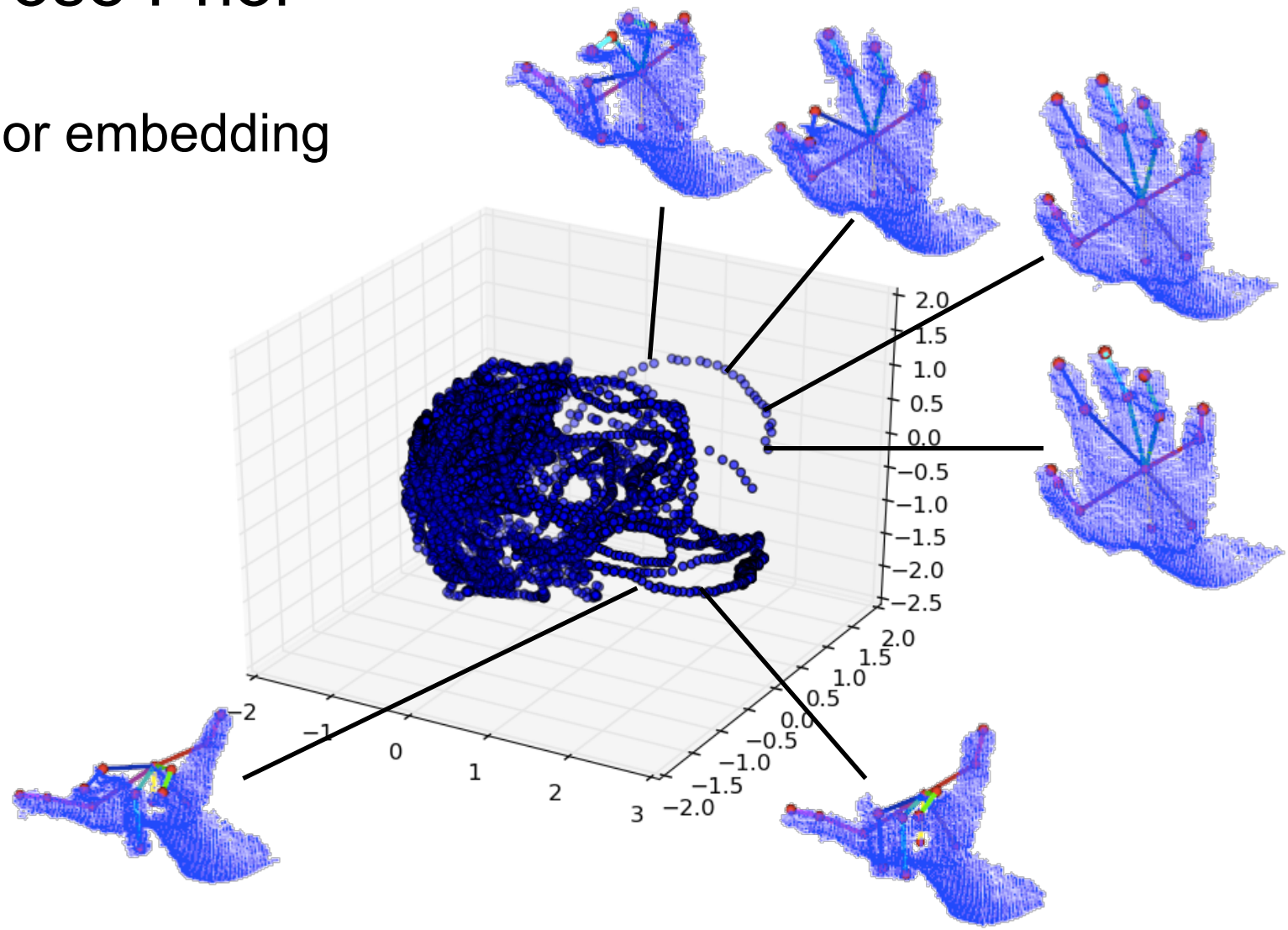
— Ground truth, — Predicted

Pose Prior

- Physical constraints → Low dimensional embedding sufficient [Wu ICCV'01]

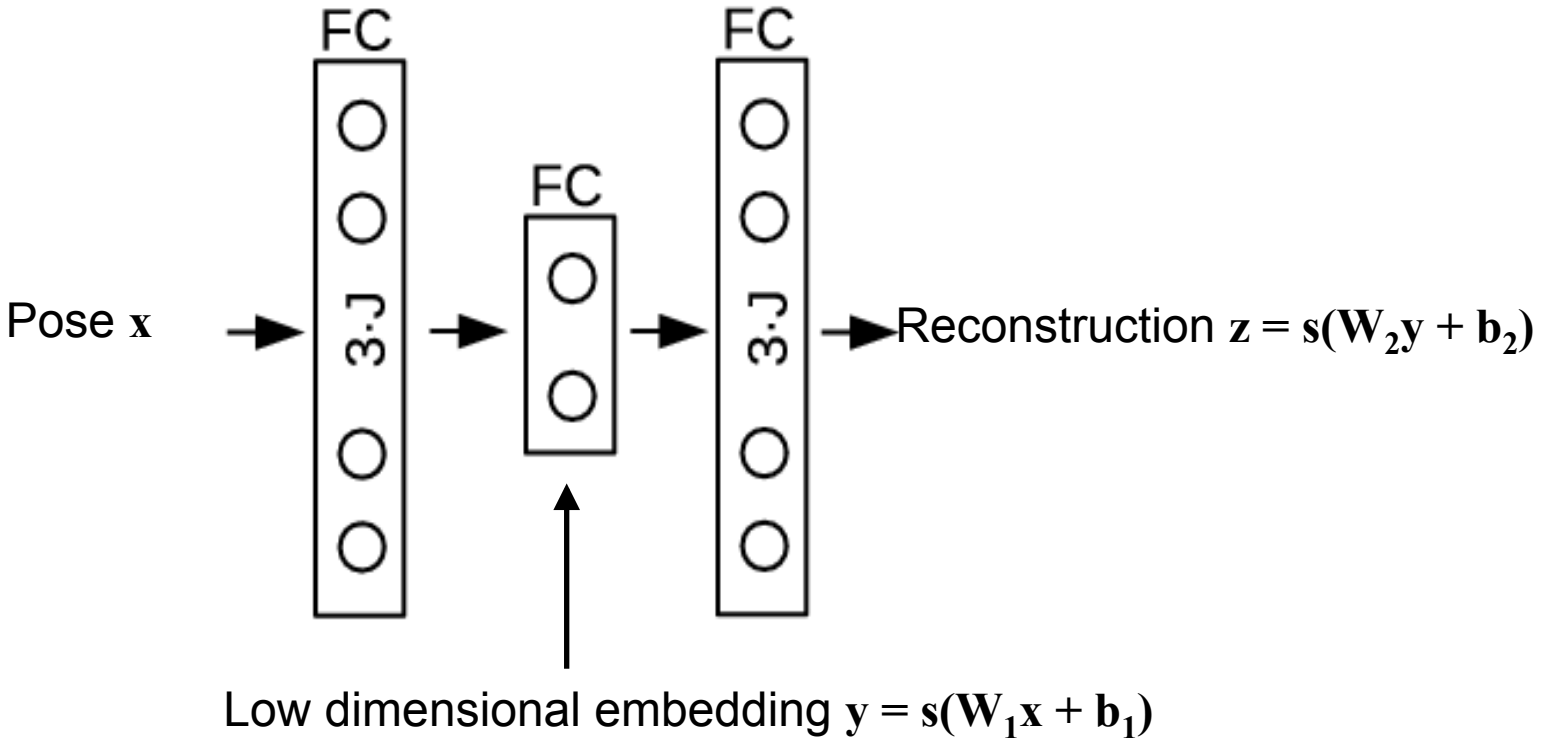
Pose Prior

3D prior embedding

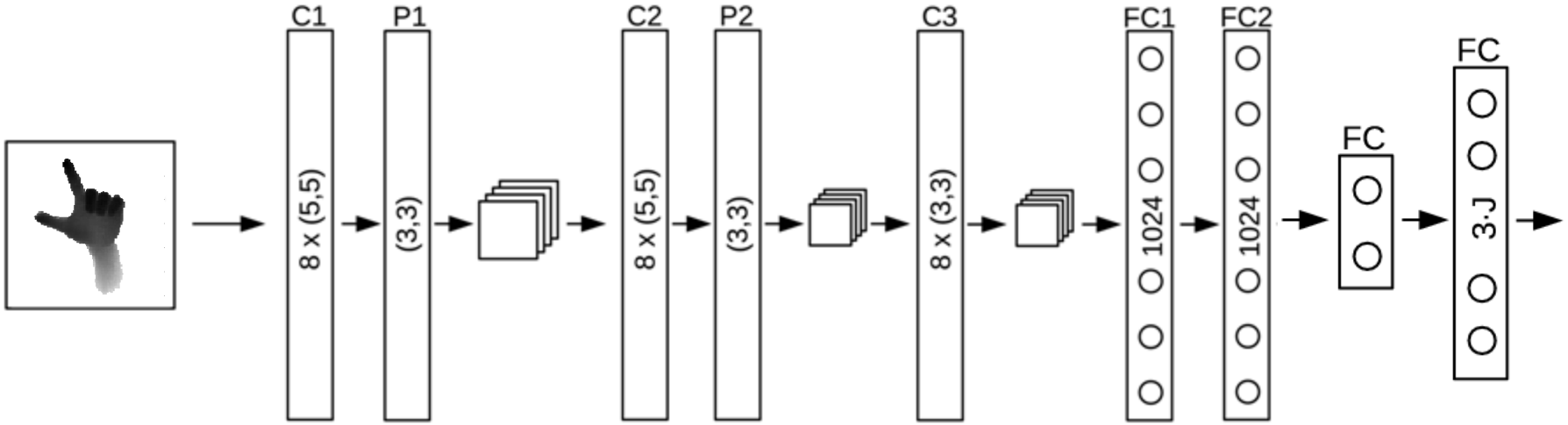


Pose Prior

- Learning a prior: Auto-Encoder network



Pose Prior

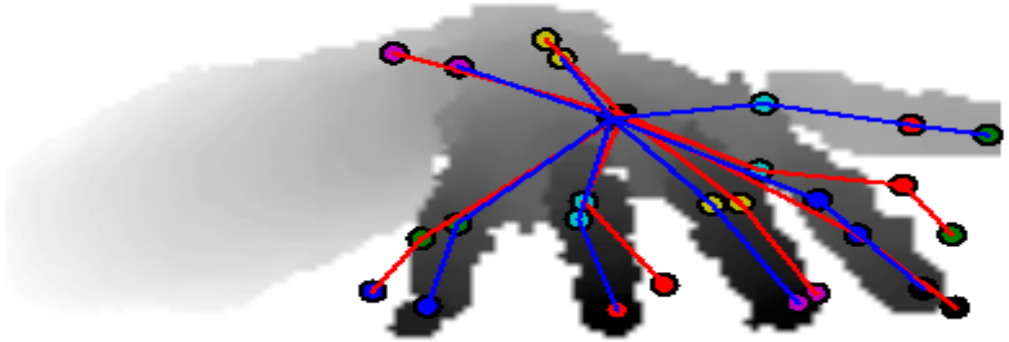


Pose Prior

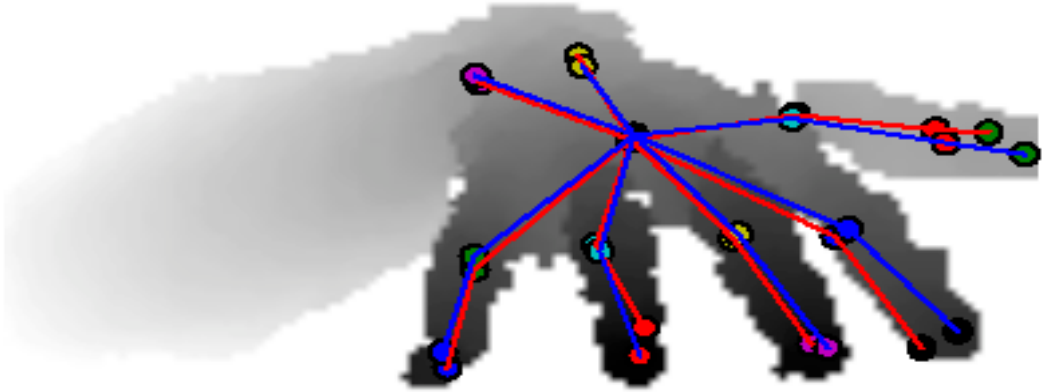
- Seamless CNN integration → speed
- Bottleneck: learn low dimensional embedding
- e.g. 48D pose space → 8D embedding
- Initialize reconstruction with PCA

Pose Prior

Deep CNN

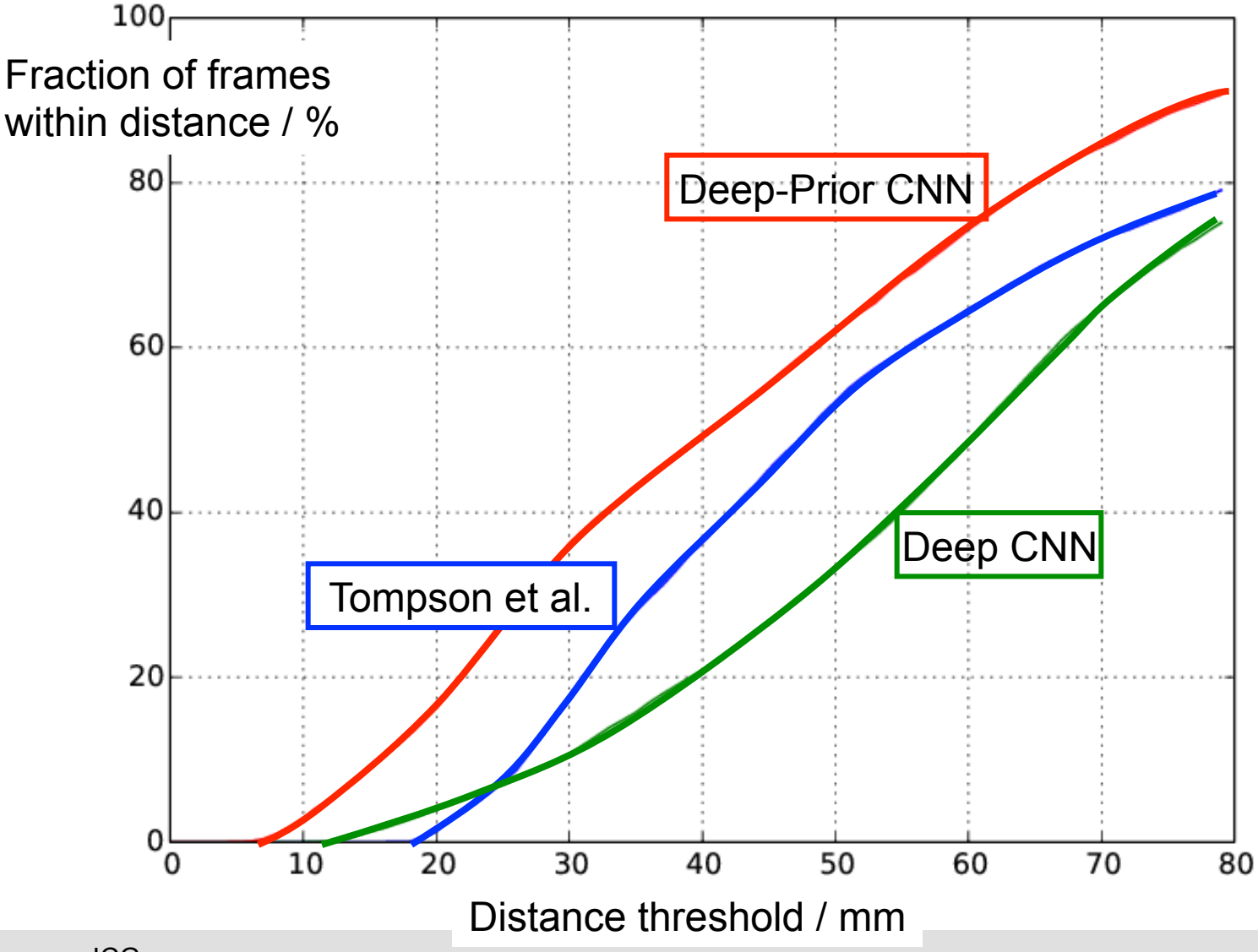


Deep-Prior CNN



— Ground truth, — Predicted

Pose Prior

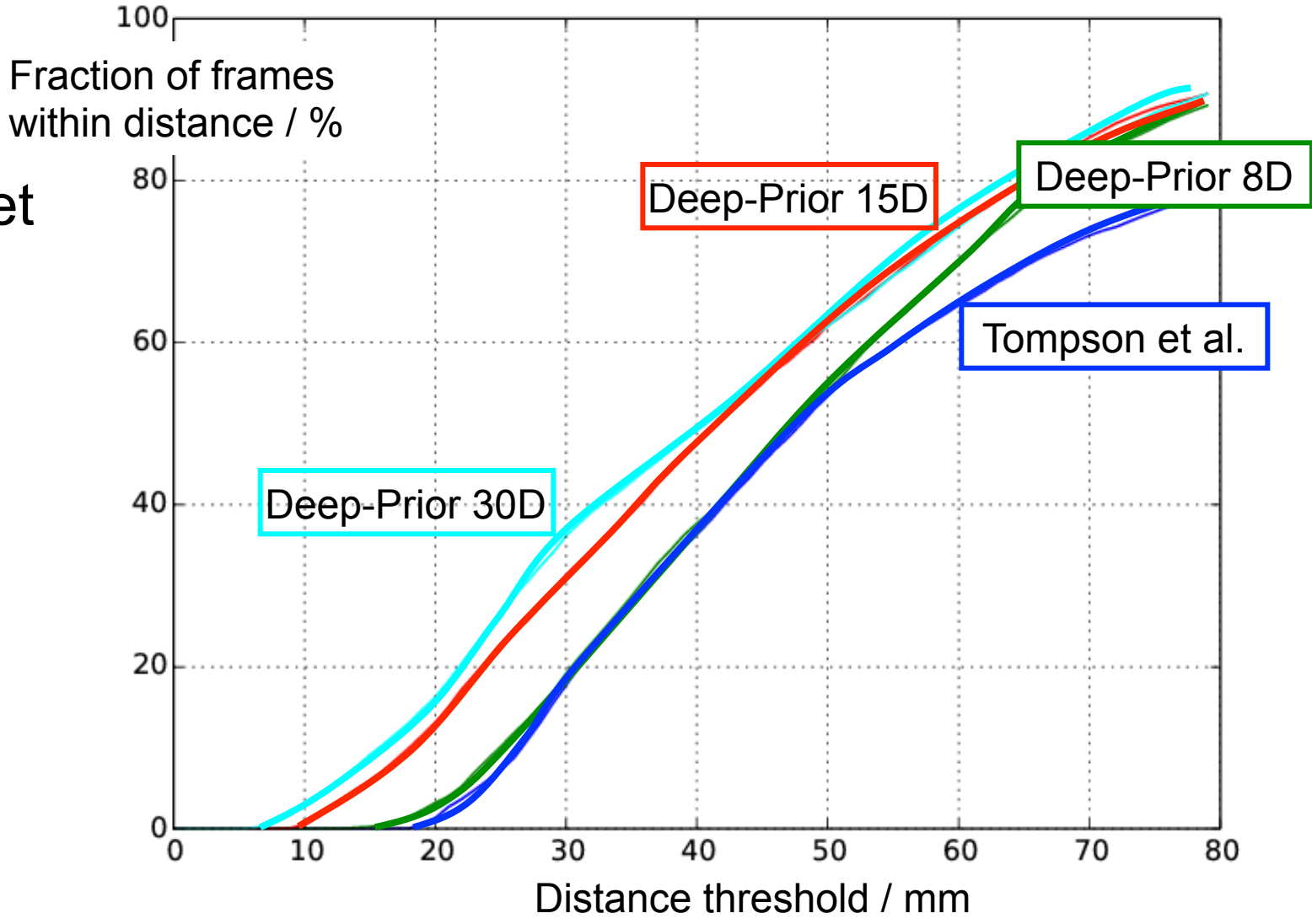

Fraction of frames
within distance / %
100
80
60
40
20
0
0
10
20
30
40
50
60
70
80
Distance threshold / mm
Deep-Prior CNN
Tompson et al.
Deep CNN

Evaluation

- Datasets
  - NYU Hand Pose Dataset [Tompson ToG'14]
    - Structured Light, 72k training samples
  - ICVL Hand Posture Dataset [Tang CVPR'13]
    - Time of Flight, 22k training samples
    - Inaccurate annotation

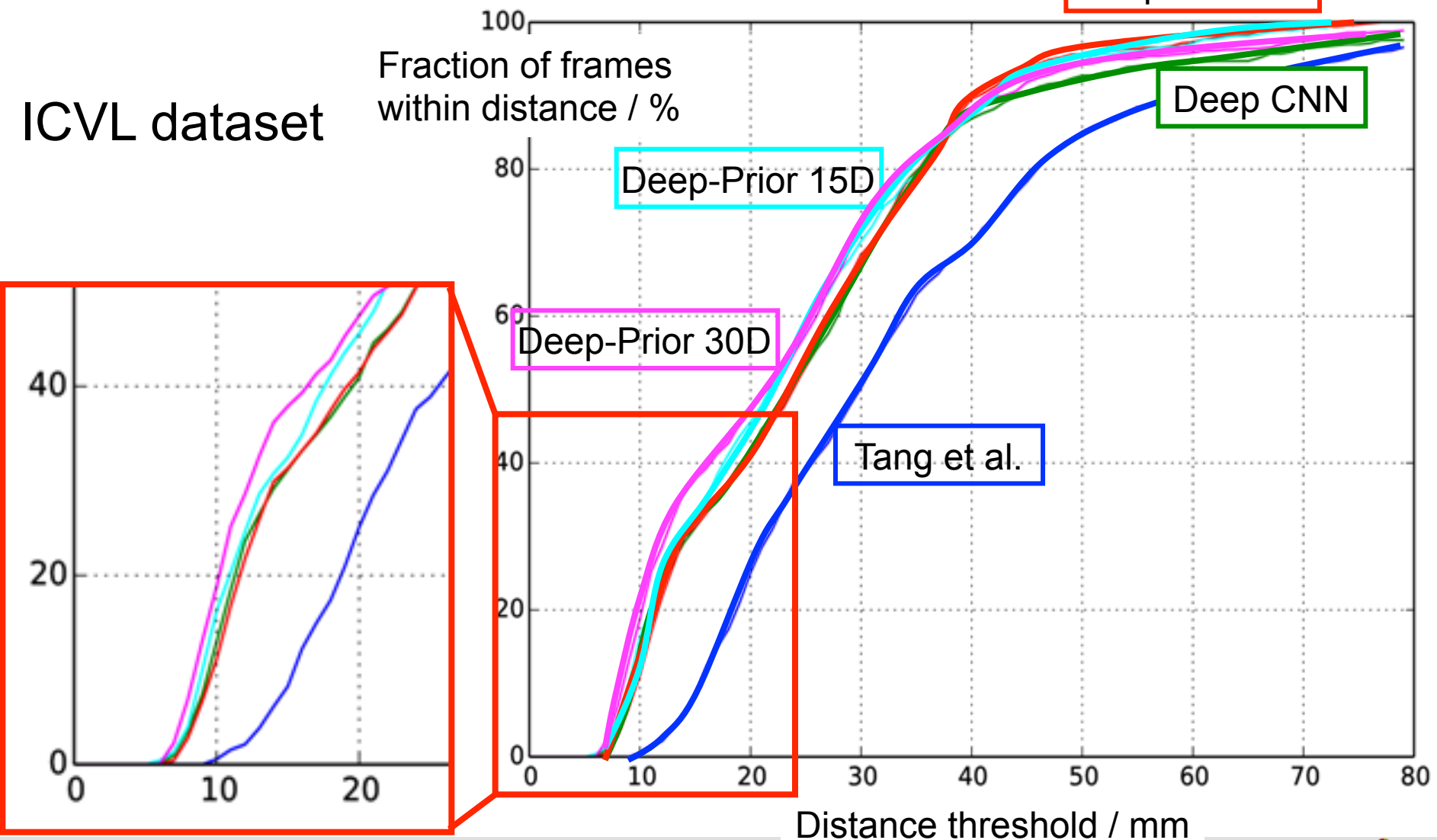
Importance of the Pose Prior

NYU dataset



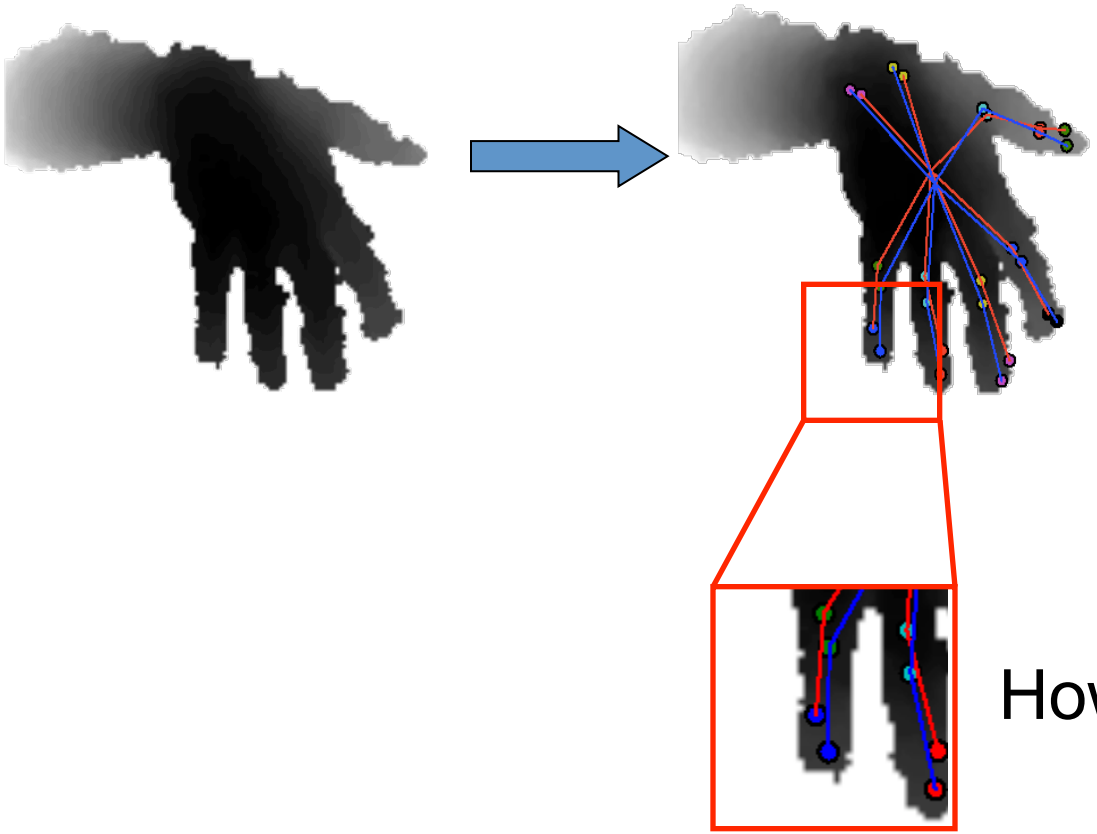
Importance of the Pose Prior

ICVL dataset



Refinement

So far: global pose estimate



How about local context?

Refinement

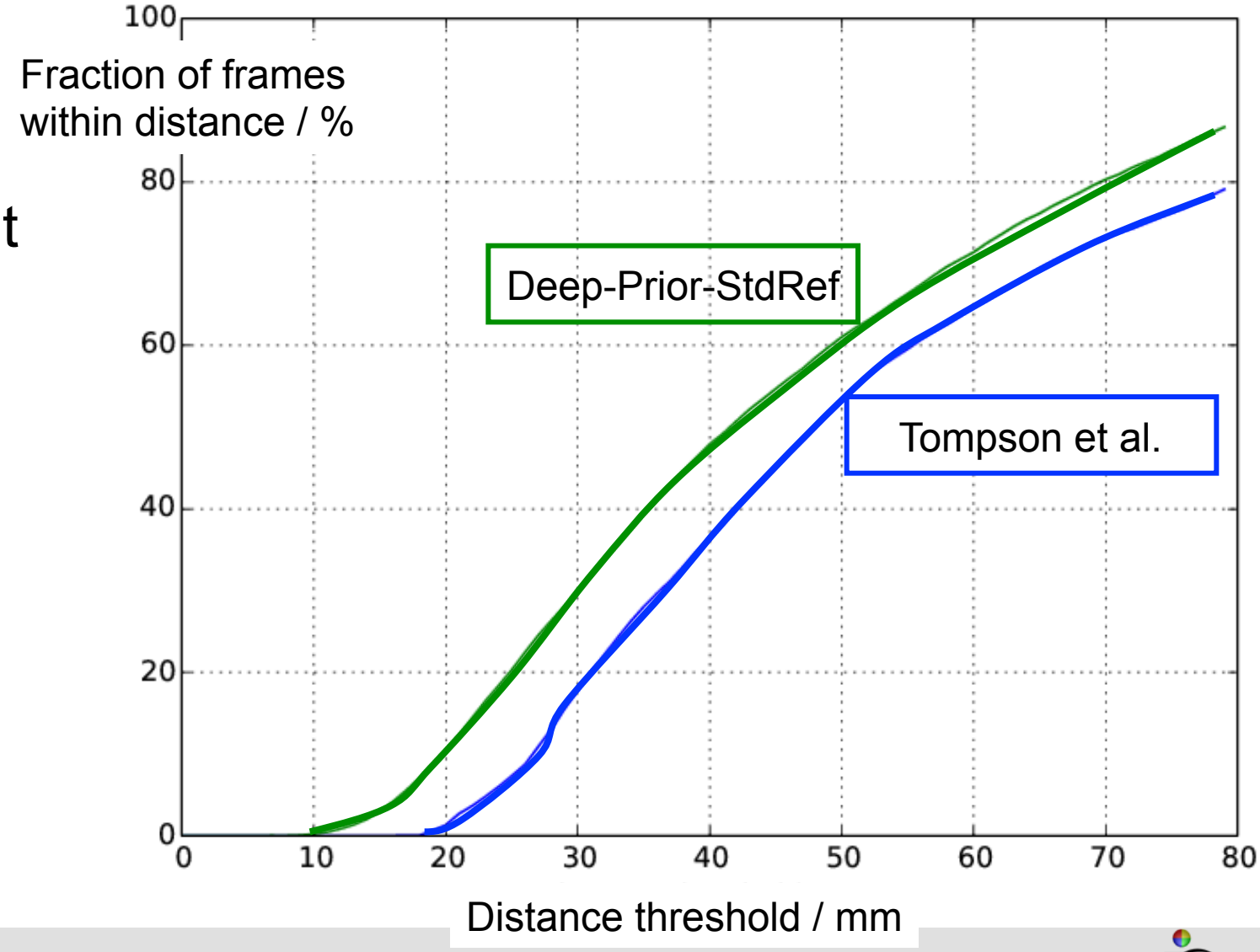
- Start from the estimates given by the previous stage
- Independent refinement for each joint
- Each regressor predicts update
- Iterate

Refinement

Straight forward: Deep CNN

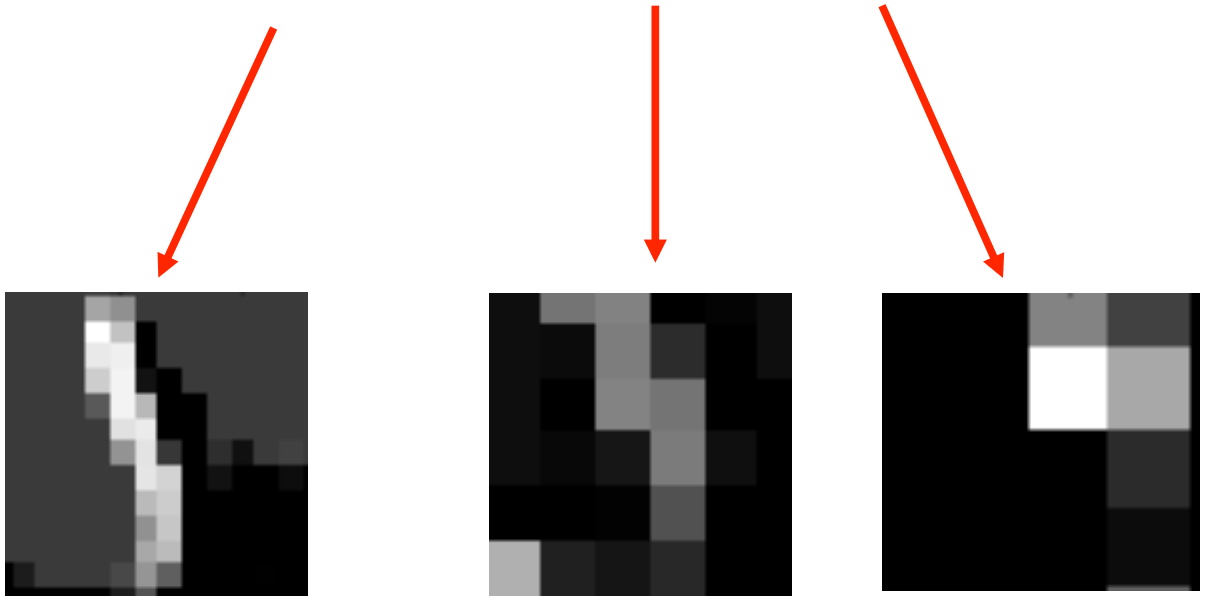
Increasing Accuracy with Pose Refinement

NYU dataset



Refinement

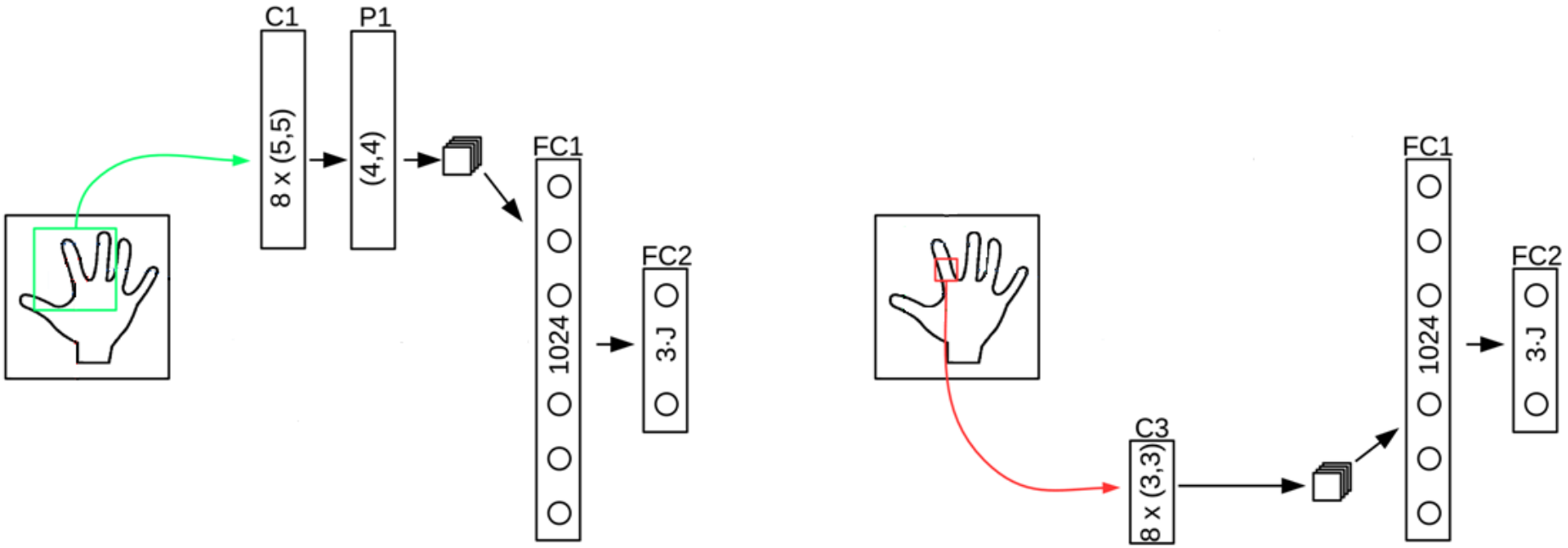
Deep CNN



Our interpretation: pooling increases uncertainty

Refinement

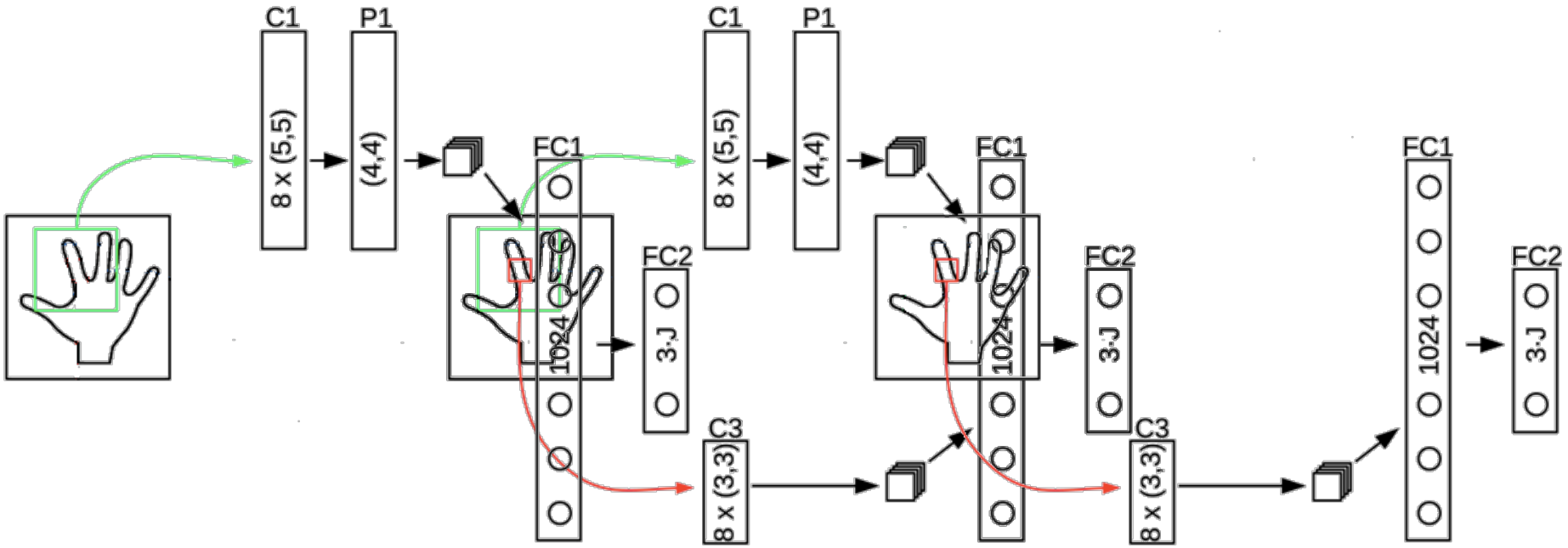
- Tradeoff pooling: uncertainty vs. #parameters



Refinement

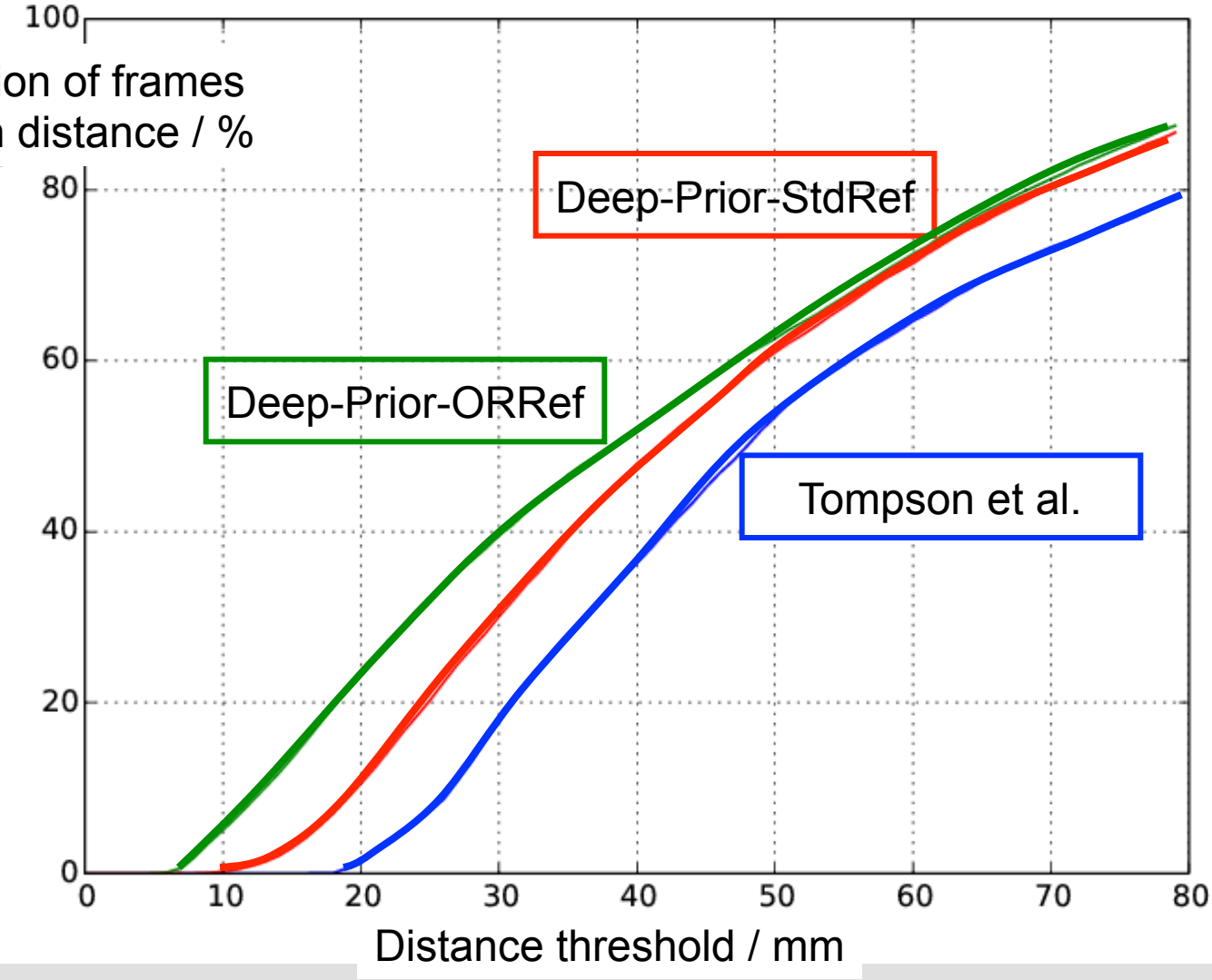
- Overlapping regions with different pooling size

→ accuracy & context



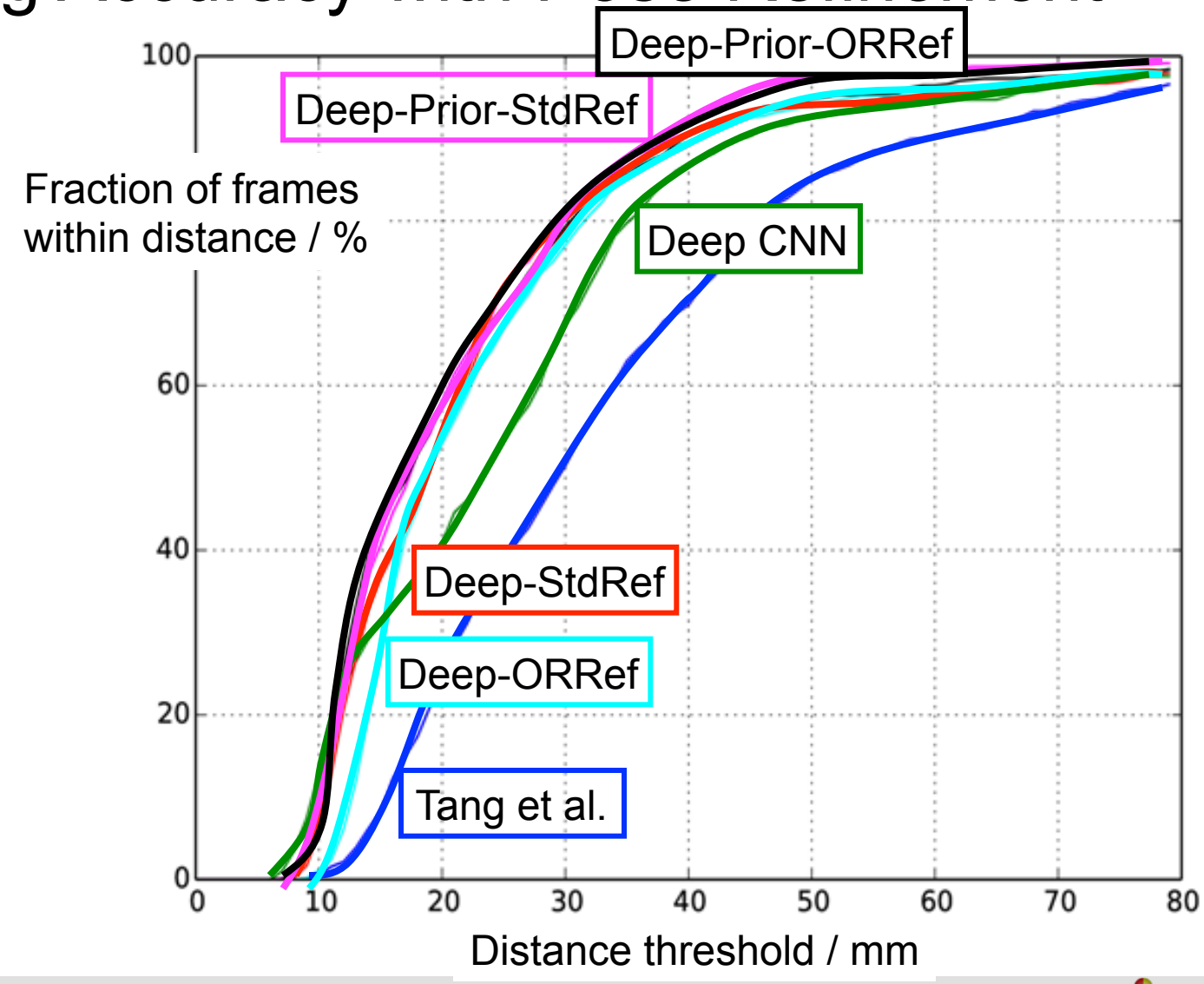
Increasing Accuracy with Pose Refinement

NYU dataset



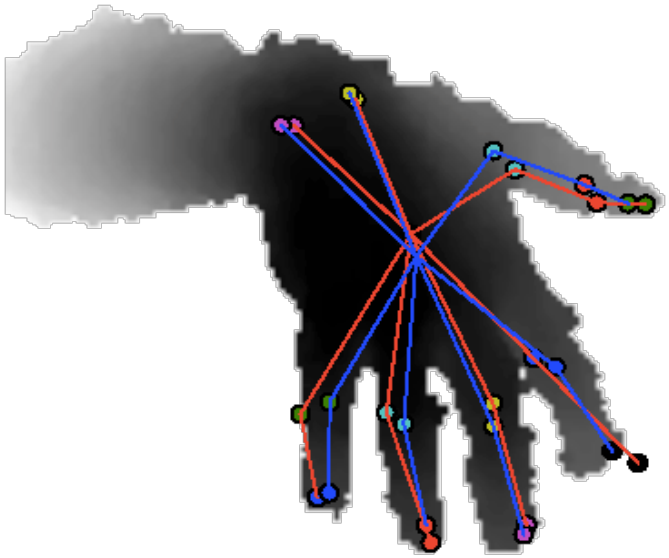
Increasing Accuracy with Pose Refinement

ICVL dataset

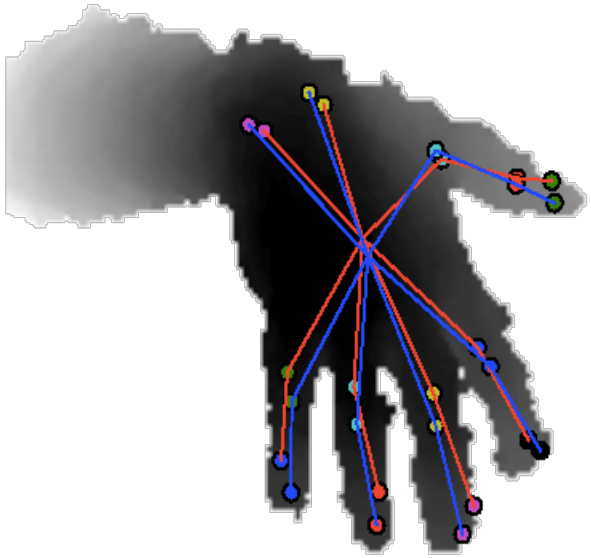


Qualitative Results - NYU

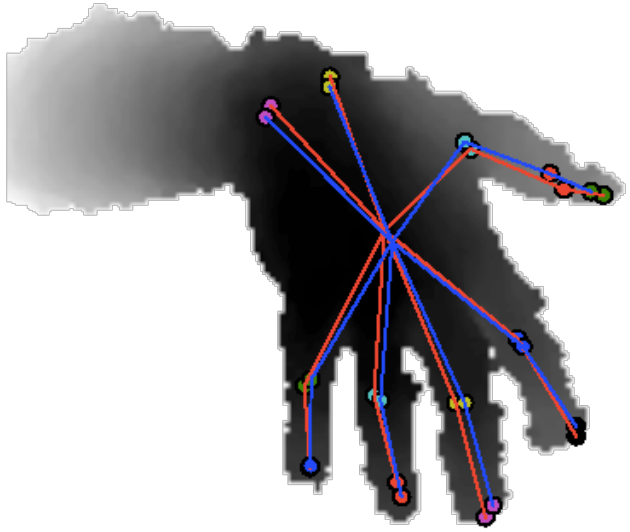
Deep CNN



Deep-Prior CNN



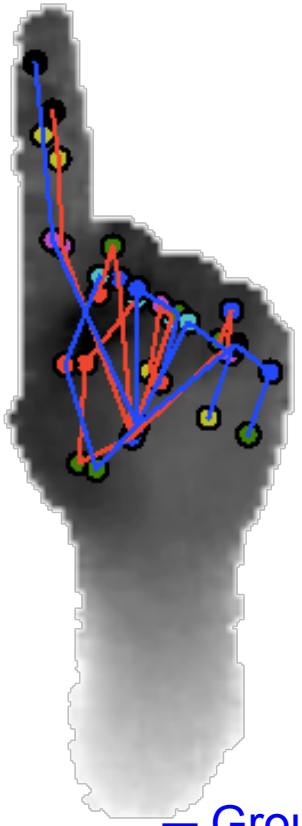
Refinement



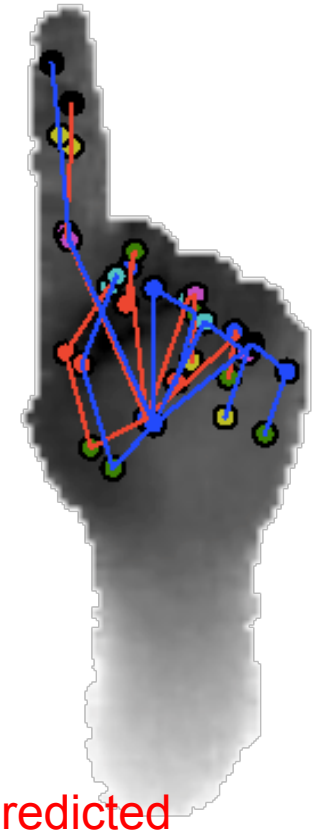
— Ground truth, — Predicted

Qualitative Results - ICVL

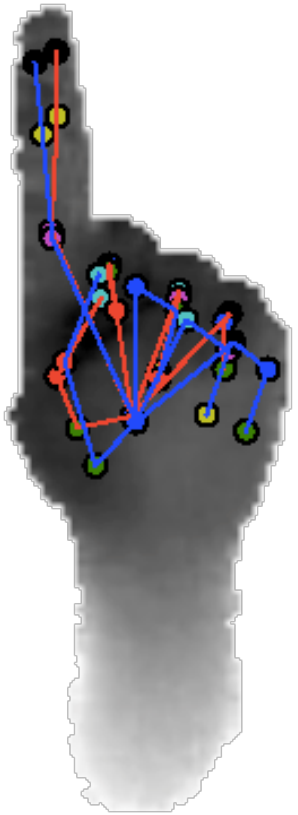
Deep CNN



Deep-Prior CNN



Refinement



— Ground truth, — Predicted

Running Times

- Up to 5000 fps on a single GPU
- Training ~5 hours

Architecture	GPU	CPU
Shallow	0.01ms	1.85ms
Deep	0.10ms	2.08ms
Multi-Scale	0.81ms	5.36ms
Deep-Prior	0.09ms	2.29ms
Refinement	2.38ms	62.91ms
[Qian CVPR'14]	-	40ms
[Tompson ToG'14]	5.6ms	-
[Tang CVPR'13]	-	16ms

Not optimized!

Conclusion

- Hand pose estimation
- Regression CNN on 3D joint positions
- Fast and high accuracy
- Pose prior integrated in CNN: increased accuracy of full pose, high speed
- Refinement: independed for each joint, no pooling for higher accuracy

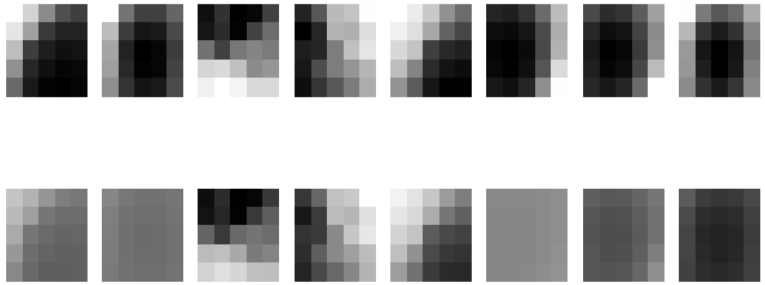
Thanks!

Questions?

{oberweger, wohlhart, lepetit}@icg.tugraz.at

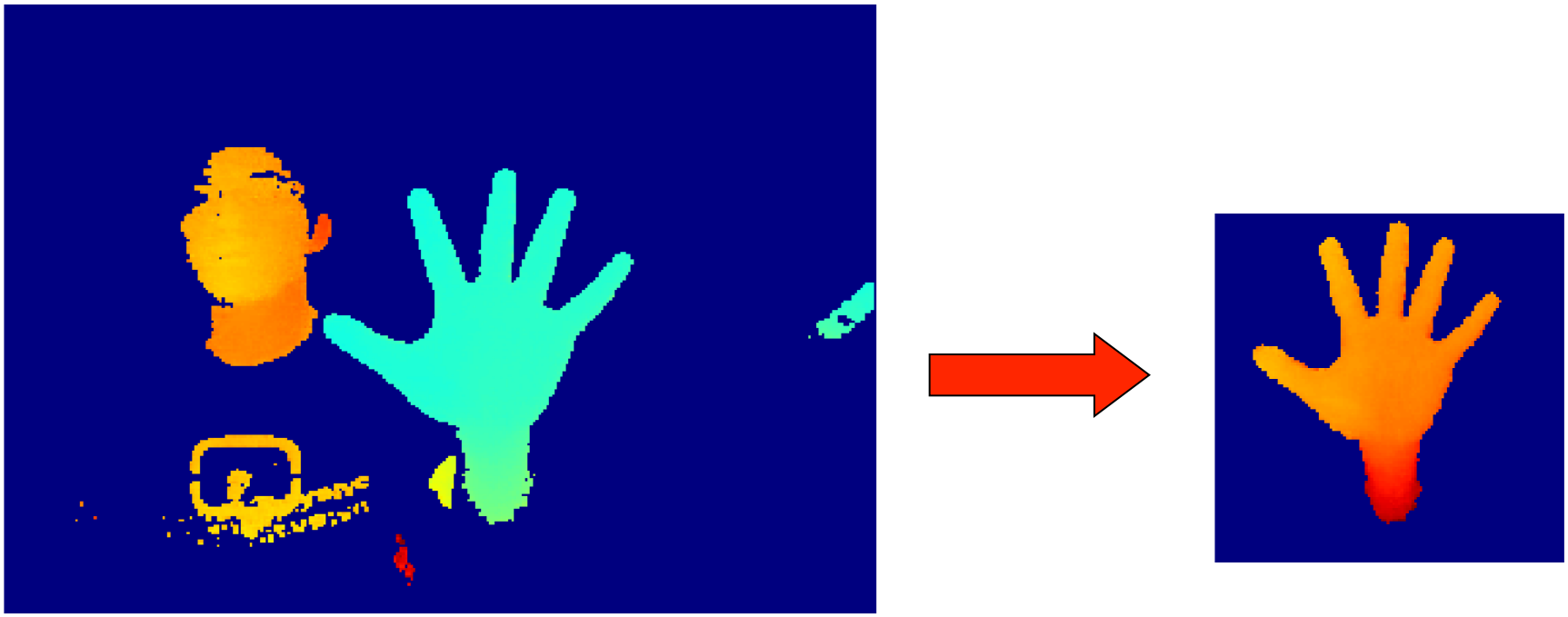
CNN

Learned Filters



Preprocessing

- Preprocessing – important for CNN
  - Hand detection
  - Cropping and normalization

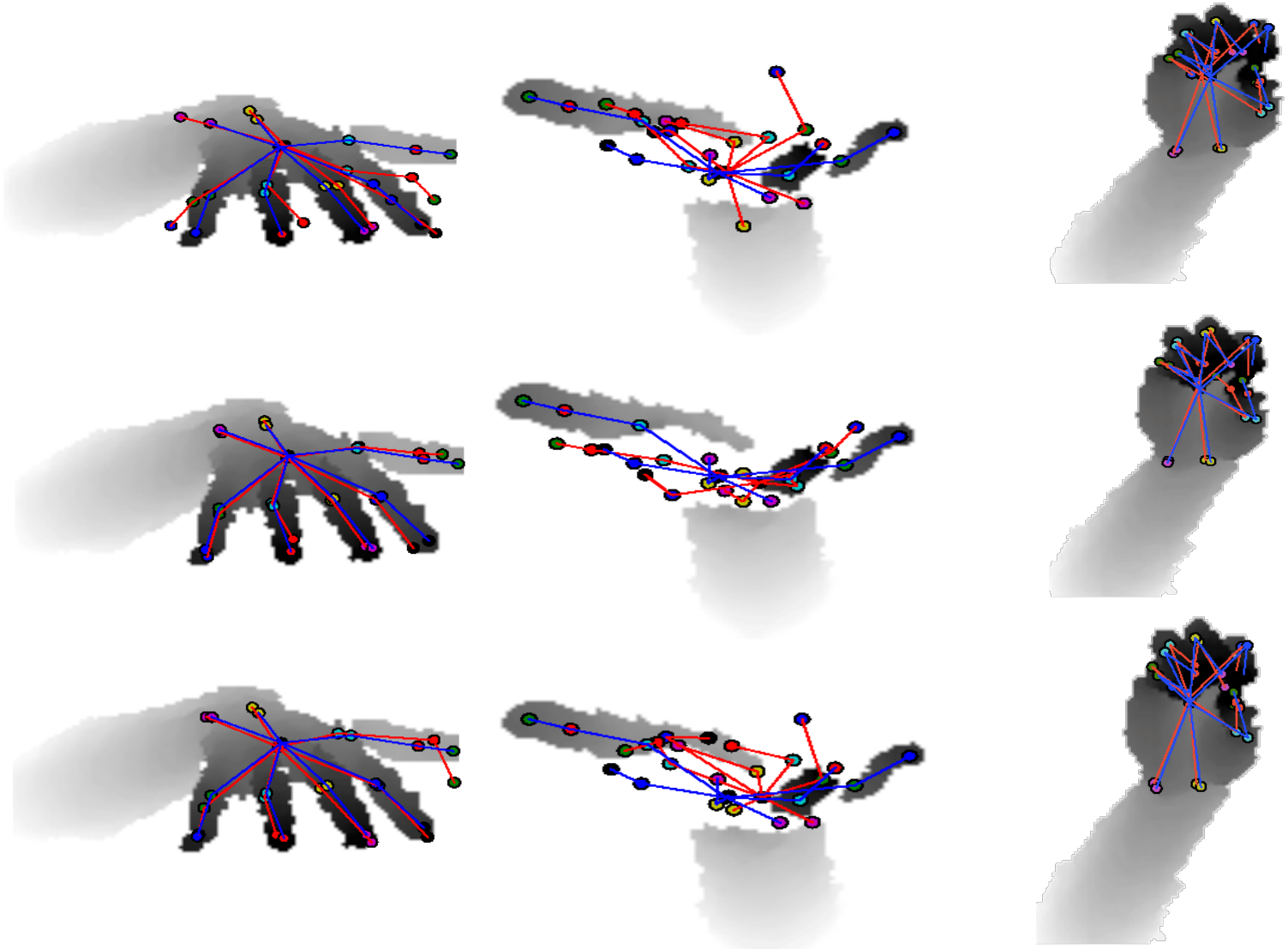


Qualitative Results - NYU

Deep

Deep-Prior

Refinement

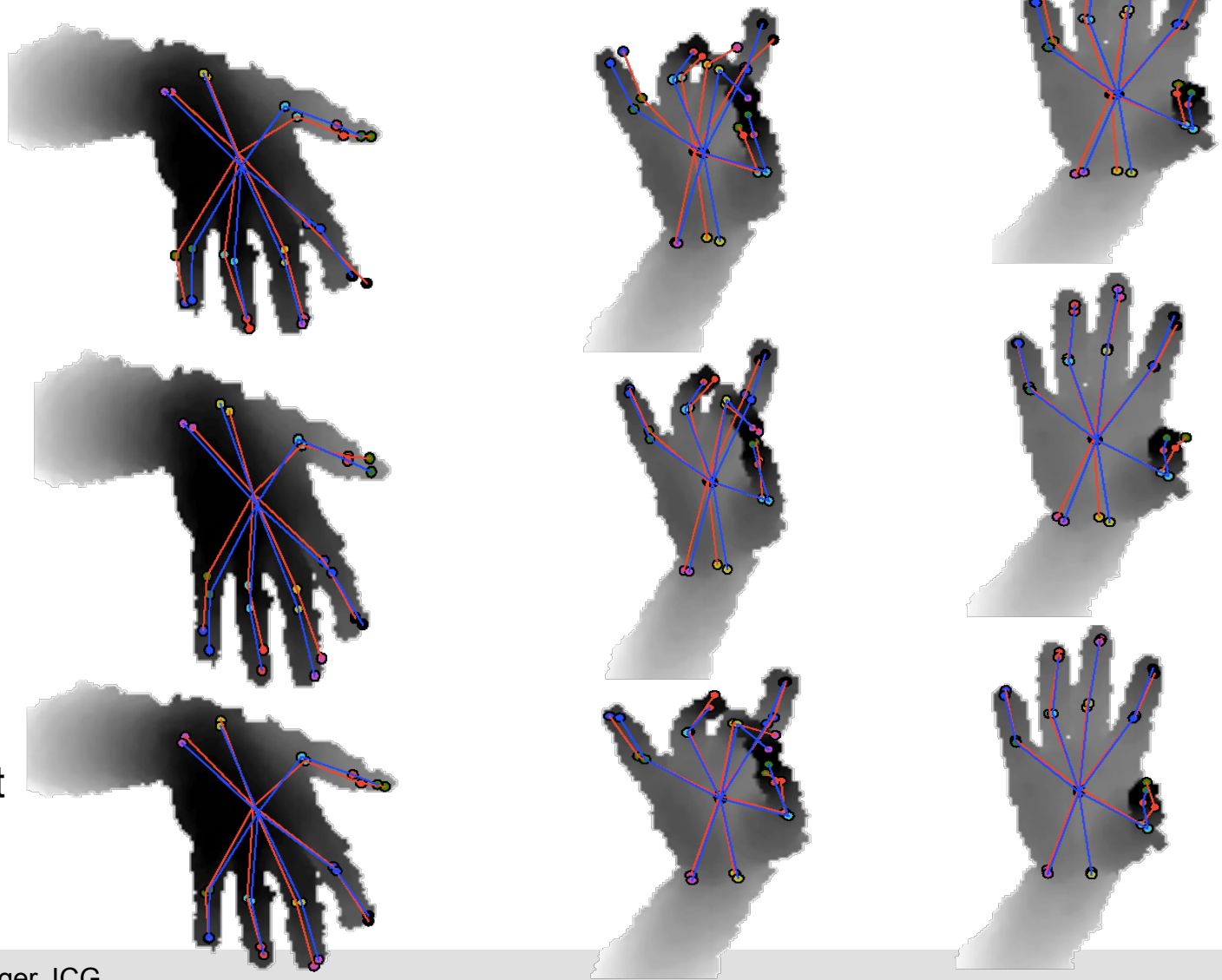


Qualitative Results - NYU

Deep

Deep-Prior

Refinement

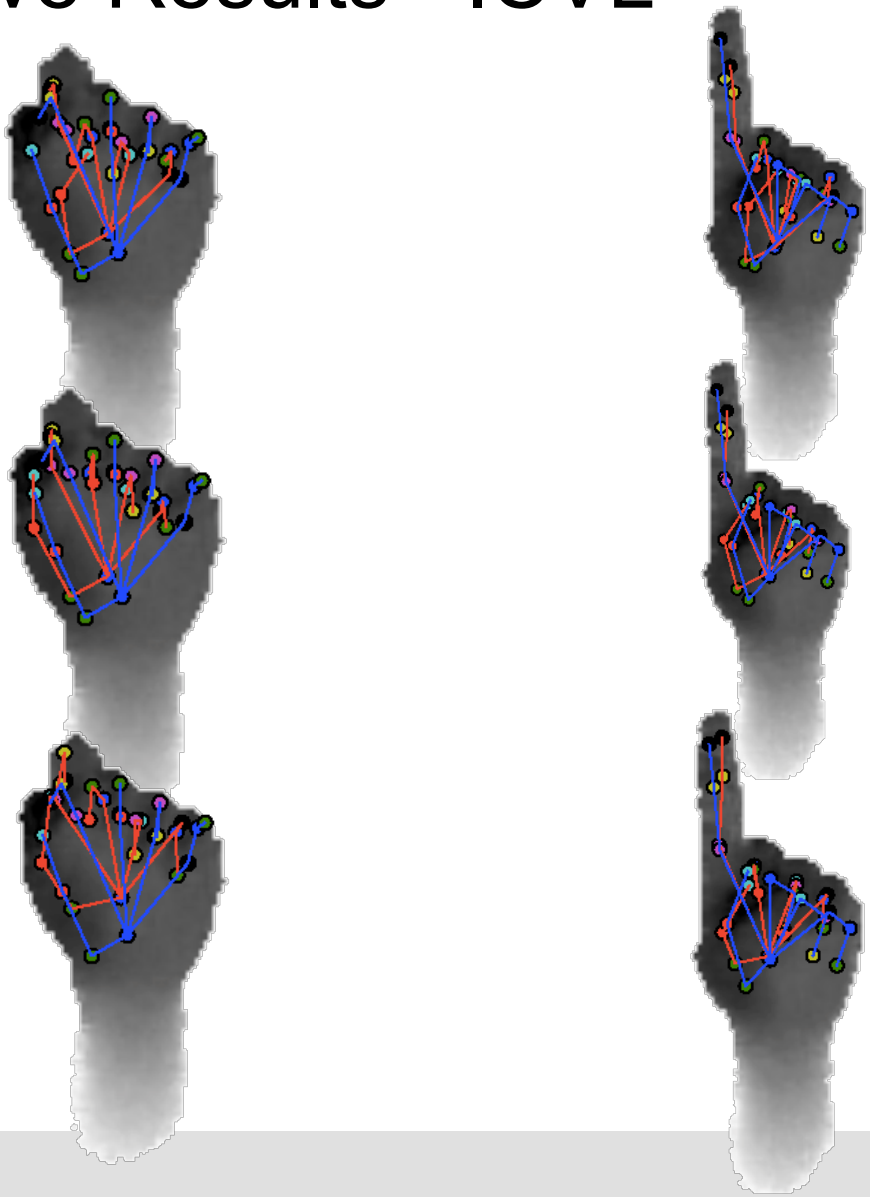


Qualitative Results - ICVL

Deep

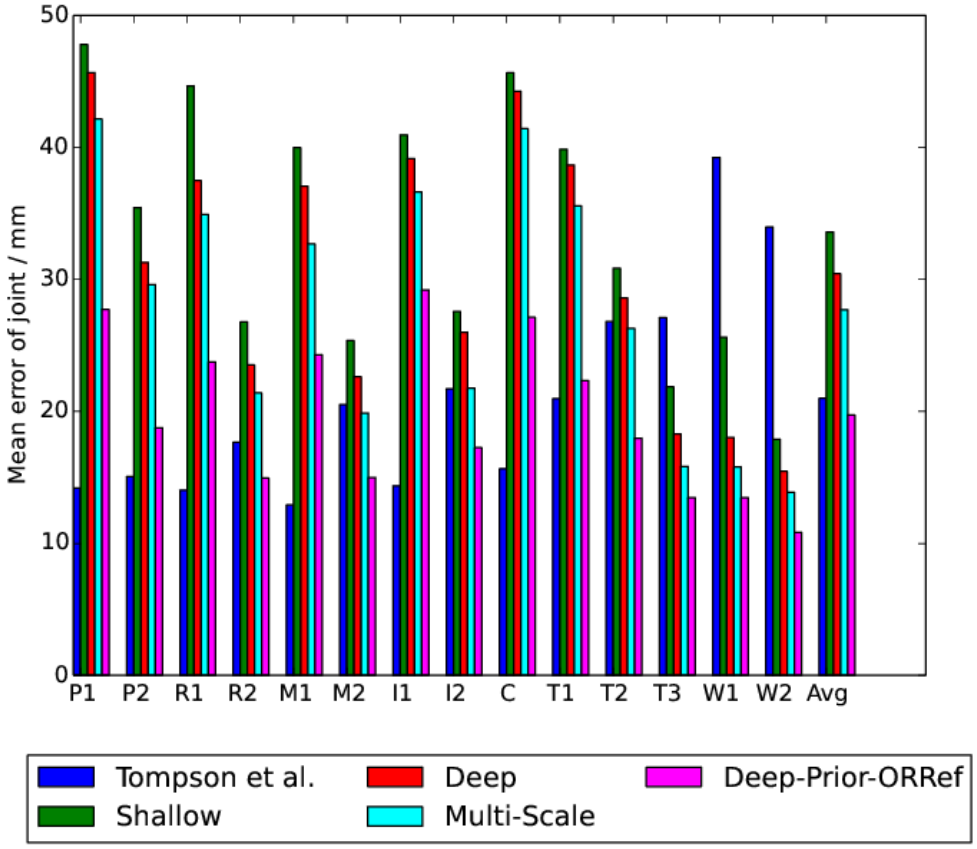
Deep-Prior

Refinement

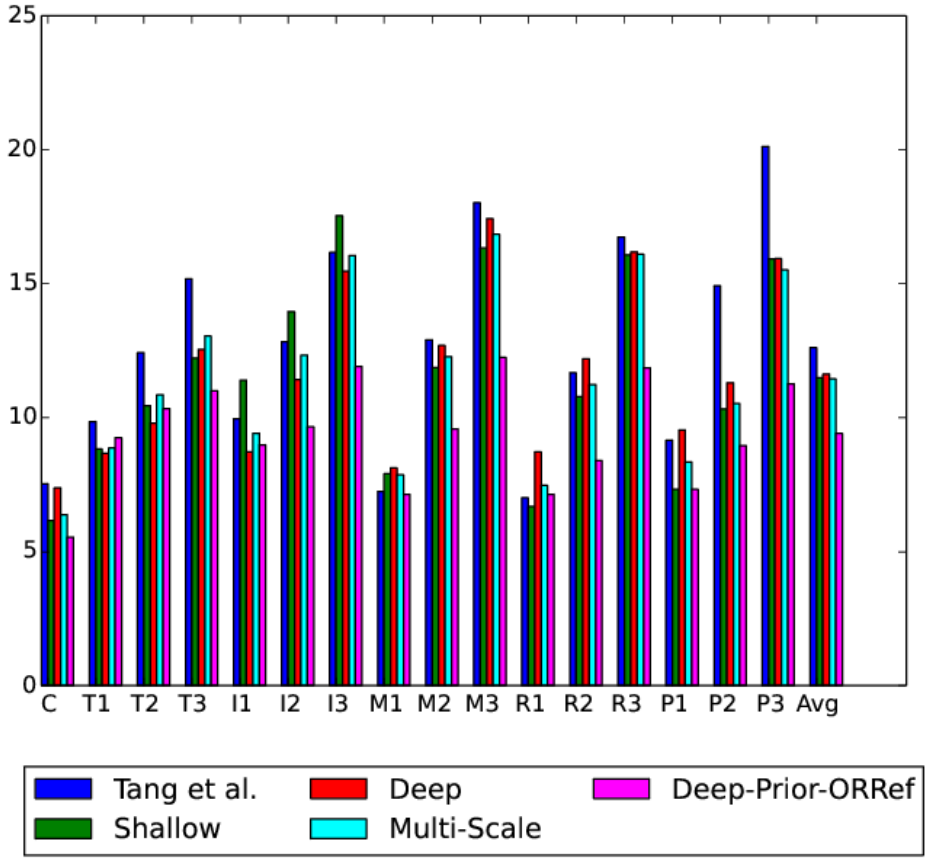


Increasing Accuracy with Pose Refinement

NYU

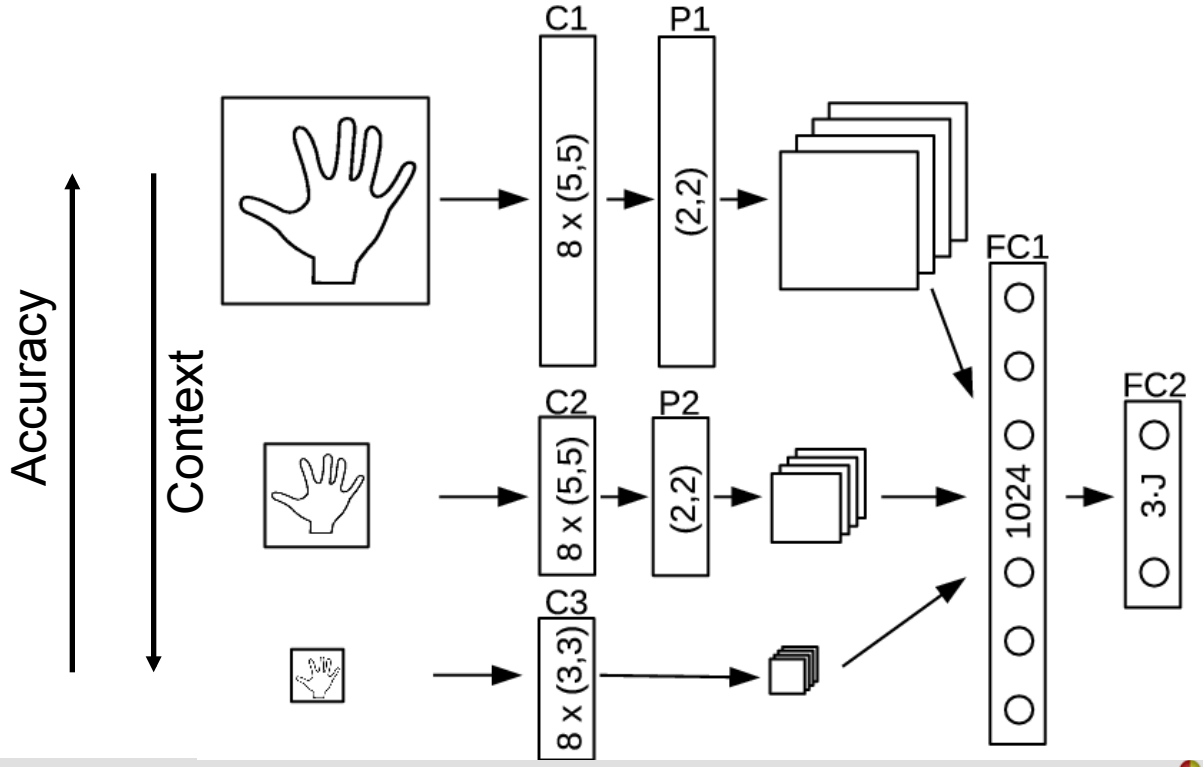


ICVL



CNN

- A first improvement: Multi-Scale CNN e.g. [Farabet TPAMI'13]
- Capture different scales



CNN

Still, we need something better!