DIY A Multiview Camera System: Panoptic Studio Teardown

How To Use Data From A Multiview System

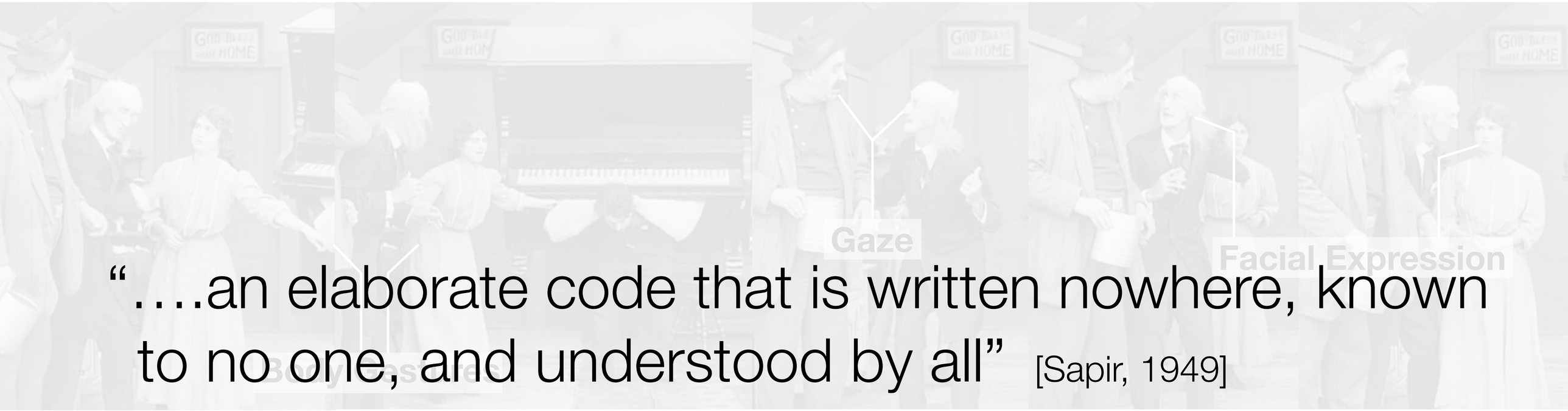
Hanbyul Joo and Tomas Simon

Robotics Institute
Carnegie Mellon University

 His Musical Career (with Charlie Chaplin) 1914

Non-Verbal Signals Convey Remarkable Information

a.k.a. Body Langauge



"….an elaborate code that is written nowhere, known to no one, and understood by all" [Sapir, 1949]

Kinesics:

"The study of the way in which certain **body movements and gestures** serve as a form of **nonverbal communication**." [Birdwhistell 1970]

The Panoptic Studio

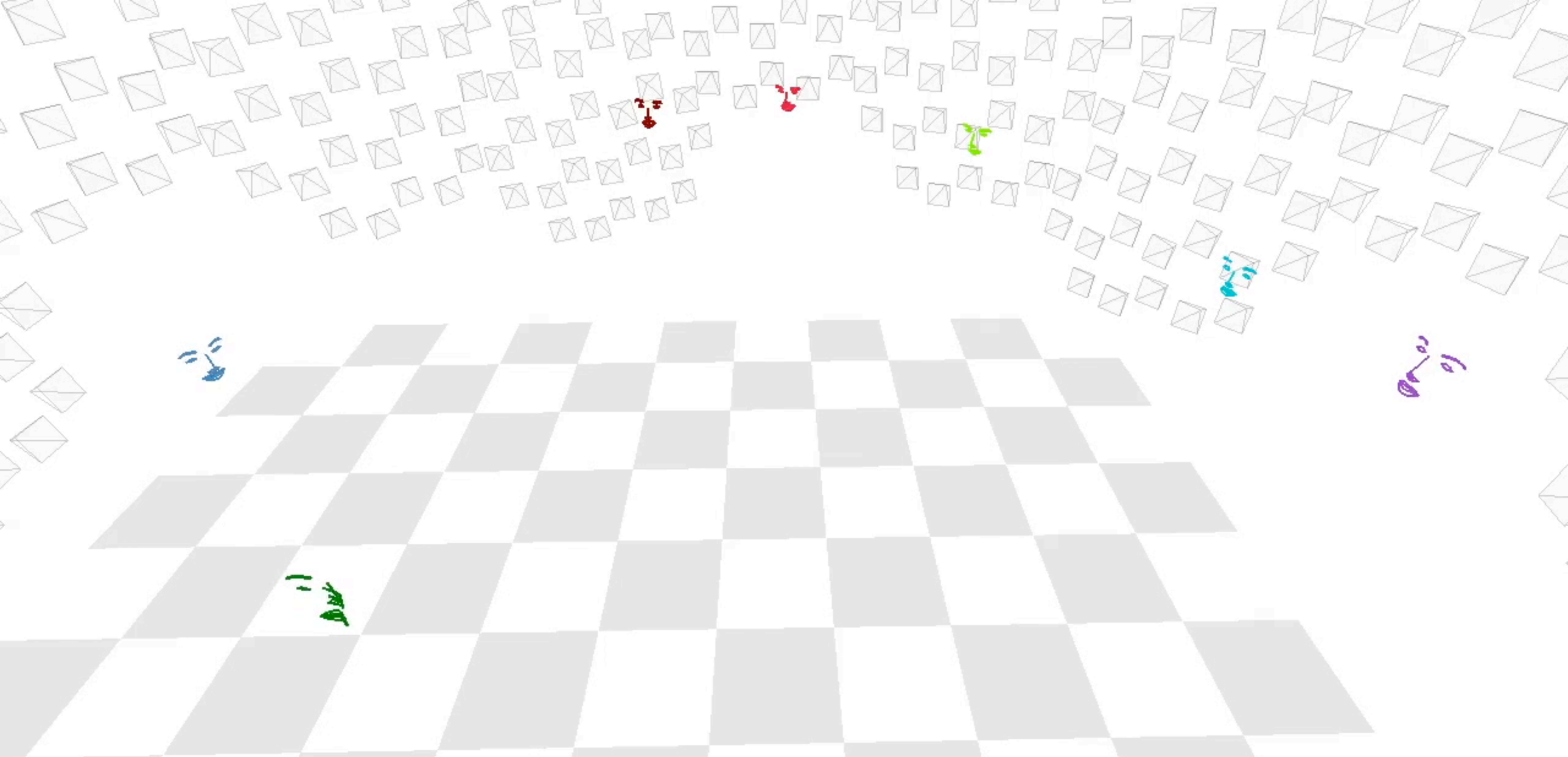
Modularized Design with 20 Panels

480 VGA Cameras
31 HD Cameras
10 Kinects

Synchronized Videos from Unique 521 Views

480 VGAs, 31HDs, and 10 RGB+Ds


Multiple naturally interacting people

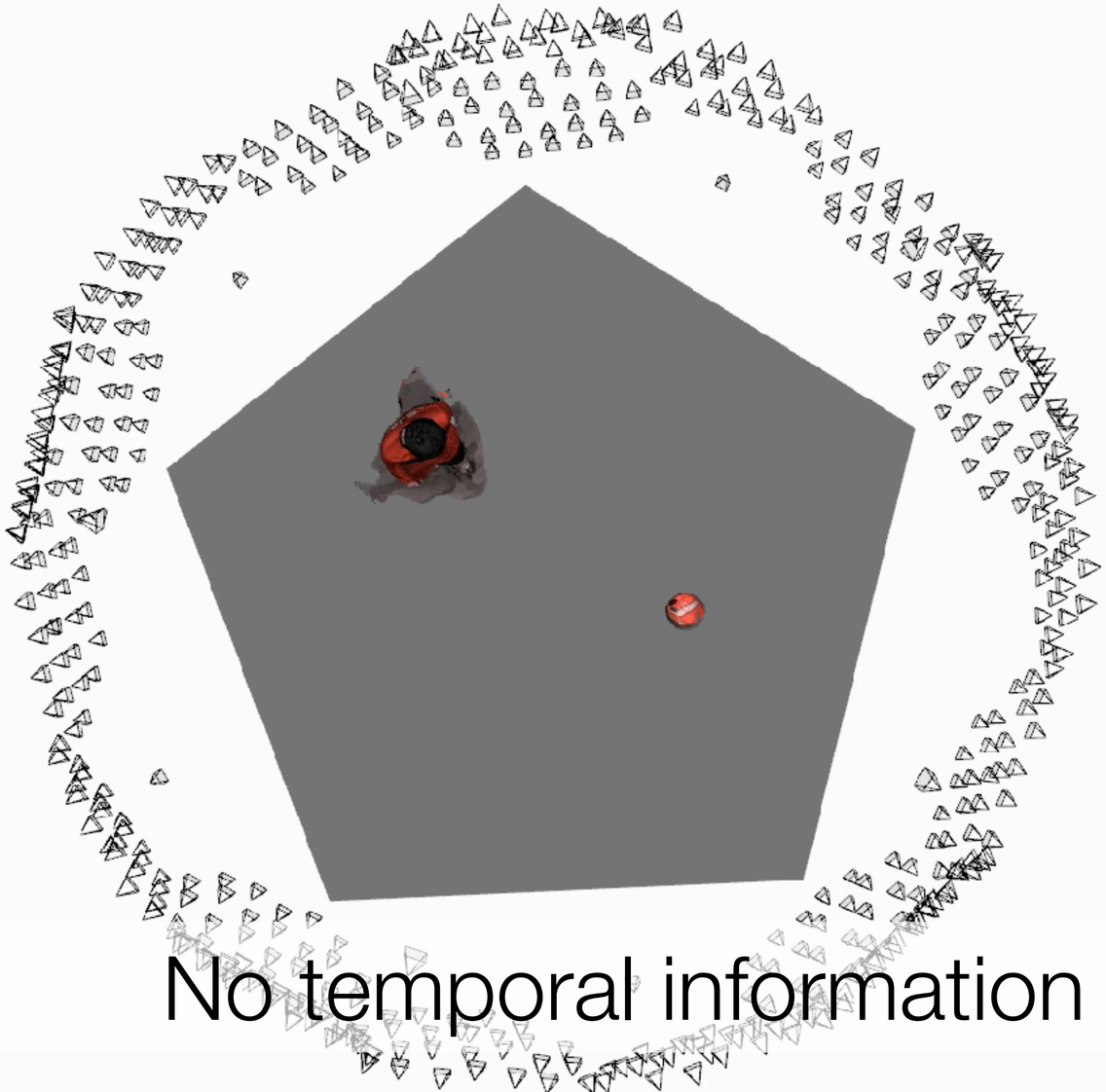


Measuring Social Signals at High-Resolution

Measuring Face, Body, Hand, and Surface Trajectories

Measuring 3D Volume

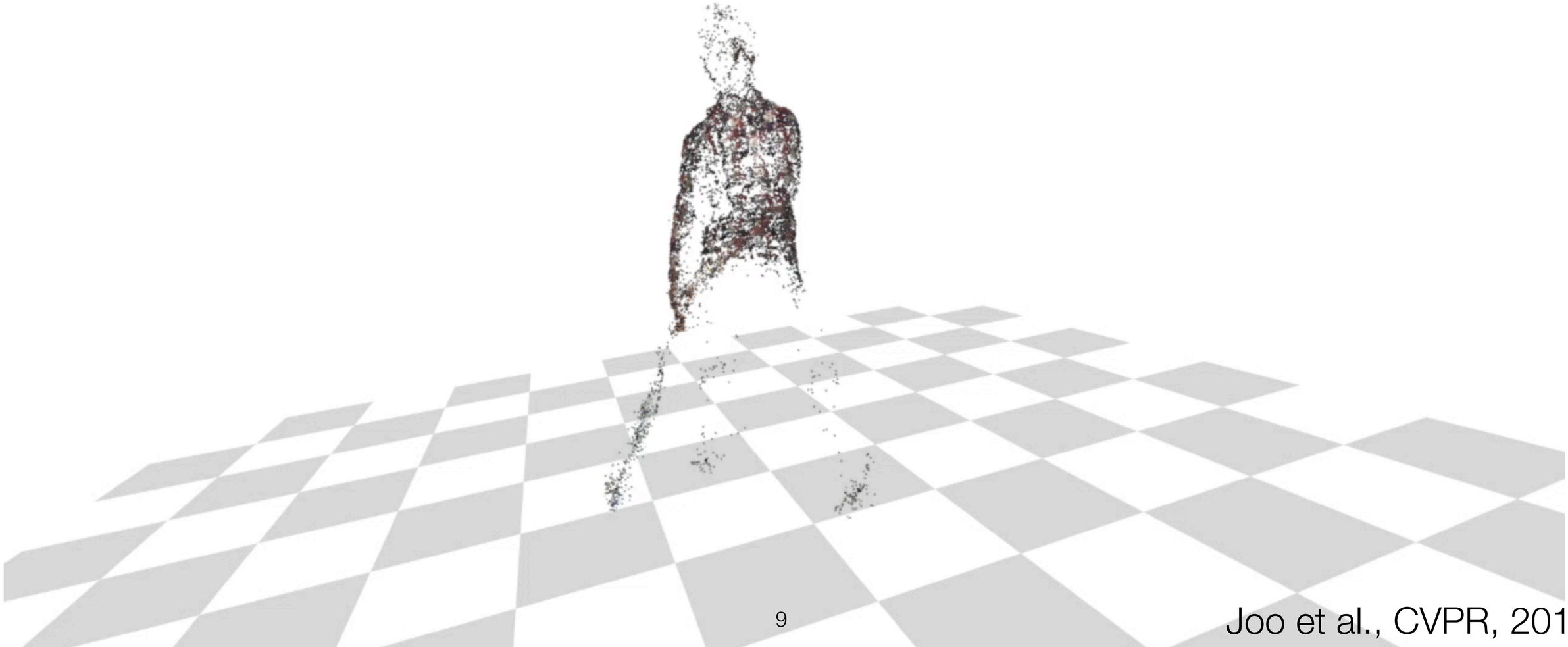
Visual Hull



No temporal information

Measuring 3D Motion

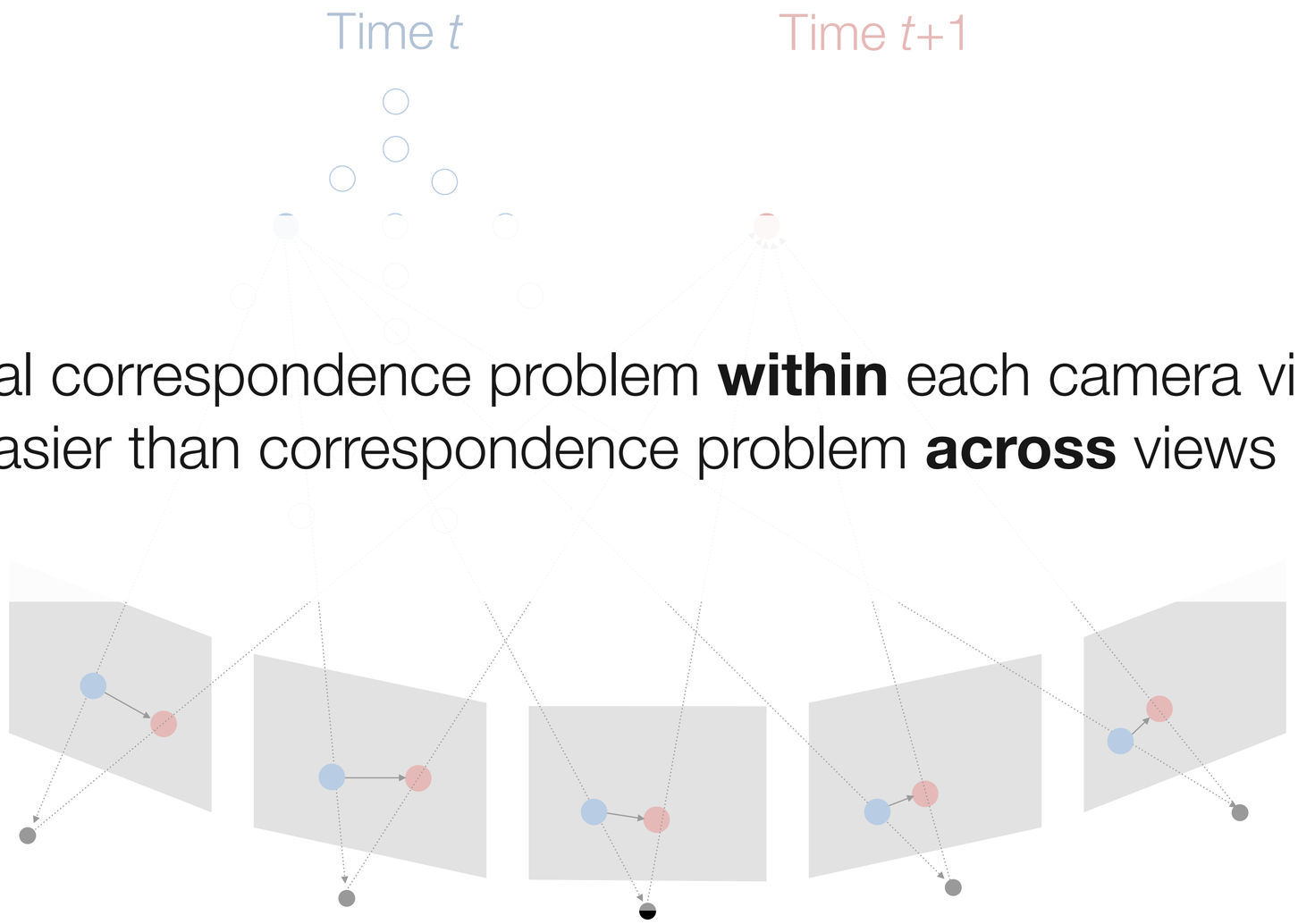
Dense Long-term 3D Trajectory Stream



Joo et al., CVPR, 2014

Measuring Dense 3D Motion

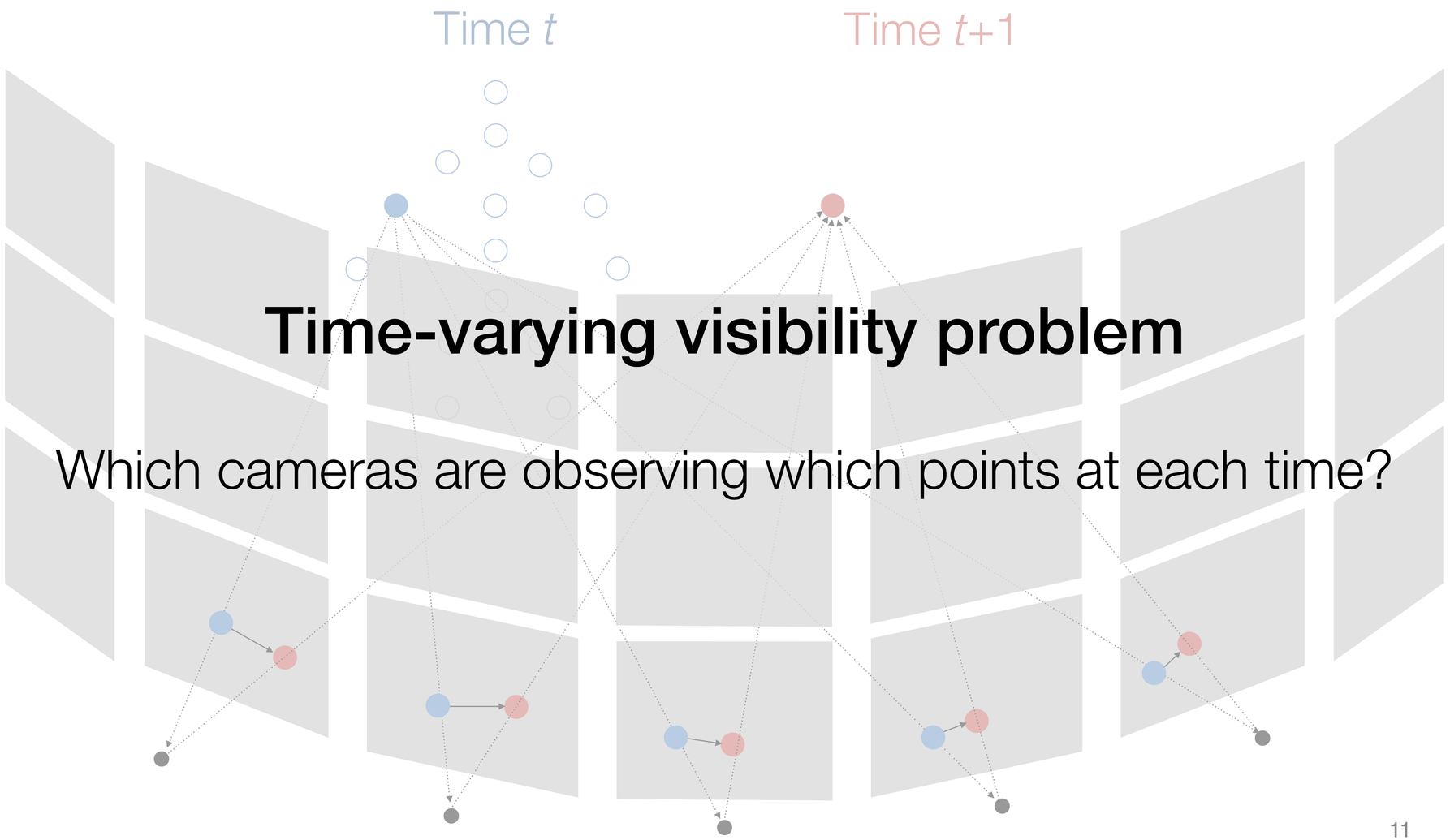
Leveraging “Flows” in A Large Number Views



Temporal correspondence problem **within** each camera view is much easier than correspondence problem **across** views

Measuring Dense 3D Motion

Key Issue To Leverage a Large Number of Views

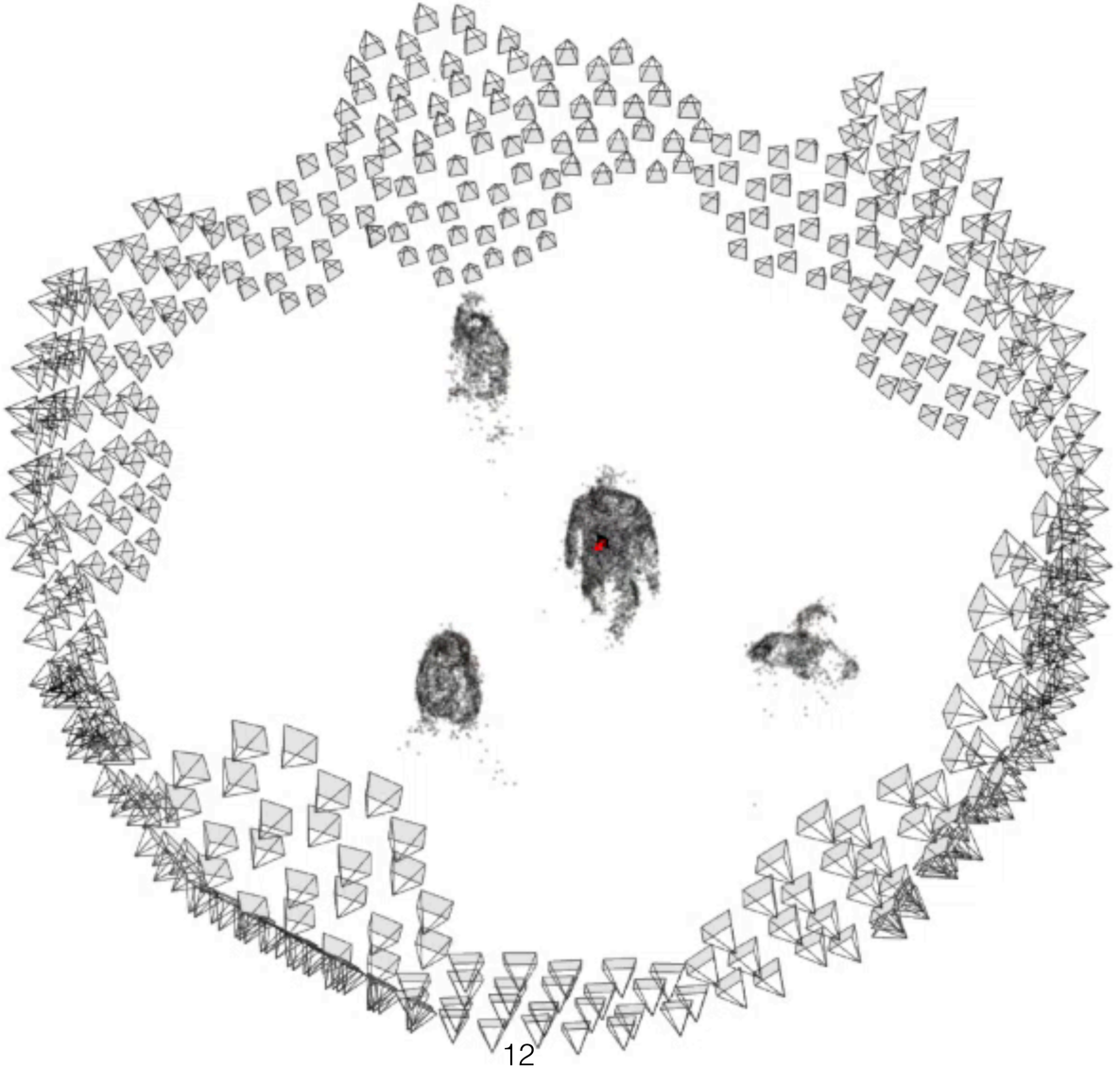


Time-varying visibility problem

Which cameras are observing which points at each time?

A Core Idea

Reasoning About Time Varying Visibility



Trajectory Stream Reconstruction

The Volleyball Sequence

Trajectory Stream Reconstruction

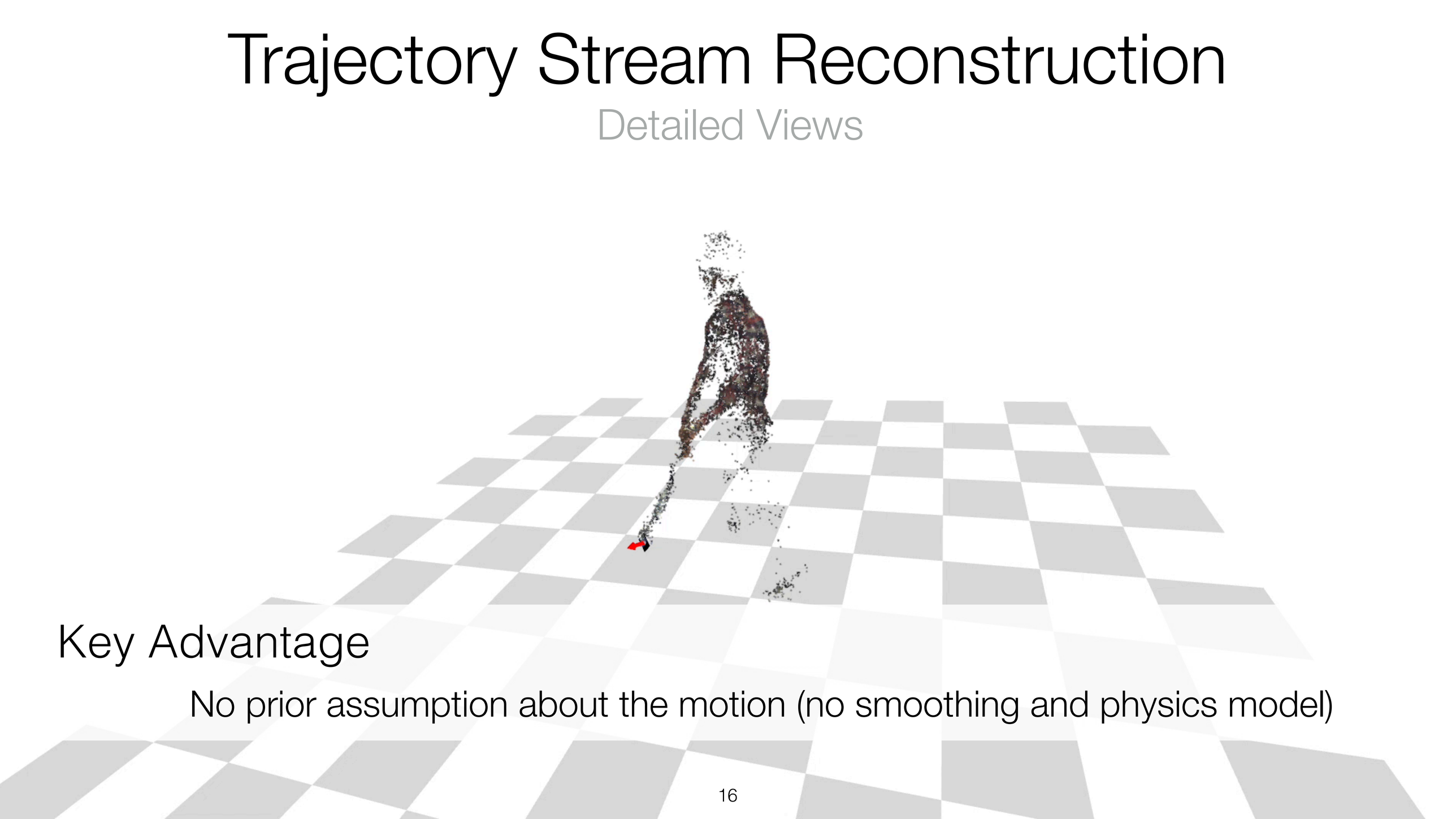
The Confetti Sequence

Trajectory Stream Reconstruction

The Fluid Motion Sequence

Trajectory Stream Reconstruction

Detailed Views

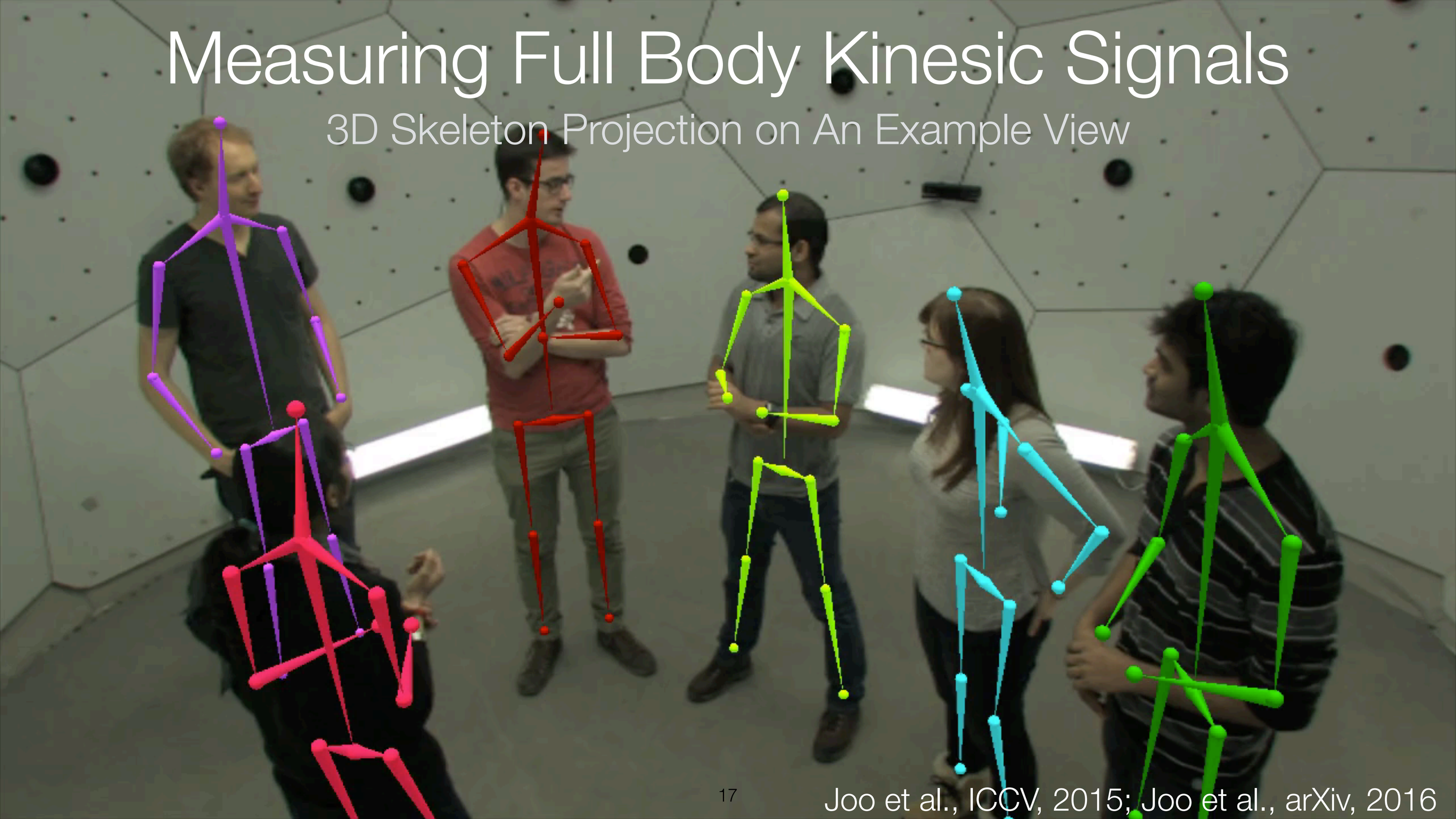


Key Advantage

No prior assumption about the motion (no smoothing and physics model)

Measuring Full Body Kinesic Signals

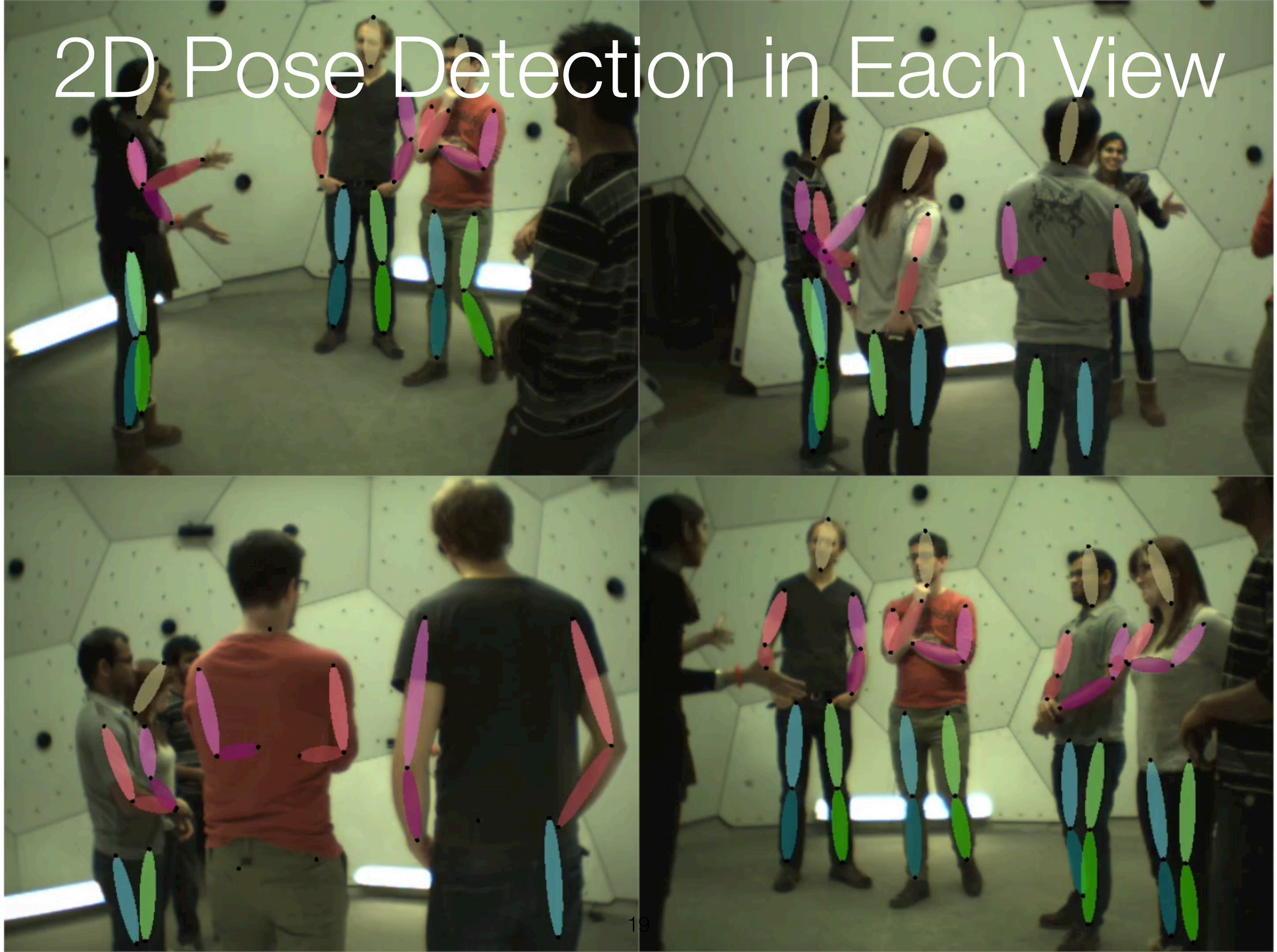
3D Skeleton Projection on An Example View



Joo et al., ICCV, 2015; Joo et al., arXiv, 2016

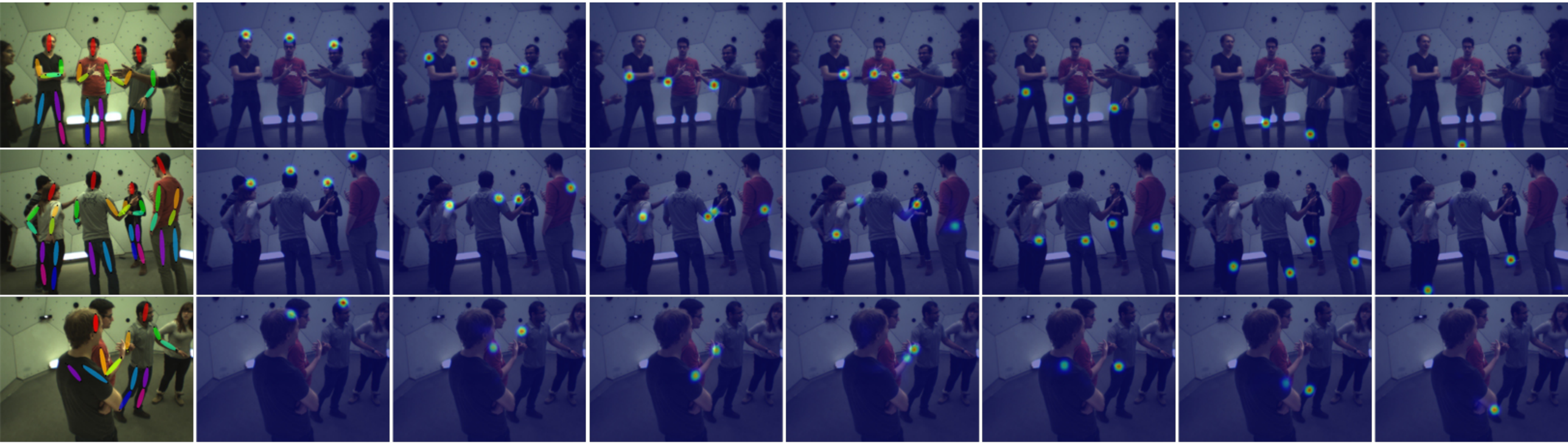
Zhe et al., Realtime Multi-person 2D Pose Estimation using Part Affinity Fields , CVPR 2017

2D Pose Detection in Each View



2D Pose Detection in Each View

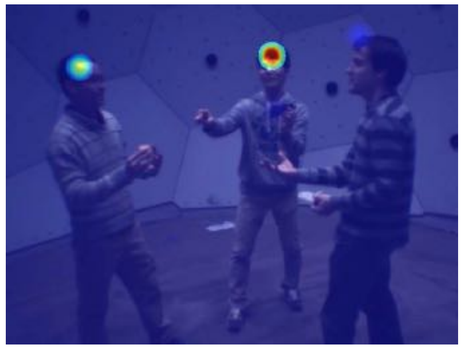
Score Map Generation



2D Pose Detection | HeadTop | Right Shoulder | Right Elbow | Right Wrist | Right Hip | Right Knee | Right Ankle

Generating 3D Node Score Maps

3D Voting from 2D Score Maps



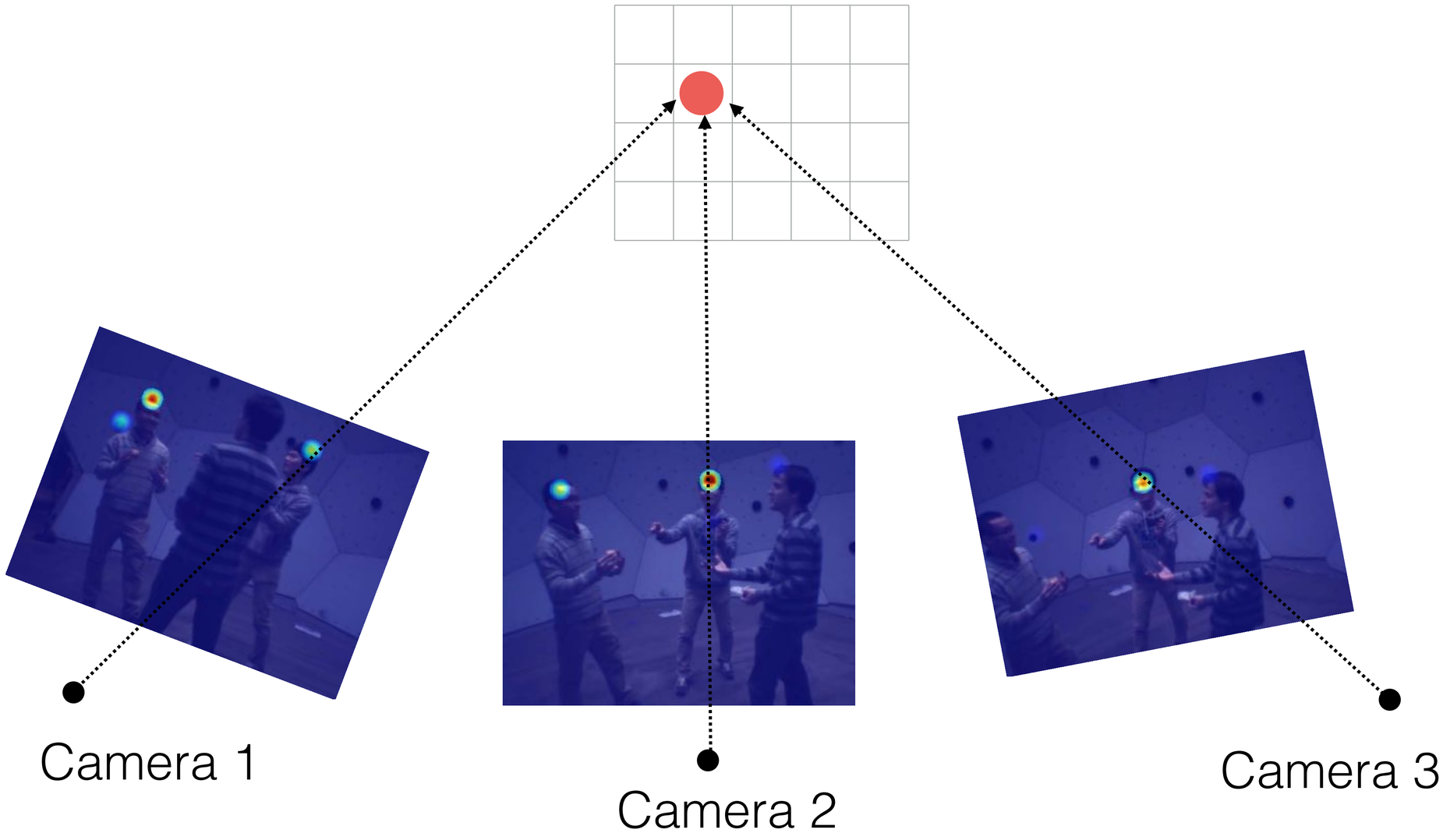
Camera 1

Camera 2

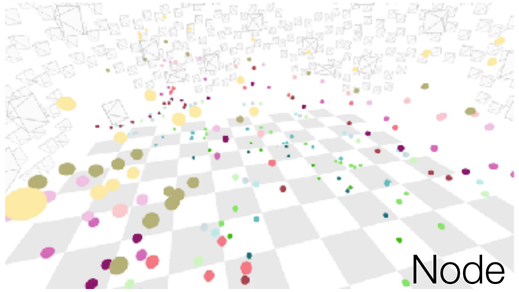
Camera 3

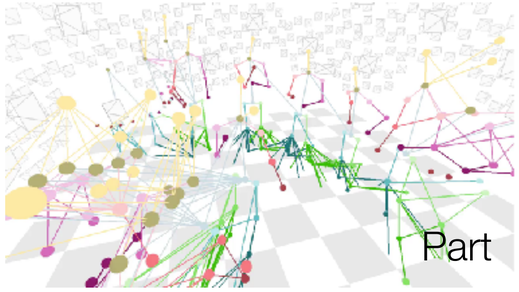
Generating 3D Node Score Maps

3D Voting from 2D Score Maps



Algorithm Flow





HeadTop/neck/bodyCenter

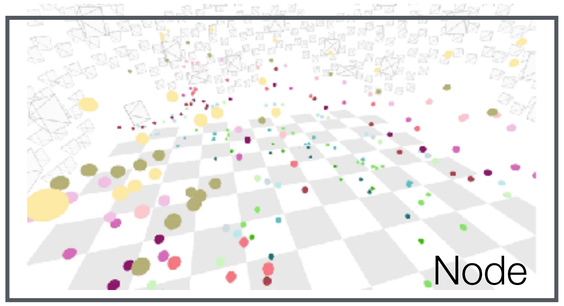
Left shoulder/elbow/wrist

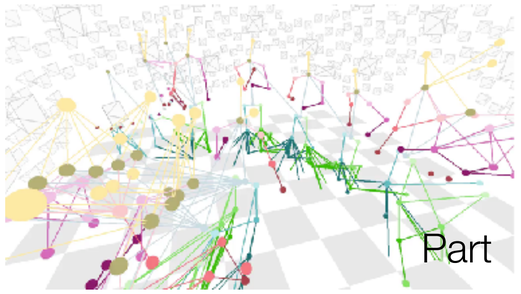
Right shoulder/elbow/wrist

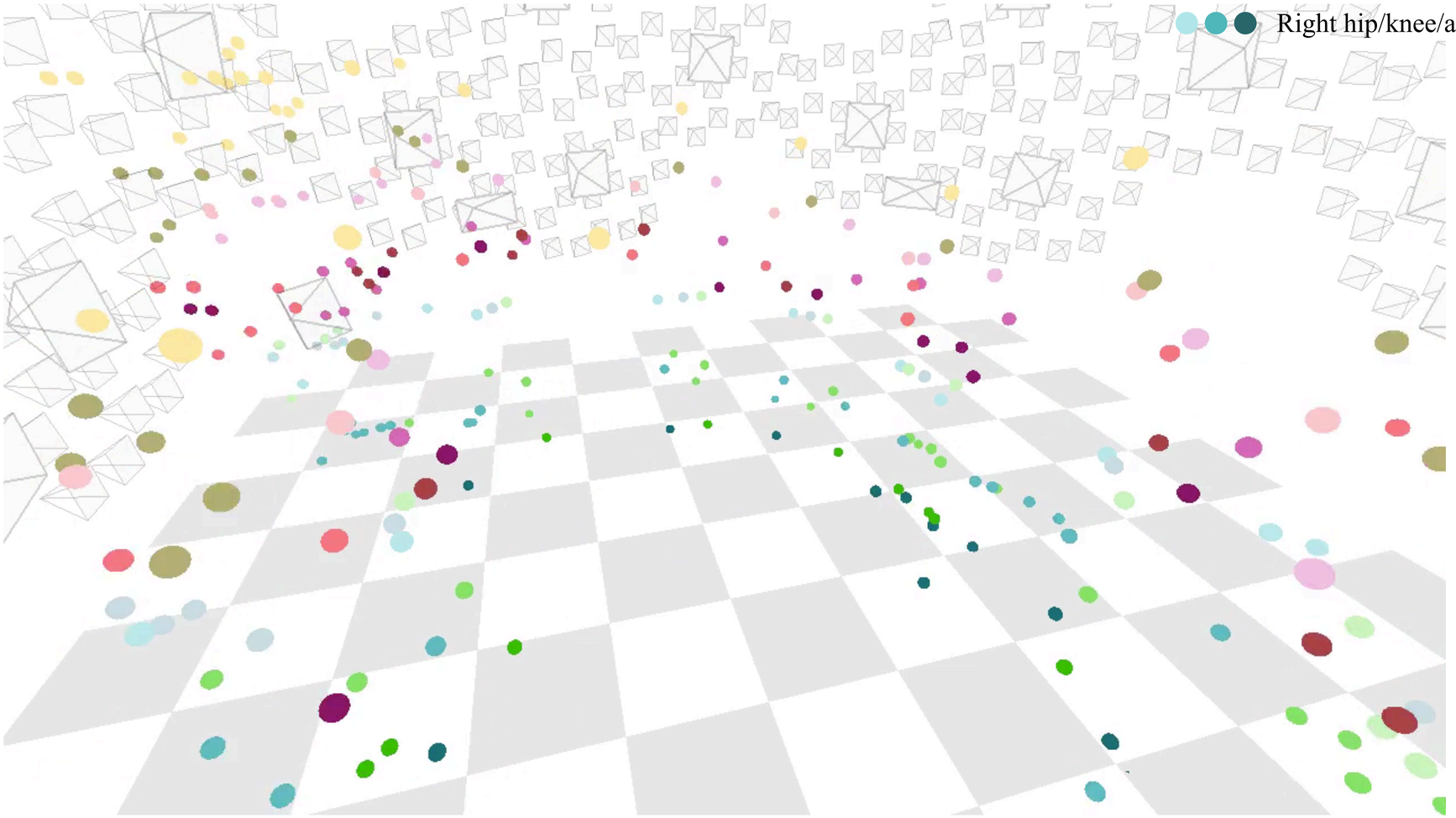
Left hip/knee/ankle

Right hip/knee/ankle

Algorithm Flow

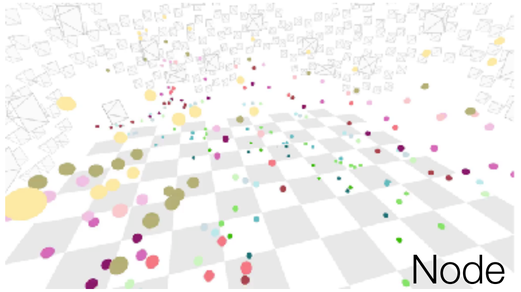


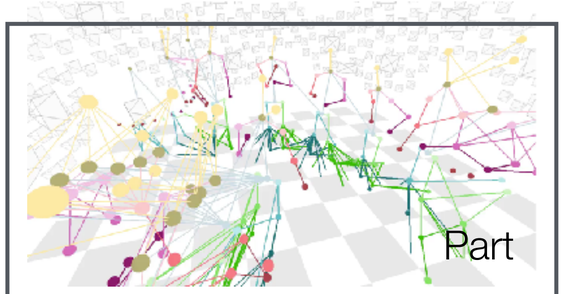


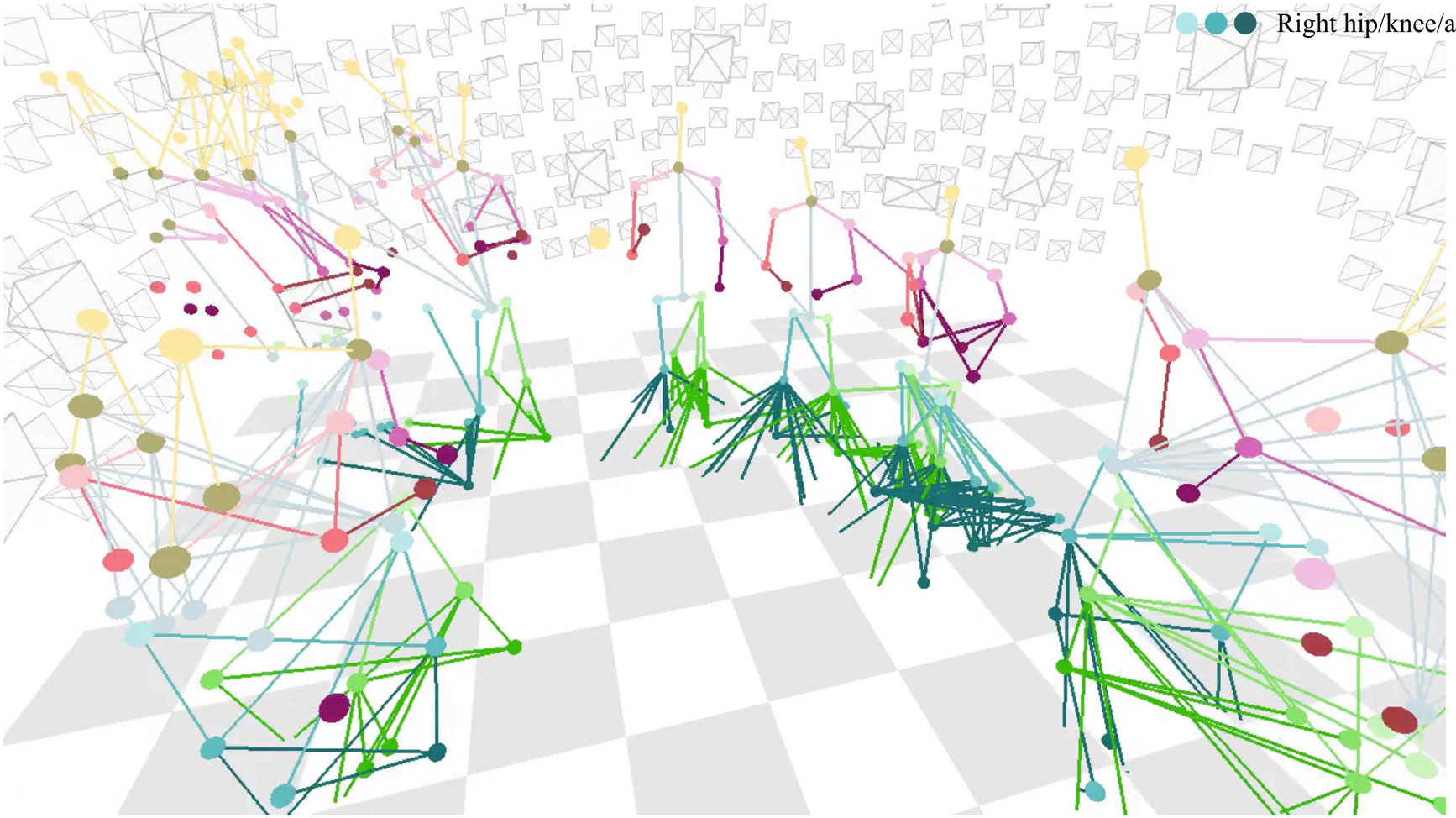


Generating “Node” Proposals

Algorithm Flow

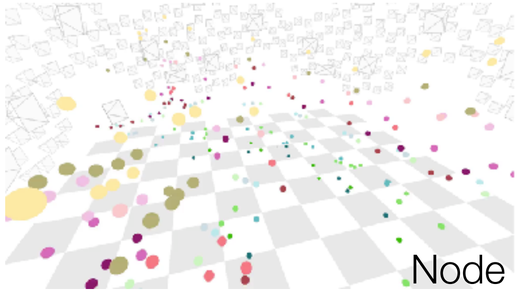


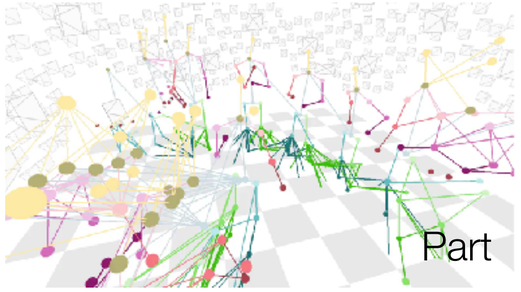


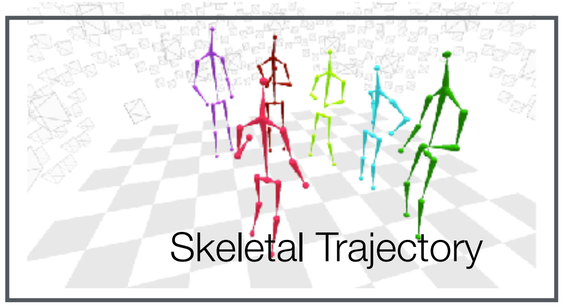


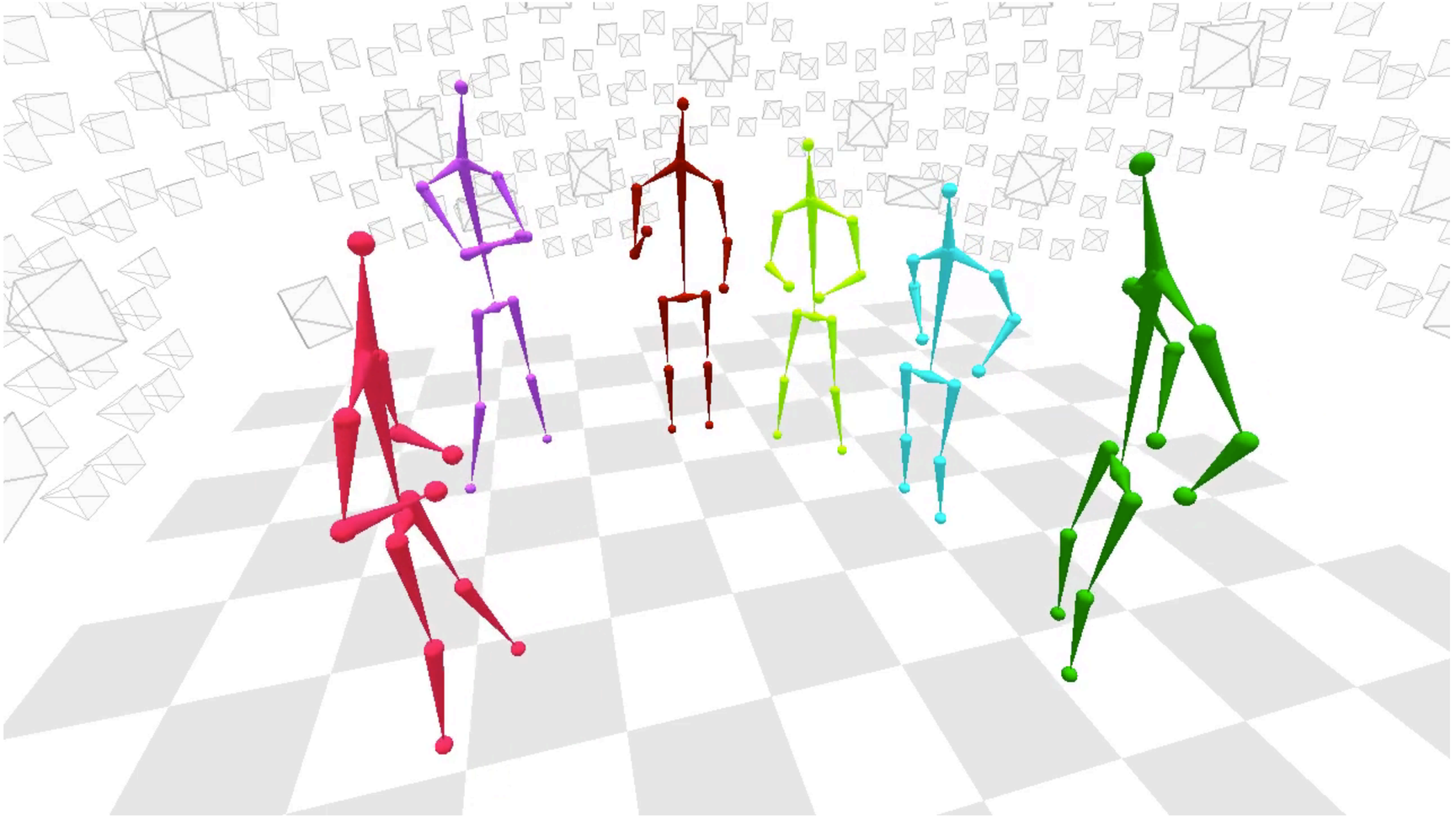
Generating “Part” Proposals

Algorithm Flow



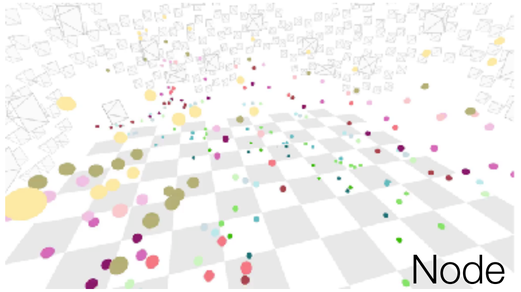


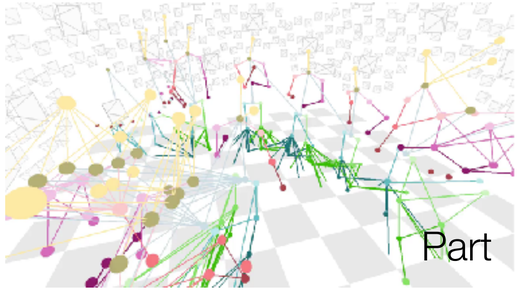


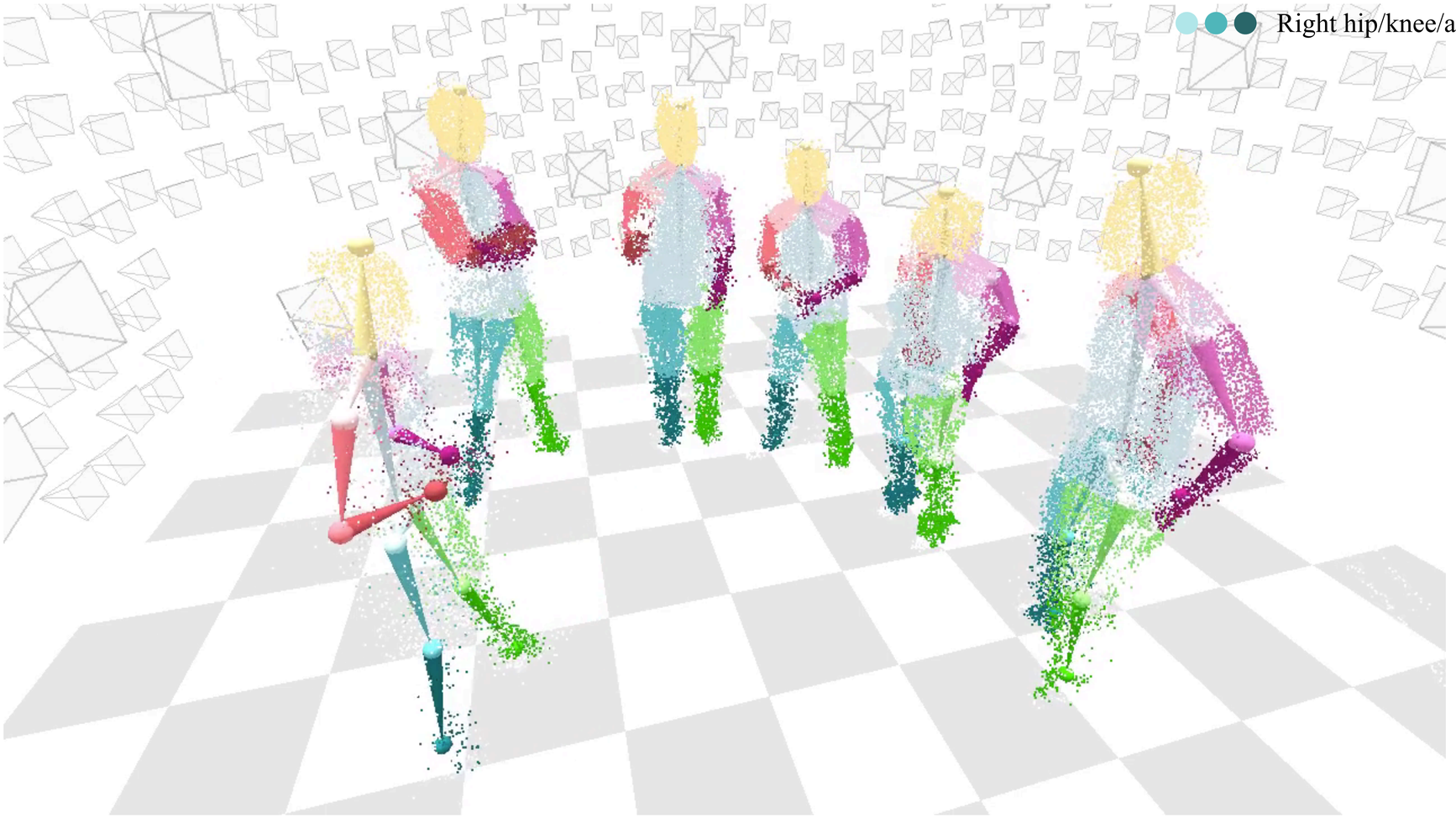


Generating “Skeletal” Proposals

Algorithm Flow

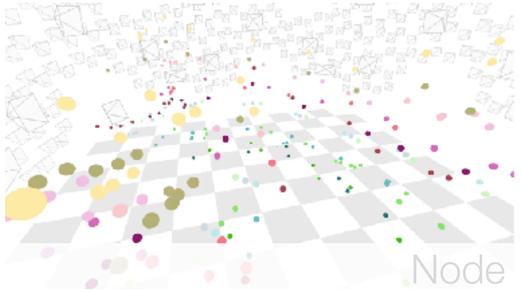






Associating with Dense 3D Trajectories
Temporal Refinement

Algorithm Flow

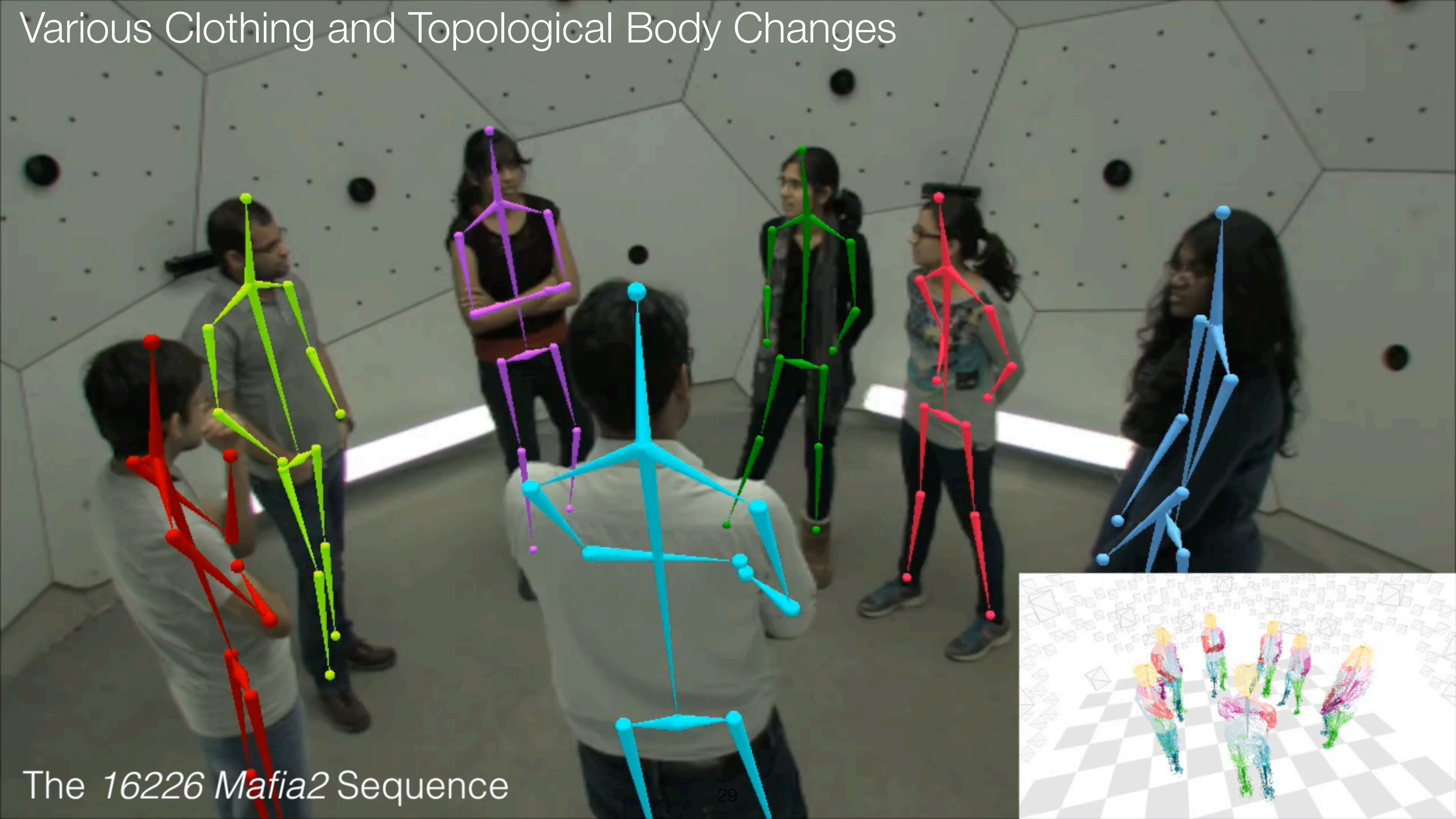


Key Advantages

- Fully automatic method
- No prior template generation
- No assumption about motion and appearance

Associating with Dense 3D Trajectories
Temporal Refinement

Various Clothing and Topological Body Changes



The *16226 Mafia2* Sequence

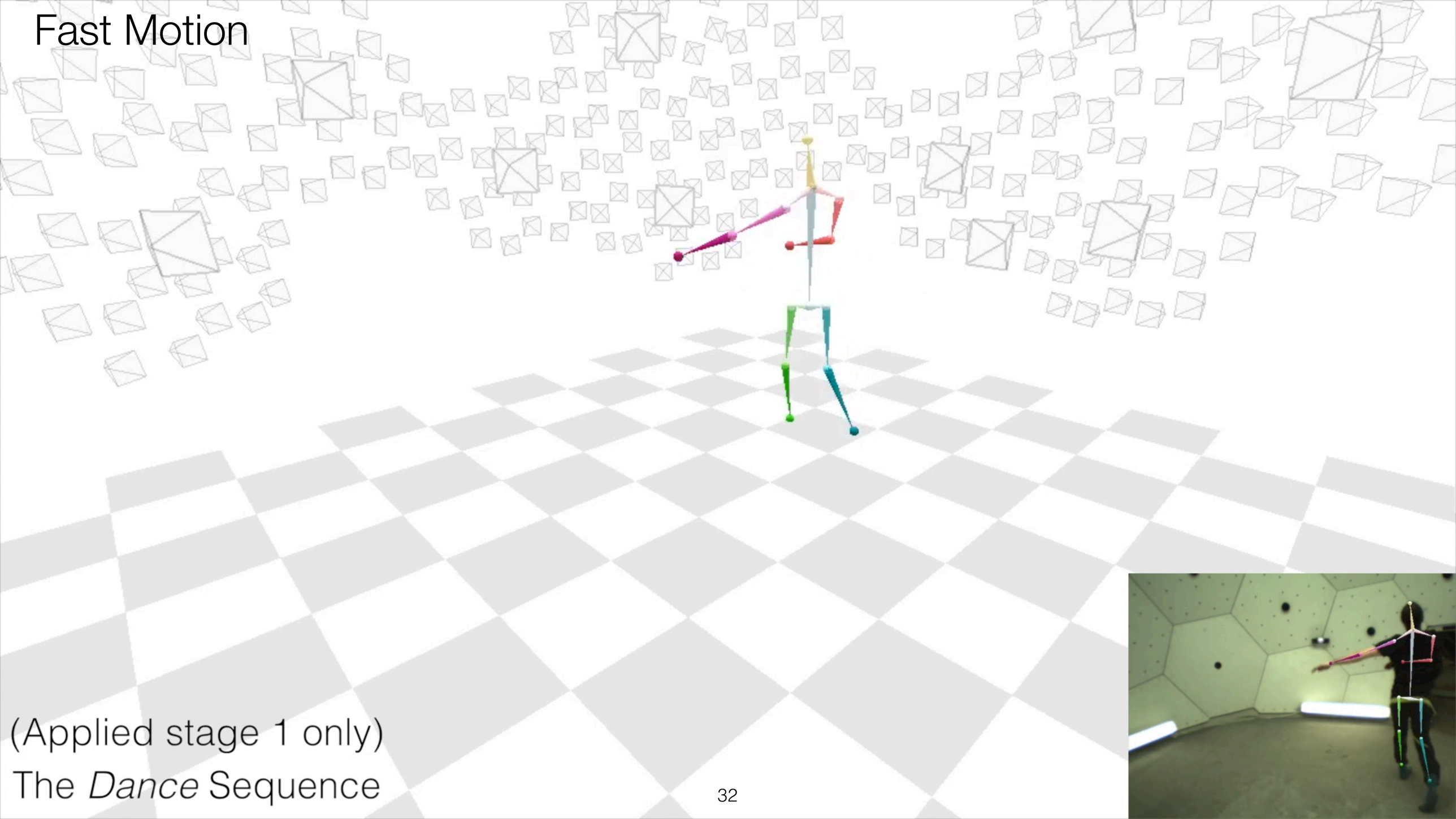
Different Size of People

The *Ian* Sequence

Severe Occlusions

The *151125 Bang* Sequence

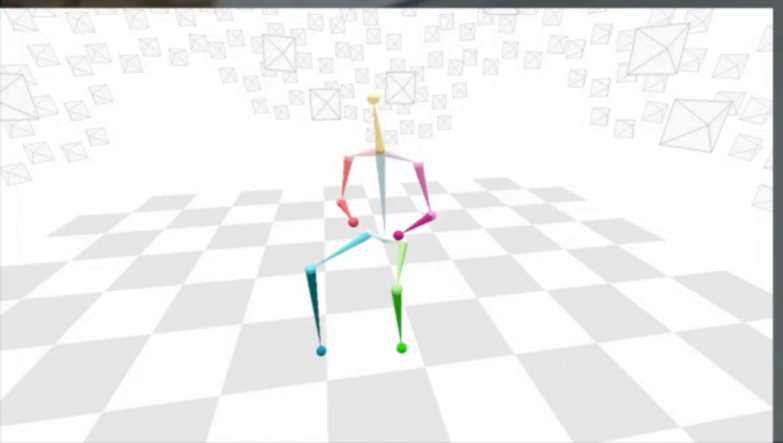
Fast Motion



(Applied stage 1 only)

The *Dance* Sequence

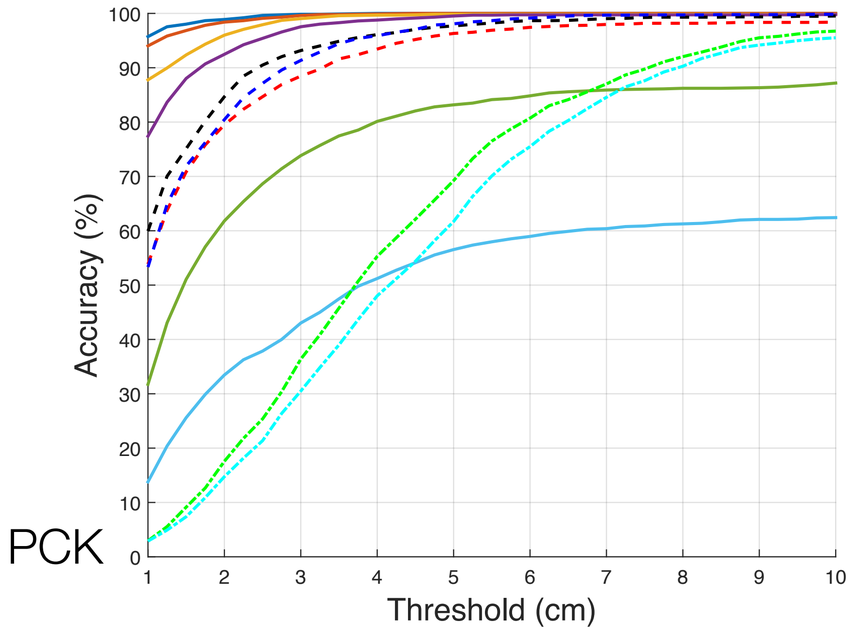
Occlusions by Other Objects



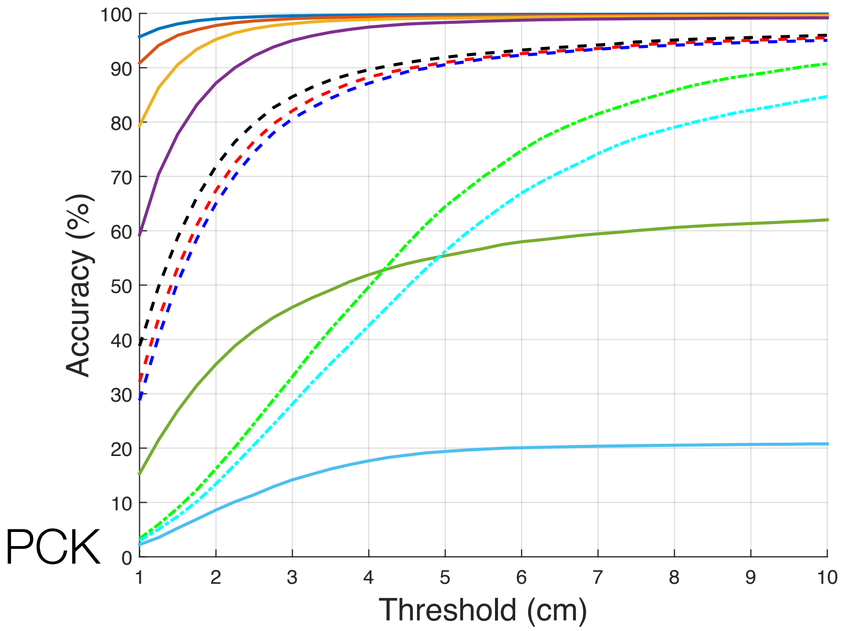
(Applied stage 1 only)
The *Drum* Sequence

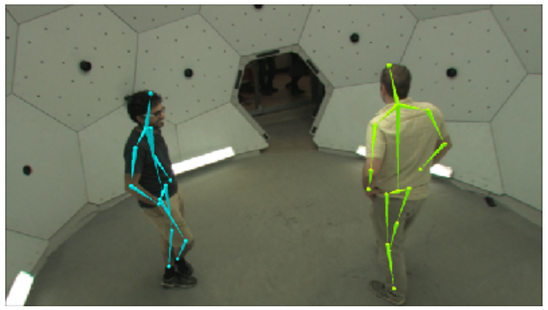
How Many Cameras Do We Need

Relation Between Scene Complexity and Number of Views

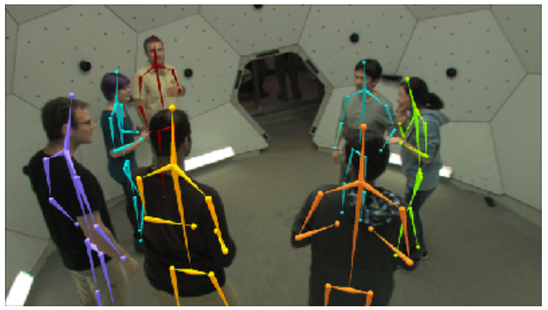


VGA 300
VGA 160
VGA 10
VGA 5





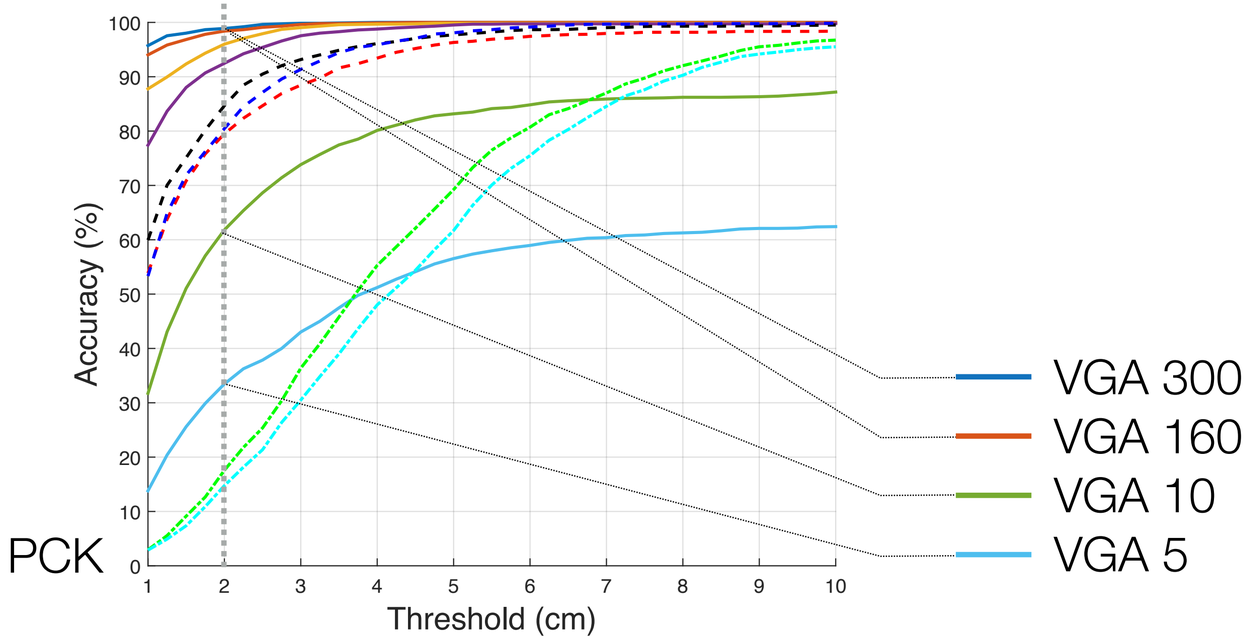
Two People

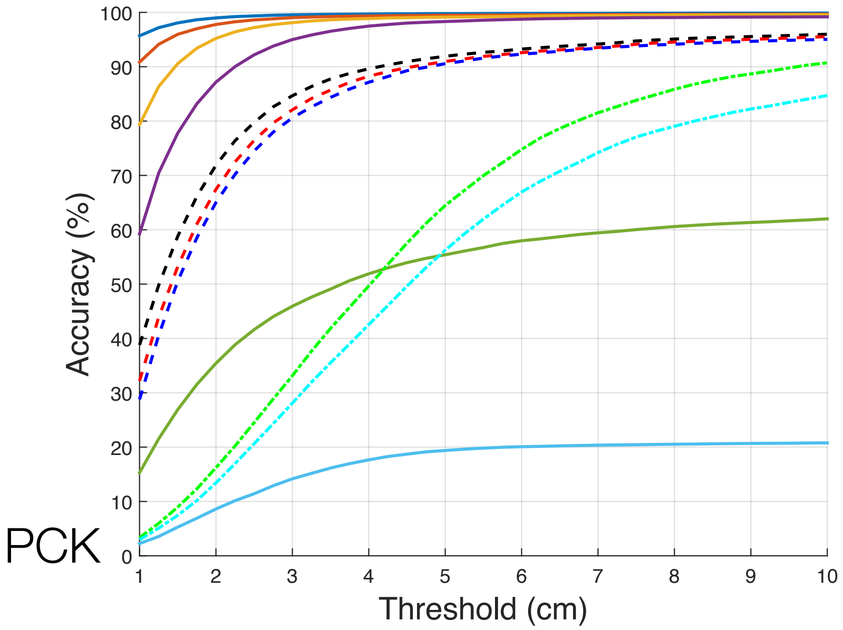


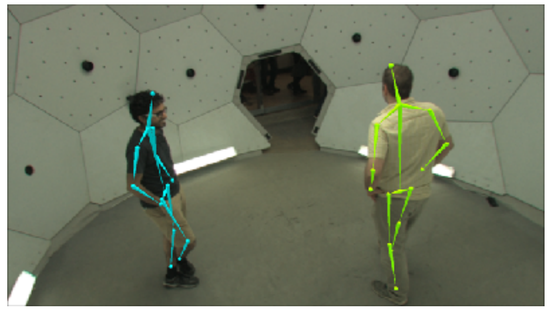
Seven People

How Many Cameras Do We Need

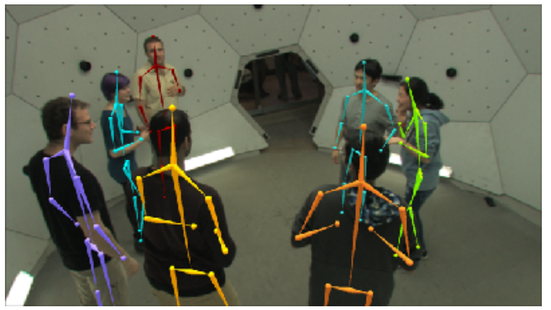
Relation Between Scene Complexity and Number of Views







Two People

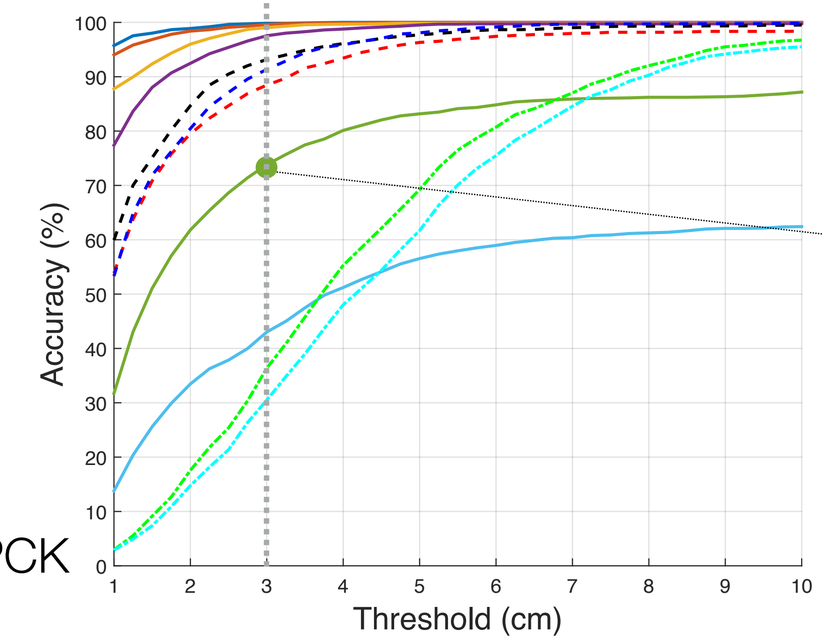


Seven People

How Many Cameras Do We Need

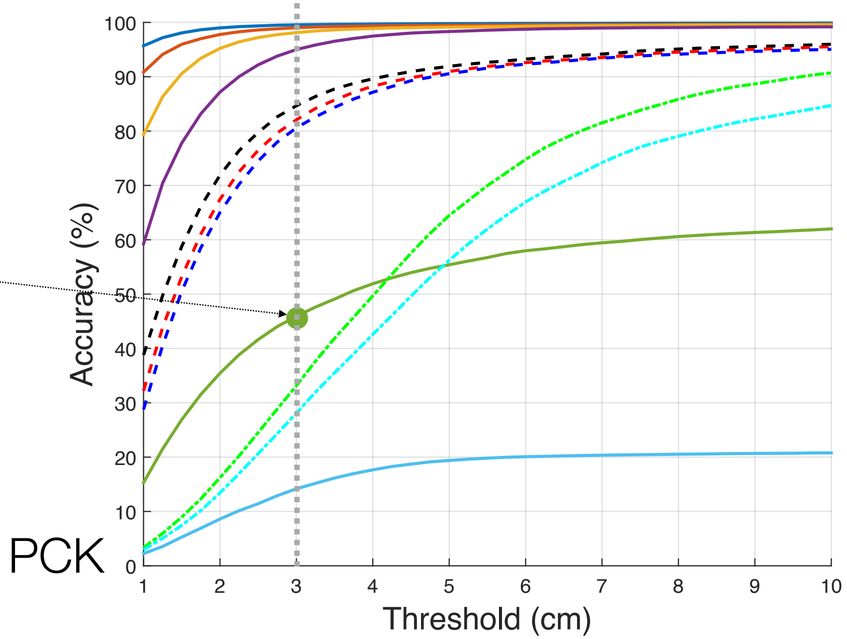
Relation Between Scene Complexity and Number of Views

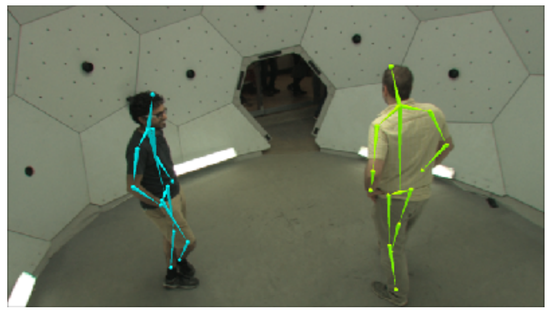
PCK



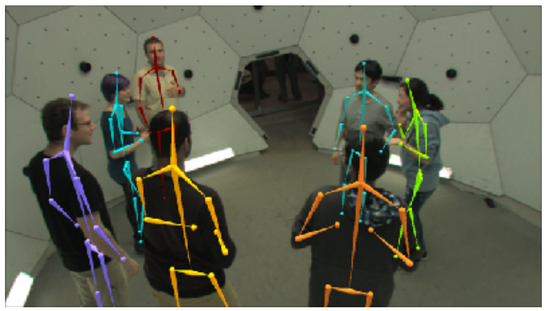
VGA 300
VGA 160
VGA 10
VGA 5

PCK





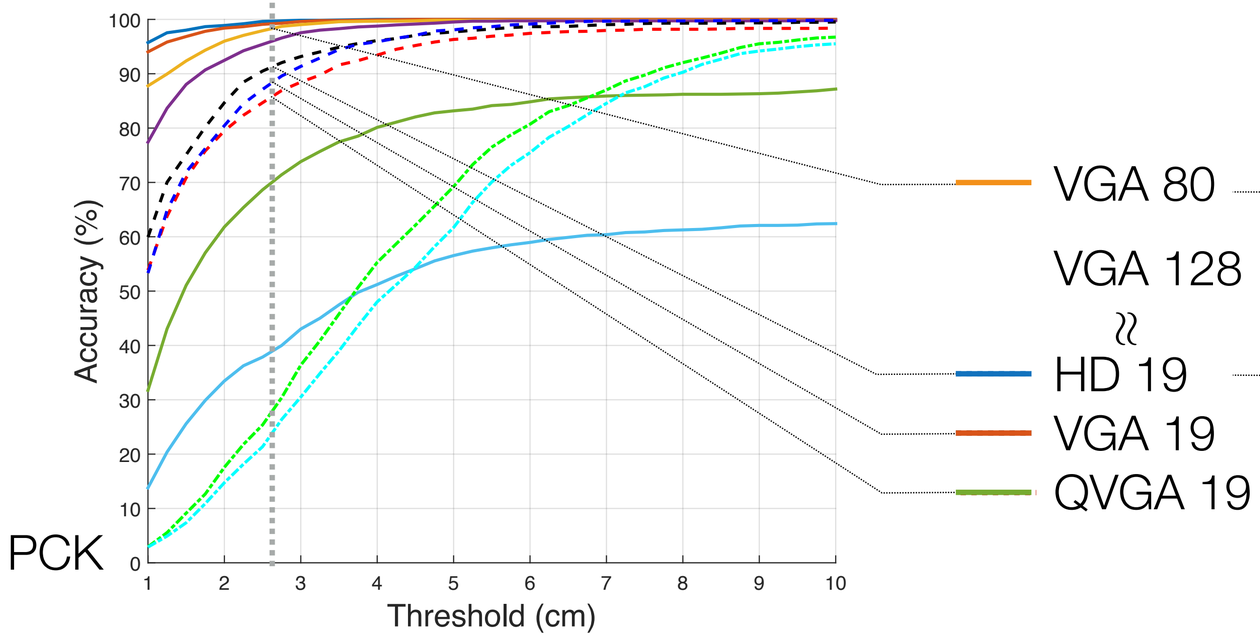
Two People



Seven People

How Many Cameras Do We Need

Relation Between Scene Complexity and Number of Views



VGA 80

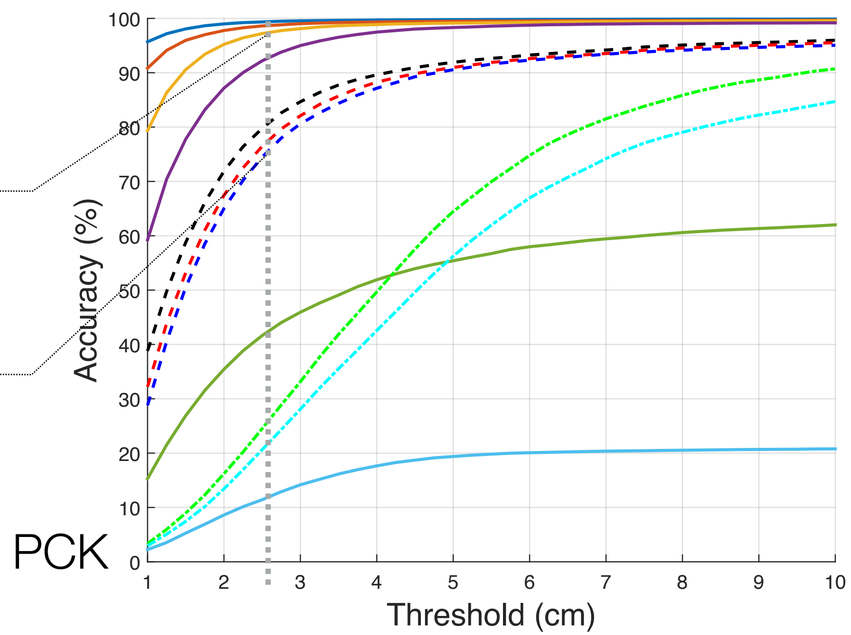
VGA 128

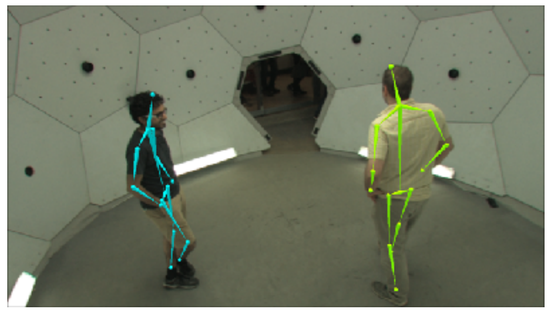
≀

HD 19

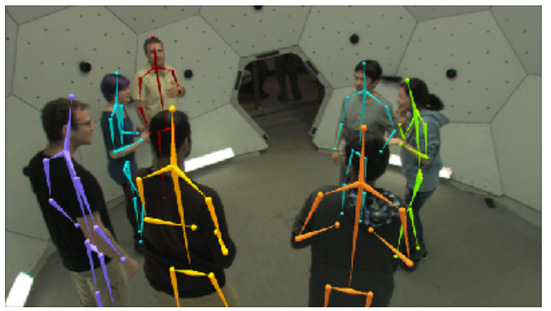
VGA 19

QVGA 19





Two People



Seven People

Are Body and Face Enough?

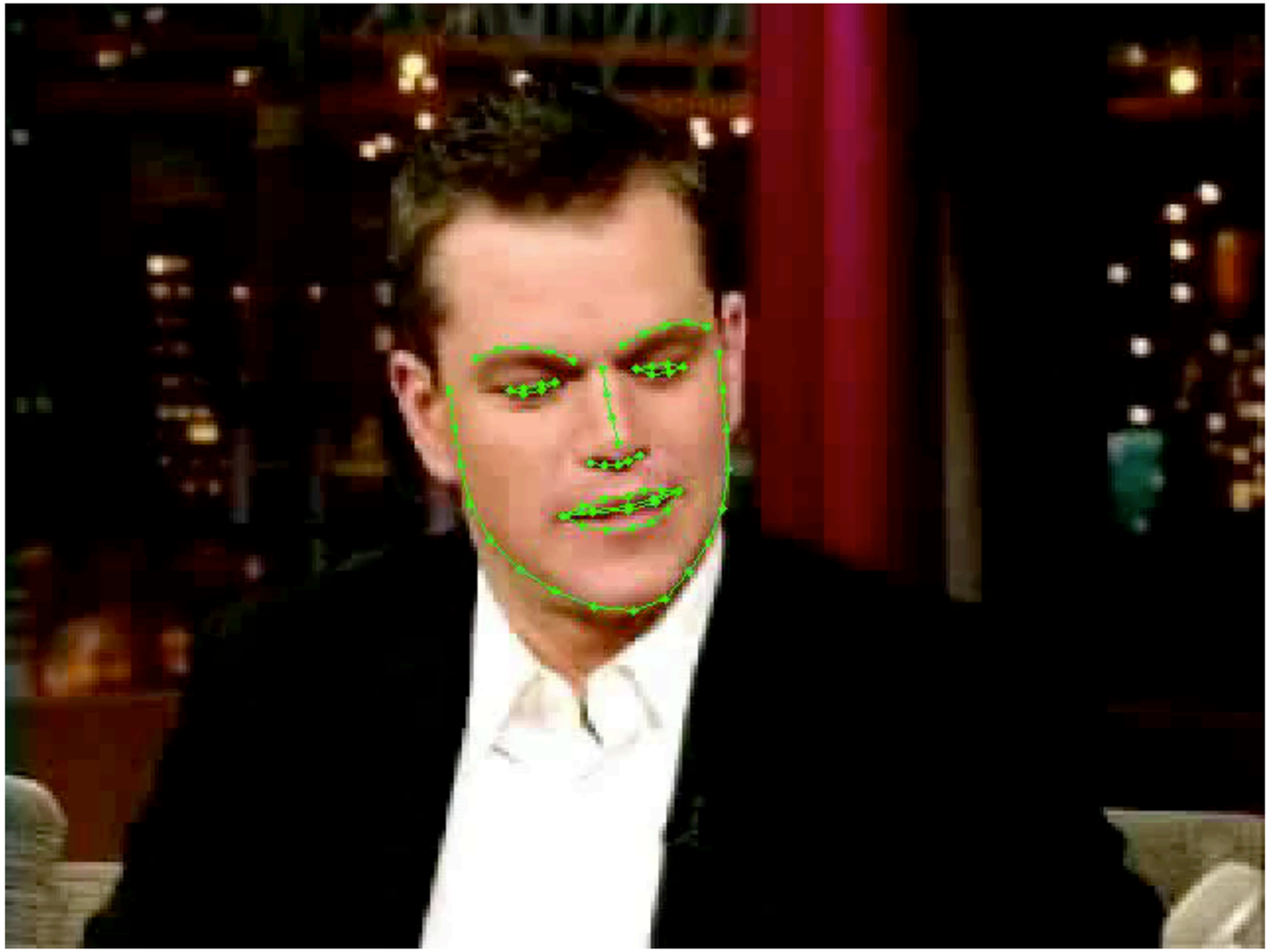
Important Nuances Are Embedded In Hands

Body Only

Face+Body

Face+Body+Hand

Face Keypoint Detectors Are Available



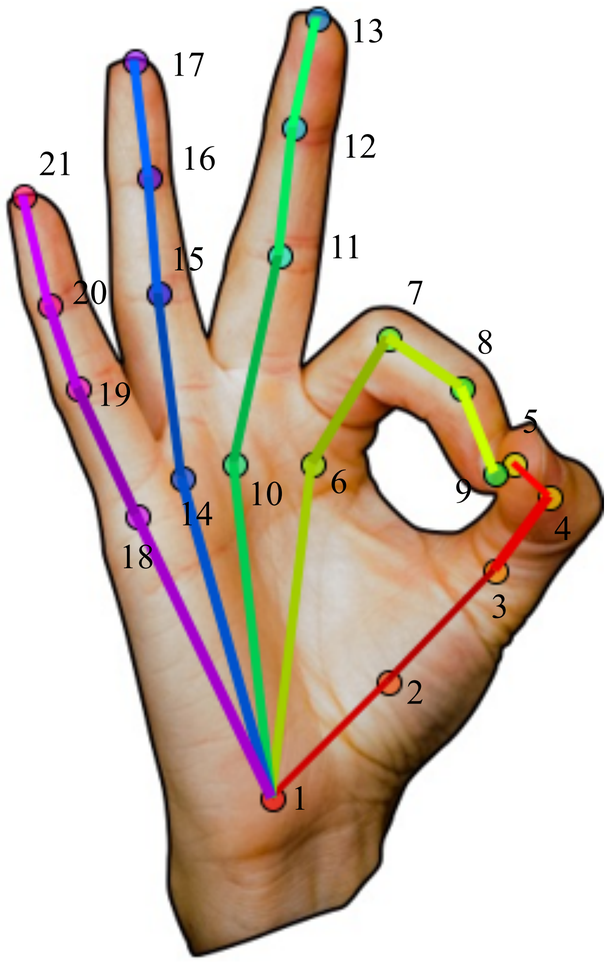
Xiong & De la Torre, *Supervised Descent Method and its Application to Face Alignment,* CVPR 2013

Body Keypoint Detectors Are Available

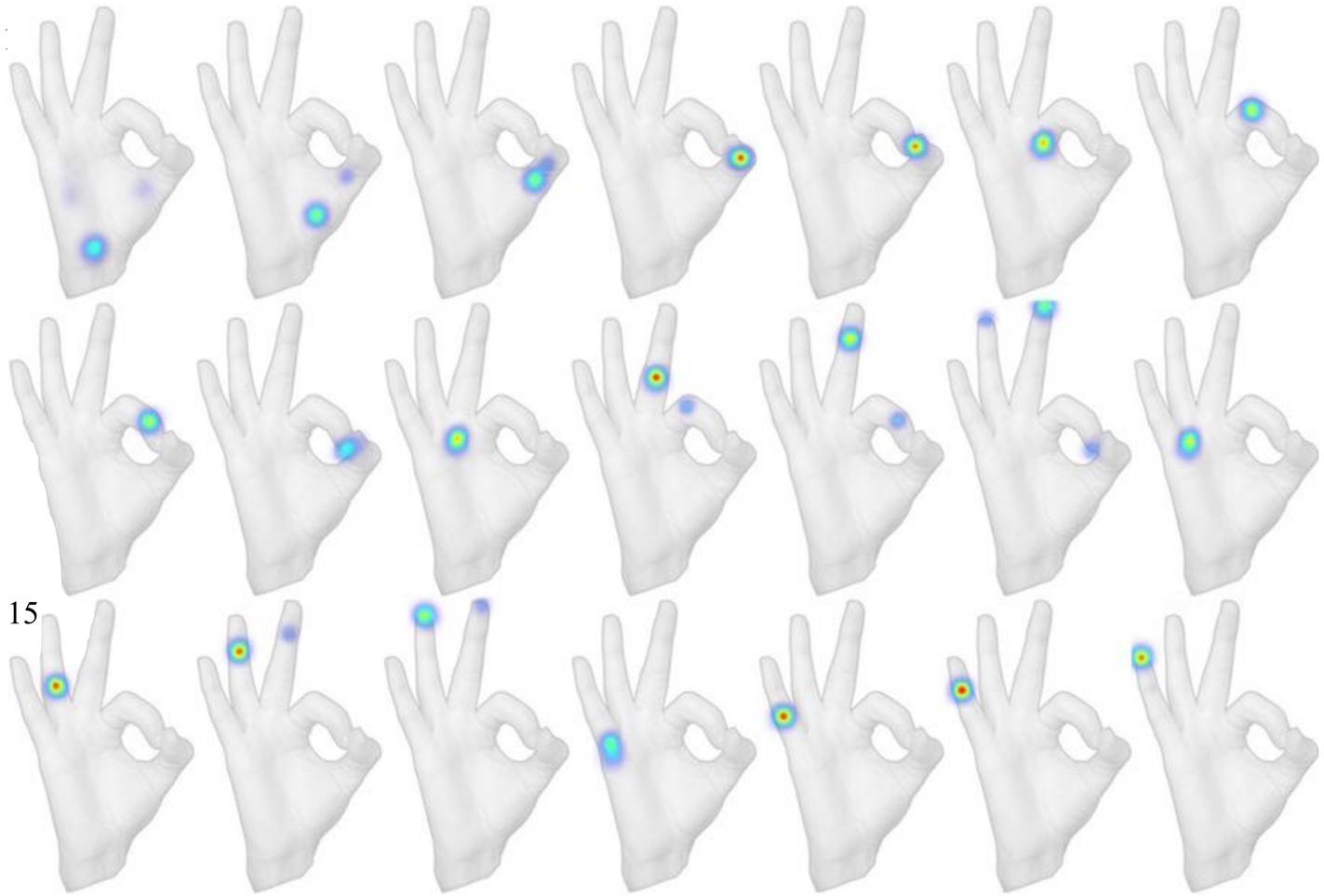
Source: https://www.youtube.com/watch?v=2DiQUX11YaY

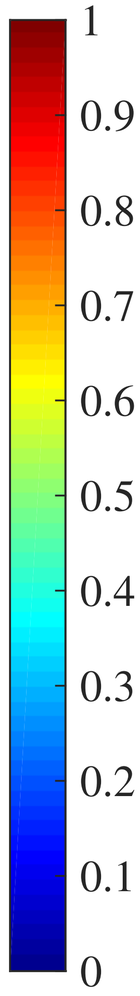
Zhe, Simon, Wei, Sheikh, CVPR 2017

How To Make A Good 2D Hand Pose Detector



Keypoints





Confidence maps

How To Make A Good 2D Hand Pose Detector

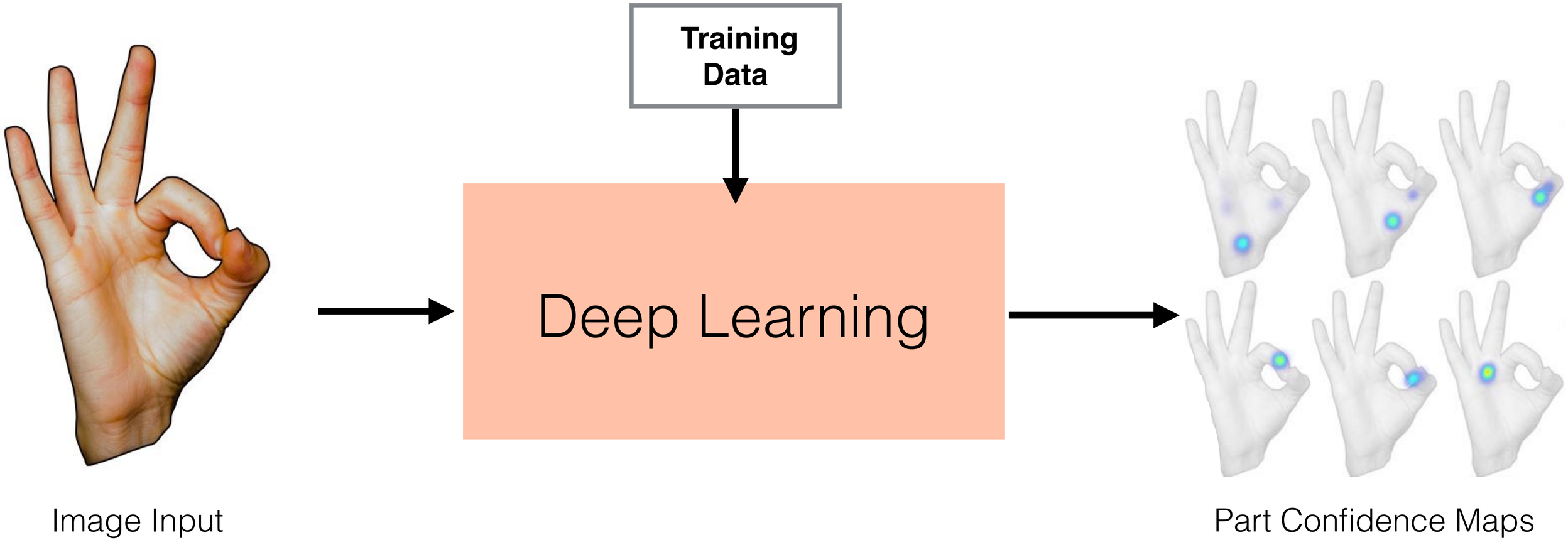
Image Input

Deep Learning

CNN du jour

Part Confidence Maps

How To Make A Good 2D Hand Pose Detector



How To Make A Good 2D Hand Pose Detector

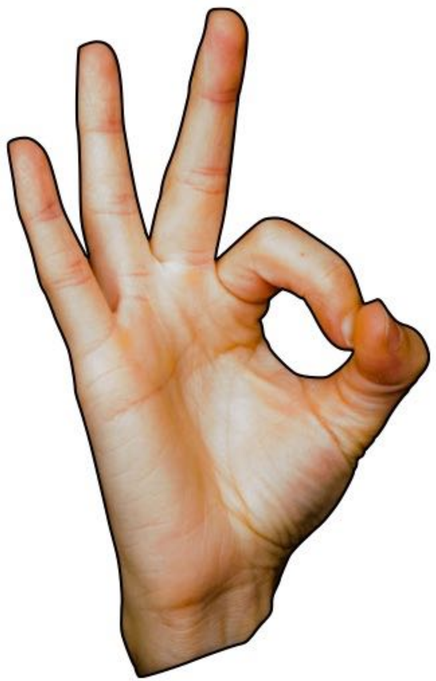


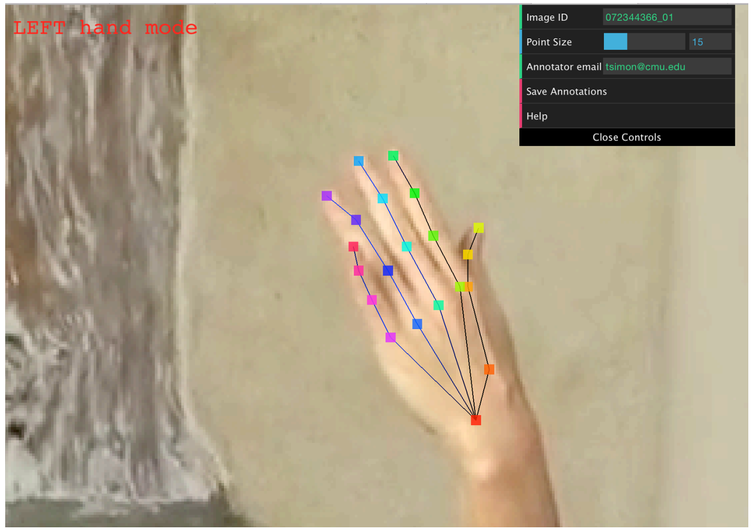
Image Input

Training Data
(Thousands of Labeled Images)

Deep Learning

Part Confidence Maps

ANNOTATORS NEEDED TO LABEL IMAGES



We are looking for people to help annotate landmarks in images and video. The ideal candidate should be consistent, self-motivated, and have great attention to detail. The position will be paid hourly at $12/hour, hours flexible.

- Work from home using any browser.

- ATTENTION TO DETAIL required.

- Proofreading and/or editing skills helpful

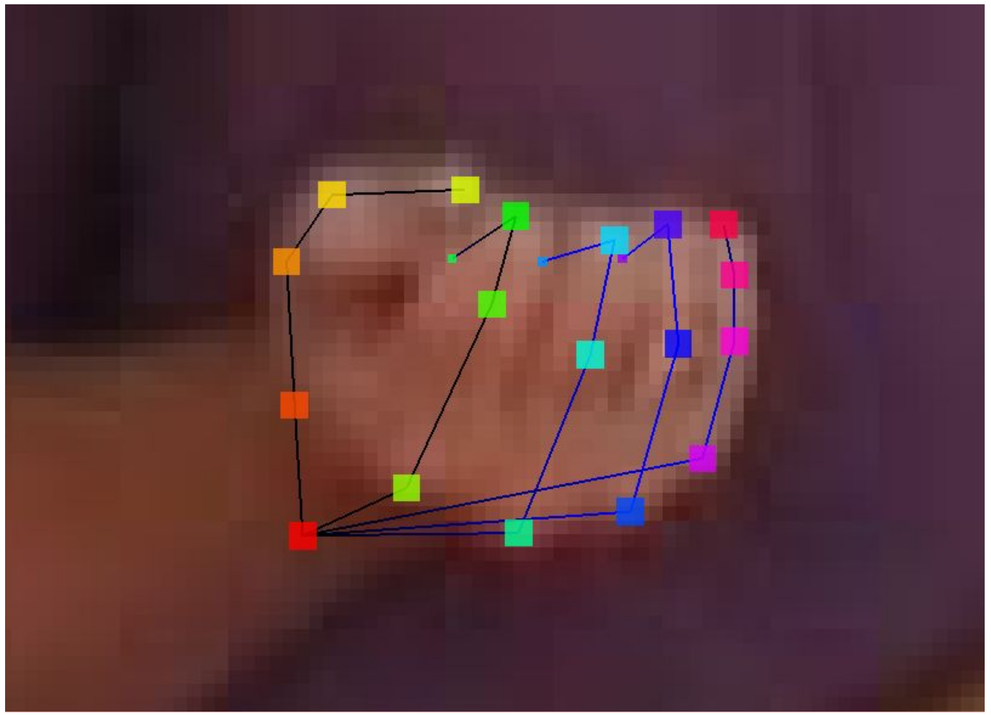
- Payment is up to $12 per hour

Contact: Tomas Simon (tsimon@cs.cmu.edu)

Tomas Simon (tsimon@cs.cmu.edu)
Tomas Simon (tsimon@cs.cmu.edu)
Tomas Simon (tsimon@cs.cmu.edu)
Tomas Simon (tsimon@cs.cmu.edu)
Tomas Simon (tsimon@cs.cmu.edu)
Tomas Simon (tsimon@cs.cmu.edu)
Tomas Simon (tsimon@cs.cmu.edu)
Tomas Simon (tsimon@cs.cmu.edu)
Tomas Simon (tsimon@cs.cmu.edu)
Tomas Simon (tsimon@cs.cmu.edu)
Tomas Simon (tsimon@cs.cmu.edu)
Tomas Simon (tsimon@cs.cmu.edu)

How To Make A Good 2D Hand Pose Detector

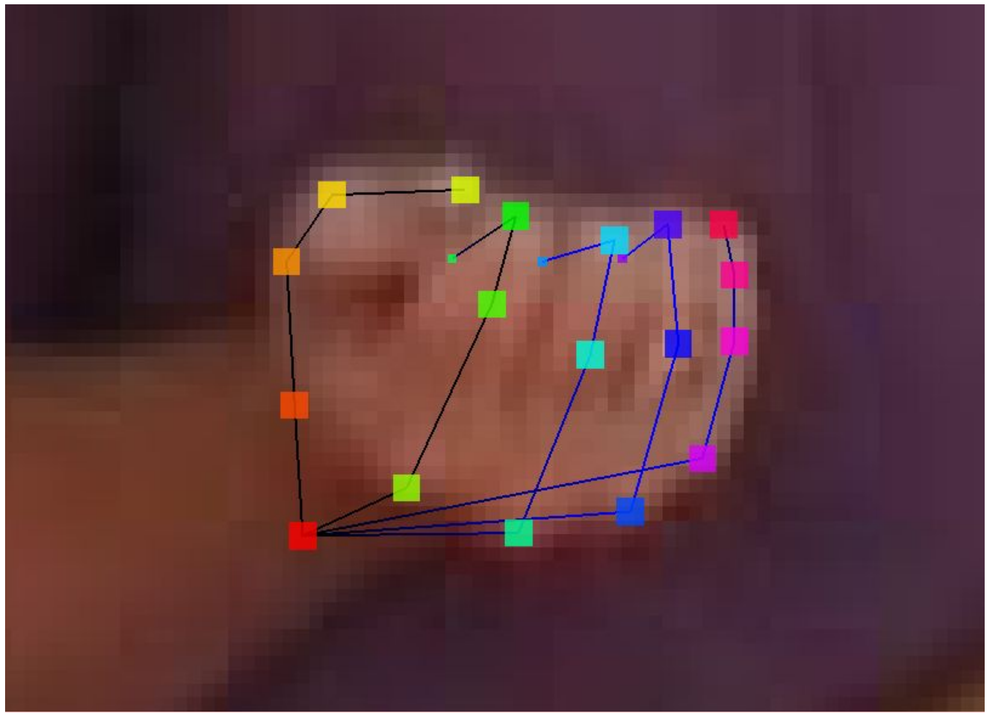
Difficulties in Labeling Hand Joints

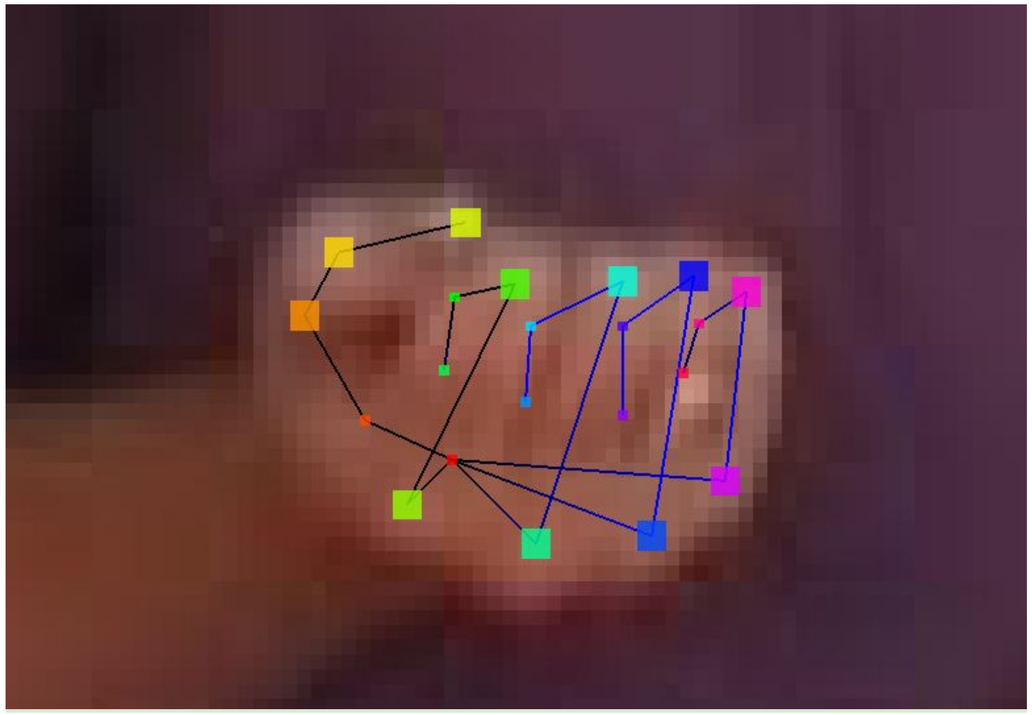


Occluded Joints are Guessed

How To Make A Good 2D Hand Pose Detector

Difficulties in Labeling Hand Joints

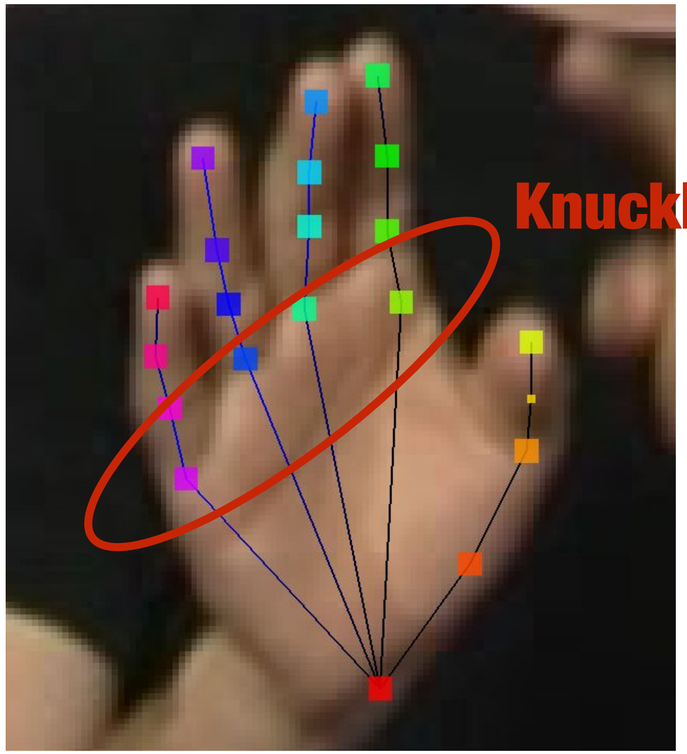


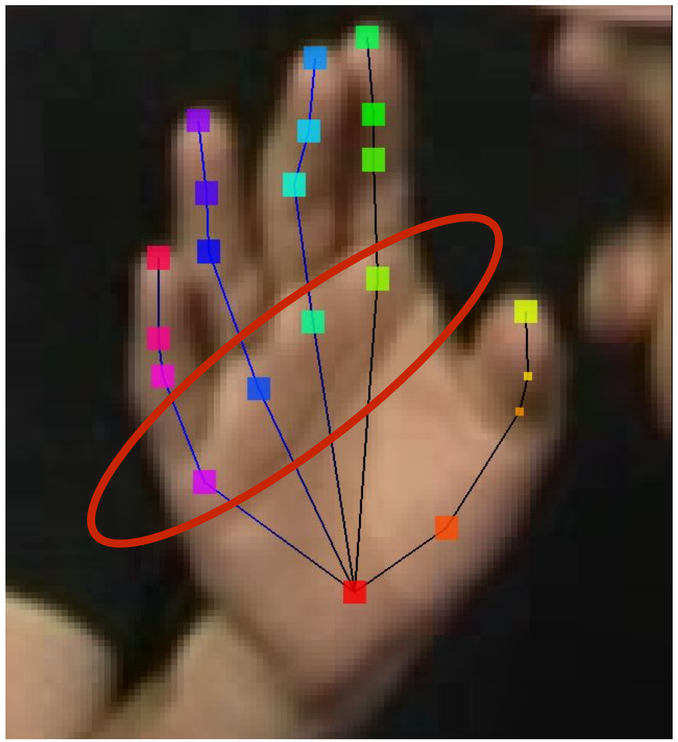


Occluded Joints are Guessed

How To Make A Good 2D Hand Pose Detector

Difficulties in Labeling Hand Joints





Internal Joints are
Hard to Localize

How To Make A Good 2D Hand Pose Detector

Synthetic Data != Real Data

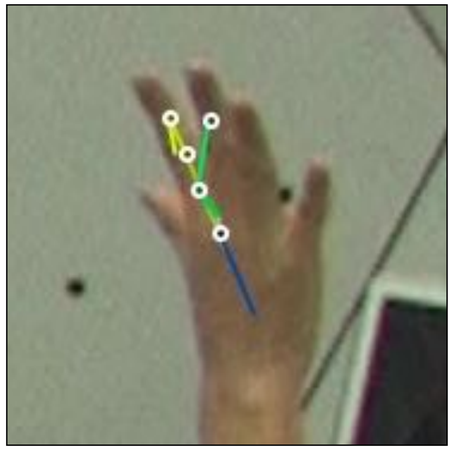
Unreal Engine + Mixamo

How To Make A Good 2D Hand Pose Detector

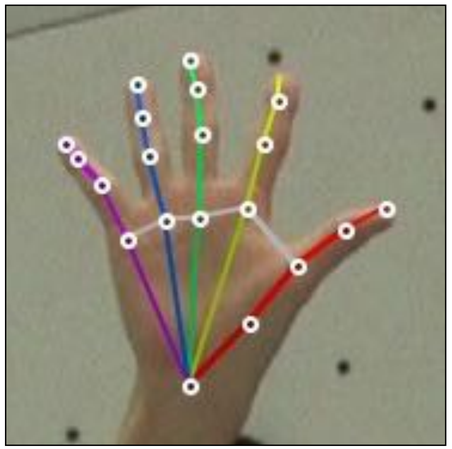
Synthetic Data != Real Data

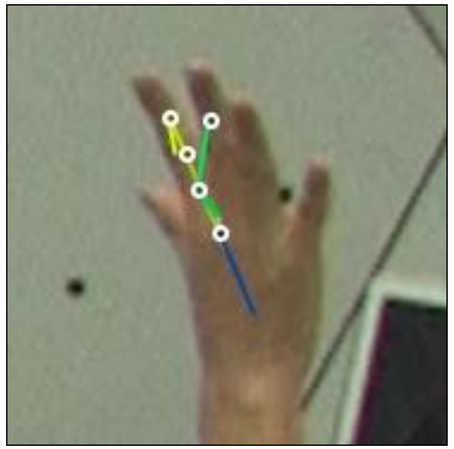
Raytraced Random Poses

Detector Trained Only on Synthetic Data

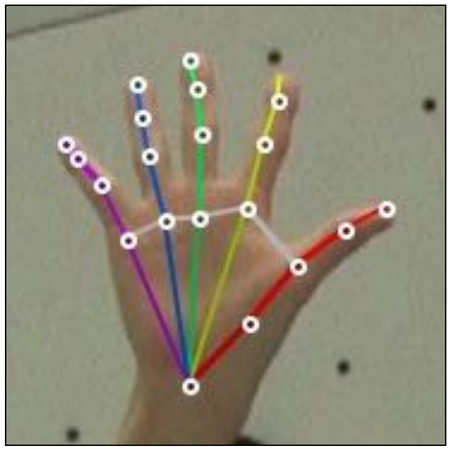


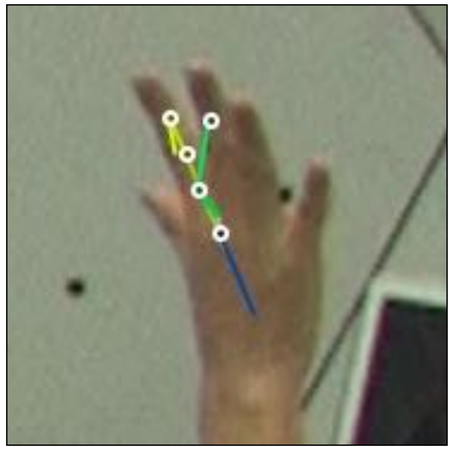
Detector Trained Only on Synthetic Data





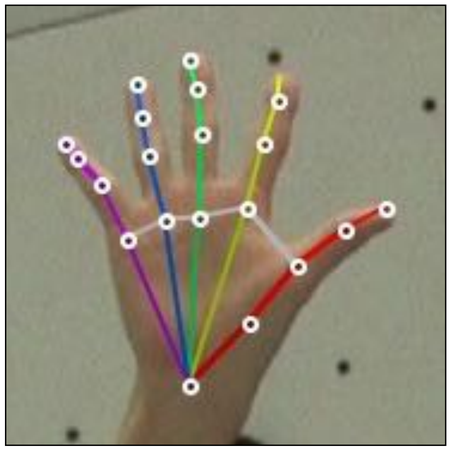
Detector Trained Only on Synthetic Data

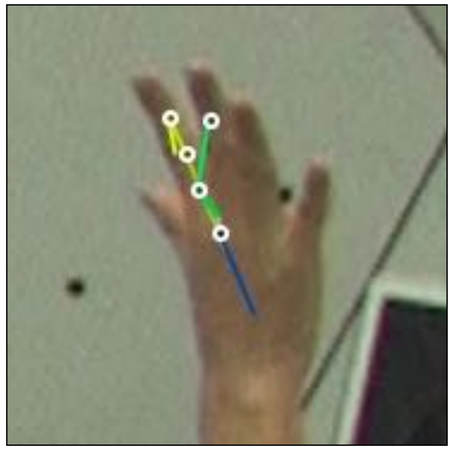




Detector Trained Only on Synthetic Data

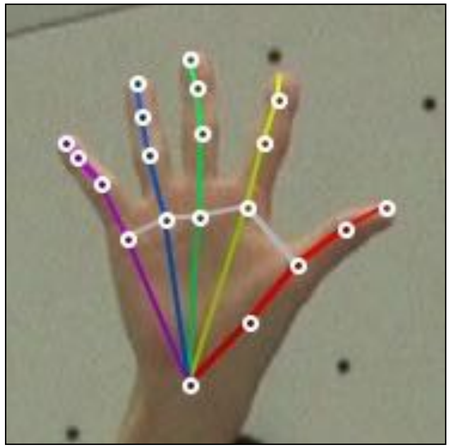
With Enough Cameras: At Least Two Good Views of Each Keypoint

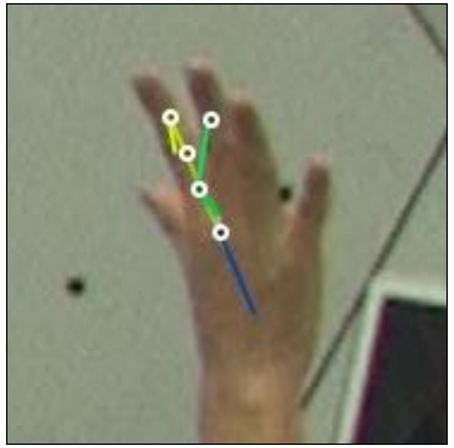




Detector Trained Only on Synthetic Data

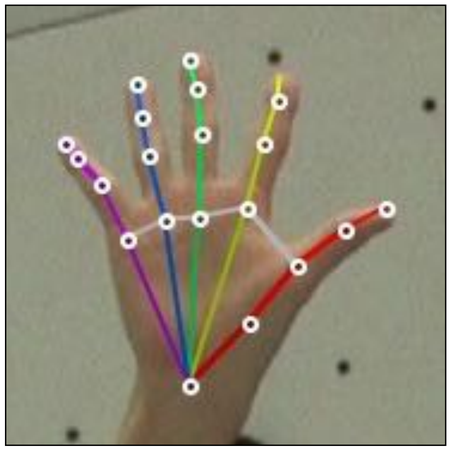
With Enough Cameras: At Least Two Good Views of Each Keypoint

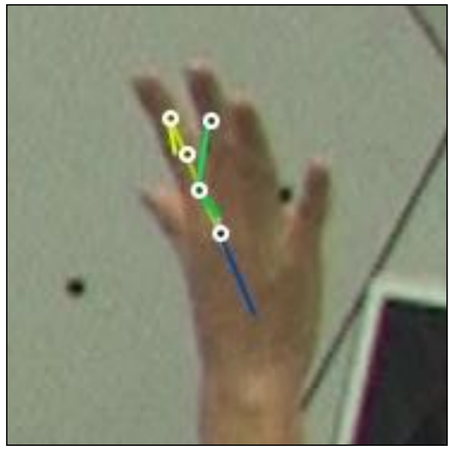




Detector Trained Only on Synthetic Data

With Enough Cameras: At Least Two Good Views of Each Keypoint



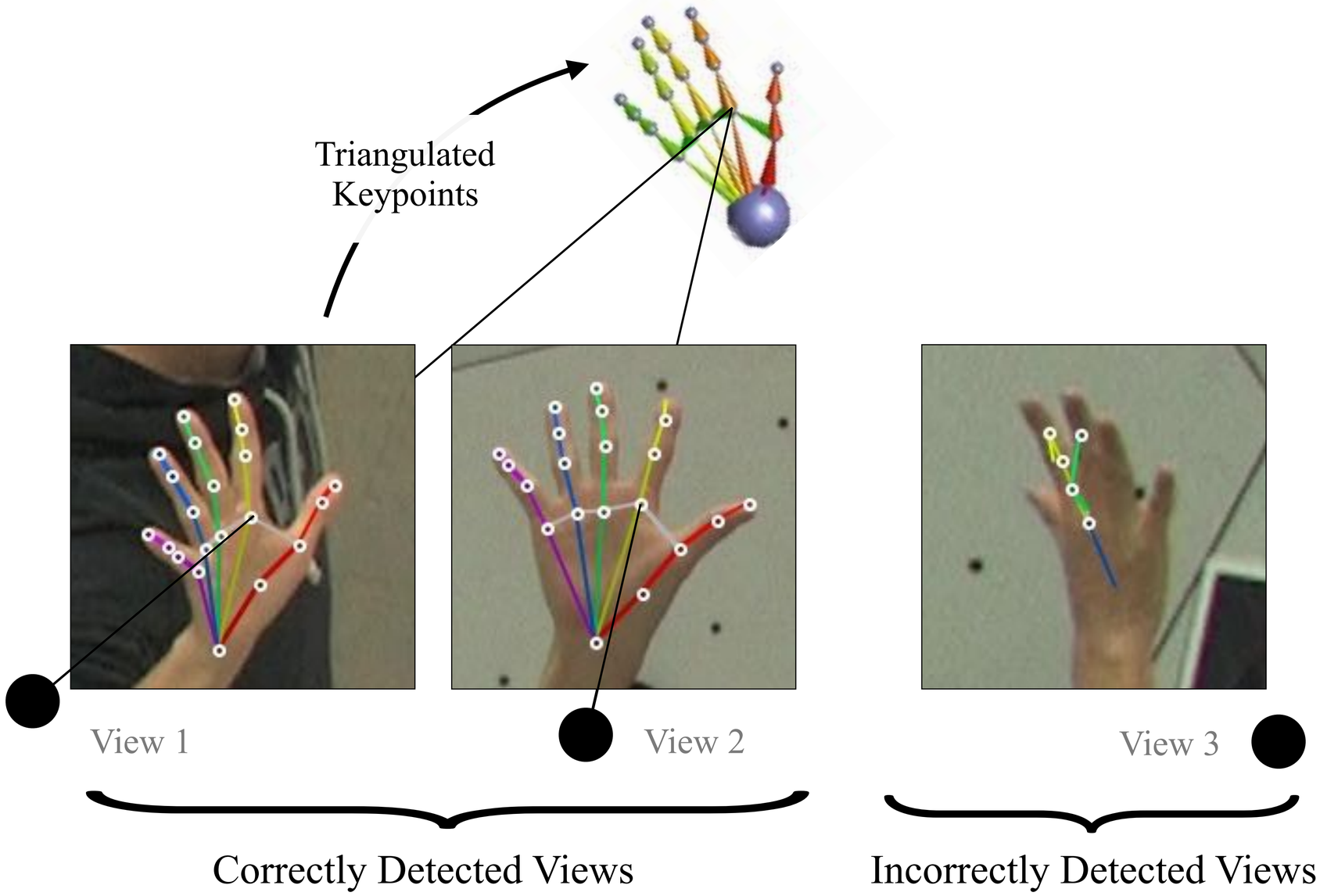


Correctly Detected Views

Incorrectly Detected Views

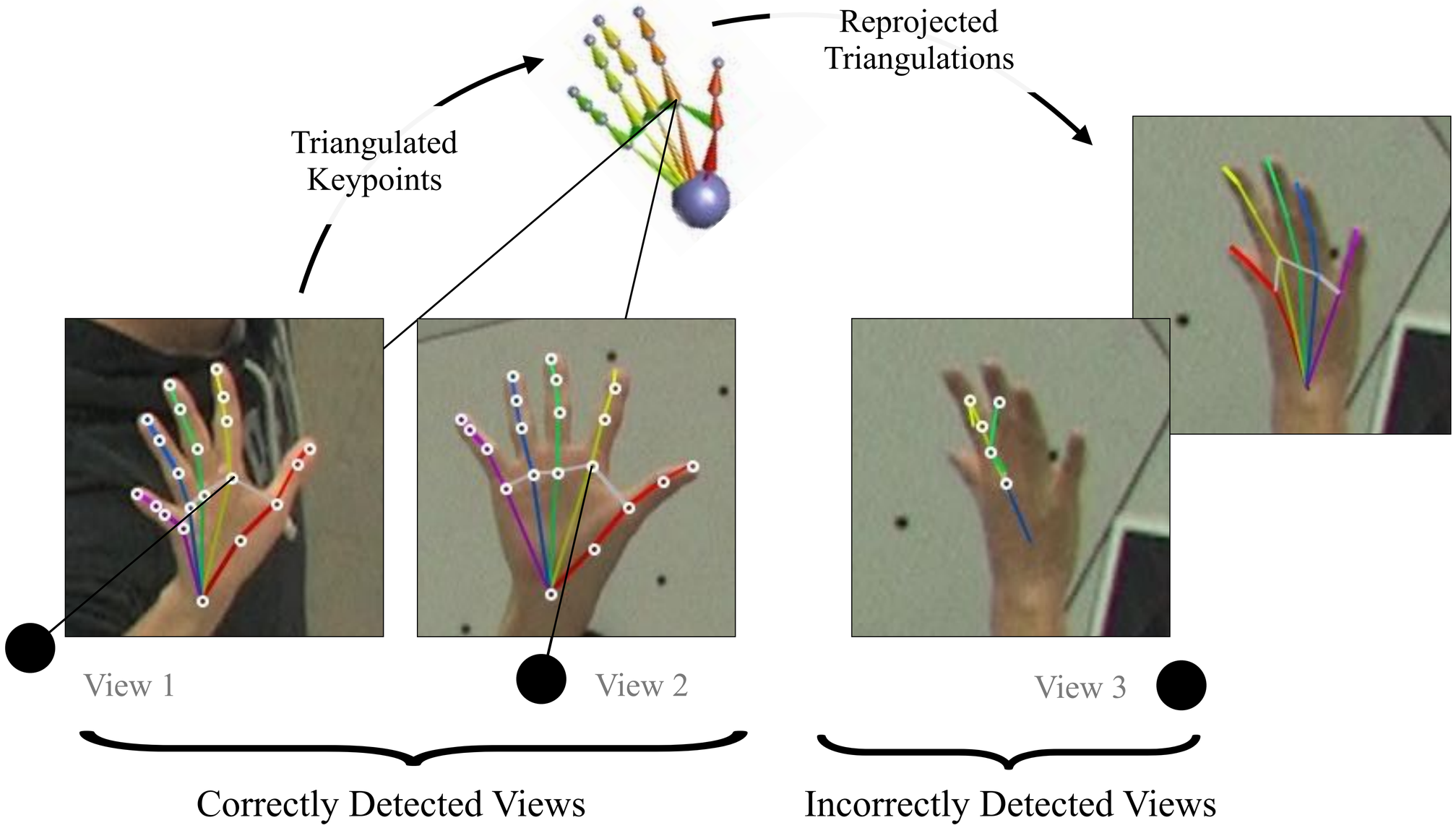
Multiview Bootstrapping

Triangulation as Supervision



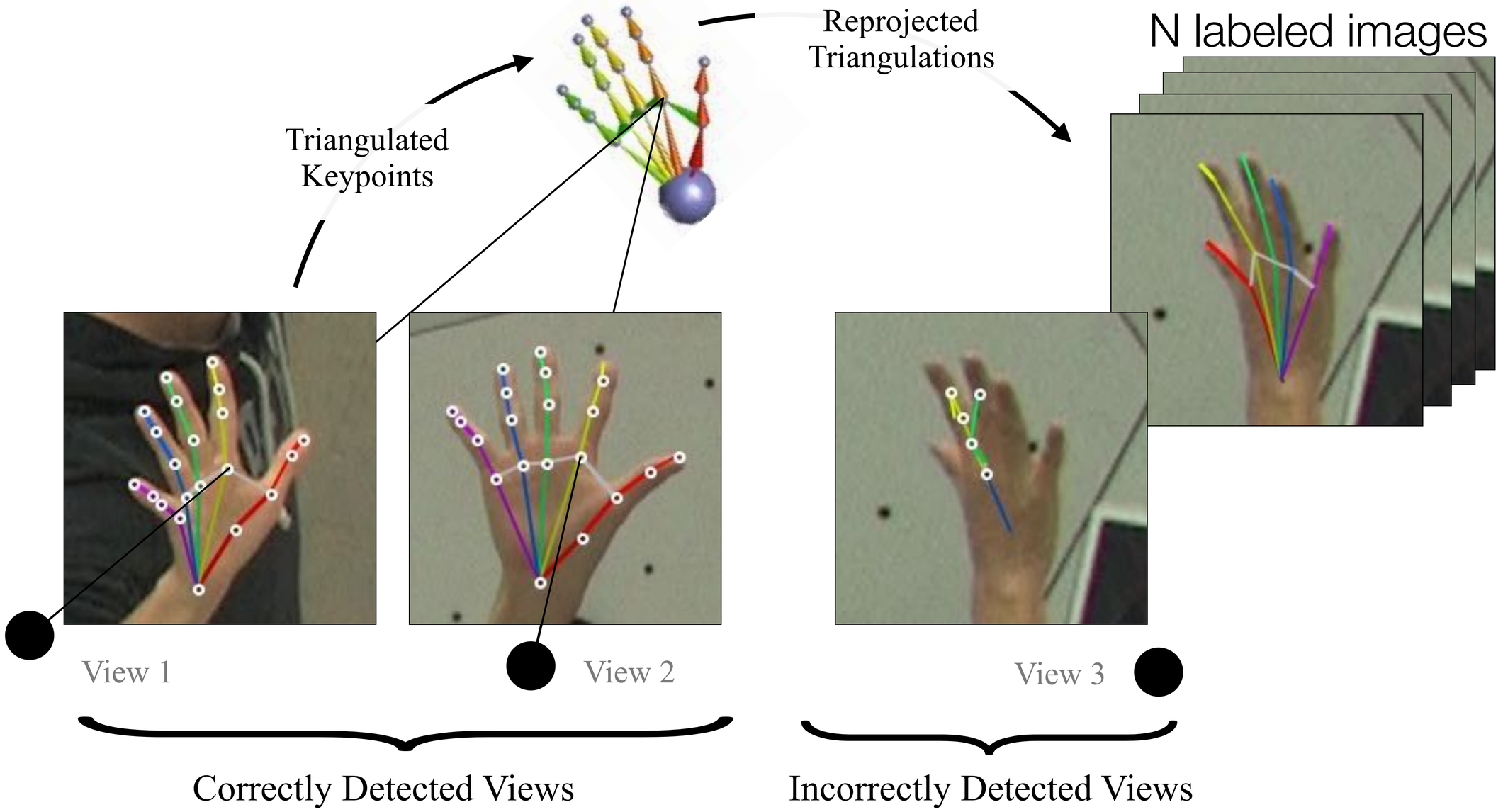
Multiview Bootstrapping

Triangulation as Supervision



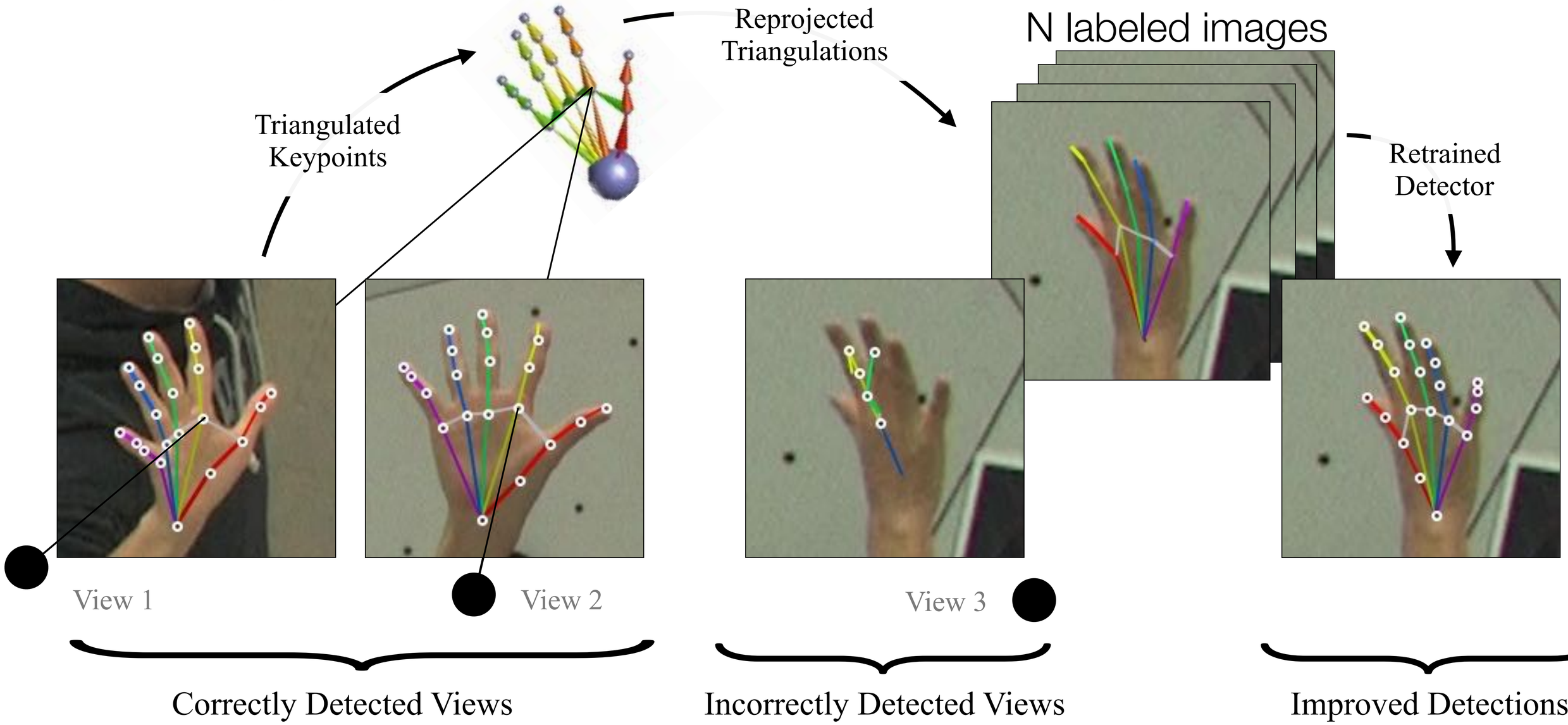
Multiview Bootstrapping

Using a Multiview System as an Annotation Machine



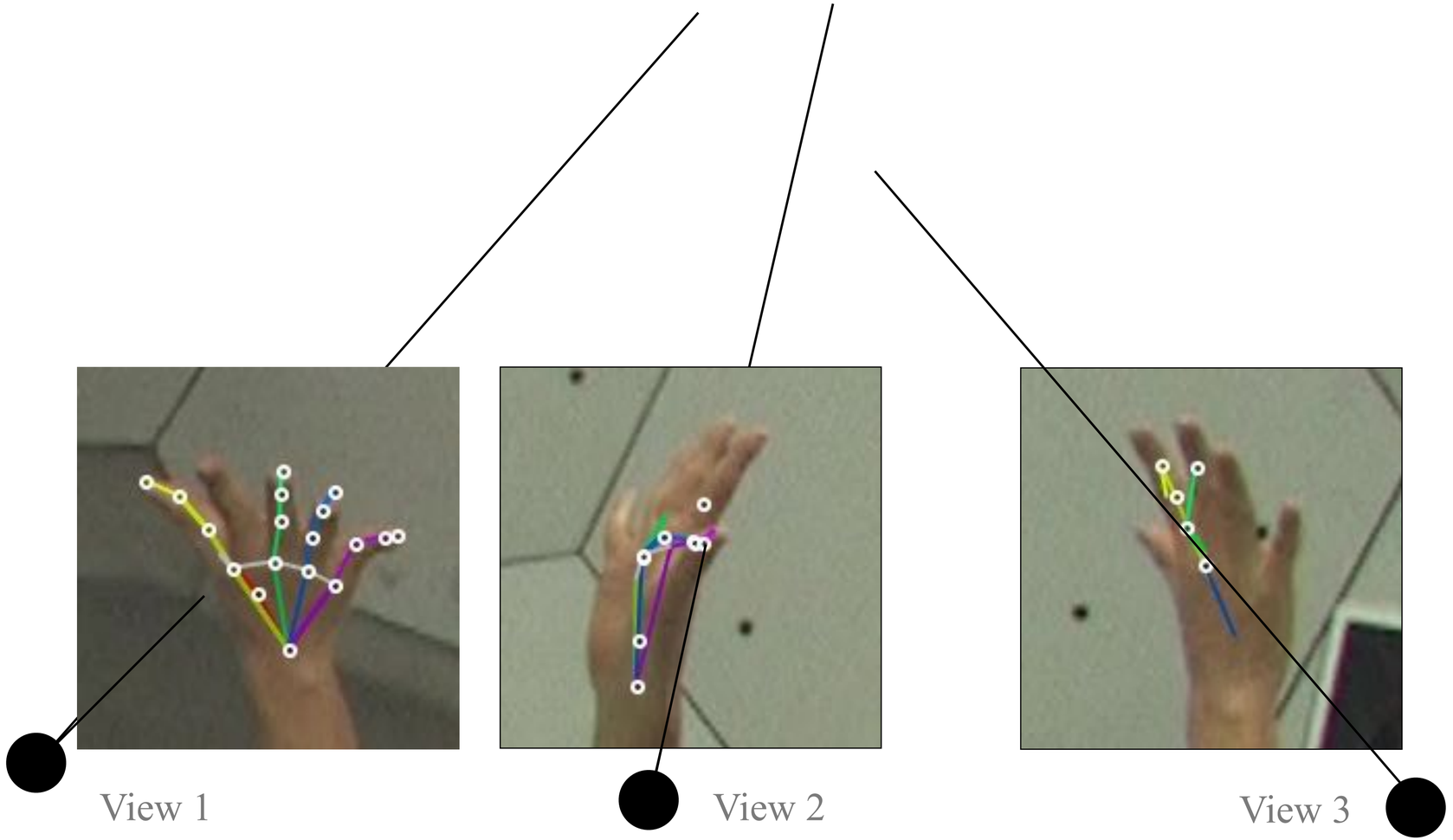
Multiview Bootstrapping

Using a Multiview System as an Annotation Machine



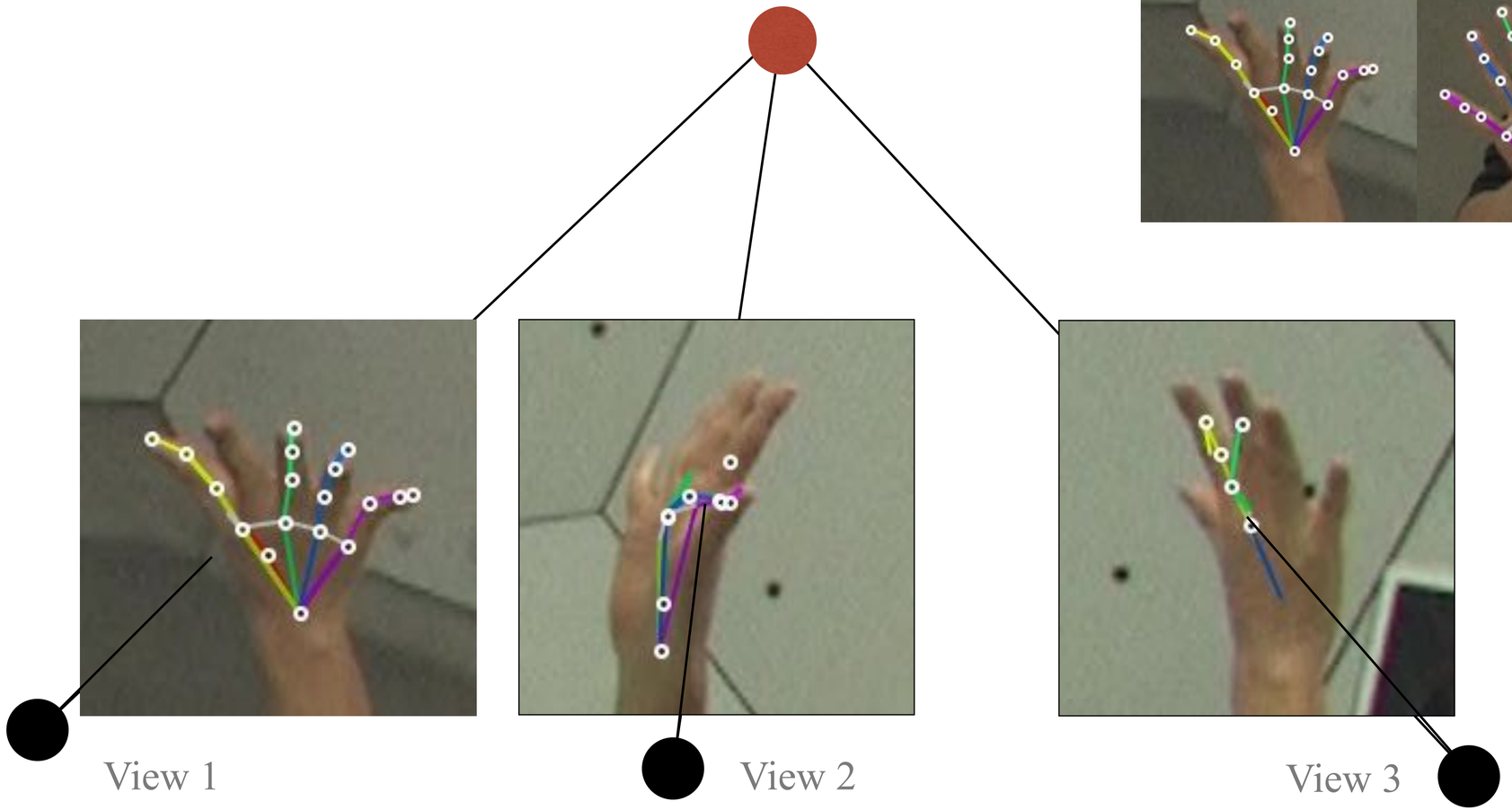
Multiview Bootstrapping

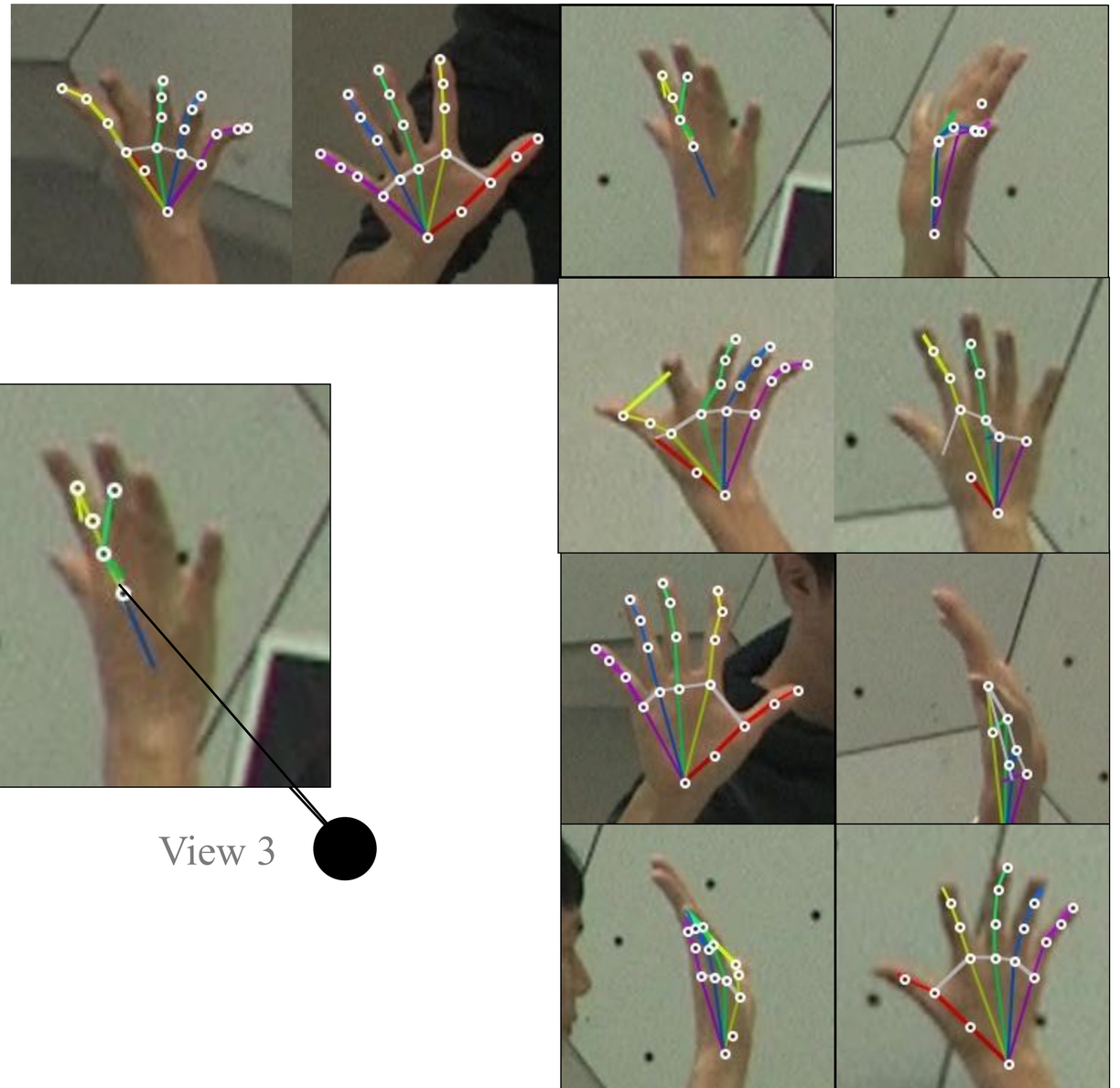
Inconsistent Detections Do Not Triangulate



Multiview Bootstrapping

False Positive When Random Triangulation

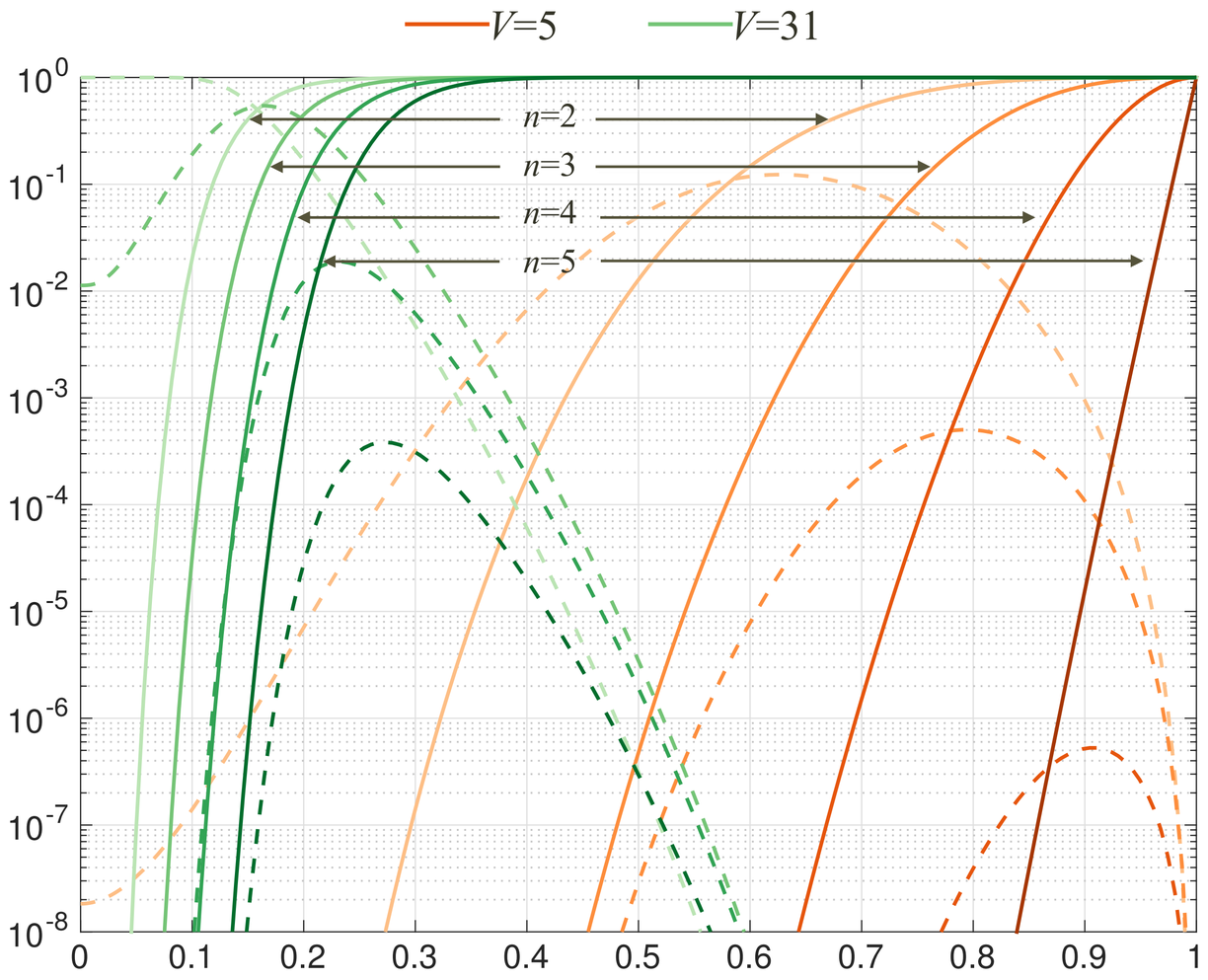




Multiview Bootstrapping

Triangulation As Supervision

True and False Positive Rates



PCK_σ

1. Initial Detector

- Trained only using rendered examples


View 1


View 1


View 2

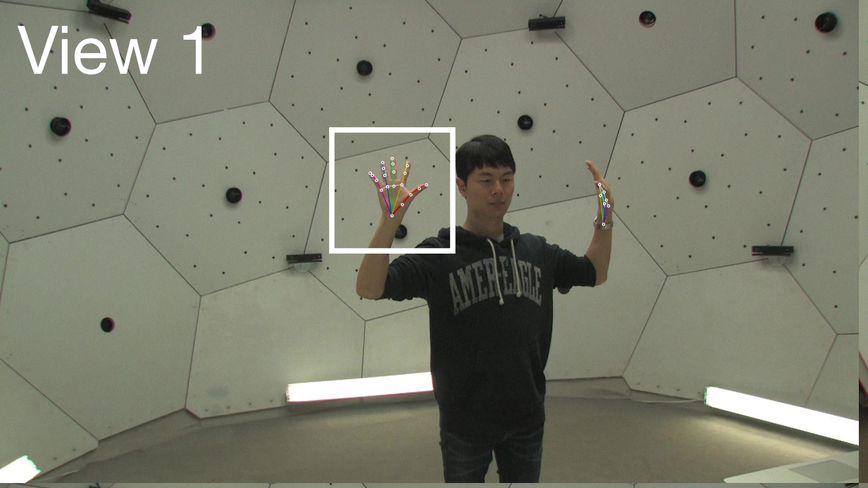

View 2

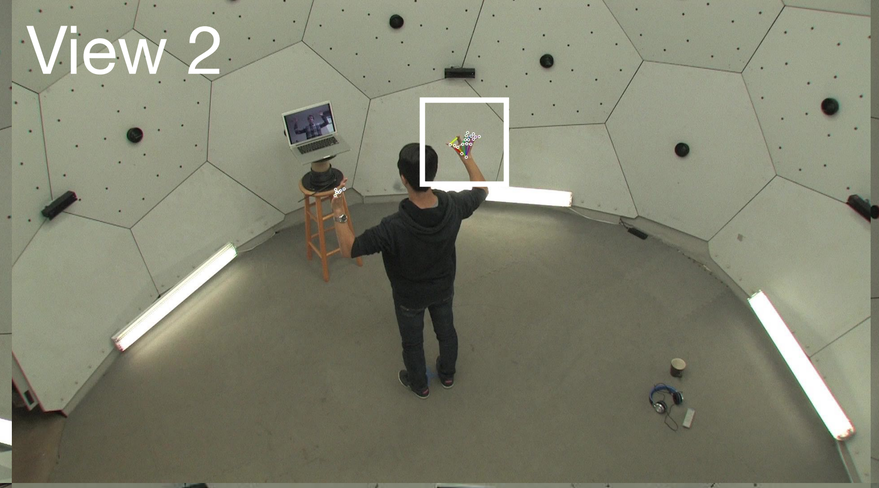

View 3


View 3

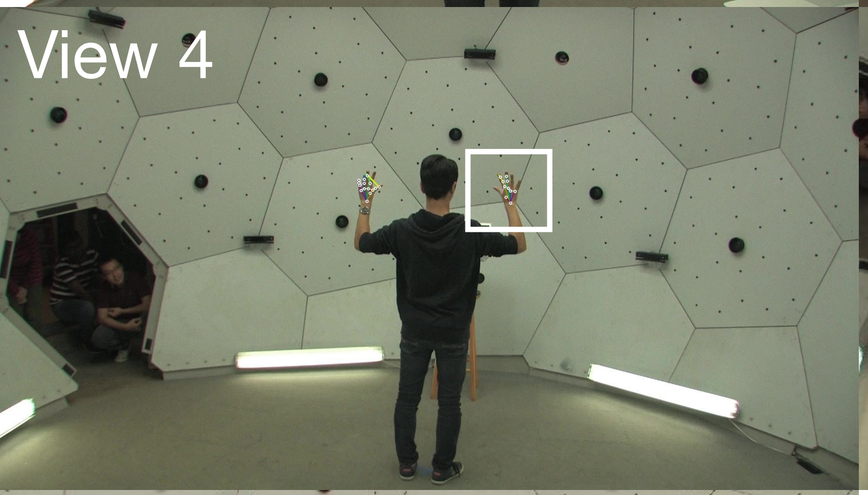

View 4

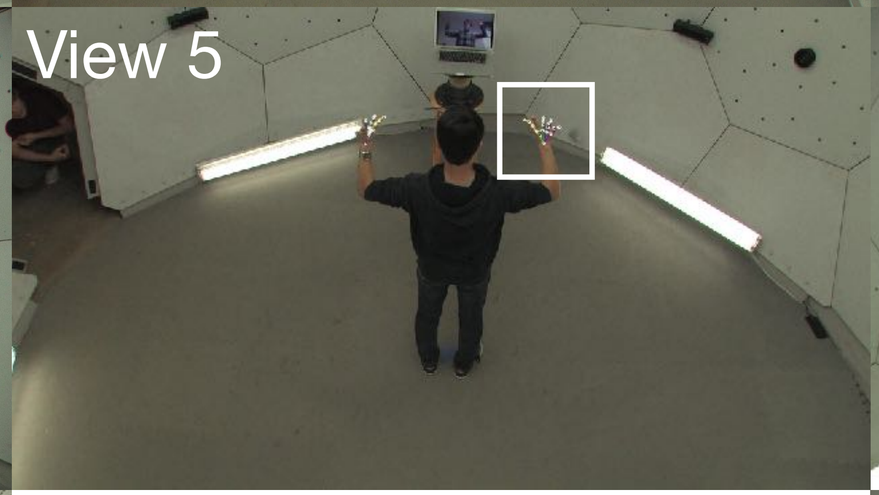

View 4

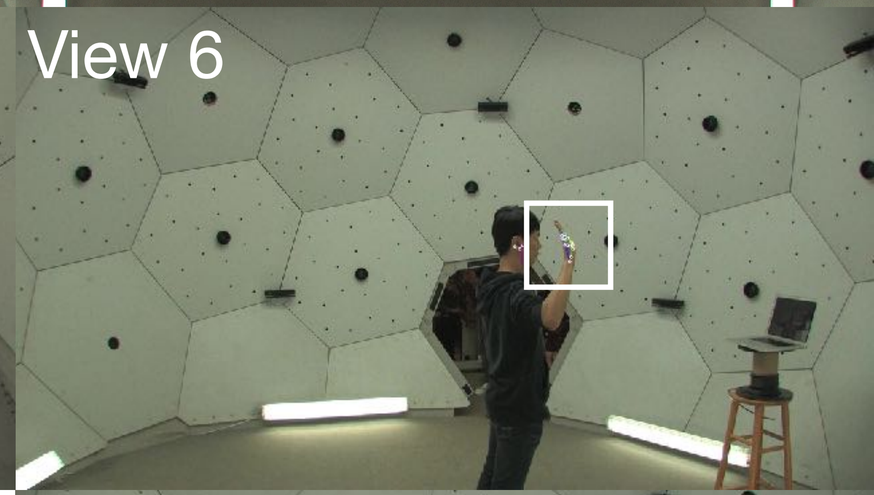

View 1


View 2


View 3

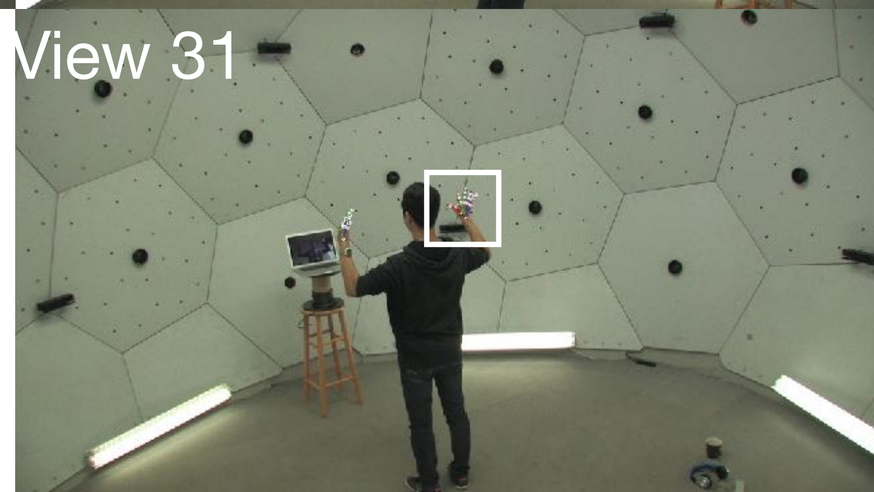

View 4


View 5

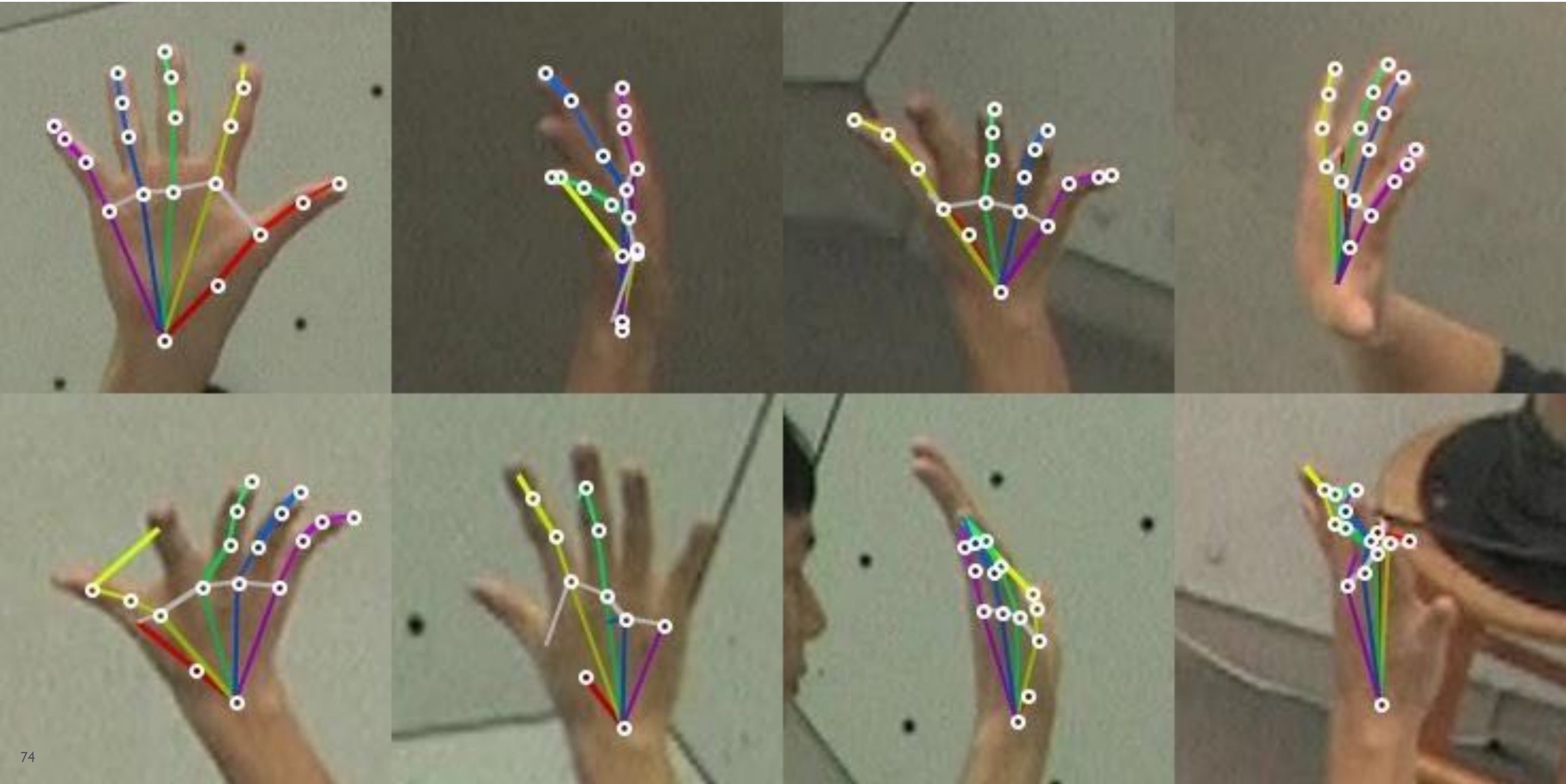

View 6


View 7

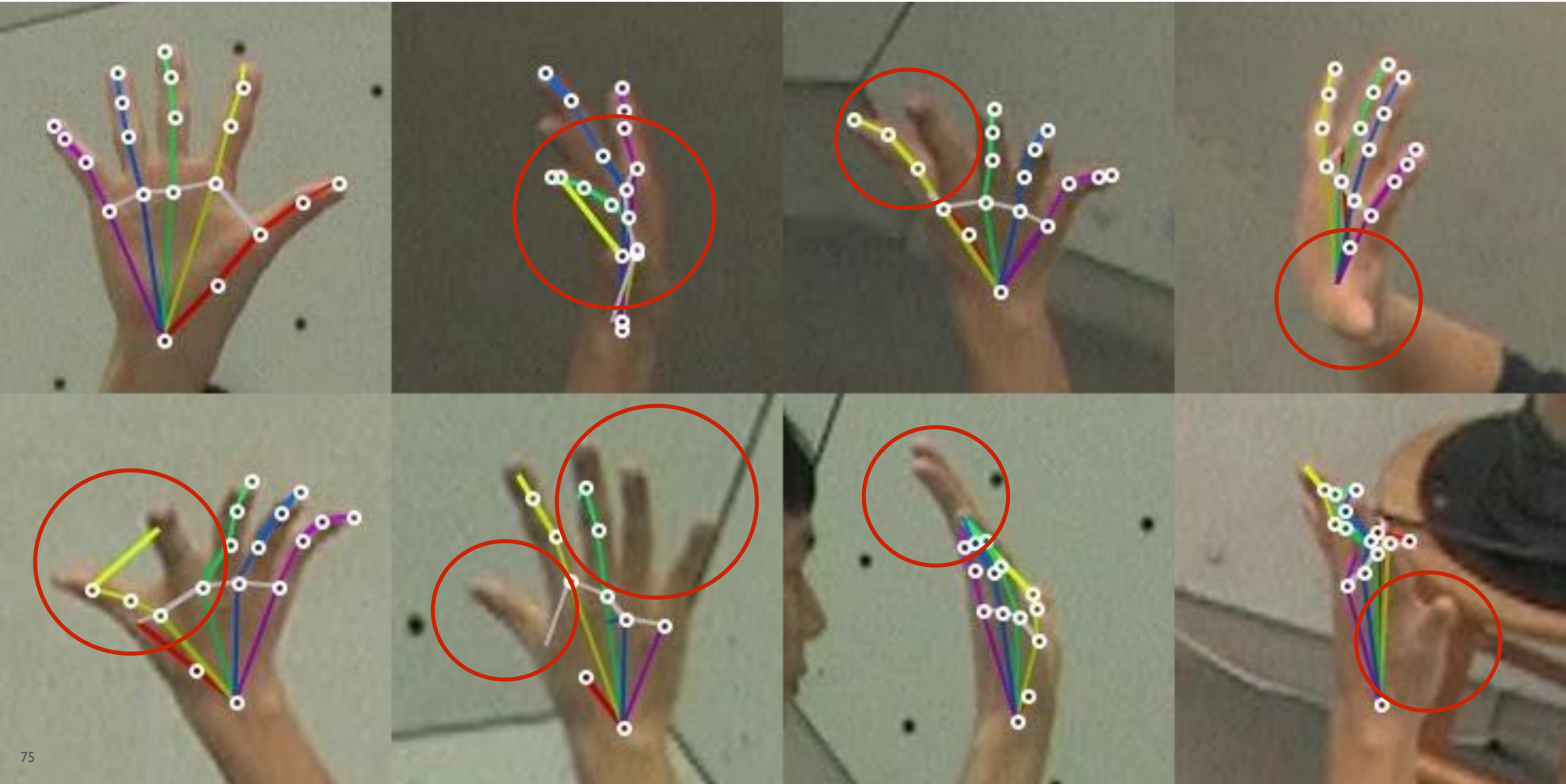
...


View 31

Initial 2D Detections (Inaccurate)

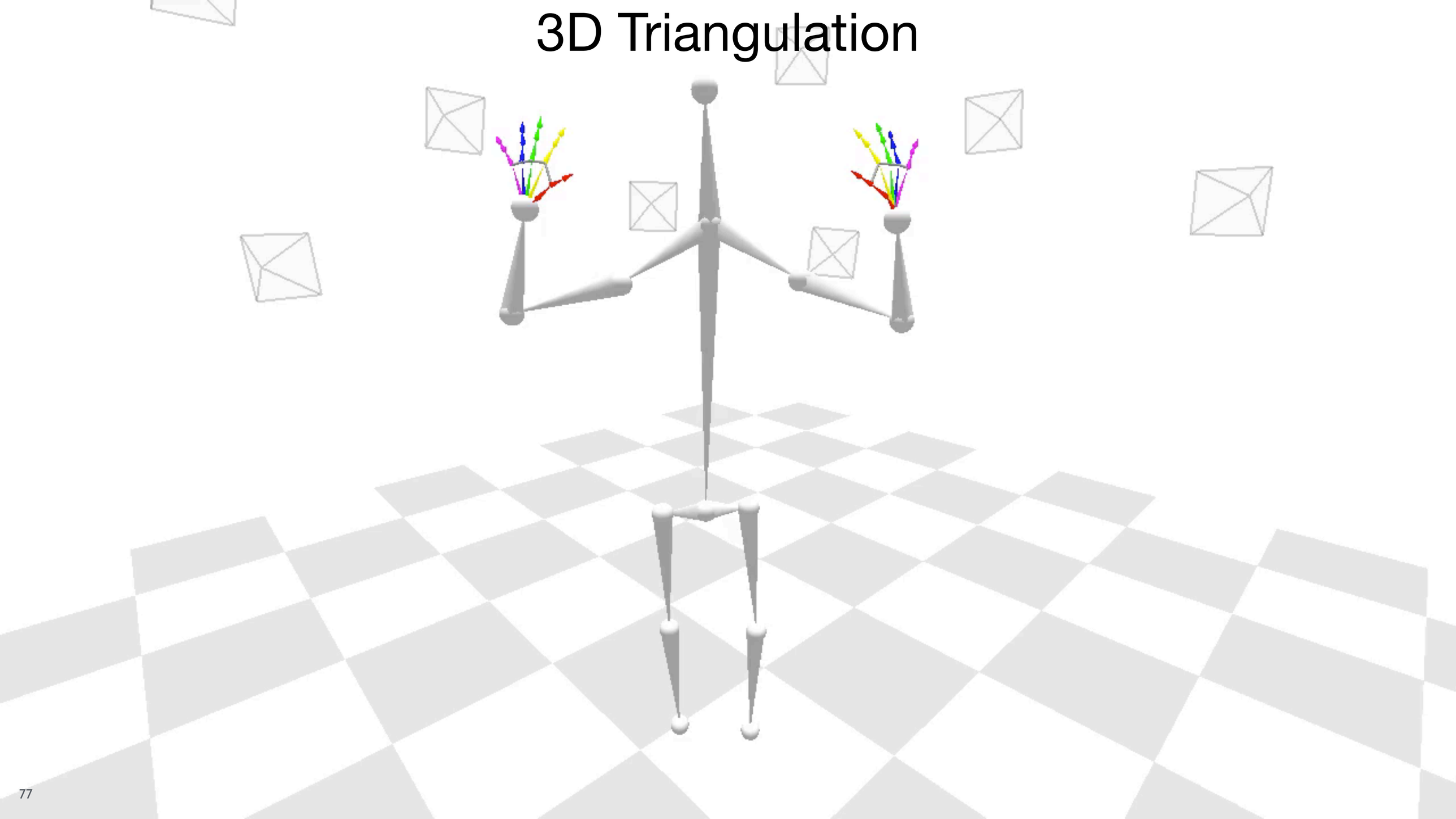


Initial 2D Detections (Inaccurate)



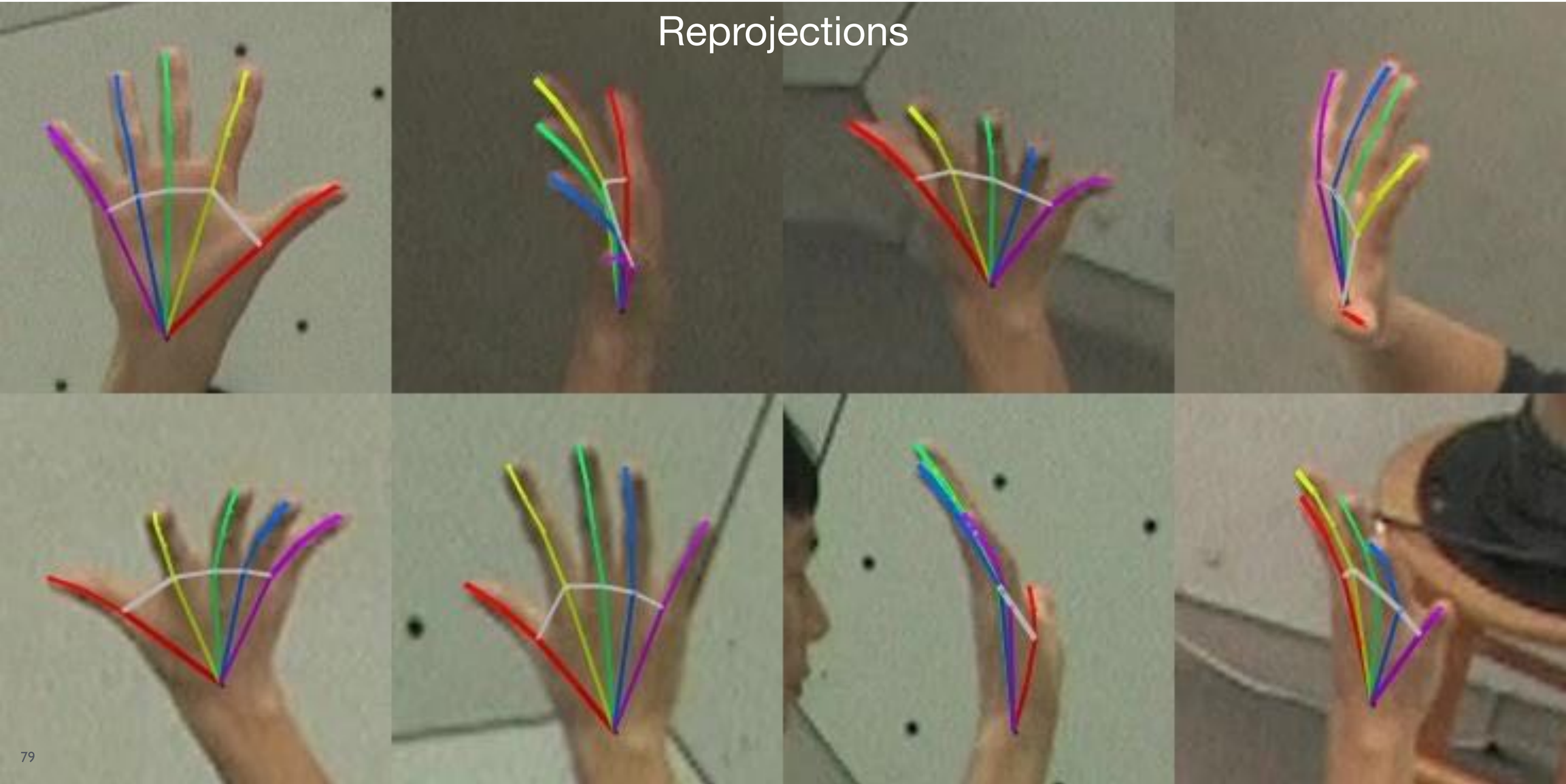
2. Robust 3D Triangulation

3D Triangulation

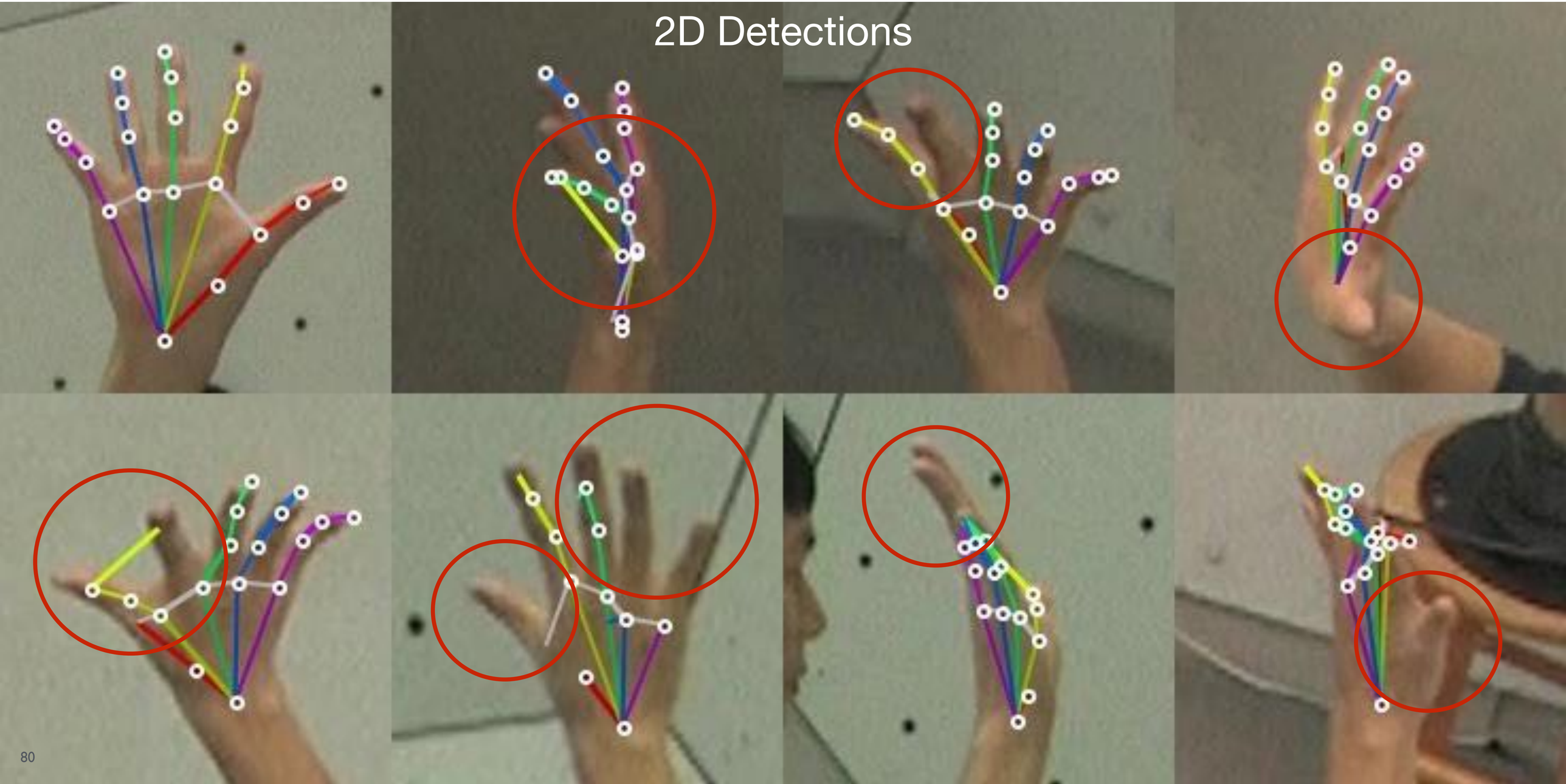


3. Reprojections

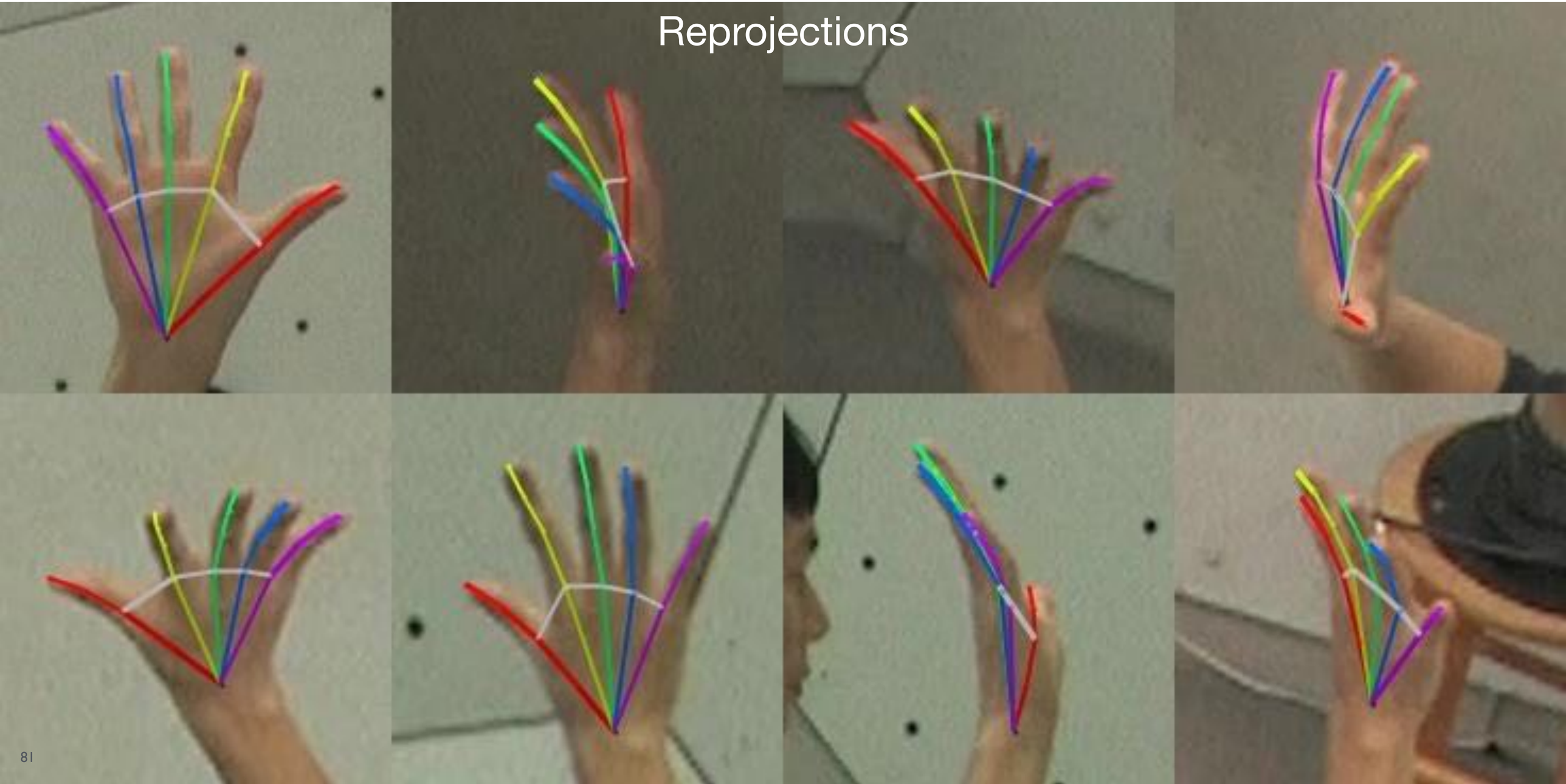
Reprojected Triangulations Are More Accurate



Reprojected Triangulations Are More Accurate

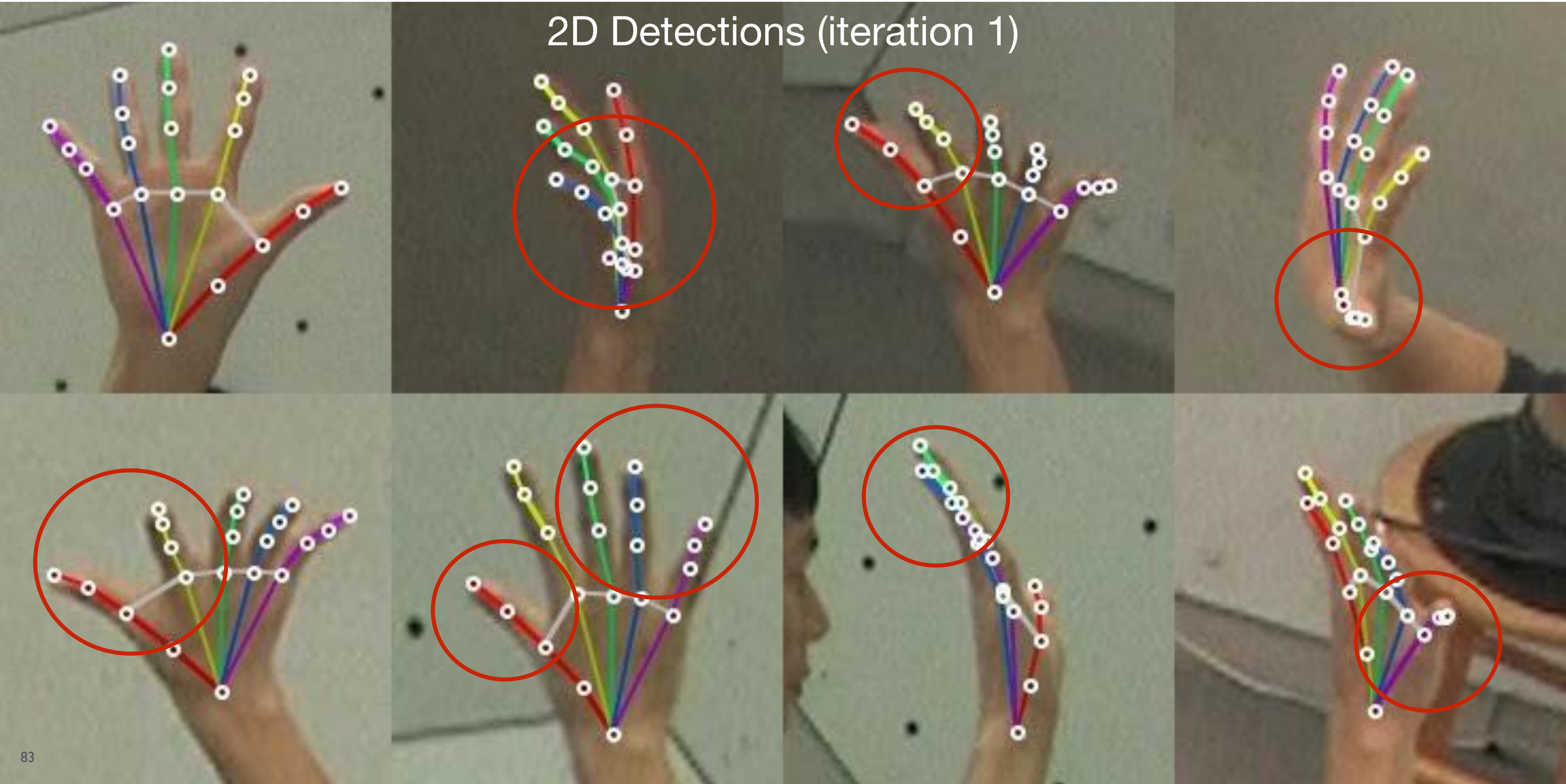


Reprojected Triangulations Are More Accurate

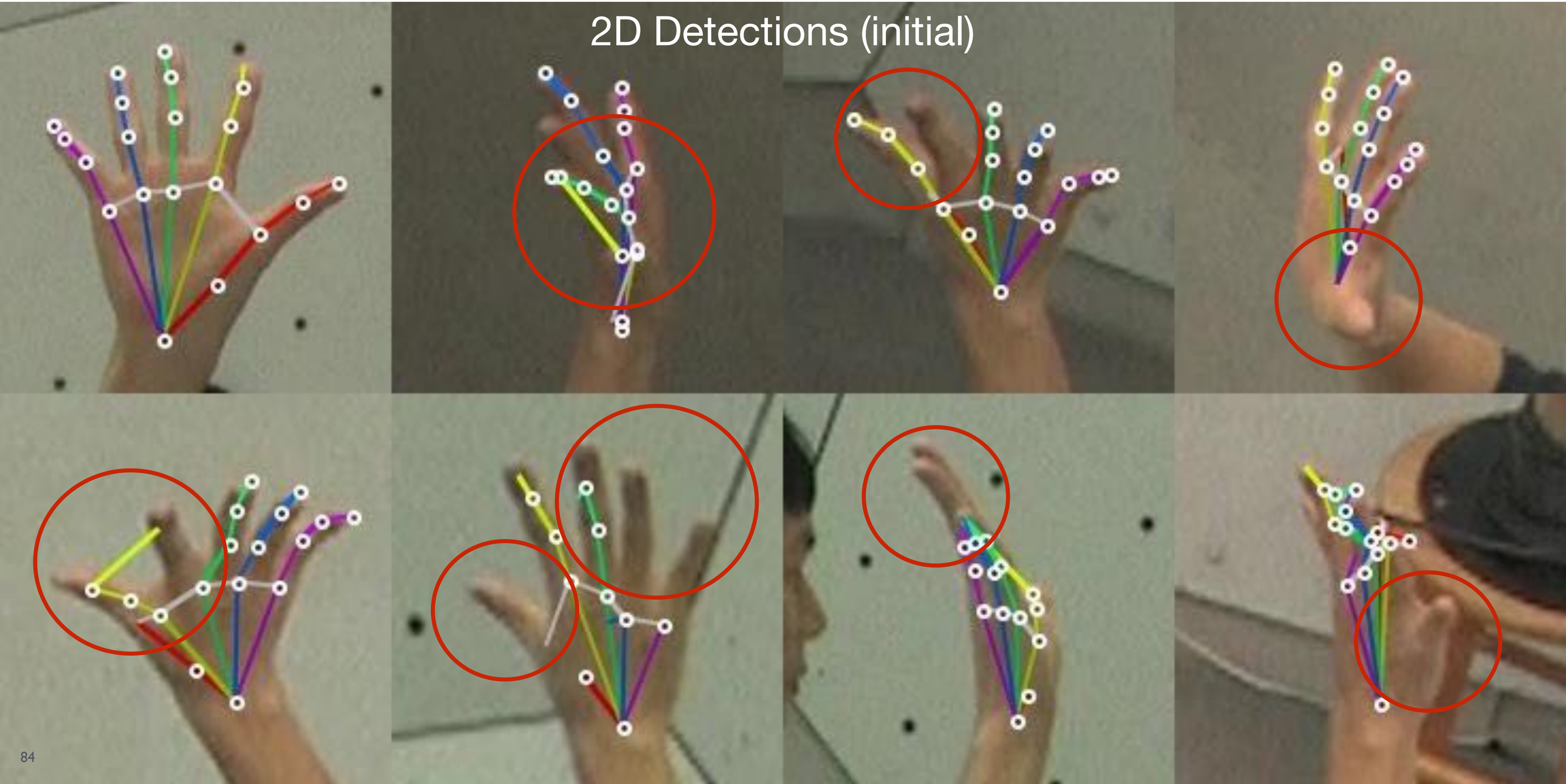


4. Retrain Detector Using Reprojected Examples

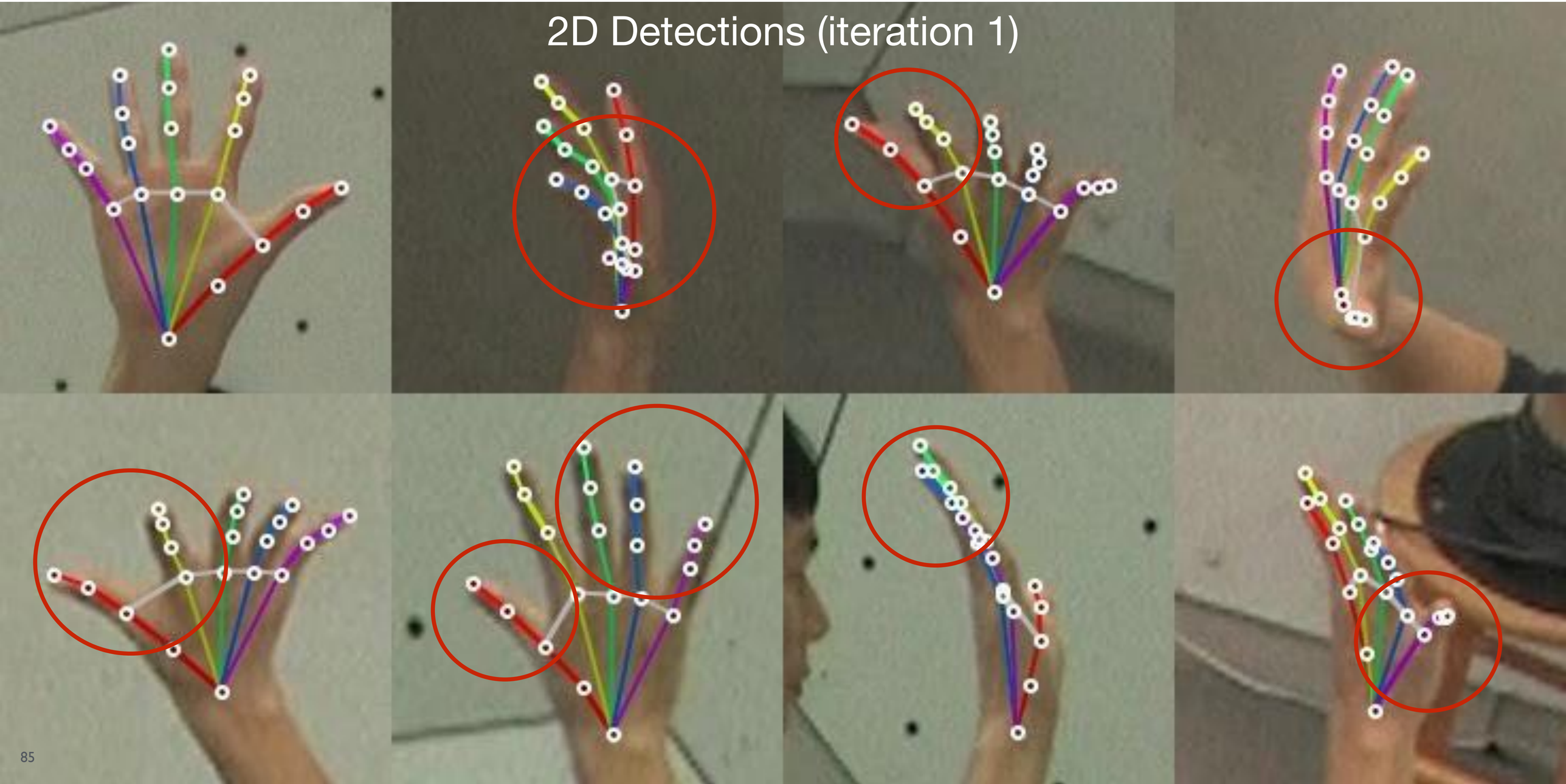
Retrained Detector Is More Accurate



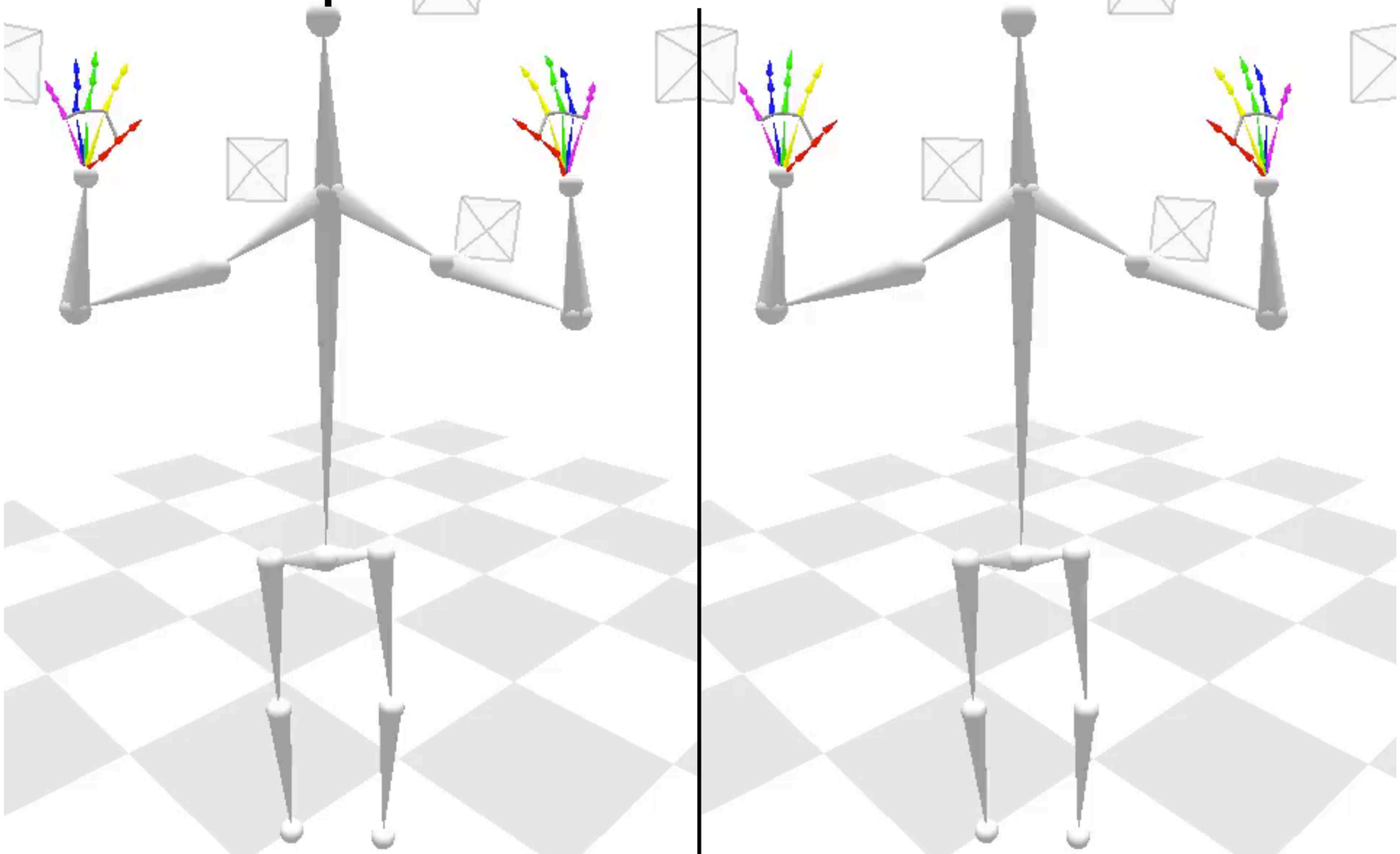
Retrained Detector Is More Accurate



Retrained Detector Is More Accurate



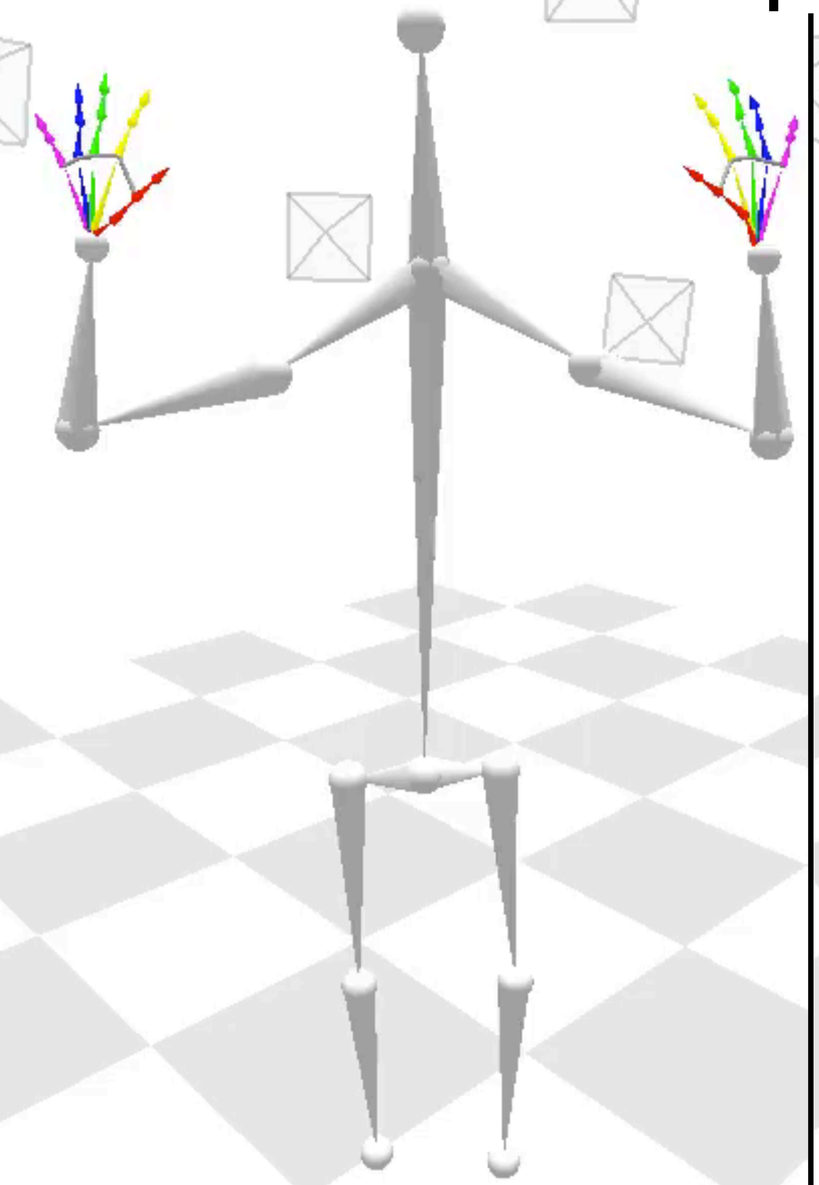
Comparison Between Iterations



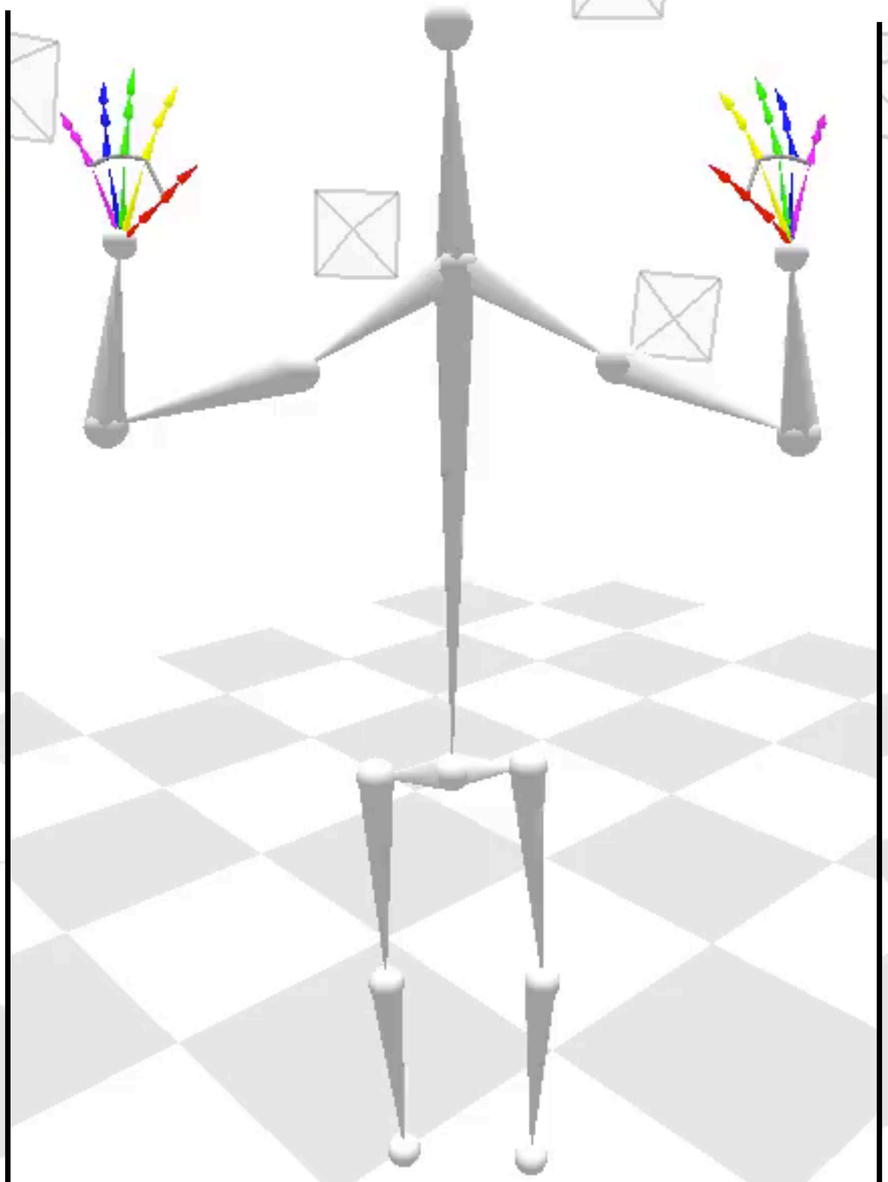
Initial (Iteration 0)

Iteration 1

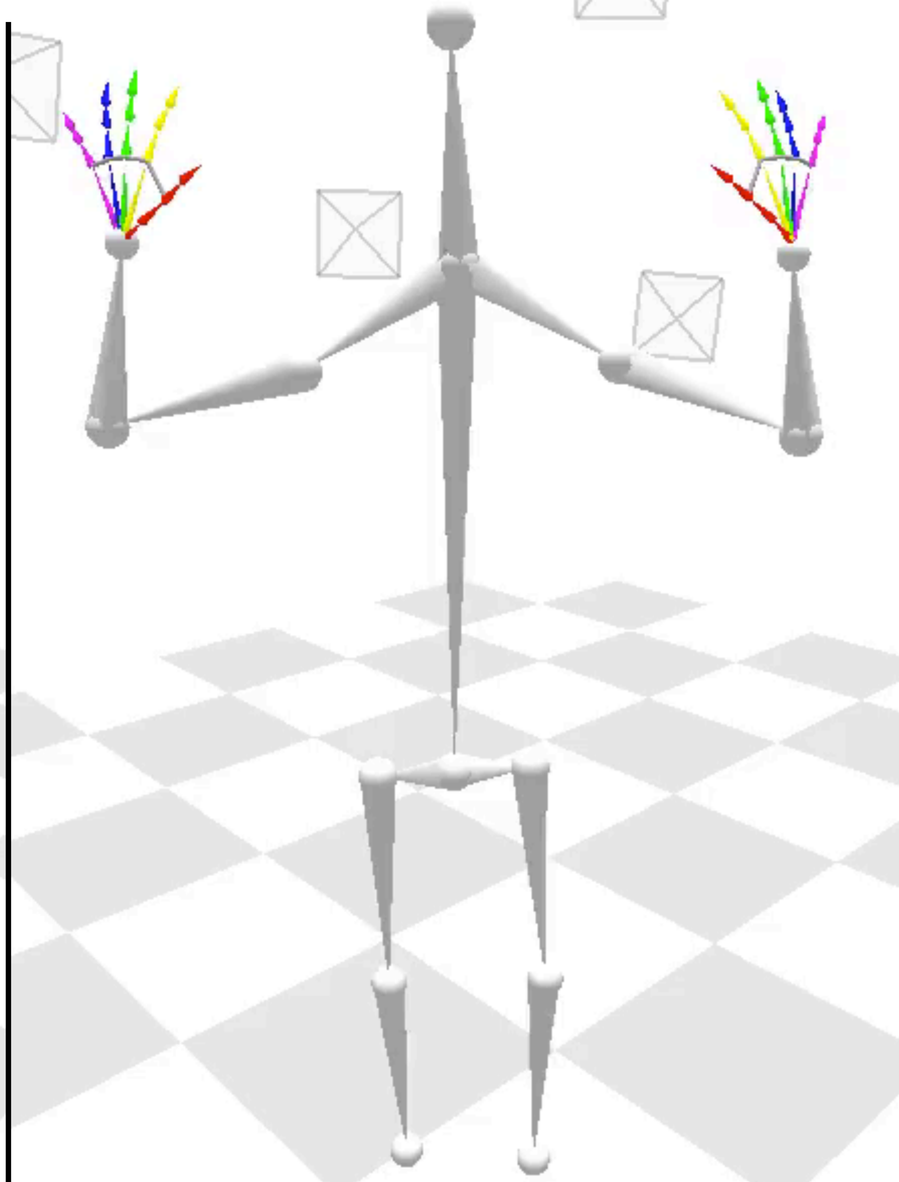
Comparison Between Iterations



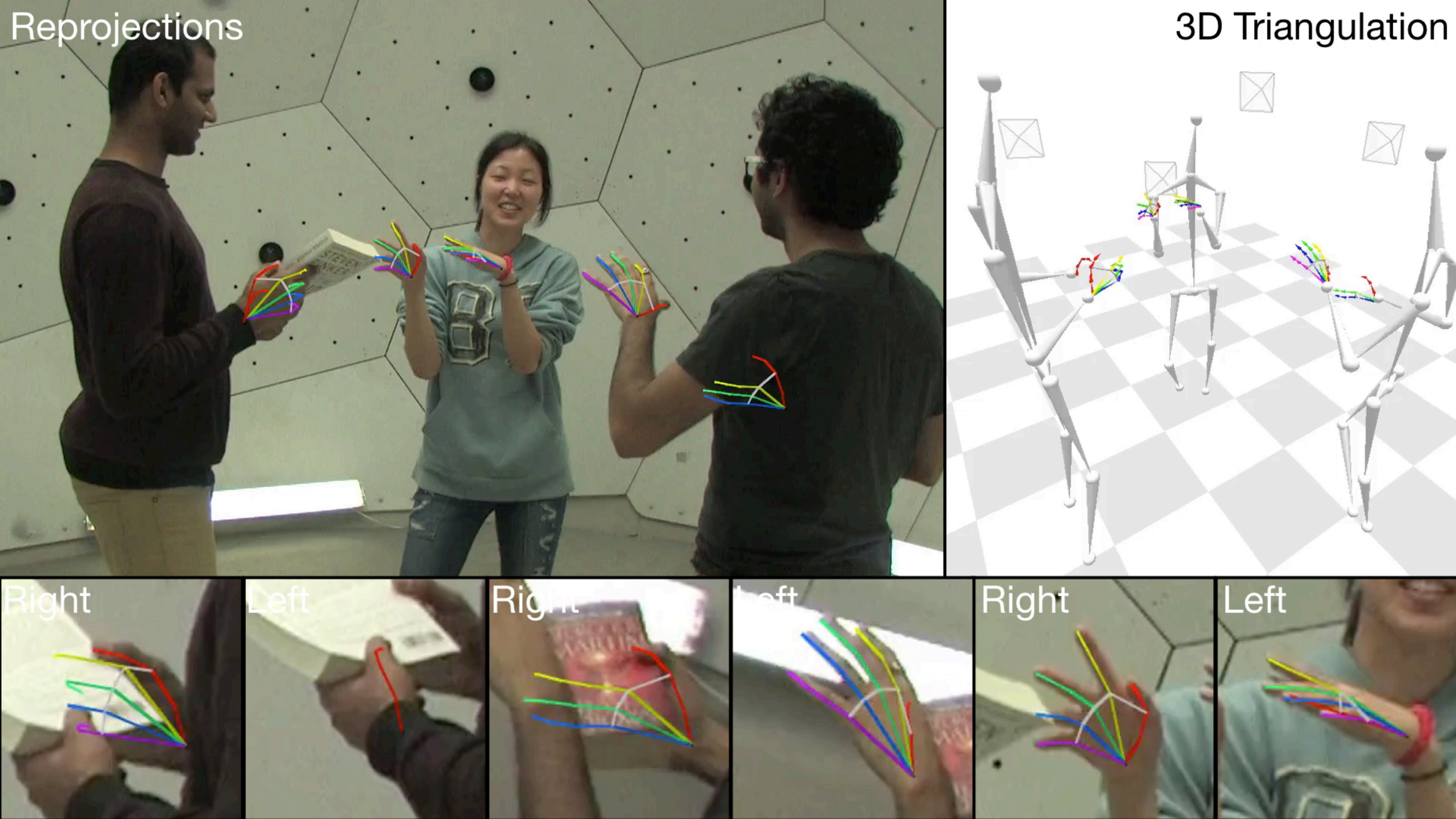
Initial (Iteration 0)

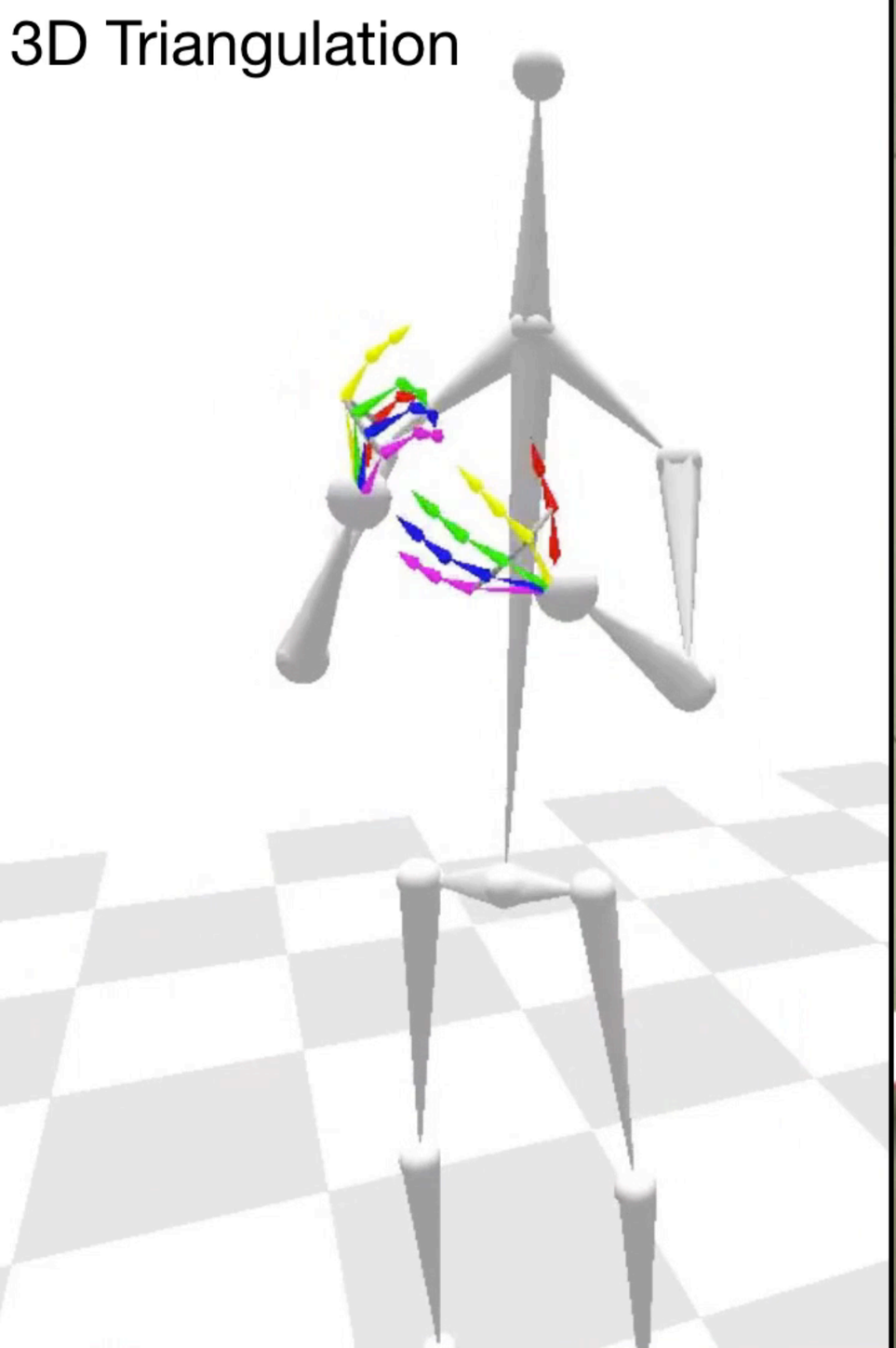


Iteration 1

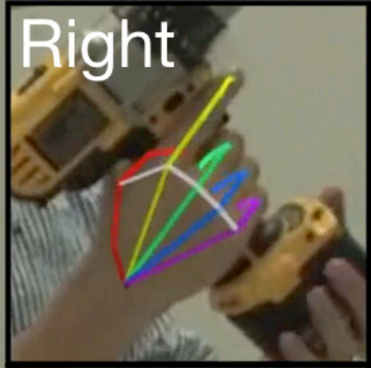


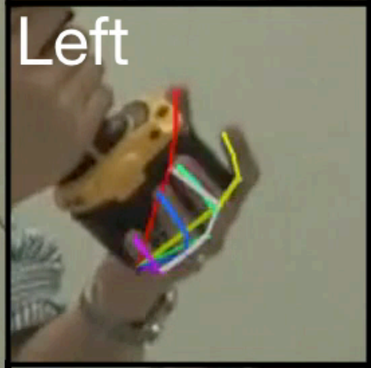
Iteration 2

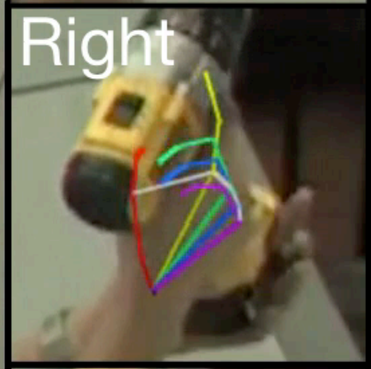

Reprojections
3D Triangulation
Right
Left
Right
Left
Right
Left

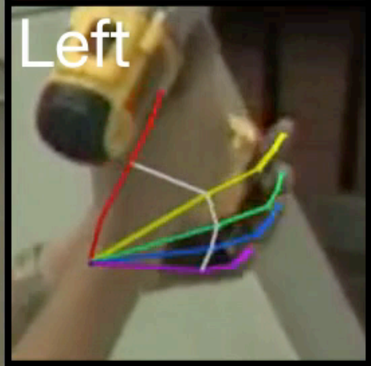

3D Triangulation


Reprojections

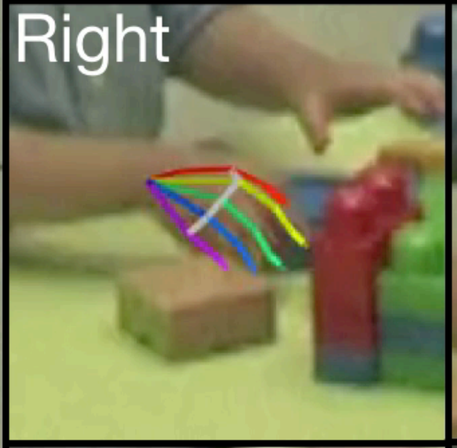

Right


Left


Right

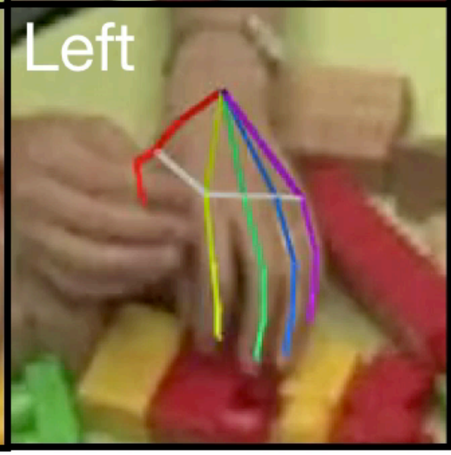

Left

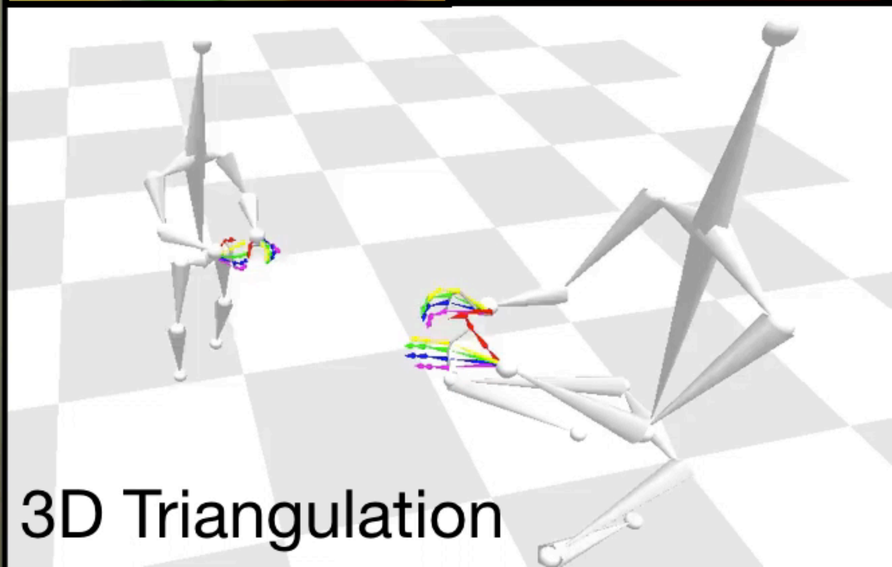

Reprojections


Right


Left


Right


Left

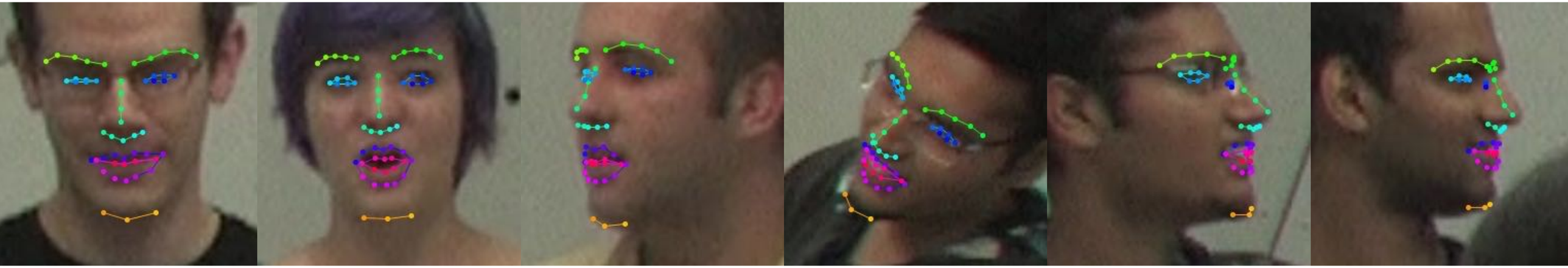

3D Triangulation

Improving View Robustness of Facial Keypoint Detectors

Initial Detections (Iteration 0 --- Manual labels MultiPIE, Helen, AFW, ...)

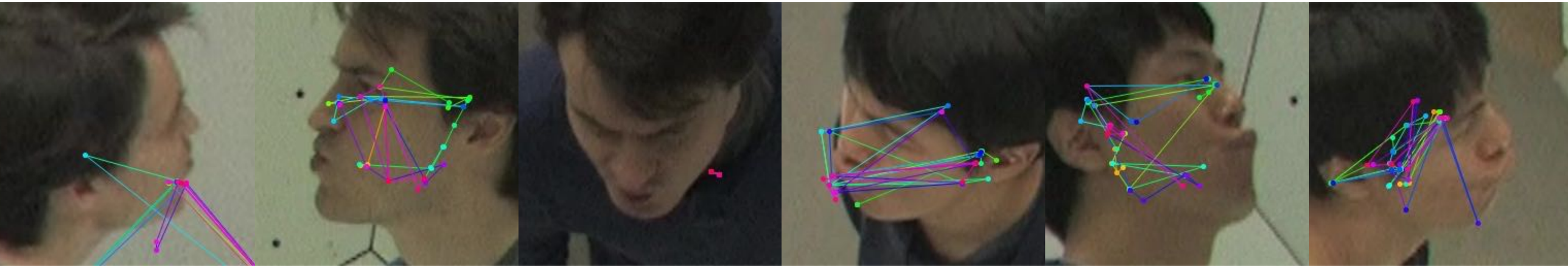
Improving View Robustness of Facial Keypoint Detectors

Initial Detections (Iteration 0 --- Manual labels MultiPIE, Helen, AFW, ...)

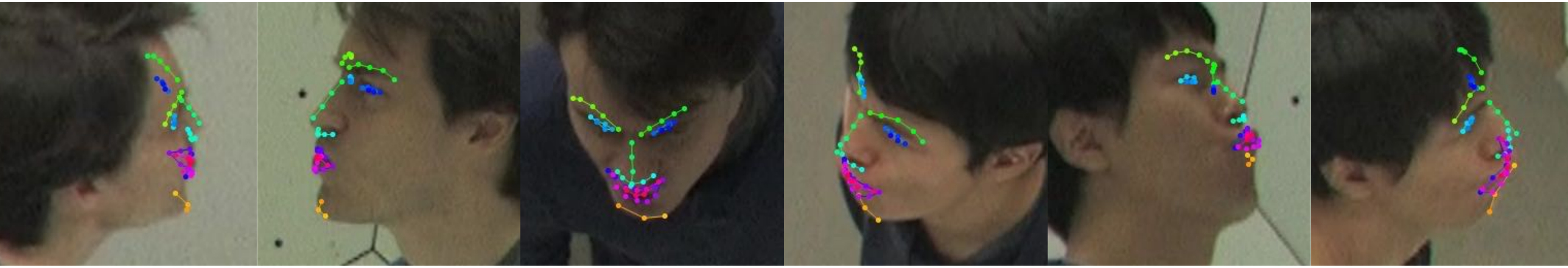


Retrained Detections (Iteration 1)

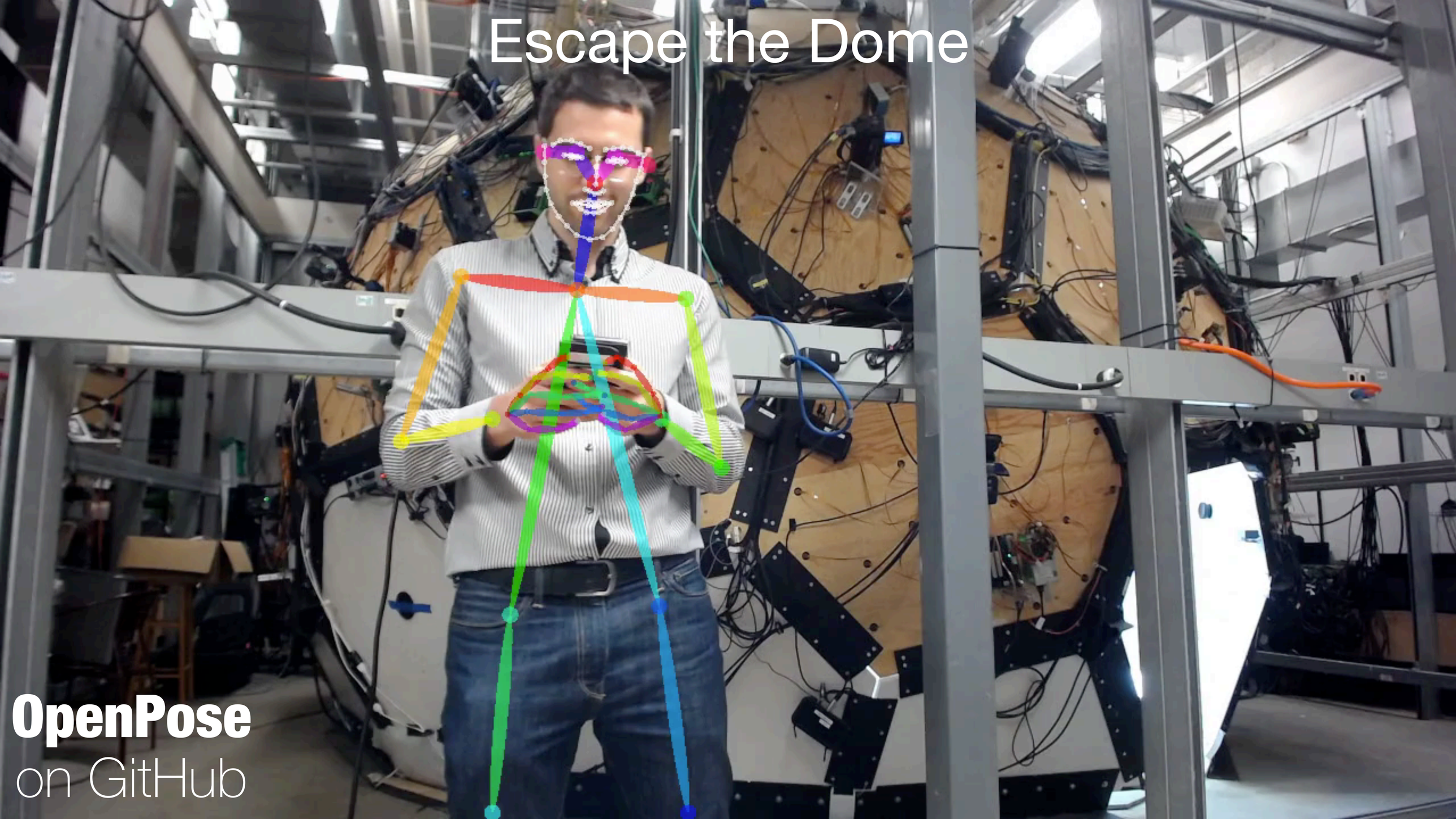
Improving View Robustness of Facial Keypoint Detectors



Initial Detections (Iteration 0 --- Manual labels MultiPIE, Helen, AFW, ...)



Retrained Detections (Iteration 1)


Escape the Dome
OpenPose
on GitHub

The Panoptic Studio Dataset

A Large Scale Dataset For Kinesic Signals Processing

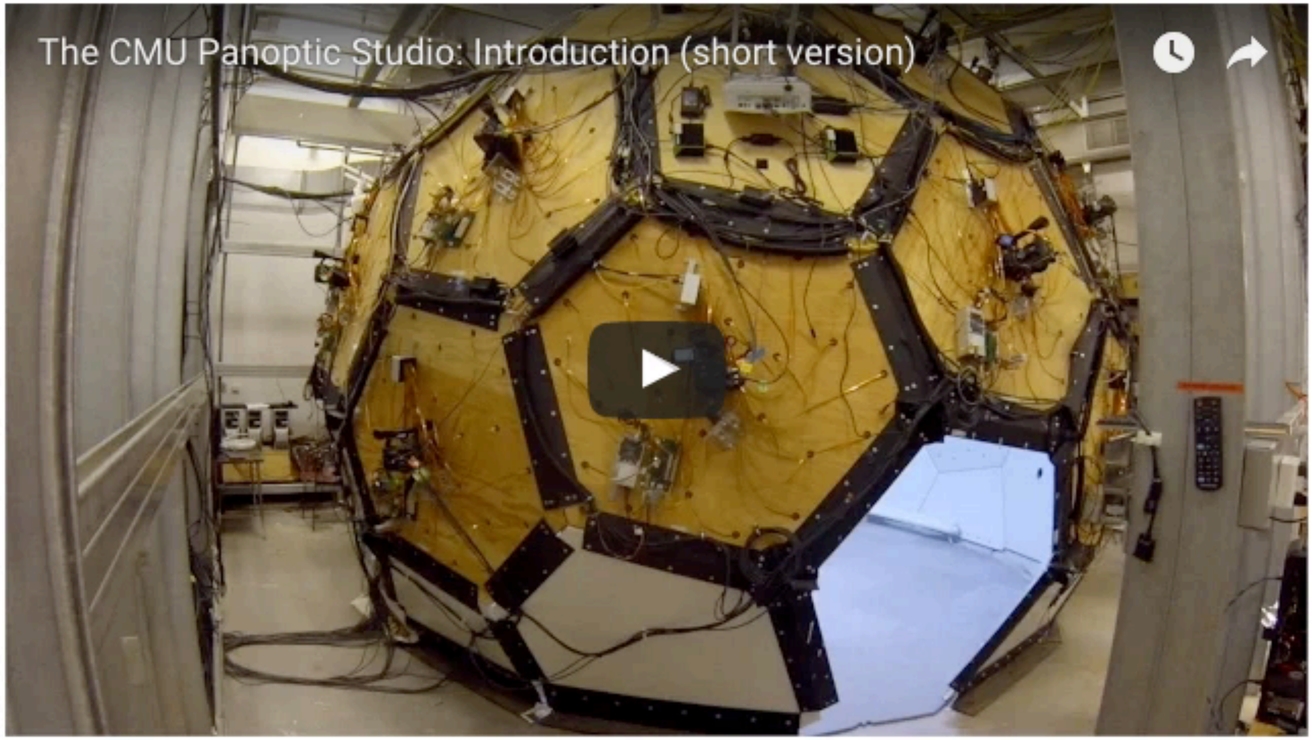
CMU Panoptic Dataset

domedb.perception.cs.cmu.edu/index.html

Apps | arxiv | CMU-Perceptual-C... | Docs | ResearchNote - Go... | Arxiv Sanity Preserver | Perceptual Computi... | Dome Database - G... | Gaze Prediction - G... | Other Bookmarks

CMU Panoptic Dataset

Home | Dataset | Collections | People | Docs & Tools | Tutorial | References



* See the full length version of this video here

Massively Multiview System

- 480 VGA camera views
- 30+ HD views
- 10 RGB-D sensors
- Hardware-based sync
- Calibration

Interesting Scenes with Labels

- Multiple people
- Socially interacting groups
- 3D body pose
- 3D facial landmarks
- Transcripts + speaker ID

Dataset Size

Currently, 65 sequences (5.5 hours) and 1.5 millions of 3D skeletons are available.

What's New

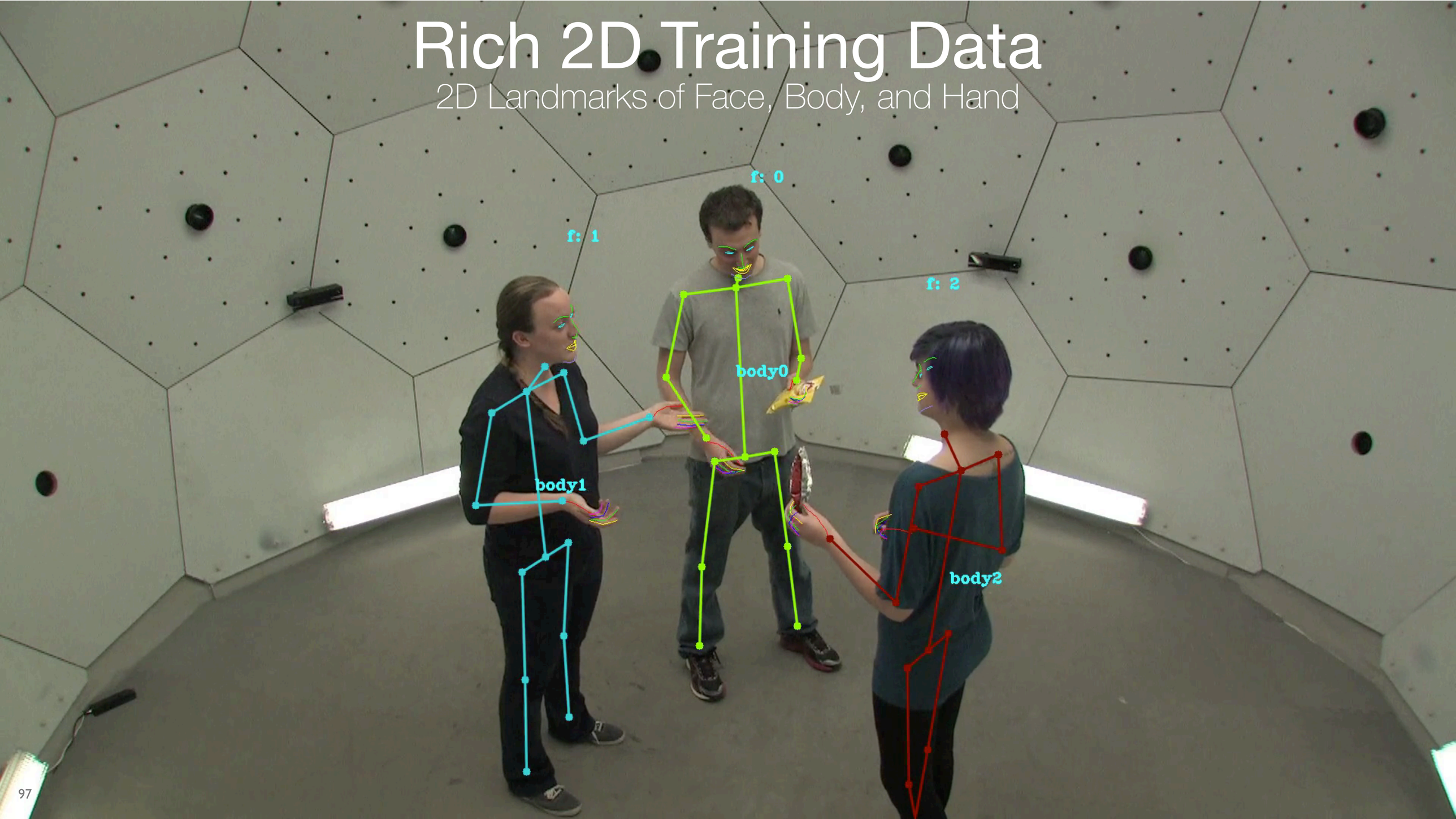
Dec. 2016 Panoptic Studio featured on The Verge. You can also see the video version here.

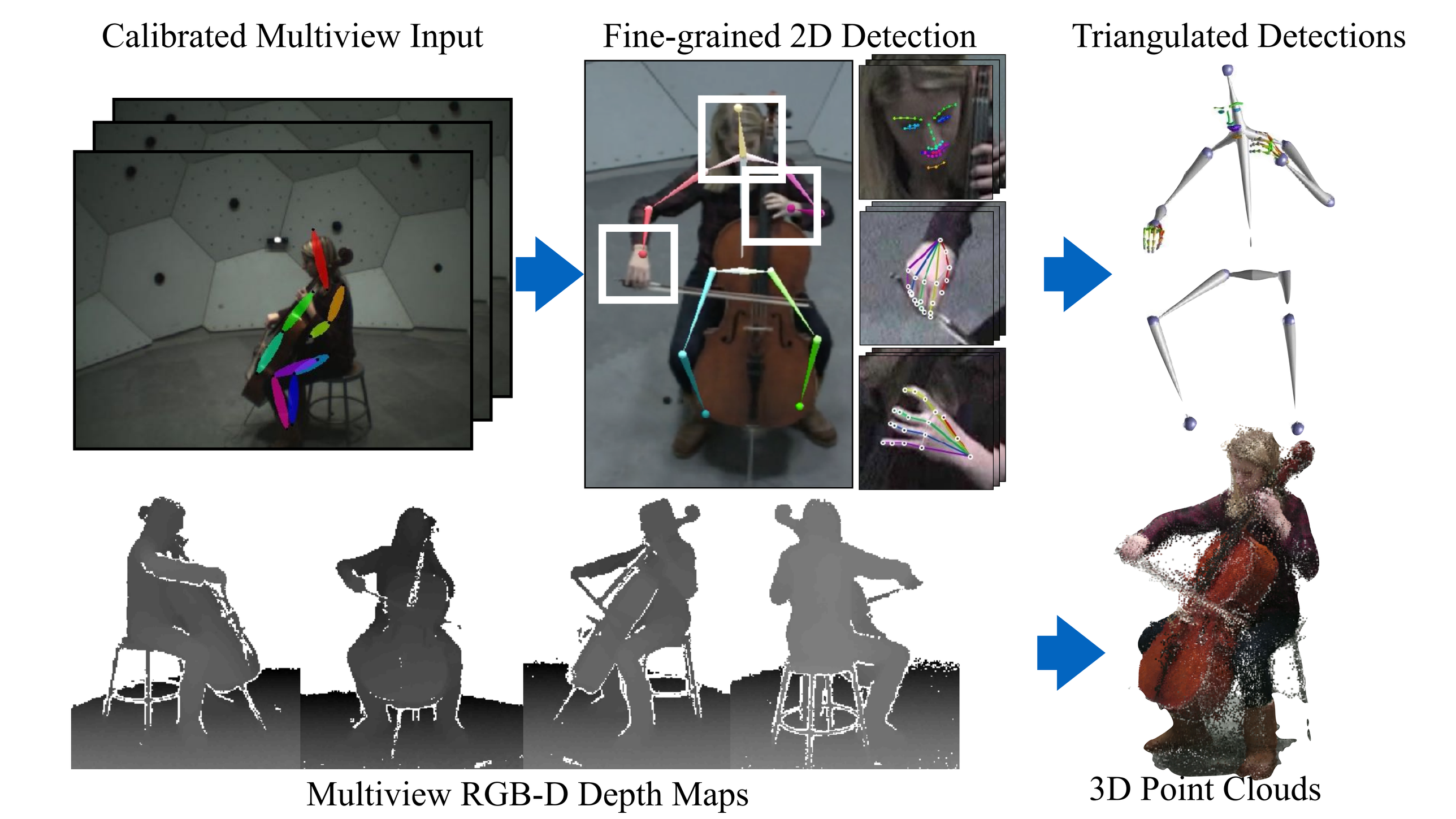
Dec. 2016 The social interaction capture paper (extended version of ICCV15) is available on arXiv.

- 12 hours of videos (500 TB)
- More than 150 individuals
- More than 300 social games

Rich 2D Training Data

2D Landmarks of Face, Body, and Hand




Calibrated Multiview Input
Fine-grained 2D Detection
Triangulated Detections
Multiview RGB-D Depth Maps
3D Point Clouds

Questions?