浙江大學
ZHEJIANG UNIVERSITY

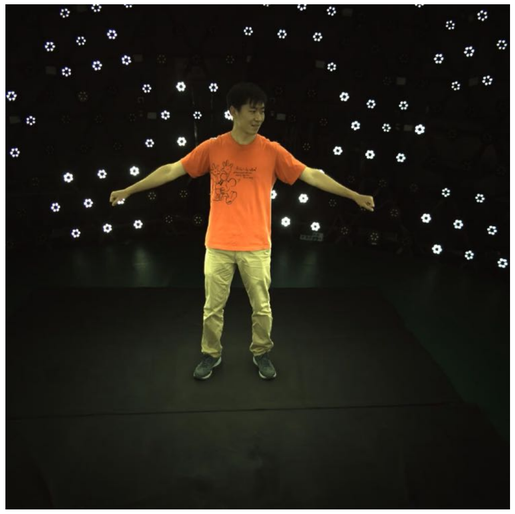
Human Motion Capture from RGB Videos

周晓巍

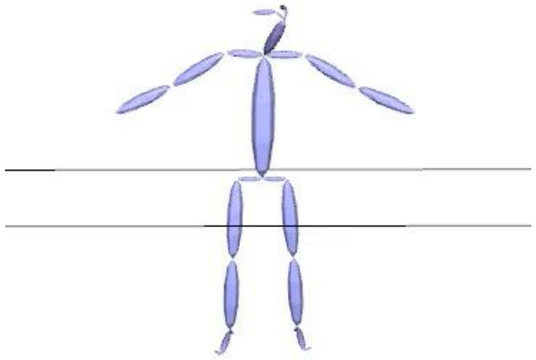
浙江大学CAD&CG国家重点实验室

Human motion capture (MoCap)

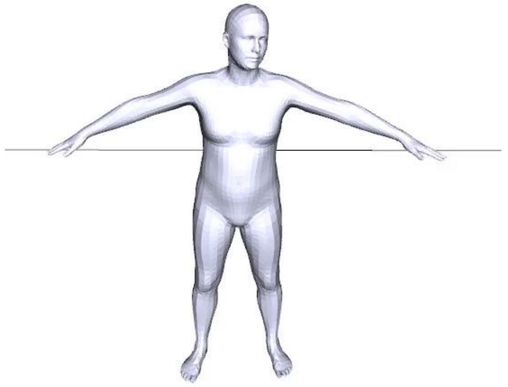
Recovering 3D human motion from sensor data



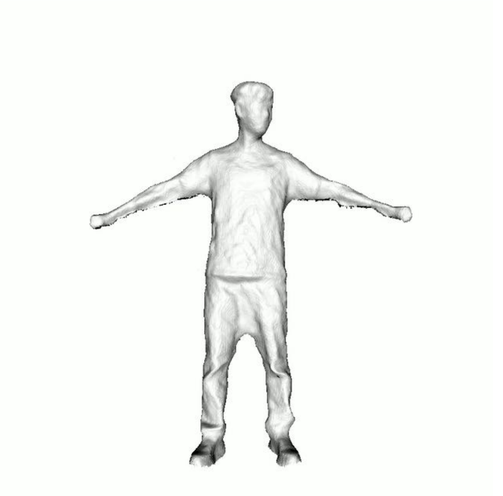
Sensor data



Skeleton



Template



Detailed surface

Applications

HCI

Source: Microsoft Kinect

Movie

Source: www.motionshadow.com

Sports

Source: ShotTracker

Telepresence

Source: Microsoft holoportation

Existing MoCap systems

Optical MoCap systems (e.g. Vicon)

- Need body markers
- Special hardware

Depth sensors (e.g. Holoportation)

- Limited sensing range
- Constrained environments

Making MoCap more practical

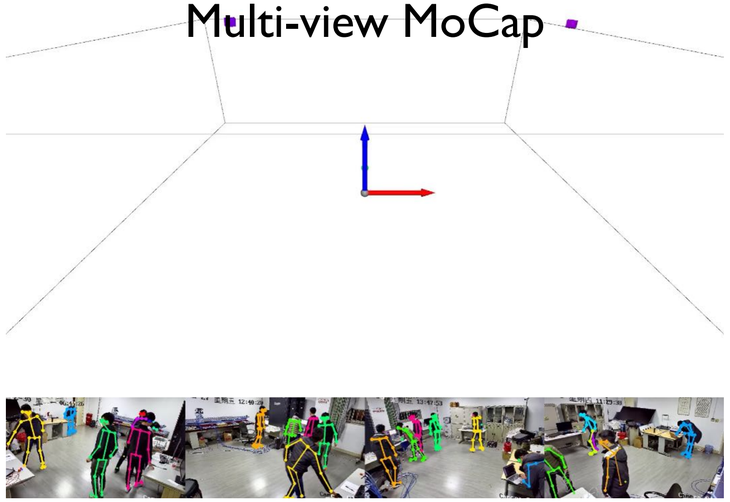
How to make MoCap systems more widely applicable?

Existing systems
- Expensive & special hardware
- Need markers
- Constrained environments

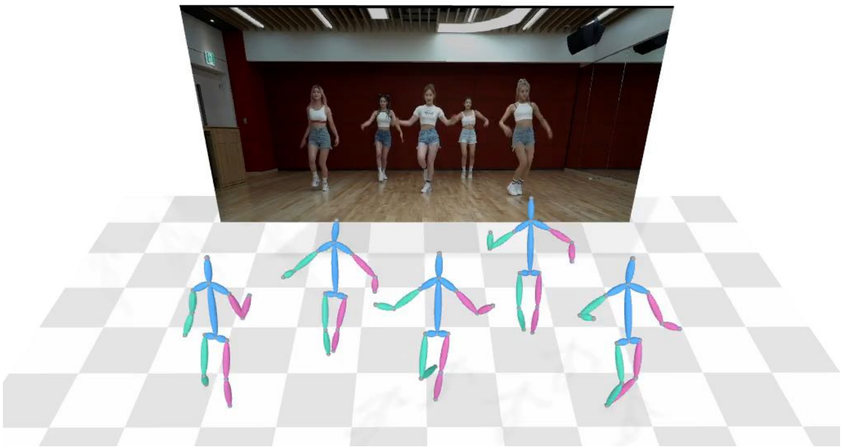
More practical:
- Few RGB cameras
- No markers
- Unconstrained environments

In this talk: MoCap from RGB videos

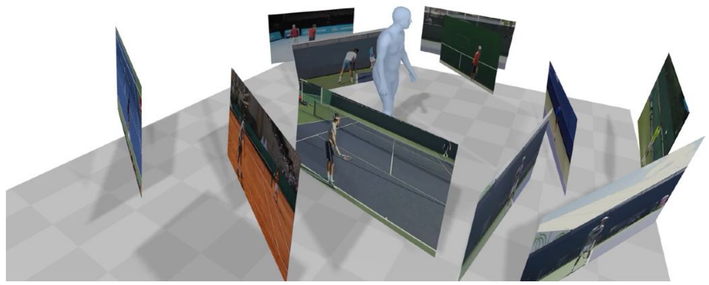
Multi-view MoCap



Single-view MoCap

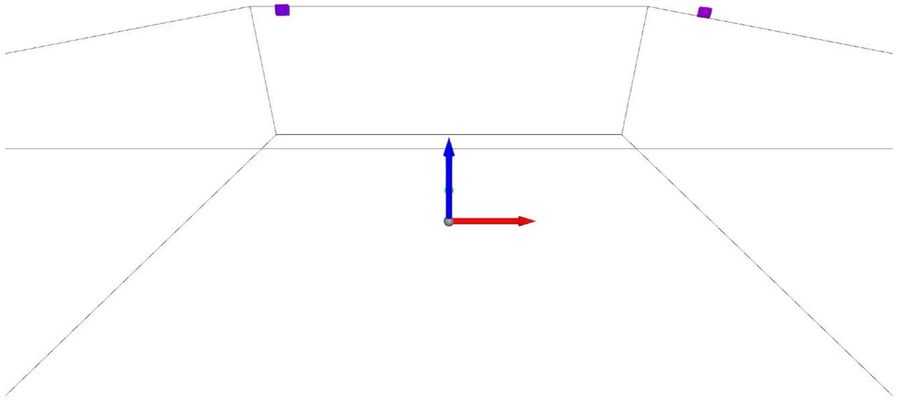


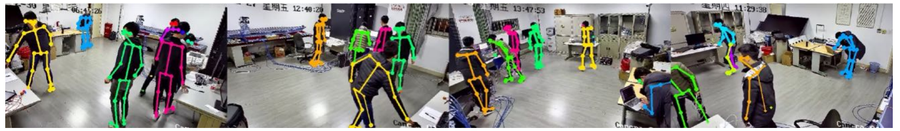
MoCap from Internet videos



Novel view synthesis

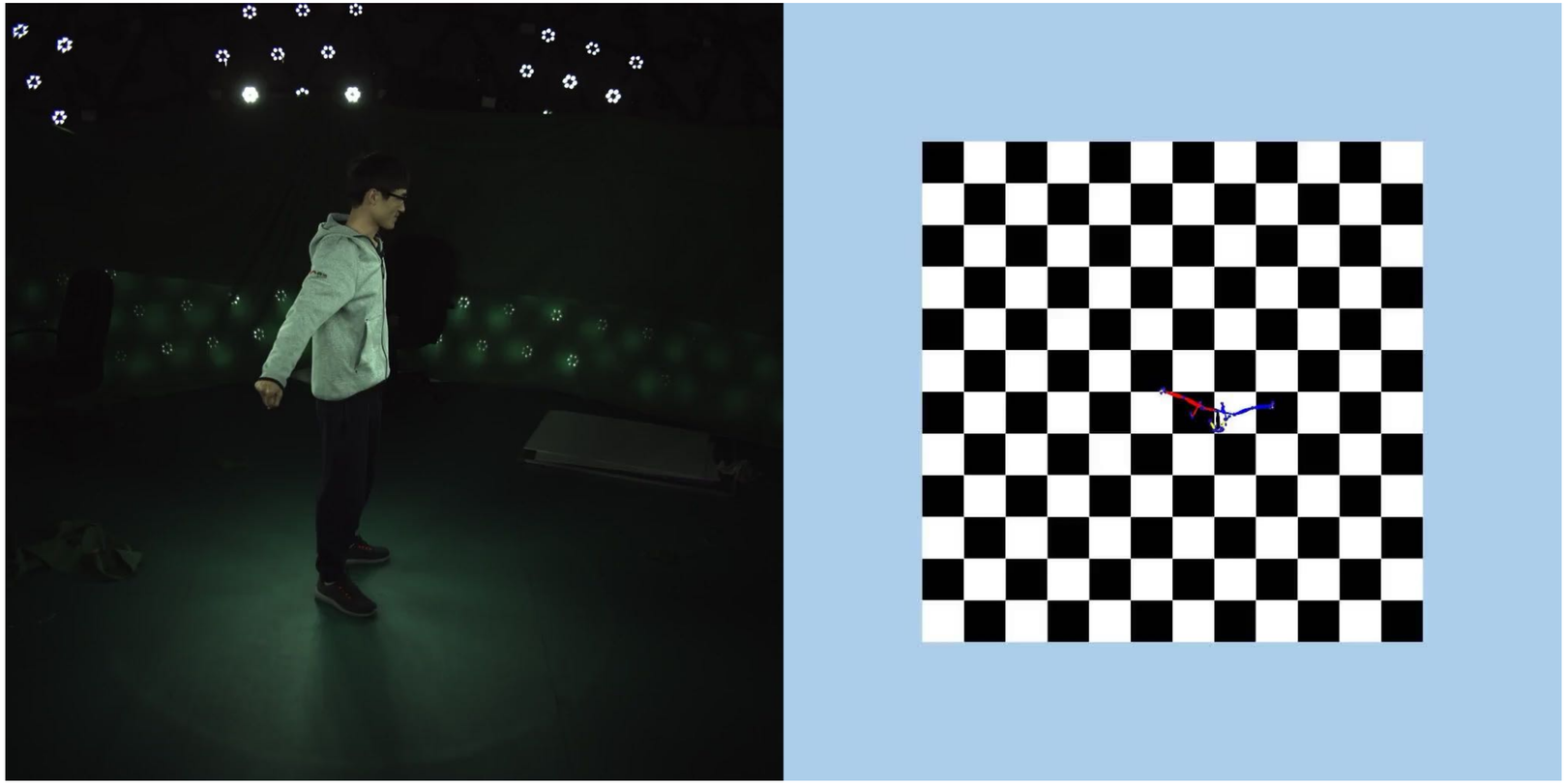
Part I: Multi-view MoCap





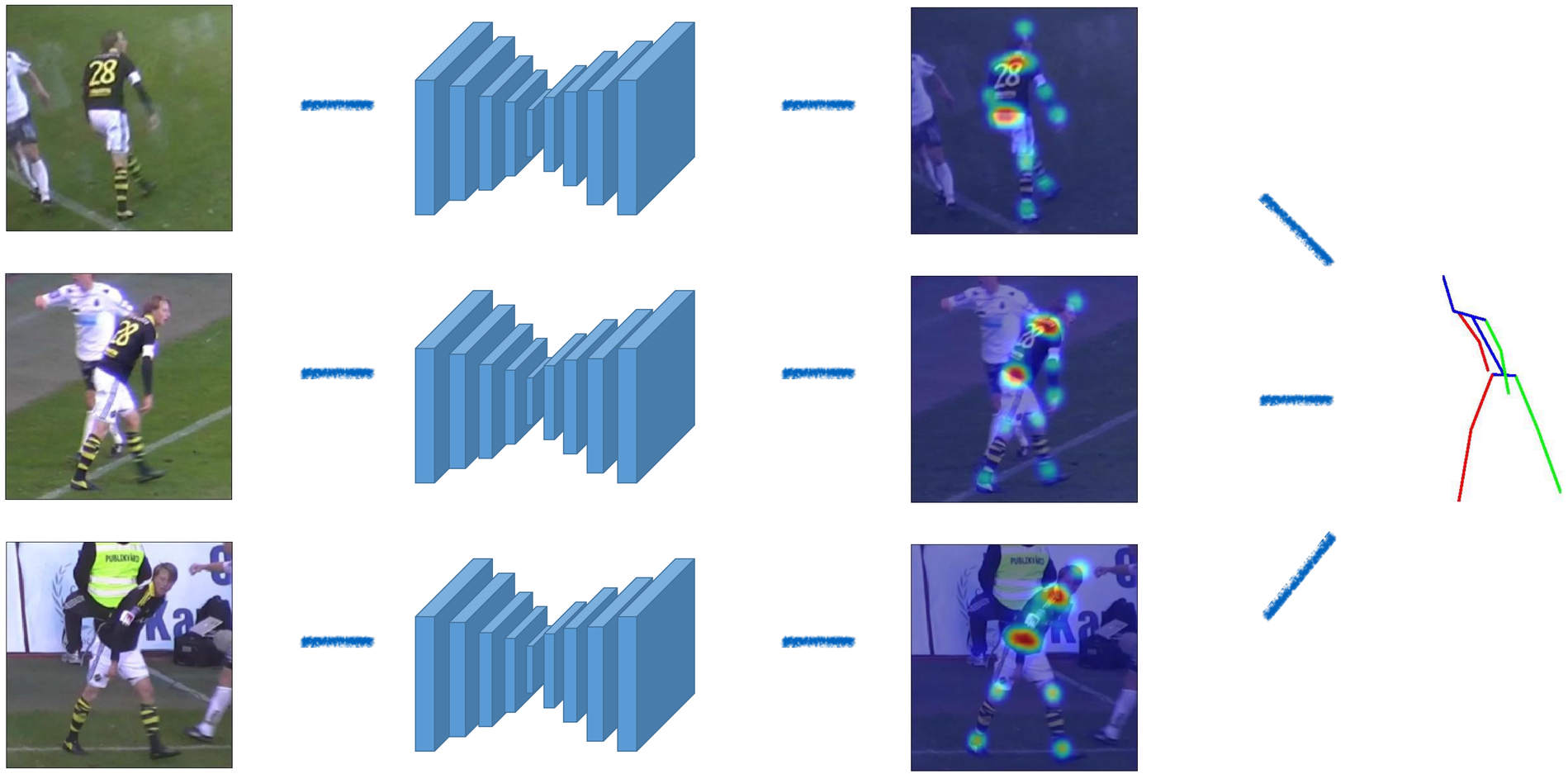
Multi-view markerless MoCap

How to get rid of markers?



Multi-view markerless MoCap

Get rid of markers by using a keypoint detector

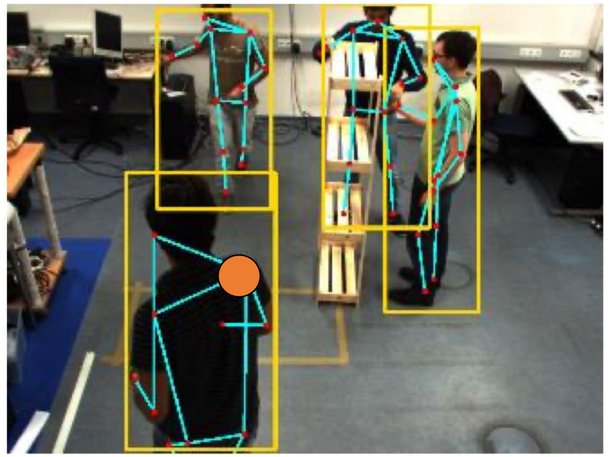


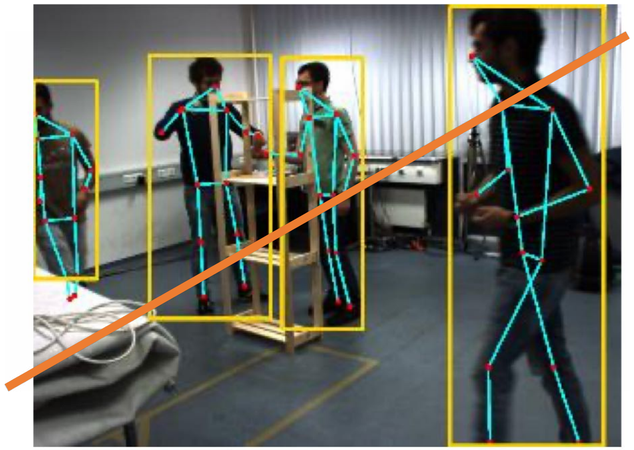
Harvesting Multiple Views for Marker-less 3D Human Pose Annotations. CVPR 2017

Multi-view markerless MoCap

How to address crowd scene?

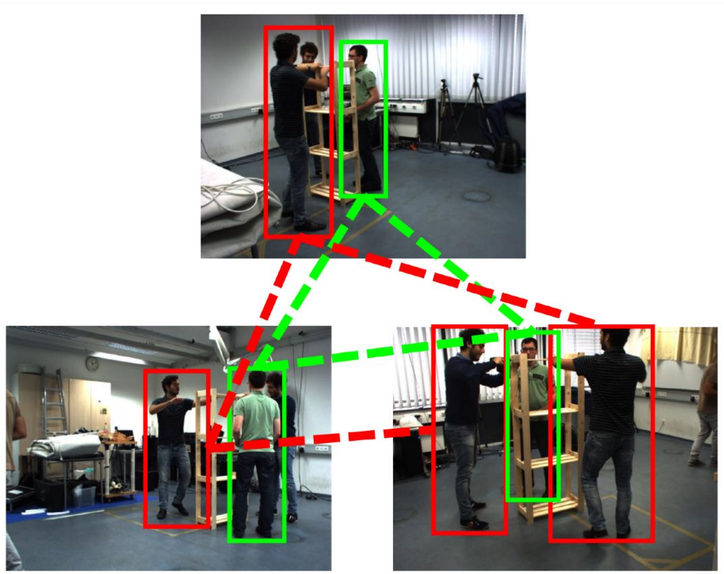
- Need to solve correspondences across views
- Challenges:
  - large viewpoint change
  - humans with similar appearance





Key ideas for multi-view matching:

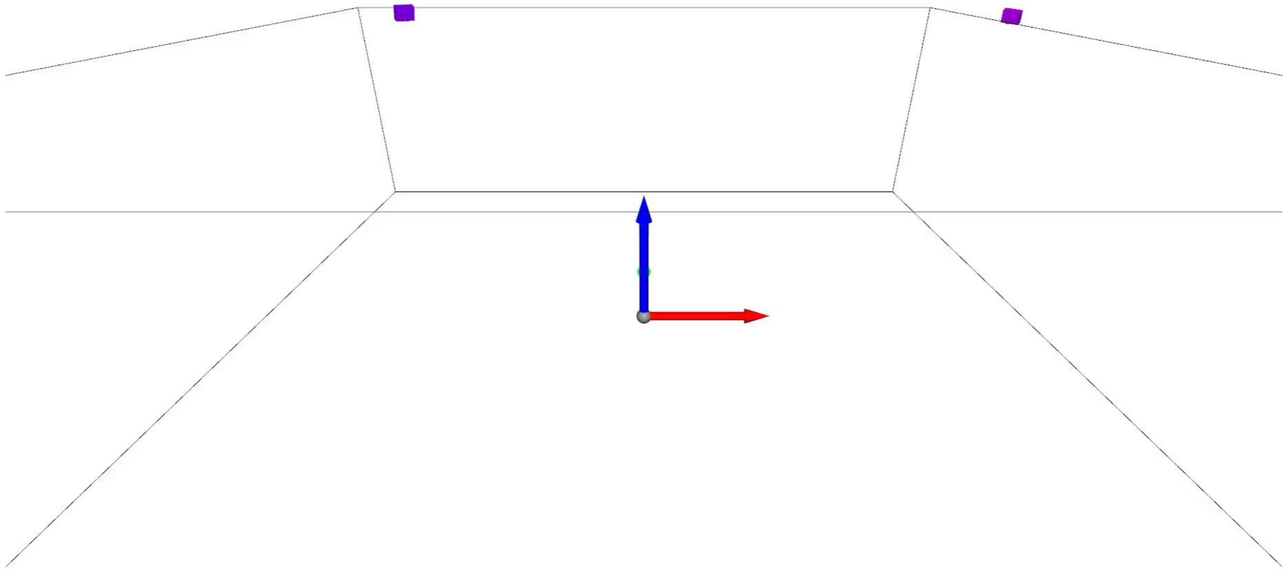
- Geometric constraints (epipolar geometry)
- Cycle consistency constraints

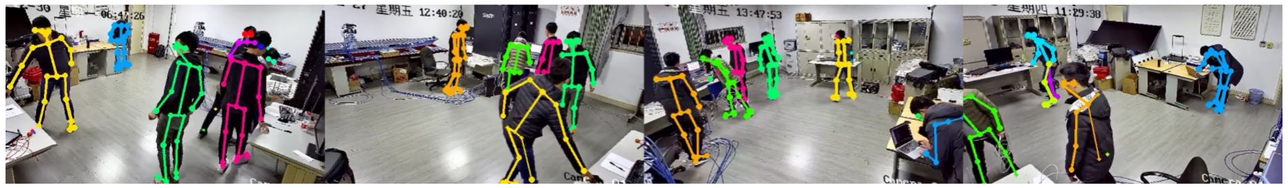


Fast and Robust Multi-Person 3D Pose Estimation from Multiple Views. CVPR 2019.

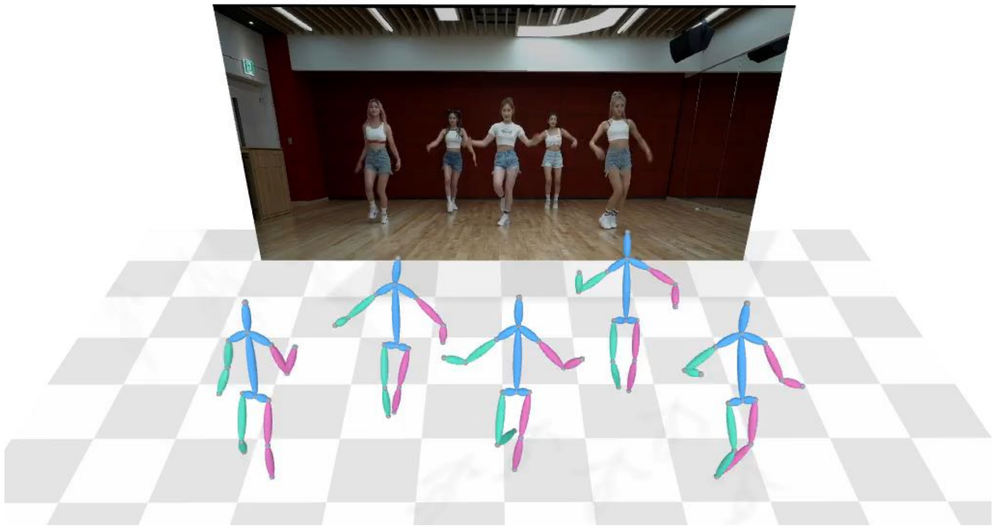
Multi-view markerless MoCap

Real-time markerless MoCap system



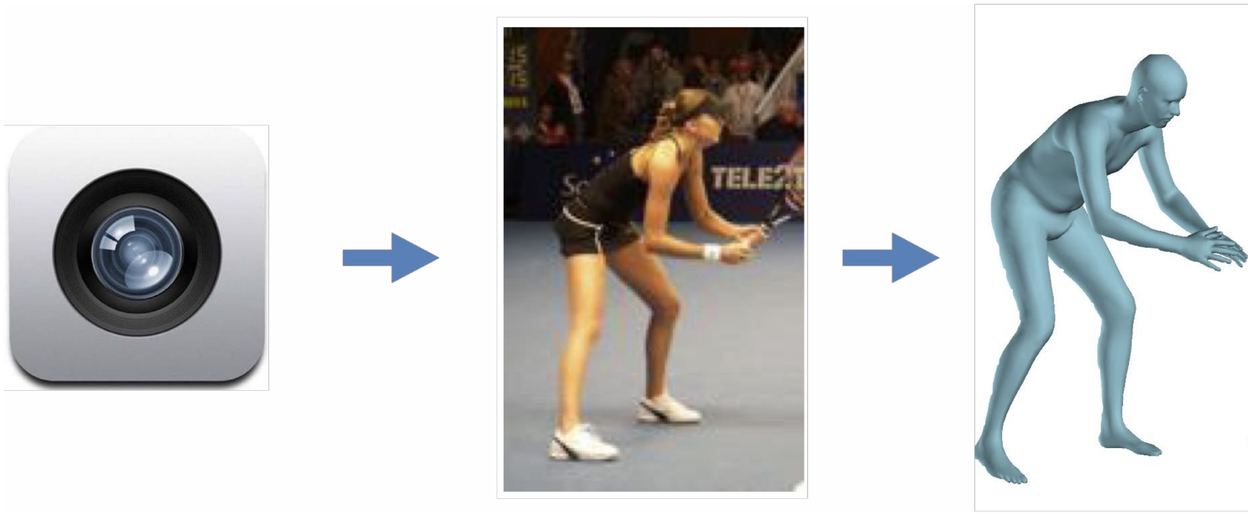


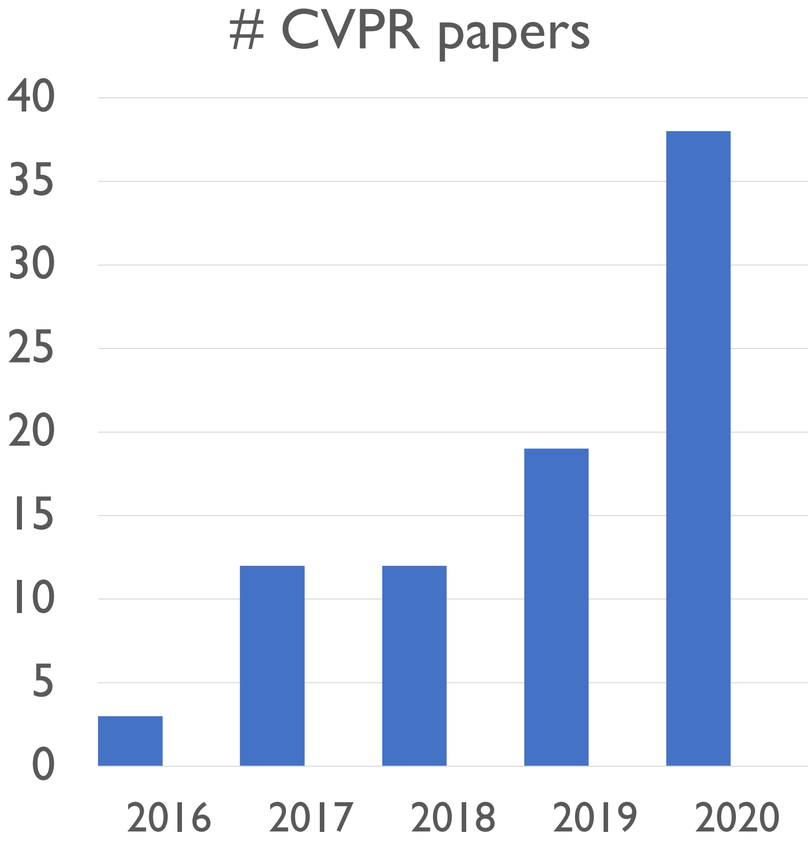
Part II: Single-view MoCap



Single-view pose estimation

Can we get rid of multiple cameras?



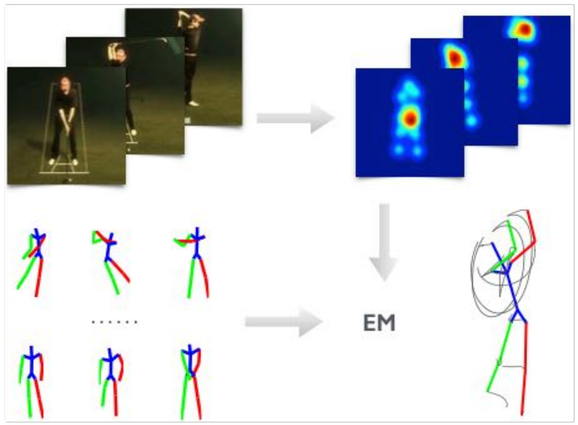


Valse2020年度进展报告：https://www.bilibili.com/video/BV1QA411Y7SD

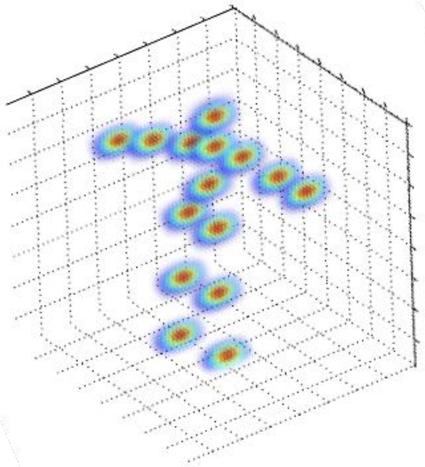
Single-view pose estimation

Different representations

Sparse representation [CVPR'16]



Volumetric representation [CVPR'17]

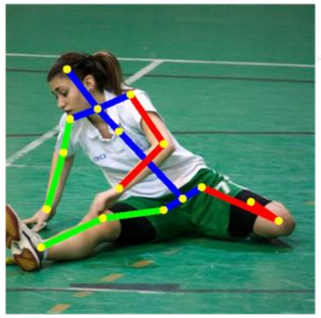


SMPL model [CVPR'18]

Weak supervisions

Multiple Views [CVPR'17]

Ordinal Depth [CVPR'18]



Z(left knee) > Z(right knee)
Z(right elbow) > Z(right wrist)
Z(left shoulder) < Z(right shoulder)
Z(right knee) < Z(left hip)
Z(left wrist) = Z(left elbow)
Z(head) > Z(right ankle)
Z(right hip) = Z(left hip)
Z(right ankle) < Z(neck)
Z(left wrist) < Z(left ankle)

Mirror symmetry

Main challenge: how to address the lack of 3D training data?

Single-view pose estimation

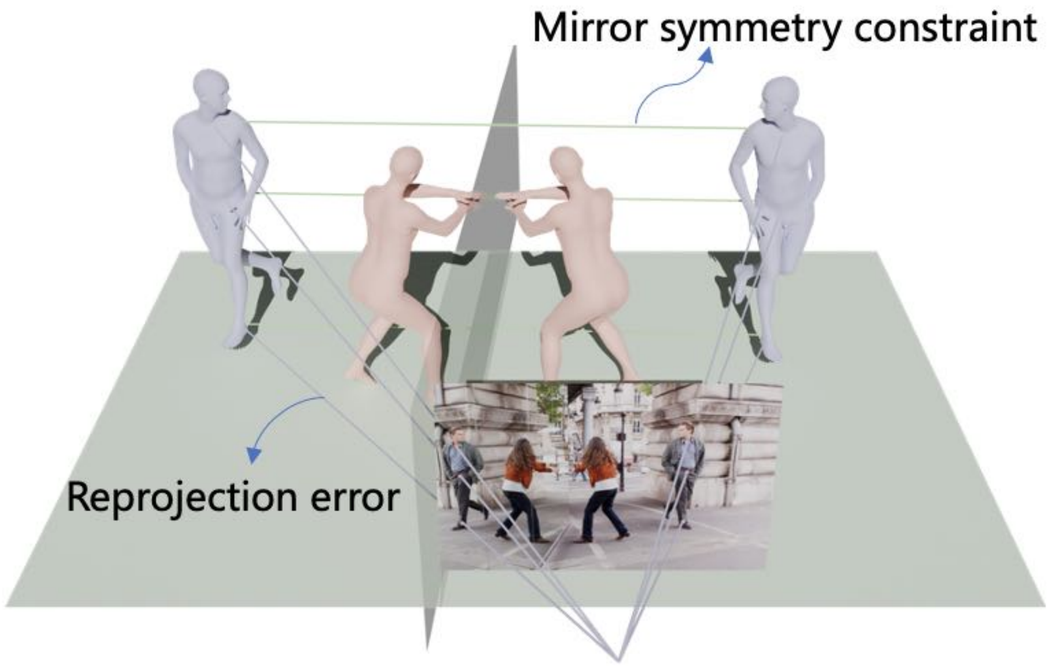
Reconstructing 3D pose estimation from mirrored human images

- Provide an additional virtual view
- Observe unseen part of the person

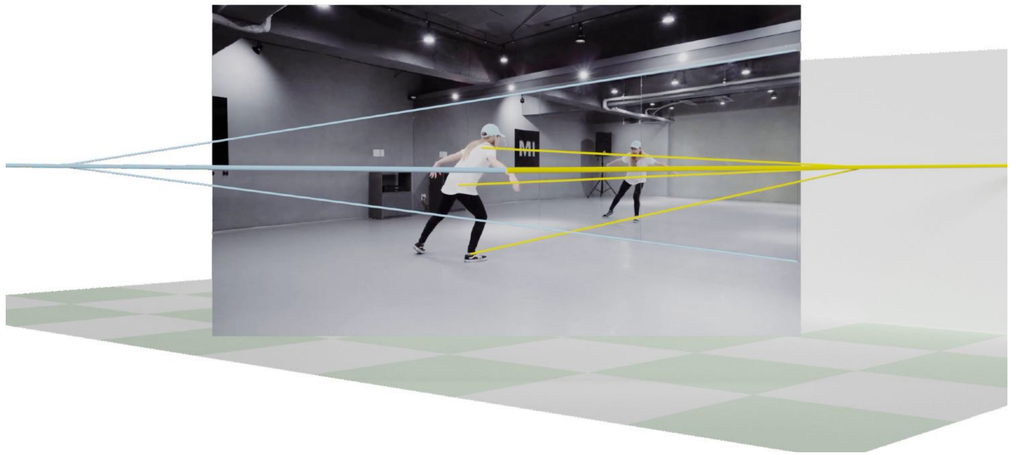
Reconstructing 3D Human Pose by Watching Humans in the Mirror. Under review.

Single-view pose estimation

Learning 3D pose estimation from mirrored human images

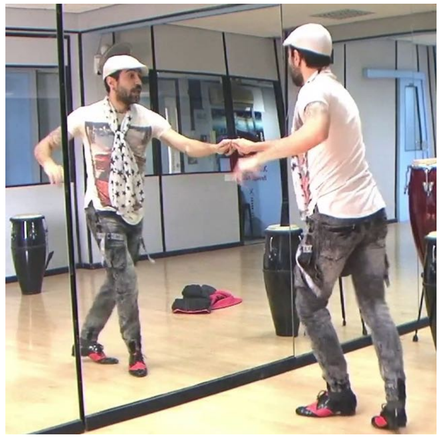


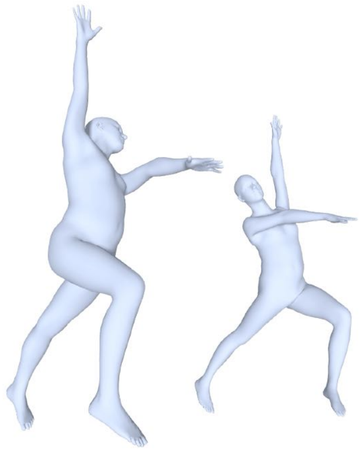
Optimize 3D poses with
mirror symmetry constraints

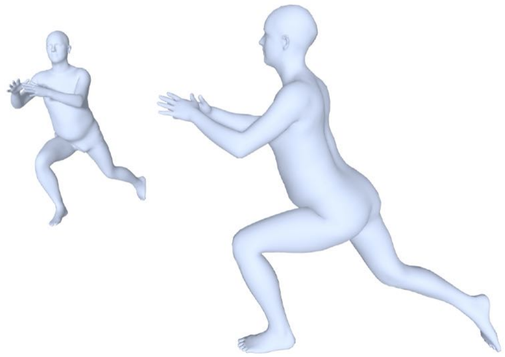


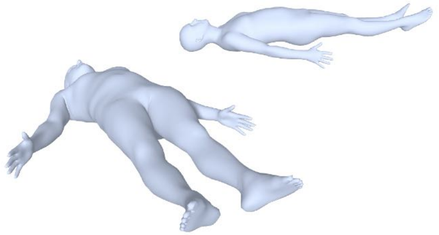
Estimate the mirror geometry
using vanishing points

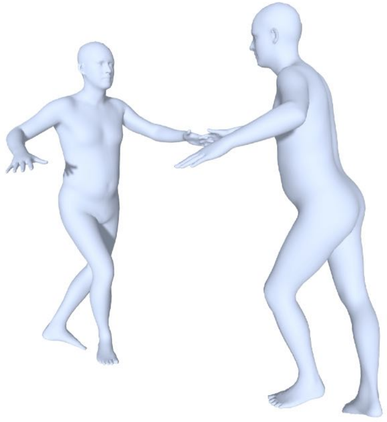
Single-view pose estimation











Single-view pose estimation

VIBE, CVPR20

Ours

Single-view pose estimation

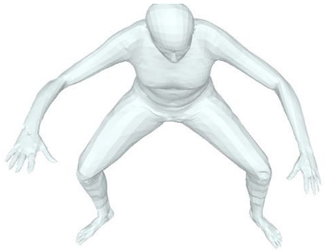
Our Mirrored-Human Dataset

Single-view pose estimation

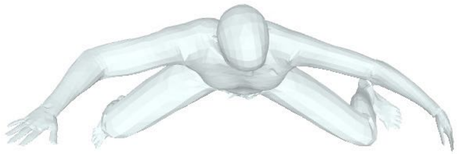
Mirrored-Human Dataset

Single-view pose estimation

Train existing single-view methods on Mirrored-Human



MeshNet



MeshNet trained on Mirrored-Human

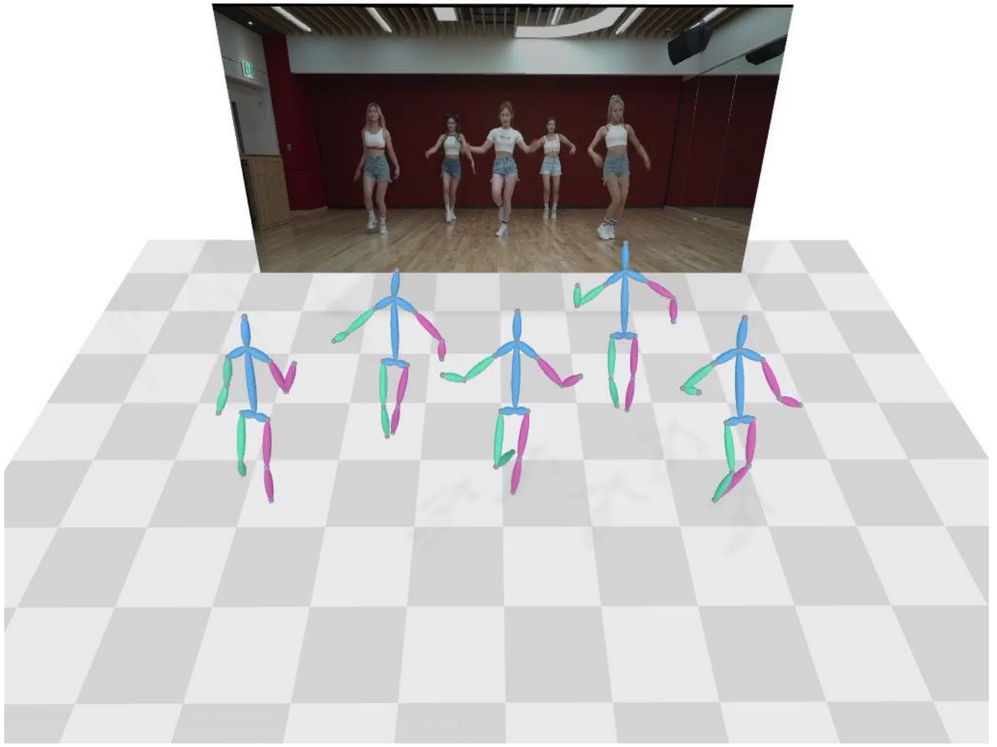
Methods	3DPW		H3.6M	
	MPJPE ↓	PA-MPJPE ↓	MPJPE ↓	PA-MPJPE ↓
HMR [19]	-	81.3	88.0	56.8
HMMR [20]	-	72.6	-	56.9
Arnab. [3]	-	72.2	77.8	54.3
CMR [23]	-	70.2	-	50.1
SPIN [22]	98.2*	59.2	62.3*	41.1
MeshNet [33]	93.2	58.6	55.7	41.7
Baseline	90.0	57.5	54.7	41.7
[33]+MiHu	**85.1**	**54.8**	**53.6**	**41.0**

Meshnet: Image-to-pixel prediction network for accurate 3d human pose and mesh estimation from a single rgb image, ECCV2020

Multi-person pose estimation

How to estimate 3D poses of multiple people from a single image?

CVPR 2020

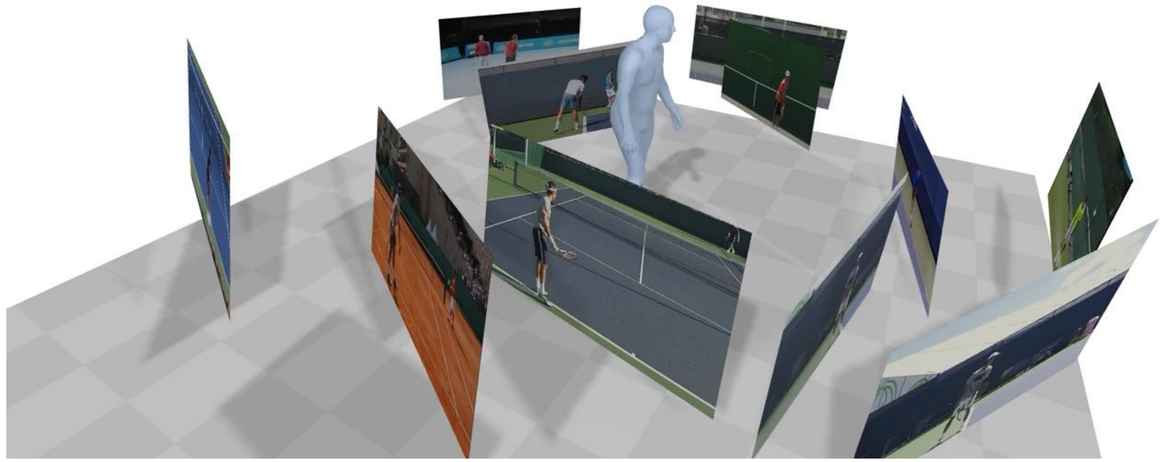


ECCV 2020

Coherent Reconstruction of Multiple Humans from a Single Image. CVPR 2020.

SMAP: Single-Shot Multi-Person Absolute 3D Pose Estimation. ECCV 2020.

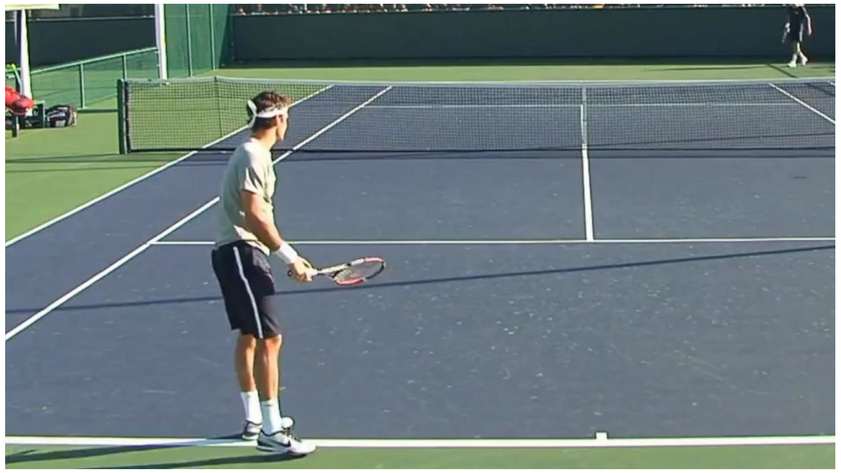
Part III: MoCap from Internet videos



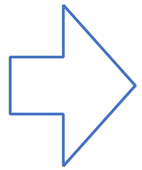
MoCap from Internet videos

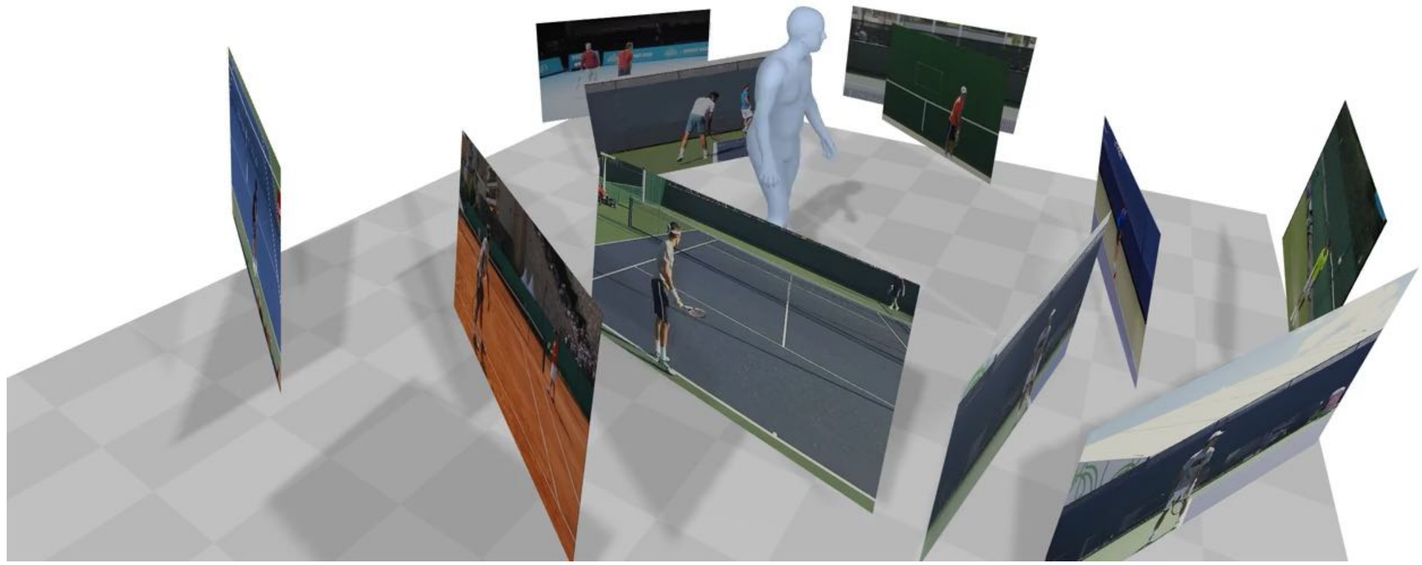
Though single-view estimation is good, it is NOT accurate enough because of

depth ambiguity & self-occlusion



MoCap from Internet videos





Motion Capture from Internet Videos. ECCV 2020.

MoCap from Internet videos

New challenges:

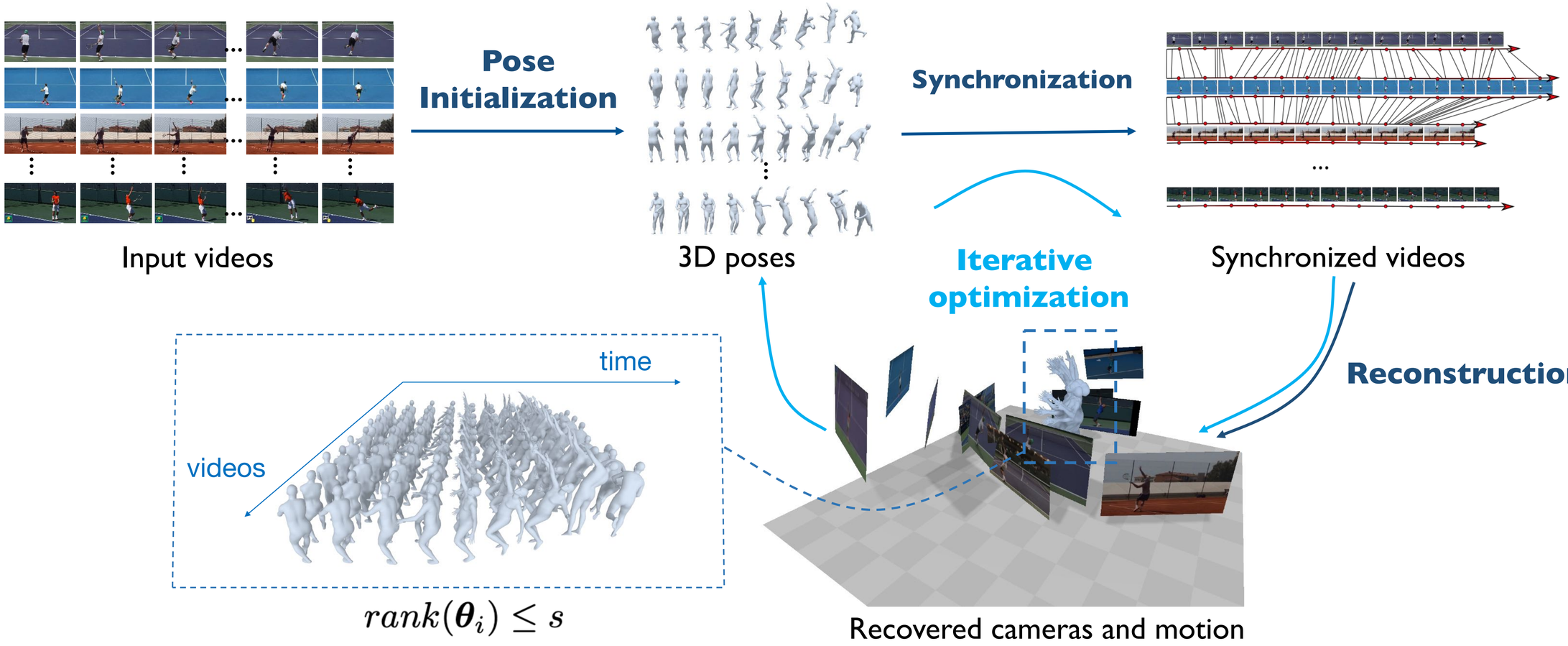
- Videos are unsynchronized
- Camera parameters are unknown
- Motions are not exactly the same

All previous multi-view reconstruction methods are inapplicable

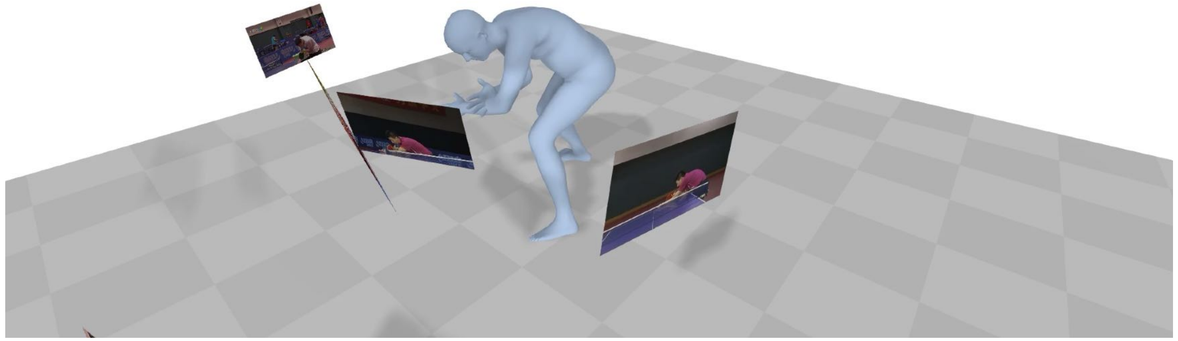
MoCap from Internet videos

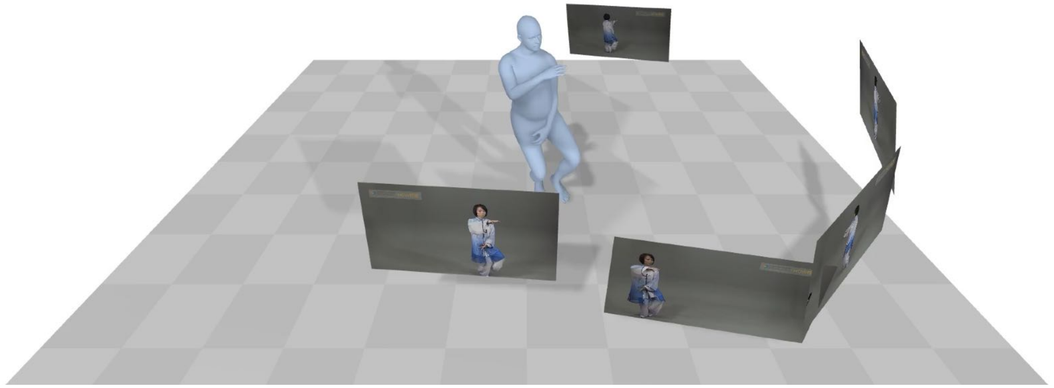
Proposed approach

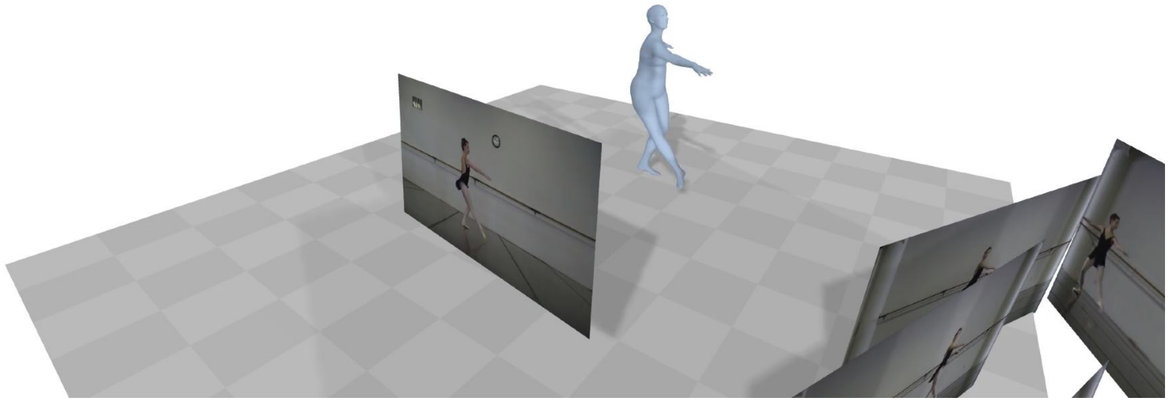
- Joint optimization of synchronizaiton, camera parameters and human motion parameters
- Model motion variation among videos by low-rank approximation



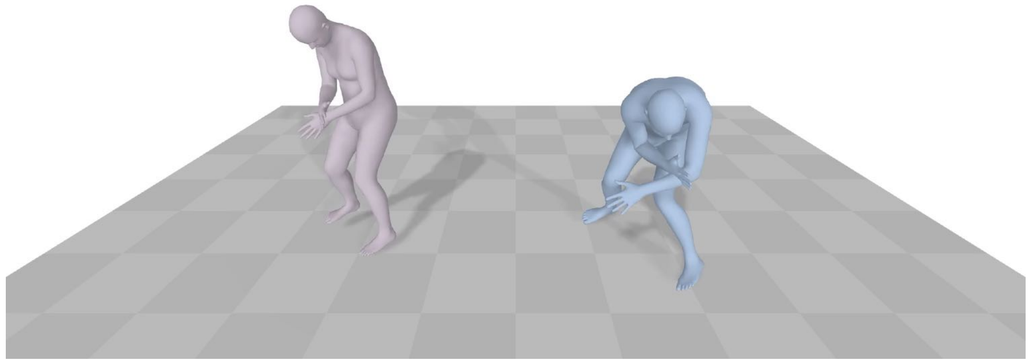
MoCap from Internet videos

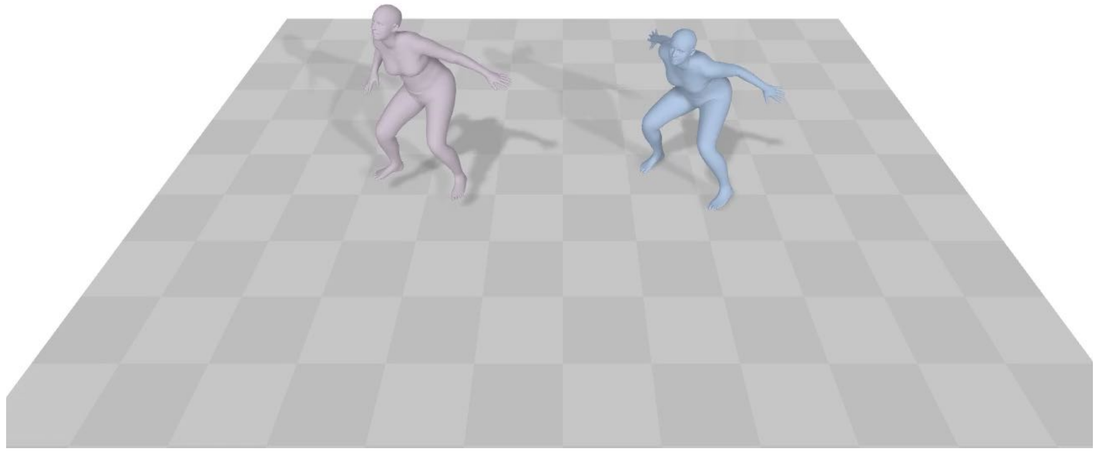






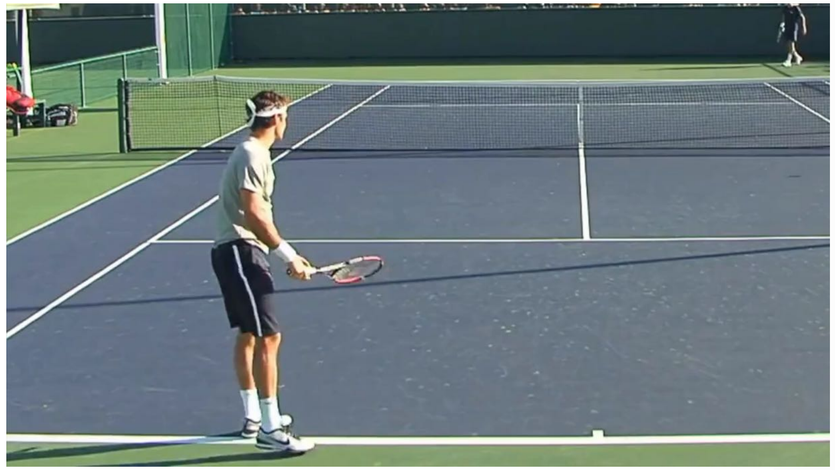
MoCap from Internet videos





单目方法　　我们的方法

MoCap from Internet videos



Part IV: Novel view synthesis for dynamic humans

Novel view synthesis

Free-viewpoint video (bullet time):

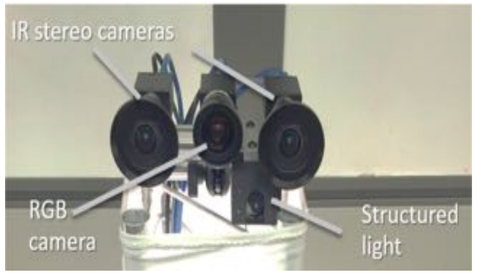
Novel view synthesis

Traidtional methods

Image-based rendering

Requiring dense camera array
Limited viewpoint range

Model-based rendering



Relying on reconstruction quality
Complicated hardware, constrained environments

Novel view synthesis

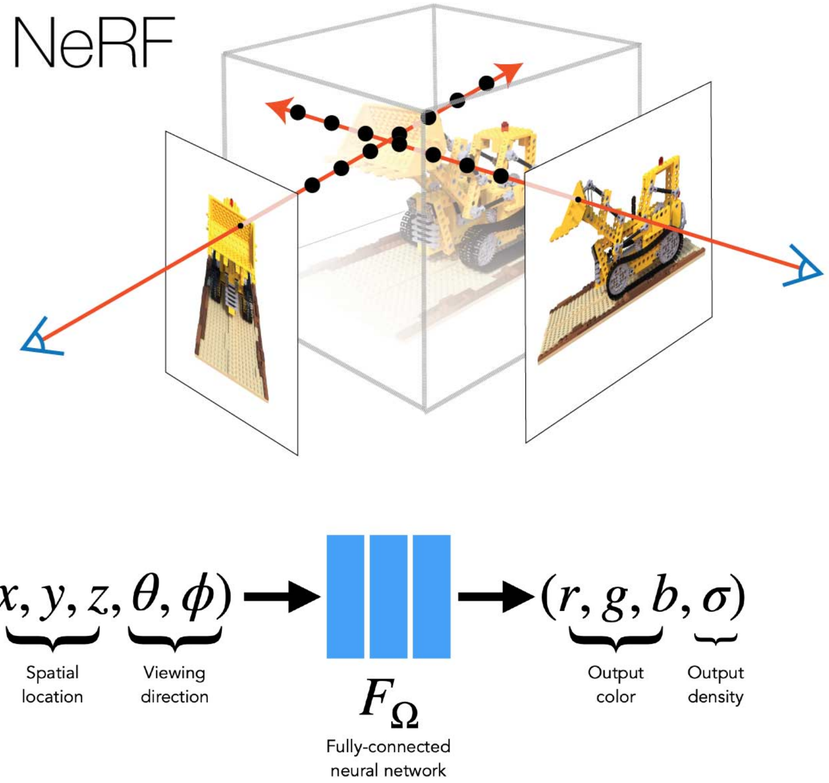
Can we generate free-viewpoint video using few RGB cameras?

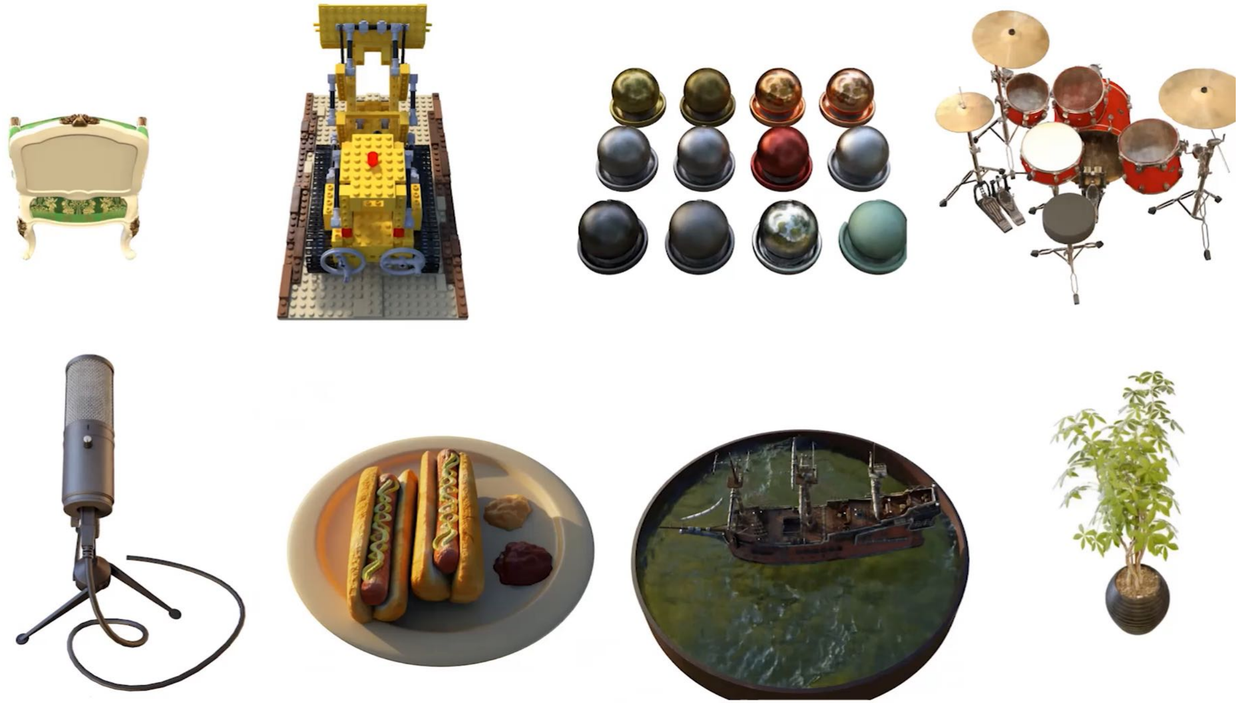
4-view video

Free-view video (bullet time)

Novel view synthesis

Recent trend: neural radiance fields (Nerf)



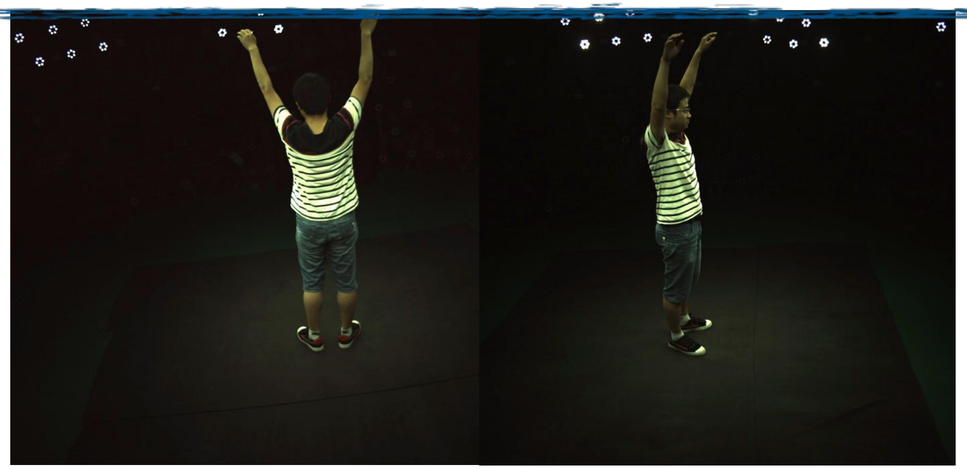


Mildenhall, Ben, et al. Nerf: Representing scenes as neural radiance fields for view synthesis. In ECCV, 2020.

Novel view synthesis

But it is ill-posed to learn the radiance fields from very sparse input views

Key idea: Integrate observations across video frames

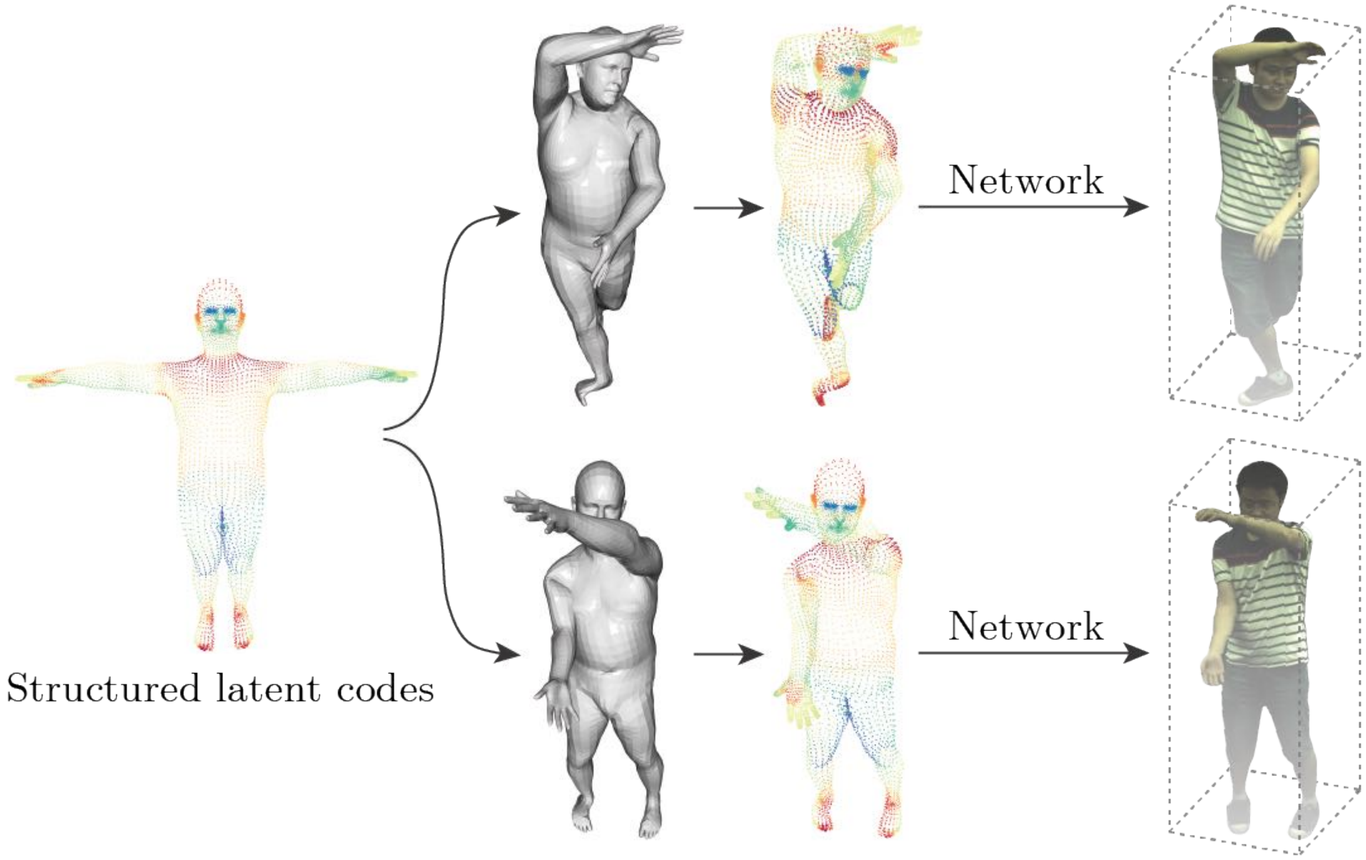


Four input views

Novel view synthesis by NeRF

Novel view synthesis

Suppose the radiance field is decoded from a set of structured latent codes



Neural Body: Implicit Neural Representations with Structured Latent Codes for Novel View Synthesis of Dynamic Humans. Under review.

Novel view synthesis

4-view video | NeRF [2] | Neural Volumes [1] | OURS

[1] Lombardi, Stephen, et al. Neural volumes: Learning dynamic renderable volumes from images. In SIGGRAPH, 2019.
[2] Mildenhall, Ben, et al. Nerf: Representing scenes as neural radiance fields for view synthesis. In ECCV, 2020.

Novel view synthesis

PIFuHD [3] OURS PIFuHD [3] OURS

[3] Saito, Shunsuke, et al. PIFuHD: Multi-Level Pixel-Aligned Implicit Function for High-Resolution 3D Human Digitization. In CVPR, 2020.

Novel view synthesis

Novel view synthesis from a monocular video

Input video

People-Snapshot [1]

OURS

[1] Alldieck, Thiemo, et al. Video based reconstruction of 3d people models. In CVPR, 2018

Novel view synthesis

Reconstruction results

Input video

People-Snapshot [1]

OURS

[1] Alldieck, Thiemo, et al. Video based reconstruction of 3d people models. In CVPR, 2018

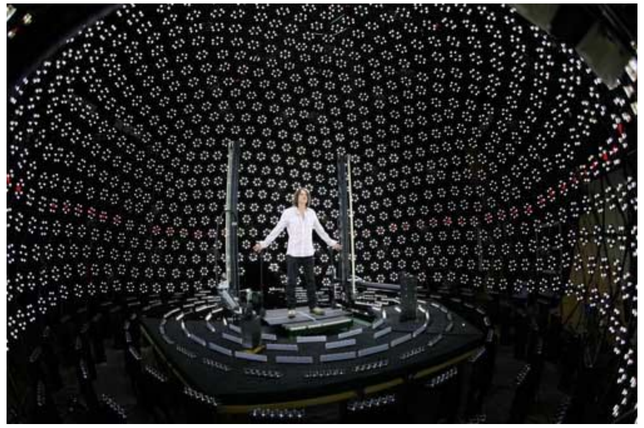
Novel view synthesis

Free-viewpoint video from only 4 cameras

Conclusion

What makes it possible?

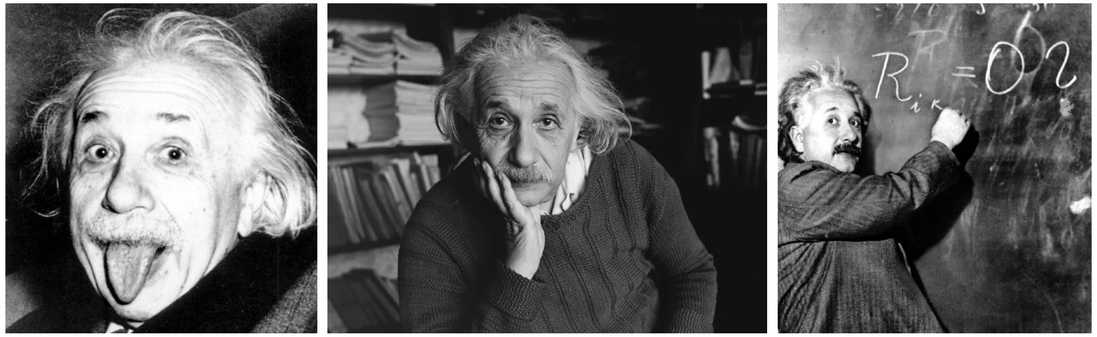
- More power tools
  - Deep learning
  - New representation
  - Differentiable rendering
- More 3D data …



Low-cost and easy-to-use capture systems

More challenges:

Large-scale & crowd scene



Reconstruction from historical data?

Acknowledgements


Sida Peng


Junting Dong


Qing Shuai


Qi Fang


Jianan Zhen


Yuanqing Zhang


Wen Jiang

Personal website: http://xzhou.me/

Github page: https://github.com/zju3dv