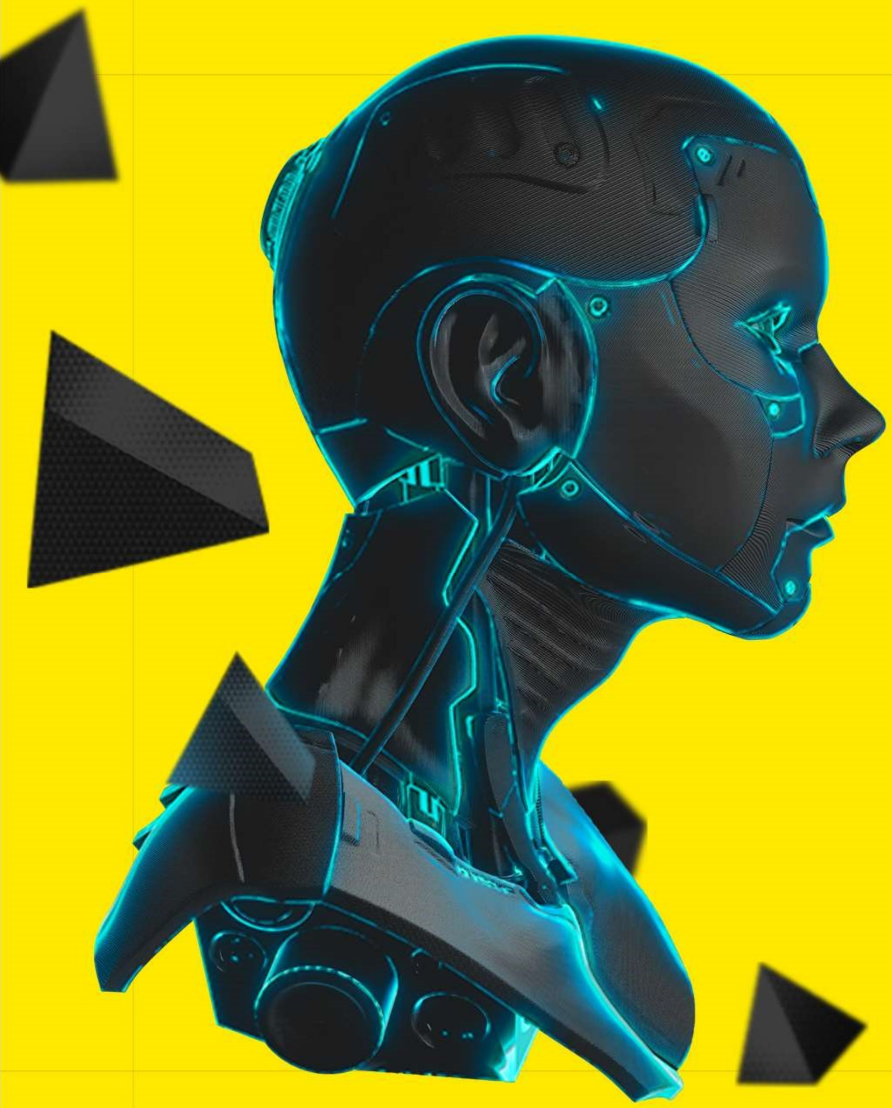


DL-BASED RECONSTRUCTION OF INTERACTION MOTIONS

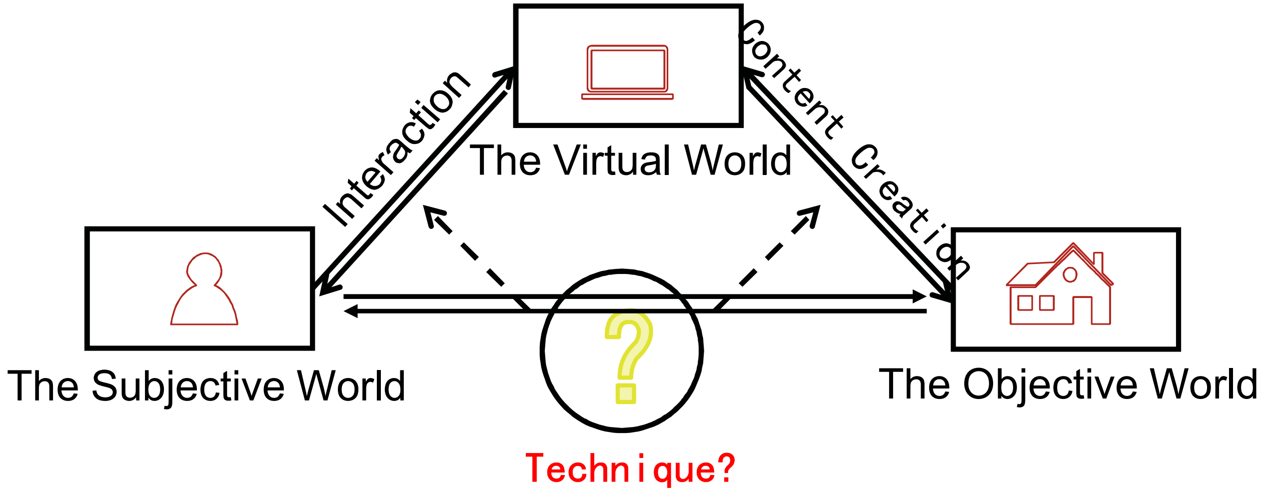
FENG XU

http://xufeng.site

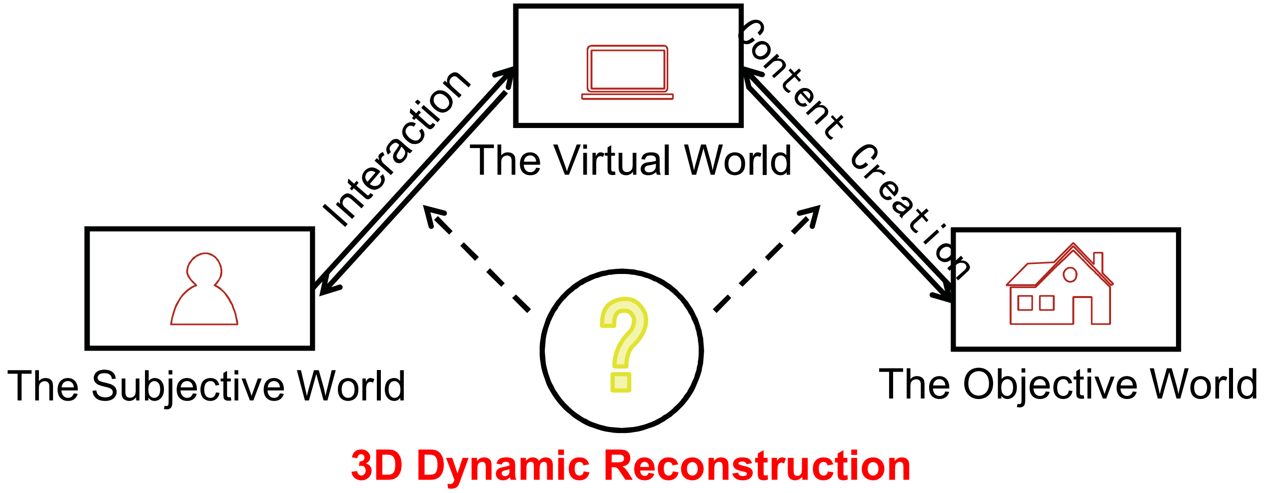
ASSOCIATE PROFESSOR, TSINGHUA

BACKGROUND

Tsinghua University



- We construct a virtual world to connect our subjective world to the objective world
- Interaction and content creation are two key issues in it

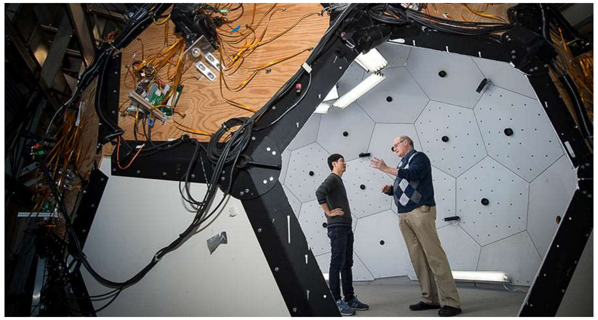


- We construct a virtual world to connect our subjective world to the objective world
- Interaction and content creation are two key issues in it

t1

CONTENT CREATION

Tsinghua University



CMU Dome

facebook 360

USC Dome

Tsinghua Dome

- ✓ High quality
- ✕ Heavy
- ✕ Complex
- ✕ Expensive

Slide 4

t1 think, 2020/11/21

Tsinghua University

Handle

Data Glove

Natural Interactions

Motion Sensor

- Heavy
  - Fixed mode
- Not Natural
- Multiple equipment

INTERACTION TECHNIQUES

清華大學
Tsinghua University

- Vision-based Interactions

Gaze

Hand

Body

- Traditional ways:

Hardware: Handle&Glove | Mouse | Marker-based

- New ways:

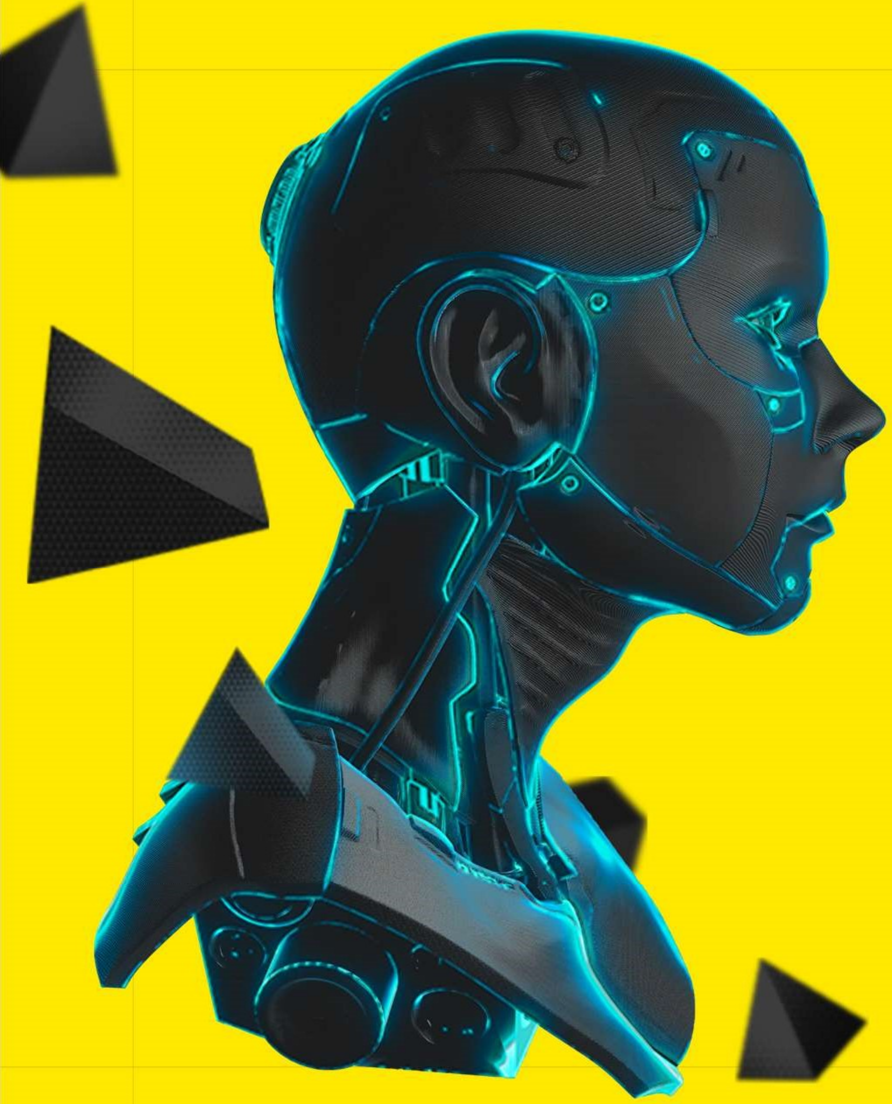
Software: Hand Recon. | Gaze Tracking | Markerless

- 3D Dynamic Reconstruction

Tasks:

Face/Gaze Recon.	Hand Recon.	Body Recon.

Goals: Use consumer Equipment to achieve real-time but high quality dynamic reconstructions of human



SIGGRAPH 2021

清華大學
Tsinghua University

SINGLE DEPTH VIEW BASED REAL-TIME RECONSTRUCTION OF HAND-OBJECT INTERACTIONS

HAO ZHANG, YUXIAO ZHOU, YIFEI TIAN, JUN-HAI YONG, and FENG XU*

BNRist and School of Software, Tsinghua University

THE PREMIER **CONFERENCE** & **EXHIBITION** IN
COMPUTER GRAPHICS & INTERACTIVE TECHNIQUES

Tsinghua University

- 3D reconstruction of hand-object interactions has many applications


[Liu et al. 2018]

Animation

[Andrychowicz et al. 2020]

Robotics

[3D Perception Lab 2019]

VR

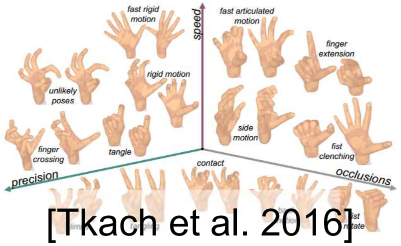
Tsinghua University

- 3D reconstruction of hand-object interactions is very challenging

lack of features

[Tagliasacchi et al. 2015]

complex motions



[Tkach et al. 2016]

Hand → +Object

high dimension/diversity

[Tzionas et al. 2016]

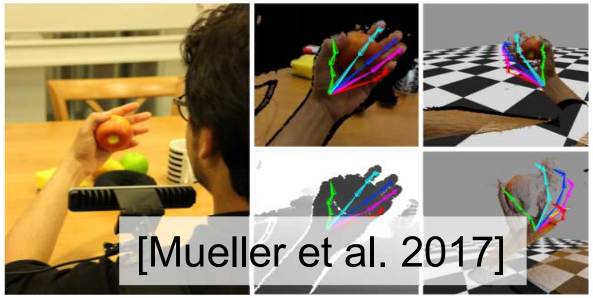
heavy occlusions

[Panteleris et al. 2015]

BACKGROUND

Tsinghua University

- Current methods have some limitations

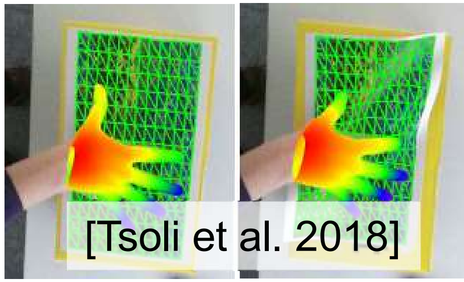

[Mueller et al. 2017]


[Petit et al. 2018]


[Tzionas et al.2016]


[Zhou et al. 2020]


[Wang et al. 2019]


[Tsoli et al. 2018]


[Zhang et al. 2019]

No Object

No Hand

Offline + Template

Two Sensors
+ Calibration

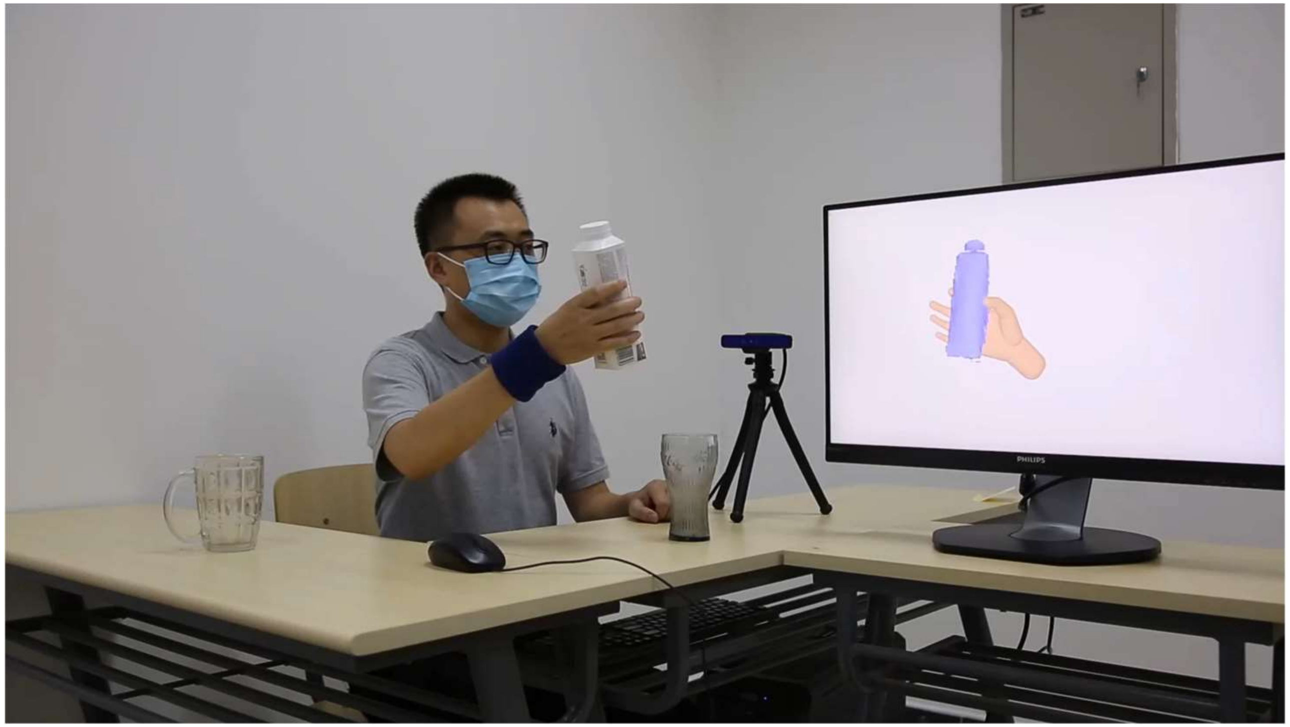
OUR WORK

Tsinghua University

- Reconstruct 3D hand-object interactions in real-time with a single depth camera

Object Model

Hand Pose

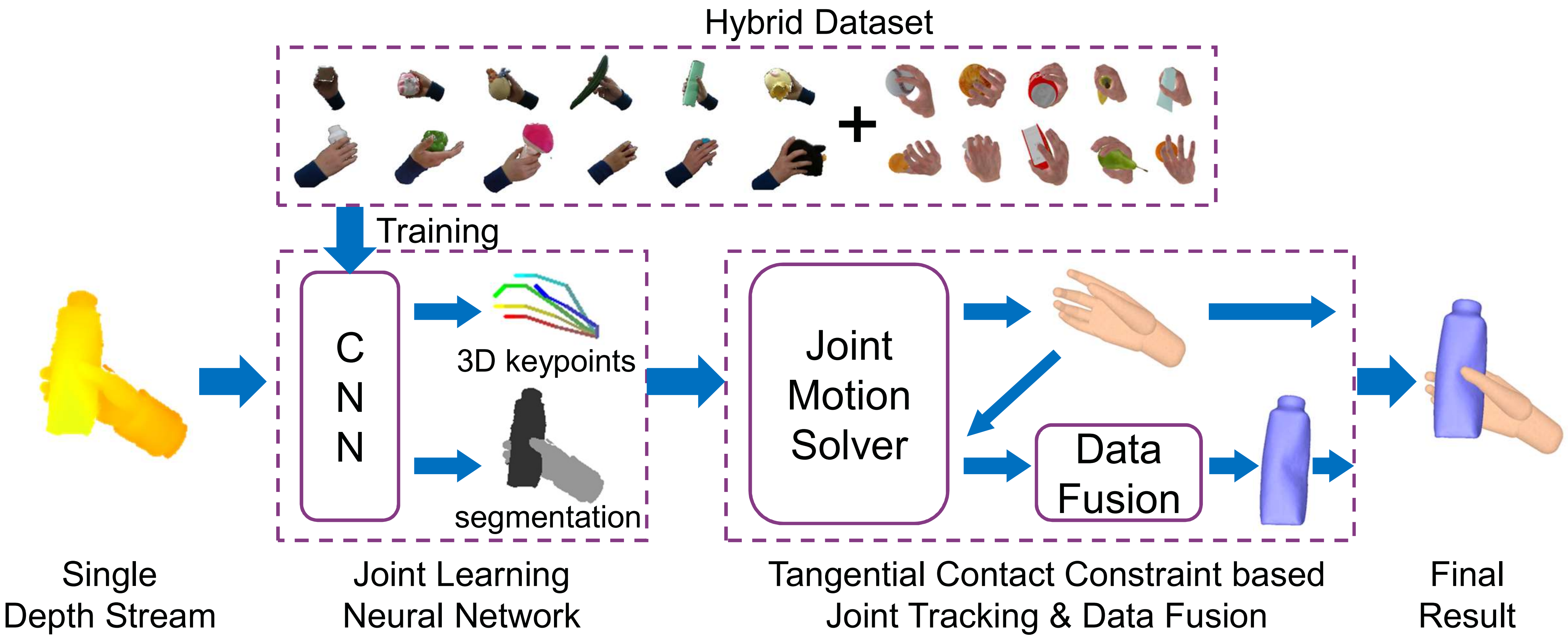


Rigid/Nonrigid Motion

Camera Move

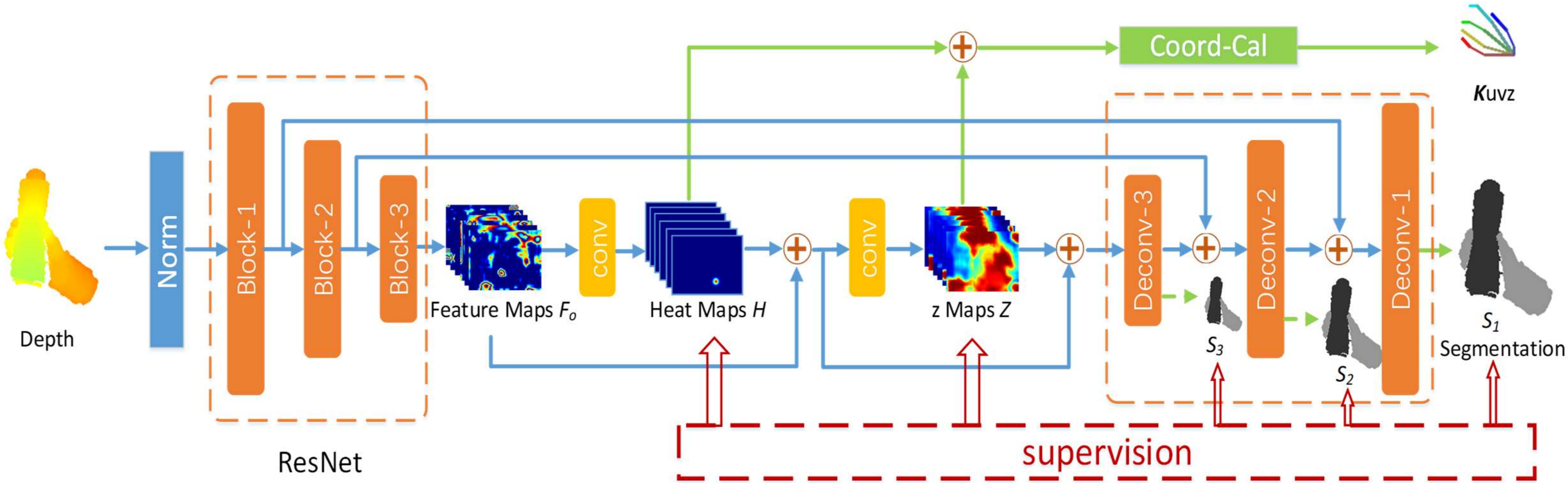
OVERVIEW

Tsinghua University



JOINT LEARNING NEURAL NETWORK

清華大學 Tsinghua University



HYBRID DATASET

Tsinghua University

Hybrid Dataset

Real Dataset

Interaction Reco.
By [Zhang et al. 2019]

+

Manually
Selection

Synthetic Dataset

C
L
A
P

HYBRID DATASET

Tsinghua University

Details of the Hybrid Dataset

	Frames	Objects	Hand	Views	Motion Type	Motion Range(mm)
Real Dataset	6703	**Number:** 12 objects **Shapes:** cube, sphere like, cylinder like, and other shapes **Size:** 4 to 22 cm **Materials:** plastic, cloth, wood, and paper	1 real hand	2 side viewpoints (vps)	Grab, hold, pinch, support, and large move	L-R: [–110 , 81] U-D: [–41, 78] N-F: [258, 510]
Synthetic Dataset	58764	**Number:** 13 objects **Shapes:** sphere, cylinder, cuboid, and other shapes **Size:** 4 to 17 cm	1 hand model	5 vps (2 side vps, 2 up-down vps, 1 frontal vps)	Grab, hold, pinch, support, and large move	L-R: [–267, 267] U-D: [–306, 219] N-F: [294, 906]

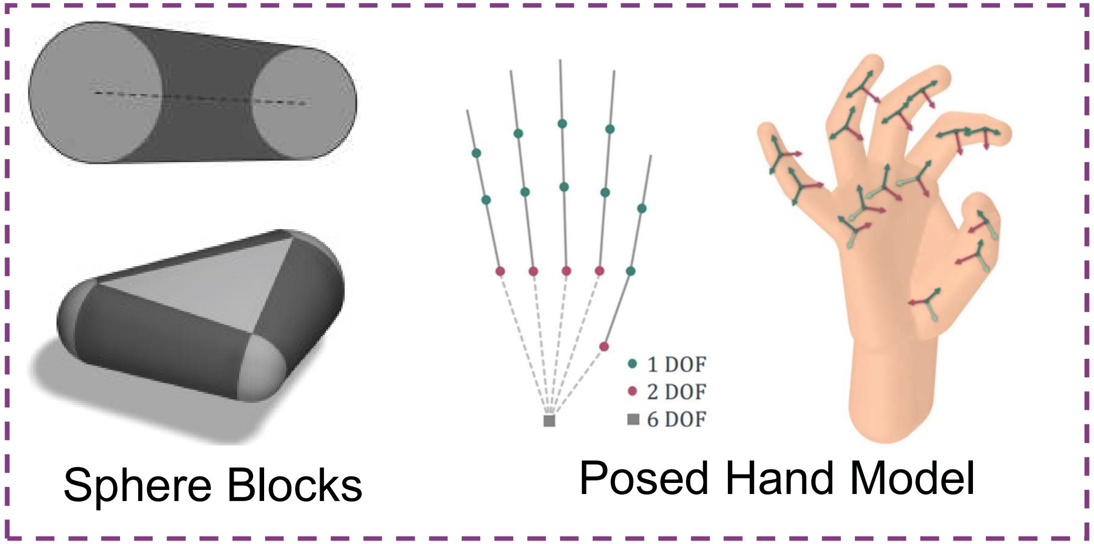
Our hybrid dataset has much diversity in object shape/size, interactive motions (pose and range)

JOINT TRACKING

Tsinghua University

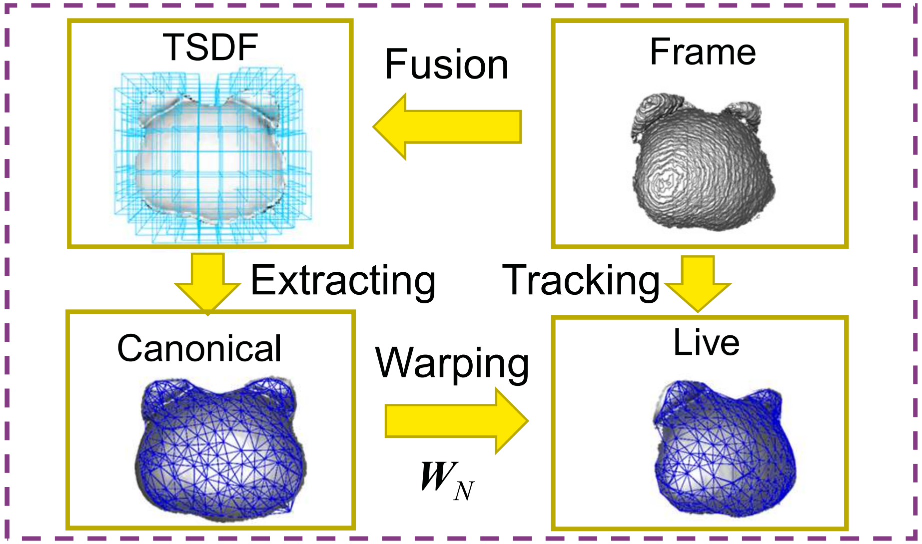
- Hand & Object modeling

[Tkach et al. 2016]



hand pose θ

[Newcombe et al. 2015]



warping field W_N

Tsinghua University 清華大學

- Total Energy

$$E_{\mathrm{tol}}(\mathcal{W}_N^t, \theta^t) = \omega_{\mathrm{uvz}} E_{\mathrm{uvz}}(\theta^t) + \omega_{\mathrm{tac}} E_{\mathrm{tac}}(\mathcal{W}_N^t) + E_{\mathrm{ori}}(\mathcal{W}_N^t, \theta^t)$$

3D Keypoints based Hand Tracking

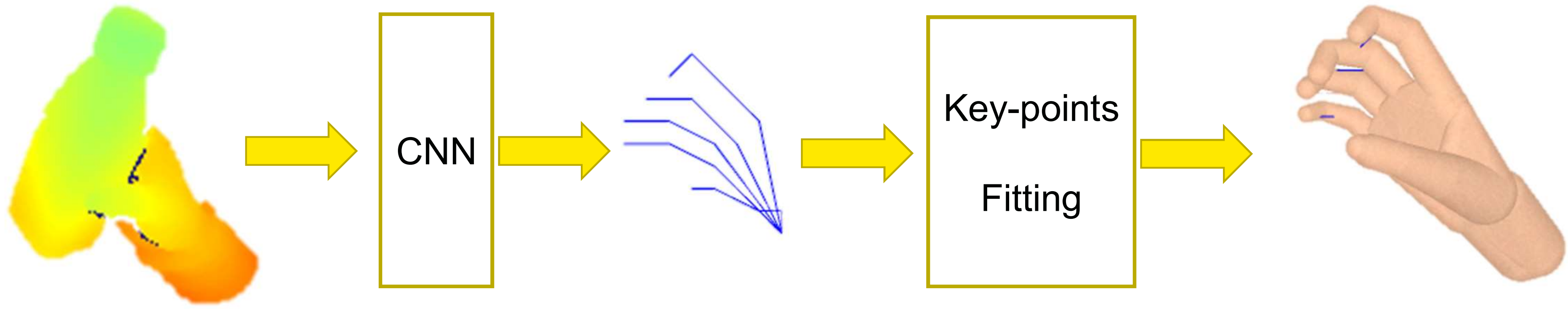
Tangential Contact Constraint based Object Tracking

[Zhang et al. 2019]

JOINT TRACKING

Tsinghua University

3D Keypoints based Hand Tracking



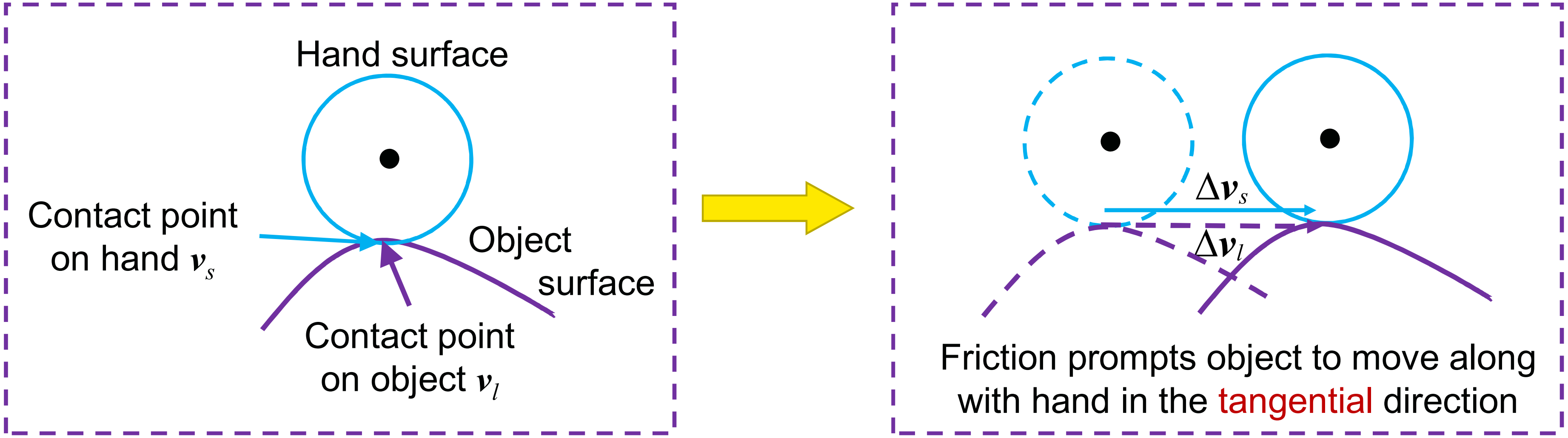
Single Depth

3D keypoints K_{uvz}

Reconstructed Hand

$$E_{uvz}(\boldsymbol{\theta}) = \left\| \boldsymbol{K}(\boldsymbol{\theta}) - \boldsymbol{K}_{uvz} \right\|_2^2$$

Tangential Contact Constraint based Object Tracking



$$E_{tac}(\boldsymbol{W}_N^t) = \sum_{(\boldsymbol{v}_l,\boldsymbol{v}_s)\in C_{tac}} \left\| P_{\mathrm{P}}[\boldsymbol{v}_l(\boldsymbol{W}_N^t) - \boldsymbol{v}_l(\boldsymbol{W}_N^{t-1})] - P_{\mathrm{P}}(\Delta\boldsymbol{v}_s)\right\|_2^2$$

Operation to extract tangential movement

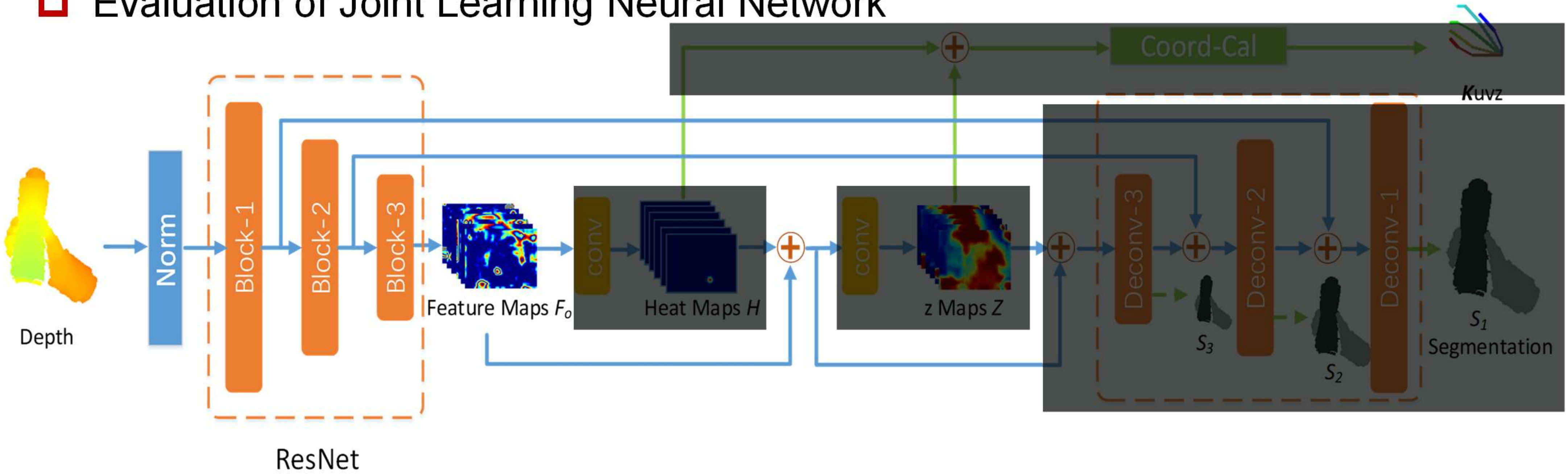
RESULTS

Evaluation of Joint Learning Neural Network

Hybrid Training Dataset		Real Test Dataset
Synthetic Dataset	Real Sub-Dataset 1	Real Sub-Dataset 2
		
13 Objects 58764 Frames	9 Objects 5249 Frames	5 Objects 1454 Frames

Tsinghua University

Evaluation of Joint Learning Neural Network



JL	KPT	SEG
3D key-points + segmentation	3D key-points prediction	segmentation prediction

- Evaluation of Joint Learning Neural Network

performances of the networks

Network	3D Error/mm	MIoU	Runtime	Trainable Var.
Our	13.1±11.2	0.943	20ms	15.49M
KPT	13.3±11.7	-	13ms	11.75M
SEG	-	0.947	17ms	14.06M

Joint learning neural network saves 1/3 runtime and half trainable variables without sacrificing accuracy

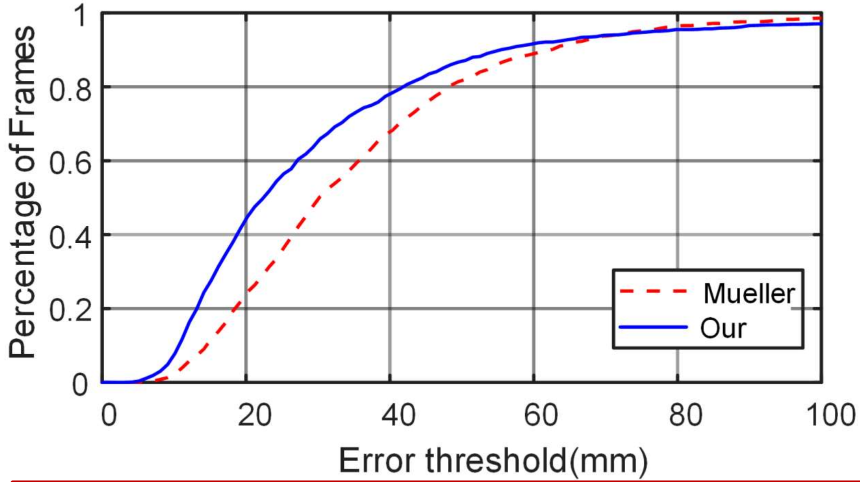
Evaluation of Joint Learning Neural Network

comparison with [Bo et al. 2020]
on hybrid dataset

	Our	[Bo et al. 2020]
MIoU	0.943	0.935
Runtime	20ms	25ms
Trainable Var.	15.49M	39.91M

Our network is slightly better on segmentation and much smaller

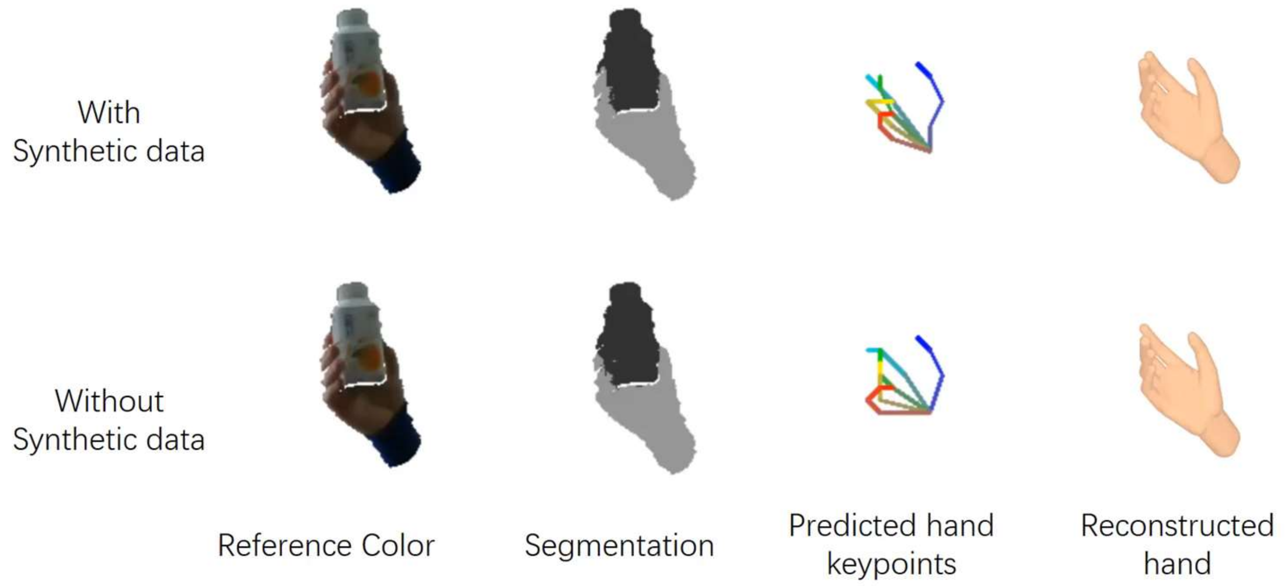
comparison with [Mueller et al. 2017]
on their dataset



Our network has comparable performance with [Mueller et al. 2017]

Evaluation of Synthetic Dataset

	3D Error/mm	MIoU
Without syn. Dataset	14.8±13.0	0.923
With syn. dataset	13.1±11.2	0.943

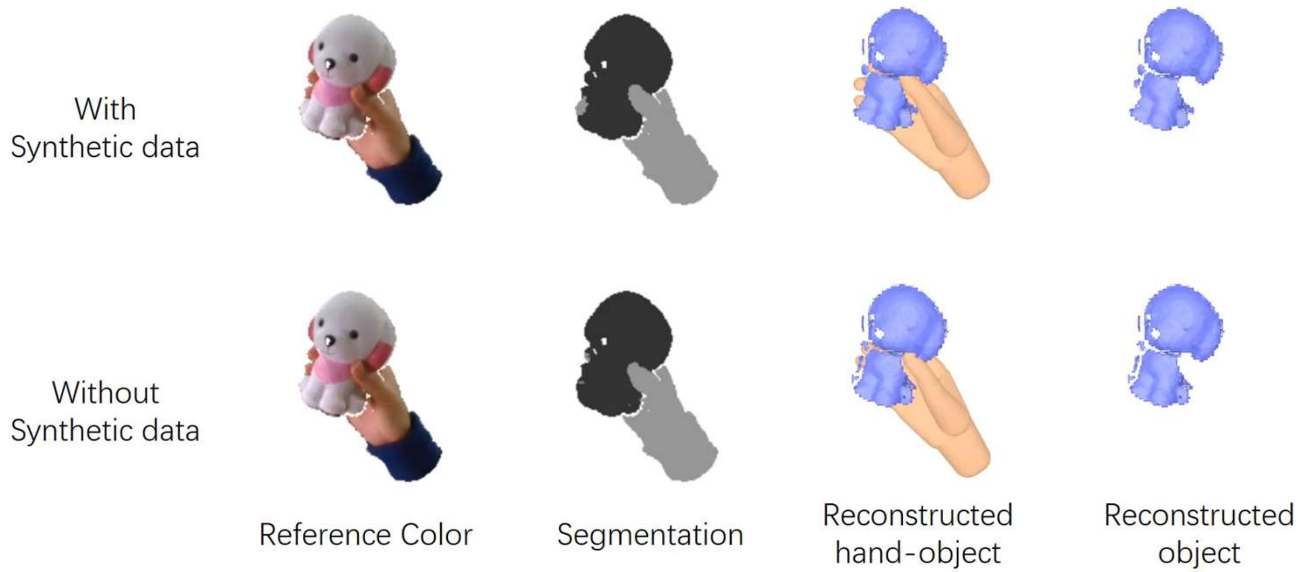


Synthetic dataset improves the performances of hand keypoints prediction and hand-object segmentation, resulting in better hand pose estimation.

RESULTS

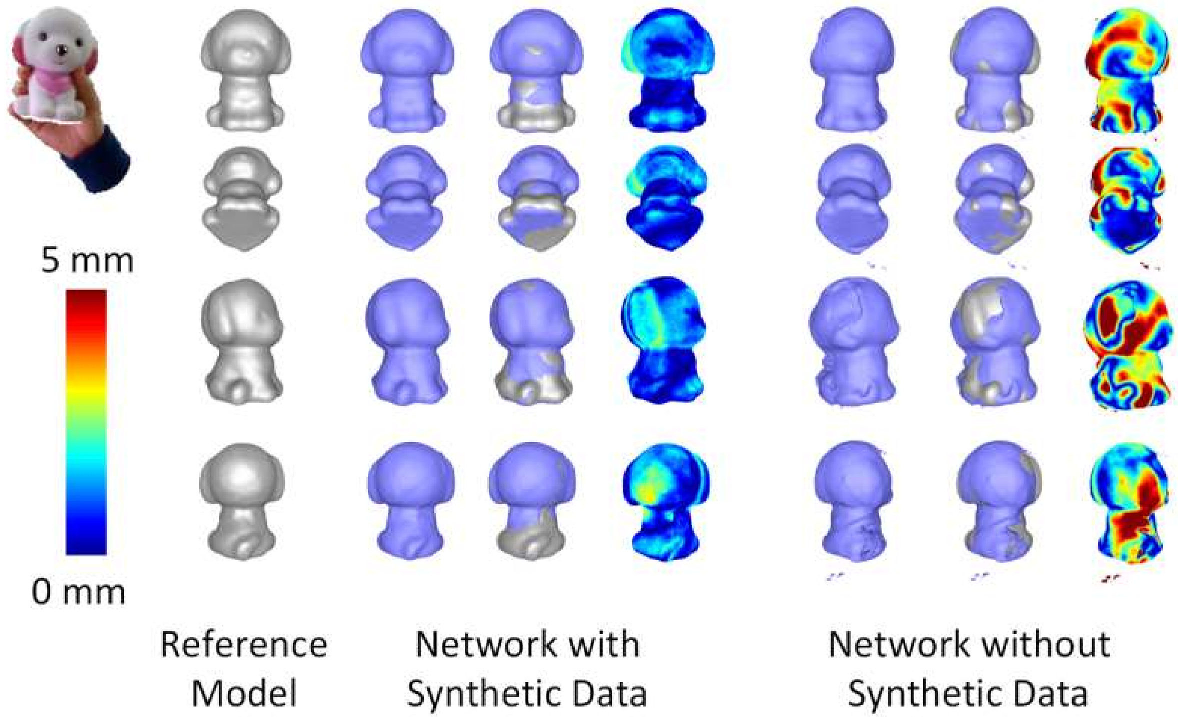
Evaluation of Synthetic Dataset

	3D Error/mm	MIoU
Without syn. Dataset	14.8±13.0	0.923
With syn. dataset	13.1±11.2	0.943



Synthetic dataset improves the performances of hand keypoints prediction and hand-object segmentation, resulting in better object reconstruction.

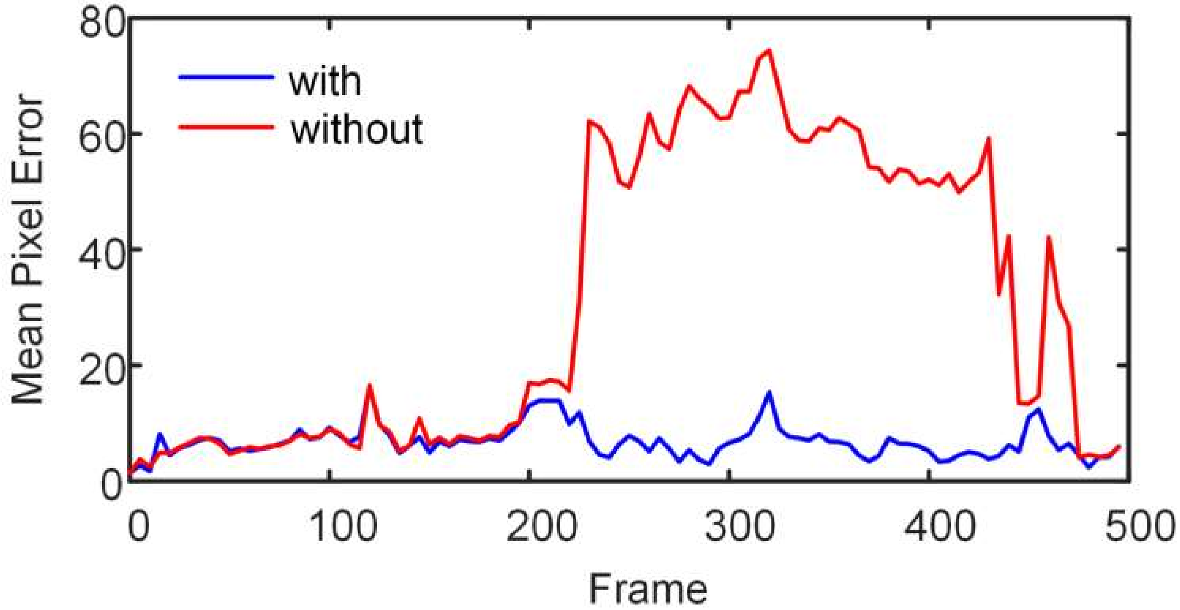
Evaluation of Synthetic Dataset



	Without syn. dataset	With syn. Dataset
Mean Distance	2.4 mm	1.0 mm

Synthetic dataset improves the model reconstruction of the object

- Evaluation of Tangential Contact Constraint



Tangential contact constraint improves the object tracking

RESULTS

Tsinghua University

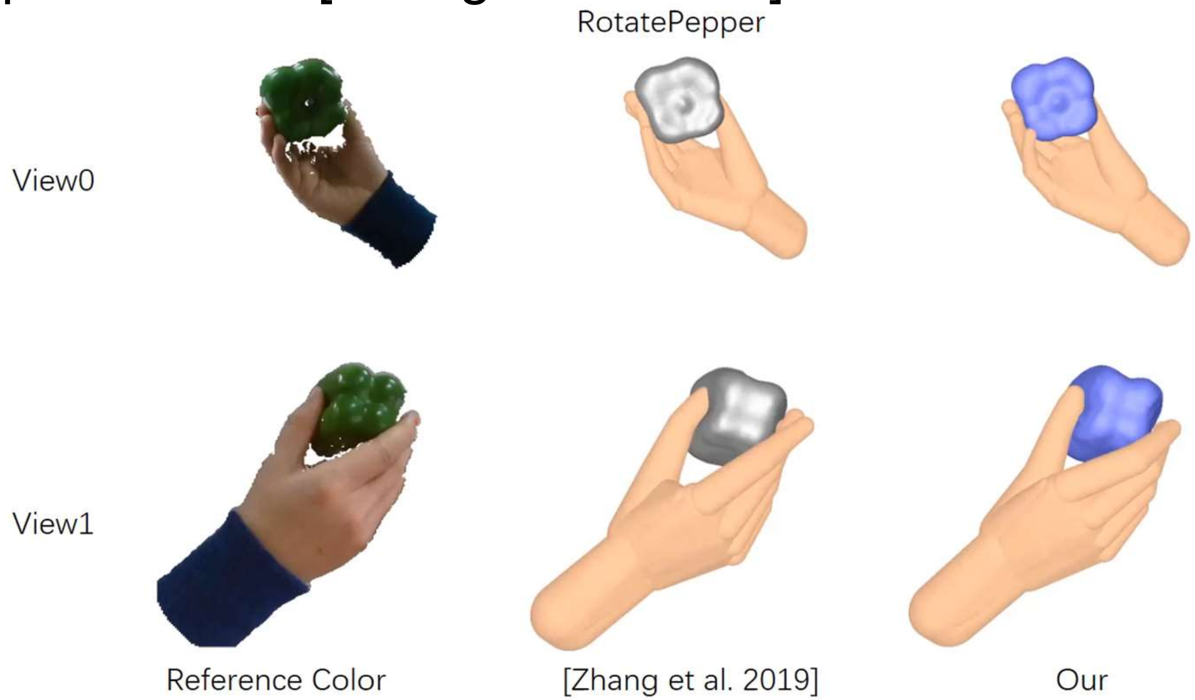
- Evaluation of Tangential Contact Constraint

Reference Color | without | with

Tangential contact constraint improves the model reconstruction of the object

RESULTS

- Comparison with [Zhang et al. 2019]



comparable results with [Zhang et al. 2019] in view0

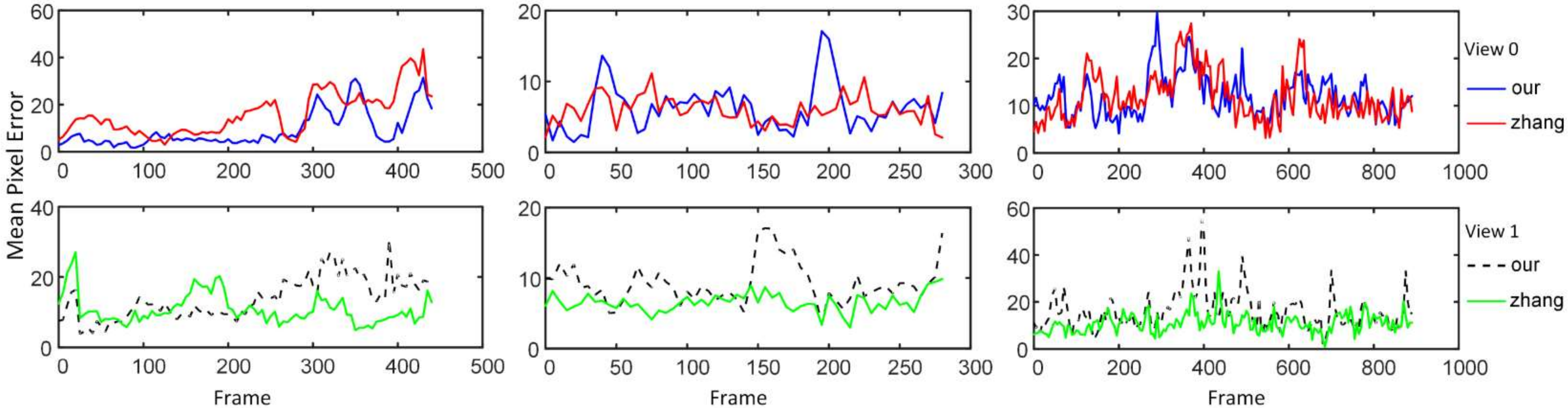
give reasonable results in view1

We only use the depth stream of view0

RESULTS

Tsinghua University

Comparison with [Zhang et al. 2019]



(a) RotatePepper (b) PourBottle (c) ReconstructCat

	View0		View1	
	Ours	Zhang et al.	Ours	Zhang et al.
RotatePepper	9.2	16.0	13.9	10.8
PourBottle	6.3	6.0	9.3	6.5
ReconstructCat	12.1	11.9	16.2	11.0

Our system achieves comparable hand tracking with [Zhang et al. 2019] in *View0*. For *View1*, we still give a reasonable result with satisfactory accuracy.

RESULTS

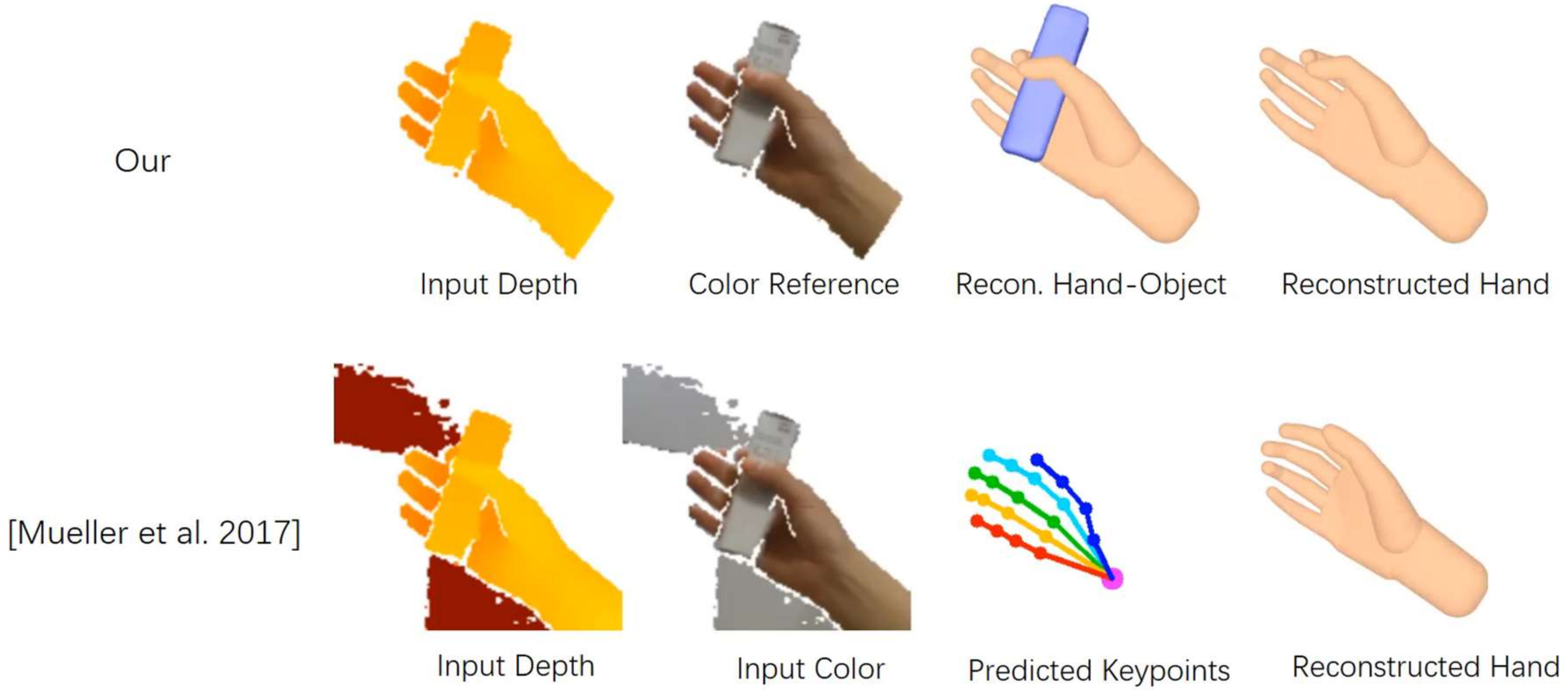
Comparison with [Zhou et al. 2020]

Our | [Zhou et al. 2020]

Input Depth | Ref. Color | Reco. Hand | Input Color | Reco. Hand

Our system gives **more robust and accurate hand poses** when interacting with a large object

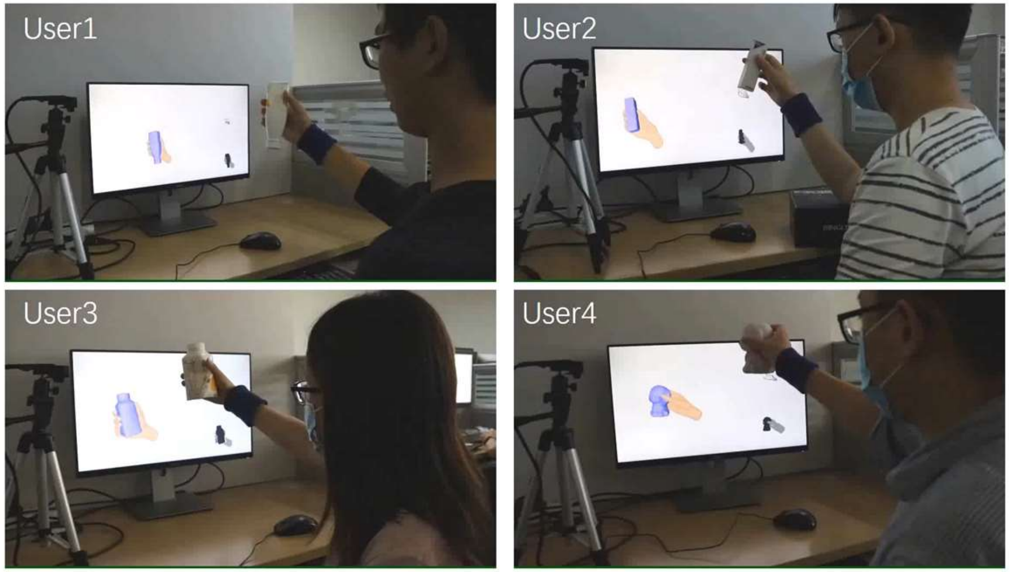
- Comparison with [Mueller et al. 2017]

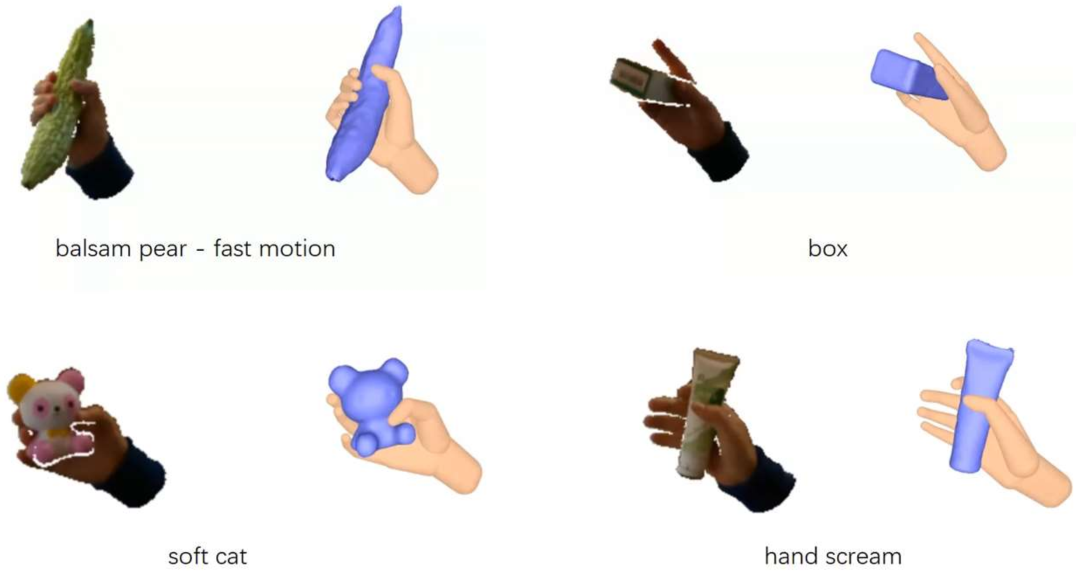


Our system gives more accurate hand poses and robust to light change

RESULTS

More Results





Our system can handle different users and different objects

RESULTS

Tsinghua University

More Results

Put Down and Pick Up

Deformation in Fusion

Object Moves In-Between the Fingers

Manipulate Object with Hole

Our system can handle challenging interactive motions

CONCLUSION

Tsinghua University

- Interaction reconstruction method with a single depth stream
- Comparable with two cameras based method
- A joint learning neural network, a hybrid dataset and a tangential contact constraint
- Robust to different users/objects, challenging interactive motions, light changes and camera moves

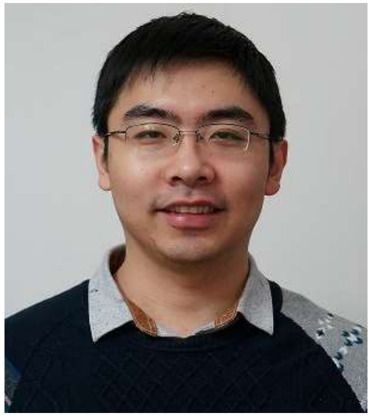
Q&A

Hao Zhang

Yuxiao Zhou

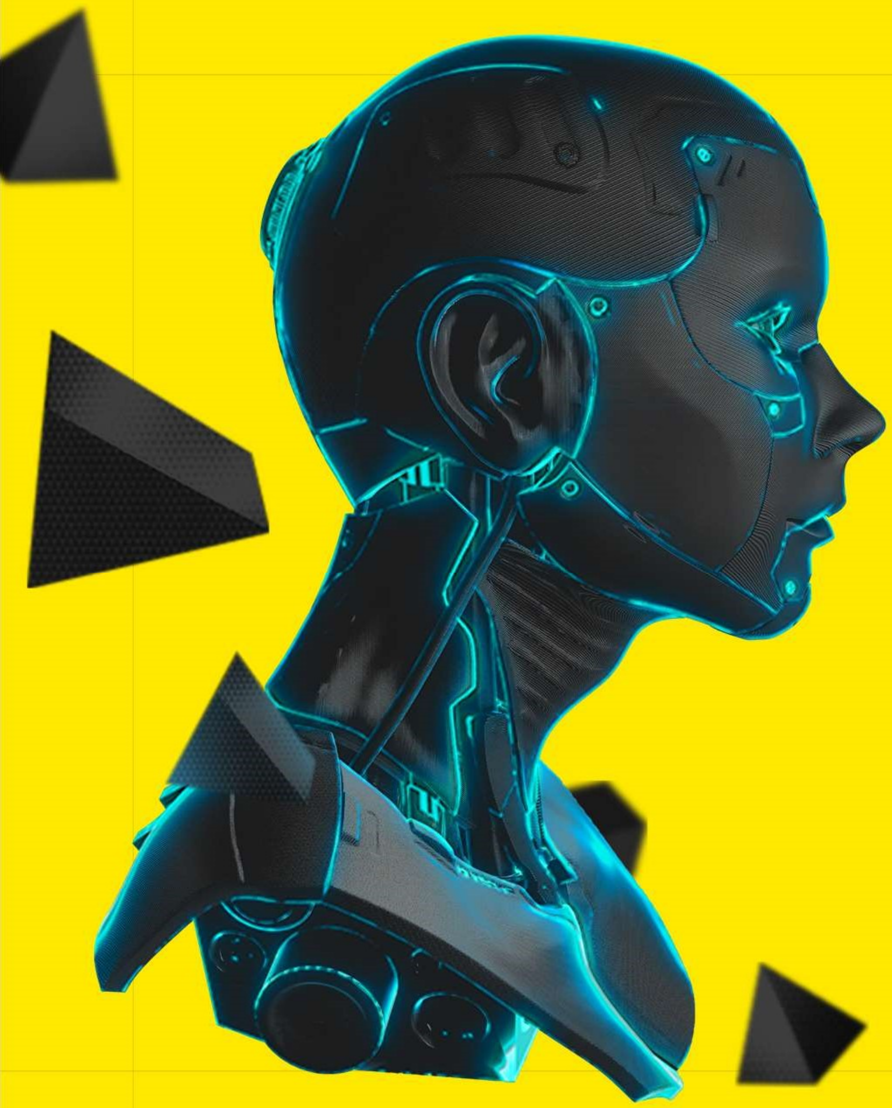
Yifei Tian

Jun-Hai Yong



Feng Xu*

Thanks for Your Attention!



SIGGRAPH 2021

清華大學
Tsinghua University

TRANSPOSE

REAL-TIME 3D HUMAN TRANSLATION AND POSE ESTIMATION WITH SIX INERTIAL SENSORS

XINYU YI, YUXIAO ZHOU, FENG XU
TSINGHUA UNIVERSITY

Paper

Project Page

THE PREMIER **CONFERENCE** & **EXHIBITION** IN
COMPUTER GRAPHICS & INTERACTIVE TECHNIQUES

LIVE DEMO


Inertial Measurement Unit Placement
the left and right lower leg, the head, and the pelvis.
Laptop Screen
Control
Root
Align
Camera
Free
Flag
Show
Use Translation

CONTRIBUTIONS

- Multi-stage body pose estimation from sparse sensors

CONTRIBUTIONS

清華大學
Tsinghua University

- Multi-stage body pose estimation from sparse sensors
  - Sensor measurements → leaf joint positions → full joint positions → pose params

CONTRIBUTIONS

- Multi-stage body pose estimation from sparse sensors
  - Sensor measurements → leaf joint positions → full joint positions → pose params
  - Better learns prior knowledge from MoCap data

CONTRIBUTIONS

Tsinghua University

- Multi-stage body pose estimation from sparse sensors
  - Sensor measurements → leaf joint positions → full joint positions → pose params
  - Better learns prior knowledge from MoCap data
  - State-of-the-art accuracy and smoothness

CONTRIBUTIONS

Tsinghua University

- Multi-stage body pose estimation from sparse sensors
  - Sensor measurements → leaf joint positions → full joint positions → pose params
  - Better learns prior knowledge from MoCap data
  - State-of-the-art accuracy and smoothness
- Fusion-based global translation estimation

CONTRIBUTIONS

清華大學
Tsinghua University

- Multi-stage body pose estimation from sparse sensors
  - Sensor measurements → leaf joint positions → full joint positions → pose params
  - Better learns prior knowledge from MoCap data
  - State-of-the-art accuracy and smoothness
- Fusion-based global translation estimation
  - A hybrid of physics rules and neural networks

CONTRIBUTIONS

- Multi-stage body pose estimation from sparse sensors
  - Sensor measurements → leaf joint positions → full joint positions → pose params
  - Better learns prior knowledge from MoCap data
  - State-of-the-art accuracy and smoothness
- Fusion-based global translation estimation
  - A hybrid of physics rules and neural networks
  - First real-time full motion capture using sparse IMUs

OUTLINE

清華大學 Tsinghua University

- Introduction
- Method
- Results

SIGGRAPH 2021

INTRODUCTION

BACKGROUND

https://www.pexels.com/photo/man-in-white-dress-shirt-wearing-black-and-white-vr-goggles-5303629/

HUMAN MOTION CAPTURE

- Human motion capture is widely used in
  - augmented reality / virtual reality / mixed reality
  - films / games / sports
- An ideal motion capture system should
  - lightweight / easy to use
  - nonintrusive
  - robust to changing environments
  - time efficient

PREVIOUS WORKS

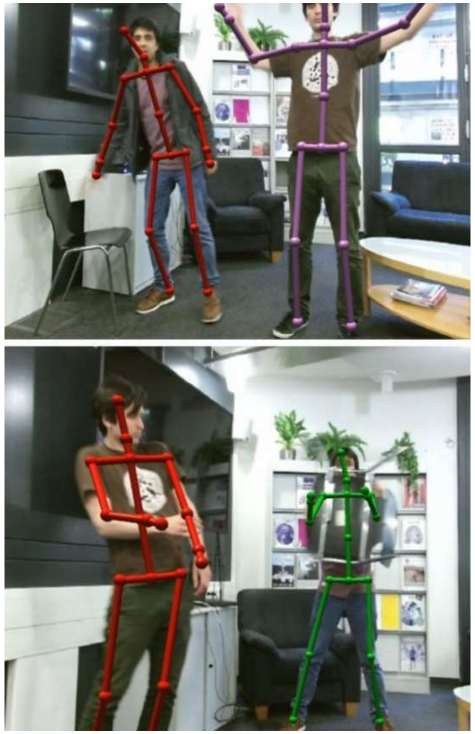
https://www.vicon.com/resources/case-studies/a-simple-motion-capture-system-delivering-powerful-results/

USING OPTICAL MARKERS

- Vicon (https://www.vicon.com/)
- Motion Analysis (https://www.motionanalysis.com/)
- OptiTrack (https://www.optitrack.com/)

PREVIOUS WORKS

清華大學 Tsinghua University



Xnect
[Mehta et al. 2020]

Monocular Real-time Full Body Capture
[Zhou et al. 2021]

DeepCap
[Habermann et al. 2020]

USING VIDEOS (MARKER-FREE)

[Chen et al. 2020]

[Mehta et al. 2020]

[Trumble et al. 2016]

[Bogo et al. 2016]

[Kolotouros et al. 2019]

[Zhou et al. 2021]

[Habermann et al. 2020]

[Habibie et al. 2019]

[Tome et al. 2018]

[Xiang et al. 2019]

[Kanazawa et al. 2019]

[Kocabas et al. 2020]

[Shimada et al. 2020]

[Xu et al. 2018]

PREVIOUS WORKS

清華大學
Tsinghua University

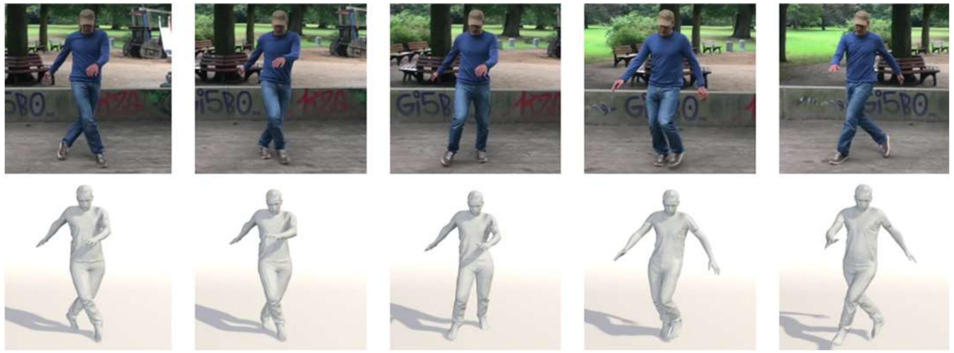
https://www.xsens.com/cases/enhancing-the-performance-of-junior-pro-tennis-athletes-with-xsens-mvn-analyze

USING DENSE INERTIAL SENSORS

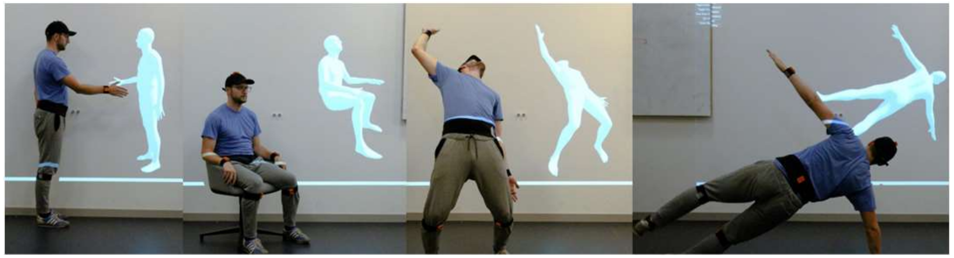
- Xsens (https://www.xsens.com/)
- Noitom (https://noitom.com/)

PREVIOUS WORKS

Tsinghua University



Sparse Inertial Poser
[Marcard et al, 2017]

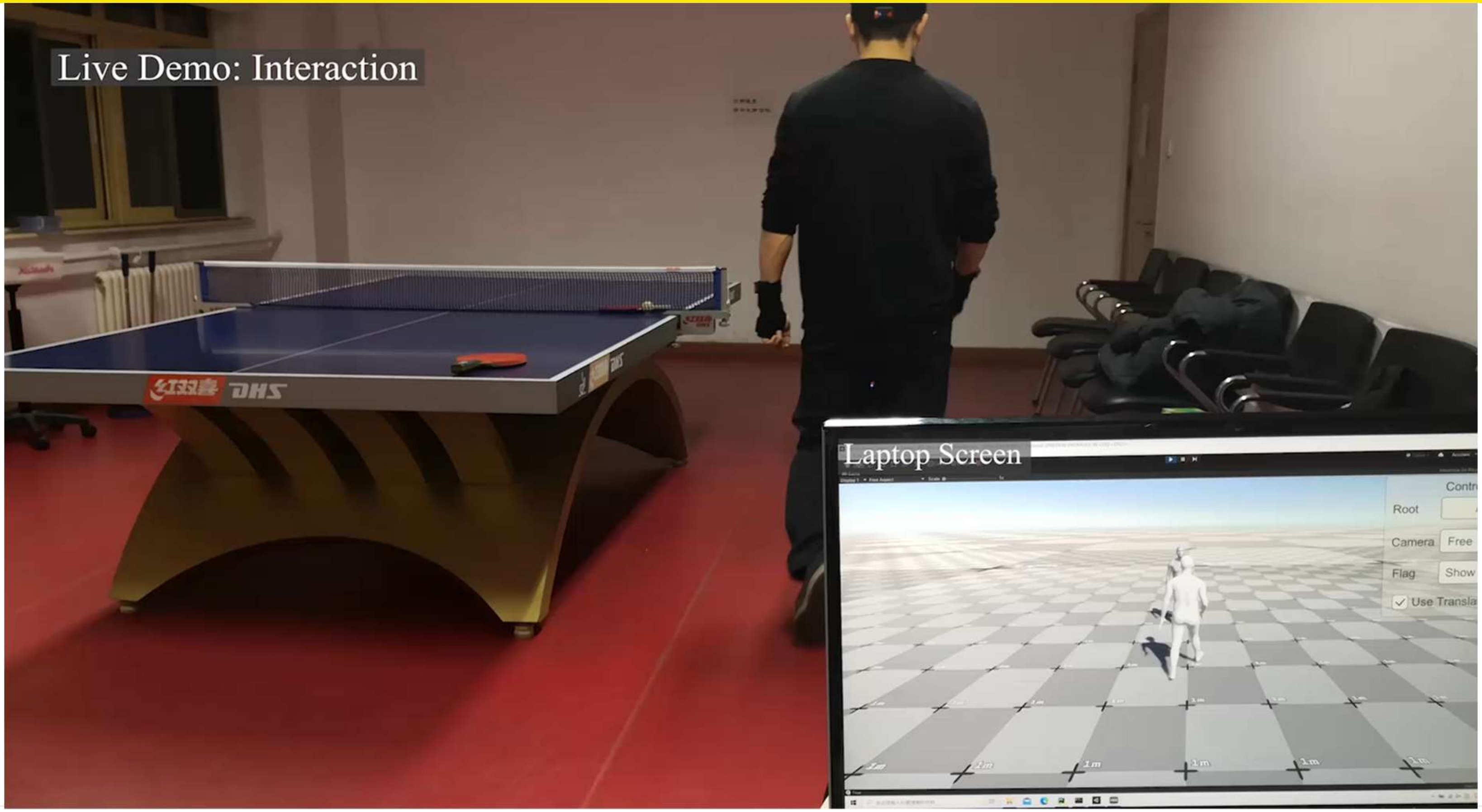


Deep Inertial Poser
[Huang et al, 2018]

USING SPARSE INERTIAL SENSORS

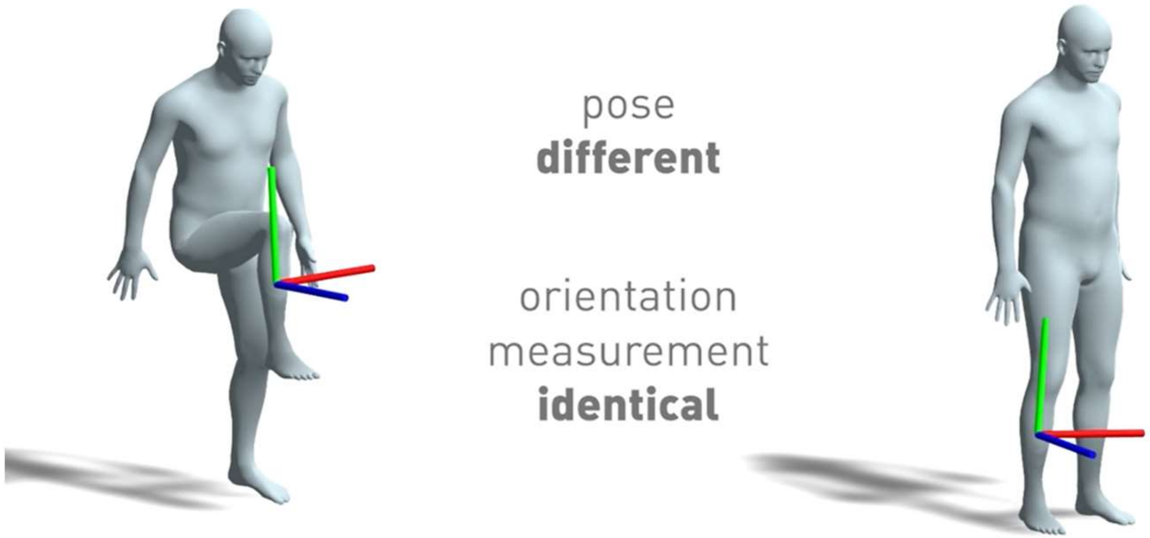
- SIP: Sparse Inertial Poser [Marcard et al, 2017]
- DIP: Deep Inertial Poser [Huang et al, 2018]

OURS: MOCAP FROM SPARSE IMUS



CHALLENGES

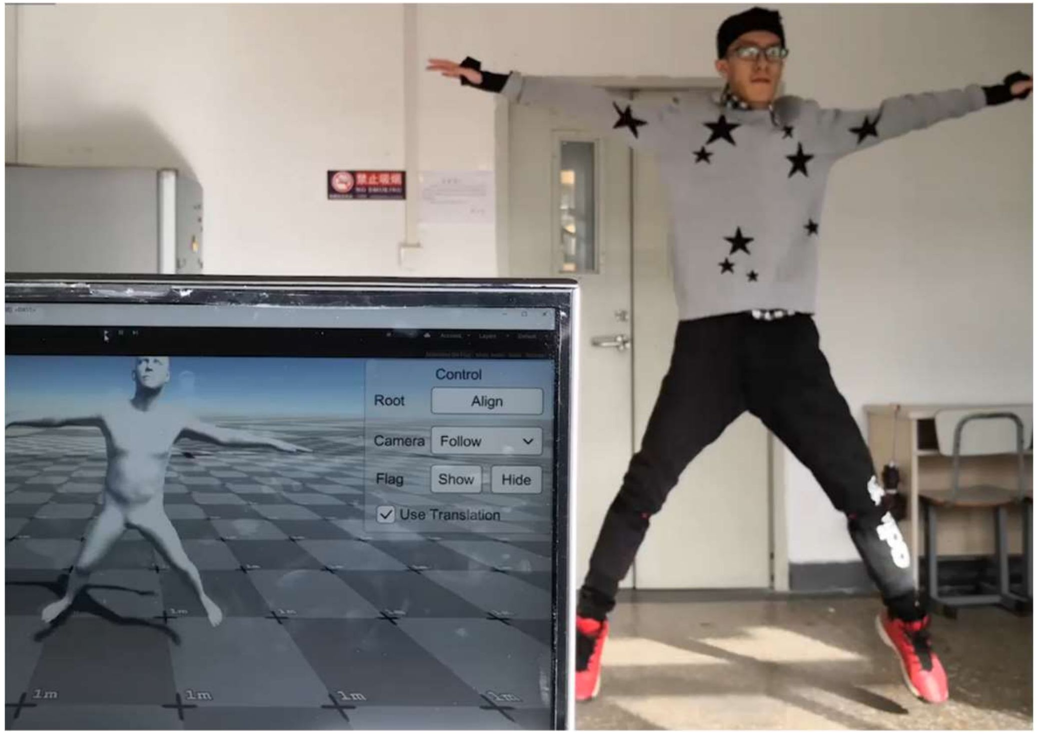
- Learning pose prior
  - IMU signals are sparse and noisy



CHALLENGES

Tsinghua University

- Learning pose prior
  - IMU signals are sparse and noisy
- Estimating global movements
  - No direct distance measurement
  - Acceleration signals are noisy

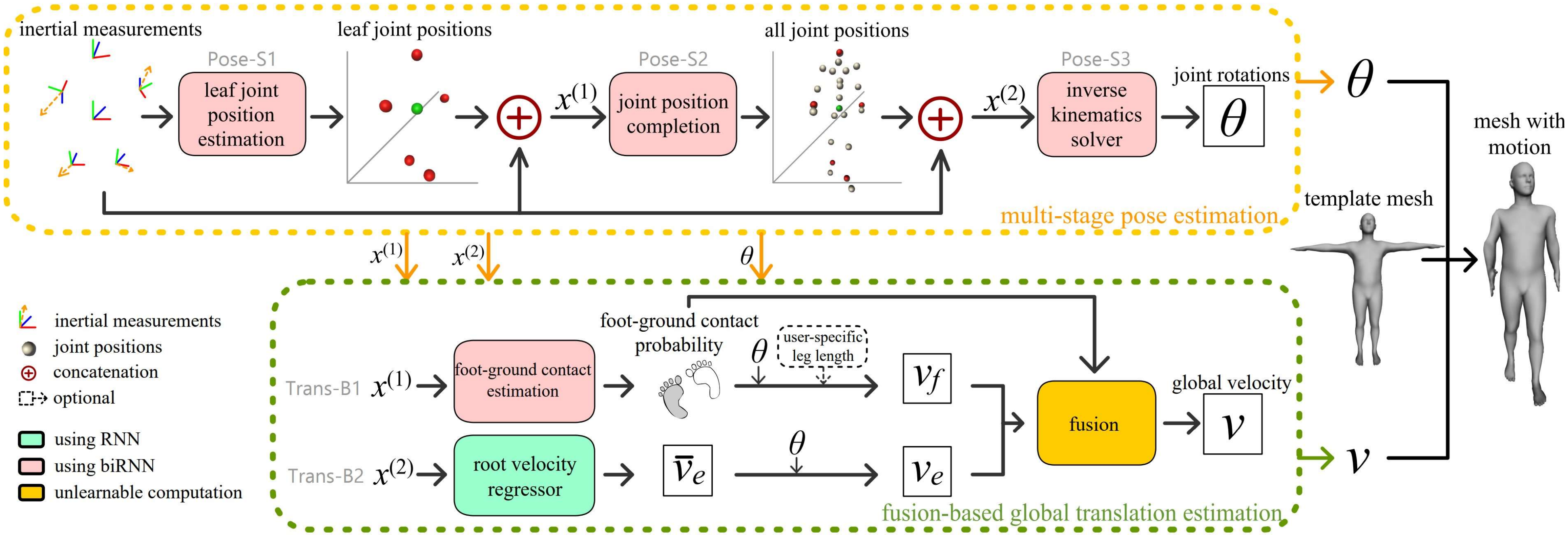


SIGGRAPH 2021

METHOD

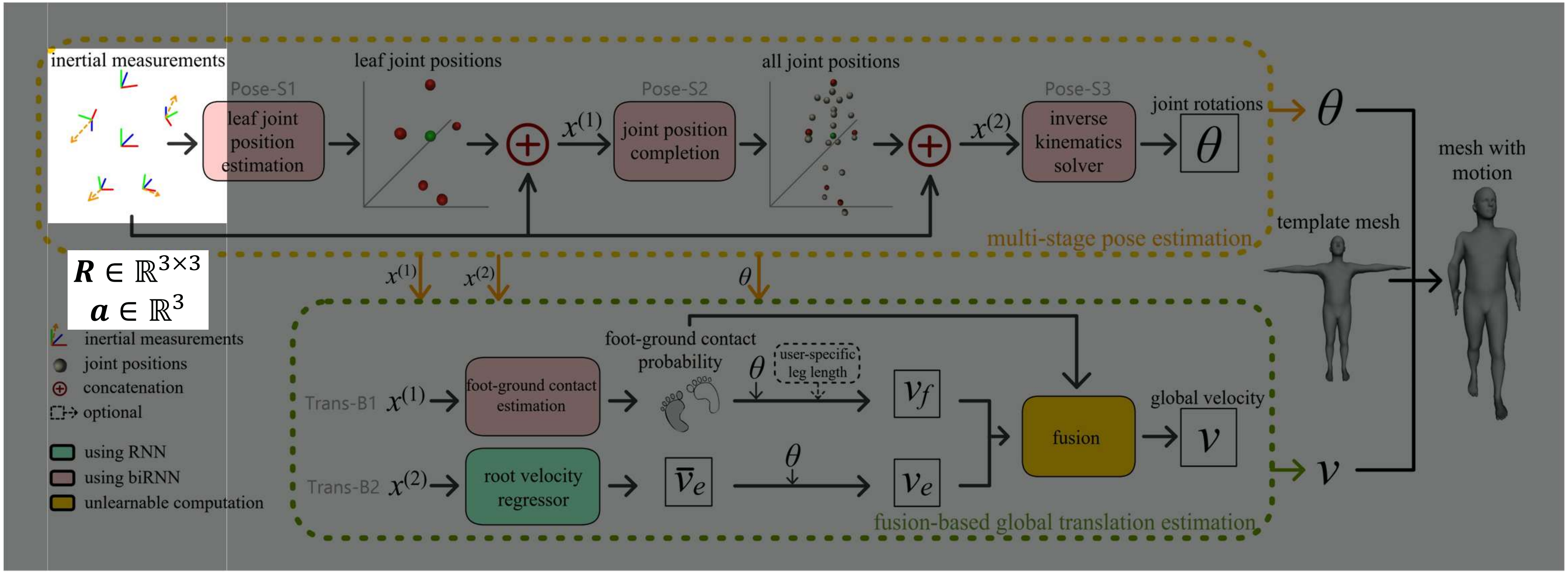
METHOD: OVERVIEW

Tsinghua University



METHOD: OVERVIEW

Tsinghua University



Input: orientations $\boldsymbol{R}$ and accelerations $\boldsymbol{a}$ of 6 IMUs

METHOD: OVERVIEW

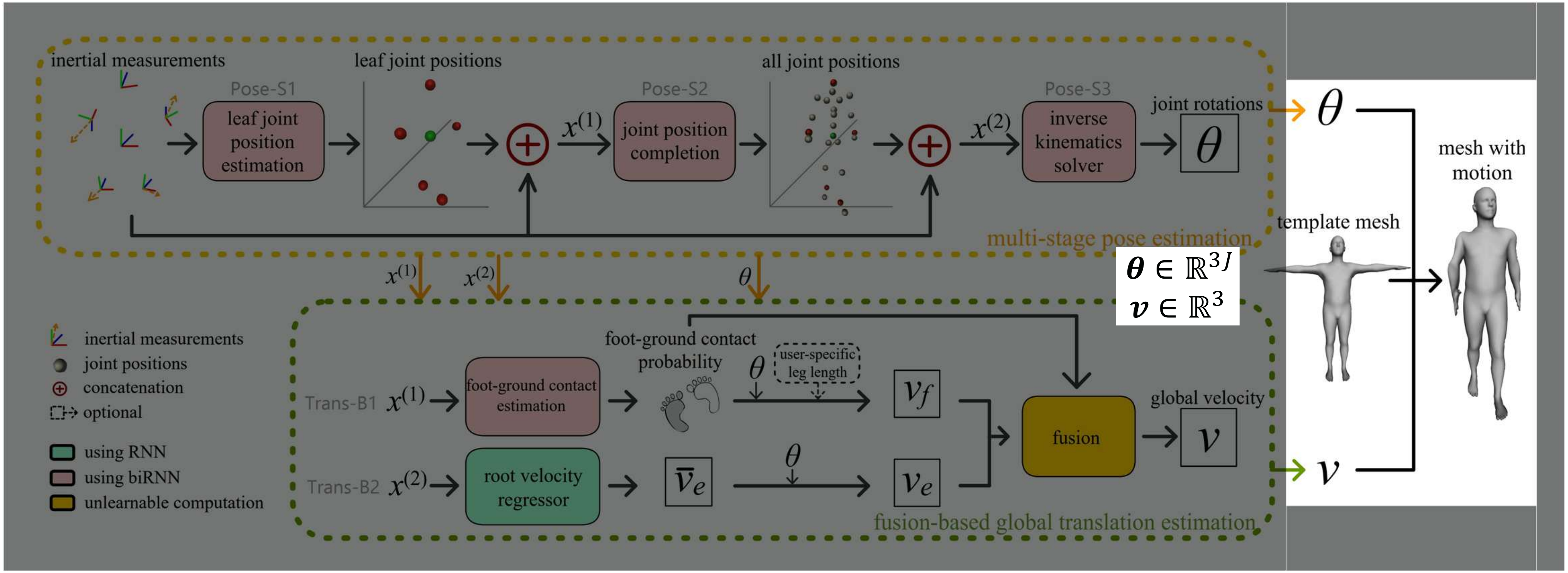
Tsinghua University

1. left forearm
2. right forearm
3. left lower leg
4. right lower leg
5. head
6. pelvis (root)

Input: orientations $\boldsymbol{R}$ and accelerations $\boldsymbol{a}$ of 6 IMUs

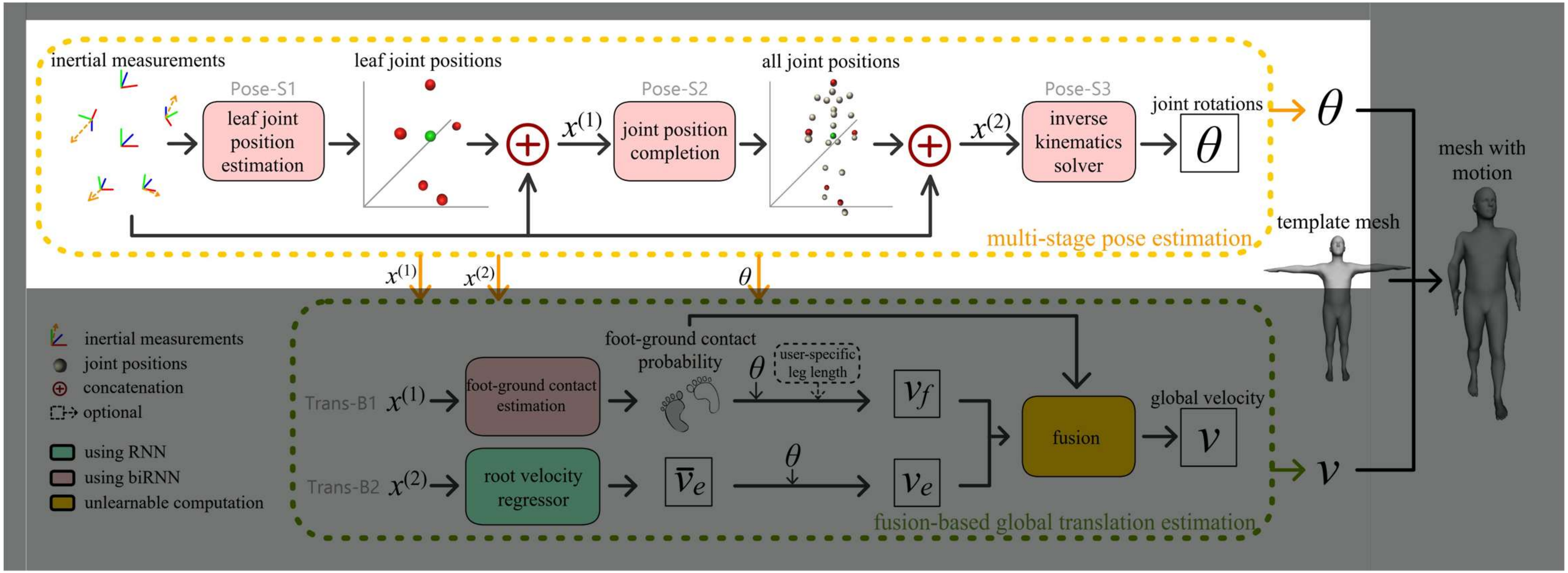
METHOD: OVERVIEW

Tsinghua University



Output: pose parameters $\boldsymbol{\theta}$ and translations $\boldsymbol{v}$ of the subject

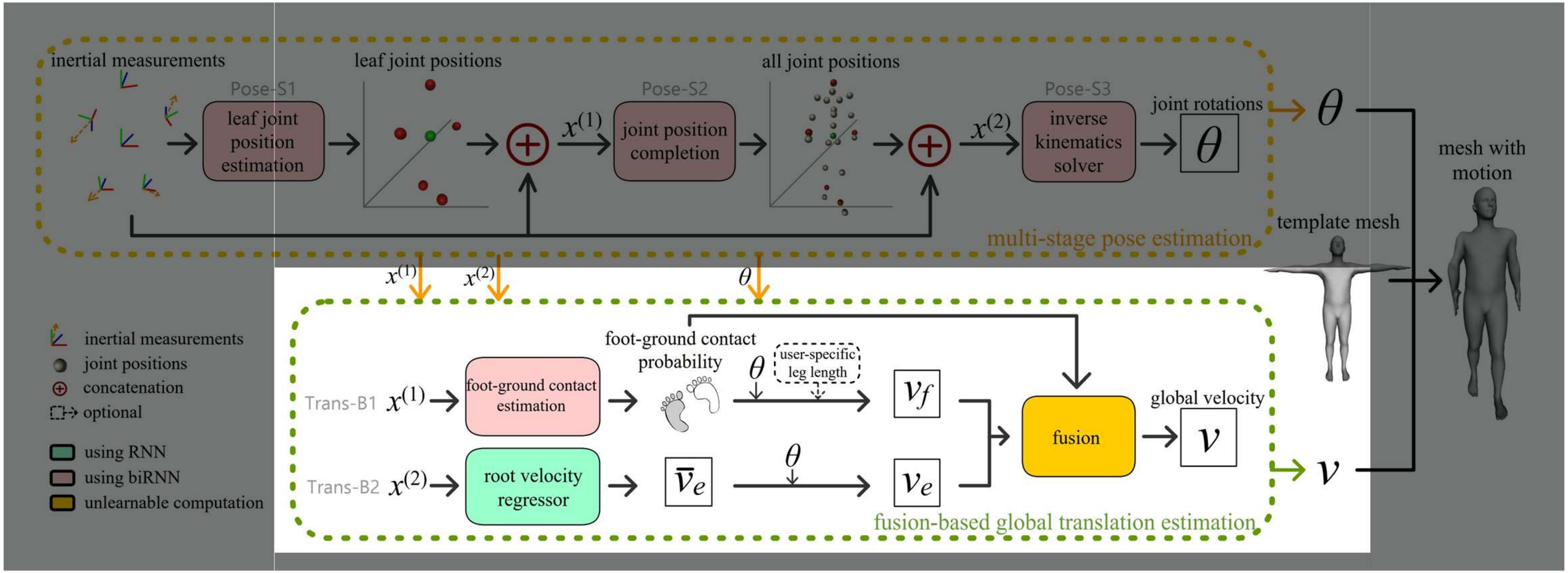
METHOD: OVERVIEW



Pose estimation subtask: pose parameters

METHOD: OVERVIEW

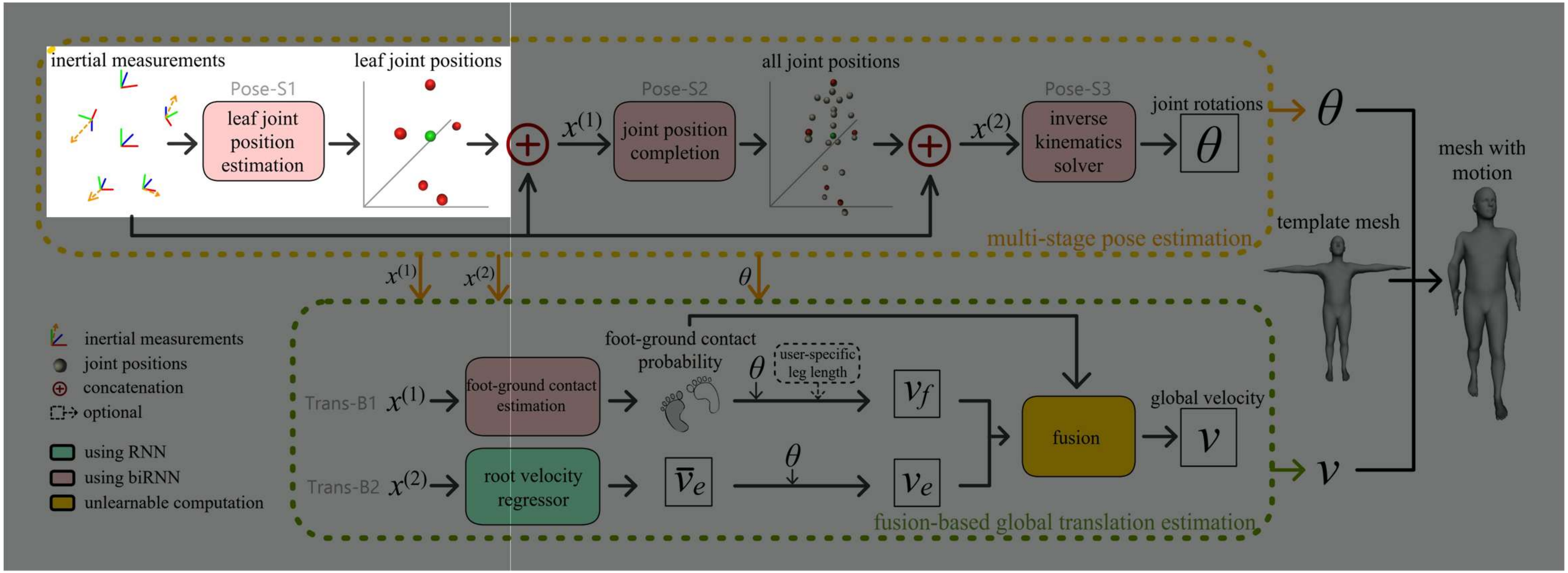
Tsinghua University



Translation estimation subtask: global translations

METHOD: MULTI-STAGE POSE ESTIMATION

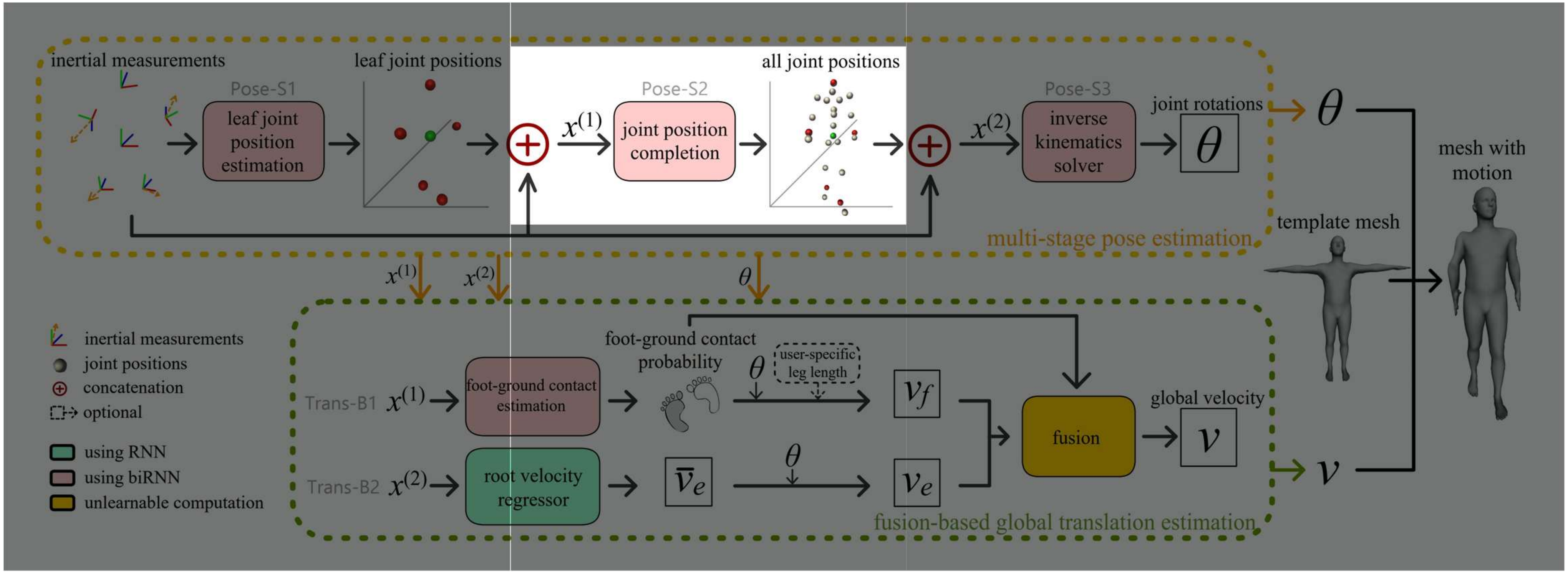
清華大學 Tsinghua University



Pose Stage 1: IMUs → leaf joint positions

METHOD: MULTI-STAGE POSE ESTIMATION

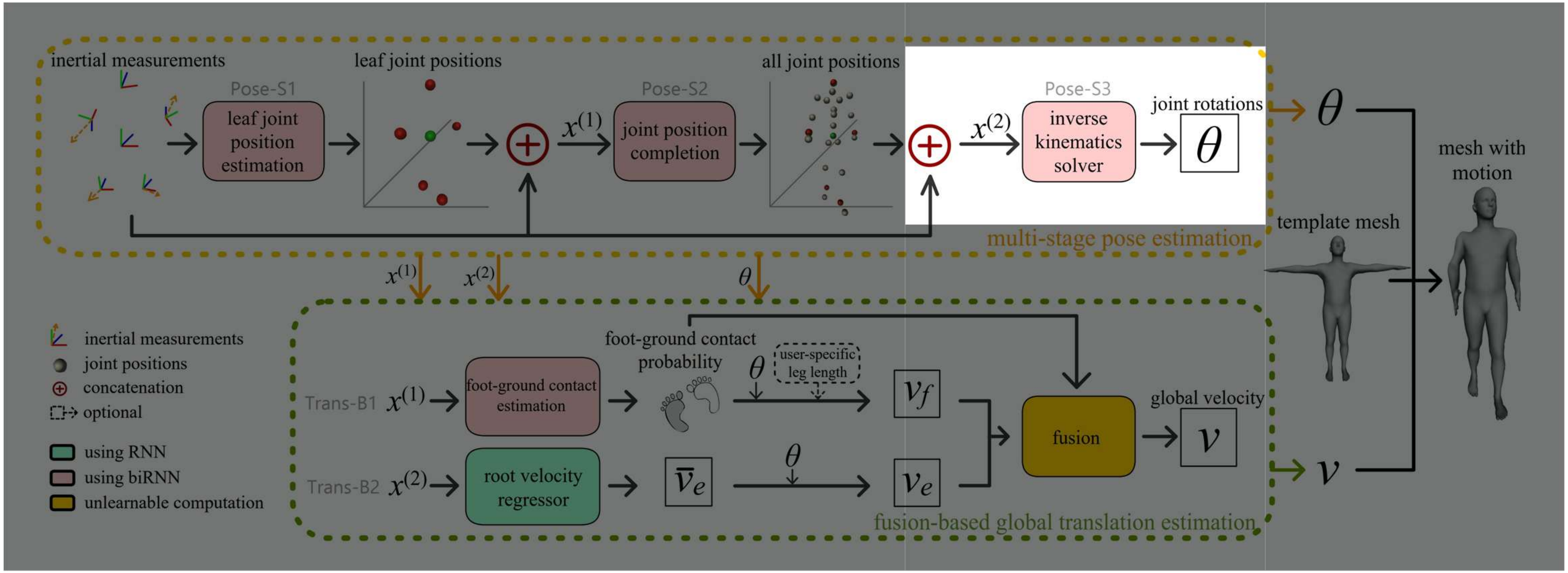
Tsinghua University



Pose Stage 2: IMUs + leaf joint positions → full joint positions

METHOD: MULTI-STAGE POSE ESTIMATION

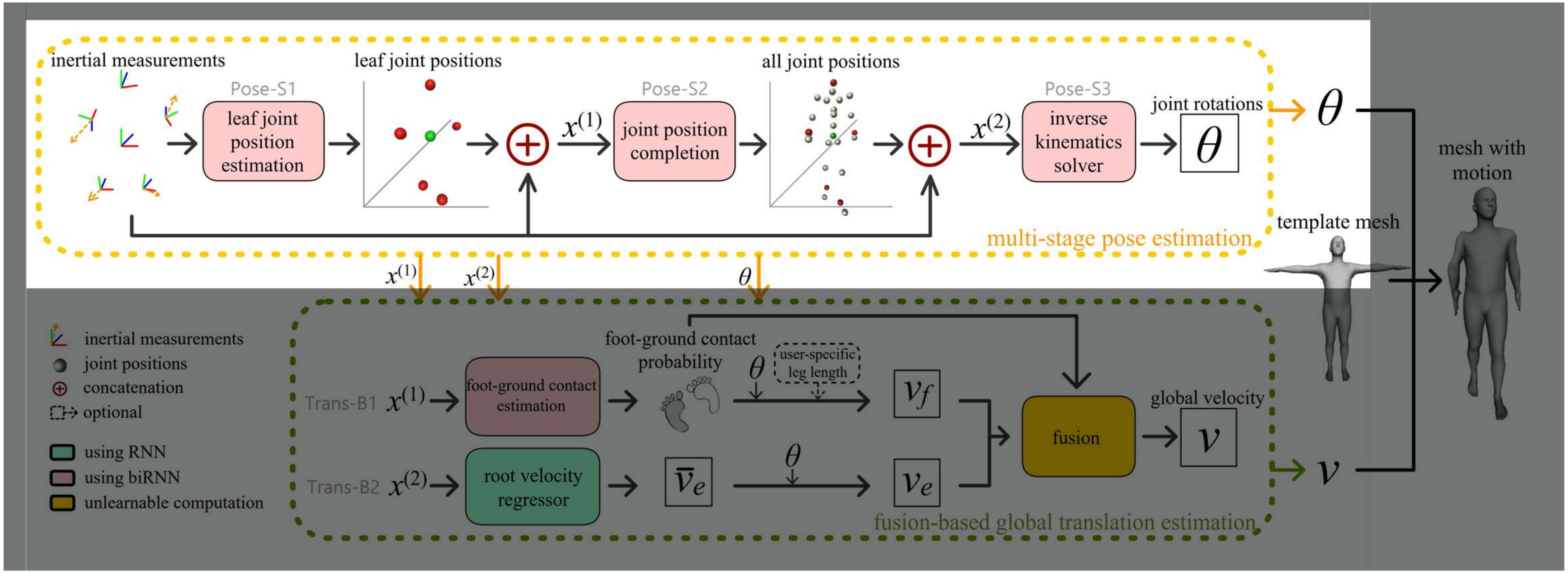
Tsinghua University



Pose Stage 3: IMUs + full joint positions → joint rotations

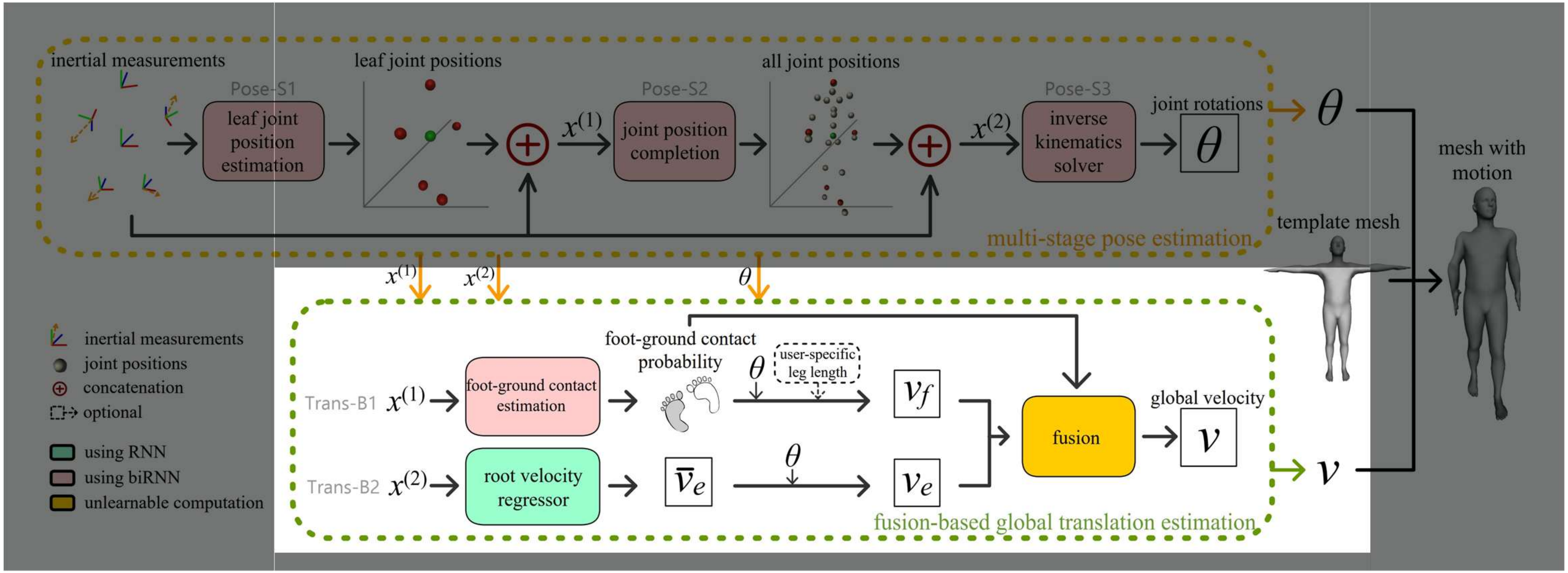
METHOD: MULTI-STAGE POSE ESTIMATION

清華大學 Tsinghua University



METHOD: FUSION-BASED TRANSLATION ESTIMATION

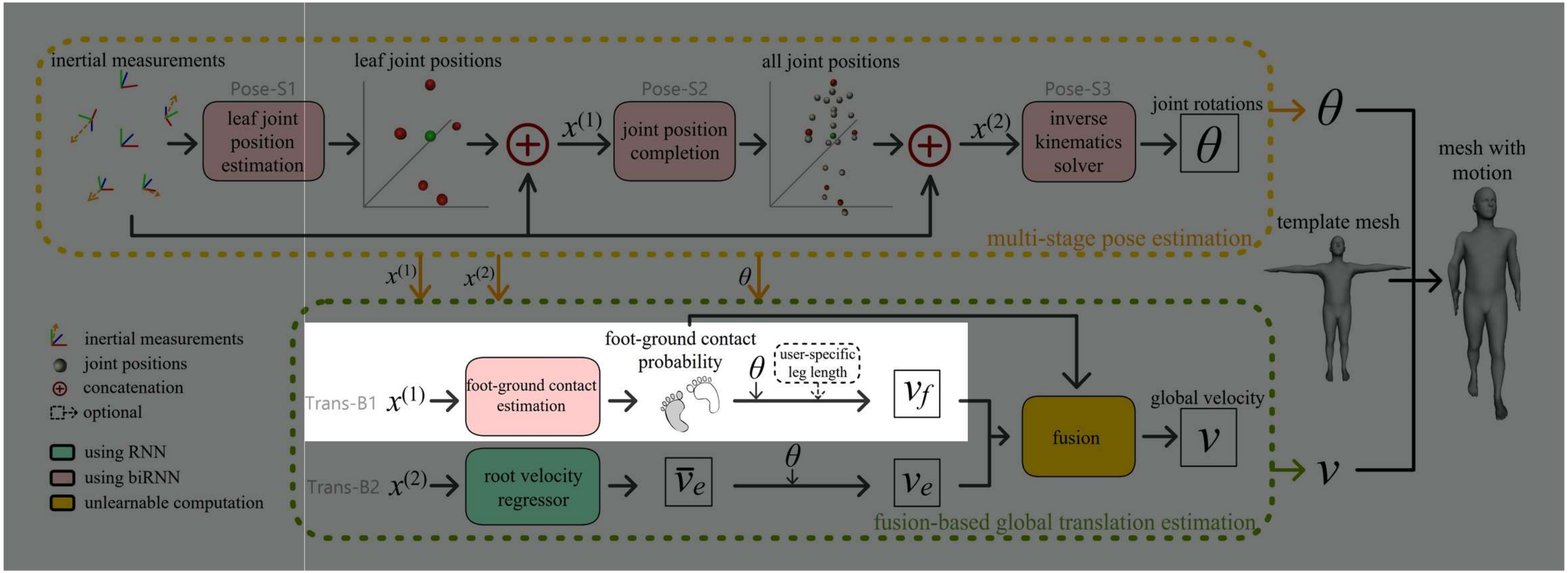
Tsinghua University



Translation estimation subtask: global translations

METHOD: FUSION-BASED TRANSLATION ESTIMATION

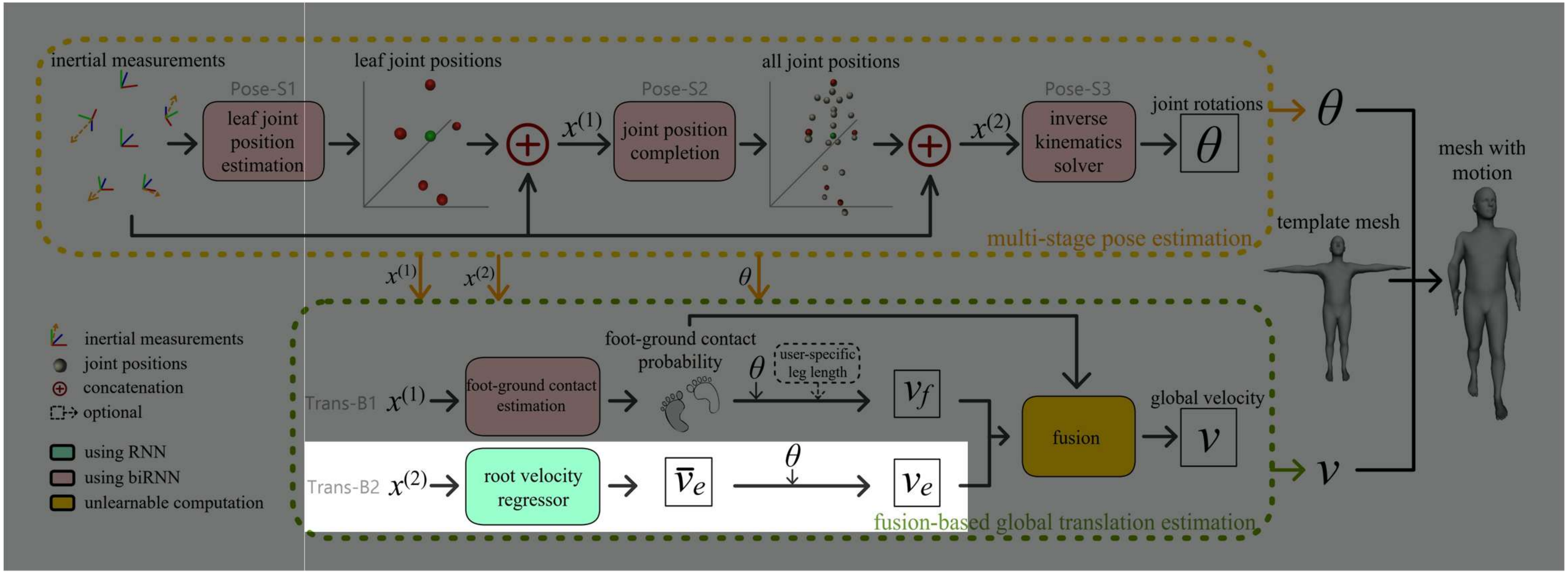
Tsinghua University



Translation Branch 1: IMUs + leaf joint positions → physics-rule-based translations

METHOD: FUSION-BASED TRANSLATION ESTIMATION

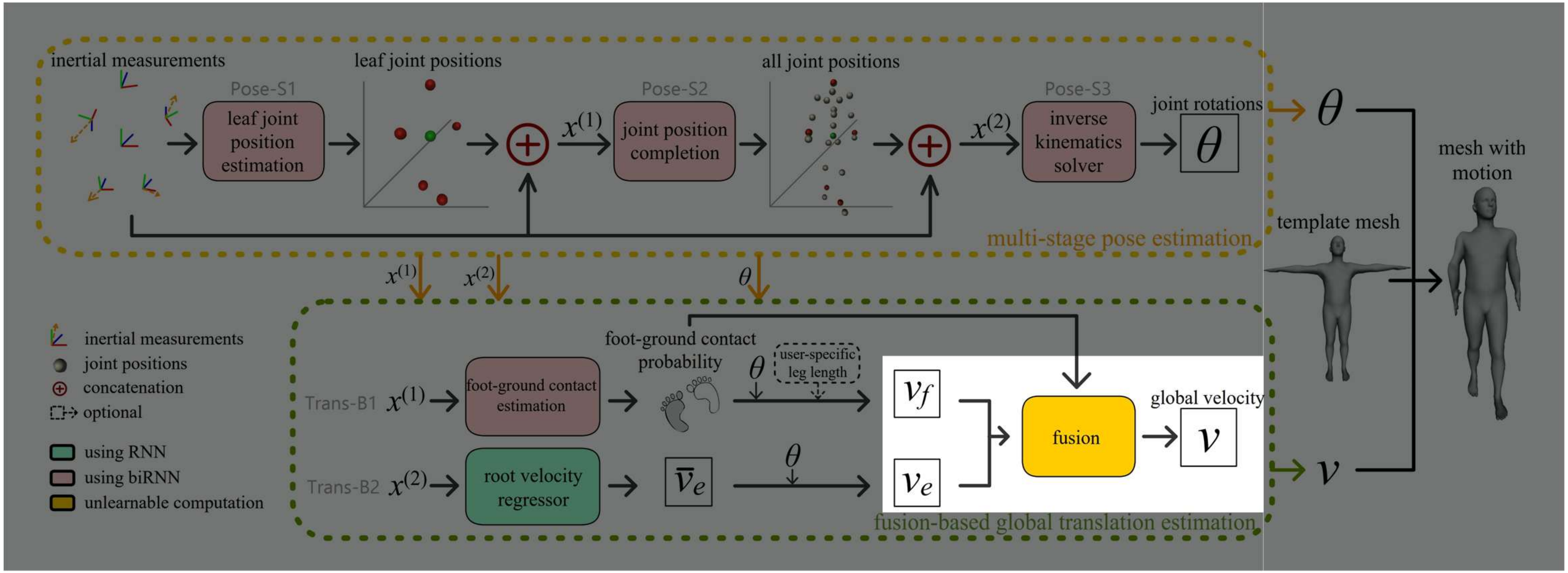
Tsinghua University



Translation Branch 2: IMUs + full joint positions → network-regressed translations

METHOD: FUSION-BASED TRANSLATION ESTIMATION

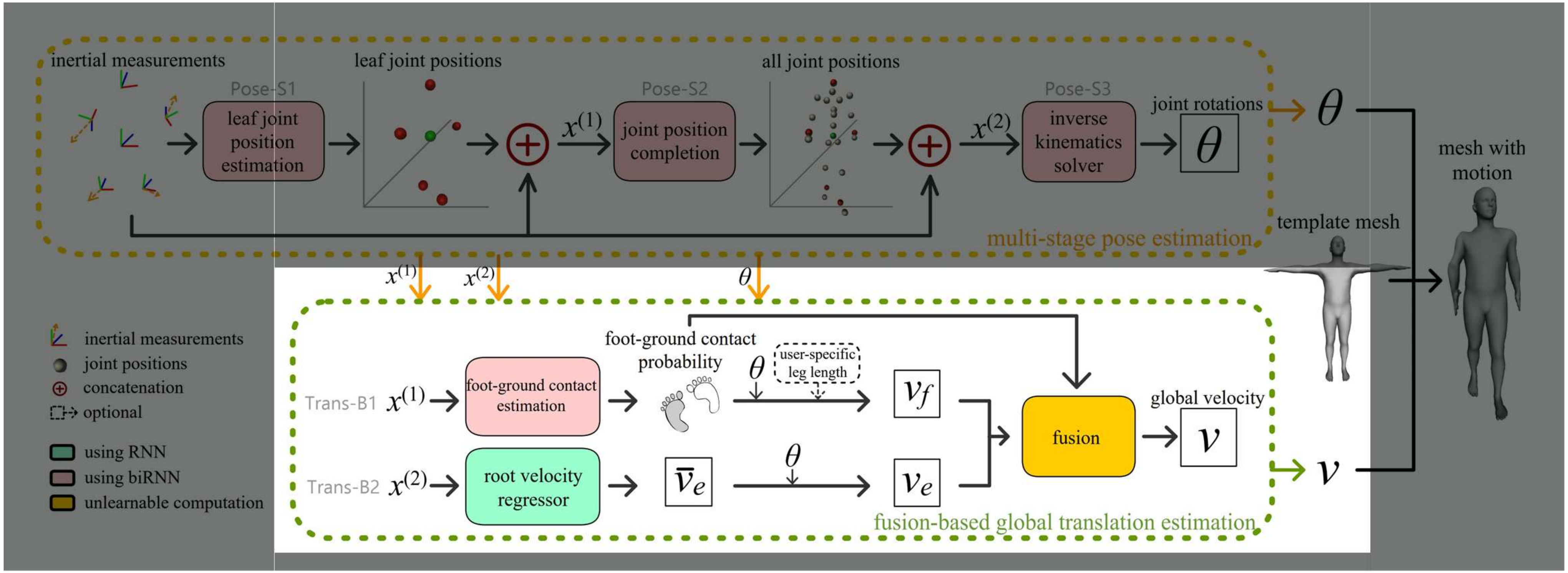
Tsinghua University



Translation Fusion: physics rule + network → final translation

METHOD: FUSION-BASED TRANSLATION ESTIMATION

Tsinghua University



Translation estimation subtask: global translations

METHOD: SUPPORTING FOOT VISUALIZATION

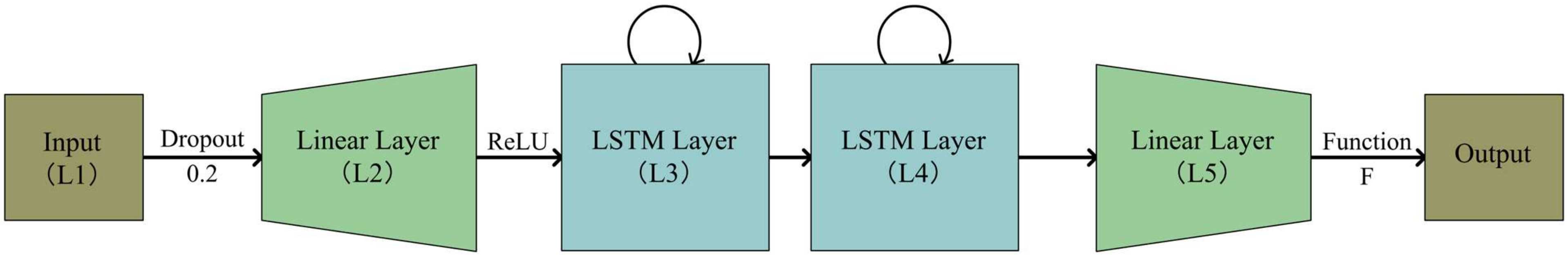
Tsinghua University

Supporting Foot Visualization I

supporting foot probability 0 1

We record the sensor measurements and run our pipeline offline to render the supporting foot predictions.

METHOD: NETWORK DETAILS



METHOD: DATA

Tsinghua University

Dataset	Pose	IMU	Translation	Contact	Minutes
DIP-IMU [Huang et al. 2018]	Y	Y	N	N	80
TotalCapture [Trumble et al. 2017]	Y[a]	Y	Y	S	49
AMASS[b] [Mahmood et al. 2019]	Y	S	Y	S	1217

[a]Provided by DIP authors [Huang et al. 2018]
[b]Down-sample into 60 fps
"Y" means that the dataset contains such information.
"N" means that the dataset does not contain such information.
"S" means that the data is synthesized from other information.

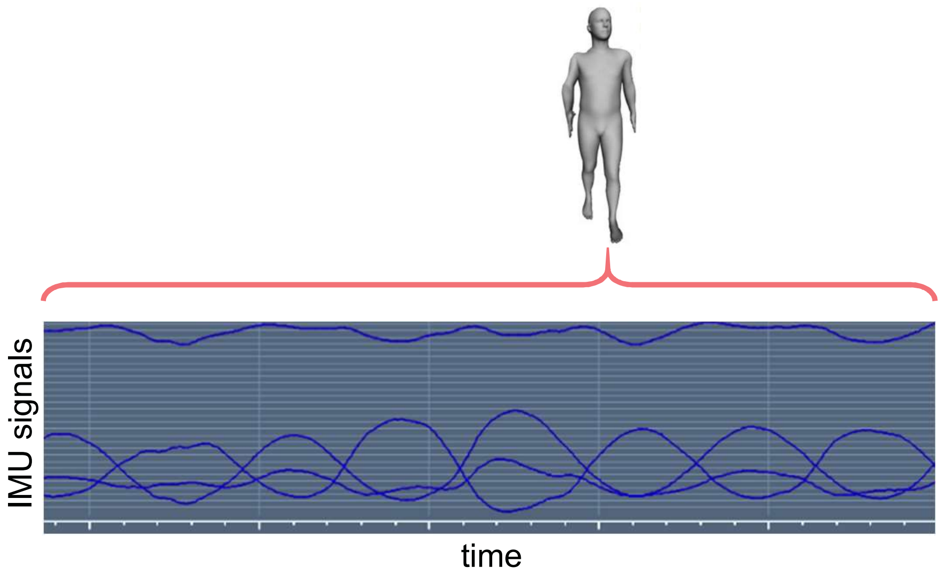
SIGGRAPH 2021

RESULTS

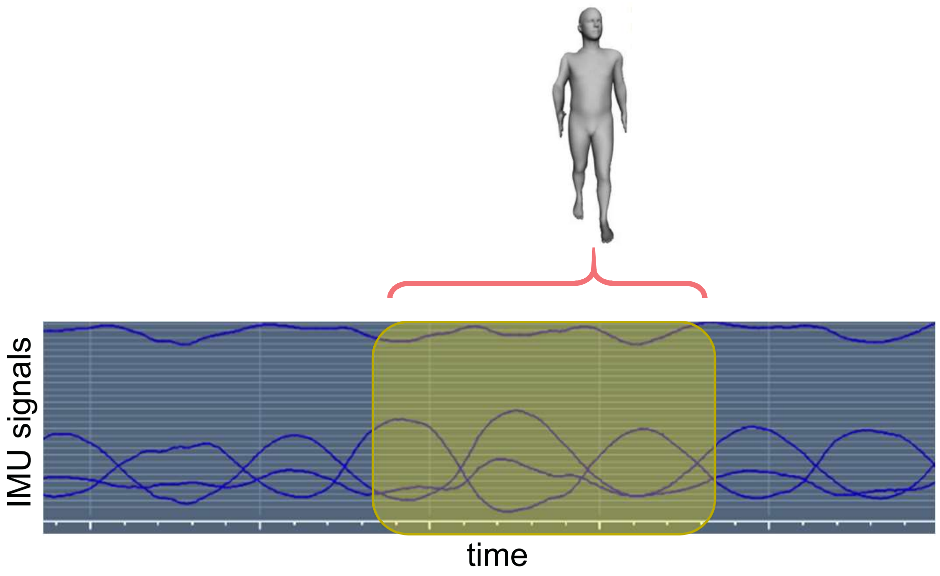
RESULTS: EVALUATION SETTINGS

Tsinghua University

OFFLINE SETTING



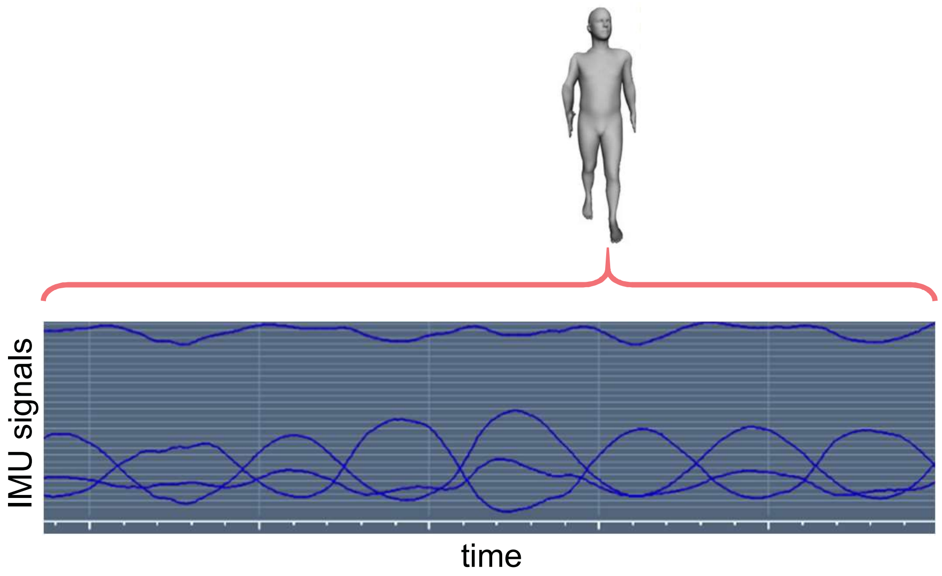
ONLINE SETTING



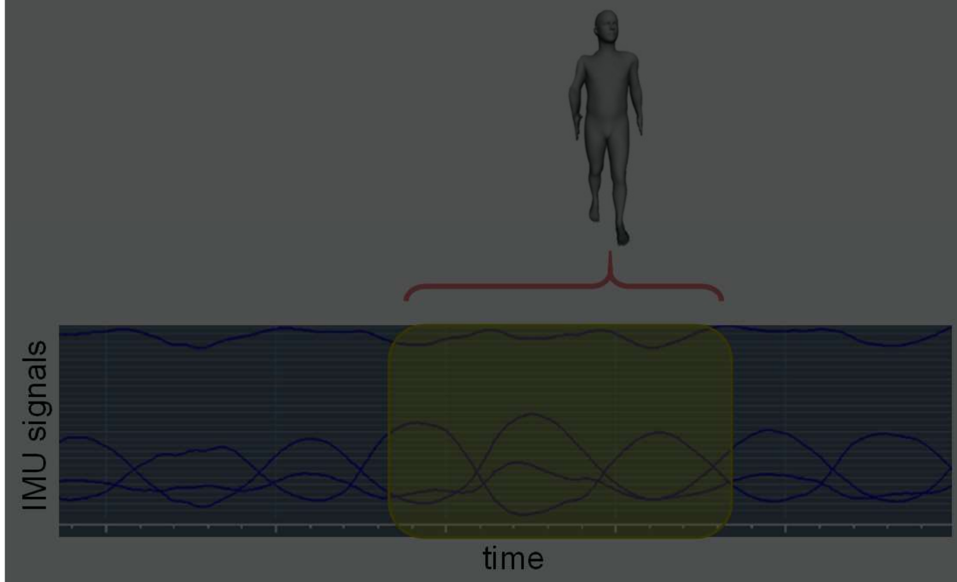
RESULTS: EVALUATION SETTINGS

Tsinghua University

OFFLINE SETTING



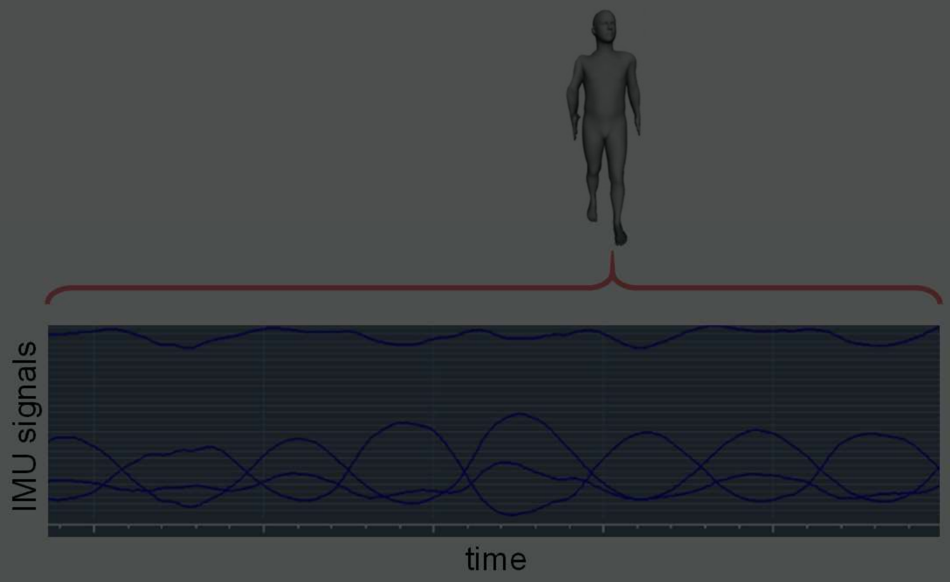
ONLINE SETTING



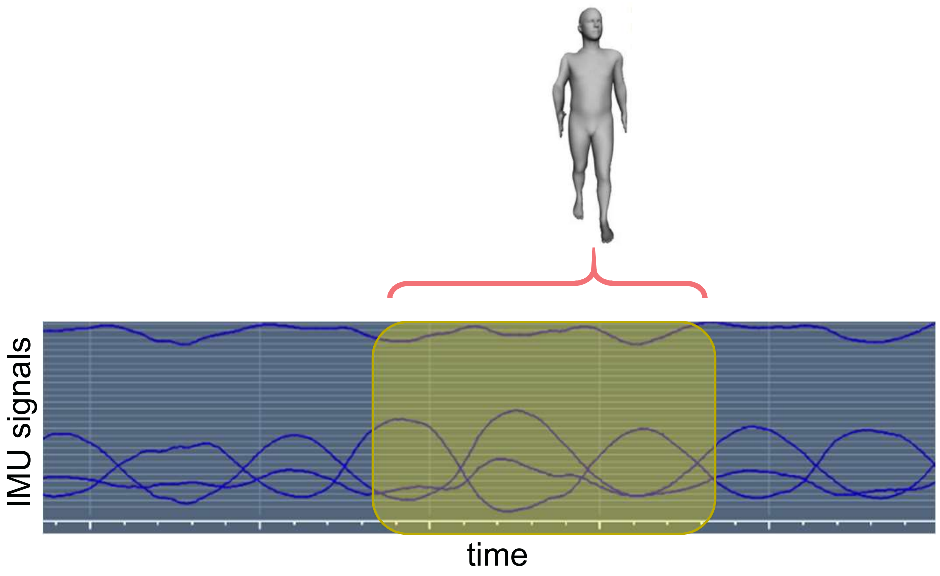
RESULTS: EVALUATION SETTINGS

Tsinghua University

OFFLINE SETTING



ONLINE SETTING



RESULTS: POSE COMPARISONS

清華大學
Tsinghua University

Offline Comparisons

	TotalCapture					DIP-IMU				
	SIP Err (deg)	Ang Err (deg)	Pos Err (cm)	Mesh Err (cm)	Jitter ($10^2 m/s^3$)	SIP Err (deg)	Ang Err (deg)	Pos Err (cm)	Mesh Err (cm)	Jitter ($10^2 m/s^3$)
SOP	23.09 (±12.37)	17.14 (±8.54)	9.24 (±5.33)	10.58 (±6.04)	8.17 (±13.55)	24.56 (±12.75)	9.83 (±5.21)	8.17 (±4.74)	9.32 (±5.27)	5.66 (±9.49)
SIP	18.54 (±9.67)	14.84 (±7.26)	7.65 (±4.32)	8.60 (±4.83)	8.27 (±17.36)	21.02 (±9.61)	8.77 (±4.38)	6.66 (±3.33)	7.71 (±3.80)	3.86 (±6.32)
DIP	18.79 (±11.85)	17.77 (±9.51)	9.61 (±5.76)	11.34 (±6.45)	28.86 (±29.18)	16.36 (±8.60)	14.41 (±7.90)	6.98 (±3.89)	8.56 (±4.65)	23.37 (±23.84)
Ours	**14.95 (±6.90)**	**12.26 (±5.59)**	**5.57 (±3.09)**	**6.36 (±3.47)**	**1.57 (±2.93)**	**13.97 (±6.77)**	**7.62 (±4.01)**	**4.90 (±2.75)**	**5.83 (±3.21)**	**1.19 (±1.76)**

Online Comparisons

	TotalCapture					DIP-IMU				
	SIP Err (deg)	Ang Err (deg)	Pos Err (cm)	Mesh Err (cm)	Jitter ($10^2 m/s^3$)	SIP Err (deg)	Ang Err (deg)	Pos Err (cm)	Mesh Err (cm)	Jitter ($10^2 m/s^3$)
DIP	18.93 (±12.44)	17.50 (±10.10)	9.57 (±5.95)	11.40 (±6.87)	35.94 (±34.45)	17.10 (±9.59)	15.16 (±8.53)	7.33 (±4.23)	8.96 (±5.01)	30.13 (±28.76)
Ours	**16.69 (±8.79)**	**12.93 (± 6.15)**	**6.61 (± 3.93)**	**7.49 (± 4.35)**	**9.44 (±13.57)**	**16.68 (±8.68)**	**8.85 (± 4.82)**	**5.95 (± 3.65)**	**7.09 (± 4.24)**	**6.11 (±7.92)**

RESULTS: POSE COMPARISONS

Tsinghua University

Offline Comparisons

	TotalCapture					DIP-IMU				
	SIP Err (deg)	Ang Err (deg)	Pos Err (cm)	Mesh Err (cm)	Jitter ($10^2 m/s^3$)	SIP Err (deg)	Ang Err (deg)	Pos Err (cm)	Mesh Err (cm)	Jitter ($10^2 m/s^3$)
SOP	23.09 (±12.37)	17.14 (±8.54)	9.24 (±5.33)	10.58 (±6.04)	8.17 (±13.55)	24.56 (±12.75)	9.83 (±5.21)	8.17 (±4.74)	9.32 (±5.27)	5.66 (±9.49)
SIP	18.54 (±9.67)	14.84 (±7.26)	7.65 (±4.32)	8.60 (±4.83)	8.27 (±17.36)	21.02 (±9.61)	8.77 (±4.38)	6.66 (±3.33)	7.71 (±3.80)	3.86 (±6.32)
DIP	18.79 (±11.85)	17.77 (±9.51)	9.61 (±5.76)	11.34 (±6.45)	28.86 (±29.18)	16.36 (±8.60)	14.41 (±7.90)	6.98 (±3.89)	8.56 (±4.65)	23.37 (±23.84)
Ours	**14.95 (±6.90)**	**12.26 (±5.59)**	**5.57 (±3.09)**	**6.36 (±3.47)**	**1.57 (±2.93)**	**13.97 (±6.77)**	**7.62 (±4.01)**	**4.90 (±2.75)**	**5.83 (±3.21)**	**1.19 (±1.76)**

Online Comparisons

	TotalCapture					DIP-IMU				
	SIP Err (deg)	Ang Err (deg)	Pos Err (cm)	Mesh Err (cm)	Jitter ($10^2 m/s^3$)	SIP Err (deg)	Ang Err (deg)	Pos Err (cm)	Mesh Err (cm)	Jitter ($10^2 m/s^3$)
DIP	18.93 (±12.44)	17.50 (±10.10)	9.57 (±5.95)	11.40 (±6.87)	35.94 (±34.45)	17.10 (±9.59)	15.16 (±8.53)	7.33 (±4.23)	8.96 (±5.01)	30.13 (±28.76)
Ours	**16.69 (±8.79)**	**12.93 (± 6.15)**	**6.61 (± 3.93)**	**7.49 (± 4.35)**	**9.44 (±13.57)**	**16.68 (±8.68)**	**8.85 (± 4.82)**	**5.95 (± 3.65)**	**7.09 (± 4.24)**	**6.11 (±7.92)**

RESULTS: POSE COMPARISONS

清華大學 Tsinghua University

Offline Comparisons

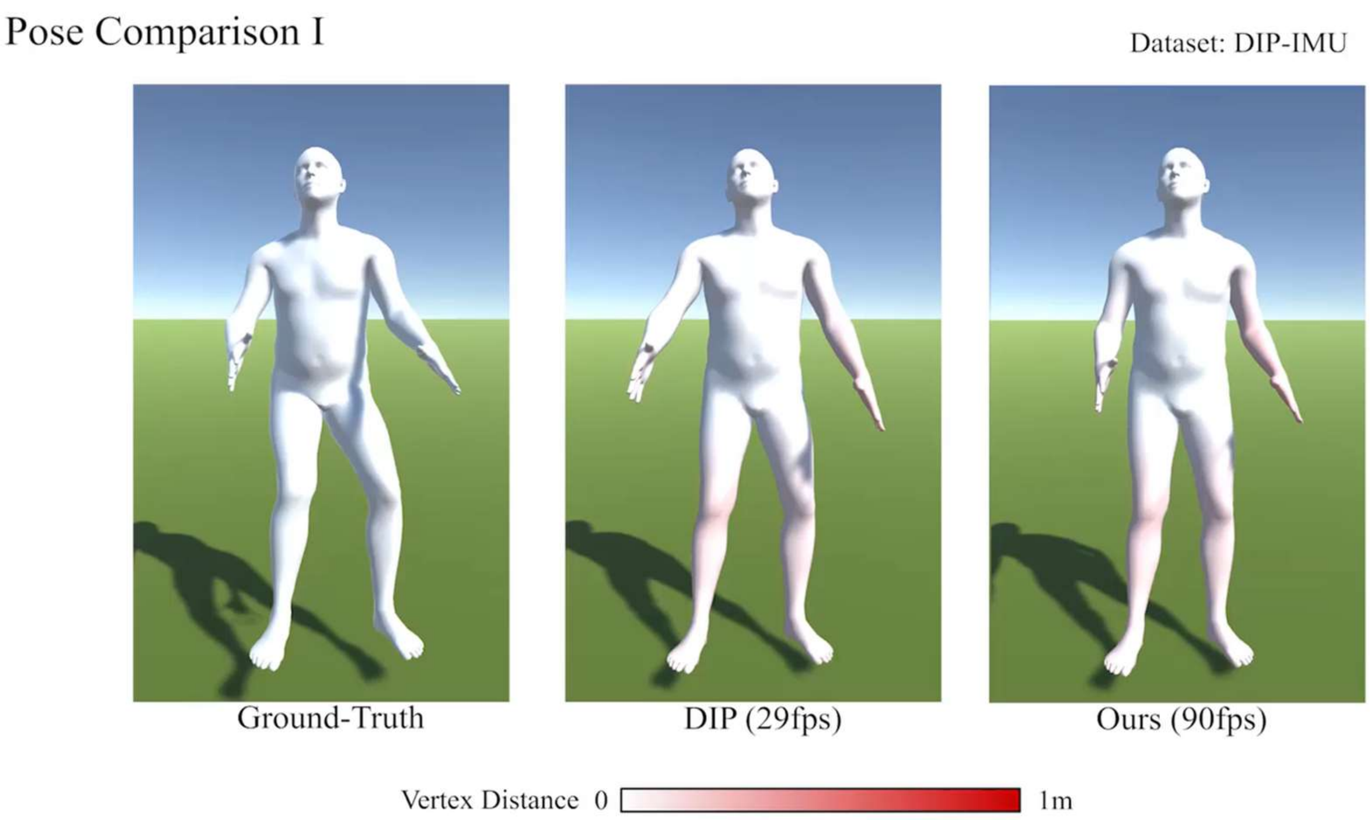
	TotalCapture					DIP-IMU				
	SIP Err (deg)	Ang Err (deg)	Pos Err (cm)	Mesh Err (cm)	Jitter ($10^2 m/s^3$)	SIP Err (deg)	Ang Err (deg)	Pos Err (cm)	Mesh Err (cm)	Jitter ($10^2 m/s^3$)
SOP	23.09 (±12.37)	17.14 (±8.54)	9.24 (±5.33)	10.58 (±6.04)	8.17 (±13.55)	24.56 (±12.75)	9.83 (±5.21)	8.17 (±4.74)	9.32 (±5.27)	5.66 (±9.49)
SIP	18.54 (±9.67)	14.84 (±7.26)	7.65 (±4.32)	8.60 (±4.83)	8.27 (±17.36)	21.02 (±9.61)	8.77 (±4.38)	6.66 (±3.33)	7.71 (±3.80)	3.86 (±6.32)
DIP	18.79 (±11.85)	17.77 (±9.51)	9.61 (±5.76)	11.34 (±6.45)	28.86 (±29.18)	16.36 (±8.60)	14.41 (±7.90)	6.98 (±3.89)	8.56 (±4.65)	23.37 (±23.84)
Ours	**14.95 (±6.90)**	**12.26 (±5.59)**	**5.57 (±3.09)**	**6.36 (±3.47)**	**1.57 (±2.93)**	**13.97 (±6.77)**	**7.62 (±4.01)**	**4.90 (±2.75)**	**5.83 (±3.21)**	**1.19 (±1.76)**

Online Comparisons

	TotalCapture					DIP-IMU				
	SIP Err (deg)	Ang Err (deg)	Pos Err (cm)	Mesh Err (cm)	Jitter ($10^2 m/s^3$)	SIP Err (deg)	Ang Err (deg)	Pos Err (cm)	Mesh Err (cm)	Jitter ($10^2 m/s^3$)
DIP	18.93 (±12.44)	17.50 (±10.10)	9.57 (±5.95)	11.40 (±6.87)	35.94 (±34.45)	17.10 (±9.59)	15.16 (±8.53)	7.33 (±4.23)	8.96 (±5.01)	30.13 (±28.76)
Ours	**16.69 (±8.79)**	**12.93 (± 6.15)**	**6.61 (± 3.93)**	**7.49 (± 4.35)**	**9.44 (±13.57)**	**16.68 (±8.68)**	**8.85 (± 4.82)**	**5.95 (± 3.65)**	**7.09 (± 4.24)**	**6.11 (±7.92)**

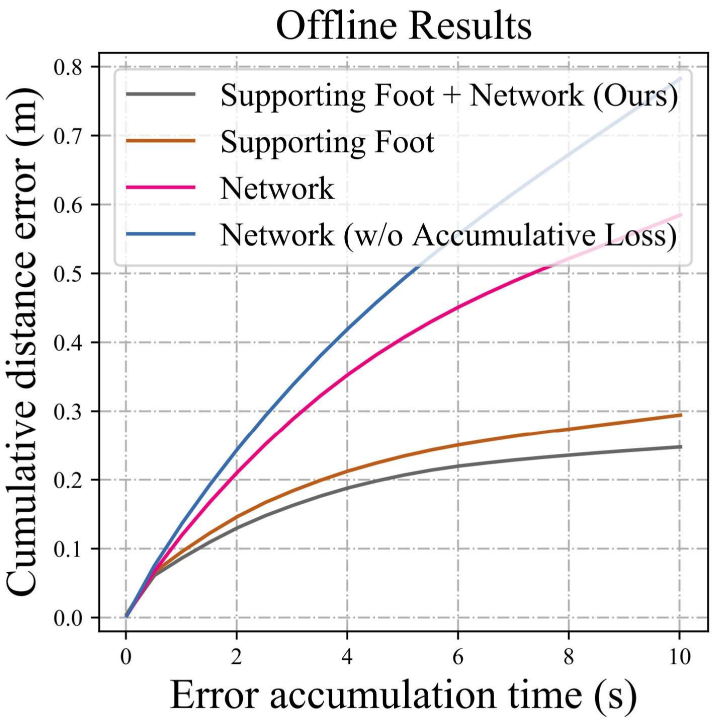
RESULTS: POSE COMPARISONS

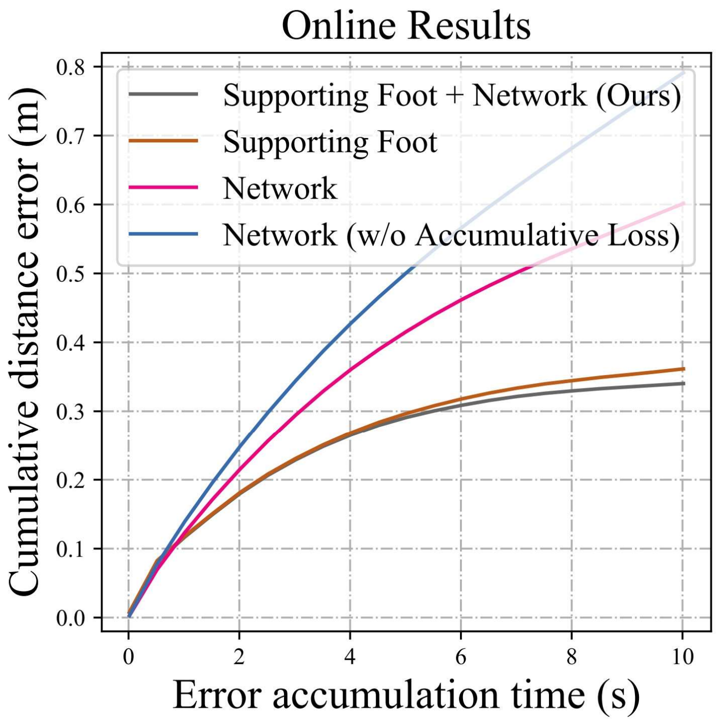
清華大學
Tsinghua University



RESULTS: TRANSLATION EVALUATIONS

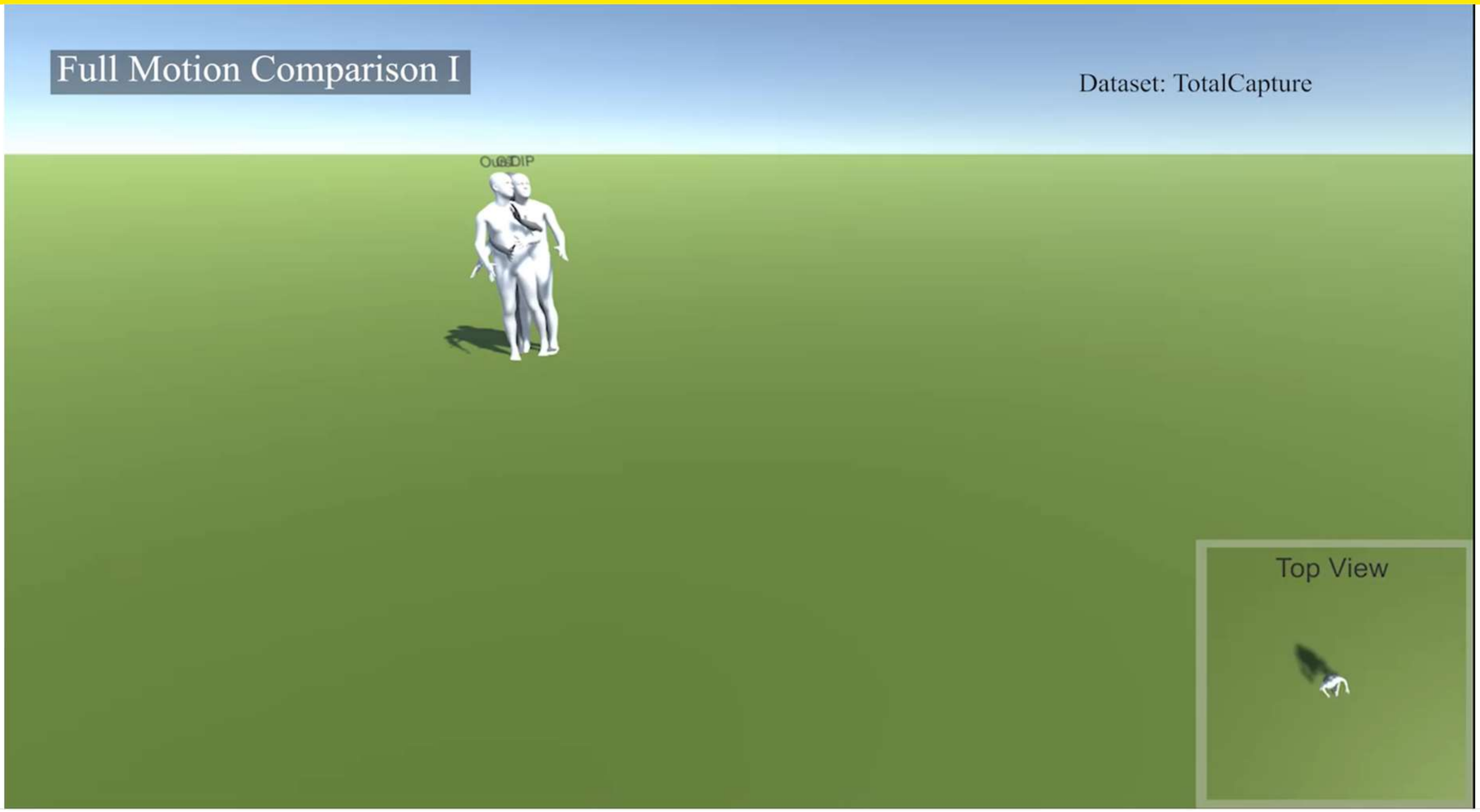
Tsinghua University





RESULTS: TRANSLATION EVALUATIONS

清華大學 Tsinghua University



RESULTS: ABLATION STUDY

Tsinghua University

Evaluation of the multi-stage pose estimation

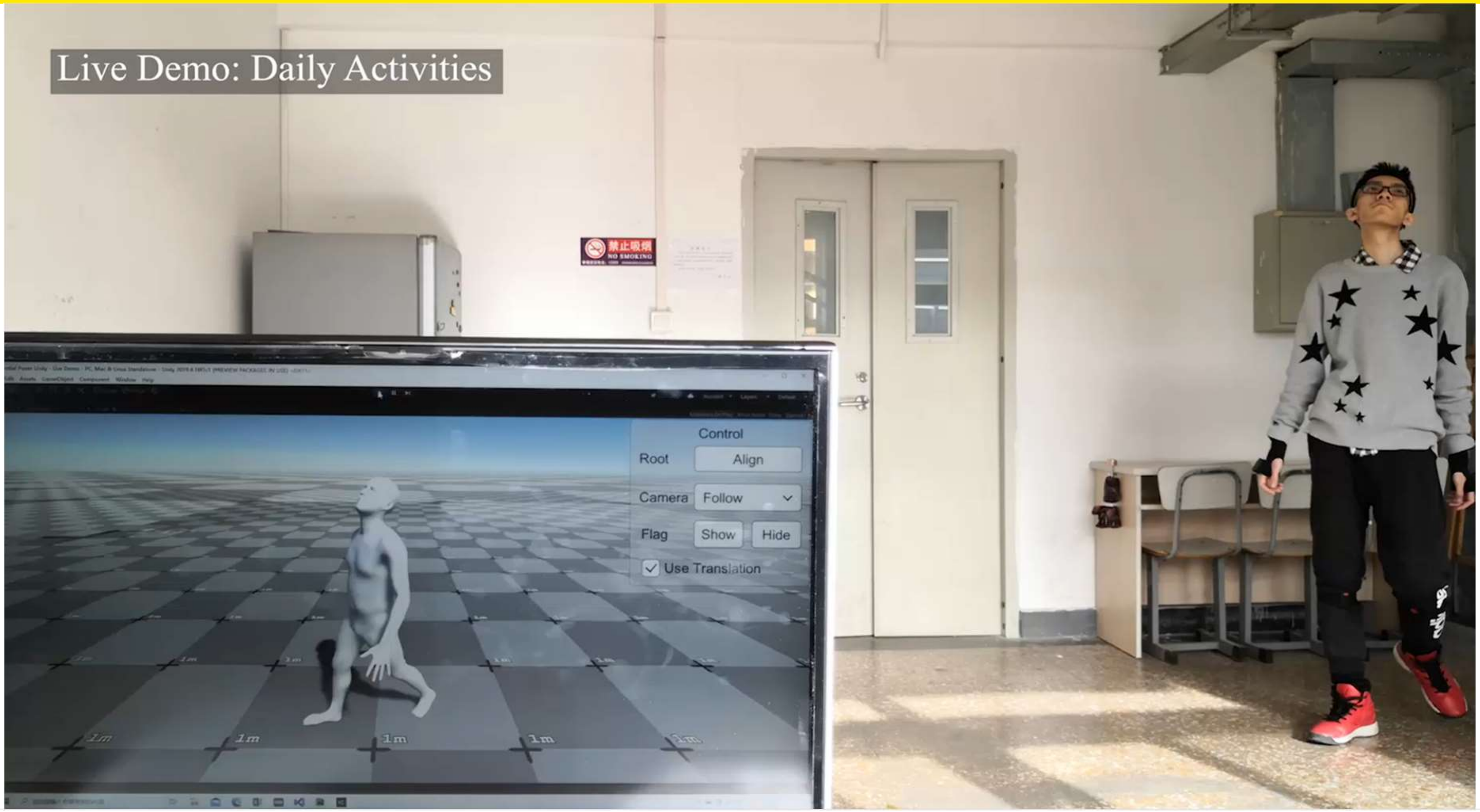
	DIP-IMU		TotalCapture	
	SIP Err (deg)	Jitter ($10^2 m/s^3$)	SIP Err (deg)	Jitter ($10^2 m/s^3$)
I→P	14.43 (±7.77)	2.50 (±3.42)	23.16(±9.00)	3.34 (±5.72)
I→LJ→P	14.35 (±7.75)	2.22 (±3.32)	17.71 (±7.89)	2.90 (±5.09)
I→AJ→P	14.29 (±7.30)	1.23 (±1.82)	19.76 (±8.05)	1.60 (±2.94)
I→LJ→AJ→P	**13.97 (±6.77)**	**1.19 (±1.76)**	**14.95 (±6.90)**	**1.57 (±2.93)**

Evaluation of the cross-layer connections of IMU data

	DIP-IMU		TotalCapture	
	SIP Err (deg)	Mesh Err (cm)	SIP Err (deg)	Mesh Err (cm)
S2 w/o IMUs	17.06 (±7.29)	6.44 (±3.38)	18.51 (±7.31)	6.84 (±3.61)
S3 w/o IMUs	15.66 (±7.53)	6.50 (±3.51)	15.75 (±7.18)	6.83 (±3.67)
Ours	**13.97 (±6.77)**	**5.83 (±3.21)**	**14.95 (±6.90)**	**6.36 (±3.47)**

RESULTS: IN-THE-WILD TEST

清華大學
Tsinghua University



Xinyu Yi

Yuxiao Zhou

Feng Xu

Thank you!

Paper

Project Page