Digital Humans

Zhizheng Liu and Joe Lin

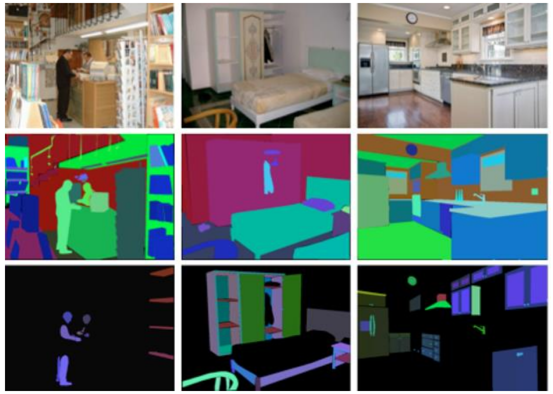
Research Goal

- Build **intelligent agents** that understand and interact with humans in outdoor environments (especially in urban contexts)
  - **Problem: Lack of human-scene and human-agent interaction data**

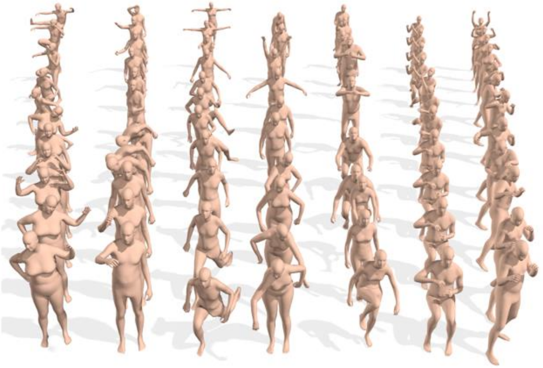
Scene

Human

Agent



ADE20k



AMASS

MetaDrive + MetaUrban

Q: How do we get human-scene interaction data and learn from it?

A: CityWalkers and PedGen

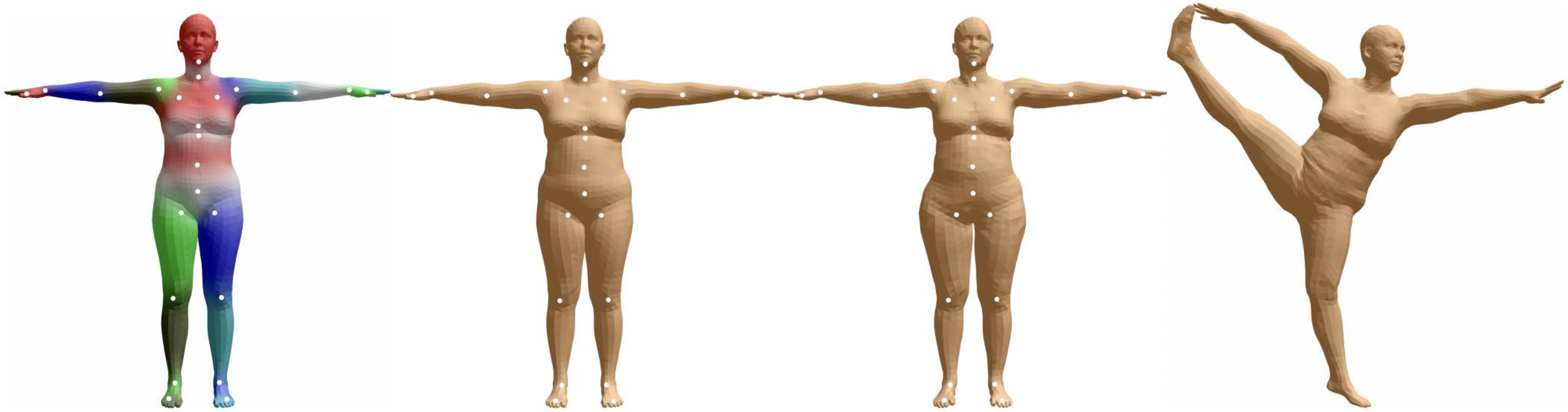
The Sneaky Gnome

Learning to Generate Diverse Pedestrian Movements from Web Videos with Noisy Labels

Zhizheng Liu, Joe Lin, Wayne Wu, Bolei Zhou

Representing Humans with SMPL

- Parameterized encoding of a 3D human mesh
  - Start with mean template shape $\bar{\mathbf{T}}$
  - Body **shape** parameters $\vec{\beta}$ (vector elements represent for ex. height, weight)
  - Body **pose** parameters $\vec{\theta}$ (rotation of each joint)
  - Global **orientation** and **translation** (pose in world space)

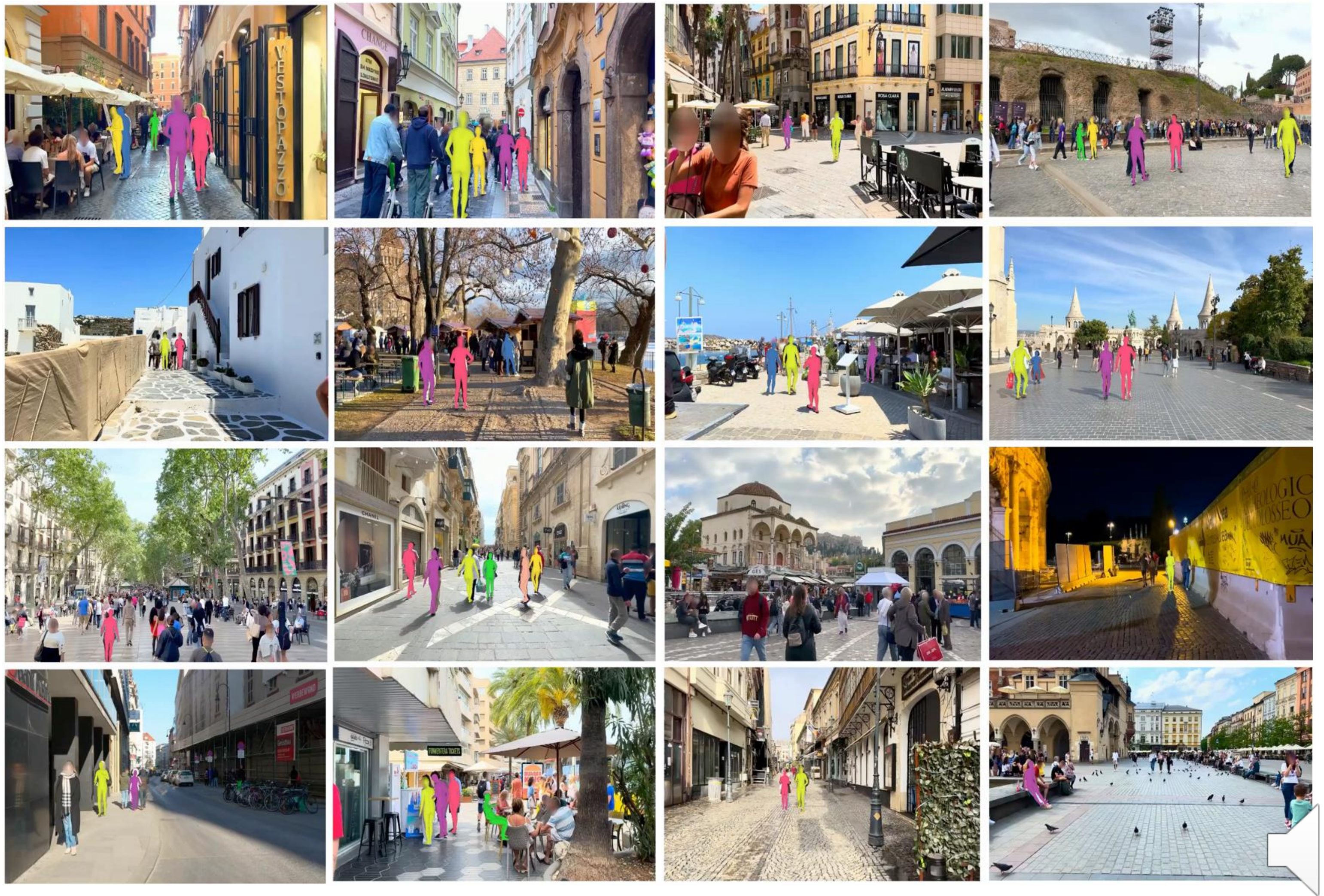


(a) $\bar{\mathbf{T}}, \mathcal{W}$ **(b)** $\bar{\mathbf{T}} + B_S(\vec{\beta}), J(\vec{\beta})$ **(c)** $T_P(\vec{\beta},\vec{\theta}) = \bar{\mathbf{T}}+B_S(\vec{\beta})+B_P(\vec{\theta})$ **(d)** $W(T_P(\vec{\beta},\vec{\theta}), J(\vec{\beta}), \vec{\theta}, \mathcal{W})$

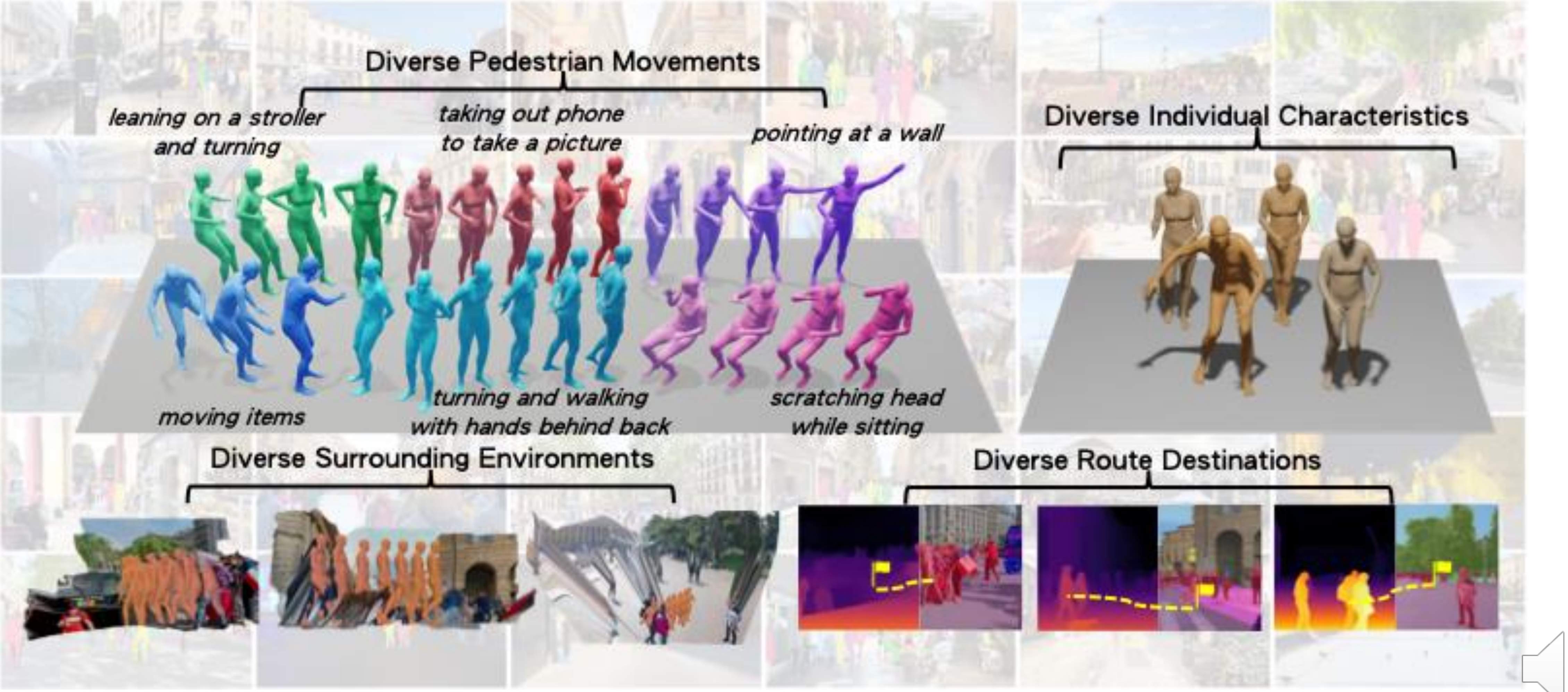
SMPL: A Skinned Multi-Person Linear Model. SIGGRAPH Asia 2015.

CityWalkers: Capturing Diverse Real-World Pedestrian Movements

- 30.8 hours of high-quality web videos with human motion pseudo-labels (SMPL)
- Diverse individual characteristics, human movements, urban environments, and route goals

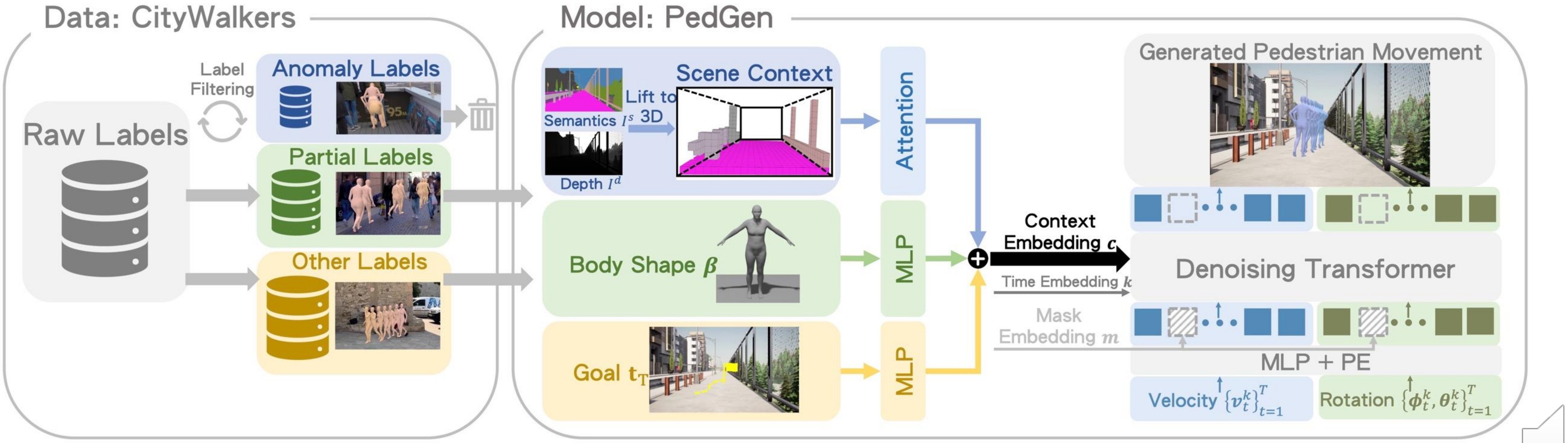


CityWalkers: Capturing Diverse Real-World Pedestrian Movements

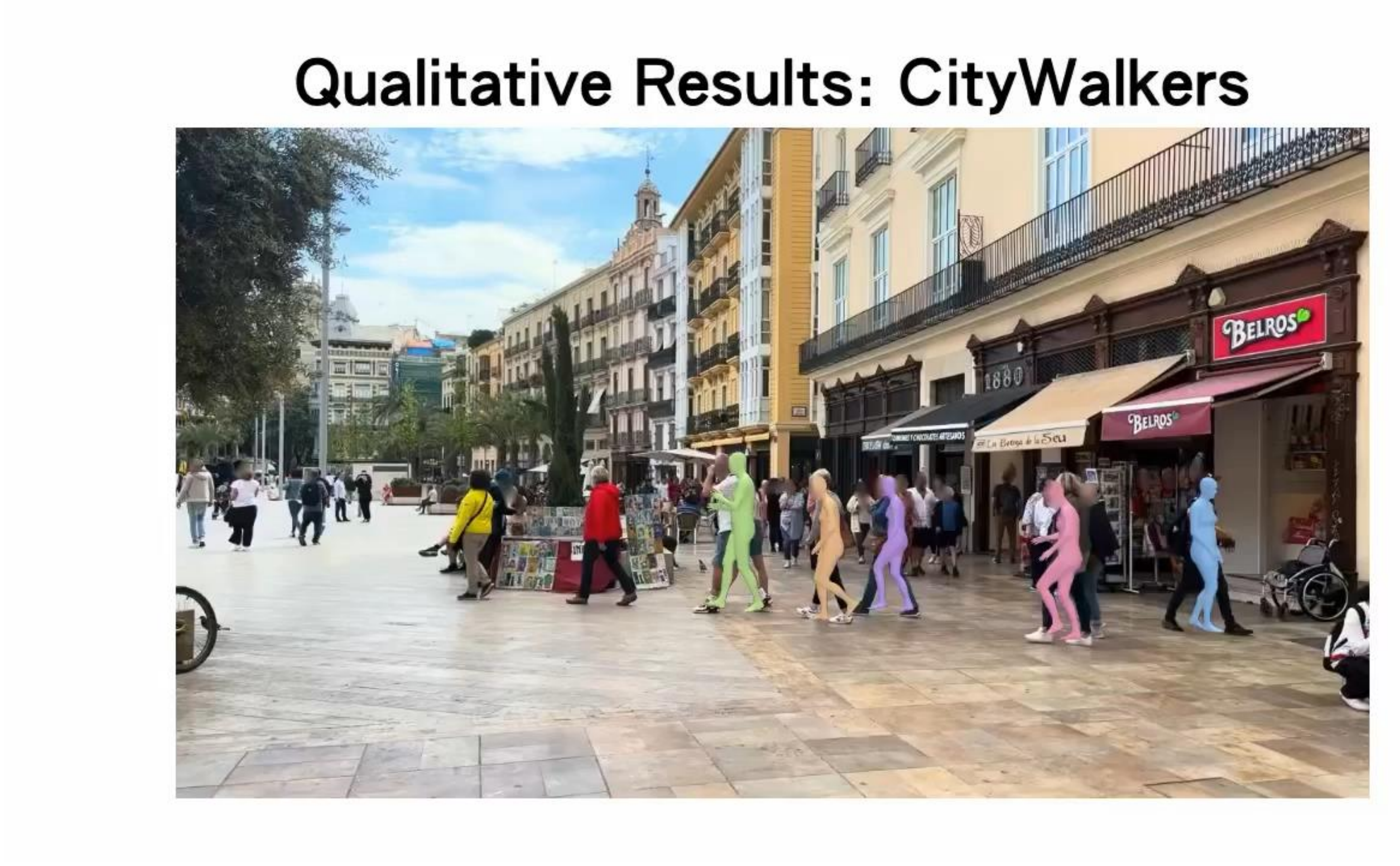

Diverse Pedestrian Movements
leaning on a stroller and turning
taking out phone to take a picture
pointing at a wall
moving items
turning and walking with hands behind back
scratching head while sitting
Diverse Individual Characteristics
Diverse Surrounding Environments
Diverse Route Destinations

PedGen: Generative Model for Context-Aware Pedestrian Motion

- Diffusion model conditioned on scene context
  - Note: Same as image diffusion models from lecture. We add noise and denoise a **motion vector** instead of an image.

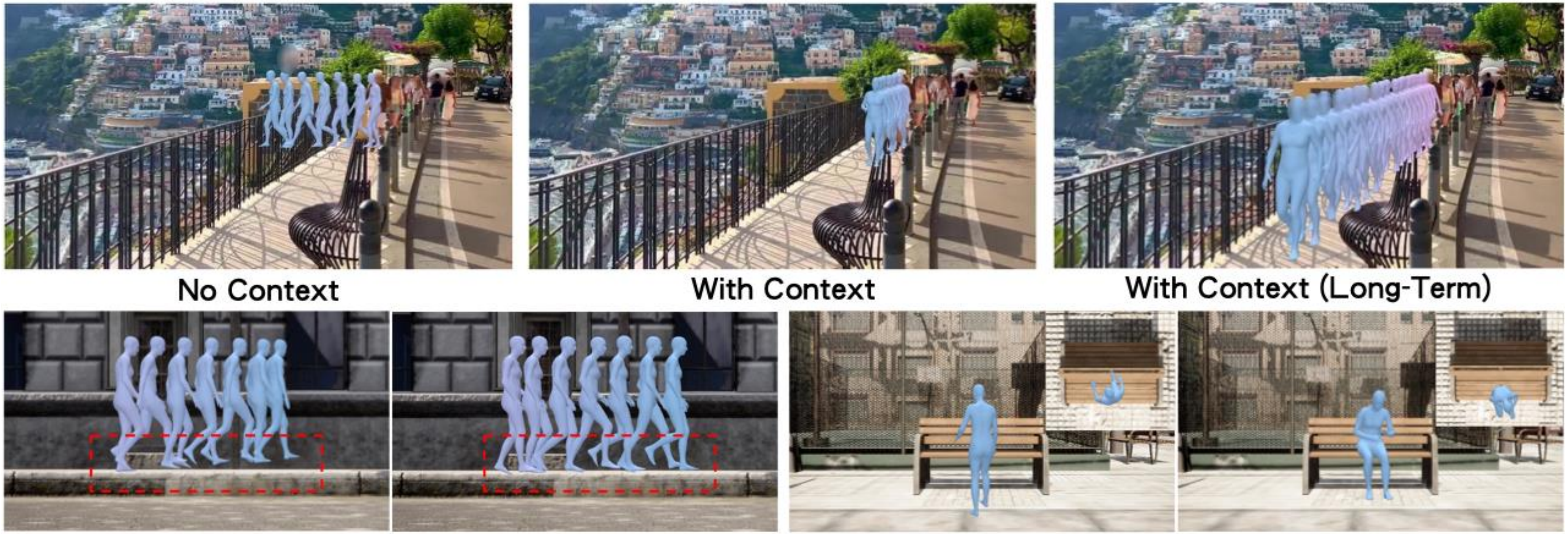


Generation Results in Real-World Scenarios and CARLA Simulation



Qualitative Results: CityWalkers

Qualitative Comparison of Training PedGen with Scene Context Factors



No Context

With Context

With Context (Long-Term)

No Context

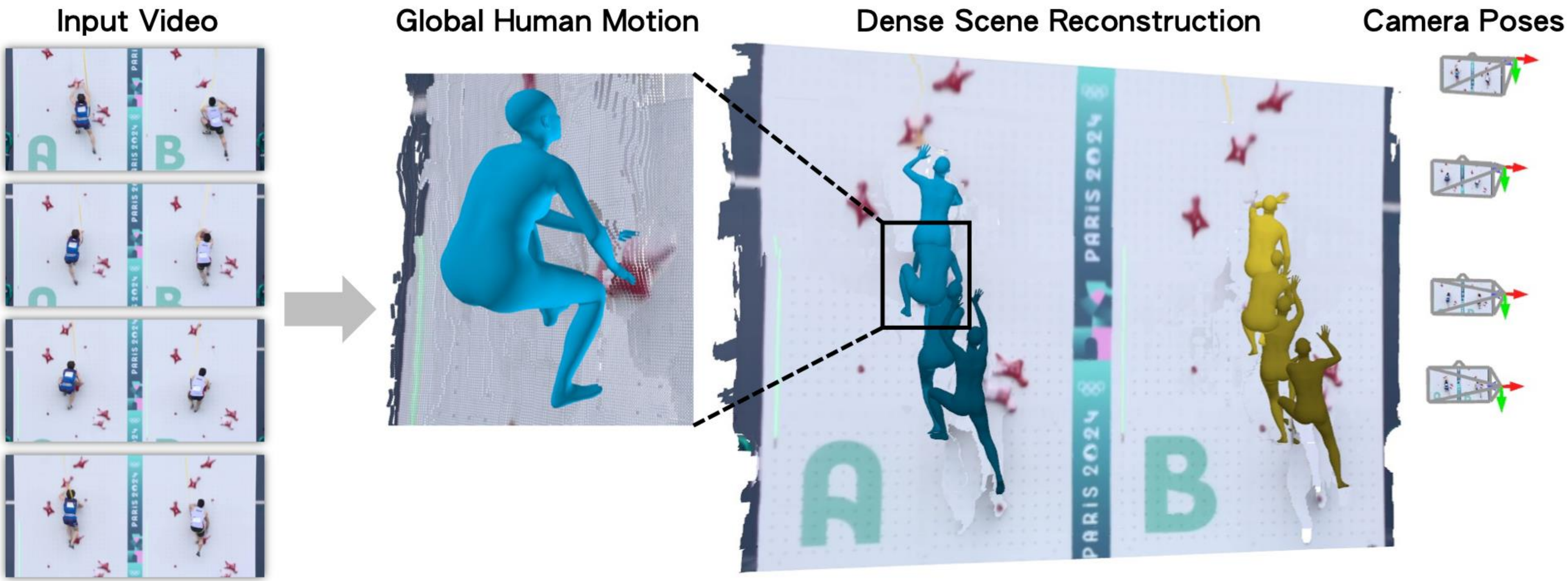
With Context

No Context

With Context

Q: How do we improve the pseudo-label quality?

A: JOSH

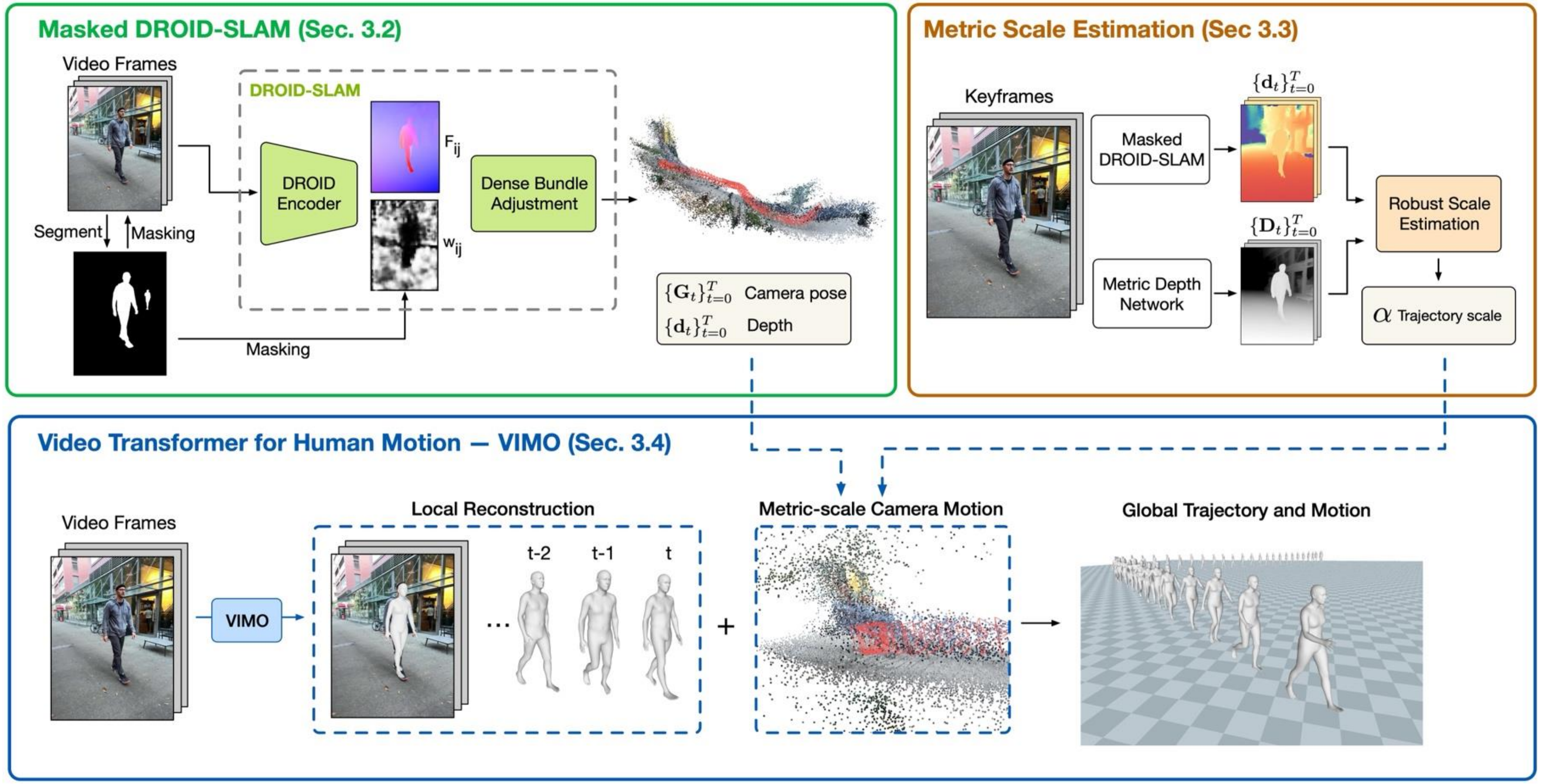


Joint Optimization for 4D Human-Scene Reconstruction in the Wild

Zhizheng Liu, Joe Lin, Wayne Wu, Bolei Zhou

Human Motion Estimation

- Predicting human movements (in SMPL) given a sequence of images
  - **TRAM** – Combines SLAM and depth estimator to predict metric-scale global trajectory



TRAM: Global Trajectory and Motion of 3D Humans from in-the-wild Videos. Arxiv 2024.

Scene Reconstruction

- Reconstructing the 3D scene given a set of images
  - Note: Most of the time we don't have access to camera intrinsics and extrinsics
- Classical CV approach
  - **Structure from Motion** (SfM) – Estimate camera poses and triangulate points with **stereo** and **epipolar** geometry
  - **COLMAP** – Dense reconstruction using SfM and bundle adjustment
  - **Problem: Computationally expensive and inaccurate with sparse scene views**

Scene Reconstruction

- Deep learning approach
  - **DUSt3R** – Dense and unconstrained stereo (two cameras) 3D reconstruction
  - **MASt3R** – Learning additional feature descriptor for efficient 3D point matching
  - **MASt3R-SfM** – Extending pipeline to videos and optimize dense depth, camera intrinsics, camera pose, global scale

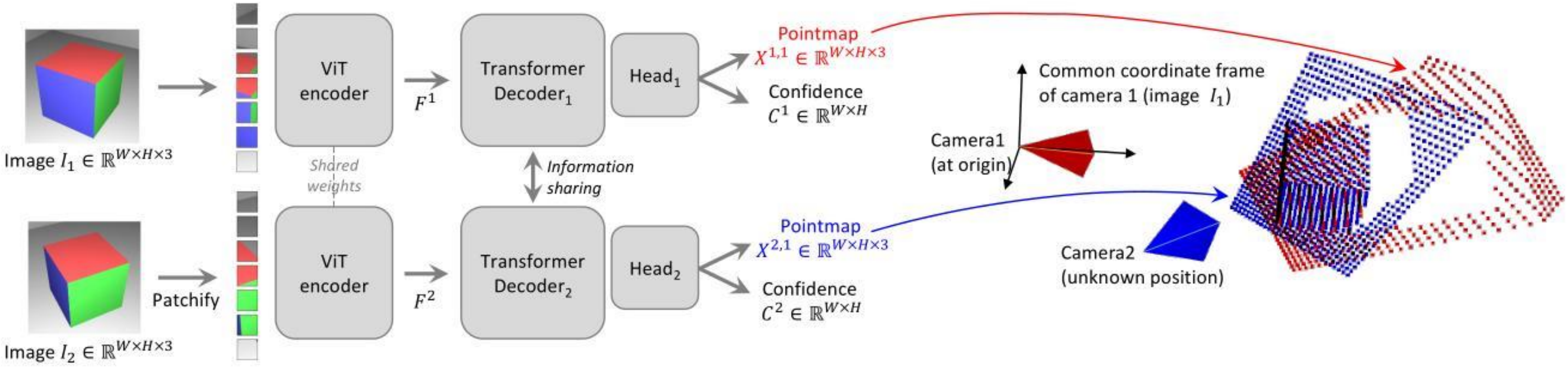
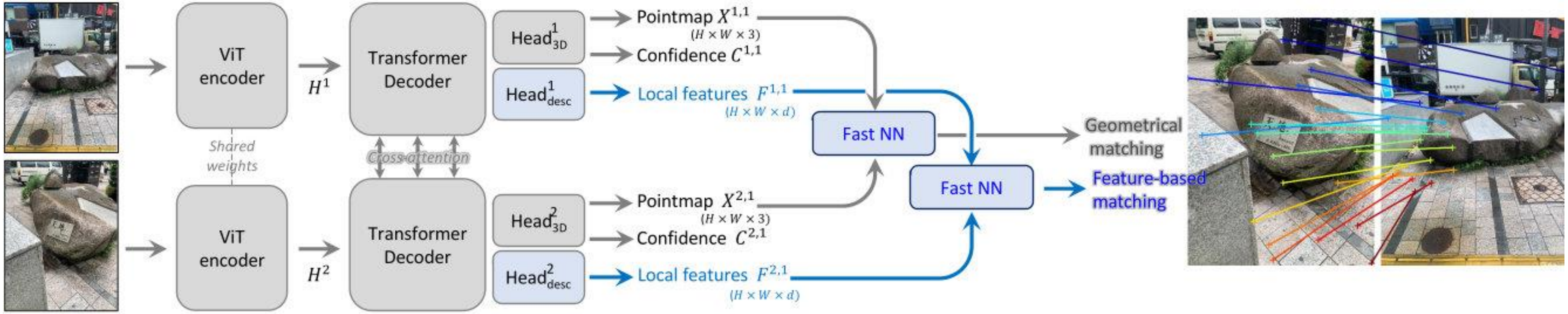


Image $I_1 \in \mathbb{R}^{W\times H\times 3}$
Image $I_2 \in \mathbb{R}^{W\times H\times 3}$
Patchify
ViT encoder
ViT encoder
Shared weights
F^1
F^2
Transformer Decoder$_1$
Transformer Decoder$_2$
Information sharing
Head$_1$
Head$_2$
Pointmap $X^{1,1} \in \mathbb{R}^{W\times H\times 3}$
Confidence $C^1 \in \mathbb{R}^{W\times H}$
Pointmap $X^{2,1} \in \mathbb{R}^{W\times H\times 3}$
Confidence $C^2 \in \mathbb{R}^{W\times H}$
Common coordinate frame of camera 1 (image I_1)
Camera1 (at origin)
Camera2 (unknown position)

DUSt3R: Geometric 3D Vision Made Easy. CVPR 2024.

Scene Reconstruction

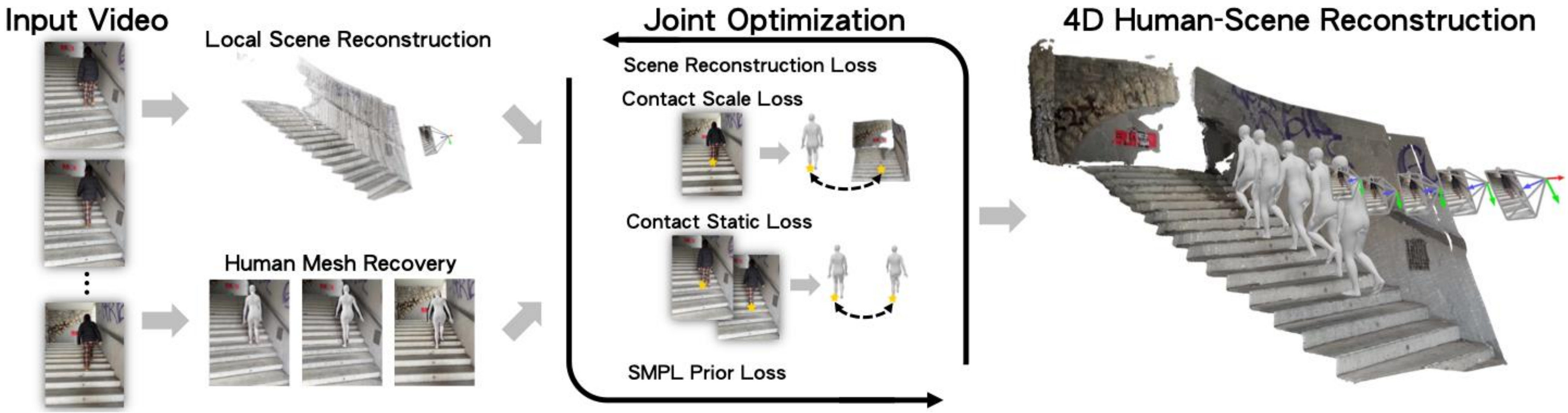
- Deep learning approach
  - **DUSt3R** – Dense and unconstrained stereo (two cameras) 3D reconstruction
  - **MASt3R** – Learning additional feature descriptor for efficient 3D point matching
  - **MASt3R-SfM** – Extending pipeline to videos and optimize dense depth, camera intrinsics, camera pose, global scale



Grounding Image Matching in 3D with MASt3R. Arxiv 2024.

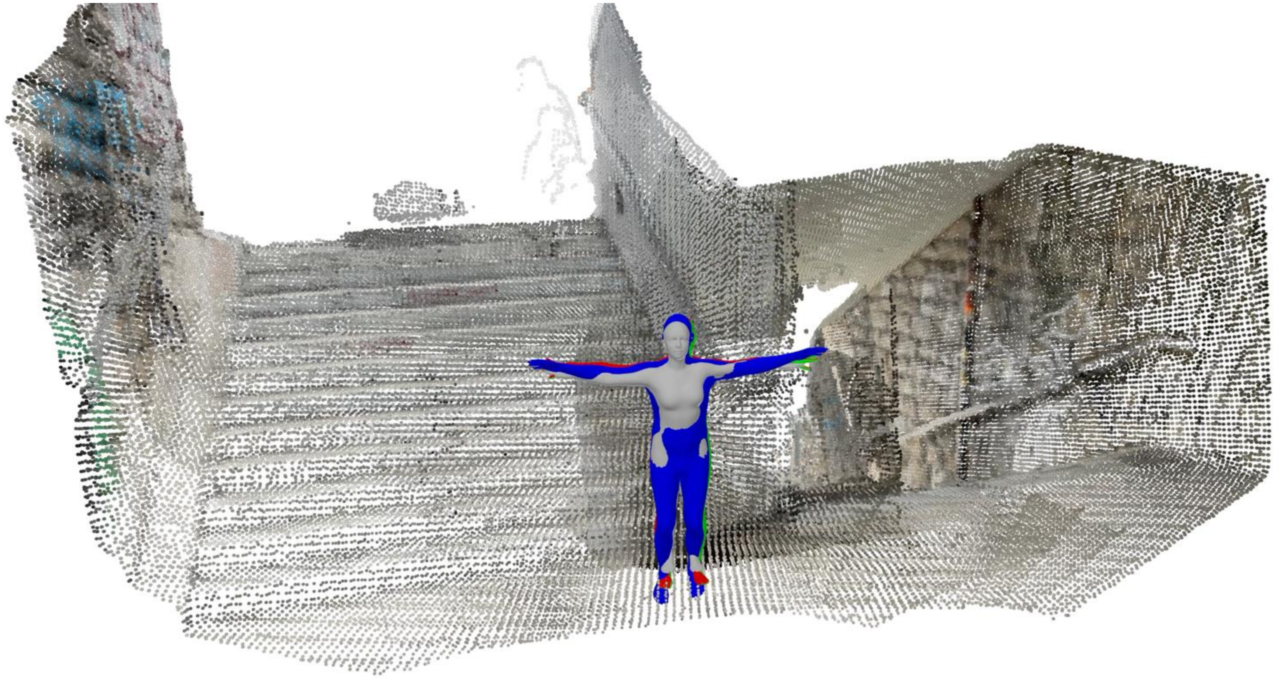
JOSH: Joint Optimization of Scene Geometry and Human Motion

- Initialize with MASt3R and TRAM + obtain pseudo-depths
- Add human into joint optimization
  - Contact Scale Loss $\mathcal{L}_{c1}$ – matches human-scene contact points
  - Contact Static Loss $\mathcal{L}_{c2}$ – ensures contact points remain stationary across time



Evaluation Results for Global Human Motion on EMDB Dataset

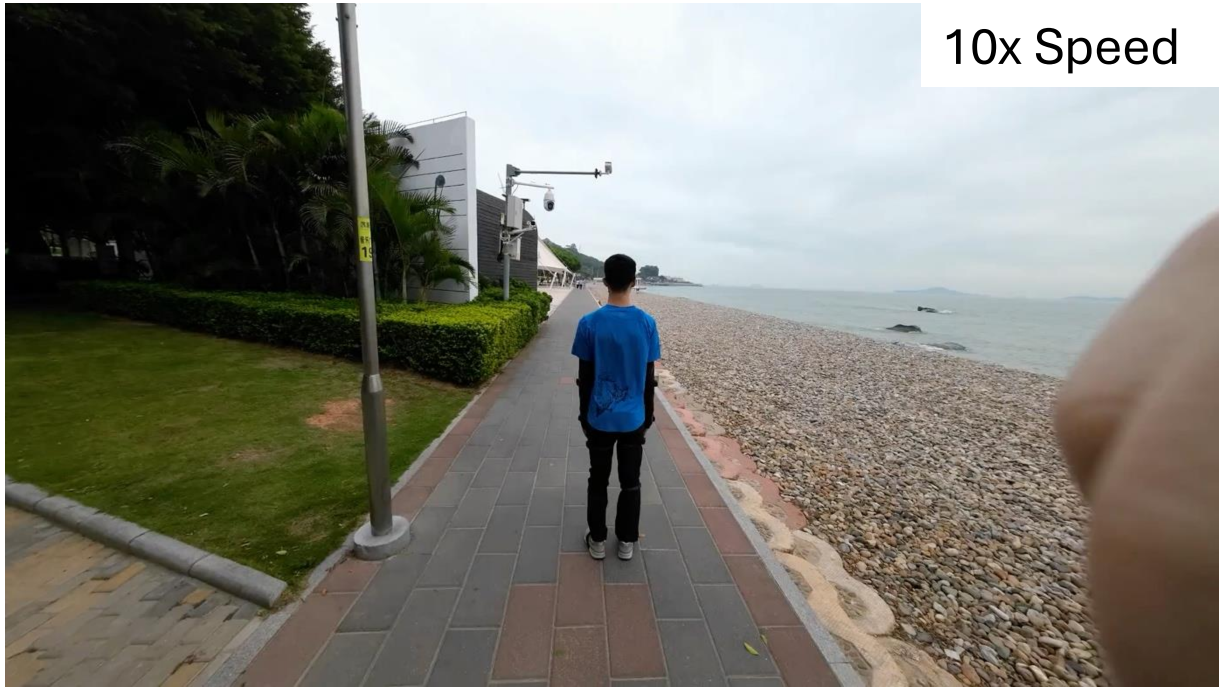
Input Video from EMDB



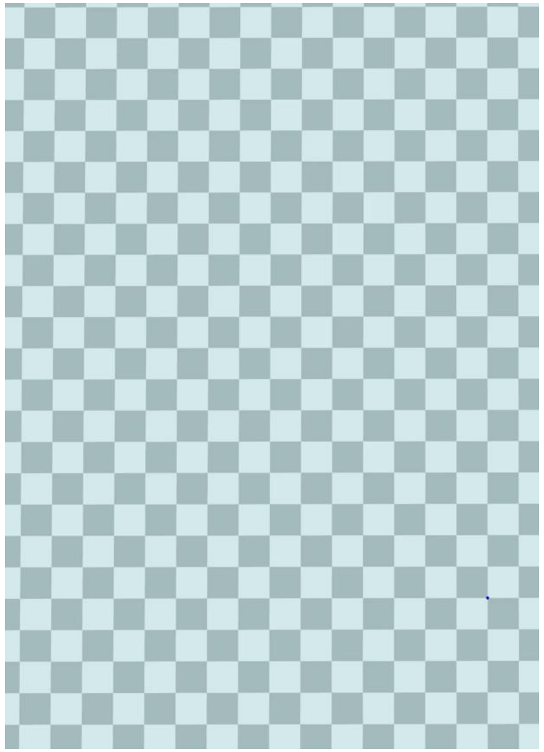
⬜ Ground Truth 🟥 JOSH (ours) 🟩 TRAM 🟦 WHAM

*Dense scene reconstruction result is from JOSH as a reference, as no GT is available for EMDB and WHAM and TRAM cannot reconstruct the scene

Evaluation Results for Global Human Motion on SLOPER4D Dataset



Input Video from SLOPER4D



⬜ Ground Truth

🟥 JOSH (ours)

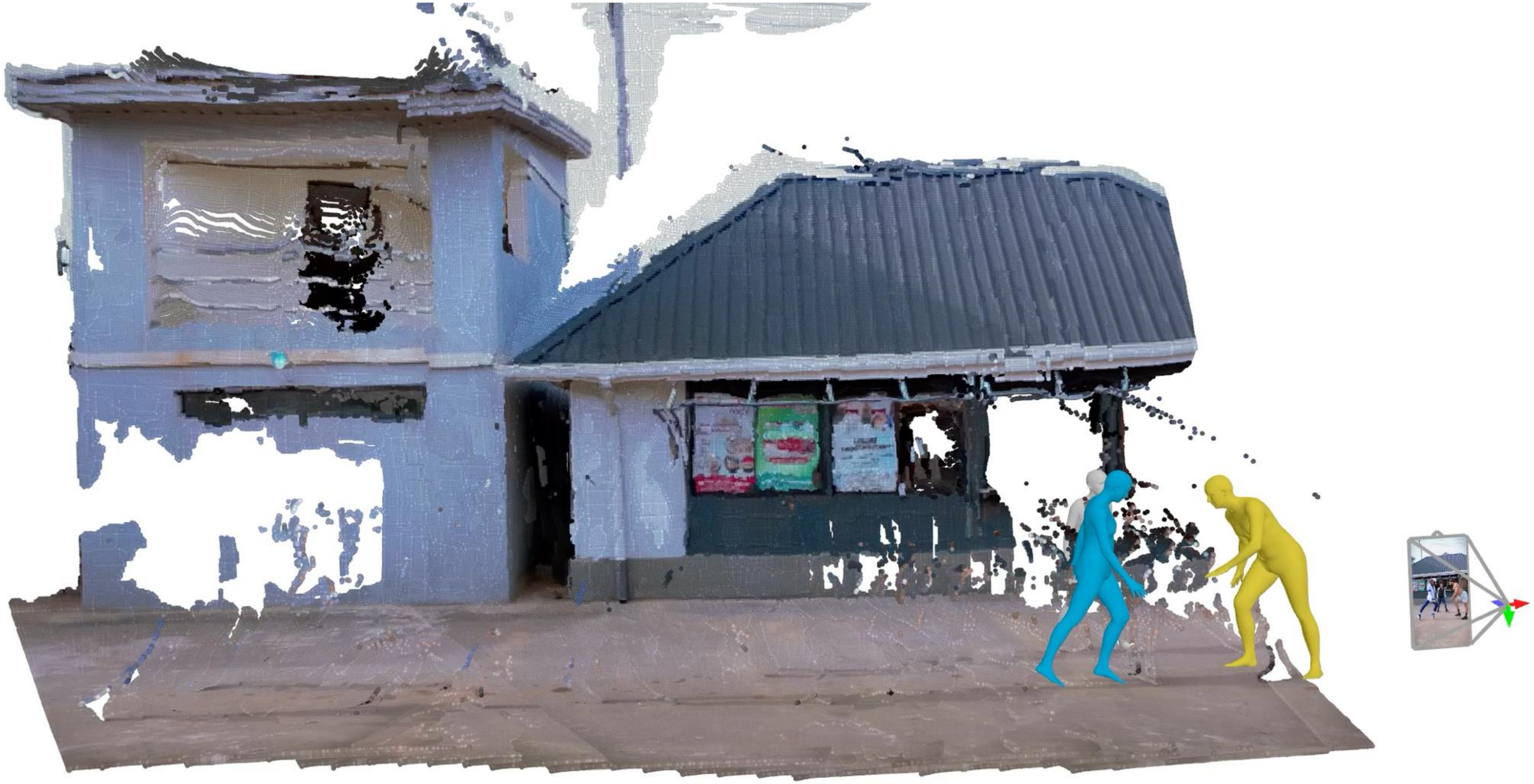
🟩 MASt3R

🟦 MonST3R

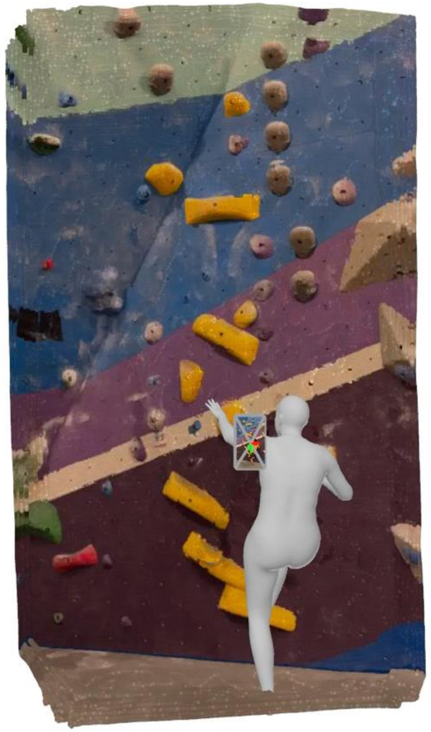
*This is a long sequence, and the checkerboard tile size is 10m x 10m

Human and Scene Reconstruction Results on in-the-wild Videos

Input Web Video



4D Human-Scene Reconstruction



Thank you for listening!

If any of this interests you, feel free to reach out to Zhizheng Liu.