Long-term Human Motion Prediction with Scene Context

- Zhe Cao, Hang Gao , Karttikeya Mangalam , Qi-Zhi Cai, Minh Vo, and Jitendra Malik

- Prakash K Naikade,
prna00001@stud.uni-saarland.de

Problem

- **Given:** *N*-step 2D human pose history $\mathbf{X}_{1:N}$ and Scene Image **I** (N^{th} video frame)
- **Predict:** next *T*-step 3D human poses together with their locations ($\mathbf{Y}_{N+1:N+T}$)

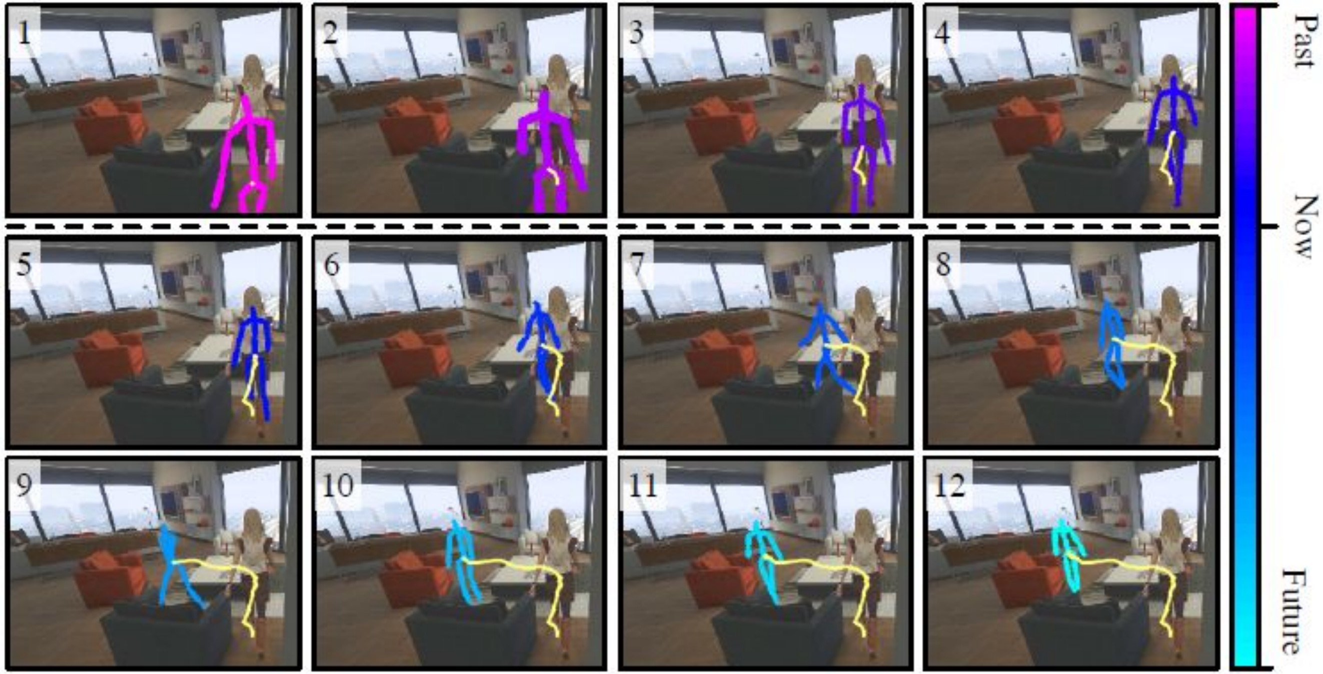


Fig 1: Problem description - Long-term 3D human motion prediction

Motivation

- Human movement is,
    1. Goal directed
    2. Constrained by environment
    3. Multimodal future

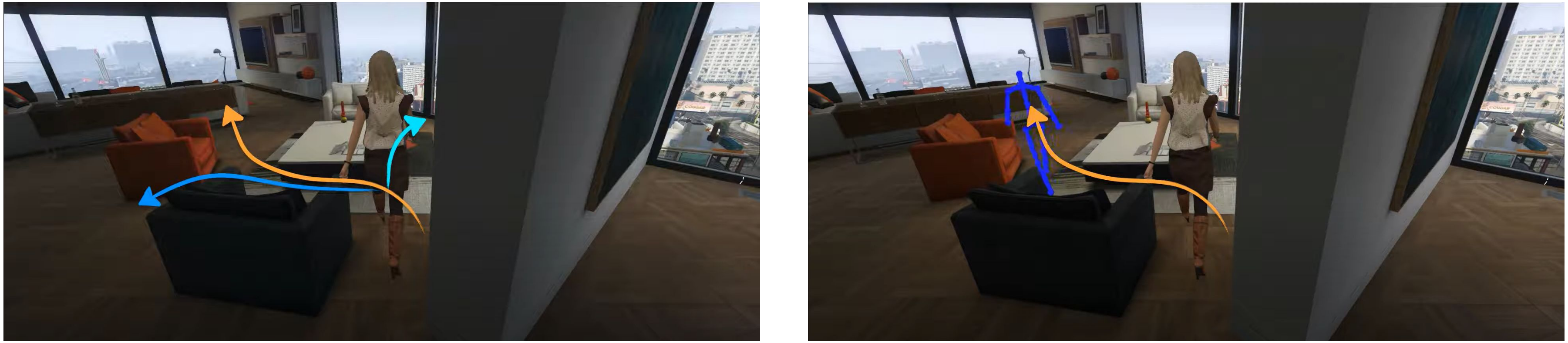


Fig 2: Predict long-term human motion with scene context

Motivation

Fig 3: Smart glass for vision impaired people

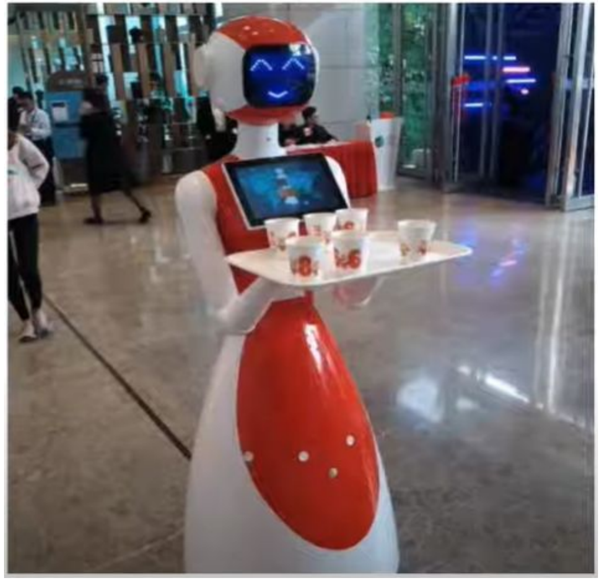


Fig 4: Home robot serves nearby passengers

Key Contribution

1. **Formulated a new task**: Long-term 3D human motion prediction with scene context in terms of 3D poses and 3D locations.

2. **GTA-IM Dataset:** Created new synthetic dataset with diverse recordings of human-scene interaction and clean annotations.

    - Renderer Scripting - To generate one million RGBD frames of 1920 × 1080 resolution
    - Labels generated automatically:
        - RGBD Video
        - 3D human pose
        - Camera pose
        - Global coordinates of paths
        - Action labels
        - Human Segmentation

3. **Developed a novel three-stage computational framework:** Framework utilizes scene context for goal-oriented motion prediction.

Proposed Solution

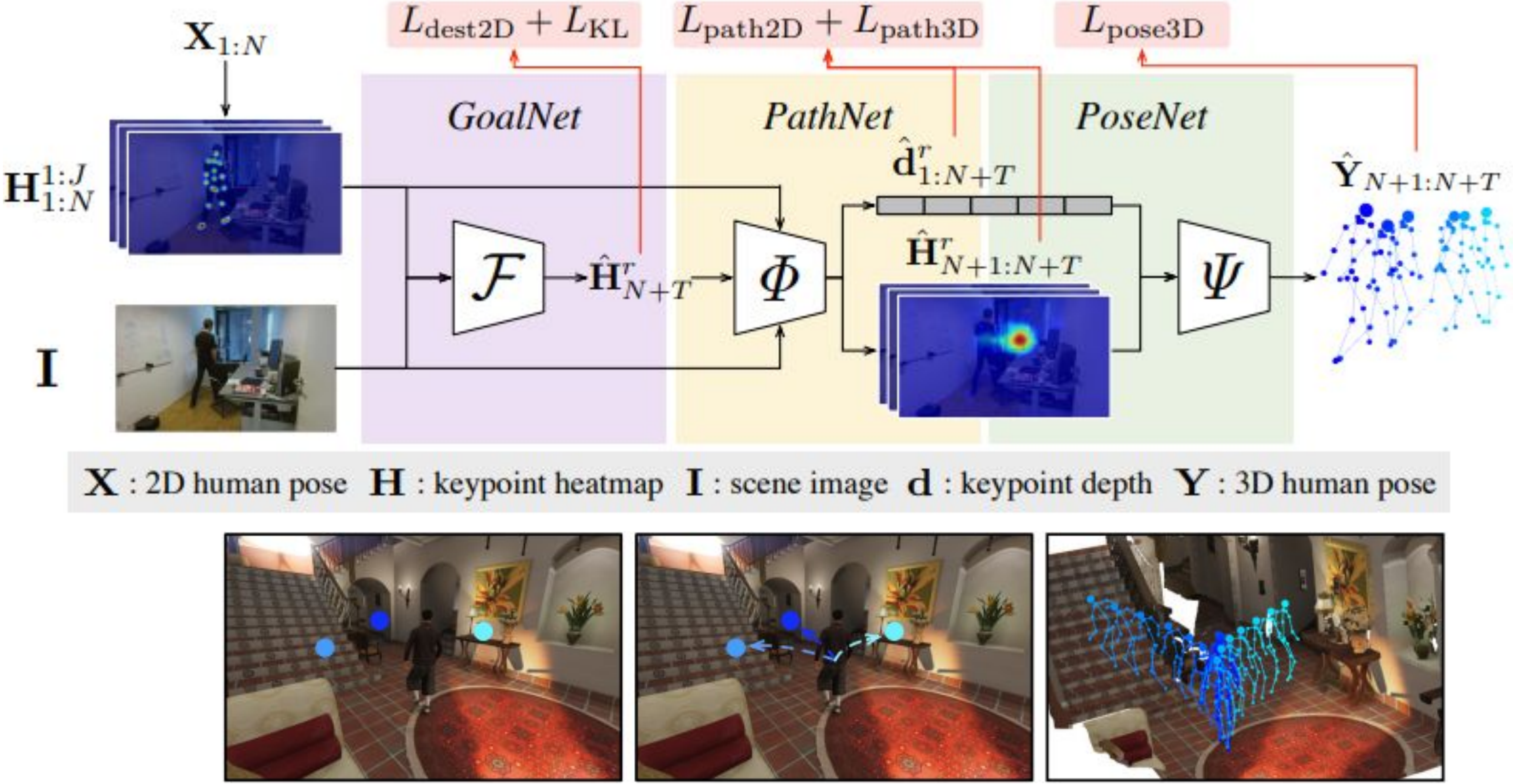

$\mathcal{L}_{\text{dest2D}} + \mathcal{L}_{\text{KL}}$
$\mathcal{L}_{\text{path2D}} + \mathcal{L}_{\text{path3D}}$
$\mathcal{L}_{\text{pose3D}}$
$\mathbf{X}_{1:N}$
GoalNet
PathNet
PoseNet
$\mathbf{H}^{1:J}_{1:N}$
$\hat{\mathbf{d}}^{r}_{1:N+T}$
$\hat{\mathbf{Y}}_{N+1:N+T}$
$\mathcal{F}$
$\hat{\mathbf{H}}^{r}_{N+T}$
Φ
$\hat{\mathbf{H}}^{r}_{N+1:N+T}$
Ψ
$\mathbf{I}$
$\mathbf{X}$: 2D human pose $\mathbf{H}$: keypoint heatmap $\mathbf{I}$: scene image $\mathbf{d}$: keypoint depth $\mathbf{Y}$: 3D human pose

Fig 4: Proposed Pipeline and Network architecture

GoalNet

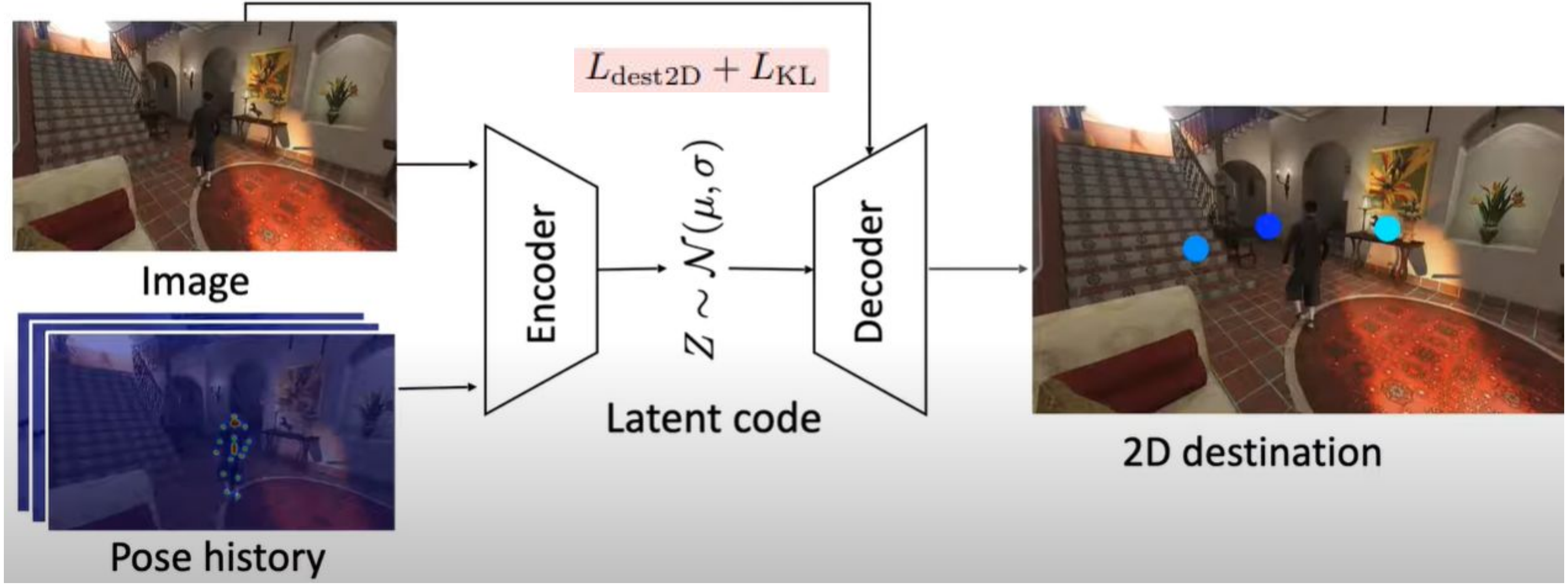
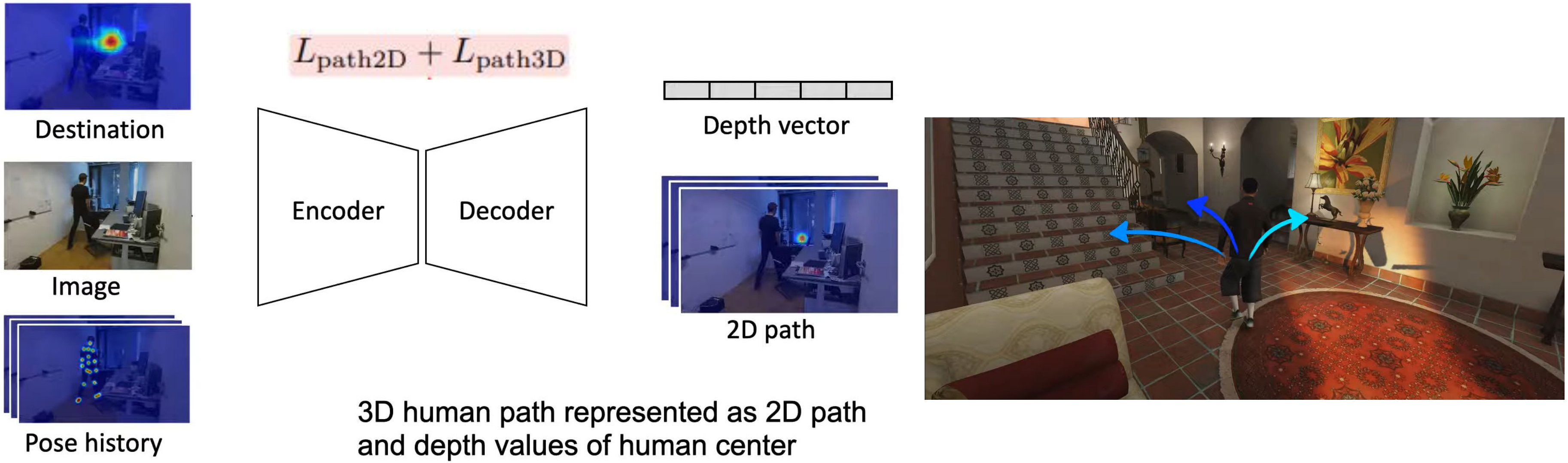

$L_{\text{dest2D}} + L_{\text{KL}}$
Image
Pose history
Encoder
$Z \sim \mathcal{N}(\mu, \sigma)$
Latent code
Decoder
2D destination

Fig 5: GoalNet - Predicting 2D Movement Destination

PathNet



3D human path represented as 2D path and depth values of human center

Fig 6: PathNet - Predicting 3D path towards each destination

PoseNet

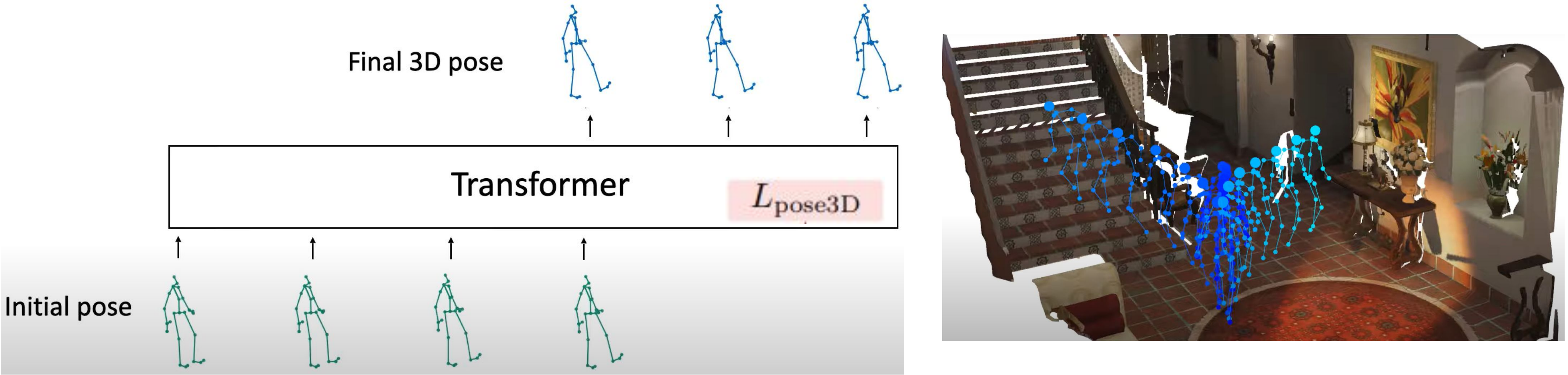


Fig 7: PathNet - Generating 3D pose following the path

Evaluation

1. **Average 3d distance between the two second long prediction and the ground truth**

Limitations

1. **Resulting 3D poses may not strictly meet all physical constraints:** Use multi-view/temporal images.

2. **Dynamic objects and multiple moving people**

3. **Naturalness and feasibility of the stochastic human motion predictions**

4. **Domain gap between synthetic and realistic image dataset**

Thank You !

Questions Please !