

PEDESTRIAN MOTION RECONSTRUCTION: A LARGE-SCALE BENCHMARK VIA MIXED REALITY RENDERING WITH MULTIPLE PERSPECTIVES AND MODALITIES

Yichen Wang¹, Yiyi Zhang^{1*}, Xinhao Hu¹, Li Niu¹, Jianfu Zhang^{1*}, Yasushi Makihara²,
Yasushi Yagi³, Pai Peng⁴, Wenlong Liao⁴, Tao He⁴, Junchi Yan¹, Liqing Zhang^{1*}

¹Department of CSE & MoE Key Lab of AI, Shanghai Jiao Tong University

²Independent Researcher, Osaka, Japan

³Department of Intelligent Media, SANKEN, Osaka University

⁴Cowa Tech. Ltd.

{rachal, yi95yi, c.sis, yanjunchi}@sjtu.edu.cn, zhang-lq@cs.sjtu.edu.cn

<https://github.com/coding-rachal/PMRDataset>

APPENDIX

A DETAILS OF DATASET COLLECTION SYSTEM

A.1 HARDWARE AND CONFIGURATION

OptiTrack MoCap System. Following industry standards, we deploy multiple optical cameras in a green screen environment (adjusting the number as needed for calibration) and transmitted motion capture data to work stations via an Ethernet network.

Work Stations. We employ 2 work stations to balance the computational load (referred to as Work Station 1 and 2, respectively). These two devices are also connected through Ethernet. Work Station 1 primarily handles Motive and MotionBuilder, while the task of Work Station 2 is to run the CARLA Dosovitskiy et al. (2017) simulation environment and other essential applications.

- **Optitrack Motive 3.1.** Motive is a compatible application for optical motion capture hardware, providing real-time motion data with high accuracy which will be processed into SMPL-X Loper et al. (2015) format.
- **Autodesk MotionBuilder 2020.** It is a 3D character animation software capable of capturing, editing, and playing back intricate animations. Our utilization is specifically focused on its powerful retargeting function.
- **Tentacle Setup.** To align the temporal information of the SMPL-X Loper et al. (2015) ground truth derived from MoCap and other data recorded in Work Station 2, we synchronize the time using Tentacle.
- **OptiTrack OpenVR Driver.** In terms of the VR HMD worn by the subjects, we use HTC Vive and replace its matched trackers with the OpenVR Driver as the positioning system, ensuring the normal operation of the HMD.
- **SteamVR.** It collaborates with the Unreal Engine to provide VR functionality, and directly participates in the data transmission of HMD device.
- **CARLA 0.9.14 on Unreal Engine 4.** CARLA Dosovitskiy et al. (2017) is a distinguished autonomous driving simulation research platform. On the CARLA platform, we construct and execute specific scenarios through scripts, recording data that includes RGB videos from third-person and egocentric perspectives, extrinsic of in-car cameras, and LiDAR data.

Construction of Virtual Scenarios and Scripts. We deploy specific traffic scenarios on CARLA Dosovitskiy et al. (2017) by executing scripts. The subject puts on a HMD and starts off as a pedestrian in an empty street. As the scripts are initiated, vehicles and sensors are generated

* Corresponding author.

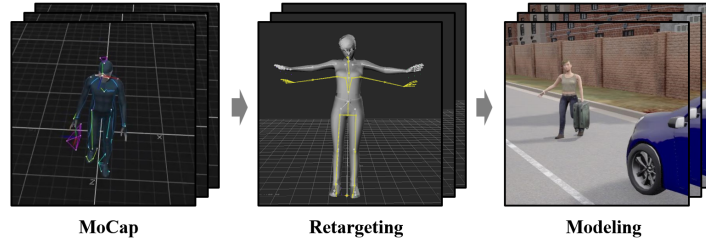


Figure 1: Pipeline of Avatar MoCap, Motion Retargeting, and Modeling. In the retargeting step, it’s recommended to ensure the best possible alignment of the two skeletons(the yellow one and the gray one). The situation here is for illustration only.

at predetermined locations around the subject, capturing his/her actions and reactions. The script includes:

- **Actors Generating.** Both vehicles and sensors are defined as actors in CARLA. Vehicles are generated at specified locations around the pedestrians’ spawn points and range of activities in the world. Sensors are placed on the vehicles at suitable positions and angles.
- **World Ticking in Synchronous Mode.** Once the vehicles and sensors are generated, the entire world will run in synchronous mode. Vehicles operate following pre-defined programs. The sensors capture data at the current timestamp on each tick and store it in files.

A.2 DATA ACQUISITION PIPELINE

Time Synchronizing. The initial step involves synchronizing the time between Work Station 2 and the MoCap system. Work Station 2 should first synchronize with Internet time to calibrate the entire capture environment. Subsequently, the Tentacle device retrieves the time from Work Station 2, and then connects to the OptiTrack MoCap Sync module, providing a 24Hz SMPTE timecode, upsampled to 120Hz in MoCap system, and exports as ground truth. During the execution of subsequent scripts, virtual world sensors record the system time of Work Station 2 by each tick. This ensures synchronization between the ground truth timestamps exported from Work Station 1 and timestamps of all other data exported from Work Station 2.

Character MoCap, Motion Retargeting and Streaming. We employ character templates with 41 markers or an extended version with 49 markers. The latter incorporates preliminary hand motion capture. The 49-marker version enhances the realism and naturalness of contact and interaction between characters and objects, as discussed later. After the creation of the subject’s skeleton in Motive on Work Station 1, the skeleton stream is transmitted to MotionBuilder on the same device. The Motive-based skeleton is then retargeted to the character skeleton in CARLA Dosovitskiy et al. (2017) definition. Finally, the character skeleton is streamed to Work Station 2’s CARLA UE4, serving as the animation source for the pedestrian character.

Object MoCap and Streaming. We introduce interactive objects into the virtual scenes (*e.g.*, suitcases and stools), each of which has a corresponding physical counterpart in the real world. The purpose is to allow subjects to be engaged in diverse and authentic actions while interacting with these objects. Specifically, we obtain 3D models of some objects from the OmniObject3D dataset Wu et al. (2023) and their physical counterparts. Reflective markers are attached to these physical objects, and rigid bodies are created for them in Motive on Work Station 1. The transformation data is then streamed to Work Station 2’s CARLA UE4 Dosovitskiy et al. (2017). To improve MoCap accuracy, markers should be attached asymmetrically to the object’s surface, and the resulting rigid body coordinates represent the centroids of these marker combinations instead of the model center. To mitigate deviations from the object model center, we simulate marker placement on the object, designate the object’s center as the centroid, and configure markers accordingly (refer to Fig. 2). Translation and rotation of characters and objects are accomplished by considering the subjects’ transformation relative to the origin of the scene in MoCap. As characters and objects are composited into the scene independently through various methods, we align their model spaces to the coincident origin and coordinate system of the world space. This alignment precisely mirrors

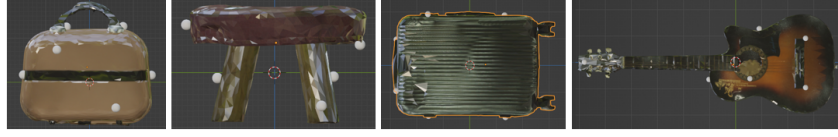


Figure 2: Simulating marks and adjusting models’ centers to facilitate reflecting transformations of real objects in the virtual environment accurately.

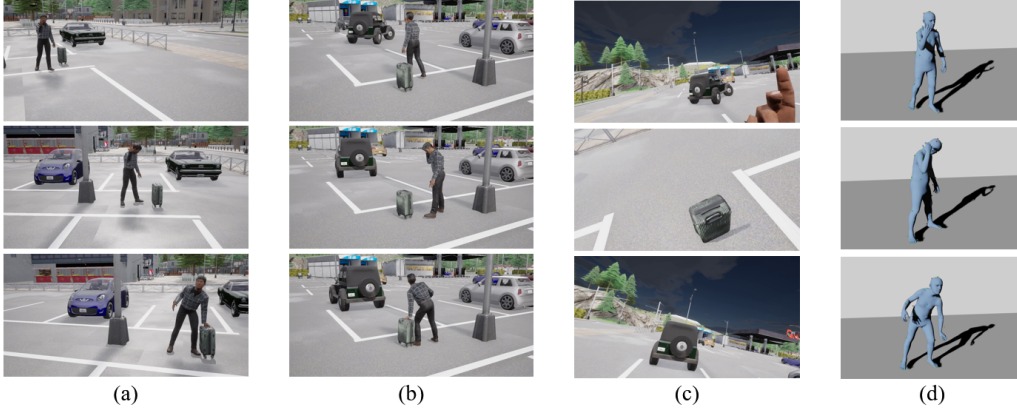


Figure 3: Visualization of collected sample sequence. (a)-(b). Two different third-person perspectives from cameras on the vehicle. (c) Egocentric perspective of the pedestrian. (d) SMPL-X groundtruth obtained from MoCap.

real-world interactions between characters and objects into the CARLA Dosovitskiy et al. (2017) environment.

Interfacing with the HMD. Similar to the process of adding objects, we create a rigid body for HMD in Motive. Then, the rigid body is streamed to the OpenVR Driver on Work Station 2 to convey data. To improve the recording effect of the pedestrian’s egocentric perspective, we consider it as a prerequisite for the HMD’s proper functionality during most of the collection. The slot at the midpoint of two eyes on the character’s skeleton serves as the actual source of positioning.

Scenario Running and Data Recording. With all the aforementioned preparations in place, the scenario script can be executed, initiating the data recording process. Initially, the subject will appear as a constructed pedestrian model in an empty street simulated by CARLA Dosovitskiy et al. (2017). As the script begins, vehicles are generated at specific locations around the subject, some equipped with sensors and others without. The vehicles follow preset logic, and sensors continuously record data. The subject is notified to commence activities based on common sense (e.g., in a scenario involving reversing into a parking space, the subject assists in vehicle reversing, employing gestures and movements akin to real-life, all of which are recorded by in-car sensors).

B DATASET DETAILS

In our research, 54 volunteers are recruited to participate in our data collection. Each volunteer is expected to choose several scenes we designed and react with our motion capture suit and VR HMD. The data acquisition protocol involved the recording of 9 sequences for each chosen scene, encompassing 3 distinct weather conditions and 3 specific car modes within the selected scene. Each recorded sequence contains 6 third-view RGB videos (with different perspectives and camera speeds), 1 egocentric video, and 2 LiDAR point sequences, as well as real-time annotation-free accurate labels (*i.e.*, (SMPL-X Loper et al. (2015), 3D/2D skeletons and bounding boxes, third-view camera extrinsic, semantic LiDAR labels, and LiDAR parameters) from MoCap and CARLA simulator. Our PMR dataset is characterized by remarkable diversity, spanning across different characters, weather

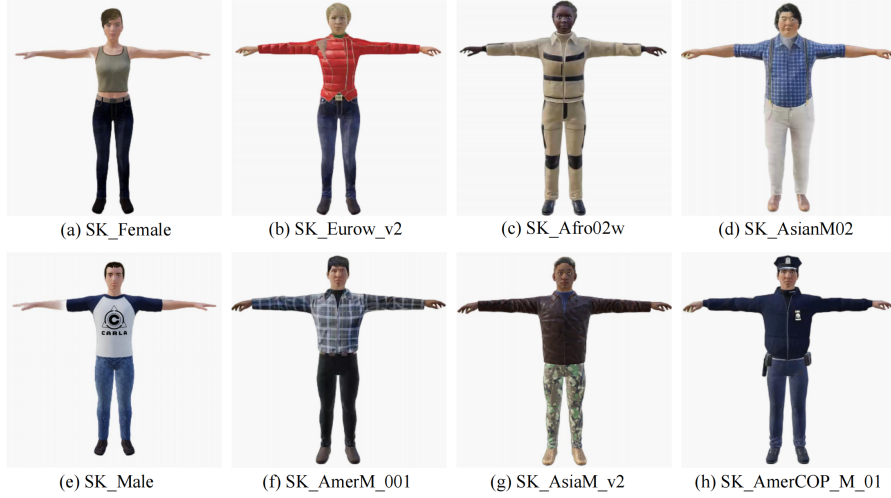


Figure 4: The characters used in our Pedestrian Motion Reconstruction (PMR) dataset.

conditions, and scenes, making it a valuable resource for exploring a wide range of scenarios and applications. For details:

- **Characters:** As depicted in Fig. 4, our virtual scenes feature a total of 8 distinct characters embodying pedestrians. Tailoring each character to the volunteer’s appearance, body shape, and gender, we tend to retarget the volunteer’s motion to one of the eight characters.
- **Weathers:** Fig. 6 provides a comprehensive display of the 15 unique weather conditions included in our PMR dataset. To emphasize the variety of our scenes, we highlight 6 representative kinds of weather in Fig. 10.
- **Scenes:** We devised a set of 12 scenarios, intending to elicit diverse reactions from volunteers within various environmental and traffic contexts. Notably, two vehicles within these scenes are equipped with a comprehensive array of sensors, including three RGB sensors and one LiDAR sensor for each car. Each scenario is characterized by three distinct car modes, such as maintaining a constant speed, braking, stopping, and others, with the aim of capturing pedestrians from various viewpoints and observing a range of reactions. Detailed visual representations of each scene and car mode are provided in Table 3, Table 4. Additionally, we have curated one representative sequence for each scene, presenting our recorded data in Fig. 6.

This detailed construction of characters, weather conditions, and urban scenarios in the PMR dataset not only enriches the dataset’s diversity but also ensures a realistic approximation of urban pedestrian dynamics, providing a robust foundation for advancing research in human pose estimation and interaction modeling.

Data Statistics: In Fig. 9, we provide a comprehensive statistical overview of our PMR dataset. The plots depict the distribution of the number of frames categorized by weather, scene, and individual ID. Additionally, distributions are showcased for the duration and frames per sequence, camera speed, and the count of human LiDAR points. Notably, our dataset accounts for occlusion-aware pose estimation, and we depict the distribution of the quantities of visible keypoints in Fig. 9 (g).

Extreme Cases: These cases primarily include the following: 1) Pedestrians falling down; 2) Serious traffic accidents; 3) Extreme weather conditions, such as dense fog and thunderstorms; 4) Pedestrians carrying bulky objects. Datasets captured in real-world conditions typically contain few samples of these low-probability scenarios. Human motions may behave abnormally in such situations, which can lead to challenges when applying current human detection and pose estimation methods. Our dataset is designed to explore these cases more comprehensively. We provide additional examples of extreme scenarios in Fig. 7.

Human Intention In our work, we have designed quantified basic modes to enable the pedestrians to interact with the cars, from which we can investigate more about the human intention and the interaction paradigm between vehicles and pedestrians. Pedestrians may act differently under the same scene. Therefore, we leverage GPT4-o to generate detailed descriptions about the human intention from image sequences. Examples are shown in Fig. 8.

C MULTI-VIEW SLAHMR

The Multi-View SLAHMR extends the original single-view framework to incorporate inputs from multiple perspectives. In contrast to the single-view SLAHMR, Multi-View SLAHMR leverages simultaneous video feeds from various angles. This approach enriches the dataset with diverse observational data, enhancing the model’s ability to reconstruct human motion accurately in world coordinates.

Single-View Setting: We first review the original single-view SLAHMR Ye et al. (2023). The original SLAHMR takes as input a video with T frames of a scene with N people. The goal is to recover the global motion of all detected people in the world coordinates system in the form of SMPL model which represent each person i at timestamp t via global orientation $\phi_t^i \in \mathbb{R}^3$, body pose $\theta_t^i \in \mathbb{R}^{22 \times 3}$, shape $\beta^i \in \mathbb{R}^{16}$, shared over all timestamps t , and root translation $\gamma_t^i \in \mathbb{R}^3$, described in:

$$\mathcal{P}_t^i = \{\phi_t^i, \theta_t^i, \beta^i, \gamma_t^i\}. \quad (1)$$

The SMPL model uses these parameters to generate human joints $\mathbf{J}_t^i \in \mathbb{R}^{3 \times 22}$ of a human body through the differentiable function M :

$$\mathbf{J}_t^i = M(\phi_t^i, \theta_t^i, \beta^i) + \gamma_t^i. \quad (2)$$

Body pose θ_t^i , shape β^i is from body-centric, while root orientation ϕ_t^i and root translation γ_t^i are different under different coordinates. To describe each person i at timestamp t in world and camera coordinate respectively, we define ${}^w\mathcal{P}_t^i$ and ${}^c\mathcal{P}_t^i$ be:

$${}^w\mathcal{P}_t^i = \{{}^w\phi_t^i, \theta_t^i, \beta^i, {}^w\gamma_t^i\}. \quad (3)$$

$${}^c\mathcal{P}_t^i = \{{}^c\phi_t^i, \theta_t^i, \beta^i, {}^c\gamma_t^i\}. \quad (4)$$

Before the main optimization process, SLAHMR estimates camera pose (**camera-to-world** transform) $\{{}_c^wR_t, {}_c^wT_t\}$ by state-of-the-art SLAM system Teed & Deng (2021), as well as each person’s per-frame pose in the camera coordinate ${}^c\hat{\mathcal{P}}_t^i$ by SOTA 3D tracking system, PHALP Rajasegaran et al. (2022). Then, SLAHMR initializes ${}^w\mathcal{P}_t^i$ using $\{{}_c^wR_t, {}_c^wT_t\}$ and ${}^c\hat{\mathcal{P}}_t^i$:

$$\begin{aligned} {}^w\phi_t^i &= {}_c^wR_t \cdot {}^c\hat{\phi}_t^i, & \beta^i &= \hat{\beta}^i \\ {}^w\gamma_t^i &= {}_c^wR_t \cdot {}^c\hat{\gamma}_t^i + \alpha \cdot {}_c^wT_t, & \theta_t^i &= \hat{\theta}_t^i, \end{aligned} \quad (5)$$

where α is also an optimization variable and initialized with 1, referring to the relative scale between the displacement of the camera and that of people.

Joint reprojection loss is the main loss during the whole optimization process, which is defined using the image observations (*i.e.*, the detected 2D keypoints x_t^i and confidence ψ_t^i) and projected 3D joints estimates:

$$E_{\text{data}} = \sum_{i=1}^N \sum_{t=1}^T \psi_t^i \rho(\Pi_K({}_c^wR_t^{-1} \cdot {}^w\mathbf{J}_t^i - \alpha \cdot {}_c^wR_t^{-1} \cdot {}_c^wT_t) - x_t^i), \quad (6)$$

$${}^w\mathbf{J}_t^i = M({}^w\phi_t^i, \theta_t^i, \beta^i) + {}^w\gamma_t^i, \quad (7)$$

where Π_K is perspective camera projection with camera intrinsics matrix $K \in \mathbb{R}^{2 \times 3}$, and ρ is the robust Geman-McClure function.

SLAHMR divides the optimization process into three stages (*i.e.*, root optimization, smooth optimization, and motion chunks optimization) to gradually introduce additional optimization variables and constraints. In the first stage, SLAHMR aligns the parameters of people in the world with the

observed 2D keypoints by optimizing only the human orientation and translation in world coordinate $\{\phi_t^w, \gamma_t^w\}$ of the human pose parameters using E_{data} :

$$\min_{\{\{\phi_t^w, \gamma_t^w\}_{t=1}^T\}_{i=1}^N} \lambda_{\text{data}} E_{\text{data}}. \quad (8)$$

In the second stage, several priors Ye et al. (2023) (i.e., smoothness prior E_{smooth} , shape prior E_{β} , and pose prior E_{pose}) are added and the optimization objectives has become the entire \mathcal{P}_t^i and camera scale α . The overall loss function is defined as:

$$\min_{\alpha, \{\{\mathcal{P}_t^i\}_{t=1}^T\}_{i=1}^N} \lambda_{\text{data}} E_{\text{data}} + \lambda_{\beta} E_{\beta} + \lambda_{\text{pose}} E_{\text{pose}} + \lambda_{\text{smooth}} E_{\text{smooth}}. \quad (9)$$

In the final stage, SLAHMR introduces a learned motion prior that better captures the distribution of plausible human motions from HuMoR Rempe et al. (2021) E_{prior} and also the environment constraints $E_{\text{env}} = \lambda_{\text{skate}} E_{\text{skate}} + \lambda_{\text{con}} E_{\text{con}}$ to prevent unrealistic foot-skate. The optimization for this stage is:

$$\min_{\alpha, \{\{\mathcal{P}_t^i\}_{t=1}^T\}_{i=1}^N} \lambda_{\text{data}} E_{\text{data}} + \lambda_{\beta} E_{\beta} + \lambda_{\text{pose}} E_{\text{pose}} + E_{\text{prior}} + E_{\text{env}} \quad (10)$$

Multi-View Setting: In our multi-view setting, the input turns to be C videos (captured from arbitrary perspectives), each with T synchronized frames of the same scene with N people. The objective remains to recover the global motion of all detected people in the world coordinates system. Transitioning SLAHMR to accommodate inputs from multi-view videos faces two key problems: (1) the “world coordinate” in SLAHMR is defined by the coordinate of camera motions estimated by SLAM; (2) most SLAM systems struggle to predict multiple independent cameras extrinsic in a consistent coordinate system. Consequently, the extrinsic estimated for multiple cameras exist under disparate coordinates, impeding our ability to ascertain the relative positions of these cameras.

In order to leverage the reprojection constraints on varied views, we need to align C camera motions under w_1, \dots, w_C into a consistent world coordinate system w_1 . Estimated human tracks are the only commonality of the C views and therefore become the only possible anchor to perform transformation.

The motion of the j -th camera c_j at time t under the j -th “world coordinate” w_j is (here we also use **camera-to-world** transform to denote camera motion):

$$\{c_j^{w_j} R_t, c_j^{w_j} T_t, c_j^{w_j} \alpha\}. \quad \text{for } j = 1, 2 \dots C, \quad (11)$$

We first go through the root optimization stage and smoothness optimization stage under C coordinates separately and derive human motions ${}^{w_j} \mathcal{P}_t^i = \{{}^{w_j} \phi_t^i, \theta_t^i, \beta^i, {}^{w_j} \gamma_t^i\}$ by 8 and 9, as well as human joints ${}^{w_j} J_t^i$ by 7. Then we assume:

Assumption C.1 *The estimated human motions ${}^{w_j} \mathcal{P}_t^i$ for $j = 1, 2 \dots C$ are relatively accurate.*

We claim the first view under the first “world coordinate” w_1 as the primary view and other views as secondary views. Convert each of the secondary camera $j (j > 1)$ under its own “world coordinates” w_j to the primary coordinates w_1 by applying Umeyama alignment Umeyama (1988) to joints ${}^{w_1} J$ and ${}^{w_j} J$:

$$\begin{bmatrix} {}^{w_1} R & {}^{w_1} T & {}^{w_1} s \\ {}^{w_j} R & {}^{w_j} T & {}^{w_j} s \end{bmatrix} = \mathcal{U}({}^{w_j} J, {}^{w_1} J) \quad (12)$$

$${}^{w_1} R_t = {}^{w_j} R \cdot {}^{w_j} R_t, \quad (13)$$

$${}^{w_1} \alpha \cdot {}^{w_1} T_t = {}^{w_j} s \cdot {}^{w_j} R \cdot {}^{w_j} \alpha \cdot {}^{w_j} T_t + {}^{w_1} T, \quad (14)$$

where \mathcal{U} denotes the Umeyama alignment algorithm.

Till now, all cameras motions ${}^{w_j} R_t, {}^{w_j} T_t$ for time $t = 1, \dots, T$ and view $j = 1, \dots, C$ are converted to the primary coordinate w_1 , denoted as ${}^{w_1} R_t, {}^{w_1} T_t$. Before we apply the reprojection loss under different views to the human motion, we add a camera optimization stage to align the human motion estimated from w_1 and the motion of camera $c_j (j > 1)$ since Ass. C.1 is a rough assumption. In this

Table 1: Optimization process comparison between the single-view SLAHMR and our modified multi-view SLAHMR.

Input	Stages	Loss	Optimization Variable
single-view	Root Optimization	E_{data}	${}^w\phi_t^i, {}^w\gamma_t^i$
	Smooth Optimization	$E_{data}, E_{\beta}, E_{pose}, E_{smooth}$	${}^w\mathcal{P}_t^i, \alpha$
	Motion Chunks Optimization	$E_{data}, E_{\beta}, E_{pose}, E_{prior}, E_{env}$	${}^w\mathcal{P}_t^i, \alpha$
multi-view	Root Optimization	${}^{w_1}E_{data}^{(1)}$	${}^{w_1}\phi_t^i, {}^{w_1}\gamma_t^i$
		\dots	\dots
		${}^{w_C}E_{data}^C$	${}^{w_C}\phi_t^i, {}^{w_C}\gamma_t^i$
	Smooth Optimization 1	${}^{w_1}E_{data}^{(1)}, {}^{w_1}E_{\beta}, {}^{w_1}E_{pose}, {}^{w_1}E_{smooth}$	${}^{w_1}\mathcal{P}_t^i, {}^{w_1}\alpha$
		\dots	\dots
		${}^{w_C}E_{data}^C, {}^{w_C}E_{\beta}, {}^{w_C}E_{pose}, {}^{w_C}E_{smooth}$	${}^{w_C}\mathcal{P}_t^i, {}^{w_C}\alpha$
	Camera Optimization	$E_{data}^{(2)}, E_{camSmooth}^{c_2}$	${}^{w_1}R_t, {}^{w_1}T_t$
		\dots	\dots
		$E_{data}^C, E_{camSmooth}^{c_C}$	${}^{w_1}R_t, {}^{w_1}T_t$
	Smooth Optimization 2	$E_{data}^{(1)}, \dots, E_{data}^C, E_{\beta}, E_{pose}, E_{smooth}$	${}^{w_1}\mathcal{P}_t^i, {}^{w_1}\alpha$
	Motion Chunks Optimization	$E_{data}^{(1)}, \dots, E_{data}^C, E_{\beta}, E_{pose}, E_{prior}, E_{env}$	${}^{w_1}\mathcal{P}_t^i, {}^{w_1}\alpha$

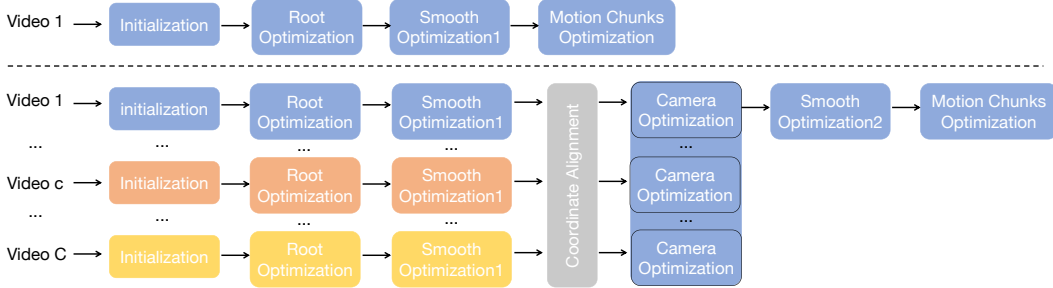


Figure 5: Optimization pipeline of single-view SLAHMR Ye et al. (2023) (top) and our modified multi-view SLAHMR(bottom). Different colors refer to different coordinates.

stage, for camera $c_j (j > 1)$, we optimize ${}^{w_1}R_t, {}^{w_1}T_t$ by the reprojection loss under view j , E_{data}^j and camera smoothness loss $E_{camSmooth}^j$:

$$\min_{\{\{{}^{w_1}R_t, {}^{w_1}T_t\}_{t=1}^T\}_{j=2}^C} \sum_{j=2}^C (\lambda_{data}^j E_{data}^j + \lambda_{camSmooth}^{c_j} E_{camSmooth}^{c_j}), \quad (15)$$

$$E_{data}^j = \sum_{i=1}^N \sum_{t=1}^T \psi_t^{i,j} \rho(\Pi_{K^{c_j}}({}^{w_1}R_t^{-1} \cdot {}^{w_1}J_t^i - {}^{w_1}\alpha \cdot {}^{w_1}R_t^{-1} \cdot {}^{w_1}T_t) - x_t^{i,j}),$$

where the subscript j means the variable is of view j , the subscript c_j means the variable is of camera j , the subscript w_j means the variable is under the j -th “world coordinate”. Subsequently, we apply the smooth optimization again, as well as the motion chunks optimization, with the 2D reprojection loss from all views:

$$\min_{w_1\alpha, \{\{{}^{w_1}\mathcal{P}_t^i\}_{t=1}^T\}_{i=1}^N} \sum_{j=1}^C \lambda_{data}^j E_{data}^j + \lambda_{\beta} E_{\beta} + \lambda_{pose} E_{pose} + \lambda_{smooth} E_{smooth}, \quad (16)$$

$$\min_{w_1\alpha, \{\{{}^{w_1}\mathcal{P}_t^i\}_{t=1}^T\}_{i=1}^N} \sum_{j=1}^C \lambda_{data}^j E_{data}^j + \lambda_{\beta} E_{\beta} + \lambda_{pose} E_{pose} + E_{prior} + E_{env}.$$

We summarize the difference between the original single-view SLAHMR and our modified multi-view SLAHMR in Table 1 and Fig. 5.

Table 2: cross-dataset evaluation on hmr using different training data.

Training data	Test on H36M (Ionescu et al., 2014)		Test on PMR		Test on 3DPW (von Marcard et al., 2018)		Test on mpi-inf-3dhp (Mehta et al., 2017)	
	MPJPE↓	PA-MPJPE↓	MPJPE↓	PA-MPJPE↓	MPJPE↓	PA-MPJPE↓	MPJPE↓	PA-MPJPE↓
100% H36M	35.287	25.069	346.068	151.195	342.373	159.088	416.254	222.917
60% H36M + 40% PMR	83.836	55.692	134.570	75.687	261.466	131.233	287.729	159.462

D CROSS-DATASET EVALUATION ON HMR

We first train BEV on 300,000 frames from the Human3.6M (H36M) dataset (Ionescu et al., 2014). Then, we randomly replace 40% of the training data with our PMR dataset to train a new model from scratch. We test the results on the test sets of both H36M and our PMR dataset, as well as two other unseen realistic datasets. The results, summarized in Table 2, indicate that the gap between different realistic datasets is also non-negligible compared to the gap between realistic and virtual datasets. Our PMR dataset can effectively enhance the diversity of datasets and reduce the gap to unseen data. The results on both BEVStereo in Section 4.4 and hmr indicate that the gap between the domain gap is actually acceptable. Our PMR dataset can effectively enhance the diversity of datasets and reduce the gap to unseen data.

E LIMITATIONS AND POTENTIAL IMPACTS

Limitations: Although we have conducted a wide range of tasks and evaluations on our dataset, there still remains scope for further exploration. A notable gap is the lack of a thorough connection between third-view and first-view perspectives, which could serve as an invaluable resource for investigating interactions between vehicles and pedestrians in autonomous driving scenarios. Moreover, an optimization-based multi-view human reconstruction pipeline is too complex and time-consuming. This is because of the inherently challenging nature of reconstructing human motion from multiple independently moving cameras, despite the optimization-based method being a direct approach leveraging state-of-the-art single-view techniques. Learning-based methods are much faster, but recovery of human motions even from a single moving camera is still an area with limited viable learning-based solutions, largely due to the scarcity of training data. Therefore, our large-scale PMR dataset can benefit future both single-view and multi-view learning-based human motion reconstruction.

Potential Impacts: Our proposed PMR database, a large-scale and annotation-free database aimed at autonomous driving scenarios, holds the potential to promote prosperity and development in the field of autonomous driving. However, it also introduces several considerable risks: (1) The widespread adoption of autonomous vehicles could lead to job losses in industries such as transportation and logistics. Drivers of taxis, trucks, and delivery vehicles may find themselves unemployed as autonomous technology replaces human operators. (2) The introduction of autonomous vehicles raises complex legal and regulatory questions. Determining liability in the event of accidents, establishing uniform safety standards, and adapting existing laws to accommodate autonomous technology present significant challenges. (3) While autonomous vehicles have the potential to reduce traffic congestion and emissions through optimized routing and platooning, they may also contribute to increased vehicle miles traveled if people choose to use them more frequently, leading to greater energy consumption and environmental damage.

F FUTURE WORK

Our data collection process is highly flexible and can be extended to other simulators, such as AirSim Shah et al. (2017). To support community efforts, we will open-source the configuration files as prior works Wang et al. (2022); Riaz et al. (2023), enabling replication of the collected scenarios and facilitating contributions to this dataset. Additionally, by integrating these simulated scenarios with real human motion dataset, we aim to further enhance the diversity and representativeness of the PMR dataset.

Moreover, incorporating personality-based behavior generation could significantly enhance the depth of human motion modeling. As part of our future research, we plan to disentangle attribute-related and

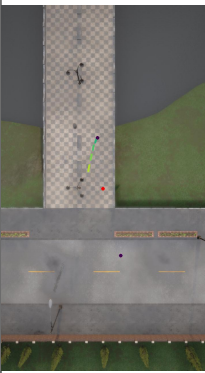
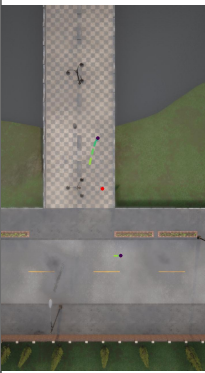

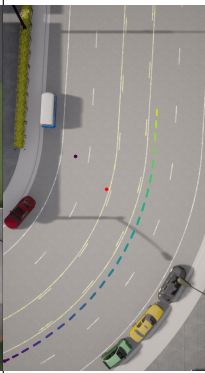
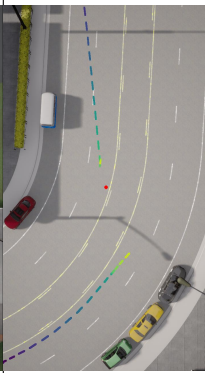

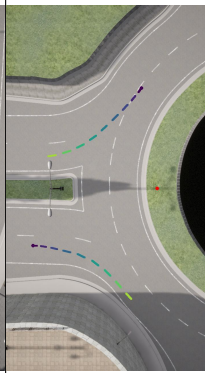
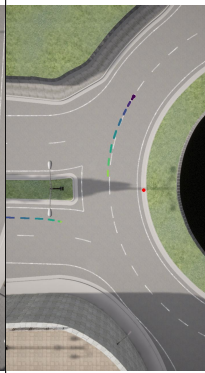
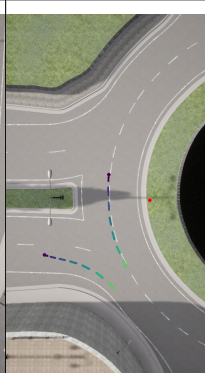



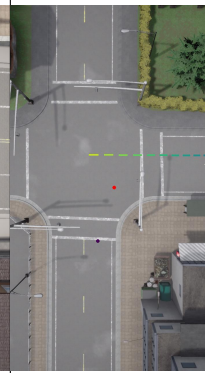
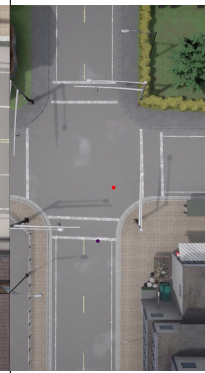
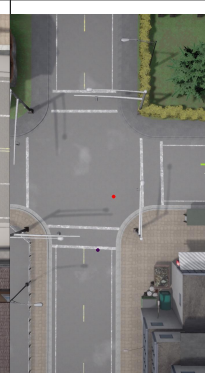



Scene No.	Scene Description	Carmode 0	Carmode 1	Carmode 2
Scene 1	Crossing the bridge. The pedestrian walks on a bridge and should be aware of the scooter behind and the car ahead. Both of these vehicles are equipped with sensors.			
Scene 2	Crossing the bent road. The pedestrian crosses a curved road, with two sensor-equipped vehicles on the left and right sides approaching them.			
Scene 3_2	Idling in the roundabout. The pedestrian performs casual actions in the green area within the roundabout. Two sensor-equipped vehicles enter the lanes of the roundabout.			
Scene 3_3	Crossing the slopped intersection. Two sensor-equipped cars are respectively in front and behind the pedestrian, and the cars may experience slipping.			
Scene 4	Witnessing collision. At the intersection, the pedestrian witnesses an accident (or narrowly avoided accident) involving two vehicles, with two other sensor-equipped cars in different directions.			
Scene 4_2	Assisting in reversing. The pedestrian guides the car in front to reverse into the parking space. The rear of this car, as well as another car beside the walker, is equipped with sensors.			

Table 3: Scene illustration of our PMR dataset (Part I). We illustrate each scene with text descriptions and three carmode diagrams. In carmode diagrams, red points represent the initial position of the pedestrian. Purple points represent the initial position of the two cars, each of which has three RGB sensors (with different perspectives) and a lidar sensor. Dashed lines (from purple to yellow) represent the cars tracks (from beginning to end).

Scene No.	Scene Description	Carmode 0	Carmode 1	Carmode 2
Scene 5	Appearing from the blind spot. The pedestrian emerges from the area obscured by a truck, with one sensor-equipped car in the left lane and another in the reverse lane.			
Scene 5_2	Flagging down vehicles at the roadside. Two parallel-moving sensor-equipped vehicles approaching, the pedestrian signals them to stop.			
Scene 5_3	Walking along the lane. The pedestrian walks on the lane to inspect the left door of a van. A sensor-equipped car is approaching from the rear and another is from the opposite lane.			
Scene 6	Seeking help. The pedestrian is in the middle of the ramp and the main road, seeking assistance from both sides. Two sensor-equipped vehicles are on the ramp and the main road respectively.			
Scene 7	Crossing the intersection. A sensor-equipped car is on the left and another is on the right of the pedestrian. There are also some other vehicles moving in the scene.			
Scene 7_2	Exiting the parking space. A sensor-equipped car is behind the pedestrian and another is on the right front.			

Table 4: Scene illustration of our PMR dataset (Part II). We illustrate each scene with text descriptions and three carmode diagrams. In carmode diagrams, red points represent the initial position of the pedestrian. Purple points represent the initial position of the two cars, each of which has three RGB sensors (with different perspectives) and a lidar sensor. Dashed lines (from purple to yellow) represent the cars’ tracks (from beginning to end).

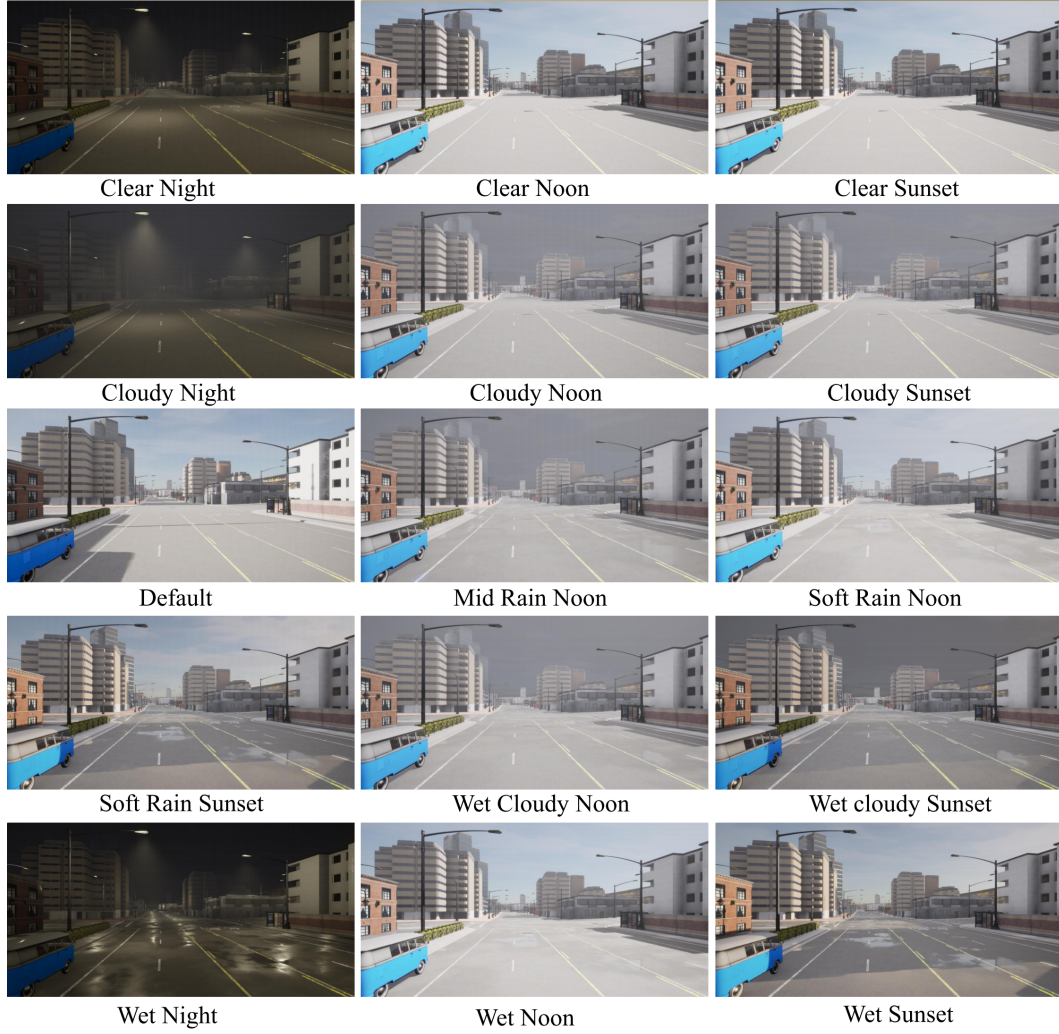


Figure 6: Weather diversity in the proposed Pedestrian Motion Reconstruction (PMR) dataset.

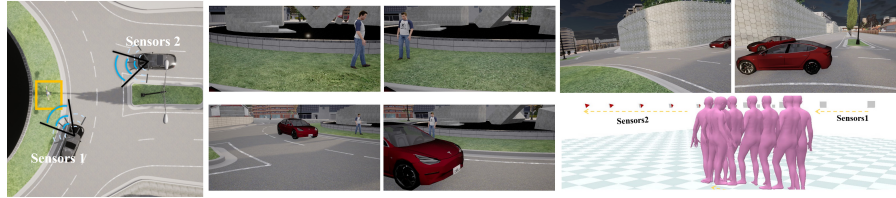
attribute-unrelated features from pedestrian behavior using a Semantics-Guided Neural Network as the feature extractor. Specifically, a disentangled bottleneck is employed to separate attribute-related and attribute-unrelated features, guided by an attribute classifier. Novel pose sequences are then generated by combining the specific attribute embeddings from one sequence with the embeddings representing the remaining attributes from another sequence. For example, if we have subject a's motion under scene 1 and subject b's motion under scene 2, we are curious about how subject 1 will react under scene 2. Preliminary results, presented in Fig. 11, demonstrate the feasibility of this approach. We believe this work has the potential to create more diverse and realistic pedestrian motion patterns.



(a) Scene 1 (Carmode 2): Crossing the bridge.



(b) Scene 2 (Carmode 0): Crossing the bent road.



(c) Scene 3_2 (Carmode 2): Idling in the roundabout.



(d) Scene 3_3 (Carmode 2): Crossing the slopped intersection.

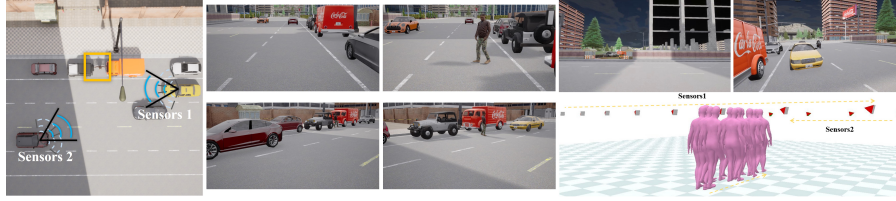


(e) Scene 4 (Carmode 0): Witnessing collision.



(f) Scene 4_2 (Carmode 2): Assisting in reversing.

Bird eye's view of each scene	3rd-view image sequence from Sensors 1	Ego-view image sequence
	3rd-view image sequence from Sensors 2	Annotations



(g) Scene 5 (Carmode 1): Appearing from the blind spot.



(h) Scene 5_2 (Carmode 0): Flagging down vehicles at the roadside.



(i) Scene 5_3 (Carmode 2): Walking along the lane.



(j) Scene 6 (Carmode 1): Seeking help.



(k) Scene 7 (Carmode 2): Crossing the intersection.



(l) Scene 7_2 (Carmode 0): Exiting the parking space.

Bird eye's view of each scene	3rd-view image sequence from Sensors 1	Ego-view image sequence
	3rd-view image sequence from Sensors 2	Annotations

Figure 6: Contents and data schematics of 12 Scenes. The bird's eye view is for illustrative purposes only. We selected 2 frames from synchronized videos of each perspective, as showcased in the image. Sensors 1 and Sensors 2 denote various sensors (multi-view RGB sensors and LiDAR sensors) in Car 1 and Car 2, respectively. Here, we have marked only one sensor in each car for clarity.



(a) multiple person / with objects



(b) with objects



(c) with objects / traffic accident



(d) multiple person / pedestrian falls / serious traffic accident



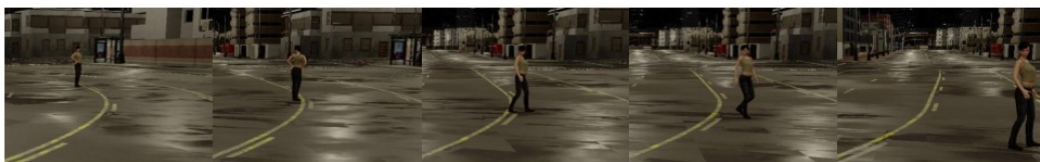
(e) interacting with the vehicle



(f) interacting with the vehicle/ pedestrian falls



(g) extreme weather condition



(h) extreme weather condition

Figure 7: Examples of extreme cases.

Prompt:

Suppose you are an agent of an autonomous vehicle. The first row of the given figure is a sequence of images captured by the in-car camera, the second row is the corresponding image sequence captured from the pedestrian's perspective, please

1. inference the pedestrian's possible intentions with the given figure,
2. predict the future motion of the pedestrian based on the figure and your inferred intentions.

Example 1:**GPT4-o Response:****1. Inference of the Pedestrian's Intentions:**

In the given sequence of images, the pedestrian (who appears to be a traffic officer) is seen making a series of hand gestures. Initially, the officer is standing still and facing the oncoming traffic.

- The first image shows the officer with one arm raised, which typically signals vehicles to stop.
- In the subsequent images, the officer's stance changes to having both hands lowered and then one hand slightly to the side, signaling the vehicles to remain stopped while possibly preparing to redirect traffic.
- The final image in the sequence shows the officer starting to point to the side, indicating an intention to direct the vehicles or pedestrians to move in a specific direction, likely to proceed or turn.

2. Prediction of the Pedestrian's Future Motion:

-Based on the series of gestures and the gradual change in posture, the officer appears to be in the process of managing traffic flow.

-The officer will likely finish signaling the current lane to stop and might move sideways to ensure vehicles or pedestrians understand the new directive.

-If the officer is about to direct traffic to proceed, it is expected they might move to the side of the road to get out of the way of the oncoming vehicles.

Therefore, the officer's next likely movement is stepping to the side to ensure a clear path for vehicles to proceed or to safely continue directing the flow of traffic. This inference is drawn from the established hand signals and body language typical of traffic management.

Example 2:**GPT4-o Response:****1. Inference of the Pedestrian's Possible Intentions:**

-The pedestrian, seen in the first row of images, appears to be moving towards the road from the left side and is gradually stepping closer to the path of an oncoming vehicle.

-Initially, the pedestrian is walking towards the road.

-In the subsequent images, the pedestrian's posture suggests an intention to cross the road. They are looking towards the approaching vehicle, likely assessing the speed and distance of the car.

-In the final image, the pedestrian is very close to the vehicle, possibly trying to make a hurried decision on whether to stop or continue crossing.

2. Prediction of the Pedestrian's Future Motion:

-Based on the sequence, the pedestrian is likely attempting to cross the road and is in the midst of assessing whether it is safe to proceed.

-Given the current position close to the vehicle, the pedestrian may either pause and wait for the vehicle to pass or make a quick dash to the other side of the road if they perceive it to be safe enough to cross before the car reaches them.

In summary, the pedestrian's intention seems to be crossing the road. Given the proximity to the vehicle, the pedestrian might make a quick decision to stop and let the car pass or dash across the road if they judge it safe to do so.

Figure 8: Examples of human intention descriptions generated from GPT4-o.

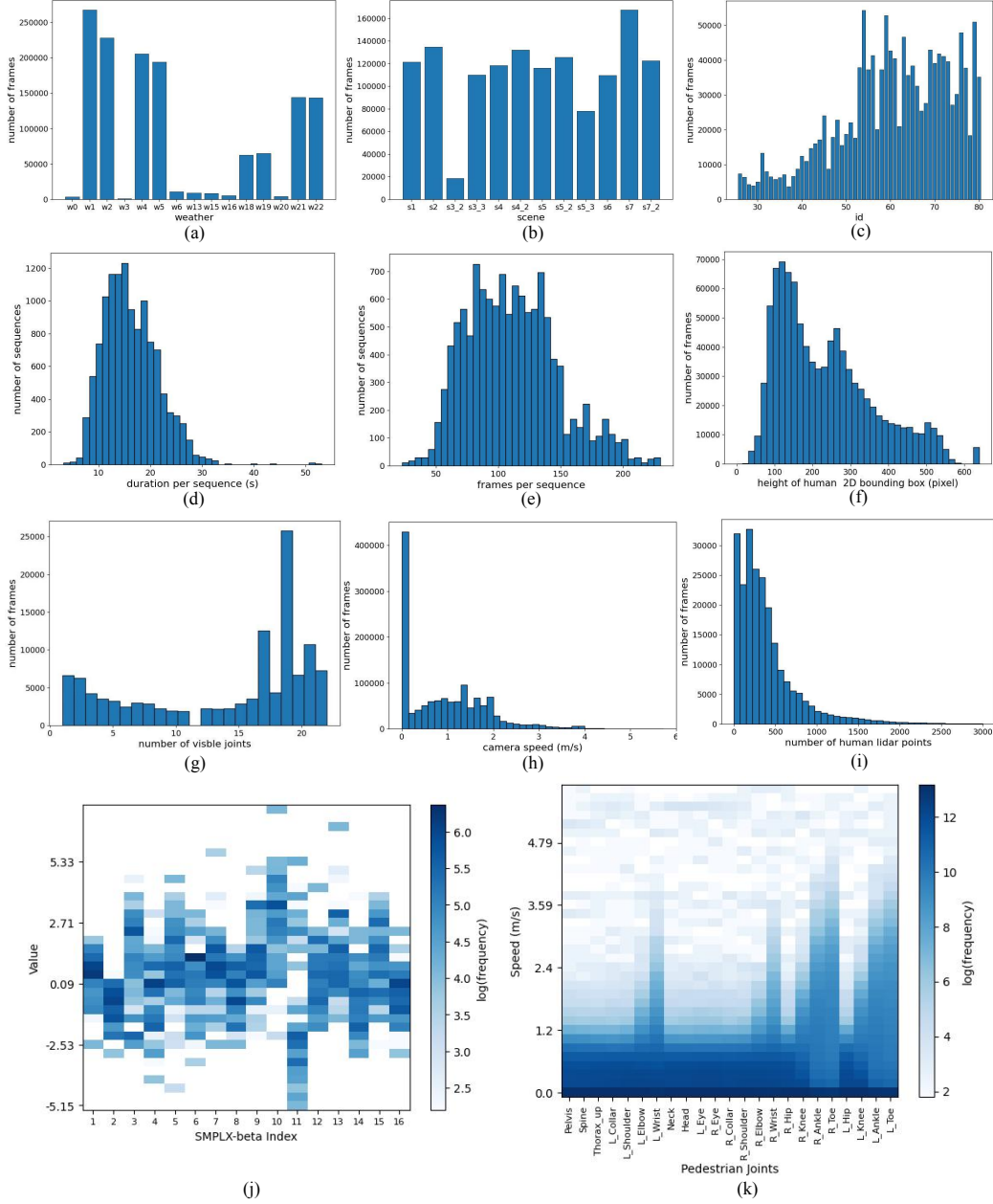


Figure 9: PMR dataset statistics. In (a), (b), and (c), we show the number of frames for different weathers, scenes, and participants(id), respectively. (d) and (e) exhibit the distribution of the duration and number of frames for a single third-view sequence. (f) shows the distribution of the height of 2D human bounding box in our captured images with a total height of 640 pixels. In (g), we statistically analyzed the distribution of visible keypoints quantities in situations where individuals were partially obscured. In (h) and (g), we shows the distribution of camera speed and number of human LiDAR points for our third-view data and LiDAR data respectively. (j) shows the distribution of pedestrians’ shape in terms of SMPL-X. (k) shows the distribution of the speed of pedestrians’ joints (the overall pedestrian’s speed can be represented by the speed of Pelvis), from which we can also find that the motion of limbs are much more flexible than other human parts. Note that we calculate the shape in (j) by sequences and the speed in (k) by frames, leading to distinct levels on the frequency.



Figure 10: Examples of scene variety under the same weather conditions.

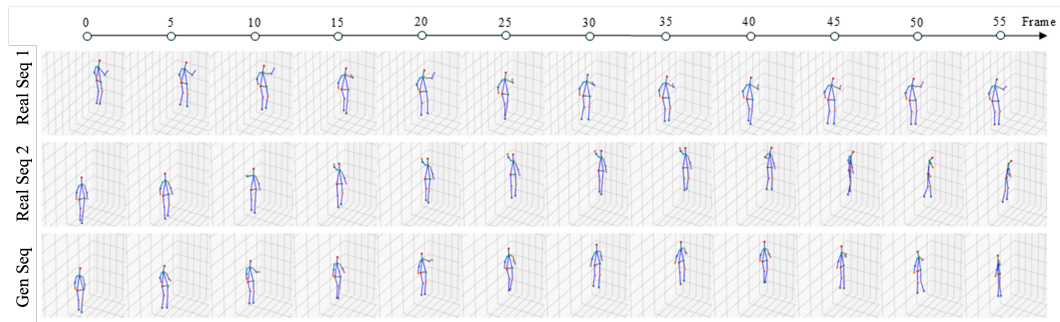


Figure 11: The first two rows show interactions with vehicles seeking help (scene 6) by Subject 1 and Subject 2 under different weather conditions from the PMR dataset. The third row presents the generated pose sequence for Subject 1 under the weather condition of Subject 2.

ACKNOWLEDGEMENTS

We appreciate all the anonymous reviewers for their constructive suggestions on polishing this paper. The work was supported by the Shanghai Municipal Science and Technology Major Project (Grant No. 2021SHZDZX0102), and the National Natural Science Foundation of China (Grant No. 62302295) in part.

REFERENCES

- Alexey Dosovitskiy, Germán Ros, Felipe Codevilla, Antonio M. López, and Vladlen Koltun. Carla: An open urban driving simulator. In *Conference on Robot Learning*, 2017.
- Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36:1325–1339, 2014.
- Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6): 248:1–248:16, October 2015.
- Dushyant Mehta, Helge Rhodin, Dan Casas, Pascal Fua, Oleksandr Sotnychenko, Weipeng Xu, and Christian Theobalt. Monocular 3d human pose estimation in the wild using improved cnn supervision. In *3D Vision (3DV), 2017 Fifth International Conference on*. IEEE, 2017. doi: 10.1109/3dv.2017.00064. URL http://gvv.mpi-inf.mpg.de/3dhp_dataset.
- Jathushan Rajasegaran, Georgios Pavlakos, Angjoo Kanazawa, and Jitendra Malik. Tracking people by predicting 3d appearance, location and pose. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2740–2749, 2022.
- Davis Rempe, Tolga Birdal, Aaron Hertzmann, Jimei Yang, Srinath Sridhar, and Leonidas J Guibas. Humor: 3d human motion model for robust pose estimation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 11488–11499, 2021.
- Muhammad Naveed Riaz, Maciej Wielgosz, Abel García Romera, and Antonio M López. Synthetic data generation framework, dataset, and efficient deep model for pedestrian intention prediction. In *2023 IEEE 26th International Conference on Intelligent Transportation Systems (ITSC)*, pp. 2742–2749. IEEE, 2023.
- S. Shah, Debadeepta Dey, Chris Lovett, and Ashish Kapoor. Airsim: High-fidelity visual and physical simulation for autonomous vehicles. In *International Symposium on Field and Service Robotics*, 2017.
- Zachary Teed and Jia Deng. Droid-slam: Deep visual slam for monocular, stereo, and rgb-d cameras. *Advances in neural information processing systems*, 34:16558–16569, 2021.
- Shinji Umeyama. An eigendecomposition approach to weighted graph matching problems. *IEEE transactions on pattern analysis and machine intelligence*, 10(5):695–703, 1988.
- Timo von Marcard, Roberto Henschel, Michael J. Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *European Conference on Computer Vision*, 2018.
- Cheng Yao Wang, Oron Nir, Sai Vemprala, Ashish Kapoor, Eyal Ofek, Daniel McDuff, and Mar Gonzalez-Franco. Citylifesim: A high-fidelity pedestrian and vehicle simulation with complex behaviors. In *2022 IEEE 2nd International Conference on Intelligent Reality (ICIR)*, pp. 11–16. IEEE, 2022.
- Tong Wu, Jiarui Zhang, Xiao Fu, Yuxin Wang, Liang Pan Jiawei Ren, Wayne Wu, Lei Yang, Jiaqi Wang, Chen Qian, Dahua Lin, and Ziwei Liu. Omniobject3d: Large-vocabulary 3d object dataset for realistic perception, reconstruction and generation. 2023.
- Vickie Ye, Georgios Pavlakos, Jitendra Malik, and Angjoo Kanazawa. Decoupling human and camera motion from videos in the wild. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21222–21232, 2023.