SMPL-Based 3D Pedestrian Pose Prediction

Anil Kunchala[1], Mélanie Bouroche[2], Lorraine D'Arcy[3] and Bianca Schoen-Phelan[4]
[1,3,4] Technological University Dublin, Ireland
[2] Trinity College Dublin, Ireland

***Abstract*— Modeling human motion is a long-standing problem in computer vision. The rapid development of deep learning technologies for computer vision problems resulted in increased attention in the area of pose prediction due to its vital role in a multitude of applications, for example, behavior analysis, autonomous vehicles, and visual surveillance. In 3D pedestrian pose prediction, joint-rotation-based pose representation is extensively used due to the unconstrained degree of freedom for each joint and its ability to regress the 3D statistical wireframe. However, all the existing joint-rotation-based pose prediction approaches ignore the centrality of the distinct pose parameter components and are consequently prone to suffer from error accumulation along the kinematic chain, which results in unnatural human poses. In joint-rotation-based pose prediction, Skinned Multi-Person Linear (SMPL) parameters are widely used to represent pedestrian pose. In this work, a novel SMPL-based pose prediction network is proposed to address the centrality of each SMPL component by distributing the network weights among them. Furthermore, to constrain the network to generate only plausible human poses, an adversarial training approach is employed. The effectiveness of the proposed network is evaluated using the PedX and BEHAVE datasets. The proposed approach significantly outperforms state-of-the-art methods with improved prediction accuracy and generates plausible human pose predictions.**

I. INTRODUCTION

Pedestrians are inherently capable of predicting changes in the surrounding environment, along with the movement of other pedestrians, and alter their path accordingly. For example, while walking, pedestrians naturally adjust their path according to the surrounding pedestrians and vehicle positions. These predictions are crucial to shape day-to-day interactions and make social life attainable [1]. Recently, there is a growing interest in pedestrian analysis for autonomous vehicle navigation systems [2], [3], [4] and urban design [5], [6]. Providing the ability to predict and understand pedestrian behaviour is paramount to create seamless integration of machines with pedestrians. Behavior analysis is often carried on a sequence of human poses representing a specific activity. Pose estimation (localizing the human body joints) and prediction (forecasting future poses using observed past poses) are primary steps in behavior analysis. For example, frame-to-frame pose estimation is required to calculate the walking speed of a pedestrian.

978-1-6654-3176-7/21/$31.00 ©2021 European Union

Human pose can be represented using absolute joint locations or relative joint rotations, with each representation having its benefits and trade-offs [7]. Joint location representation describes the human body as a collection of 2D or 3D joint locations and expresses body features using connections between joints and locations. Absolute joint location representation is widely used for pose estimation and prediction [1], [8], [9], [10], [11], [12], [13], [14], [15]. However, absolute joint locations do not constrain the full degree of freedom of each joint and are often paired with prediction errors such as bone stretching [7].

The relative-rotations-based approach describes poses as a collection of 3D joint rotations where each rotation is relative to its parent in a kinematic tree. Furthermore, 3D joint rotations are also used to regress the statistical 3D wireframe to represent the human pose. The generative Skinned Multi-Person Linear (SMPL) model [16] is used widely in joint rotation-based pose representation [17], [18], [19], [20], [21], [2], [22], [12]. SMPL models the human body as a 3D mesh using pose, shape, and translation parameters. The shape parameters represent individual variations of human body proportions, pose parameters are used to define 3D body shape articulation using 3D relative rotation joint angles including global rotation, and translation parameters are used to indicate global translation of 3D mesh.

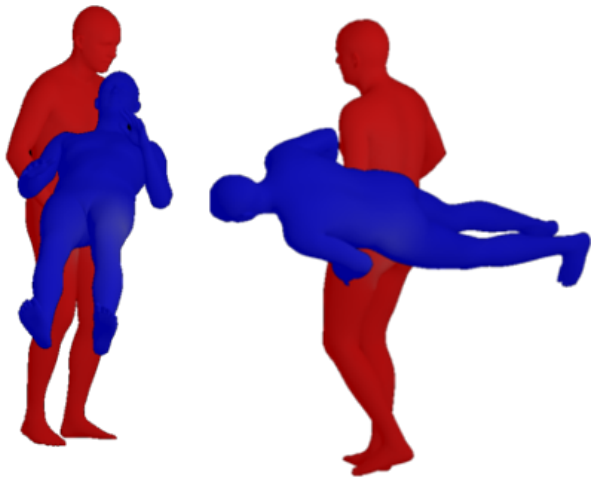

Fig. 1: Unnatural poses generated by joint-rotation-based approaches. The ground truth wireframes are represented in red and the predicted 3D wireframes are represented in blue.

Recently, a few joint rotation-based approaches have been

proposed for pose prediction [2], [7] and activity recognition [22]. Most of these approaches consider all joints as single vector input to the model. Using single vector input with average error enforces the model to distribute equal weights for all joints thereby ignoring the various impact of important joints (like the global rotation joint) on the body pose. The equal weight distribution of joints may generate large errors in important joints which severely impact the final pose in qualitative results as shown in Fig 1.

In this paper, this challenge is addressed by introducing the SMPL-based Recurrent Neural Network trained in an adversarial fashion. In contrast to previous models [2], [7], the novel SMPL-based architecture is proposed to address different components in the SMPL parameters and distribute the network weights among them to address the centrality of each component within a final pose. Adversarial training [23] is used to discriminate against unnatural poses while allowing natural ones. The novel approach presented in this paper is similar to the recurrent pose prediction networks [2], [10], and adversarial networks [23], [14]. However, this work go beyond existing technique in multiple ways:

- To our knowledge, this is the first paper to incorporate adversarial training in SMPL pose prediction. We designed a multi-layer perception generative adversarial network to penalize the network for unnatural poses while allowing natural ones.
- The SMPL-based architecture is proposed to address the centrality of global rotation and translation parameters with respect to the pose parameters.

We present experimental results on the proposed network using 1) PedX [24], an in the wild image dataset collected in urban intersections, 2) BEHAVE [25], a large scale image dataset with different scenario's of people acting out various interactions, and 3) HYBRID, which consists of a combination of both PEDX and BEHAVE datasets. We also compare the effects of both adversarial network and SMPL aware architecture on the proposed model and show that the proposed model outperforms state-of-the-art baseline methods.

The remaining paper is organized as follows: Section II explores related work in pose prediction, generative adversarial networks, and 3D wireframe representation of the human body. Section III describes the proposed SMPL-based architecture and adversarial SMPL-based recurrent neural network, including implementation details. Section IV describes baseline methods, datasets, and the experimental setup. Section V presents the evaluation metrics, qualitative and quantitative results. Finally, section VI concludes the proposed work and delineates the potential future work. The source code for this work is made available at https://github.com/anilkunchalaece/ADV-SA-LSTM

II. RELATED WORK

This section reviews the existing literature in 3D wireframe representation of human body, pose prediction, and generative adversarial networks.

A. 3D Wireframe Representations of the Human Body

Skinned Linear Person Model (SMPL)[16] is a generative model that represents the 3D wireframe mesh of the human body as a function of shape, pose and translation parameters. SMPL pose parameters consist of following components: 1) The relative rotation of 23 SMPL joints in the axis angle representation, 2) Three root orientation parameters in the axis angle representation, 3) The Translation parameters consist of global translation in x, y and z axes, and 4) The Shape parameters consist of the first 10 principal component analysis shape space parameters. The SMPL model is widely used in pose and shape estimation networks [17], [18], [19], [20], [21]. In [18], [17], [26], 2D key points extracted from image features are used to regress a 3D wireframe mesh from a single image. In addition, a discriminator network is used to determine whether the generated pose and shape parameters are natural. [20] proposed an improved model from [18], by including a pose prior to generate pose, hand pose and facial expressions from a single image. In contrast to image based models, Video Inference for Human Body Pose and Shape Estimation (VIBE) [19] proposed a temporal network architecture to generate 3D motion from monocular video. VIBE utilizes a convolution neural network pre-trained on a single image [17], followed by a temporal encoder and a motion discriminator to capture the sequential nature of motion. Due to the scarcity of 3D wireframe datasets for pedestrians in real world scenarios, this work utilizes VIBE to generate SMPL 3D wireframes using readily available image datasets[24], [25].

B. Pose Prediction

Pose prediction exploits the sequential nature of pose data across frames. A large number of approaches use Recurrent Neural Networks (RNN) as they can model temporal dependencies [10], [12], [13], [27], [2], [14], [28]. For example, a stacked LSTM[2], [10] and encoder-recurrent-decoder[10] architectures are proposed for 2D joint-location based prediction and joint-rotation based pose predictions[27], [22]. In addition to the temporal features, spatial features are also modelled using structured prediction layer[12] and bidirectional RNNs[28]. To further improve the model to generate naturally plausible future joint locations, an adversarial networks [13], [14] are also introduced in recurrent encoder-decoder architectures. More recently, [2] proposed the Bio-LSTM architecture with bio-mechanical constraints-based loss functions. This architecture adapts and extends 3LR-LSTM [10] to predict SMPL pose and translation parameters instead of joint locations. The Bio-LSTM architecture considers all SMPL parameters as a single vector input to the model. The single vector input forces the model to adjust its weights based on the average error over all components. This approach, however, ignores the centrality of the different components, which leads to higher error rates in qualitative results. To address this, we propose a novel SMPL-based architecture to utilize the centrality of each SMPL component. In contrast to Bio-LSTM [2], the proposed architecture consists of a dedicated independent

LSTM layer for each SMPL component to model spatial and temporal dependencies of individual SMPL components.

C. Generative Adversarial Networks

Generative Adversarial Networks (GANs) [29] contain both a generator network to generate images, and a discriminator network to distinguish between generated and ground truth images. Both of these networks are trained in an adversarial fashion. The GANs are traditionally used for generative tasks Recently, recurrent GANs have been trained in an adversarial manner to improve the quality of pose and shape estimation networks [19], [17], [30] and pose joint-location based pose prediction models[14], [13]. By adapting adversarial training introduced in pose estimation[19], an adversarial SMPL-based network is proposed to constrain network from generating unnatural poses thereby improving the overall network performance.

III. PROPOSED SYSTEM

The goal of the proposed system is to predict the 3D SMPL pose parameters in the future frame for a given past 'n' poses. The input and output of the model are the SMPL parameters representing the 3D pedestrian pose in each frame. For a given dataset, pre-processing utilizes VIBE to extract SMPL pose parameters. Fig 2 depicts the architecture of proposed model. For a given past 'n' SMPL pose parameters, each SMPL component is fed to a dedicated LSTM layer in the SMPL-based pose predictor network. During the training, each LSTM layer will optimize its weights based on the temporal dependencies of the respective components. The outputs of these LSTM layers will be applied to the fully connected layer(FC) to capture the spatial dependencies between SMPL components. The SMPL-based architecture is used to model spatial and temporal relations for the given pose parameters. However, the generator may still produce unnatural poses while reducing the average error. To constrain the network to generate naturally plausible poses and to further optimize the generator weights, a discriminator network is proposed. The output of the SMPL-based pose predictor plus its ground truth act as inputs to the discriminator network. The discriminator will penalize the generator network for unnatural poses thereby further optimizing its weights to enforce the network to generate naturally plausible poses.

This section first introduces the pre-processing followed by the proposed SMPL-based recurrent neural network and adversarial SMPL-based RNN. Finally, the implementation details are presented, specifying detailing all hyperparameters used in the implementation.

A. Pre-Processing

During the pre-processing step, the pedestrians in input video frames are converted to a gender-neutral SMPL 3D meshes as illustrated in Fig 3. The SMPL model is defined as $M(\theta, \beta) \in \mathbb{R}^{6890\times 3}$ where $\theta \in \mathbb{R}^{3\times K}$ is relative rotation of $K = 23$ joints in axis angle representations and $\beta \in \mathbb{R}^{10}$ is shape's space parameters. The function $M(\theta, \beta)$ generates a triangulated mesh with 6,890 vertices. This is achieved by shaping template vertices conditioned on pose θ and shape β via forward kinematics. When the desired pose is achieved, surface deforming is performed using linear blend skinning.

The pre-processing step utilizes the human body and the shape estimation network (VIBE) [19] to generate a 3D mesh for a given set of frames. The majority of real world applications process a sequence of images, rather than a single image. In contrast to existing single image processing models [31], [21], VIBE proposes a temporal neural network to process a sequence of frames. The temporal network allows continuity of movement in consecutive frames and forces the model to generate naturally-feasible pose sequences ignoring invalid poses. In addition to the temporal network, VIBE employs an adversarial training approach to predict the SMPL body model parameters from an input video. In VIBE, a weak prospective camera is utilized to estimate the global rotation $\delta \in \mathbb{R}^3$ and camera parameters $\gamma = [\tau, s]$. Where $\tau \in \mathbb{R}^2$ represents translation and $s \in \mathbb{R}^2$ represents the scale in the original image space.

For given frames $\{F_t\}_{t=0}^{T}$ with N people, VIBE outputs $\sum_{i=1}^{N}[(P_1^i, P_2^i, ..P_t^i), \beta^i]$ where $P = [\delta, \theta, \gamma]$ is a vector of the global rotation, joints rotation, and translation parameters at time step t for the i^{th} person. The average shape for the i^{th} person is given by β^i, and we assumes that a person's shape does not change across frames.

B. SMPL-based Recurrent Neural Network

SMPL pose parameters for a person at time step 't' can be described as a vector of a global rotation angle(p_{δ_t}), 23 joints rotation angles(p_{θ_t}), and translation parameters(p_{γ_t}) as shown in (1)

$$p_t = [p_{\delta_t}, p_{\theta_t}, p_{\gamma_t}] \tag{1}$$

Traditional 2LR-LSTM [2] considers p_t as a single vector input for a given model. The output of the 2LR-LSTM is given by:

$$\hat{p}_{t+1} = 2LR\left(P_t; W_{2lr}\right) \tag{2}$$

where W_{2lr} denotes the weight matrix for the 2LR-LSTM. For a given input P_t, 2LR-LSTM distribute the network weights equally among all the SMPL components. The proposed SMPL-based Recurrent Neural Network architecture(SA-LSTM) extends 2LR-LSTM using a dedicated LSTM layer for each SMPL component.

For 'n' given past SMPL pose parameters,

$$P = [p_t, p_{t-1}, .., p_{t-n}] \tag{3}$$

the individual SMPL components for global rotation(P_δ), joints rotations(P_θ), and translation parameters(P_γ) are given by:

$$P_\vartheta = \left[p_{\vartheta_t}, p_{\vartheta_{t-1}}, .., p_{\vartheta_{t-n}}\right] \quad \forall\ \vartheta \in (\delta, \theta, \gamma) \tag{4}$$

In SA-LSTM, each SMPL component(δ, θ, γ) is used as an input to the dedicated independent LSTM layer($l_\delta, l_\theta, l_\gamma$). The output of each LSTM layer is given by :

$$pl_\vartheta = l_\vartheta\left(P_\vartheta; W_\vartheta\right) \quad \forall\ \vartheta \in (\delta, \theta, \gamma) \tag{5}$$

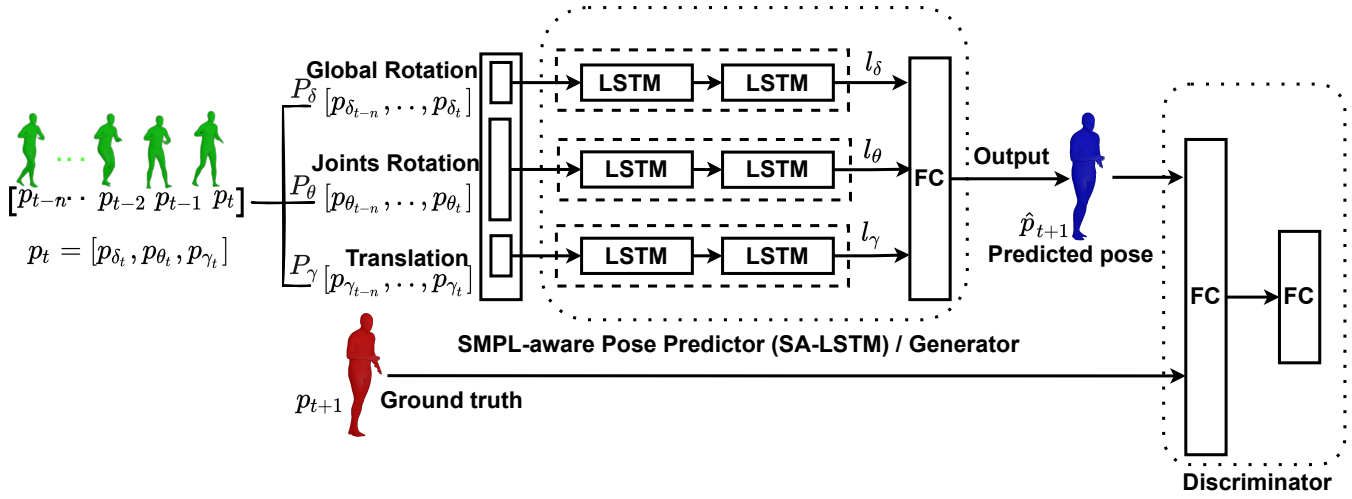

Fig. 2: Proposed Adversarial SMPL-based Recurrent Neural Network Architecture. Each SMPL component vector is applied to dedicated LSTM layers of SMPL-based pose predictor. The SMPL-based pose predictor network is used as a generator in adversarial training. For the given past 'n' poses (represented in green), Pose predictor network will generate a future pose (blue). The predicted future pose and ground truth mesh (red) are used as inputs to the discriminator network

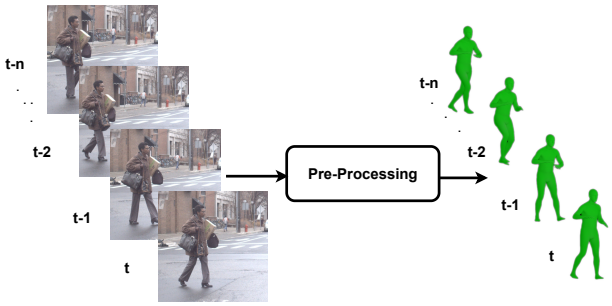

Fig. 3: An illustration of data pre-processing. Images are processed using VIBE [19], a state-of-the art pose and shape estimation network to extract SMPL parameters for each pedestrian.

where $pl_\vartheta, W_\vartheta$ denotes outputs and weights of LSTM layer $l_\vartheta \ \forall \ \vartheta \in (\delta, \theta, \gamma)$. These dedicated LSTM layers allow the network to learn the temporal dependencies of each SMPL component independently. The output of all these layers will be concatenated and applied as input to the fully connected layer to capture the spatio-temporal relation between SMPL components as shown in Fig. 2. The output of the SA-LSTM is given by:

$$\hat{p}_{t+1} = FC\left(\left(pl_\delta + pl_\theta + pl_\gamma\right); W_{fc}\right) \tag{6}$$

W_{fc} denotes the weight matrix of fully connected layer(FC). In SA-LSTM network, the network weights are distributed i.e $W_{SA} = [W_\delta, W_\theta, W_\gamma, W_{fc}]$ for each component to capture the spatial and temporal dependencies between SMPL components. The SA-LSTM is trained using the Mean Absolute Error (MAE) loss function given by:

$$L_{SA} = \mid P_{t+1} - \hat{P}_{t+1} \mid \tag{7}$$

C. Adversarial SMPL-based RNN

The SA-LSTM architecture with MAE loss encourages the network to predict the future 3D pose ($\hat{P}_{t+1}$) based on the temporal dependencies of the past 'n' poses. Naturally implausible poses, however, may still minimize the network loss. To constrain the network to generate naturally plausible poses, an adversarial SMPL-based RNN (ADV-SA-LSTM) is proposed. The ADV-SA-LSTM consists of a generator(G) and a discriminator(D) networks trained in an adversarial manner. The discriminator network is used to penalize the unnatural poses while allowing the natural poses and SA-LSTM is used as a generator as shown in Fig. 2.

1) Generator: In a traditional GAN framework, the generator network is used to capture data distribution and generate samples to fool the discriminator network. In proposed ADV-SA-LSTM, the goal of the generative network is to learn temporal mapping from past 'n' poses and predict the future pose. In contrast to traditional GAN's, the pose prediction network SA-LSTM is used as a generator network. As training progresses, a generator capable to generate more accurate poses, which are used to fool the discriminator network. The adversarial loss of the generator L_{adv} is given by:

$$L_{Adv} = \mathbb{E}_{\rho \sim \hat{P}_{t+1}}[(D(\rho) - 1)^2] \tag{8}$$

where $\hat{P}_{t+1}$ is the manifold of generated SMPL parameters. The adversarial loss from the discriminator (L_{Adv}) is combined with the SA-LSTM loss (L_{SA}) to train the generator. The generator total loss is given by:

$$L_G = \xi_{sa} L_{SA} + \xi_g L_{Adv} \tag{9}$$

where ξ_{sa} and ξ_g are the hyper-parameters to control the weights of the SA-LSTM and adversarial losses respectively. The L_{SA} loss penalizes the network according to the

predicted and ground truth poses, whereas the L_{Adv} loss constrains network to output plausible poses.

2) Discriminator: The discriminator network is used to distinguish the real data from generated data and forces the network to output valid data. The inputs for discriminator are manifold of ground truth (p_{t+1}) and predicted poses ($\hat{p}_{t+1}$). From the inputs, the discriminator should learn whether the generated pose corresponds to a manifold of the natural pose or not. A multi-layer perception is used as discriminator with the last layer outputting a value between 0 and 1, describing the probability of the pose belong to the manifold of natural plausible poses. The Adversarial loss of the discriminator is given by:

$$L_D = \mathbb{E}_{\rho\sim P_{t+1}}[(D(\rho)-1)^2] + \mathbb{E}_{\rho\sim \hat{P}_{t+1}}[(D(\rho))^2] \quad (10)$$

where $\hat{P}_{t+1}$ and P_{t+1} are manifold of generated and ground truth SMPL parameters respectively.

D. Implementation Details

For the given past 'n' poses where each pose is represented with 76 SMPL parameters, and a batch size B, the input shape of the network will be $[B, n, 76]$. The input to the SA-LSTM (Generator) will be distributed across the three SMPL component layers, with shape of (n_δ, n_θ and n_γ) as $[B, n, 3]$,$[B, n, 69]$ and $[B, n, 4]$ respectively. Each LSTM layer in l_δ and l_γ consists of 2 hidden units and the l_θ LSTM layers are defined with 8 hidden units (we experimented with multiple combinations of hidden layers and found out given combination generated better results). The discriminator network architecture consists of a single MLP layer with 250 neurons with tanh activation function. The final layer predicts a single fake or real probability for each sample using a sigmoid activation function.

IV. EVALUATION

In this section, baseline models will be presented first, followed by a brief description of the datasets and finally the experimental setup.

A. Baseline models

The proposed ADV-SA-LSTM network is compared with several baseline methods derived from the start-of-the-art 2LR-LSTM [10], [2] architecture. The 2LR-LSTM is a two-layer stacked LSTM recurrent neural network architecture followed by a fully connected layer. In addition to pose parameters, frame difference parameters [32] are also used as the model inputs. For a given pose parameters P (see (3)), the frame difference input vector is given by:

$$P_{fd} = [(p_{t-1} - p_t), (p_{t-2} - p_{t-1}), .., \big(p_{t-(n-1)} - p_{t-n}\big)] \quad (11)$$

For the frame difference vector P_{fd}, SMPL components vectors defined as:

$$P_{fd_\vartheta} = [\big(p_{\vartheta_{t-1}} - p_{\vartheta_t}\big), \big(p_{\vartheta_{t-2}} - p_{\vartheta_{t-1}}\big), .., \big(p_{\vartheta_{t-(n-1)}} - p_{\vartheta_{t-n}}\big)] \quad \forall\ \vartheta \in (\delta, \theta, \gamma) \quad (12)$$

By combining the different model inputs and training objective functions, the following baseline methods from [10], [2] are derived to compare effectiveness of the proposed model:

2LR-P-MSE: The 2LR-LSTM architecture with pose parameters as model inputs and MSE loss function as a training objective.

2LR-FD-MSE: The 2LR-LSTM architecture with frame difference parameters as model inputs and MSE loss function as a training objective

2LR-P-MAE: The 2LR-LSTM architecture with pose parameters as model inputs and MAE loss function as a training objective

2LR-FD-MAE: The 2LR-LSTM architecture with frame difference parameters as model inputs and MAE loss function as a training objective

B. Datasets

For ease of comparison with state-of-the-art[2] and to demonstrate model performance in both real and simulated environment, we report results from the PedX[24] and BEHAVE[25] datasets, as well as combination of the two.

PedX: The PedX dataset [24] is a collection of more than 10K images from three road intersections with heavy pedestrian traffic in Michigan, USA. Images are collected using two stereo RGB camera pairs with six frames per second (FPS). This dataset contains both original and gamma rectified images of four cameras of the same scene. In this work, a single camera (specified as grn43E3) data is processed using VIBE [19] and the extracted SMPL parameters are used as inputs to the model.

BEHAVE: The BEHAVE dataset [25] consists of more than 76K images with multiple scenarios of people acting out various interactions. The images are collected using commercial tripod mounted camcorder at 25 FPS.

HYBRID: The HYBRID dataset is the combination of data from the PEDX and BEHAVE datasets. To avoid the bias due to the substantial variation in the number of images, we combined the PedX dataset with the equal number of pose parameters randomly sampled from the BEHAVE dataset. The number of samples in the HYBRID dataset are approximately 20K.

Both the PedX and BEHAVE datasets are captured using a fixed camera. The PedX dataset contains images from real scenarios where the subjects may or may not be aware of whether they are being recorded. In contrast, the BEHAVE dataset is recorded by actors performing various actions. Our framework will therefore be evaluated against both real and simulated environments with various lighting and background conditions.

C. Experimental Setup

All the experiments are performed using batch size B=50, n=4 past poses, and epochs of size 500. The value 'n' of past poses was selected assuming that a pedestrian generally completes walking cycle in 5-6 frames [2]. An Adam optimizer [33] is used with a learning rate of 0.0003 and 0.0006

for the generator and the discriminator respectively. Finally, weighting coefficients for the generator loss are $\xi_{sa} = 0.40$ and $\xi_g = 1.0$. VIBE is used to process all datasets using the following configuration : batch size=6, vibe batch size = 100, and a minimum number of frames=6. Each dataset is divided into 70% training, 15% validation, and 15% testing sets.

V. RESULTS AND DISCUSSION

This section first introduces the evaluation metrics and then presents the quantitative and qualitative results. The quantitative results are the average results from three random initializations of pose parameters.

A. Evaluation Metrics

The output of the proposed models are 76 parameters including 23 joint rotation angles in axis angle format ($23 \times 3 = 69$), three global rotation angles, and four camera parameters. Note that the shape parameters are considered constant from frame to frame. All models are evaluated using the standard vertex-to-vertex RMSE error (Vertex), Mean Per Joint Position Error (MPJPE) [34], and Mean Per Joint Angle Errors (MPJAE) [35]. The predicted and ground truth 24 joint rotation angles (including the global rotation) are used to calculate MPJAE. From the predicted SMPL parameters, a 3D wireframe mesh is regressed [18] to calculate the Vertex and MPJPE metrics. The standard evaluation metrics Vertex, MPJPE and MPJAE calculates the average errors across all SMPL components. In order to evaluate the effectiveness of the proposed SMPL-based architecture on global rotation, the following evaluation metrics are also introduced:

GR-E: The angular difference between the global rotation in the predicted and ground truth SMPL parameters.

JA-E: The Mean Per Joint Angle Error for 23 joints excluding the global rotation angle.

TR: The Mean Translation Error between predicted and ground truth SMPL parameters.

TABLE I: Pose prediction results on PedX dataset. Units for Vertex and MPJPE are $\times 10^{-3}$m

Model	Vertex	MPJPE	MPJAE	JA-E	GR-E	TR
2LR-P-MSE	635.54	421.81	8.51	5.68	73.58	202.43
2LR-FD-MSE	579.48	368.98	7.63	5.07	66.46	91.69
2LR-P-MAE	437.18	219.92	6.29	4.80	40.69	157.06
2LR-FD-MAE	380.76	162.91	5.54	4.41	31.66	82.22
SA-LSTM	**353.58**	**135.46**	**5.54**	**4.69**	**25.08**	**81.20**
ADV-SA-LSTM	**276.46**	**69.81**	**5.18**	**4.72**	**15.84**	**82.63**

TABLE II: Pose prediction results on BEHAVE dataset. Units for Vertex and MPJPE are $\times 10^{-3}$m

Model	Vertex	MPJPE	MPJAE	JA-E	GR-E	TR
2LR-P-MSE	550.84	329.32	5.59	3.30	58.25	126.03
2LR-FD-MSE	473.54	276.16	4.79	2.76	51.42	72.69
2LR-P-MAE	327.58	126.25	3.86	2.84	27.24	117.68
2LR-FD-MAE	302.78	111.88	3.55	2.63	24.64	72.32
SA-LSTM	**289.91**	**93.80**	**3.49**	**2.69**	**21.97**	**71.71**
ADV-SA-LSTM	**289.06**	**91.72**	**3.47**	**2.69**	**21.51**	**71.74**

TABLE III: Pose prediction results on HYBRID dataset. Units for Vertex and MPJPE are $\times 10^{-3}$m

Model	Vertex	MPJPE	MPJAE	JA-E	GR-E	TR
2LR-P-MSE	567.48	345.02	6.25	3.89	60.65	158.18
2LR-FD-MSE	475.26	279.14	5.24	3.19	52.29	76.40
2LR-P-MAE	360.05	152.89	4.40	3.24	31.25	104.32
2LR-FD-MAE	307.59	107.88	3.82	2.96	23.58	74.38
SA-LSTM	**280.43**	**83.89**	**3.77**	**3.06**	**20.11**	**74.32**
ADV-SA-LSTM	**277.99**	**82.43**	**3.77**	**3.03**	**20.75**	**74.41**

B. Results

Table I presents the quantitative results for the PedX dataset. and shows that the proposed SA-LSTM and ADV-SA-LSTM methods are able to produce improved results compared to the baseline methods. Models with frame difference as inputs perform better than those with pose parameters as inputs. The improved performance might be due to the ability of the frame difference inputs to capture the temporal difference in consecutive frames. Furthermore, training the model with the MAE loss function yields better results compared to the MSE loss function. For this reason, the SA-LSTM model uses the frame difference parameters as the model input parameters and MAE as training objective and ADV-SA-LSTM also uses MAE as L_{SA}.

Tables II and III present the results for the BEHAVE and HYBRID datasets respectively. For all baseline methods, it can be observed that the models trained on PedX are underperforming compared to the models trained with the BEHAVE and HYBRID datasets. This may be due to the difference in the number of training and testing samples across datasets and camera parameters (The PedX dataset is collected using 6 FPS, and the BEHAVE dataset is collected with 25 FPS). Regardless of the camera parameters and training samples, the proposed ADV-SA-LSTM and SA-LSTM models yield improved results across datasets compared to state-of-the-art baseline models.

SMPL-based Architecture: It can be observed that proposed SA-LSTM and ADV-SA-LSTM architectures are able to minimize the global rotation angle error (GR-E) due to introduction of dedicated layer. Although JA-E is slightly increased, it has a minimum impact on the MPJAE and overall model performance compared to baseline models.

Qualitative Results: Fig. 4 shows some of the 3D pose prediction results on the PEDX, BEHAVE, and HYBRID datasets. The ground truth and predicted 3D wireframe meshes are represented in red and blue respectively. The global rotation angle have a huge impact on the rendered 3D mesh position due to its centrality. The proposed ADV-SA-LSTM model successfully minimizes the global rotation angle error along with 23 joint rotation errors, and enforces the model to predict naturally feasible human poses.

VI. CONCLUSION AND FUTURE WORK

This work presents a novel SMPL-based recurrent neural network for 3D pedestrian pose prediction. While the current joint-rotation-based pose prediction methods work reasonably well, most of them ignore the centrality of

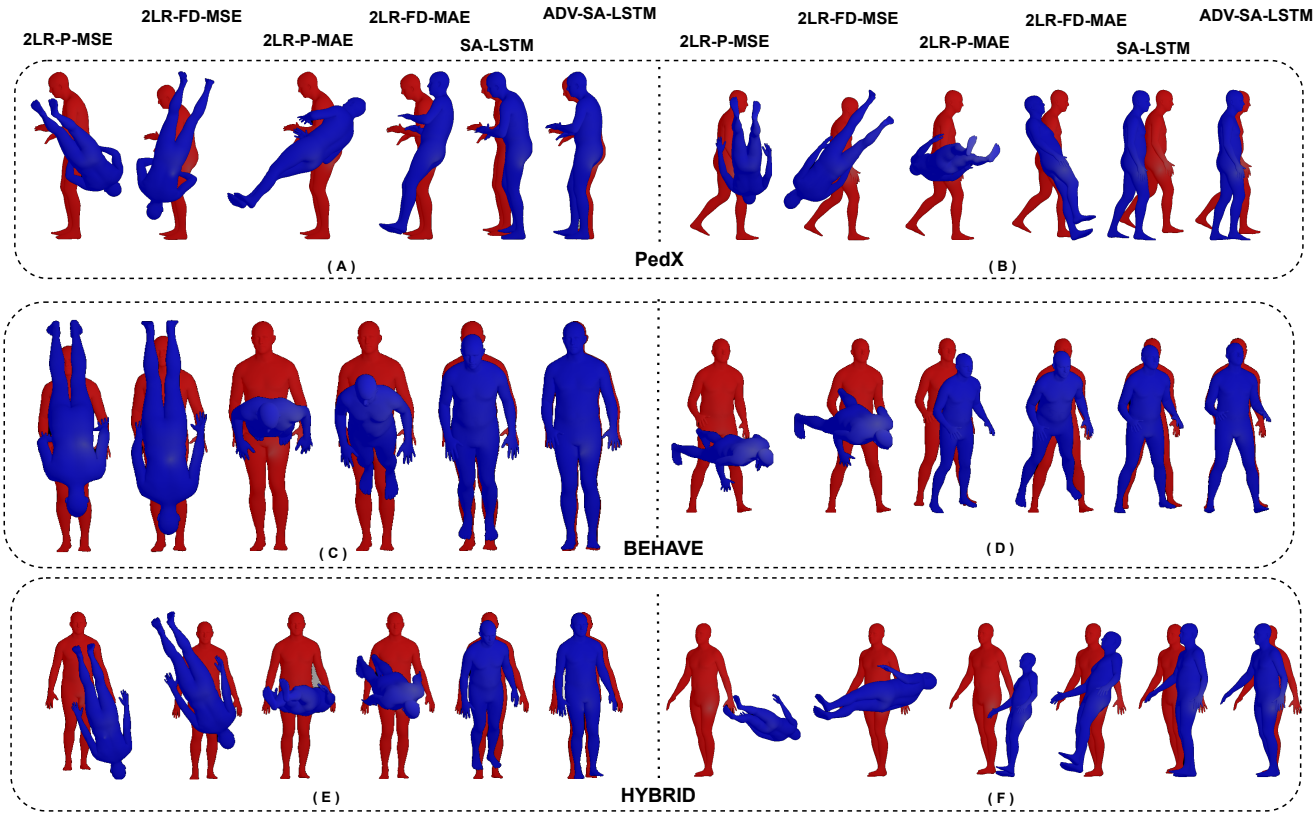

Fig. 4: Qualitative 3D pose prediction results comparison for PedX [24], BEHAVE [25], and HYBRID datasets. The Ground truths are represented in red and the predicted poses are in blue. Two sets of results ((A,B) for PedX, (C,D) for BEHAVE and (E,F) for HYBRID) are presented for each dataset with each showing qualitative results for 2LR-P-MSE,2LR-FD-MSE,2LR-P-MAE,2LR-FD-MAE,SA-LSTM and ADV-SA-LSTM models starting from left.

global rotation and translation parameters with respect to pose parameters. Ignoring the centrality of each component forces the model to assign equal weights for all components, resulting in high error accumulation in the kinematic chain leading to qualitative errors. The proposed architecture extends current models by introducing the novel SMPL-based architecture to address the centrality and network weight distribution. Furthermore, a discriminator network is proposed to constrain the network to generate naturally feasible poses.

The results show that the proposed SMPL-based architecture is able to reduce the average error term introduced by SMPL components and achieves improved performance compared to the current state-of-the-art. Future work include modelling inter joint spatial dynamics using graph neural networks to further improve the model performance. In addition, current work can be extended by introducing a novel dataset with increased pedestrian presence across multiple cameras to validate the model performance in the comprehensive temporal domain.

VII. ACKNOWLEDGMENTS

This work was conducted with the financial support of the Science Foundation Ireland Centre for Research Training in Digitally-Enhanced Reality (d-real) under Grant No. 18/CRT/6224.

REFERENCES

[1] Hsu-kuang Chiu, Ehsan Adeli, Borui Wang, De-An Huang, and Juan Carlos Niebles. Action-agnostic human pose forecasting. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1423–1432. IEEE, 2019.

[2] Xiaoxiao Du, Ram Vasudevan, and Matthew Johnson-Roberson. Bio-LSTM: A Biomechanically Inspired Recurrent Neural Network for 3-D Pedestrian Pose and Gait Prediction. *IEEE Robotics and Automation Letters*, 4(2):1501–1508, April 2019.

[3] M. Sami Zitouni, Andrzej Sluzek, and Harish Bhaskar. Towards understanding socio-cognitive behaviors of crowds from visual surveillance data. *Multimedia Tools and Applications*, 79(3-4):1781–1799, January 2020.

[4] Houman Hediyeh, Tarek Sayed, and Mohamed H Zaki. The use of gait parameters to evaluate pedestrian behavior at scramble phase signalized intersections. page 12, 2014.

[5] Alexandra Willisô, Nathalia Gjersoe, Catriona Havard, Jon Kerridge, and Robert Kukla. Human movement behaviour in urban spaces: implications for the design and modelling of effective pedestrian environments. page 24.

[6] Mohamed H Zaki and Tarek Sayed. Automated Analysis of Pedestrian Group Behavior in Urban Settings. *IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS*, 19(6):10, 2018.

[7] Dario Pavllo, Christoph Feichtenhofer, Michael Auli, and David Grangier. Modeling human motion with quaternion-based neural networks. *International Journal of Computer Vision*, pages 1–18, 2019.

[8] Erwin Wu and Hideki Koike. Futurepose - mixed reality martial arts training using real-time 3d human pose forecasting with a rgb camera.

In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1384–1392, 2019.

[9] Sam Toyer, Anoop Cherian, Tengda Han, and Stephen Gould. Human pose forecasting via deep markov models. In *2017 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, pages 1–8, 2017.

[10] Katerina Fragkiadaki, Sergey Levine, Panna Felsen, and Jitendra Malik. Recurrent network models for human dynamics. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4346–4354, 2015.

[11] Maosen Li, Siheng Chen, Yangheng Zhao, Ya Zhang, Yanfeng Wang, and Qi Tian. Dynamic Multiscale Graph Neural Networks for 3D Skeleton Based Human Motion Prediction. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 211–220, Seattle, WA, USA, June 2020. IEEE.

[12] Emre Aksan, Manuel Kaufmann, and Otmar Hilliges. Structured Prediction Helps 3D Human Motion Modelling. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7143–7152, Seoul, Korea (South), October 2019. IEEE.

[13] Emad Barsoum, John Kender, and Zicheng Liu. HP-GAN: Probabilistic 3D Human Motion Prediction via GAN. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1499–149909, Salt Lake City, UT, USA, June 2018. IEEE.

[14] Liang-Yan Gui, Yu-Xiong Wang, Xiaodan Liang, and José M. F. Moura. Adversarial Geometry-Aware Human Motion Prediction. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision – ECCV 2018*, volume 11208, pages 823–842. Springer International Publishing, Cham, 2018. Series Title: Lecture Notes in Computer Science.

[15] Ye Yuan and Kris Kitani. Ego-pose estimation and forecasting as real-time pd control. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10082–10092, 2019.

[16] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *ACM transactions on graphics (TOG)*, 34(6):1–16, 2015.

[17] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7122–7131, 2018.

[18] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J Black. Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In *European conference on computer vision*, pages 561–578. Springer, 2016.

[19] Muhammed Kocabas, Nikos Athanasiou, and Michael J Black. Vibe: Video inference for human body pose and shape estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5253–5263, 2020.

[20] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10975–10985, 2019.

[21] Yu Sun, Yun Ye, Wu Liu, Wenpeng Gao, Yili Fu, and Tao Mei. Human mesh recovery from monocular images via a skeleton-disentangled representation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5349–5358, 2019.

[22] Raviteja Vemulapalli and Rama Chellapa. Rolling rotations for recognizing human actions from 3d skeletal data. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4471–4479, 2016.

[23] Yu Chen, Chunhua Shen, Xiu-Shen Wei, Lingqiao Liu, and Jian Yang. Adversarial PoseNet: A Structure-Aware Convolutional Network for Human Pose Estimation. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 1221–1230, Venice, October 2017. IEEE.

[24] Wonhui Kim, Manikandasriram Srinivasan Ramanagopal, Charles Barto, Ming-Yuan Yu, Karl Rosaen, Nick Goumas, Ram Vasudevan, and Matthew Johnson-Roberson. Pedx: Benchmark dataset for metric 3d pose estimation of pedestrians in complex urban intersections. *IEEE Robotics and Automation Letters*, 2019.

[25] Scott Blunsden and RB Fisher. The behave video dataset: ground truthed video for multi-person behavior classification. *Annals of the BMVA*, 4(1-12):4, 2010.

[26] Yu Sun, Qian Bao, Wu Liu, Yili Fu, Michael J Black, and Tao Mei. Monocular, one-stage, regression of multiple 3d people. *arXiv preprint arXiv:2008.12272*, 2020.

[27] Dario Pavllo, David Grangier, and Michael Auli. QuaterNet: A Quaternion-based Recurrent Model for Human Motion. *arXiv:1805.06485 [cs]*, July 2018. arXiv: 1805.06485.

[28] Yong Du, Wei Wang, and Liang Wang. Hierarchical recurrent neural network for skeleton based action recognition. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1110–1118, Boston, MA, USA, June 2015. IEEE.

[29] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *arXiv preprint arXiv:1406.2661*, 2014.

[30] Ching-Hang Chen, Ambrish Tyagi, Amit Agrawal, Dylan Drover, Stefan Stojanov, and James M Rehg. Unsupervised 3d pose estimation with geometric self-supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5714–5724, 2019.

[31] Angjoo Kanazawa, Jason Y Zhang, Panna Felsen, and Jitendra Malik. Learning 3d human dynamics from video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5614–5623, 2019.

[32] Chaohui Zhan, Xiaohui Duan, Shuoyu Xu, Zheng Song, and Min Luo. An improved moving object detection algorithm based on frame difference and edge detection. In *Fourth international conference on image and graphics (ICIG 2007)*, pages 519–523. IEEE, 2007.

[33] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[34] 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(7):1325–1339, 2013.

[35] 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 *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 601–617, 2018.