Disentangling Human Dynamics for Pedestrian Locomotion Forecasting with Noisy Supervision

Karttikeya Mangalam1,3*, Ehsan Adeli1, Kuan-Hui Lee2, Adrien Gaidon2, Juan Carlos Niebles1
1 Stanford University  2 Toyota Research Institute  3 University of California, Berkeley
mangalam@cs.berkeley.edu {eadeli, jniebles}@cs.stanford.edu {kuan.lee, adrien.gaidon}@tri.global

Abstract

We tackle the problem of Human Locomotion Forecasting, a task for jointly predicting the spatial positions of several keypoints on human body in the near future under an egocentric setting. In contrast to the previous work that aims to solve either the task of pose prediction or trajectory forecasting in isolation, we propose a framework to unify these two problems and address the practically useful task of pedestrian locomotion prediction in the wild. Among the major challenges in solving this task is the scarcity of annotated egocentric video datasets with dense annotations for pose, depth, or egomotion. To surmount this difficulty, we use state-of-the-art models to generate (noisy) annotations and propose robust forecasting models that can learn from this noisy supervision. We present a method to disentangle the overall pedestrian motion into easier to learn subparts by utilizing a pose completion and a decomposition module. The completion module fills in the missing key-point annotations and the decomposition module breaks the cleaned locomotion down to global (trajectory) and local (pose keypoint movements). Further, with Quasi RNN as our backbone, we propose a novel hierarchical trajectory forecasting network that utilizes low-level vision domain specific signals like egomotion and depth to predict the global trajectory. Our method leads to state-of-the-art results for the prediction of human locomotion in the egocentric view.

1. Introduction

Pedestrians are one of the most vulnerable and prevalent entities in self-driving scenarios. The ability to predict their dynamics in the near future can assist in making proper decisions for immediate next action the vehicle needs to take. Forecasting human locomotion is useful in several downstream tasks for self-driving cars such as reasoning about pedestrian intent, path planning and reactive control. It is also relevant for social robots that require anticipating future human movements for collision avoidance and navigation. Human dynamics or locomotion can be defined in terms of the joint spatial movement of several keypoints on the human body. Forecasting future human dynamics can lead to foretelling certain undesirable activities like falling, which can then be planned for. It is the final product of a complex interaction between large scale trajectorial motion and finer body limb movements. In previous works, often one of these two critical components is studied. In this work, we propose to predict human locomotion by disentangling the global and local components. An additional challenge to this task is the scarcity of human annotated pedestrian pose datasets in egocentric view. To this end, we use off-the-shelf estimation models to generate ‘noisy’ ground-truth to train our model.

Developing computational methods for modeling human dynamics and forecasting how the pose might change in the future is itself an extremely challenging task. The first level of complexity comes from the inherent multimodal nature of pedestrian locomotion distribution. The space of possi-
ble future locomotion is both complex and uncertain even conditioned on the observed history. Furthermore, in real scenarios, the pedestrians often occlude with each other or other objects in the scene. Moreover, obtaining full annotations of the dynamics (trajectory and pose) is a very intensive task. Therefore, in contrast to previous works [1, 2, 3, 4, 5], applying fully-supervised methods is hardly possible for real-world in-the-wild applications.

In this work, we address the challenge of predicting human locomotion in the wild without explicit human labelled ground-truth supervision. We start with noisy machine generated supervision for forecasting using off the shelf pose detectors like OpenPose [6, 7, 8]. We then propose a forecasting pipeline (Fig. 1) to complete the missing detections and somewhat denoise the detected poses. Furthermore, we identify the complexities associated with learning to forecast human locomotion in the wild in an end-to-end fashion and introduce a disentangling method to overcome this problem. While almost all of the previous work for prediction of human dynamics [1, 2, 9, 10] target the problem of predicting global (i.e., trajectories [11, 12, 13]) and local movements (i.e., human pose and joints motion [1, 3]) separately, we propose to unify the previous works on future pedestrian pose and trajectory prediction under the single problem of locomotion forecasting. To achieve this goal, we propose a sequence-to-sequence pose forecasting model (see Fig. 2) that build upon this decomposition. Lastly, we show that with these advancements our proposed method achieves better results than several other previous works and confirm our hypothesis of disentangling human dynamics for pedestrian locomotion forecasting in the wild.

In summary, our contributions are two-fold. First, we propose the pose completion and decomposition modules that uses disentanglement of global and local motion to reduce the overall forecasting complexity and enables learning under noisy supervision. Second, we present the pose prediction module consisting of a novel egocentric trajectory prediction network based on the Encoder-recurrent-Decoder architecture that utilizes domain specific settings of egomotion to forecast these different granularity streams. They are merged for the final locomotion forecasting.

2. Related Work

Trajectory Prediction: The literature for trajectory prediction can be split into two major categories of work: (1) works that aim to predict human trajectories from the points of view of either other humans in the scene (human-human) [4, 14, 15, 16] or from a top down view [4, 15, 16, 17, 18]; and (2) works that predict the paths of other vehicles on the road from the viewpoint of a car (car-car) [5, 19].

The previous work on human-human trajectory prediction often solves the problem from a top-down prospective i.e., bird’s eye view. For instance, Social LSTM [4] jointly predicted the paths of all the people in a scene by taking into account the common sense rules and social conventions that humans typically utilize as they navigate in shared environments. Social GAN [15] defined a recurrent sequence-to-sequence model using Generative Adversarial Networks [20] (GAN) that observes motion histories to predict future behavior with spatial pooling thereby tackling the multi-modality associated with motion prediction. In [17], Zou et al. proposed to infer the latent factors of human decision-making process in an unsupervised manner by extending the Generative Adversarial Imitation Learning framework to anticipate future paths of pedestrians (human-car) from the top-down view. In [16], SoPhie introduces an attentive GAN that predicts individuals’ paths by taking into account physical constraints, i.e., scene context and other peoples’ trajectories. Several work model the human trajectories with social forces [21] and attraction repulsion models using functional objects [22]. Also, approaches using probabilistic modelling of trajectories with Gaussian processes [23] and mixture models [24] have been employed. Among the few works for trajectory prediction in the human’s egocentric view, [14] predicted trajectories of other pedestrians in an egocentric setting with a tri-stream CNN architecture.

Car trajectory prediction from egocentric view were done by explicitly modeling other agents (cars) along with motion with respect to the camera. Yao et al. [25] proposed a multi-stream recurrent neural network (RNN) encoder-decoder model for predicting vehicle location and scale using pixel-level observations. DESIRE [19] was proposed to generate several possible trajectories for vehicles using both scene and vehicles, which were then ranked in an Inversion of Control (IOC) framework to choose the most likely one. Similarly, Huang et al. [5] introduced an uncertainty-aware method for prediction of self-trajectories.

One of the most crucial tasks for self-driving car or social robots is to predict future dynamics of pedestrians (i.e., car-human task). However, recent work focuses on human-human or car-car prediction, both of which domains lack either egocentric cues like monocular depth cue and large camera motion due to ego-vehicle or pedestrian centered features like human pose and intention.

Pose Forecasting: A number of recent work predicted 3D poses using fully annotated data with exact 3D positions. These works included Structural RNN [3], convolutional sequence-to-sequence model with encoder-decoder architecture [26], and human motion prediction with adversarial geometry-aware based on geodesic loss [27]. Some other works introduced models that can predict poses in both 2D and 3D. Triangular-Prisms RNN (TP-RNN) [1] proposed an action-agnostic pose forecaster, and Martinez et al. [2] proposed a method based on residuals and a careful design of RNNs for human dynamics. These works concluded that zero-velocity is a very strong baseline for pose forecasting.
3. Proposed Method

We frame the task of forecasting human locomotion in egocentric view (of the vehicle) as a sequence-to-sequence problem. Suppose \( p_t \) denotes the human pose at time \( t \), which contains \( d \) two-dimensional joints, i.e., \( p_t = \{ (u_i^t, v_i^t) \}_{i=1}^d = \{ u_i \}_{i=1}^d \). Forecasting human locomotion is then defined as: given \( t_p \) previous poses, \( \{ p_t \}_{t=t_p+1}^{t_f} \equiv \{ p_{t-t_p}, \ldots, p_t \} \), predict the position of each of these joints for \( t_f \) future frames. In other words, predict the sequence \( \{ p_t \}_{i=t+1}^{t_f} \equiv \{ p_{t+1}, \ldots, p_{t+t_f} \} \). Our method is illustrated in Fig. 2. We start with noisy pose estimates generated using state of the art pre-trained models (Section 3.1). The missing and low confidence joints are filled in using the pose completion module (Section 3.2). The completed poses are then split into global and local streams (Section 3.3) that separate concurrent motions. The prediction modules forecast the future streams as outlined in Section 3.5. Finally, these streams are merged to predict future pedestrian locomotion.

3.1. Machine Generated Noisy Supervision

**Keypoint Detection** We use state-of-the-art models for *multiple-person keypoint detection* module to autonomously generate dense but noisy frame-level supervision for human poses \( p_t \). To the best of our knowledge, there are no existing datasets with manually annotated pose data that can be used for the task of human locomotion forecasting in the wild. So, we use pre-trained models on task-specific datasets to infer pose and depth on our datasets. For each frame in the video, we create masked frames where everything except the pedestrians are masked out using the human labelled bounding box annotations. These masked frames are then processed through OpenPose [6, 7, 8], a pre-trained pose detection model to generate pose labels for every pedestrian in the frame. We notice that using masked full frames for gen-
Monocular Depth Estimation We autonomously estimate depth in a monocular camera using SuperDepth [35]. SuperDepth extends a subpixel convolutional layer for depth super-resolution, where high-resolution disparities are synthesized from their corresponding low-resolution convolutional features. We train the depth estimation model in a self-supervised fashion as proposed in [35]. Fig. 4 shows qualitative examples of the depth estimation by SuperDepth.

Egomotion Estimation We use the state-of-the-art-model unsupervised model [36] for autonomously estimating the camera motion that occurs between consecutive frames due to the movement of the egovehicle. However, since the method proposed in [36] can reliably estimate the scene transformation matrix $T^j_i$ between frames $i$ and $j$ only frame very close to each other ($|i - j| \leq 5$ in a 30 fps video) and the time horizons relevant for prediction are much larger ($\sim 30$ frames), we chain the estimate between different overlapping short windows to obtain $T^i_{t+k}$ for $k > 5$.

3.2. Pose Completion

Off-the-shelf pre-trained pose detectors often result in noisy detection of keypoints. In addition, the small scale and off co-planar orientation of pedestrians in real-world 2D videos, the pose detector may miss detecting some keypoints even for a completely visible occlusion-free pedestrians in the frame. Hence, we propose a pose completion network for completing the detected human poses obtained from OpenPose. This processing has a two-fold benefit. First, it suppresses noise by filling in the low confidence output with better estimates. Second and importantly, it enables us to decompose the motion (Section 3.3) with noisy data. Separating uncompleted global and local components of motion would be perplexing as the joints flicker frequently.

Network Architecture We use an over-complete symmetric autoencoder similar to [33, 32]. The network is only trained on the subset of the total training data that has been assigned high confidence scores by Openpose. In other words, the poses are filtered by thresholding on OpenPose confidence scores to only include the examples with $c^i_j > \alpha_c, \forall i \in \{1, \ldots d\}$. These high confidence examples are then used to train a symmetric autoencoder with dropout that embeds the $2d$-dimensional vectors $p_t$ to the latent dimension $d_{ne}$. Then, it maps them back to the denoised detections $\hat{p}_t$. Training with dropout in the input layer on high confidence examples, helps model the effect of missing data. Furthermore, supervising the loss function with the ground-truth on these good examples and the information bottleneck in the form of a narrow layer in the middle of network allows the model to learn to reconstruct the complete pose even if some of the joints are missing. The training procedure details are in Section 4.2.

Completion The trained autoencoder is then used to estimate the full pose on all the low confidence noisy detections.

![Figure 4: Examples of the input (left col.) and the output (right col.) of the monocular depth estimation module [35].](Image 427x109 to 543x174)
3.3. Disentangling Pedestrian Locomotion into Global and Local Streams

After noise suppression using the autoencoder and filling in the missing joints, we propose to disentangle the overall pose locomotion sequence \( \{p_j\}_{j=t-t_p+1} \) into two streams. A global motion stream encodes the overall rigid body motion of the pedestrian \( \{p_j^{\text{global}}\}_{j=t-t_p+1} \) and a local motion stream encodes the motion of the body with respect to the global stream \( \{p_j^{\text{local}}\}_{j=t-t_p+1} \). We hypothesize that these separate streams capture different granularities of simultaneous concurrent motion and need to be modelled and forecasted separately. The global motion stream \( p^{\text{global}} \) models large scale movements of the pedestrian position with respect to the camera, such as those caused by the trajectory or egocentric motion of the camera mounted on the moving car etc. The local motion stream captures the effect of depth change on the overall pose size and the movement of different joints of the pedestrian with respect to the global motion stream such small repetitive motions like swinging of arms.

Our proposal to disentangle global and local motion is motivated by the relative difference in the nature and magnitude of motion exhibited by these streams. Consider for example, the actual path \( \{p_j\}_{j=t-t_p+1} \) traced by the elbow though space. In the original formulation this trajectory is quite complex, consisting of several swirls generated from gyrating around the shoulder and drifting with the rest of the body. In addition, this disentangling allows to significantly reduce the overall complexity, since each of the streams now model a much simpler and easier to predict motion.

We propose to use the neck joint sequence denoted by \( (u^i_j, v^i_j)_{j=t-t_p+1} \) as a representation of the global stream, because the neck is the most widely observed joint in the dataset. Note that this representation is possible only after filling in the missing data as discussed in Section 3.2. The local stream \( p^{\text{local}} \) (denoted by \( \vec{u} \)) is then derived by reparameterizing the original stream \( p_t \) (denoted by \( \vec{v} \)) as

\[
\vec{u}_t^i = (u_t^i, v_t^i) - (u_t^i, v_t^i)
\]

\[
\forall i \in \{1, \ldots, d\}, i \neq \tilde{i}, t \in \{t-t_p+1, \ldots, t\}
\]

3.4. Quasi-RNN Encoder-Decoder

The Quasi-Recurrent Neural Network [34] forms the backbone of the sequence-to-sequence learning structure described in Section 3.5. QRNNs consist of alternating convolutional and recurrent pooling module and is designed to parallelize efficiently better than vanilla LSTMs. Similar to the findings of Bradbury et al. [34], we found that Quasi-RNN trains faster (825 ms/batch) compared to LSTM (886 ms/batch) on a single GPU under same parameters and yields faster convergence for similar model capacity.

Network Architecture Our quasi-RNN encoder-decoder has \( N \) layers of alternate convolutions and recurrent pooling, both in the input encoder and the output decoder. As illustrated in Figs. 3 and 7, the recurrent pooling is a thin aggregation function applied to the convolutional activations. The encoder churns through the latent representation of the \( t_p \) previous poses and encodes the necessary information into a context vector. This vector is then consumed by the
QRNN decoder to forecast the future $t_f$ poses mapped back to the same latent space. The dotted arrow represents the use of the technique of teacher forcing [37] for training recurrent networks for motion prediction, as also proposed by [2].

3.5. Forecasting the Future

After being split into separate granularities, the global and local streams are then forecasted separately with different prediction modules as described in the following.

Local Stream Forecasting As shown in Fig. 7, the filled in and decomposed pose $\{p_j^{\text{local}}\}_{j=t-t_p+1}^{t}$ is used as the input to the pose prediction module. This module comprises of a spatial encoder with the latent dimension $d_{ae}$. The weights of this spatial encoder are separately trained using the autoencoder while the complexity of the latent space is similar. The forecasting is processed in the latent space with $N_{\text{local}}$ layers of the QRNN Encoder Decoder module, briefly described in Section 3.4. We use the latent space to forecast because as reported in [32, 10] and also as confirmed by our pose completion module experiments, the human pose lies on a low dimensional manifold because of the various kinematic constraints enforced by the human body. Thus, forecasting in this lower dimensional denser space makes the prediction easier for the quasi RNN module. The predicted latent pose is then mapped back into the image space with the spatial decoder to forecast $\{p_j^{\text{local}}\}_{j=t}^{t+f}$

Global Stream Forecasting As mentioned in Section 3.3, the global stream captures the effects of camera movement, perspective distortion from changing depth, and also the large-scale motion of target itself. Distortions due to the egocentric view, like perspective effects, can be implicitly learned using absolute coordinate trajectories while training. However, such an approach adds unnecessary complexity to the input distribution and hence is data inefficient. Instead, we propose to predict residuals from the first observed positions instead of forecasting absolute coordinates. In particular, we learn to predict the global stream from separately processed low level vision signals (monocular depth, camera egomotion). Furthermore, since these signals are machine generated and noisy, instead of using classical computer vision methods to employ these features like re-projecting the future positions with estimated camera pose matrices to cancel out egomotion effects, we propose a frame level encoder as depicted in Fig.3. The frame-level encoder is a feedforward neural network, which at sequence frame $\alpha$ takes as input observed coordinates ($\tilde{a}_\alpha$), estimated monocular depth $d_\alpha$ of joint $i$, and a flattened $3 \times 4$ transformation matrix $T_\alpha$ estimating the change in the 6D pose between the sequence and frame $\alpha$. The frame-level encoder architecture is motivated by the classical computer vision algorithm of projecting a point $(p_1, p_2, p_3)$ to a new point $(p'_1, p'_2, p'_3)$ under a transformation matrix $T$ and hence is shared across different time steps. The output of encoder is a 2 dimensional vector $\tilde{x}_\alpha$, which is then pooled across time steps to be fed into the sequence level encoder. The rest of the architecture is similar to local stream forecasting module but with $N_{\text{global}}$ layers of QRNN.

Loss Metric Both streams are trained with the $f_\ell$ loss between the predicted pose and the original pose (before completion) weighted by the confidence scores $c_{i_j=t+1}$ of OpenPose in the original detection.

3.6. Recombining the Streams

These future predictions are finally recombined to forecast the overall human locomotion $\{p_j\}_{j=t}^{t+f}$ as follows:

$$\hat{p}_j = \begin{cases} p_j^{\text{global}} & \text{for } i = \hat{i} \\ p_j^{\text{global}} + p_j^{\text{local}} & \text{for } i \neq \hat{i} \end{cases}$$

$\forall j \in \{t+1, \ldots, t+f\} \text{ and } \forall i \in \{1, \ldots, d\}$. This operation forms the inverse of the splitting procedure (Section 3.3).

4. Experiments

JAAD Dataset [38]: Joint Attention in Autonomous Driving is a dataset for modelling attention in the context of autonomous driving scenarios. The dataset contains 346 videos with 82032 frames, where the videos are the recorded with front-view camera under various scenes, weather, and lighting conditions. It is also one of the very few autonomous driving datasets that has temporally dense pedestrian bounding box annotations. For example, popular egocentric datasets like KITTI [39], Berkeley DeepDrive [40], and CityScapes [41] do not contain human bounding boxes in a dense per-frame fashion. Similarly, PedX [42] and Mapillary [43] has pedestrian bounding box annotations, but these are separate images and not videos. The videos in JAAD range from short 5 second clips to longer 15 second videos shot through wide angle camera mounted inside the car in centre of windshield. It contains 64 clips of resolution $1920 \times 1080$ pixels and rest 293 of resolution $1280 \times 720$ pixels. Employing the procedure mentioned in 3.1, a little under 1500 non-overlapping 30 frame length human locomotion examples are extracted which are then split into train and test sets while taking care not to have the same pedestrian’s trajectories in both groups.

4.1. Experimental Settings

We use a $d = 25$ pose keypoints as extracted from OpenPose. Empirically, the value of $c_\alpha$ is set to 0.25 for all our following experiments. The pose completion network is trained with the ADAM optimizer [44] with learning rate $10^{-3}$ and dropout 0.5 applied to the input layer. The local stream network is trained with ADAM too with learning rate of $10^{-4}$ and $N_{\text{local}} = 4$. The global stream network is trained on the residuals between consecutive frames with ADAM at learning rate 0.001 and $N_{\text{global}} = 2$. We
Table 1: Average KDE for baselines and our proposed method. Despite their success at modelling other sequence learning tasks, their direct application in predicting human locomotion in egocentric view fail due to the complicated interaction between the local and global structures.

<table>
<thead>
<tr>
<th>Method</th>
<th>Decomposition</th>
<th>KDE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Last Observed-Velocity</td>
<td>-</td>
<td>195.5</td>
</tr>
<tr>
<td>Constant-Velocity</td>
<td>-</td>
<td>48.9</td>
</tr>
<tr>
<td>Zero-Velocity</td>
<td>-</td>
<td>40.4</td>
</tr>
<tr>
<td>LSTM-ED [45]</td>
<td>✗</td>
<td>24.4</td>
</tr>
<tr>
<td>TCNN [46]</td>
<td>✗</td>
<td>31.8</td>
</tr>
<tr>
<td>GRU-ED [47]</td>
<td>✗</td>
<td>23.8</td>
</tr>
<tr>
<td>Structural RNN [3]</td>
<td>✗</td>
<td>31.7</td>
</tr>
<tr>
<td>Ours</td>
<td>✓</td>
<td>10.9</td>
</tr>
</tbody>
</table>

Table 2: Ablation Study on our method showing the relative importance of different components.

<table>
<thead>
<tr>
<th>Stream</th>
<th>Completion</th>
<th>Disentanglement</th>
<th>KDE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Global</td>
<td>✗</td>
<td>-</td>
<td>6.8</td>
</tr>
<tr>
<td>Global</td>
<td>✓</td>
<td>-</td>
<td>5.6</td>
</tr>
<tr>
<td>Local</td>
<td>✗</td>
<td>-</td>
<td>9.5</td>
</tr>
<tr>
<td>Local</td>
<td>✓</td>
<td>-</td>
<td>6.3</td>
</tr>
<tr>
<td>Both</td>
<td>✗</td>
<td>✗</td>
<td>22.1</td>
</tr>
<tr>
<td>Both</td>
<td>✓</td>
<td>✗</td>
<td>18.2</td>
</tr>
<tr>
<td>Both</td>
<td>✓</td>
<td>✓</td>
<td>15.4</td>
</tr>
<tr>
<td>Both</td>
<td>✓</td>
<td>✓</td>
<td>10.9</td>
</tr>
</tbody>
</table>

use Keypoint Displacement Error (KDE) defined as the average \( l_2 \) distance between the predicted and the ground truth keypoints to measure model’s performance, defined as 

\[
\text{KDE} = \frac{1}{n} \sum_{i=1}^{n} \| \hat{p}_i - p_i \|_2 /	f
\]

Naturally, Mean KDE measures the Keypoint displacement error per human joint. For global stream, we use the same formulation but with \( d = 1 \) in contrast to \( d = 25 \) for local stream.

4.2. Pose Completion Results

As described in Section 3.2, we use a symmetrical over complete auto-encoder for pre-processing noisy detections. Our model comprises of a compressing module that encodes the \( 2d = 50 \) dimensional pose into \( d_{ae} = 10 \) dimensions, in three equally proportionate non linear layers. Symmetrically, the decoder decompreses the latent pose to the original 50 dimensional space. A dropout of 0.5 dropout is applied to input layer and training is done only on poses with all keypoints having confidence more than \( c_i \). This is trained as mentioned in Section 4.1 to yield 9.6 unweighed Mean KDE on high confidence pose detections on JAAD dataset. In practice, this means that on average the module is able to predict a missing joint within an \( |c| \) ball of radius 10 pixels centered at the joint’s true location in a frame. Considering that the average resolution of frames in JAAD is 1408 × 790 pixels and the pose completion network is not trained to regress on residuals but instead predicts the absolute coordinates in original frame resolution, this means that the error is about 1% – 1.5% of the frame dimensions and quite tolerable for our forecasting task.

4.3. Stream Forecasting

We compare our proposed method with several other baselines trained from scratch with the same input (i.e. uncompleted pose estimates from Openpose) and output data. These methods are briefly described below:

Encoder Decoder Architectures: We train the popular Encoder Decoder architecture for sequence-to-sequence learning with several backbones like LSTM [48, 45], Gated Recurrent Units [47] and recently introduced Quasi RNN [34]. All the models are trained from scratch on the same input data and confidence values but without any explicit spatial encoding. Also, Quasi RNN is the most parallelizable and train much faster than other recurrent counterparts. Hence, it forms the backbone of our method.

Temporal Convolutional Neural Networks: Lea et al. [46] recently proposed Temporal Convolutional Networks which have also shown promising performance on a variety of sequence-to-sequence tasks. We found that TCNN trained with the same input signals, are the fastest to train and perform the best among deep learning methods without decomposition but are still much worse than the simple zero-velocity baseline. While TCNN perform well before decomposition, the future time horizon length \( t_f \) is constrained by the input sequence length \( t_p \) and does not allow flexible predictions as with other recurrent architectures.

Strutural RNN[3]: SRNN has also achieved state of the art performance on several sequence prediction tasks including pose forecasting. However, we observe that its capacity is much larger than other sequence-to-sequence models because of larger number of trainable parameters that does not generalize well to our task (limited data regime).

Uniform Velocity: We also compare our method with simple yet powerful baselines [9, 49] based on zero, constant, and last observed velocity assumptions. In Zero Velocity, the future is forecasted assuming no movement from the last observed position. Similarly, in constant-velocity/last-observed-velocity, the forecasted velocity is assumed to be identically equal to the average observed velocity/last-
observed-velocity and this assumption is rolled out in time to predict future positions.

**Results Discussion:** Table 1 reports the Average KDE achieved by these approaches for predicting the pedestrian locomotion for \( t_f = 15 \) frames into the future given the previous \( t_p = 15 \) frames. Our method performs better than several widely used sequence-to-sequence models and also state-of-the-art pose forecasting models like Structural RNN [3]. Figure 8 shows some qualitative results from our method for pedestrian locomotion prediction.

**Ablation:** We perform the ablation study on our proposed method to investigate the importance of each component. As Table 2 indicates, both disentanglement and completion are quite central to our method. Also, the significant increase in performance conferred by decomposing the streams confirms our hypothesis that disentangling is essential to locomotion forecasting and reduces the complication of overlapping simultaneous movements. Also noteworthy is the decrease in performance due decomposing without pose completion. This backs up our proposition that completion is essential to disentangling and data is subject to severe method-induced noise without pose completion.

**Time Horizon Experiments:** We extensively test our novel egocentric global stream (trajectory) forecasting module for various time horizons against the best performing baseline from Table 2. Table 3 shows these trends for time horizons from 165 ms to 1 second. We observe that as the horizon increases and the trajectories become longer, our method outperforms the baseline by an increasing and larger margin. This is because our model accounts for the camera motion, which can accumulate over longer time horizons leading to a large error in egocentric motion agnostic methods.

**5. Discussion**

Taking into consideration the results discussed in the last section, we would like to revisit our original hypothesis and discuss final remarks that also highlight the new findings.

**Disentanglement:** We hypothesize that disentangling the human dynamics into overlapping but simpler streams helps overall prediction. This is corroborated with the results and ablation experiments discussed before. We further suspect that disentanglement performed in our proposed manner is useful in predicting motion dynamics of any object, cognizant or inanimate. The overall prediction difficulty of the spatially disentangled parts is lesser than that of the whole.

**Finetuning Pose Detection:** The state of the art pose detection network is pretrained on general purpose pose datasets and is not adapted to this setting. The pose completion network finetunes these detection conditioning on the information that the majority of human poses for pedestrian locomotion prediction are upright, as is the case when pedestrian is standing, walking or running. Thus, it implicitly also works as an adaptation step to use pose regularities specific to the setting of locomotion prediction. Additionally, it fills in the missing data because as discussed in Section 3.2, it is the existence of the keypoints and not so much their accuracy that helps the decomposition.

**6. Conclusion**

We propose a human locomotion forecasting pipeline that disentangles the complex pedestrian locomotion into simpler tasks of global and local motion prediction. We also present a pose completion module and an Encoder-Recurrent-Decoder style pose prediction network. The pose completion module disentangles the overlapping motion and fills in the missing detected joints. Furthermore, the pose prediction network forecasts these streams separately which are finally merged for the final prediction. We demonstrate that our method improves the overall performance & achieves state-of-the-art short term human locomotion forecasting in egocentric view.

**Acknowledgments** Toyota Research Institute (TRI) provided funds to assist the authors with their research. This article solely reflects the opinions and conclusions of its authors and not TRI or any other Toyota entities.
References


