A Dual Perspective of Human Motion Analysis -3D Pose Estimation and 2D Trajectory Prediction

Mayssa Zaier, Hazem Wannous, Hassen Drira, Jacques Boonaert

A Dual Perspective of Human Motion Analysis - 3D Pose Estimation and 2D Trajectory Prediction

Mayssa Zaier[1], Hazem Wannous[1], Hassen Drira[2], Jacques Boonaert[1]
[1] IMT Nord Europe, University of Lille, CNRS, UMR 9189 - CRIStAL, F-59000 Lille, France *
[2] ICube CNRS UMR 7357, University of Strasbourg, Strasbourg, France

Abstract

Anticipating human motion based on given sequences is a challenging and crucial task in computer vision and machine learning, enabling machines to understand human behaviors effectively. Precise prediction of human pose and motion trajectory holds great significance for various applications, including autonomous driving, robotics, and virtual reality. This paper presents a novel approach to address the interconnected tasks of estimating human motion, represented as 3D poses or 2D trajectories, and predicting future motions using 2D images and human pose/position sequences jointly. We propose an encoder-decoder architecture that leverages Transformer networks with a self-attention mechanism, utilizing visual context features, combined with an LSTM to model human motion kinematics. Our approach demonstrates consistent and remarkable improvements over existing methods, both quantitatively and qualitatively. Extensive experiments conducted on diverse public datasets, such as GTA-IM and PROX for 3D human pose estimation, and ETH and UCY combined datasets for 2D trajectory prediction, showcase that our method substantially reduces prediction errors compared to the current state-of-the-art methods.

1. Introduction

Predicting human motion accurately plays a pivotal role in various fields, such as computer vision, autonomous driving [13, 9], intelligent robots [19], human-robot collaboration [32, 28], virtual reality [38], and can occur in different environments, such as streets, airports and sports arenas. The task involves two main aspects: 3D poses estimation and 2D trajectories prediction. These tasks are interrelated and essential for understanding human movements and enabling machines to interact intelligently with humans. In the context of 3D poses estimation, the goal is to focus on reconstructing the three-dimensional positions and orientations of human joints from a given sequence of data, such as images or videos. On the other hand, 2D trajectories prediction aims to forecast future paths or movements of humans in a two-dimensional space based on their past behaviors.

Human motion can be interpreted and represented in various ways, including kinematics, joints graph, video frames, and moving from one point to another as a mass point, each reflecting distinct understandings of human motion and leading to diverse predictions of it. Three primary types of motion trajectory prediction are discussed in the literature: 2D motion trajectory prediction [2, 14, 41], video prediction [55, 45], and 3D poses sequence prediction [10, 33, 31, 56, 5, 34, 11].

To accurately predict human behavior, it is crucial to consider not only the movements of individuals but also the physical surroundings, including obstacles, trees, buildings, and any other scene elements. The combination of these factors makes forecasting human behavior in crowded settings a formidable undertaking. For instance, the complex nature of human intentions which act as internal stimuli that can influence behavior. Additionally, external factors in the physical world can also actively or passively impact human motion. Previous approaches have explored various methods for integrating scene information, including static scene information extraction [22, 41, 43] and dynamic spatial and temporal context utilization [44, 7]. However, these models often suffer from limitations concerning memory and computational complexity.

Remarkable progress has been made in predicting human pose sequences with the emergence of deep learning. While predicting human motion, many methods tend to overlook the fact that humans move around in three-dimensional spaces, which leads to ignoring the crucial interactions between humans and their surroundings. Zhang et al. [56], detect human bounding boxes across various time frames and use the changing appearance of the human form to generate their predictive signal, but they do not incorporate the background image. Mao et al. [31] use the skeleton's spa-

*Corresponding Author: Hazem Wannous: hazem.wannous@imt-nord-europe.fr

tial configuration explicitly in the deep architecture itself, which provides valuable information that leads to more accurate predictions. The joints of the skeleton are related to each other in terms of their angles and distances, and introducing this information to the deep model constrains its output from producing random points far from the ground truth. More other methods for estimating 3D human motion such as *TR* [47], *VP* [36], and *LTD* [31] have also been evaluated for this task. Although significant advancements have been achieved, these approaches have certain limitations since they exclusively rely on kinematic data and neglect to consider information from alternative sources. Nevertheless, incorporating the scene context by including an embedding of a 2D scene image, a 3D scene, or a specific object as an additional input to the model, is considered to be an advantageous approach. This issue has recently addressed by introducing a novel long-term trajectory prediction problem that focuses on deriving 3D poses from 2D poses [5] . To accomplish their task, they proposed a three-stage deep framework named *GPP-Net*, which utilizes a single scene image and 2D pose histories, the framework sequentially predicts 3D human pose sequences along each path. Meanwhile, Zhe et al. [5] utilize the visual signal of the observed sequence to incorporate information about the environment's objects and geometry, which helps to eliminate ambiguity in some prediction scenarios. Recently, Mohamed et al. [34] formulate the problem as an end-to-end spatio-temporal graph while learning a suitable graph adjacency kernel to capture the interaction among the joints of the skeleton by exploiting their spatial configuration. Additionally, the authors fuse the vision signal into the model. While previous methods have shown remarkable progress in predicting human pose sequences using deep learning, they often overlook the fact that humans move in three-dimensional spaces, leading to limited consideration of interactions with their surroundings.

In this article, we propose a new approach which: first, addresses the two tasks: 3D poses estimation and 2D trajectories prediction. Second, it leverages the scene context to enhance both 2D trajectory and 3D human pose predictions. We argue that the necessary information for human motion and interaction, along with environmental constraints, can be effectively extracted from the video itself, rather than individual frames. This is because kinematic trajectories are derived from videos and cannot contain more information than the video. As a result, we adopt a self-attention mechanism to capture relevant features from the video context and investigate its interaction with human motion kinematics in spatial and temporal dimensions. The dynamic scene context is introduced in many existing trajectory prediction methods. However, complex architectures like 3DCNN [44, 59] are often used to describe the context. Having a simple architecture using a 2D CNN combined with a transformer, it allows to: (1) capture the dynamic context of the scene by taking into account the observed poses / trajectories and the video streams, (2) better understanding of the scene taking advantage of static elements and moving objects, (3) demonstrate improved performance in scenarios with rapid motion changes, predicting sharp turns and avoiding moving obstacles.

2. Related Work

This section provides an overview of the existing research in the field of estimating human motion, which encompasses two primary aspects: 3D poses estimation and 2D trajectories prediction.

2D trajectory prediction. Early methods for trajectory prediction [16, 1, 51, 40] used physical forces and Hidden Markov models to model social-scene interactions. However, data-driven approaches using neural networks that capture multi-modal interactions between the scene and agents have become dominant. Recent works focus on RNN-based architectures [23, 3, 58, 22] to encode interactions between humans, but RNNs struggle with spatio-temporal interactions. Graph representations [29, 53, 20, 35] capture social interactions, but some methods lack environmental context understanding. Other approaches incorporate human-environment interaction models [42, 18, 43] with visual features [50, 8] and dynamic 3D scene information [44]. Transformer-based methods [53, 53, 46] are seen as potentially more suitable for trajectory forecasting, but solely relying on past trajectories may fail to detect unpredictable sharp turns, suggesting the need to incorporate additional information, such as environmental configuration. Our work focuses on predicting individual pedestrian motion, achieving the best performance on challenging benchmarks. Context-aware trajectory prediction models aim to incorporate physical scene information, such as crosswalks and roads. Previous methods have proposed extracting and integrating static scene information [22, 42, 43], while recent models have explored dynamic spatial and temporal context [7, 44]. However, these models suffer from limitations related to memory and computational complexity. For instance, employing 3D-CNNs can be computationally expensive and require large amounts of memory due to processing volumetric data [44].

3D pose estimation. The challenge of predicting 3D human motion involves estimating both the pose and skeleton of the target subject, which is more complex than 2D pose estimation due to a larger 3D pose space and increased ambiguities. Various approaches have been proposed to address this challenge. Some methods directly infer 3D coordinates from 2D RGB images through joint points regression and

point detection tasks [26], while others explore intermediate 2D pose approximations [6] instead of directly inferring the 3D pose from an image. However, these approaches may not fully consider the image context, leading to inconsistencies in predicted human motion with the scene. To address this limitation, some works utilize the static 3D layout of the scene to improve pose estimation from monocular images by considering environmental constraints [15]. Additionally, recent studies define tasks for predicting 3D human motion over extended periods while accounting for the surrounding environment and employ scene context for achieving goal-oriented motion prediction [5]. CNN graphs and deep graph networks have also been effective in classifying human skeleton actions and predicting 3D human motion based on past movements [52, 17, 25, 60].

State-of-the-art works address these two applications of human motion analysis and estimation separately. We propose in this work an approach which, on the one hand tackles the two problems, and on the other hand addresses the limitations of existing methods, in particular those relating to the modality of motion and context description.

3. Approach

3.1. Problem Formulation

The problem addressed in this study is the prediction of human motion using two main inputs: data and video context. Our model's estimation of movement depends on the specific target representation, which can be either 3D pose prediction or 2D trajectory prediction. The data can be in the form of 2D skeleton sequences or coordinate position sequences, based on the target representation. The challenge lies in effectively integrating the 2D pose information and video context to achieve accurate predictions for each task. The first input, denoted as $u_t^{(p)}$, defines the 2D position of human p at frame t following this formula: $u_t^{(p)} = [(x_t^{(p)}, y_t^{(p)}), ..., (x_t^{(p)}, y_t^{(p)})]$ Given that we observe trajectories and scenes from frame 1 to t_{obs}, the goal is to predict the paths of movements in frames ranging from the next frame (t_{obs}+1) to the final frame (t_f). For a pedestrian p, the sequence of observed and future positions are respectively denoted by $\tau_{obs}^{(p)} = (u_1^{(p)}, ..., u_{t_{obs}}^{(p)})$ and $\tau_f^{(p)} = (u_{t_{obs}+1}^{(p)}, \cdots, u_{t_f}^{(p)})$. The sequence of observed frames is denoted by $\nu_o = (\mathrm{I}_1, ..., \mathrm{I}_{\mathrm{t_{obs}}})$, which consists of video frames of the scene. Building such a model presents challenges due to the complexity and diversity of human motions and the scarcity of accurate data. To address these issues, we employ an LSTM-based 2D pose / position encoder, a transformer-based visual encoder, and a decoder for predicting either the 2D trajectory or the 3D human poses.

3.2. Architecture

Our architecture is expressed an encoder-decoder model for both 2D trajectory prediction and 3D pose prediction (Figure 1). The encoder part is composed of a *landmark encoder*-based LSTM responsible of encoding the motion from 2D skeleton joints matrices, and a *context encoder* which analyse the image through a block of ResNet followed by a Transformer encoder. Finally, a *decoder* take as input the concatenation of the two previous encoders, and predicts both 2D trajectory and/or 3D skeleton joints representation of human motion.

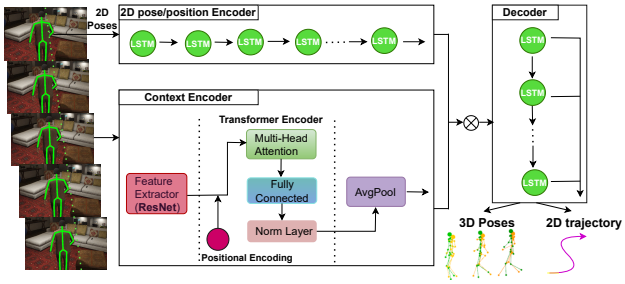

Figure 1: Our proposed approach.

2D pose/position encoder. This encoder uses the observed trajectory of the target human as input, which includes the starting position and the relative displacements between consecutive frames. This format was chosen to better capture the similarities between similar trajectories that start at different points. The input vectors are transformed using a fully-connected network, which power one-layer LSTM, to model the dependencies between different coordinates of the observed trajectory. The outputs from the LSTM units capture the latent kinematic trajectory (2D human coordinates) or landmarks (2D human poses).

Context encoder. This encoder is used to extract the visual information of each human. In fact, the observed video contains information about the physical constraints of all humans present in the scene. This encoder is composed of a ResNet followed by a transformer network [47]. The ResNet encodes spatial information from the video frames while the Transformer encode the temporal relationship while focusing on regions relevant to each human using the multi-head attention mechanism. It is composed of a Multi-Head attention mechanism, a fully connected layer and a normalization layer. $MultiAtt(x) = (Att(x) \oplus ... \oplus Att(x))W^O$, where $Att(x) = \sigma(\frac{Q_i K_i}{d_k})V_i$ where $Q_i = xW_i^Q$, $K_i = xW_i^K$, $V_i = xW_i^V$ are independent linear projections of x into a d-feature space, d is a scaling factor.

Some missing information present here as no recurrence neither convolution in the encoder attention mechanism. To overcome that, a *Positional Encoding* information is added with the same shape as $encoder\_input$ to inject information on the relative or absolute position of the frame in the set of videos. Thus, $T(x) = AvgPool(ET(x + PE))$, where ET is the encoder transformer mechanism function which can be defined as: $ET(x) = Norm(x + FC(MultiAtt(x)))$, PE is the aforementioned Positional Encoding, FC is a fully connected layer, and $Norm$ is a normalization layer.

Decoder. The decoder takes as input the combination of the two resulted previous vectors. After, it is passed to a maxpool layer followed by a linear layer to understand the relation between these. Then, the result will be feed to the LSTM as hidden vector, followed by a MLP. We aim to predict human motion using two main inputs: 2D human poses / position and video context. The desired output depends on the specific target problem. In the case of 3D pose prediction, our goal is to forecast the 3D poses of the human subjects. Conversely, for 2D trajectory prediction, the objective is to predict the trajectory of the target person in the 2D space. The challenge lies in effectively leveraging the 2D pose / position information and video context to achieve accurate and reliable predictions for the respective tasks.

Loss function. The loss function is defined depending on the form of data to be predicted. For 3D pose estimation, a sub-sequence of 2D body pose (21 2D joints) is observed in order to predict a sub-sequence of 3D body poses (21 3D joints). As in [34], the idea is to use the loss functions to fix the distance and the angle between two joints, in addition to the typical *mean square error*. For the angle, we use the *cosine similarity* between the joints. For the length between two joints, we use the *l2-norm*. In other words, the loss function have 3 components: MSE between the joints and the prediction, a consistency loss that make sure the intra-distance between joints are consistent and that the angles between the joints are consistent. Also known as, we don't want to end up with weird angles and/or weird skeleton bone lengths. The loss function can be defined as: $\mathcal{L}_{model} = \lambda_1 \mathcal{L}_{cos} + \lambda_2 \mathcal{L}_{norm} + \mathcal{L}_{mse}$, where λ_1 and λ_2 are parameters to tune, so we set them to 0.01.

For the 2D trajectory prediction, our loss function consists of two components - the mean-squared loss and a regularization term called $\mathcal{L}reg$, which regulates the smoothness of future trajectories, as prior work [44]. In training our network, we use the following loss function: $\mathcal{L}model = \mathcal{L}mse + \lambda \mathcal{L}reg$, where λ is a regularization parameter. We kept the value of λ fixed at 0.5 in our experiments to avoid restricting the model's ability to capture sudden changes in the target pedestrians' trajectory. $\mathcal{L}mse$ is calculated as the average of the squared differences between predicted and observed values, while $\mathcal{L}reg$ is calculated as the sum of Euclidean distances between each step of the predicted trajectory and a line fitted to the observed trajectory. In our experiments, we sample 20 future trajectories and select the top 5 trajectories that are closest to the ground-truth to calculate $\mathcal{L}_{mse}$. Through empirical observation, we have found that this approach enables our network to converge more quickly while producing more accurate predictions.

4. Experiments

4.1. Datasets

Two datasets are used to evaluate our approach for 3D pose prediction: GTA Indoor Motion (GTA-IM) [5] and Proximal Relationships with Object eXclusion (PROX) [15].

GTA-IM. GTA-IM is collected in an indoor environments emphasizing human-scene interactions. It contains HD RGB-D image sequences of 3D human pose and camera pose annotations, and large diversity in human appearances, indoor environments, camera views, and human activities. Following the experimental protocol proposed by [5, 34], we spit the data into two subsets, with 80% (8 scenes) for training and 20% (2 scenes) for evaluation purposes.

PROX. PROX is collected with the Kinect-One sensor, and contains 12 different 3D scenes, including 3d human body skeleton, and RGB sequences of 20 subjects moving in and interacting with the scenes. Following [5, 34], we split this dataset into 52 training sequences for training and 8 sequences for test.

Regarding 2D trajectory prediction, our approach was evaluated on well-established public human-trajectory datasets, namely ETH [37] and UCY [23] datasets, which are widely-used benchmarks for pedestrian motion prediction.

ETH and UCY combined. These datasets were acquired from surveillance videos of pedestrians walking on sidewalks and annotated with location coordinates. The datasets contain real-world pedestrian trajectories in top-view coordinates expressed in meters, with rich human-human and human-object interaction scenarios. The acquisition was done using a fixed camera on 5 different scenarios captured at 2.5 Hz, and an image is annotated with pedestrian positions every 0.4 seconds. ETH/UCY consists of five different scenes. The ETH dataset consists of two scenes (*ETH* and *Hotel*) taken from a bird's eye view, with hundreds of pedestrian trajectories engaged in walking activities. The UCY dataset provides three scenes (*Zara1*, *Zara2*, and *Univ*) taken from a bird's eye view with standing/walking activities.

4.2. Metrics

For evaluation, we choose several appropriate metrics to measure the error of the 3D path and pose human motion prediction. The main metric employed is *Mean Per Joint Position Error* (*MPJPE*) or *3D pose error*. It is calculated for a frame t and a skeleton consisting of J joints by the following formula (1), where P the prediction, and Y the ground truth.

$$MPJPE(t) = \frac{1}{J}\sum_{j=1}^{J} ||P_j^t - Y_j^t||_2 \quad (1)$$

For the *3D path error*, we use the same formula but only for a fixed joint that is the center of the skeleton joints. In fact, these two metrics are used to compute the evaluation of the error with the time step, i.e. after the quarter, half, one half, and full 3D joint estimation. Similar to existing works, other metrics are also used for comparison, *Average Displacement Error (ADE)*, *Final Displacement Error (FDE)*, and *Stability Error (STB)*. ADE is used to measure the overall performance of the pose and the path on the entire predicted sequence. On the other hand, FDE is an indicator of the final time step prediction for path and pose error. Finally, STB is used to detect the variation of the path and pose MPJPE error over different time steps. This metric is mainly used to check the divergence over the prediction time steps, not the accuracy of the model.

For the 2D trajectory prediction, similar to existing works [3, 14, 4, 42, 43, 53, 22, 18, 20, 59, 24], our method is evaluated using two widely used metrics in the field, namely the *(ADE)* and the *(FDE)*. ADE is defined as the average L2 distance (in meters) between the actual trajectory and the predicted trajectory at each time step of the trajectory from $T_{\text{obs}+1}$ to T_{pred} on average over all pedestrians. FDE is defined as the Euclidean distance (in meters) between the ground truth (actual position) and the prediction (predicted position) at the last time step of the prediction T_{pred}, averaged over all pedestrians. Formally, $ADE = \frac{\sum_{i=1}^{n}\sum_{t=T_{\text{obs}+1}}^{T_{\text{pred}}} ||\hat{Y}_t^i - Y_t^i||}{n*T}$, and $FDE = \frac{\sum_{i=1}^{n} ||\hat{Y}_{T_{\text{pred}}}^i - Y_{T_{\text{pred}}}^i||}{n}$ Where n represents the number of pedestrians, $\hat{Y}_t^i$ are the predicted coordinates for pedestrian i at time t, Y_t^i are the real future positions, and $||$ is the Euclidean distance. T_{pred} is the final predicted timestep. T is the prediction horizon.

4.3. Implementation details

In GTA-IM, the initial learning rate is set at 0.01, while in the PROX dataset, it is 0.03. The model is trained for 400 epochs, with the learning rate decreasing by 0.2 every 200 epochs. A batch size of 128 is utilized. We use 1 second of observation (5 frames) and 2 seconds for predictions (10 frames) following the settings of prior works for comparison. Table 1 shows the different model parameters chosen in the experiments. The model is trained for 400 epochs, a batch size of 128 and a learning rate of 0.01. Table 1 shows the different model parameters chosen in the experiments.

Table 1: Model details.

Input	Image size	Output	Hidden size	Embed Size	Visual out	Landmark out
42	90*160	63	256	64	256	256

For the 2D trajectory prediction, we trained the entire network with a batch size of 40 for 400 epochs, using stochastic gradient descent (SGD) optimizer with a learning rate scheduler and two mean squared error (MSE) loss functions. The learning rate is adjusted every 40 steps with an initial learning rate of 0.01 and the maximum gradient value is clipped to 1 to prevent gradient explosion. We adopted the teacher force strategy and used our proposed loss function with a value of $\lambda = 0.5$ to achieve faster convergence as prior work [44]. During training, we generated 20 output trajectory samples and used the 5 samples with the lowest loss value to train the model. We use 3.2 second of observation (8 frames) and 4.8 seconds for predictions (12 frames) following the settings of prior works for comparison. Our model is implemented in Pytorch on Ubuntu with an NVIDIA TITAN RTX GPU and 24 GB RAM memory.

4.4. Results on 3D pose prediction

Baseline. There exists very few prior works that predicts 3D human pose with global movement using 2D pose sequence as input. Thus, as a reference, we use the results from methods reported in [5]. (1) TR [47]: utilizes transformer architecture to predict the 3D human pose directly from 2D input data, treating the entire problem as a single-stage sequence to sequence task. (2) TR [47]+VP [36]: a combination of the transformer and a 2D-to-3D human pose estimation method. It is constructed by first predicting the future 2D human pose using [47] from inputs and then lifting the predicted pose into 3D using [36]. (3) VP [36]+LTD[31]: a combination between a 2D-to-3D human pose estimation method [36], and a 3D human pose estimation based on graph [31]. None of these three baselines take into account the context of the scene. (4) GPP-Net [5]: is composed of three stages and incorporates various concepts, including VAEs, as well as customized stages for predicting both paths and poses. (5) SG [34]: concerns the prediction of the 3D human pose prediction based on *Skeleton Graph*, using graph CNNs with self learning adjacency matrix and formulating the problem as a spatio-temporal graph. Note that there are two variants of SG : with/without the context (designed as +/-C), in order to observe the impact of context, especially on the PROX dataset.

Quantitative Analysis. Tables 2-5 show the result obtained from both GTA-IM and PROX datasets, employing

the provided metrics. Specifically, Table 2 and 3 display the assessment outcomes for the forecasted path and 3D pose on the GTA-IM and PROX datasets respectively. Table 4 and 5 illustrate the FDE, ADE, and STB outcomes obtained through various techniques on the GTA-IM and PROX datasets respectively.

Table 2: Evaluation results in GTA-IM dataset (errors are in mm).

	3D path error				3D pose error			
Time step (s)	0.5	1	1.5	2	0.5	1	1.5	2
TR[47]	277	352	457	603	291	374	489	641
TR[47]+VP[36]	157	240	358	494	174	267	388	526
VP[36]+LTD[31]	124	194	276	367	121	180	249	330
Ours (-C)	222	221	**222**	**226**	299	291	291	**295**
GPP-Net (+C) [5]	**104**	**163**	219	297	**91**	**158**	237	328
SG (+C) [34]	154	**163**	172	186	198	209	217	230
Ours (+C)	167	168	**169**	**167**	208	211	**212**	**213**

Our approach demonstrates better performance than previous methods across several metrics. At first sight, the STB metric of 1 and 4 for GTA-IM and PROX dataset respectively shows that our model is stable for long term prediction. This result could be due to the trick of predicting displacement instead of just the coordinates, in order to better capture the similarities between close trajectories starting at different positions. For the 3D path error and the 3D pose error in *GTA-IM*, only for long term, our approach starts to outperform the other methods. FDE is very closed to ADE in our case, which is due to the stability of the prediction over time. If we want to choose short-term prediction, *GPP-Net* [5] is the best choice, and for a long term prediction, our model is the best. In the other hand, *PROX* dataset was primarily recorded in a vacant laboratory setting, resulting in limited diversity in terms of visual characteristics. As a result, it lacks the same level of richness as the GTA-IM dataset, which has implications for the diversity of backgrounds and camera angles. Due to these limitations, the visual signal is inherently less informative, which can result in increased ambiguity in predictions. These limitations of the PROX dataset will impact our results as seen it in the table 3 and 5 (without context, our results are better than those with context). A similar discussion was raised in [5, 34]. For the PROX dataset using the vision signal resulted in a divergence of the results. As seen in table 4, the errors in short-term prediction are even larger than in long-term prediction. Further investigation is needed to understand the reasons behind this behavior. Despite this, we are mostly better than the other methods and these results on the PROX dataset demonstrate the resilience of our approach in scenarios involving imprecise 2D pose estimations. Also, we can note that the 3D pose error was the lowest compared

Table 3: Evaluation results in PROX dataset. Error are in mm. C: to indicate, using context or not.

	3D path error				3D pose error			
Time step (s)	0.5	1	1.5	2	0.5	1	1.5	2
TR[47]	487	583	682	783	512	603	698	801
TR[47]+VP[36]	262	358	461	548	297	398	502	590
VP[36]+LTD[31]	**194**	**263**	332	394	**216**	**274**	335	394
Ours (-C)	306	303	302	302	291	286	**285**	**286**
GPP-Net (+C)[5]	**189**	**245**	317	389	**190**	**264**	335	406
SG (+C) [34]	353	353	355	360	358	362	365	370
Ours (+C)	350	346	**342**	**343**	362	351	350	**352**

Table 4: Results of FDE, ADE and STB in GTA-IM dataset are reported in mm. C: to indicate, using context or not.

Time step (s)	FDE	ADE	STB
TR[47]	622	436	147
TR[47]+VP[36]	510	326	150
VP[36]+LTD [31]	349	230	98
Ours (-C)	**260**	**258**	**2**
GPP-Net (+C) [5]	313	200	93
SG (+C) [34]	208	192	11
Ours (+C)	**190**	**189**	**1**

Table 5: Results of FDE, ADE and STB in PROX dataset are reported in mm. C: to indicate, using context or not.

Time step (s)	FDE	ADE	STB
TR[47]	792	644	126
TR[47]+VP[36]	569	427	126
VP[36]+LTD[31]	394	300	82
Ours(-C)	**294**	**295**	**2**
GPP-Net (+C) [5]	398	292	90
SG (+C) [34]	365	359	3
Ours(+C)	**347**	**349**	**4**

to prior works, suggesting our model (-C) effectiveness in predicting 3D poses, even without the context. However, it had the highest error for the 3D path indicating a trade-off between the objectives of path and poses, which may be influenced by the lack of visual signal. Additionally, as seen in table 2 and 4, when the vision signal was used on the GTA-IM dataset, there was a significant improvement in both path and pose prediction performance mainly in long term. However, when the same signal was used on the PROX dataset, the results diverged. Additionally, even without context, the results for GTA were superior, emphasizing the importance of context in the prediction as we explained in the introduction.

Qualitative Analysis. Figure 2 visualizes the predicted 3D

human poses, presenting predictions for 10 human poses in two motion scenarios: walking and sitting. The joint estimation for walking sequence Figure 2-(top) seems very close to the ground truth and even the model becomes better with time steps, which is a concretization of the result obtaining in the Table 2, when the error decreases with time steps. In general, this prediction seems to be good, with the exception of the two last predicted joints, where the feet cross each other. For the sitting sequence in Figure 2-(down), we can observe a total lack of prediction, especially for the feet. This error may be due to the lack of training data on the sitting motion. This result is obtained despite that during training we use the consistency loss, which tunes the angle between two consecutive joints. Here for instance between two successive feet joint, the angle is tuned, but the problem comes from the angle between the feet and the back joint. For the evaluation, even if the prediction is not the best, the metric used will only compute the distance between the prediction and the target joins, so no loss related to angle is evaluated.

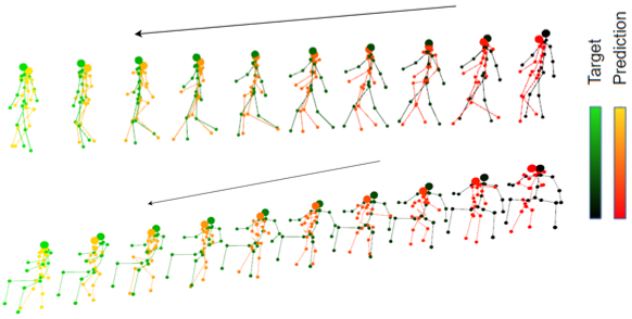

Figure 2: Predicted 3D human pose compared to the ground truth. The arrow indicates the time direction. (top) Example of walking motion prediction. (down) Example of sitting motion prediction.

4.5. Results on 2D trajectory prediction

To address the problem of pedestrian trajectory forecasting (2D), we used the same encoder-decoder architecture. It is bimodal in the way that it is composed of two parallels encoder branches, see **figure 1**.

Baseline. We compare the effectiveness of our proposed approach with state-of-the-art models by conducting a comparative analysis. We evaluate our approach against six *deterministic baselines*, which are linear regression, LSTM, Social-LSTM [3], Social ATTN [48], TrafficPredict [29], and SR-LSTM [58]. For evaluation, we generate 20 predictions for each observed trajectory and select the prediction closest to the ground truth. This evaluation technique enables us to examine the multi-modality and diversity of the predictions. We also compare our approach against various *generative baselines*, including Social-GAN [14], DESIRE [22], FSGAN [21], SoPhie [42], Trajectron [18], MATF [59], NEXT [27], Social-BiGAT [20], Social-STGCNN [35], Social ways [4], and PECNet [30], M2P3 [39], SGN LSTM [57], Trajectron++ [43], Introvert [44], GroupNet [49], Transformer-TF [12], STAR [53], and AgentFormer [54]. These models use various approaches such as LSTM, GAN, spatio-temporal graph convolutional neural networks, and transformers to predict human trajectories.

Evaluation method. For benchmarking purposes, we follow a similar evaluation method as previous studies (See Table 6). When evaluating trajectory forecasting models, the chosen *time horizon* is crucial due to varying object speeds. The choice of time horizon should depend on the class of objects being considered. To ensure fairness in comparisons, we observe each training trajectory for 8 times-steps (3.2 seconds) and then evaluate performance oven the next 12 time-steps (4.8 seconds). To fully utilize the datasets during training, we adopt a *leave-one-out* evaluation strategy, common in prior research. We train our model on four sets of data and evaluate it on the remaining set. We repeat this process for all the 5 sets.

Quantitative Analysis. The main results, shown in Table 6, highlight the superior performance of *ours** over state-of-the-art (SOTA) deterministic models in terms of overall performance, with ADE/FDE scores of 0.23/0.37. Moreover, our stochastic model exhibits a significant performance improvement over all SOTA models by a large margin. Our primary focus lies in the comparison of these stochastic models. The results are presented against state-of-the-art approaches as mentioned above, using the *best-of-20 protocol*, which involves sampling 20 possible future trajectories and selecting the one with the best test performance.
Our proposed method achieves outstanding performance, ranking the first among state-of-the-art methods. In particular, on the FDE metric, our method significantly outperforms existing algorithms on 4 out of 5 datasets, achieving the best average error of 0.33. On the ADE metric, the proposed method outperforms existing algorithms on 3 out of 5 datasets and achieves an average ADE error of 0.19 across all 5 datasets. The University dataset has higher displacement errors compared to other datasets, making it challenging to predict future trajectories accurately. Our method remains comparable to other existing approaches but outperforms all the dense interaction-based methods like *S-GAN*, *Sophie*, *S-BiGAT*, *S-STGCNN*, and *Social Ways*. The *Hotel* dataset has many pedestrians waiting for trains, resulting in limited motion. Therefore, most methods, including ours, achieve relatively small displacement errors by predicting small motions accurately. Our proposed method achieves the lowest FDE (0.10) and ADE (0.15) errors on this dataset. The *ETH* dataset often produces larger dis-

Method	Performance ADE/FDE ↓ (m)					
	Univ	Zara1	Zara2	Hotel	ETH	Avg
Linear*	0.82/1.59	0.62/1.21	0.77/1.48	0.39/0.72	1.33/2.94	0.79/1.59
LSTM*	0.61/1.31	0.41/0.88	0.52/1.11	0.86/1.91	1.09/2.41	0.70/1.52
Social-LSTM* [3]	0.67/1.40	0.47/1.00	0.56/1.17	0.79/1.76	1.09/2.35	0.72/1.54
Social-ATTN* [48]	0.33/3.92	0.20/0.52	0.30/2.13	0.29/2.64	**0.39**/3.74	0.30/2.59
TrafficPredict* [29]	3.31/6.37	4.32/8.00	3.76/7.20	2.55/3.57	5.46/9.73	3.88/6.97
SR-LSTM* [58]	0.51/1.10	0.41/0.90	0.32/0.70	0.37/0.74	0.63/1.25	0.45/0.94
Ours*	**0.25/0.41**	**0.16/0.27**	**0.17/0.26**	**0.10/0.14**	0.47/**0.76**	**0.23/0.37**
DESIRE [22]	0.59/1.27	0.41/0.86	0.33/0.72	0.52/1.03	0.93/1.94	0.53/1.11
Social-GAN [14]	0.60/1.26	0.34/0.69	0.42/0.84	0.72/1.61	0.81/1.52	0.58/1.18
FSGAN [21]	0.54/1.14	0.35/0.71	0.32/0.67	0.43/0.89	0.68/1.16	0.46/0.91
SoPhie [42]	0.54/1.24	0.30/0.63	0.38/0.78	0.76/1.67	0.70/1.43	0.54/1.15
Trajectron [18]	0.54/1.13	0.43/0.83	0.43/0.85	0.35/0.66	0.59/1.14	0.47/0.92
MATF [59]	0.44/0.91	0.26/0.45	0.26/0.57	0.43/0.80	1.01/1.75	0.48/0.90
Next [27]	0.60/1.27	0.38/0.81	0.31/0.60	0.30/0.59	0.73/1.65	0.46/1.00
Social-BiGAT [20]	0.55/1.32	0.30/0.62	0.36/0.75	0.49/1.01	0.69/1.29	0.48/1.00
Social-STGCNN [35]	0.44/0.79	0.34/0.53	0.30/0.48	0.49/0.85	0.64/1.11	0.44 / 0.75
Social Ways [4]	0.55/1.31	0.44/0.64	0.51/0.92	0.39/0.66	0.39/0.64	0.46/0.83
PECNet [30]	0.35/0.60	0.22/0.39	0.17/0.30	0.18/0.24	0.54/0.87	0.29/0.48
M2P3 [39]	0.64/1.34	0.45/0.95	0.37/0.79	0.54/1.13	1.04/2.16	0.60/1.27
Transformer-TF [12]	0.35/0.65	0.22/0.38	0.17/0.32	0.18/0.30	0.61/1.12	0.31/0.55
STAR [53]	0.31/0.62	0.26/0.55	0.22/0.46	0.17/0.36	0.36/0.65	0.26/0.53
AgentFormer [54]	0.25/0.45	0.18/0.30	**0.14**/0.24	0.14/0.22	0.45/0.75	0.23/0.39
Trajectron++ [43]	0.30/0.54	0.25/0.41	0.18/0.32	0.18/0.28	0.67/1.18	0.32/0.55
SGN LSTM [57]	0.48/1.08	0.30/0.65	0.26/0.57	0.63/1.01	0.75/1.63	0.48/0.99
Introvert [44]	**0.20/0.32**	0.16/0.27	0.16/0.25	0.11/0.17	0.42/0.70	0.21/0.34
GroupNet [49]	0.26/0.49	0.21/0.39	0.17/0.33	0.15/0.25	0.46/0.73	0.25/0.44
Ours	0.23/0.40	**0.15/0.26**	**0.14/0.23**	**0.10/0.15**	**0.35/0.62**	**0.19/0.33**

Table 6: The average/final displacement error (ADE/FDE) metrics for several methods compared to our model are shown. Lower is better. The models with * have deterministic outputs. All the stochastic models sample 20 possible trajectories and report the best result using a *best-of-20 protocol*. All models observe 8 frames and forecast the subsequent 12 frames.

placement errors, which is a common occurrence among many models, due to lower frequency of video frames and kinematic data. However, our method achieves the lowest ADE/FDE errors on the ETH dataset, showing the effectiveness of our approach in capturing and incorporating information about the movements and behaviors of neighboring pedestrians. Our proposed method outperforms transformer-based methods on the *Zara1* dataset, which is the least structured dataset in the benchmark. The dataset mainly consists of straight lines. The proposed method achieves an ADE similar to the best method [54] (0.14) and also achieves the lowest FDE among all the methods compared (0.23). As seen, our approach outperforms previous Transformer-based methods such as *Transformer-TF*, *STAR*, and *AgentFormer* on the ETH and UCY datasets. Overall, our model offers a competitive alternative to graph-based methods [20, 35] and has the potential to improve trajectory prediction accuracy.

Qualitative Analysis. The qualitative outcomes, depicted in Figure 3, display the accuracy of our trajectory prediction on multiple videos from the ETH and UCY datasets, providing visual evidence of its effectiveness in accurately predicting pedestrian trajectories. The examples presented illustrate various scenarios, including *human-human interaction*, *human-space interaction*, and *avoiding obstacles*. For instance, in the top left example, our model successfully predicts the target pedestrian's trajectory through the store's left-side door. Additionally, in the bottom left example, our method accurately predicts the target person's avoidance of a tree and continues straight towards the train. In the top right example, our model captures a human-human interaction, wherein the target pedestrian slows down before a group of standing people, bypasses them from the left side. These instances demonstrate our method's ability to effectively predict future pedestrian positions, particularly in crowded scenes with complex interactions.

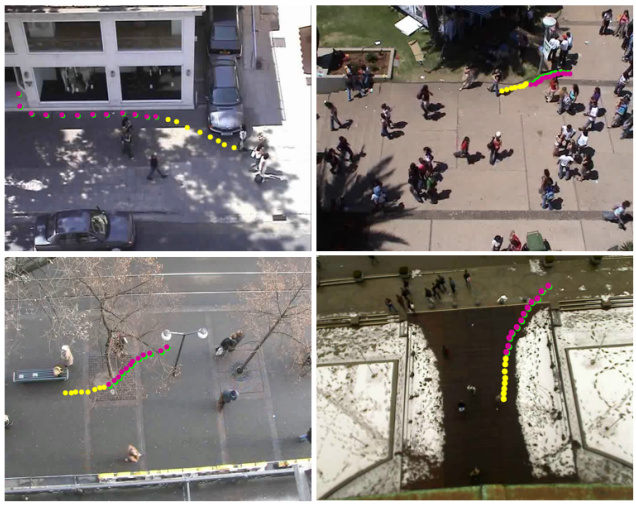

Figure 3: Illustration of the prediction trajectories. Yellow dots represents the past observed while violet & green dots represent our prediction and the ground truth.

5. Conclusion

In this paper, we presented a novel approach for predicting future human motion, encompassing both 3D pose estimation and 2D trajectory prediction, depending on the specific target representation. Our general encoder-decoder architecture combined transformer networks with self-attention mechanisms and LSTM to jointly capture contextual information and motion sequences. By incorporating image video scene context and a sequence of 2D positions/body poses, our model demonstrated remarkable performance across diverse benchmark datasets, including GTA, PROX, and the combined ETH and UCY datasets. The extensive benchmarking underscored the superiority of our approach over existing state-of-the-art methods, particularly excelling in long-term prediction tasks essential for real-world applications. For future work, we propose exploring multi-modal data fusion, integrating additional modalities such as depth information or semantic segmentation maps, along with attention mechanisms and transfer learning.

References

[1] Alexandre Alahi, Vignesh Ramanathan, and Li Fei-Fei. Socially-aware large-scale crowd forecasting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2203–2210, 2014. 2

[2] Alahi, Alexandre and Goel, Kratarth and Ramanathan, Vignesh and Robicquet, Alexandre and Fei-Fei, Li and Savarese, Silvio. Social LSTM: Human Trajectory Prediction in Crowded Spaces. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 961–971, June 2016. 1

[3] Alahi, Alexandre and Goel, Kratarth and Ramanathan, Vignesh and Robicquet, Alexandre and Fei-Fei, Li and Savarese, Silvio. Social LSTM: Human Trajectory Prediction in Crowded Spaces. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 961–971, June 2016. 2, 5, 7, 8

[4] Javad Amirian, Jean-Bernard Hayet, and Julien Pettré. Social Ways: Learning Multi-Modal Distributions of Pedestrian Trajectories with GANs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2019. 5, 7, 8

[5] Zhe Cao, Hang Gao, Karttikeya Mangalam, Qi-Zhi Cai, Minh Vo, and Jitendra Malik. Long-term human motion prediction with scene context. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pages 387–404. Springer, 2020. 1, 2, 3, 4, 5, 6

[6] Ching-Hang Chen and Deva Ramanan. 3d human pose estimation= 2d pose estimation+ matching. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7035–7043, 2017. 3

[7] Hao Cheng, Wentong Liao, Xuejiao Tang, Michael Ying Yang, Monika Sester, and Bodo Rosenhahn. Exploring Dynamic Context for Multi-path Trajectory Prediction. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 12795–12801. IEEE, 2021. 1, 2

[8] Chiho Choi and Behzad Dariush. Looking to relations for future trajectory forecast. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 921–930, 2019. 2

[9] Nemanja Djuric, Vladan Radosavljevic, Henggang Cui, Thi Nguyen, Fang-Chieh Chou, Tsung-Han Lin, Nitin Singh, and Jeff Schneider. Uncertainty-aware short-term motion prediction of traffic actors for autonomous driving. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2095–2104, 2020. 1

[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. 1

[11] Tomohiro Fujita and Yasutomo Kawanishi. Future pose prediction from 3d human skeleton sequence with surrounding situation. *Sensors*, 23(2):876, 2023. 1

[12] Francesco Giuliari, Irtiza Hasan, Marco Cristani, and Fabio Galasso. Transformer Networks for Trajectory Forecasting. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 10335–10342. IEEE, 2021. 7, 8

[13] Renshu Gu, Gaoang Wang, and Jenq-Neng Hwang. Efficient multi-person hierarchical 3d pose estimation for autonomous driving. In *2019 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, pages 163–168. IEEE, 2019. 1

[14] Agrim Gupta, Justin Johnson, Li Fei-Fei, Silvio Savarese, and Alexandre Alahi. Social GAN: Socially Acceptable Trajectories with Generative Adversarial Networks. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2255–2264, 2018. 1, 5, 7, 8

[15] Mohamed Hassan, Vasileios Choutas, Dimitrios Tzionas, and Michael J Black. Resolving 3d human pose ambiguities with 3d scene constraints. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2282–2292, 2019. 3, 4

[16] Dirk Helbing and Peter Molnar. Social Force Model for Pedestrian Dynamics. *Physical Review E*, 51(5):4282–4286, May 1995. 2

[17] Zhen Huang, Xu Shen, Xinmei Tian, Houqiang Li, Jianqiang Huang, and Xian-Sheng Hua. Spatio-temporal inception graph convolutional networks for skeleton-based action recognition. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 2122–2130, 2020. 3

[18] B. Ivanovic and Marco Pavone. The Trajectron: Probabilistic Multi-Agent Trajectory Modeling With Dynamic Spatiotemporal Graphs. *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2375–2384, 2019. 2, 5, 7, 8

[19] Hema S Koppula and Ashutosh Saxena. Anticipating human activities using object affordances for reactive robotic response. *IEEE transactions on pattern analysis and machine intelligence*, 38(1):14–29, 2015. 1

[20] Vineet Kosaraju, Amir Sadeghian, Roberto Martín-Martín, Ian D. Reid, Hamid Rezatofighi, and Silvio Savarese. Social-BiGAT: Multimodal Trajectory Forecasting using Bicycle-GAN and Graph Attention Networks. In *Advances in Neural Information Processing Systems*. Neural Information Processing Systems (NIPS), 2019. 2, 5, 7, 8

[21] Parth Kothari and Alexandre Alahi. Human trajectory prediction using adversarial loss. In *Proceedings of the 19th Swiss Transport Research Conference, Ascona, Switzerland*, pages 15–17, 2019. 7, 8

[22] Namhoon Lee, Wongun Choi, Paul Vernaza, Christopher Bongsoo Choy, Philip H. S. Torr, and Manmohan Chandraker. DESIRE: Distant Future Prediction in Dynamic Scenes with Interacting Agents. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2165–2174, 2017. 1, 2, 5, 7, 8

[23] Alon Lerner, Yiorgos Chrysanthou, and Dani Lischinski. Crowds by Example. *Computer Graphics Forum*, 26(3):655–664, 2007. 2, 4

[24] Jiachen Li, Hengbo Ma, and Masayoshi Tomizuka. Conditional Generative Neural System for Probabilistic Trajectory Prediction. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 6150–6156. IEEE, 2019. 5

[25] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 214–223, 2020. 3

[26] Sijin Li and Antoni B Chan. 3d human pose estimation from monocular images with deep convolutional neural network. In *Computer Vision–ACCV 2014: 12th Asian Conference on Computer Vision, Singapore, Singapore, November 1-5, 2014, Revised Selected Papers, Part II 12*, pages 332–347. Springer, 2015. 3

[27] Junwei Liang, Lu Jiang, Juan Carlos Niebles, Alexander G Hauptmann, and Li Fei-Fei. Peeking into the future: Predicting future person activities and locations in videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5725–5734, 2019. 7, 8

[28] Ruixuan Liu and Changliu Liu. Human motion prediction using adaptable recurrent neural networks and inverse kinematics. *IEEE Control Systems Letters*, 5(5):1651–1656, 2020. 1

[29] Yuexin Ma, Xinge Zhu, Sibo Zhang, Ruigang Yang, Wenping Wang, and Dinesh Manocha. Trafficpredict: Trajectory prediction for heterogeneous traffic-agents. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 6120–6127, 2019. 2, 7, 8

[30] Karttikeya Mangalam, Harshayu Girase, Shreya Agarwal, Kuan-Hui Lee, Ehsan Adeli, Jitendra Malik, and Adrien Gaidon. It Is Not the Journey but the Destination: Endpoint Conditioned Trajectory Prediction. In *ECCV*, 2020. 7, 8

[31] Wei Mao, Miaomiao Liu, Mathieu Salzmann, and Hongdong Li. Learning trajectory dependencies for human motion prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9489–9497, 2019. 1, 2, 5, 6

[32] Wei Mao, Miaomiao Liu, Mathieu Salzmann, and Hongdong Li. Multi-level motion attention for human motion prediction. *International journal of computer vision*, 129(9):2513–2535, 2021. 1

[33] Martinez, Julieta and Black, Michael J. and Romero, Javier. On Human Motion Prediction Using Recurrent Neural Networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4674–4683, 2017. 1

[34] Abduallah Mohamed, Huancheng Chen, Zhangyang Wang, and Christian Claudel. Skeleton-graph: long-term 3d motion prediction from 2d observations using deep spatio-temporal graph cnns. *arXiv preprint arXiv:2109.10257*, 2021. 1, 2, 4, 5, 6

[35] Abduallah Mohamed, Kun Qian, Mohamed Elhoseiny, and Christian Claudel. Social-STGCNN: A Social Spatio-Temporal Graph Convolutional Neural Network for Human Trajectory Prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14424–14432, 2020. 2, 7, 8

[36] Dario Pavllo, Christoph Feichtenhofer, David Grangier, and Michael Auli. 3d human pose estimation in video with temporal convolutions and semi-supervised training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7753–7762, 2019. 2, 5, 6

[37] S Pellegrini, A Ess, K Schindler, and L van Gool. You'll never walk alone: Modeling social behavior for multi-target tracking. In *IEEE 12th International Conference on Computer Vision*, pages 261–268, Kyoto, Sept. 2009. IEEE. 4

[38] Sebastian Pohl, Armin Becher, Thomas Grauschopf, and Cristian Axenie. Neural network 3d body pose tracking and prediction for motion-to-photon latency compensation in distributed virtual reality. In *Artificial Neural Networks and Machine Learning–ICANN 2019: Image Processing: 28th International Conference on Artificial Neural Networks, Munich, Germany, September 17–19, 2019, Proceedings, Part III 28*, pages 429–442. Springer, 2019. 1

[39] Atanas Poibrenski, Matthias Klusch, Igor Vozniak, and Christian Müller. M2P3: multimodal multi-pedestrian path prediction by self-driving cars with egocentric vision. In *Proceedings of the 35th Annual ACM Symposium on Applied Computing*, pages 190–197, Brno Czech Republic, 2020. ACM. 7, 8

[40] Alexandre Robicquet, Amir Sadeghian, Alexandre Alahi, and Silvio Savarese. Learning Social Etiquette: Human Trajectory Understanding In Crowded Scenes. In *Computer Vision – ECCV 2016*, volume 9912, pages 549–565. Springer International Publishing, 2016. Series Title: Lecture Notes in Computer Science. 2

[41] Amir Sadeghian, Vineet Kosaraju, Ali Sadeghian, Noriaki Hirose, Hamid Rezatofighi, and Silvio Savarese. Sophie: An attentive gan for predicting paths compliant to social and physical constraints. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1349–1358, 2019. 1

[42] Amir Sadeghian, Vineet Kosaraju, Ali Sadeghian, Noriaki Hirose, S. Hamid Rezatofighi, and Silvio Savarese. SoPhie: An Attentive GAN for Predicting Paths Compliant to Social and Physical Constraints. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2019*, pages 1349–1358, United States of America, 2019. IEEE, Institute of Electrical and Electronics Engineers. 2, 5, 7, 8

[43] Tim Salzmann, Boris Ivanovic, Punarjay Chakravarty, and Marco Pavone. Trajectron++: Dynamically-Feasible Trajectory Forecasting With Heterogeneous Data. *arXiv:2001.03093 [cs]*, Jan. 2021. 1, 2, 5, 7, 8

[44] Nasim Shafiee, Taskin Padir, and Ehsan Elhamifar. Introvert: Human Trajectory Prediction via Conditional 3D Attention. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16810–16820, Nashville, TN, USA, June 2021. IEEE. 1, 2, 4, 5, 7, 8

[45] Mingyi Shi, Kfir Aberman, Andreas Aristidou, Taku Komura, Dani Lischinski, Daniel Cohen-Or, and Baoquan Chen. Motionet: 3d human motion reconstruction from monocular video with skeleton consistency. *ACM Transactions on Graphics (TOG)*, 40(1):1–15, 2020. 1

[46] Tong Su, Yu Meng, and Yan Xu. Pedestrian Trajectory Prediction via Spatial Interaction Transformer Network. In *2021 IEEE Intelligent Vehicles Symposium Workshops (IV Workshops)*, pages 154–159, 2021. 2

[47] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. 2, 3, 5, 6

[48] Anirudh Vemula, Katharina Muelling, and Jean Oh. Social Attention: Modeling Attention in Human Crowds. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4601–4607, Oct. 2018. 7, 8

[49] Chenxin Xu, Maosen Li, Zhenyang Ni, Ya Zhang, and Siheng Chen. GroupNet: Multiscale Hypergraph Neural Networks for Trajectory Prediction with Relational Reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6498–6507, 2022. 7, 8

[50] Hao Xue, Du Q. Huynh, and Mark Reynolds. SS-LSTM: A Hierarchical LSTM Model for Pedestrian Trajectory Prediction. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1186–1194, Mar. 2018. 2

[51] Kota Yamaguchi, Alexander C. Berg, Luis E. Ortiz, and Tamara L. Berg. Who are you with and where are you going? In *CVPR*, pages 1345–1352, Colorado Springs, CO, USA, June 2011. IEEE. 2

[52] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018. 3

[53] Cunjun Yu, Xiao Ma, Jiawei Ren, Haiyu Zhao, and Shuai Yi. Spatio-Temporal Graph Transformer Networks for Pedestrian Trajectory Prediction. In *European Conference on Computer Vision – ECCV 2020*, pages 507–523. Springer International Publishing, 2020. 2, 5, 7, 8

[54] Ye Yuan, Xinshuo Weng, Yanglan Ou, and Kris M Kitani. AgentFormer: Agent-Aware Transformers for Socio-Temporal. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9813–9823, 2021. 7, 8

[55] Han Zhang, Jianming Wang, and Hui Liu. Video-based reconstruction of smooth 3d human body motion. In *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*, pages 42–53. Springer, 2021. 1

[56] Jason Y Zhang, Panna Felsen, Angjoo Kanazawa, and Jitendra Malik. Predicting 3d human dynamics from video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7114–7123, 2019. 1

[57] Lidan Zhang, Qi She, and Ping Guo. Stochastic trajectory prediction with social graph network. *CoRR*, abs/1907.10233, 2019. 7, 8

[58] Pu Zhang, Wanli Ouyang, Pengfei Zhang, Jianru Xue, and Nanning Zheng. SR-LSTM: State Refinement for LSTM Towards Pedestrian Trajectory Prediction. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12077–12086, 2019. 2, 7, 8

[59] Tianyang Zhao, Yifei Xu, Mathew Monfort, Wongun Choi, Chris L. Baker, Yibiao Zhao, Yizhou Wang, and Ying Nian Wu. Multi-Agent Tensor Fusion for Contextual Trajectory Prediction. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 12126–12134, 2019. 2, 5, 7, 8

[60] Chongyang Zhong, Lei Hu, Zihao Zhang, Yongjing Ye, and Shihong Xia. Spatio-temporal gating-adjacency GCN for human motion prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6447–6456, 2022. 3