TRiPOD: Human Trajectory and Pose Dynamics Forecasting in the Wild

Vida Adeli[1], Mahsa Ehsanpour[2], Ian Reid[2], Juan Carlos Niebles[3],
Silvio Savarese[3], Ehsan Adeli[3], Hamid Rezatofighi[3,4]
[1]*Ferdowsi University of Mashhad* [2]*University of Adelaide*
[3]*Stanford University* [4]*Monash University*

Abstract

Joint forecasting of human trajectory and pose dynamics is a fundamental building block of various applications ranging from robotics and autonomous driving to surveillance systems. Predicting body dynamics requires capturing subtle information embedded in the humans' interactions with each other and with the objects present in the scene. In this paper, we propose a novel TRajectory and POse Dynamics (nicknamed TRiPOD) method based on graph attentional networks to model the human-human and human-object interactions both in the input space and the output space (decoded future output). The model is supplemented by a message passing interface over the graphs to fuse these different levels of interactions efficiently. Furthermore, to incorporate a real-world challenge, we propound to learn an indicator representing whether an estimated body joint is visible/invisible at each frame, e.g. *due to occlusion or being outside the sensor field of view. Finally, we introduce a new benchmark for this joint task based on two challenging datasets (PoseTrack and 3DPW) and propose evaluation metrics to measure the effectiveness of predictions in the global space, even when there are invisible cases of joints. Our evaluation shows that TRiPOD outperforms all prior work and state-of-the-art specifically designed for each of the trajectory and pose forecasting tasks.*

1. Introduction

The ability to forecast human movements (pose dynamics and trajectory) in time is an essential component for many real-world applications, including robotics [38, 48], healthcare [32], detection of perilous behavioral patterns in surveillance systems [37, 53].

While this problem sounds interesting, it is extremely challenging in real-world scenes due to the different factors involved. Humans are intuitively social agents, able to effortlessly conceive a detailed level of semantics from the scene, which contributes to making swift decisions for their next movements. To accurately forecast their trajectory and pose dynamics, one primary factor is the interactions between people in the scene and the influences their joints have on each other. For example, consider a tennis-playing scene, when the opponent starts serving and hits a stroke, the other person is probable to take a ready position in the near future (*e.g.* see the purple agent in Fig. 1). Besides, the objects involved in the scene can provide informative clues for future prediction. For instance, when the person observes the ball in the tennis example, he/she would take a striking pose to return it. However, the movements of all the persons in the scene are not always highly correlated with each other nor the humans to objects. For instance, in Fig. 1, the pose and motion of tennis players will be barely affected by the skateboarder or his skateboard. This

Figure 1. An example of a real-world scene containing different levels of interactions (human to human and human to objects). The top-left graph shows the weighted interaction graphs between humans (blue edges) and between humans and objects in the scene (green edges). The top-right graph illustrates the evolved social interactions over time in the future. The red arrows indicate an example of a relation being intensified over time.

defines different levels of interactions that need to be discovered by the forecasting model. In addition, these different levels of interactions can change over time, *i.e*. getting strengthen or weaken. In Fig. 1, lines thickened in future human-human graph (indicated by red arrows), show that the skateboarder's movements increasingly correlate with the parent and the kid (passengers), while some others lose correlations over time. Finally, a person might move outside the sensor field-of-view or be a partially/fully occluded by an object. In these cases, it is important to have an indication of visibility/invisibility for each prediction, which can be interpreted as its reliability score, conducive for the applications such as navigation safety and a collision risk assessment for an autonomous robot/vehicle.

Existing solutions often neglect some of these challenging factors and hence fall short when applied to real-world *in-the-wild* scenarios. Pose dynamics forecasting methods [14, 40, 41, 58] *mostly* forecast the changes in joints with respect to a center position, ignoring the global position changes. They often do not effectively model *all* the informative environmental and social interactions in the scene either. Similarly, the influence of individual joints is usually overlooked in trajectory forecasting [22, 27]. Moreover, existing frameworks often assume that all tracks and/or body joints are always observable in the past and future, which is an impractical assumption in many real-world scenarios.

To address these challenges, we push the current state of existing solutions for human pose dynamics and trajectory forecasting one step forward toward more practical scenarios in-the-wild by considering all these factors together. To this end, similar to other works that use attentional graphs for various purposes [27, 33], we model the input *skeleton body joints*, the *social human-human* and *human-object interactions* with different attention graphs. Since, these two types of information are different by nature, we give an effective solution to fuse them and as well make them insensitive to their choice of order by applying an *iterative message passing*. Furthermore, on account of the fact that humans may retain their influences on each other consistently in future, we do not content with only representing the history of interactions, but also we preserve their spatio-temporal attentional relationships by modeling them also in future prediction phase. To overcome the problem of accumulative error in sequential models for long-term sequences and to speed up the convergence, we take a curriculum learning approach to train our model. Finally, since there is no proper benchmark dataset for such real-world problem, we introduce a new *benchmark* by repurposing existing datasets and introducing relevant evaluation *metrics*.

In summary, the main contributions of our paper are to 1) propose a model that considers all the mentioned challenges together by (*i*) modeling the human skeleton, social and human-objects interactions through different dense and sparse graph incorporating attention, (*ii*) introducing a **message passing** approach to efficiently fuse different level of interactions, (*iii*) dynamically modelling the spatio-temporal attentional human interactions during **decoding phase**, (*iv*) addressing the concept of **joint invisibility or body disappearance** in trajectory and pose dynamics forecasting problem, (*v*) suggesting a curriculum learning strategy to compensate accumulating error in recurrent models, 2) introduce **proper evaluation metrics** and a new **benchmark** for this real-world problem.

2. Related work

A. Pose dynamics forecasting. Generally, pose dynamics forecasting aims to predict the future human pose coordinates in which global motion (trajectory) is excluded. Early approaches modeled human dynamics by utilizing hand-crafted features and applying probabilistic graphical models [60, 61]. Recently, deep sequence-to-sequence models [8, 14, 19, 31, 41, 57, 58] have been used to capture such dynamics. Following the success of RNNs in capturing temporal dependencies, these models have been extensively used in capturing pose dynamics [13, 14, 18, 41, 46]. Since future forecasting of human pose dynamics is not a deterministic task, some works have utilized VAEs and GANs [5, 25, 57, 65, 66, 68] and some focusing on the scene context [11, 15]. With the popularity of recently proposed transformers [54, 2], Mao et al. [40] introduced an attention-based motion extraction model that aggregates current motion with its history. Likewise, in [10], a transformer-based architecture is used for capturing the spatio-temporal correlations of the human pose. All the aforementioned works are limited to only predicting local dynamics since global motion is subtracted from the human body joint coordinates. Further, interactions between joints in skeleton level and individual level are not modeled or captured. We argue that simultaneously capturing global and local dynamics is essential in forecasting reliable and robust 2D and 3D human poses and in general fine-grained human understanding. Moreover, human joints' movements are tightly coupled in the skeleton level and between interacting individuals. The problem is best formulated in a social manner.

B. Human trajectory predictions. The goal of human trajectory prediction is to predict a set of 2D coordinates for each human characterizing its global motion. Human social interactions in crowds have always been considered an important cue for predicting humans' global trajectories, which were dominantly ignored by pose dynamic frameworks. Its literate goes back to pre-deep learning era when hand-crafted features were mainly used [4, 24, 43, 45, 47, 64]. Although being successful, these works are task-dependent and require domain expert knowledge to carefully design hand-crafted rules. Recent deep data-driven models [3, 17, 23, 33, 35, 39, 49, 34, 6] used a recurrent neural network and a social pooling layer on

top to capture spatio-temporal social feature representation to predict the future trajectory of each individual. More recently, graph-structured models have been used to model human global motion and the existing interactions [15, 28, 30, 33, 50, 44, 59]. [28, 33] used graph attention networks [55] to model social interactions. Trajectron [30] proposed a graph-structured model that predicts many potential future trajectories. Trajectron++ [50] also proposed a graph-based model that incorporates environmental information such as semantic maps and integrated with robotic planning. In [21] a transformer-based method is proposed. Some other works improve the accuracy by incorporating novel trainable modules. For instance, [26] proposed a neural motion message passing model to explicitly model the directed interactions between actors, [16] proposed a trajectory proposal network to ensure safe and multimodal predictions and [52] proposed reciprocal learning to train forward and backward networks. While performing well, all these works lack modeling detailed human joints dynamics. Capturing and forecasting such fine-grained human body motions, *i.e.* human poses, is essential for safe autonomous agents navigating through humans.

C. Pose dynamics and trajectory forecasting. As discussed earlier, learning fine-grained human joint dynamics as well as global motion are major components of a human understanding model that lead to development of a safe agent navigating in a crowd [38]. Recently, there have been some attempts to tackle the two problems in a unified manner. [11] released a new synthetic dataset from a game engine and focused on utilizing scene context to tackle the unified task. [1] unified human pose and trajectory forecasting in a socially-aware manner. Despite the novelty in the formulation of the problem by these works, social interactions are ignored or modelled in a basic way, making these works unable to handle invisible joints and complex in-the-wild scenarios. Here, we encode social interactions spatially and temporally via attention networks and message passing, and explicitly modeling human-objects interactions.

3. Trajectory and Pose Dynamics Forecasting

Humans are, by nature, social agents with complex interactions with not only other similar agents but also different parameters of the scene. All of these interactions and the conception of an agent from its surroundings build its actions, forming its future body pose and trajectory. Generally, the problem of joint human trajectory and pose dynamics forecasting can be defined as estimating the person's most probable future pose and trajectory given their prior history. Needless to say that when it comes to prediction in-the-wild, many other environmental factors come into play in addition to the individual's history. Likewise, our goal is to model the complex human-human and human-object interactions in a way that can also predict all the joint visibility indicators in the future.

Problem Definition. Formally, assuming the past global history of a person $p \in \mathcal{P}$ as $\mathbf{X}^p_{1:\tau_o} = \{\mathbf{x}^p_1, \mathbf{x}^p_2, ..., \mathbf{x}^p_{\tau_o}\}$, where $\mathbf{x}^p_t \in \mathbb{R}^F$ with F as the number of parameters describing the state of all the joints for person p, our goal is to predict the set of poses $\mathbf{Y}^p_{+1,+\tau_f}$ for the future τ_f frames.

$$\mathbf{Y}^p_{+1,+\tau_f} = \{\mathbf{y}^p_{\tau_o+1}, \mathbf{y}^p_{\tau_o+2}, ..., \mathbf{y}^p_{\tau_o+\tau_f}\}, \quad \forall p \in \mathcal{P} \quad (1)$$

where $\mathbf{y}^p_t \in \mathbb{R}^F$. Throughout the paper, we consider any arbitrary variable, *e.g.* ϕ_t being defined at time t and $\phi_{t_1:t_2} = \{\phi_{t_1}, \phi_{t_1+1}, ..., \phi_{t_2}\}$ between times t_1 and t_2 and hold same convention for all variables. We also use the $+t$ notation as an indication of the future time. In our case, the state of each joint $k \in K$, is specified by 3 major indicators, *i.e.* offset $\Delta\boldsymbol{\ell}$ (temporal location velocities), absolute locations $\boldsymbol{\ell}$ and joint visibility score s, which is a binary value being 0 if the joint is invisible, *i.e.* $\mathbf{x}^p_t = \{(\Delta\boldsymbol{\ell}^p_t(k), \boldsymbol{\ell}^p_t(k), s^p_t(k)) \mid k{=}1 : K\}$, where $\Delta\boldsymbol{\ell}, \boldsymbol{\ell} \in \mathbb{R}^d$; $s \in [0, 1]$. $\mathbf{y}^p_t$ is defined the same as $\mathbf{x}^p_t$ (details are available in supplementary material). Unlike existing methods on pose forecasting, we use inputs in the original space, which means that the poses are not centered and contain the global trajectory. The importance of modeling these two sources is demonstrated in [1].

TRiPOD Model: Our TRajectory and POse Dynamics (TRiPOD) model consists of multiple components (Fig. 2) and sub-components, described in detail as follows.

A. Attentional Human Pose History. The dynamics of human body skeleton are the primary information that convey important knowledge for modeling the past history of pose and trajectory and also their prediction in future. This emphasizes the importance of how this information is represented to the pose forecasting models.

Most earlier methods in pose forecasting [14, 31, 42] utilize the joint coordinates, or other primary information upon coordinates such as velocities, to form a raw feature vector as their input. However, doing so ignores the significance of the natural connectivities in human body skeletons. Inspired by recent works [36, 67], we model skeleton pose as a graph, leveraging joint connections. However, as the influence of joints on each other is not uniform, we use an attentive graph encoder to model them. The inputs to this pose attention graph $ATT_{In}(.)$ are state information of each body joint in each input frame $\mathbf{X}^p_t$.

$$\mathcal{G}^p_{In,t} = ATT_{In}(\mathbf{X}^p_t; \mathbf{W}_{In}) \quad (2)$$

where $\mathbf{W}_{In}$ is the set of the parameters of input pose attention graph. Accordingly, the output would be $\mathcal{G}^p_{In,t}$, which is a body pose representation, attending over different joint interactions. Then, an encoder RNN_{en} (*e.g.* LSTM) encodes the past history of each person's skeleton graph, up to time step t as Eq. (3), with $\mathbf{W}_{en}$ as the encoder's parameters and $\boldsymbol{h}_{en,0:\tau_o-1}$ as hidden states.

$$\mathcal{Z}^p = RNN_{en}(\mathcal{G}^p_{In,1:\tau_o}, \boldsymbol{h}_{en,0:\tau_o-1}; \mathbf{W}_{en}) \quad (3)$$

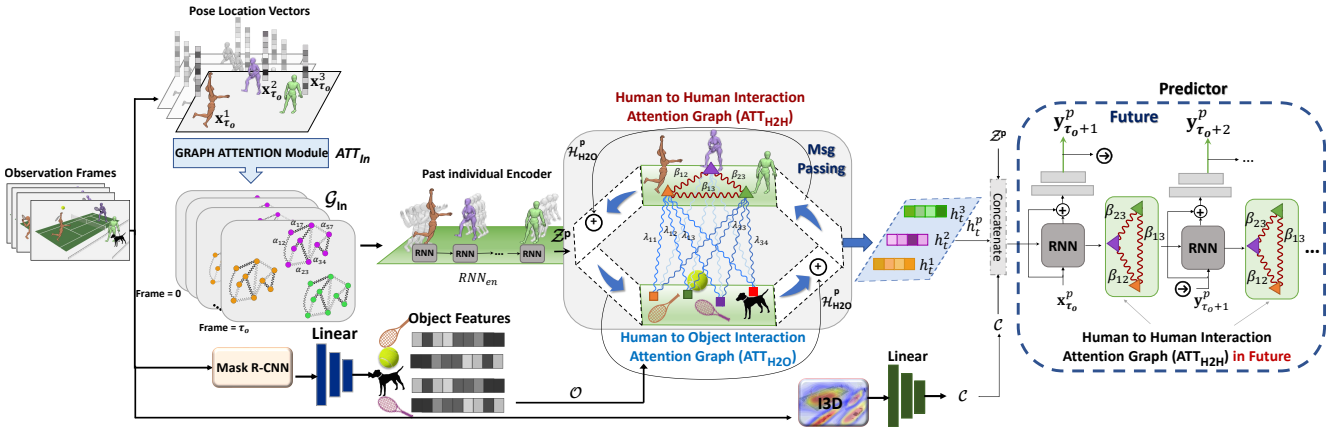

Figure 2. An overview of the TRiPOD model proposed for human pose dynamics and trajectory forecasting in-the-wild. First, the history of poses are initially encoded using an attentive graph upon the body skeleton joints. Then, the encoded history is used to model interactions in Human to Human (H2H) and a Human to Object (H2O) attention graphs through a couple of iterative message passing. The future poses are then predicted, with the aid of the refined social interactions at every steps in the future.

resulting in the encoded past global pose history $\mathcal{Z}^p$.

B. Object and Global Scene Features. As alluded earlier, to conceive the high-level semantics in the scene, the model should understand interactions between humans and objects and the scene context since the humans' pose is highly correlated to them. For this purpose, an object detector is used to extract the objects in the scene in the last observation frames, *i.e.* τ_o. The final object representations ($\mathcal{O}$) is then obtained by passing the object feature vectors including visual feature, geometrical information, and its class label, through few embedding layers. Regarding the holistic scene features ($\mathcal{C}$) of all observation frames, we use a spatio-temporal model to represent the sequence and then the feature vectors are passed through a couple of embedding layers to form the final features.

C. Human to Object Attention Module. Since all the humans are not completely interrelated with all the objects in the scene, we also want the model to learn these different levels of interactions. To achieve this, we utilize a graph attention module (ATT_{H2O}) for encoding the human to object (H2O) interactions, which takes as input the encoded past representation of person $\mathcal{Z}^p$ and the described object features $\mathcal{O}$ and outputs the H2O encoded interaction $\mathcal{H}^p_{H2O}$.

$$\mathcal{H}^p_{H2O} = ATT_{H2O}(\mathcal{Z}^p, \mathcal{Z}^{P\setminus p}, \mathcal{O}; \mathbf{W}_{H2O}) \qquad (4)$$

where $\mathcal{P} \setminus p$ represents all the people in the scene excluding person p and $\mathbf{W}_{H2O}$ represents the parameters of H2O attention graph (Blue wavy links in Fig. 2).

D. Social Attention Module. Next, to augment the level of semantic in human pose dynamics forecasting in real-world, we encode humans' social interactions. [28, 33] used graph attention networks [55] to model social interactions. However, they incorporated it with simplistic inputs, without considering object interactions. Similar to the H2O attention module since different persons have different levels of interaction in the scene, their relation is modeled with an attention graph ATT_{H2H}, taking each person's representation produced by the ATT_{H2O} graph and outputs a representation $\mathcal{H}^p_{H2H}$ containing weighted social interactions. Similarly, $\mathbf{W}_{H2H}$ is the parameters of H2H attention graph (Red wavy links in Fig. 2).

$$\mathcal{H}^p_{H2H} = ATT_{H2H}(\mathcal{H}^p_{H2O}, \mathcal{H}^{\mathcal{P}\setminus p}_{H2O}; \mathbf{W}_{H2H}) \qquad (5)$$

E. Message Passing. So far, we have two different sources of information (human-human and human-object interaction) that are different by nature, even the types of effects they have are distinct. Thereby, an efficient approach is required to combine these two sources. We apply an iterative message passing inspired by approaches primarily proposed for obtaining useful information of molecular data [20]. However, we reformed it to employ in our problem, to combine this two types of information effectively and make the framework invariant to their choice of order. We extend the message passing concept to share information internally and between nodes of two different attention graphs.

In simplistic terms, we describe our message passing module on the two undirected graphs of ATT_{H2O} and ATT_{H2H}. Assume we have a set of node features f, where f is the nodes in ATT_{H2O} or ATT_{H2H} graphs, and edge features of $e_{pu} \in \{\mathbf{W}_{H2H}, \mathbf{W}_{H2O}\}$, each node is then allowed to exchange information with its neighbours through a couple of N time steps. Firstly, the node feature would become updated through a run of the H2O attention graph and is then fed to the H2H attention graph. We repeat this message passing process for a specified number of times N. After one such step, each node state would gain a primary perception of its immediate neighbors. Then, repeating more steps enhances these perceptions by incorporating second-order information and so on. Our message passing involves two main procedures: message passing (Eq. (6)) and the node update (Eq. (7)). During the message passing, each person's node's hidden state is updated based on the message m^p_{n+1} while that of objects remains intact.

$$m^p_{n+1} = \underset{u \in neighbors(p)}{ATT}(f^p_n, f^u_n, e_{pu}), \qquad (6)$$

where $ATT \in \{ATT_{H2O}, ATT_{H2H}\}$, $p \in \{\mathcal{P}\}$ and $u \in \{\mathcal{O}, \mathcal{P}\}$.

$$f_{n+1}^{p} = U_n(f_n^p, m_{n+1}^p), \tag{7}$$

u denotes the neighbors of p in H2O & H2H graphs and U_n is the update function at step n, here an average function.

F. Future Social Interactions. To accurately predict future poses, only incorporating historical human social interactions is not sufficient. The model should also dynamically reconsider social interactions in the future, leveraging other people's actions during the same window of time. This valuable source of information, in addition to the loss supervision, can effectively improve training and performance during inference, as shown in the experimental section. This interactive decoding is mainly ignored by the previous works. We address this issue by retaining the future interactions through the attentional H2H graph. Formally, after encoding all the previous individual and social history, scene and human-object interactions into a single representation for each person, the corresponding features are used as the input hidden state of a decoder predictor ($\boldsymbol{h}_{dec,0}^p = f_N^p$) to generate the set of future poses recursively after applying an embedding function ψ.

$$y_{+t+1}^p = \psi\left(RNN_{dec}(y_{+t}^p, \boldsymbol{h}_{dec,+t}^p; \mathbf{W}_{dec}); \mathbf{W}_{\psi}\right) \tag{8}$$

Where $t \in (0, ..., \tau_f - 1)$ and y_{+t+1}^p is the output global pose predicted for person p at time $\tau_o + t + 1$. Then the persons' representations are refined by the social attention graph forming the hidden state of the next time step (Eq. (9)) and the whole process continues until time step τ_f.

$$\boldsymbol{h}_{+t+1}^p = ATT_{H2H}(\boldsymbol{h}_{dec,+t}^p, \boldsymbol{h}_{dec,+t}^{\mathcal{P}\setminus p}; \mathbf{W}_{H2H}) \tag{9}$$

G. Training Strategies. A common problem in pose forecasting methods is that in the training phase, the model cannot recover from its accumulating errors at each time step and therefore, feeding this error as the input to the next step propagates it throughout the network and results in a large discrepancy between prediction and ground-truth poses in long-term. To address this problem, we *first* make the final prediction to consider both the input and output of the RNN decoder at each time step using a skip connection to retain the output's continuity and can recover from the error. *Second*, we employ the concept of curriculum learning [9] and adopted it to train our model, which is starting with easier sub-tasks and gradually increasing the difficulty level of the tasks. This approach expedites the speed of convergence. Hence, we divide our future pose prediction problem for τ_f frames into $\frac{\tau_f}{\omega}$ sub problems where ω is the number of frames injected at each step. The model is first trained on the first sub-frames of prediction and after that, it learned this sub-task, then the second set of sub-frames are added to be trained. Note that the loss is calculated based on the new injected frames and the previous ones at each step.

Training Loss: As is proven in [1], naturally the two problems of body pose dynamics forecasting and trajectory forecasting are highly correlated and should be approached jointly. Hence, we define a joint loss function in the global data coordinates and use the three described source of information as input (*i.e.* offset $\Delta\boldsymbol{\ell}$, absolute locations $\boldsymbol{\ell}$ and invisibility indicator s). For the first two, we minimize the norm (the MSE ℓ_2) error values of the ground-truth $(\Delta\boldsymbol{\ell}, \boldsymbol{\ell})$ and the prediction $(\hat{\Delta\boldsymbol{\ell}}, \hat{\boldsymbol{\ell}})$. For the visibility score, we employ a Binary Cross Entropy loss. In training mode, if a joint is invisible in the truth, no gradient based on MSE loss ($\mathcal{L}_{\ell 2}$) is calculated for it, while the visibility loss ($\mathcal{L}_{BCE}$) still penalizes the predictions. This concept is implemented by setting the value of loss to zero for that joints using a visibility mask $\mathcal{M}$ and a normalization factor η. Considering Θ as the collections of all weights, *i.e.* $\Theta = \left(\mathbf{W}_{In}, \mathbf{W}_{en}, \mathbf{W}_{H2O}, \mathbf{W}_{H2H}, \mathbf{W}_{dec}, \mathbf{W}_{\psi}\right)$, we use the following to train our model.

$$\Theta^* = \underset{\Theta}{\operatorname{argmin}}\ \mathbb{E}_{p,t}[\mathcal{L}(\mathbf{y}_t^p, \hat{\mathbf{y}}_t^p)] \tag{10}$$

$$\mathcal{L}(\mathbf{y}_t^p, \hat{\mathbf{y}}_t^p) = \frac{1}{\eta}\Big(\mathcal{L}_{\ell 2}(\Delta\boldsymbol{\ell}_t^p, \hat{\Delta\boldsymbol{\ell}}_t^p) + \mathcal{L}_{\ell 2}(\boldsymbol{\ell}_t^p, \hat{\boldsymbol{\ell}}_t^p)\Big) \times \mathcal{M} + \mathcal{L}_{BCE}\left(s_t^p, \hat{s}_t^p\right) \tag{11}$$

4. Benchmarking

There is no standard dataset available that can provide a fair pipeline for this problem, considering all the mentioned challenges. Furthermore, there is no metric for pose dynamics and trajectory forecasting, which accounts for joints' invisibility cases (such as occlusion or being outside the scene). We form a standard assessment platform as a benchmark *(i)* using existing multi-person datasets by repurposing them for human pose dynamics forecasting and, *(ii)* by proposing new metrics, taking the both source of errors, *i.e.* predicted joint locations and visibility indicators, into account. Our benchmark is available at http://somof.stanford.edu/.

4.1. Metrics Generally, consistent with prior work [1, 40, 41], the fundamental procedure to report our evaluation results is based on the Mean Per Joint Position Error (MPJPE) [29], which is the average Euclidean distance ($d_{\ell 2}$) between ground-truth and estimated joint positions (but in our case in the global coordinate), averaged over number of persons in the sequence at each frame $i \in \{\tau_o+1, ..., \tau_o+\tau_f\}$. However, since we introduce the concept of invisible joints, we propose other types of metrics accounting for those cases.

Visibility-Ignored Metric (VIM). This metric is the simple MPJPE metric except that the invisible joints (if exist) are not penalized and are simply discarded by considering truth.

Visibility-Aware Metric (VAM). The second metric is proposed for performance evaluation in the presence of joint invisibility. Here, the goal is to calculate the distance of every joint per person in each time between the ground-truth and the prediction. When assuming the possibility of a joint being invisible, for every predicted joint, q, and its equivalent in ground-truth, g, three possible scenarios can be pre-

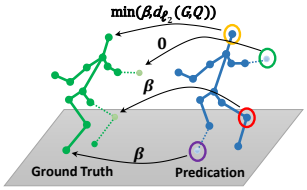

Figure 3. Illustration of VAM (joint visibility-aware metric). Pale joints are the missing joints in ground-truth (green) and prediction (blue) skeletons. The values on top of arrows are the penalty values. For other joints, the penalty is calculated by the minimum of error distance and cutoff β.

sumed (see Fig. 3): 1) Both joints are invisible. 2) One joint is invisible in either the prediction or ground-truth. 3) Both joints are visible. These cases can be modeled by two singleton sets for each joint in ground-truth ($G = (\emptyset$ or $\{g\})$) and prediction ($Q = (\emptyset$ or $\{q\})$). To do so, inspired by the concept of miss-distance in OSPA metric [51], which is a distance metric (by mathematical definition) for comparing two sets of point patterns, we define $d_o^{\beta}(G, Q)$ as the distance between the two Singleton sets as follows:

$$d_o^{\beta}(G,Q) = \left(d^{(\beta)}(G,Q)^2 + \beta^2 |c_g - c_q|^2\right)^{\frac{1}{2}}, \qquad (12)$$

where the $c_g \in \{0,1\}$ and $c_q \in \{0,1\}$ as the cardinality of the two singleton sets, $d^{(\beta)}(G,Q) = \min(\beta, d_{\ell 2}(\{g\},\{q\}))$ is the distance between two joints (if both visible) cut off at $\beta > 0$. Note $d^{(\beta)}(G,Q)$ is zero if any of the sets are empty.

The value of $d_o^{\beta}(G,Q)$ for all three possible scenarios are shown in Fig. 3. Finally, the Visibility-Aware Metric (VAM) d_v for all persons' K joints is

$$d_v = \frac{1}{\alpha} \sum_{p \in \mathcal{P}} \sum_{G,Q \in K} d_o^{\beta}(G^p, Q^p) \qquad (13)$$

where α is the normalization variable which can be defined as the summation of maximum cardinality of the prediction and ground-truth sets over each joint and each person: $\alpha = \sum_{p \in \mathcal{P}} \sum_{g,q \in K} max(c_g^p, c_q^p)$.

Visibility Score Metric. As the third metric, we evaluate the model only on the visibility scores s. For this purpose, we apply two criteria, the Intersection over Union (IoU) and the F1-score measure for all the joints in future frames, averages over the number of joints and the number of persons.

4.2. Data Since the proposed method's main objective is to model and predict human poses in-the-wild, the choice of the dataset used for evaluation should also conform to the criteria in the real world. To this end, we re-purpose the recently released 3D Poses in the Wild dataset (3DPW) [56] and the PoseTrack [7] to create a standard pipeline for human pose dynamics and trajectory forecasting, to unify both communities, and to create a platform with proper data splits and metrics to ensure a fair comparison between different approaches. These datasets reasonably provide us with the unconstrained set of information for complex real-world scenarios and contain both pose annotations and global trajectory data. The PoseTrack containing poses with invisible joints enables us to reconsider the occlusion and disappearing individuals problem in pose dynamics forecasting, which is essential to reflect the trustworthiness of the predictions. Details are elaborated in the supplement.

5. Experiments

In this section, we evaluate the performance of the TRiPOD model on the proposed benchmark and compare it against state-of-the-art methods. We further conduct ablation study and provide some qualitative results.

We use a one layer sequence-to-sequence model for encoding and decoding poses, with LSTM modules with a hidden dimension of 256. To model the attention mechanism in graph, we utilize the graph attention networks (GATs) [55] which is dense in case of input pose and H2H module and sparse for H2O in which only humans and objects are linked. Also, the social graph in decoding phase has shared parameters with the H2H graph. To extract the objects, a mask-rcnn-R-50-FPN-3x model, pre-trained on COCO with box AP of 41.0 [62] is used and for spatio-temporally representing the scene context, the I3D model [12] pre-trained on Kinetics is employed. Then, different two layer embedding modules are applied to the object and I3D features. The hyper-parameters are selected through experiments on a validation set (details are available in supplementary material). To report the results, each experiment is performed three time and their average values are reported.

5.1. Quantitative Results

A. Baselines. Generally, the two problems of pose dynamics and trajectory forecasting had been commonly treated as isolated problems by the community and thereby the number of approaches that jointly model these information are limited. Consequently, to investigate the effectiveness of the proposed method, we break down the problem and retrain the task with the models in each community separately and then combine their results in prediction. Various works have been conducted in these two community, however, we try to select the most popular and recent state-of-the-art methods that are conceptually similar to our problem and could be simply applied. Ultimately, we select [40, 41] and [3, 22, 27] as the most popular and recent state-of-the-art methods in human pose dynamics and trajectory forecasting, respectively, and did our best to fairly retrain these methods by effectively setting up their parameters and prepare data in compliance to how should be used to obtain the best results (data preparation details for baselines are available in supplementary material). We also compare against SC-MPF [1] that considers the two problems jointly. Note, other prior works do not mainly consider the joint problem.

Joint Evaluation: We first compare the results jointly in global space and in the next step represent the comparison in each problem separately. For *3DPW*, the results are reported based on visibility ignore metric (VIM), since it does not have invisible joint cases (in this case, VIM acts the

Table 1. Error rate in **3DPW** (in cm) and **PoseTrack** (in pixel). In each column the best obtained result is highlighted with boldface typesetting.

	3DPW					**PoseTrack**									
	VIM (Invisibility ignored)					**VIM (Invisibility ignored)**					**VAM (Invisibility considered $\beta = 200$)**				
	prediction time in milliseconds					prediction time in milliseconds					prediction time in milliseconds				
	100	240	500	640	900	80	160	320	400	560	80	160	320	400	560
Center pose Trajectory															
PF-RNN [41] + S-LSTM [3]	73.2	126.79	180.03	201.75	277.53	87.05	103.35	129.19	138.66	160.96	101.37	114.09	133.88	145.39	160.49
PF-RNN [41] + S-GAN [22]	68.40	119.72	172.73	195.88	263.05	84.74	98.94	121.35	129.55	150.16	99.96	111.51	129.3	140.08	154.09
PF-RNN [41] + ST-GAT [27]	67.12	116.53	164.61	189.82	250.88	80.93	95.72	119.03	127.66	149.44	96.16	109.06	127.5	137.75	152.49
Mo-Att [40] + S-LSTM [3]	65.24	109.67	168.94	200.16	268.14	84.45	101.63	121.16	135.48	157.48	100.02	113.89	130.41	144.27	158.24
Mo-Att [40] + S-GAN [22]	63.41	106.25	161.89	193.98	258.51	81.33	97.45	118.74	125.78	147.12	99.21	109.56	129.08	139.25	152.47
Mo-Att [40] + ST-GAT [27]	62.41	94.59	153.24	188.02	249.91	78.14	93.75	115.61	119.31	140.83	97.16	107.42	125.36	136.04	149.78
Joint Trajectory & Pose															
SC-MPF [1]	45.44	73.73	129.23	159.47	208.31	21.41	39.92	66.32	77.73	93.41	82.74	95.67	106.5	113.19	133.78
TRiPOD	**31.04**	**50.8**	**84.74**	**104.05**	**150.41**	**15.36**	**26.32**	**46.45**	**57.94**	**71.78**	**50.62**	**60.77**	**79.69**	**80.07**	**96.98**

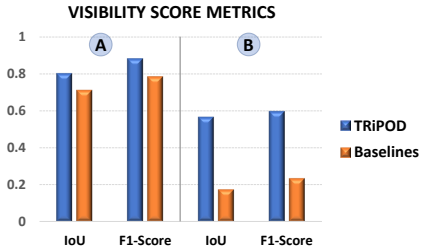

Figure 4. Comparison of visibility score metrics. (A): All data considered. (B): Joints with at least one case of disappearance in future are considered.

same as simple MPJPE metric). Table 1 evidences that we achieve the best results for joint pose dynamics and trajectory forecasting on 3DPW. For *PoseTrack*, since the dataset contains cases in which the joints are occluded or poses disappear by the persons leaving the scene, we employ all the three proposed metrics (see subsection 4.1) for its evaluation. Table 1 shows that we consistently outperform other methods in both ignored and considered joint visibility metrics (VIM and VAM) in PoseTrack. Fig. 4 demonstrates the evaluation results for the visibility score metric. The visibility scores s for baselines are considered to be always true, since they assume all joints are visible during the whole past and future. The goal is to investigate the performance of the model in recognizing visible/invisible cases. To do so, The *IoU* and *F1-score* of the binary vectors (s) are calculated (as described in subsection 4.1). Then two approaches are adopted: **A)** The whole data is used in the metric evaluation. Generally, in PoseTrack, 27.28% and 28.82% of joints are invisible in the observation and the future frames, respectively. **B)** Since the visible cases are more frequent, to better show the gap between the performance of TRiPOD and the baselines in predicting invisibility, we perform evaluations only on joints with at least one future case of invisibility. This experiment shows the performance difference between methods when some joints disappear in some future frames. Fig. 4 shows that TRiPOD is able to estimate joint invisibility, and this claim can be better seen when always-present joints are not considered in the evaluation.

Separate Evaluation: For more comparison, we also evaluate the TRiPOD model on center pose and trajectory independently and compare results with baselines in each community. Fig. 5 illustrates that our TRiPOD model achieves the lowest error rate (VIM) for center pose prediction in both datasets and in trajectory forecasting performs in par with ST-GAT in 3DPW and outperforms others in PoseTrack (a more challenging dataset with invisible joints).

Table 2. Ablation study on **3DPW** based on VIM (ignored invisibility). Each notation is defined as: $\mathcal{C}$: Scene context, $\mathcal{P}$: Input pose representation, tensors (T) or attention graph (G). $\mathcal{H}$: Social module (max operation (M) or attention graph (G)). $\mathcal{O}$: Human-object graph. $\mathcal{M}$: Message passing. $\mathcal{FH}$: Human interactions in future. **CL**: Curriculum Learning.

	$\mathcal{C}$	$\mathcal{P}$	$\mathcal{H}$	$\mathcal{O}$	$\mathcal{M}$	$\mathcal{FH}$	prediction time in milliseconds				
							100	240	500	640	900
S-MPF [1]	✗	T	M	✗	✗	✗	52.89	89.27	146.2	176.98	249.18
SC-MPF [1]	✓	T	M	✗	✗	✗	45.44	73.73	129.23	159.47	208.31
Baseline 1	✓	T	G	✗	✗	✗	39.74	64.44	106.13	128.36	181.32
Baseline 2	✓	G	G	✗	✗	✗	33.99	54.57	93.75	114.75	167.32
Baseline 3	✓	G	G	✓	✗	✗	32.64	52.73	91.25	111.9	166.68
Baseline 4	✓	G	G	✓	✓	✗	32.85	52.64	88.77	108.38	161.72
TRiPOD	✓	G	G	✓	✓	✓	**31.56**	**51.97**	**86.53**	**107.52**	**153.12**
TRiPOD(CL)	✓	G	G	✓	✓	✓	**31.04**	**50.8**	**84.74**	**104.05**	**150.41**

Results Discussion: The results in Table 1 and Fig. 5 reveal that although using the combination of two state-of-the-art methods in each community (Mo-Att+ST-GAT) can improve results, the outputs for the naive joint learning method SC-MPF proves that the tasks of pose and trajectory forecasting are interrelated and results can be significantly improved when they are modeled jointly. Finally, the TRiPOD shows its superiority by jointly modeling the two tasks and incorporating effectively different levels of historical and futures interactions in the scene and allowing the model to be aware of the possibility of joint invisibility.

B. Ablation Study. In particular, we examine each component's contributions in TRiPOD by performing an ablation study on the 3DPW in Table 2. The first two rows are the results for the SC-MPF baseline that uses a max-pooling social operation, and the second set of results (baselines 1 to 4) are for experiments in which each component is added to the model one by one, showing the effect of each module in the final performance. The results indicate that every module improves the prediction results which evidences the benefits of exploiting different levels of semantic from the scene both in observation and prediction. We also performed an ablation study on the number of iterations for message pass-

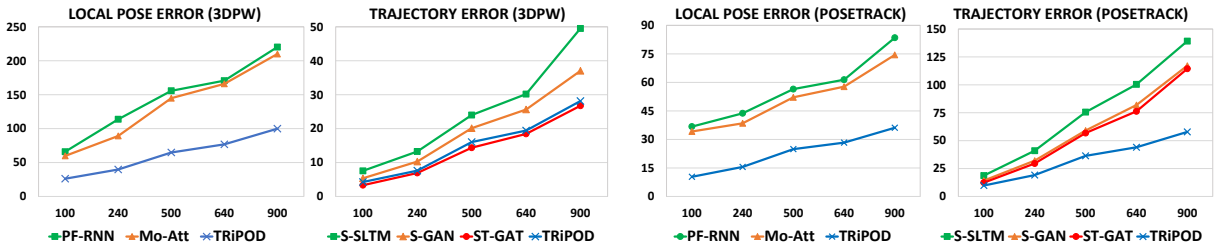

Figure 5. The VIM error rate in each pose dynamic and trajectory forecasting problem separately, for 3DPW and PoseTrack in different times.

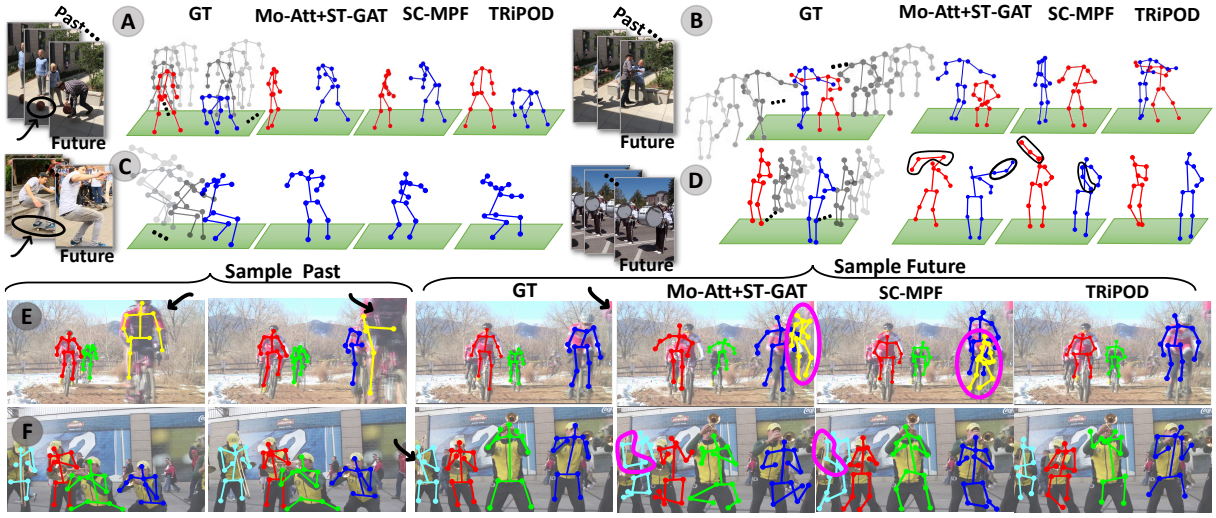

Figure 6. Qualitative results of TRiPOD, SC-MPF and Mo-Att+ST-GAT. (A) and (B) are samples taken from 3DPW and the others are from PoseTrack.

ing and the results reported in this table are based on three iterations (the results for this ablation are available in the supplementary material). Finally, the best performance is obtained by training using a curriculum learning scheme.

5.2. Qualitative Results

To better understand the contribution of the TRiPOD model in improving the understanding of different interactions in the scene, occlusion, or termination of pose existence, we visualize the prediction results for a number of samples, comparing the TRiPOD against SC-MPF and Mo-Att+ST-GAT outputs. Fig. 6. (*A*) and (*C*) illustrates cases that evidence the effect of interpreting the interactions between humans and objects in the scene. Being aware of the objects and the person's history, the model could effectively predict the final pose in the future very close to that of the truth compared to the baselines' prediction. Similarly, case (*B*) shows the same effect when the interactions between humans are modeled effectively in TRiPOD. Sample (*D*) further evidences the importance of the model being aware to estimate occlusion. Instead of outputting improper outlier predictions (joints in black curves in (*D*)), the model is capable of recognizing occlusion cases with the help of both occlusion handling indicators and also the interpretation of the object and its location in the scene. Finally, the two bottom cases (*E* and *F*) show an agent leaving the scene or the joints being out of the camera sight. TRiPOD is favorably capable of handling such cases. The pink curves indicate such faulty predictions in sample future prediction.

6. Conclusion

In this paper, we proposed a model for joint human pose dynamics and trajectory forecasting in-the-wild. Instead of only paying attention to the individual's history, our model considers different levels of semantics and interactions in the scene by attentionally modeling skeleton pose, social and human-object interactions through different graphs, and incorporating global context. The model also reinforces the future predictions, letting them be socially inter-correlated in the future in each time-step. Our method is also able to handle occlusion and pose disappearance cases. The accumulative error problem in long-term sequences is effectively handled through training model in a curriculum concept. Finally, we introduce a benchmark and relevant metrics to jointly solve the pose dynamics and trajectory forecasting problem in more realistic scenarios. Our experiments demonstrated that our TRiPOD model outperforms state-of-the-art methods in this problem. Directions for future works can be defined as incorporating 3D information (when camera parameters are available) and considering multi-modal future predictions.

References

[1] Vida Adeli, Ehsan Adeli, Ian Reid, Juan Carlos Niebles, and Hamid Rezatofighi. Socially and contextually aware human motion and pose forecasting. *IEEE Robotics and Automation Letters*, 5(4):6033–6040, 2020. 3, 5, 6, 7

[2] Emre Aksan, Peng Cao, Manuel Kaufmann, and Otmar Hilliges. A spatio-temporal transformer for 3d human motion prediction. *arXiv e-prints*, pages arXiv–2004, 2020. 2

[3] Alexandre Alahi, Kratarth Goel, Vignesh Ramanathan, Alexandre Robicquet, Li Fei-Fei, and Silvio Savarese. Social lstm: Human trajectory prediction in crowded spaces. In *CVPR*, pages 961–971, 2016. 2, 6, 7, 12

[4] Alexandre Alahi, Vignesh Ramanathan, and Li Fei-Fei. Socially-aware large-scale crowd forecasting. In *CVPR*, 2014. 2

[5] Sadegh Aliakbarian, Fatemeh Sadat Saleh, Mathieu Salzmann, Lars Petersson, and Stephen Gould. A stochastic conditioning scheme for diverse human motion prediction. In *CVPR*, pages 5223–5232, 2020. 2

[6] Javad Amirian, Jean-Bernard Hayet, and Julien Pettré. Social ways: Learning multi-modal distributions of pedestrian trajectories with gans. In *CVPRW*, pages 0–0, 2019. 2

[7] Mykhaylo Andriluka, Umar Iqbal, Eldar Insafutdinov, Leonid Pishchulin, Anton Milan, Juergen Gall, and Bernt Schiele. Posetrack: A benchmark for human pose estimation and tracking. In *CVPR*, pages 5167–5176, 2018. 6, 12

[8] Emad Barsoum, John Kender, and Zicheng Liu. Hp-gan: Probabilistic 3d human motion prediction via gan. *CVPRW*, 2018. 2

[9] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48, 2009. 5

[10] Yujun Cai, Lin Huang, Yiwei Wang, Tat-Jen Cham, Jianfei Cai, Junsong Yuan, Jun Liu, Xu Yang, Yiheng Zhu, Xiaohui Shen, et al. Learning progressive joint propagation for human motion prediction. In *ECCV*, 2020. 2

[11] Zhe Cao, Hang Gao, Karttikeya Mangalam, Qi-Zhi Cai, Minh Vo, and Jitendra Malik. Long-term human motion prediction with scene context. In *ECCV*, 2020. 2, 3

[12] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, pages 6299–6308, 2017. 6

[13] Yu-Wei Chao, Jimei Yang, Brian Price, Scott Cohen, and Jia Deng. Forecasting human dynamics from static images. In *CVPR*, 2017. 2

[14] Hsu-kuang Chiu, Ehsan Adeli, Borui Wang, De-An Huang, and Juan Carlos Niebles. Action-agnostic human pose forecasting. In *WACV*, 2019. 2, 3

[15] Enric Corona, Albert Pumarola, Guillem Alenya, and Francesc Moreno-Noguer. Context-aware human motion prediction. In *CVPR*, pages 6992–7001, 2020. 2, 3

[16] Liangji Fang, Qinhong Jiang, Jianping Shi, and Bolei Zhou. Tpnet: Trajectory proposal network for motion prediction. In *CVPR*, pages 6797–6806, 2020. 3

[17] Tharindu Fernando, Simon Denman, Sridha Sridharan, and Clinton Fookes. Soft+ hardwired attention: An lstm framework for human trajectory prediction and abnormal event detection. *Neural networks*, 2017. 2

[18] Katerina Fragkiadaki, Sergey Levine, Panna Felsen, and Jitendra Malik. Recurrent network models for human dynamics. In *ICCV*, pages 4346–4354, 2015. 2

[19] Partha Ghosh, Jie Song, Emre Aksan, and Otmar Hilliges. Learning human motion models for long-term predictions. *International Conference on 3D Vision*, 2017. 2

[20] Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals, and George E Dahl. Neural message passing for quantum chemistry. *ICML*, 2017. 4

[21] Francesco Giuliari, Irtiza Hasan, Marco Cristani, and Fabio Galasso. Transformer networks for trajectory forecasting. *arXiv preprint arXiv:2003.08111*, 2020. 3

[22] Agrim Gupta, Justin Johnson, Li Fei-Fei, Silvio Savarese, and Alexandre Alahi. Social gan: Socially acceptable trajectories with generative adversarial networks. In *CVPR*, pages 2255–2264, 2018. 2, 6, 7, 12

[23] Agrim Gupta, Justin Johnson, Li Fei-Fei, Silvio Savarese, and Alexandre Alahi. Social gan: Socially acceptable trajectories with generative adversarial networks. In *CVPR*, pages 2255–2264, 2018. 2

[24] Dirk Helbing and Peter Molnar. Social force model for pedestrian dynamics. *Physical review E*, 51(5):4282, 1995. 2

[25] Jonathan Ho and Stefano Ermon. Generative adversarial imitation learning. In *Advances in neural information processing systems*, pages 4565–4573, 2016. 2

[26] Yue Hu, Siheng Chen, Ya Zhang, and Xiao Gu. Collaborative motion prediction via neural motion message passing. In *CVPR*, pages 6319–6328, 2020. 3

[27] Yingfan Huang, Huikun Bi, Zhaoxin Li, Tianlu Mao, and Zhaoqi Wang. Stgat: Modeling spatial-temporal interactions for human trajectory prediction. In *ICCV*, pages 6272–6281, 2019. 2, 6, 7, 12

[28] Yingfan Huang, Huikun Bi, Zhaoxin Li, Tianlu Mao, and Zhaoqi Wang. Stgat: Modeling spatial-temporal interactions for human trajectory prediction. In *ICCV*, pages 6272–6281, 2019. 3, 4

[29] 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, 2014. 5

[30] Boris Ivanovic and Marco Pavone. The trajectron: Probabilistic multi-agent trajectory modeling with dynamic spatiotemporal graphs. In *ICCV*, pages 2375–2384, 2019. 3

[31] Ashesh Jain, Amir R Zamir, Silvio Savarese, and Ashutosh Saxena. Structural-rnn: Deep learning on spatio-temporal graphs. In *CVPR*, pages 5308–5317, 2016. 2, 3

[32] Łukasz Kidziński, Bryan Yang, Jennifer L Hicks, Apoorva Rajagopal, Scott L Delp, and Michael H Schwartz. Deep neural networks enable quantitative movement analysis using single-camera videos. *Nature communications*, 11(1):1–10, 2020. 1

[33] Vineet Kosaraju, Amir Sadeghian, Roberto Martín-Martín, Ian Reid, Hamid Rezatofighi, and Silvio Savarese. Social-bigat: Multimodal trajectory forecasting using bicycle-gan

and graph attention networks. In *NeurIPS*, pages 137–146, 2019. 2, 3, 4

[34] Parth Kothari, Sven Kreiss, and Alexandre Alahi. Human trajectory forecasting in crowds: A deep learning perspective. *arXiv preprint arXiv:2007.03639*, 2020. 2

[35] Namhoon Lee, Wongun Choi, Paul Vernaza, Christopher B Choy, Philip HS Torr, and Manmohan Chandraker. Desire: Distant future prediction in dynamic scenes with interacting agents. In *CVPR*, pages 336–345, 2017. 2

[36] 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 *CVPR*, pages 214–223, 2020. 3

[37] Wen Liu, Weixin Luo, Dongze Lian, and Shenghua Gao. Future frame prediction for anomaly detection–a new baseline. In *CVPR*, pages 6536–6545, 2018. 1

[38] Karttikeya Mangalam, Ehsan Adeli, Kuan-Hui Lee, Adrien Gaidon, and Juan Carlos Niebles. Disentangling human dynamics for pedestrian locomotion forecasting with noisy supervision. *WACV*, 2020. 1, 3

[39] Karttikeya Mangalam, Harshayu Girase, Shreyas Agarwal, Kuan-Hui Lee, Ehsan Adeli, Jitendra Malik, and Adrien Gaidon. It is not the journey but the destination: Endpoint conditioned trajectory prediction. *ECCV*, 2020. 2

[40] Wei Mao, Miaomiao Liu, and Mathieu Salzmann. History repeats itself: Human motion prediction via motion attention. In *ECCV*, 2020. 2, 5, 6, 7, 12

[41] Julieta Martinez, Michael J. Black, and Javier Romero. On human motion prediction using recurrent neural networks. In *CVPR*, 2017. 2, 5, 6, 7, 12

[42] Julieta Martinez, Rayat Hossain, Javier Romero, and James J Little. A simple yet effective baseline for 3d human pose estimation. In *ICCV*, 2017. 3

[43] Ramin Mehran, Alexis Oyama, and Mubarak Shah. Abnormal crowd behavior detection using social force model. In *CVPR*, pages 935–942, 2009. 2

[44] Abduallah Mohamed, Kun Qian, Mohamed Elhoseiny, and Christian Claudel. Social-stgcnn: A social spatio-temporal graph convolutional neural network for human trajectory prediction. In *CVPR*, pages 14424–14432, 2020. 3

[45] Stefano Pellegrini, Andreas Ess, Konrad Schindler, and Luc Van Gool. You'll never walk alone: Modeling social behavior for multi-target tracking. In *ICCV*. 2

[46] Davis Rempe, Leonidas J Guibas, Aaron Hertzmann, Bryan Russell, Ruben Villegas, and Jimei Yang. Contact and human dynamics from monocular video. In *ECCV*, pages 71–87. Springer, 2020. 2

[47] Alexandre Robicquet, Amir Sadeghian, Alexandre Alahi, and Silvio Savarese. Learning social etiquette: Human trajectory understanding in crowded scenes. In *ECCV*, pages 549–565, 2016. 2

[48] Christoph Rösmann, Malte Oeljeklaus, Frank Hoffmann, and Torsten Bertram. Online trajectory prediction and planning for social robot navigation. In *2017 IEEE International Conference on Advanced Intelligent Mechatronics (AIM)*, pages 1255–1260. IEEE, 2017. 1

[49] 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 *CVPR*, 2019. 2

[50] Tim Salzmann, Boris Ivanovic, Punarjay Chakravarty, and Marco Pavone. Trajectron++: Dynamically-feasible trajectory forecasting with heterogeneous data. *ECCV*, 2020. 3

[51] Dominic Schuhmacher, Ba-Tuong Vo, and Ba-Ngu Vo. A consistent metric for performance evaluation of multi-object filters. *IEEE transactions on signal processing*, 56(8):3447–3457, 2008. 6

[52] Hao Sun, Zhiqun Zhao, and Zhihai He. Reciprocal learning networks for human trajectory prediction. In *CVPR*, pages 7416–7425, 2020. 3

[53] Shuai Tang, Mani Golparvar-Fard, Milind Naphade, and Murali M Gopalakrishna. Video-based motion trajectory forecasting method for proactive construction safety monitoring systems. *Journal of Computing in Civil Engineering*, 34(6):04020041, 2020. 1

[54] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, pages 5998–6008, 2017. 2

[55] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. *International Conference on Learning Representations*, 2018. 3, 4, 6

[56] 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 *ECCV*, pages 601–617, 2018. 6, 12

[57] Jacob Walker, Kenneth Marino, Abhinav Gupta, and Martial Hebert. The pose knows: Video forecasting by generating pose futures. In *ICCV*, pages 3352–3361, 2017. 2

[58] Borui Wang, Ehsan Adeli, Hsu kuang Chiu, De-An Huang, and Juan Carlos Niebles. Imitation learning for human pose prediction. In *ICCV*, 2019. 2

[59] Chengxin Wang, Shaofeng Cai, and Gary Tan. Graphtcn: Spatio-temporal interaction modeling for human trajectory prediction. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3450–3459, 2021. 3

[60] Jack M Wang, David J Fleet, and Aaron Hertzmann. Gaussian process dynamical models for human motion. *IEEE transactions on pattern analysis and machine intelligence*, 30(2):283–298, 2008. 2

[61] Di Wu and Ling Shao. Leveraging hierarchical parametric networks for skeletal joints based action segmentation and recognition. In *CVPR*, pages 724–731, 2014. 2

[62] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019. 6

[63] Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? In *ICLR*, 2019. 13

[64] 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, 2011. 2

[65] Xinchen Yan, Akash Rastogi, Ruben Villegas, Kalyan Sunkavalli, Eli Shechtman, Sunil Hadap, Ersin Yumer, and

Honglak Lee. Mt-vae: Learning motion transformations to generate multimodal human dynamics. In *ECCV*, pages 265–281, 2018. 2

[66] Ye Yuan and Kris Kitani. Dlow: Diversifying latent flows for diverse human motion prediction. In *ECCV*, 2020. 2

[67] Long Zhao, Xi Peng, Yu Tian, Mubbasir Kapadia, and Dimitris N Metaxas. Semantic graph convolutional networks for 3d human pose regression. In *CVPR*, pages 3425–3435, 2019. 3

[68] Yi Zhou, Zimo Li, Shuangjiu Xiao, Chong He, Zeng Huang, and Hao Li. Auto-conditioned recurrent networks for extended complex human motion synthesis. In *ICLR*, 2018. 2

A. Discussion

Why two separated H2H and H2O graphs? In this section, we discuss the reasons for considering the human to human (H2H) and human to objects (H2O) as two different graphs. *First*, these two sources of information are naturally different and the type of information and influences obtained from them are also disparate. Therefore, considering them as similar nodes of a single graph is not intuitively a sensible practice. *Second*, densely connecting these two different types of information as a single huge graph and training them all together makes it difficult for the model to converge, increases the model's complexity and the overall computation. Besides, the quality of the final features obtained are not necessarily effective. Therefore, a better practice is to consider the H2H and H2O as two different graphs but devising a solution to effectively fuse these two sources of information and their effects (described as iterative message passing in the paper).

B. Benchmark Data Details

Here, we provide more details about the two datasets that we used and re-purposed to create our human pose dynamics and trajectory forecasting benchmark.

3D Poses in the Wild (3DPW) [56]: The recently released 3DPW is a challenging outdoor dataset captured using IMU sensors, with a moving camera and consists of 60 long video clips divided into 3 train, test and validation splits. We divided the video clips into multiple non-overlapping 30-frame shorter sequences sampling over every two frames resulting in 342 sequences and to investigate the importance of predicting pose dynamics and trajectories in complex scenarios, we only consider the multi-person sequences containing social interactions. We use the 3 provided splits, However, switched the train and test splits since the number of sequences in test have become larger after the aforementioned preprocess. The body poses are in world coordinate and the results are reported in centimeter (cm). In 3DPW, the pose annotations are represented by 3D locations of 24 body joints. Since some of the joints, such as fingers and toes, are not important for the current problem, we limit our selection to a subset of 13 main body joints including the neck, shoulders, elbows, wrists, knees, hips, and ankles. In 3DPW, we feed 1000ms of past history into the model and the goal is to predict the next 1000ms of future data.

PoseTrack [7]: The PoseTrack is a large-scale multi-person dataset which covers a diverse variety of interactions including person-person and person-object in dynamic crowded scenarios. In PoseTrack, pose annotations are provided for 30 consecutive frames centered in the middle of the sequence. The pose forecasting in this dataset is challenging because of the wide variety of human actions in real-world scenarios and the large number of individuals in each sequences with large body motions and a high number of occlusions and disappearing individuals cases. Since this dataset contains cases with huge portion of joints being invisible during the time, we perform some preprocess steps to make it practicable for the current problem. We maintained only those persons that are not completely invisible in all the observation frames (means at least some partial past history should be available for a person to enable the model forecasts its future). Moreover, there were some faulty, inaccurate annotations in the dataset that we did our best to refine them. The overall number of sequences is 516 which are from the training split of this dataset. We use 60% of these sequences for training our model and the rest were split equally for validation and test. We use a set of 14 joints in 2D space defining the poses including the head, neck, shoulders, elbows, wrists, knees, hips, and ankles. The data being used is in image coordinate and therefore the results are reported in pixel. In PoseTrack sequences, we trained our model by observing the past 560ms frames and learning to minimize the prediction error over the next 560ms.

C. Input Data Types

As mentioned in the paper, we used both the offset and absolute positions as the model's input data. We practically investigated that using both offset and absolute provides the best results. The reason is that although the offset is zero mean and improves the training process, a small error in offset prediction can deviate significantly from the absolute value in high dimensions or in a long time horizon. On the other hand, the absolute is not zero mean value but keeps offset error bounded to the absolute position. Considering both information together can recompense the mutual errors.

D. Baselines Setups

The Posetrack containing invisible joints entails some initial setups for the baselines (center pose [41, 40] or trajectory forecasting[3, 22, 27]) to make it possible for them to be trained on this dataset. For training the baselines with both datasets, pose information is first centered by subtracting the neck position from every joint and the pose dynamics forecasting methods [41, 40] are trained on the local poses of the datasets. Simultaneously, the trajectory, considered as neck positions, is also learned by the three state-of-the-art trajectory forecasting methods [3, 22, 27]. Then, during prediction, we add the trajectory predictions to the local pose to obtain the global poses (results in paper, Table 1).

Moreover, to train the baselines on the PoseTrack, which contains invisible joints, we perform a similar procedure we take for training our model which means if a joint disappears from ground-truth during training, no gradient for that joint is calculated. Besides, as the neck position is required

for centering the pose for pose dynamics forecasting baselines, we tried our best to refine the dataset manually, to have a good estimation of neck in occluded cases and for other cases that the agent leaves the scene we completely discard the pose. During back propagation we simply ignore these samples (do not calculate loss for them) and in test time, we use the centered poses obtained from refined neck as input and the output is whatever model predicts. Important to note that we use the refined data only for centering the pose for input and the evaluation is performed with the original data.

For the reported SC-MPF results in Table 1, we used the original SC-MPF code and metrics (requested from the authors). However, the PoseTrack data used in the SC-MPF paper is a very smaller subset of the dataset to ensure all joints for all persons in the selected sequences are fully visible as they did not model joint invisibility. We removed those assumptions from the input dataset, creating more realistic benchmarks, and used the whole dataset for the evaluation.

E. Experimental Settings

Regarding the objects used for H2O graph, we represent each object with four main features: 1) the extracted visual feature obtained from the detector 2) together with its location defined as the center location of the extracted bounding box, 3) the height and width of the bounding box, normalized over the sequence resolution and 4) the object class label as the final feature. The final object representations are obtained by passing these features through multiple MLP layers of sizes 5000, 1024 and 256. Similarly, The embedding dimensions of the MLP used for the context are 512 and 256. The hyper-parameters are selected through experiments on the validation set. We applied an initial learning rate of $5e^{-5}$ with a decay factor of 0.95 and an Adam optimizer and the step size of 2 frames being injected in each step of curriculum learning to train the model. The cut off value (β) is set to be 200 pixels. The GATs used are all single layer with 3 heads. Each experiment is performed three times and their average values are reported.

F. Additional Results

Here we provide the results for an ablation study on the number of steps performed in the iterative message passing. Table 3 shows the results. As expected, when the number of message passing iterations increase the performance first improves and then starts declining. This is commonly explored by prior graph-based learning literature [63], a crucial aspect of the graph-level representation learning is that node representations become refined and more global with the increase of the number of iterations. Therefore, it is essential to find the sufficient number of iterations for the best performance, as outlined herein.

Table 3. Error rate for ablation study on **3DPW** dataset (in cm) using different number of message passing iterations.

Message Passing	milliseconds					
#iterations	100	240	500	640	900	AVG
1 iteration	32.49	52.71	90.39	110.51	163.46	89.91
2 iterations	32.43	52.6	89.06	109.14	159.54	88.55
3 iterations	**31.56**	**51.97**	**86.53**	**107.52**	**153.12**	**86.14**
4 iterations	33.58	52.98	91.21	111.75	163.63	90.63

Table 4. Error rate for ablation study on **3DPW** dataset (in cm) using a sparse or dense graph as input skeleton representation.

Input representation	milliseconds					
	100	240	500	640	900	AVG
Sparse Graph	33.81	53.01	89.49	110.44	158.65	89.08
Dense Graph	**31.56**	**51.97**	**86.53**	**107.52**	**153.12**	**86.14**

We also investigated the effect of using a sparse or dense graph as the input skeleton representation, which is connecting the human joints in compliance with the nature of human body skeleton or representing them as fully connected graphs and letting the model to learn their relationships.The results for this study is illustrated in Table 4. The results indicate that the model can perform better when it learns the human joint relations by itself rather than sparse natural connections. This verifies the fact that the relationship between joints of an individual is not a simple hierarchical connection but every joint can have a segregated effect on each of the other joints directly.