TRiPOD: Human Trajectory and Pose Dynamics Forecasting in the Wild

Vida Adeli¹, Mahsa Ehsanpour², Ian Reid², Juan Carlos Niebles³, Silvio Savarese³, Ehsan Adeli³, Hamid Rezatofighi⁴

¹Ferdowsi University of Mashhad  ²University of Adelaide  ³Stanford University  ⁴Monash University

https://somof.stanford.edu

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 propose 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

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.
barely affected by the skateboarder or his skateboard. This 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, 64, 65, 67] 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, 63]. 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 interactions. 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 modeled 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 model of 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 \( X_{\mathcal{P}}^{t_{\tau}} = \{ \mathbf{x}_p^t, \mathbf{x}_p^{t+1}, \ldots, \mathbf{x}_p^{t_{\tau}} \} \), 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 \( Y_{\mathcal{P}}^{t_{\tau}+1} \) for the future \( t_{\tau} \) frames.

\[
Y_{\mathcal{P}}^{t_{\tau}+1} = \{ y_{\mathcal{P}}^{p,t_{\tau}+1}, y_{\mathcal{P}}^{p,t_{\tau}+2}, \ldots, y_{\mathcal{P}}^{p,t_{\tau}+1} \}, \quad \forall p \in \mathcal{P} \quad (1)
\]

where \( y_{\mathcal{P}}^{p,t} \in \mathbb{R}^F \). Throughout the paper, we consider any arbitrary variable, e.g. \( \phi_t \) being defined at time \( t \) and \( \phi_{t_1:t_2} = \{ \phi_{t_1}, \phi_{t_1+1}, \ldots, \phi_{t_2} \} \) 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 \( \Delta \ell \) (temporal location velocities), absolute locations \( \ell \) and joint visibility score \( s \), which is a binary value being 0 if the joint is invisible, i.e. \( \mathbf{x}_k^t = \{ (\Delta \ell_k^t(k), \ell_k^t(k), s_k^t(k)) \mid k = 1 : K \} \), where \( \Delta \ell, \ell \in \mathbb{R}^d; \quad s \in [0, 1] \). \( y_{\mathcal{P}}^{p,t} \) is defined the same as \( \mathbf{x}_k^t \)

(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, 66], 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 \( \mathbf{X}_{In} \).

\[
G_{In,t}^p = ATT_{In}(\mathbf{X}_{In}^p, \mathbf{W}_{In}) \quad (2)
\]

where \( \mathbf{W}_{In} \) is the set of the parameters of input pose attention graph. Accordingly, the output would be \( G_{In,t}^p \), 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 \( h_{en:0: \tau_{en}-1} \) as hidden states.

\[
Z = RNN_{en}(G_{In,1: \tau_{en}}, h_{en:0: \tau_{en}-1}; \mathbf{W}_{en}) \quad (3)
\]

13392
resulting in the encoded past global pose history $Z^p$.

**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., $\tau_o$. 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 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 $Z^p$ and the described object features $\mathcal{O}$ and outputs the H2O encoded interaction $\mathcal{H}^p_{H2O}$.

$$\mathcal{H}^p_{H2O} = ATT_{H2O}(Z^p, Z^{P\setminus p}, \mathcal{O}; \mathcal{W}_{H2O})$$  \hspace{1cm} (4)

where $P \setminus p$ represents all the people in the scene excluding person $p$ and $\mathcal{W}_{H2O}$ 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-the-wild, 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, $\mathcal{W}_{H2H}$ is the parameters of H2 attention graph (Red wavy links in Fig. 2).

$$\mathcal{H}^p_{H2H} = ATT_{H2H}(\mathcal{H}^p_{H2O}, \mathcal{H}^p_{H2O}; \mathcal{W}_{H2H})$$  \hspace{1cm} (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 these 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 $e_{pu} \in \{\mathcal{W}_{H2H}, \mathcal{W}_{H2O}\}$, 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.

$$m^p_{n+1} = ATT_{u \in \text{neighbors}(p)}(f^p_n, f^n_u, e_{pu}),$$  \hspace{1cm} (6)
where \( ATT \in \{ ATT_{H2O}, ATT_{H2H} \}, p \in \{ P \} \) and \( u \in \{ O, P \} \).

\[
f_{n+1}^p = U_n(f_n^p, m_{n+1}^p),
\]

(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 \( h_{dec,0}^p = f_n^p \) to generate the set of future poses recursively after applying an embedding function \( \psi \).

\[
y_{t+1}^p = \psi \left( RNN_{dec}(y_{t}^p, h_{dec,4t}^p, W_{dec}); W_{\psi} \right)
\]

(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 \( \tau_f \).

\[
h_{t+1}^p = ATT_{H2H}(h_{dec,4t+1}^p, h_{dec,4t+1}^p; W_{H2H})
\]

(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 \( \tau_f \) frames into \( \sum \omega \) sub problems where \( \omega \) 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 \( \Delta \ell \), absolute locations \( \ell \) and invisibility indicator \( s \)). For the first two, we minimize the norm (the MSE \( \ell_2 \)) error values of the ground-truth \( (\Delta \ell, \ell) \) and the prediction \( (\Delta \hat{\ell}, \hat{\ell}) \). 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 \( (L_{\ell_2}) \) is calculated for it, while the visibility loss \( (L_{BCE}) \) still penalizes the predictions. This concept is implemented by setting the value of loss to zero for that joints using a visibility mask \( M \) and a normalization factor \( \eta \). Considering \( \Theta \) as the collections of all weights, i.e. \( \Theta = (W_{In}, W_{en}, W_{H2O}, W_{H2H}, W_{dec}, W_{\psi}) \), we use the following to train our model.

\[
\Theta^* = \arg\min_{\Theta} \mathbb{E}_{p,t}[L(y_t^p, \hat{y}_t^p)]
\]

(10)

\[
L(y_t^p, \hat{y}_t^p) = \frac{1}{\eta} (L_{\ell_2}(\Delta \ell_t, \Delta \hat{\ell}_t) + L_{BCE}(\ell_t, \hat{\ell}_t)) \times M
\]

(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, available at https://somof.stanford.edu, (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.

**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, \( g \), and its equivalent in ground-truth, \( m \), three possible scenarios can be presumed (see Fig. 3): 1) Both joints are invisible. 2) One joint
is invisible in either the prediction or ground-truth. 3) Both joints are invisible. 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 \(d^V_\beta(G, Q)\) as the distance between the two Singleton sets as follows:

\[
d^V_\beta(G, Q) = d^{(\beta)}(G, Q)^2 + \beta^2 |c_g - c_q|^2, \quad (12)
\]

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 \(d^V_\beta(G, Q)\) for all three possible scenarios are shown in Fig. 3. Finally, the Visibility-Aware Metric (VAM) \(d_v\) for all persons’ \(K\) joints is

\[
d_v = \frac{1}{\alpha} \sum_{p \in P} \sum_{G, Q \in K} d^V_\beta(G^p, Q^p) \quad (13)
\]

where \(\alpha\) 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:

\[
\alpha = \sum_{p \in P} \sum_{g, q \in K} \max(c^p_g, c^p_q).
\]

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 re-train 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 same as simple MPIPE 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 dis-
appear 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 visibility 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 gap between the performance of TRiPOD and 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.

**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 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 passing 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.

---

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

<table>
<thead>
<tr>
<th>Metric</th>
<th>Center pose</th>
<th>Trajectory</th>
</tr>
</thead>
<tbody>
<tr>
<td>3DPW</td>
<td>prediction time in milliseconds</td>
<td>100</td>
</tr>
<tr>
<td>VIM (Invisibility ignored)</td>
<td>80</td>
<td>160</td>
</tr>
<tr>
<td>PoseTrack</td>
<td>prediction time in milliseconds</td>
<td>80</td>
</tr>
<tr>
<td>VAM (Invisibility considered $\beta = 200$)</td>
<td>27</td>
<td></td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Component</th>
<th>Prediction time in milliseconds</th>
<th>100</th>
<th>240</th>
<th>500</th>
<th>640</th>
<th>900</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline 1</td>
<td>$\times$</td>
<td>$T$</td>
<td>$G$</td>
<td>$\times$</td>
<td>$\times$</td>
<td>$90.96$</td>
</tr>
<tr>
<td>Baseline 2</td>
<td>$\times$</td>
<td>$G$</td>
<td>$\times$</td>
<td>$\times$</td>
<td>$33.99$</td>
<td></td>
</tr>
<tr>
<td>Baseline 3</td>
<td>$\times$</td>
<td>$G$</td>
<td>$\times$</td>
<td>$\times$</td>
<td>$32.64$</td>
<td></td>
</tr>
<tr>
<td>Baseline 4</td>
<td>$\times$</td>
<td>$G$</td>
<td>$\times$</td>
<td>$\times$</td>
<td>$32.85$</td>
<td></td>
</tr>
<tr>
<td>TRiPOD (CL)</td>
<td>$G$</td>
<td>$G$</td>
<td>$\checkmark$</td>
<td>$\checkmark$</td>
<td>$31.56$</td>
<td></td>
</tr>
<tr>
<td>TRiPOD (CL)</td>
<td>$G$</td>
<td>$G$</td>
<td>$\checkmark$</td>
<td>$\checkmark$</td>
<td>$31.04$</td>
<td></td>
</tr>
</tbody>
</table>
Figure 5. The VIM error rate in each pose dynamic and trajectory forecasting problem separately, for 3DPW and PoseTrack in different times. For the disjoint problems the VIM becomes same as the FDE metric for trajectory information.

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.

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) illustrate 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. 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 incorporating 3D information (when camera parameters are available) and considering multi-modal future predictions.

Acknowledgement. J.C. Niebles and E. Adeli would like to thank Panasonic for their support.
References


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


[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


