IEEE Robotics and Automation Letters (RAL) paper presented at the
2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 25-29, 2020, Las Vegas, NV, USA (Virtual)

Socially and Contextually Aware Human Motion and Pose Forecasting

Vida Adeli[1], Ehsan Adeli[2], Ian Reid[3], Juan Carlos Niebles[2], Hamid Rezatofighi[2,3]

***Abstract*— Smooth and seamless robot navigation while interacting with humans depends on predicting human movements. Forecasting such human dynamics often involves modeling human trajectories (global motion) or detailed body joint movements (local motion). Prior work typically tackled local and global human movements separately. In this paper, we propose a novel framework to tackle both tasks of human motion (or trajectory) and body skeleton pose forecasting in a unified end-to-end pipeline. To deal with this real-world problem, we consider incorporating both scene and social contexts, as critical clues for this prediction task, into our proposed framework. To this end, we first couple these two tasks by i) encoding their history using a shared Gated Recurrent Unit (GRU) encoder and ii) applying a metric as loss, which measures the source of errors in each task jointly as a single distance. Then, we incorporate the scene context by encoding a spatio-temporal representation of the video data. We also include social clues by generating a joint feature representation from motion and pose of all individuals from the scene using a social pooling layer. Finally, we use a GRU based decoder to forecast both motion and skeleton pose. We demonstrate that our proposed framework achieves a superior performance compared to several baselines on two social datasets.**

I. Introduction

Forecasting human motion and body pose can be conducive in many real-world problems, *e.g.*, in prediction of hazardous or anomalous behaviour for a smart surveillance system, in fine-grained anticipation of future activities, or in a collision avoidance system for an autonomous vehicle or a mobile robot navigating through a crowd.

The existing solutions for human skeleton pose forecasting focus on accurate prediction of intricate body joint movements as highly structured problem [1], [2]. These approaches often concentrate on the problem of natural body joint prediction only while ignoring the global body motion reflecting these pose changes. For example, if a person is running, it is not only important to predict a skeleton pose representing this natural behaviour in future, it is also crucial to estimate the entire body displacement consistent with the predicted skeleton motions. There exists also a parallel research thread focusing on forecasting human motion or trajectories [3], [4], by modeling it as prediction of a set of locations over time. However, this projected information may not be sufficient for the problems which require a detailed analysis of human body motions. Intuitively, the problem of human body pose forecasting and global motion prediction are inevitably correlated and should not be approached independently. In this paper, we aim to tackle these two problems in a unified framework.

This research was partially supported by Mindtree, Oppo, and Panasonic.
[1]Faculty of Engineering, Ferdowsi University of Mashhad, Mashhad, 9177948974, Iran. vida.adeli@mail.um.ac.ir
[2]Department of Computer Science, Stanford University, Stanford, CA 94035, USA {eadeli, jniebles}@cs.stanford.edu
[3]School of Computer Science, University of Adelaide, Australia {hamid.rezatofighi, ian.reid}@adelaide.edu.au

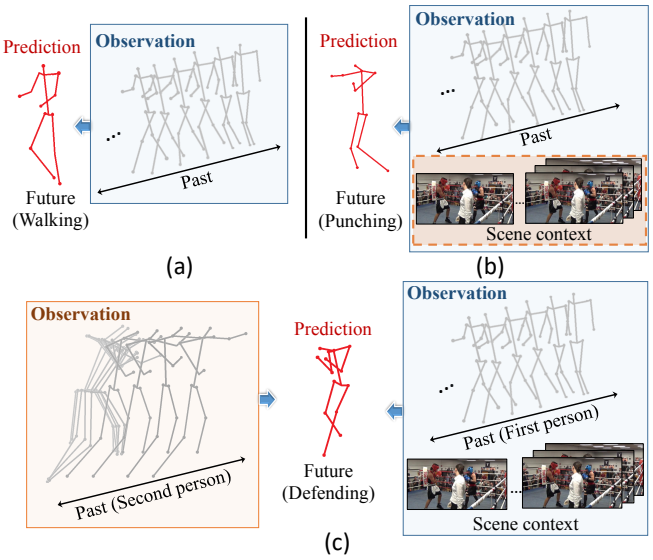

Fig. 1. It is very challenging, even for human, to observe only a history of skeleton pose to predict the future pose and behaviour. In all three scenarios (a)-(c), the same motion and skeleton pose are provided as the observation, while in (b) the scene context (as a video) and in (c) both scene contexts and social interactions (by observing the others' motion and skeleton pose) are additionally provided. It is intuitive both scene contexts and social clues are required to accurately predict the future human motion and pose.

To accurately predict human skeleton pose and global motion, a model cannot solely rely on the history of body joints locations as input data. The physical context of the scene can provide important clues for this task. For example, considering an ice skating field, it is very likely that the body pose and motion of a person in this scene represent a sliding, but not diving or jogging, activity. Therefore, incorporating this context as an informative clue into the pipeline would be helpful for tackling the problem.

The social context in a scene can also provide valuable information for a pose and motion anticipating model. In a real scene including individuals with different social connections and interactions, future gesture and pose of each person can be influenced by the other people activities and poses. For instance, consider a scene captured from a boxing game; if one of the boxers starts punching, it is predictable the opponent will hold a defending posture in near future (see Fig. 1). Therefore, it is intuitive to build a framework which encodes these social interactions.

We aim to push the existing research thread in the community for human behavior prediction one step forward toward more practical scenarios, where human global motion and skeleton pose are jointly predicted while considering all social and contextual clues in the scene. Our proposed framework jointly incorporates a) the history of human trajectory and skeleton pose in the past using a recurrent

based neural network encoder, *i.e*., Gated Recurrent Units (GRU) [5], b) the physical context of the scene by encoding a spatio-temporal representation from the visual data, *i.e*., all past frames of a video stream, using the state-of-the-art I3D backbone [6] and c) the social interaction between all individuals body and pose motion using a social pooling module. Furthermore, contrary to the majority of the prior works [2], [7], our model does not incorporate action labels during training and only builds one single action-agnostic model. We evaluate our proposed framework on two social datasets, created from NTU RGB+D 60 [8] and PoseTrack [9], and demonstrate its superior performance against relevant baselines.

In summary, the main contributions of the paper are: (1) We propound the new problem of joint prediction of human **global** motion and body **pose** and introduce a proper evaluation metric (and also a loss) to measure the performance of both tasks in a single metric. (2) We propose an end-to-end learning framework for the task, which considers both **social** and visual **context** of the scene by incorporating justifiable learning modules such as a) GRU as a temporal encoder of skeleton and body motion, b) a social pooling layer (permutation invariant function) to generate a social feature, c) I3D backbone as context encoder and d) GRU decoder to forecast global motion and body skeleton pose. (3) Our framework achieves a **superior performance** in comparison with several relevant baselines and state-of-the-art models on two social datasets, including a 2D (PoseTrack [9]) and a 3D (NTU RGB+D 60 [8]) dataset.

II. Related work

In this section, we review relevant literature on video based motion, trajectory, and pose forecasting as well as the prior work on social motion modeling.

A. Video-based forecasting

Prediction and forecasting of visual data based on videos was initially defined as extrapolating pixels values of the future frames [10]–[15]. Although these methods successfully predicted future frame(s) pixel values, it remains unclear how such a low-level future forecasting can be sufficient for analysis of events and scenes in the future. On the other hand, recent work focused on anticipating activities [16]–[22] or trajectories [23]–[25] (more details in the next subsection). Some other works have proposed predicting future semantic segmentation in videos [26]–[28]. In this work, we instead present a method that predicts the human dynamics, including global motion and local pose data. Such information can be used for detailed future action understanding.

B. Motion and pose forecasting

Human dynamics can be best modeled via global motion and detailed local joint locations (referred to as human skeleton or pose) [29]–[33]. Human dynamics are previously modelled in images [34] by two major types of methods in videos including state transition models (such as graphical models) [35], [36] or more recently sequence-to-sequence deep learning methods [1], [2], [7], [32], [37]–[39]. Chao *et al*. [34] introduced a method for pose forecasting in 3D on static images, Barsoum *et al*. [39] used Wasserstein GAN [40] in a probabilistic setting, Walker *et al*. [38] used variational autoencoders, Fragkiadaki *et al*. [41] proposed architectures based on LSTM and Encoder-Recurrent-Decoder methods, Yan *et al*. [42] and Zhao *et al*. [43] proposed methods that could predict longer into the future, Martinez *et al*. [1] introduced a designed RNN for human pose prediction, Chiu *et al*. [32] utilized a multi-layer hierarchical recurrent architecture, and Wang *et al*. proposed to use imitation learning and specifically Generative Adversarial Imitation Learning (GAIL) [44] to capture human dynamics. However, all these works only predict the local dynamics of the pose data, *i.e*., they subtract global human motion from the joint coordinates and only predict these local movements. We argue that predicting both global and local motion information at the same time has better implications in terms of translating to real-world applications. Although a more challenging task, such modeling of the problem can leverage information from both global and local motion patterns and help better prediction of realistic 2D or 3D poses. Furthermore, all these previous works aim at only predict the joint coordinates of each single human in the scene. They ignore the social forces that other humans may deduce or neglect the context of the environment.

C. Social models for motion predictions

Even before the deep learning era, modeling human social interactions was a popular research topic among the community. Traditional models use hand-crafted features and rules such as "social forces" in order to model and predict human social interactions [23], [45]–[49]. However, the usability of these hand-crafted rules is restricted to their abstract level and the domain experts' information. Recently, due to the rise of popularity in development of autonomous driving systems and social robots and also with the recent developments in model-free data-driven frameworks for learning these social interactions, the problem of social human motion forecasting has received significant attention from both academia and industry. The main idea behind the modern socially-aware motion prediction approaches is to use a recurrent based neural networks along with a social pooling layer (a symmetrical function) to encode a social representation from this spatiotemporal data before forecasting the future path for each person [3], [4], [50]–[53]. However the problem considered in these works is only about the task of social trajectory forecasting, *i.e*., prediction of a set of 2D locations over time. This simplified problem is lacking a detailed representation for this human body motions, which is crucial for human activity and behaviour anticipation. In a recent work, Joo *et al*. [54] proposed a method for predicting social signals in triadic social interaction scenarios, in which the social signals (like facial expression and body position) of two other individuals are processed during the same time window to estimate the target person's signals.

In this paper, we aim to tackle this shortage by unifying the prediction of human motion and pose while incorporating social and scene contexts into the proposed framework.

III. Social motion and pose forecasting model

We define the task of motion and pose forecasting on a video with T frames containing N persons, where each person includes L body joints. Note that we drop the index of the video for a better readability. Throughout the paper,

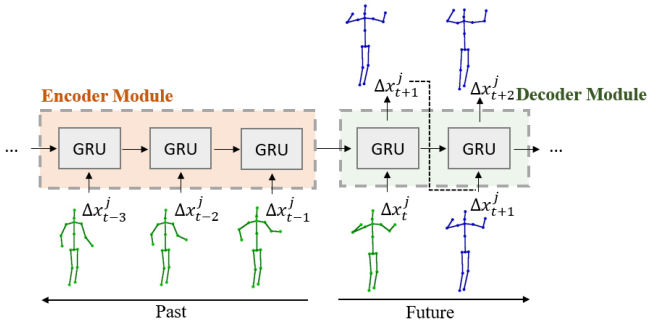

Fig. 2. Sequence-to-sequence modeling of the human pose forecasting problem, using gated recurrent units. The network consists of an encoder module that encodes the historical (past) sequence of poses and a decoder module to recover the future poses. This modeling is for each human in the scene separately (*i.e.*, it is blind to social or context-related forces).

we use $\mathbf{X}_o^j = \left(x_1^j, x_2^j, \ldots, x_t^j\right)$ to represent all observed values of body joint locations for j^{th} person in a global coordinate up to time t, where $x_t^j \in \mathbb{R}^L$ is a vector of all body joint coordinates at each time. Similarly, we use $\mathbf{X}_f^j = \left(x_{t+1}^j, x_{t+2}^j, \ldots, x_{t+T}^j\right)$ to demonstrate the future joints positions, up to $T-(t+1)$ time-steps into the future.

A. Forecasting motion and pose jointly

The dominant trend for forecasting the human pose is to predict the future position of body joints relative to the body center (neck or torso) [1], [2], [7], [32]. Intuitively, it is an easier task for a model to only learn the body joint movements respected to a centered body skeleton, making the predicted values (as offsets), zero mean variables. Similarly, with the same logic, the global body motion is often modelled as an offset prediction task from the last predicted position instead of predicting their absolute coordinates [52]. To forecast both components jointly, one trivial extension is to estimate an offset value for each component independently, *i.e.*, one offset for the entire body motion and one set of offsets for each body joints motion relative to the body. However, there are two potential problems with this approach: I) this setting will make the behaviour of joints and body motions uncorrelated since their values are predicted independently, and II) the evaluation and ranking of overall task will become complicated as each component has its own error. To address this, we first suggest a very simple, but an intuitive, metric which encodes both source of errors jointly by a single value. To this end, we consider the real-world setting of movements of joints in space and time in *the original space*. Therefore, we define the evaluation metric as a norm error, *e.g.*, the MSE or ℓ_2 error, of the ground-truth ($\mathbf{X}_f$) and predicted ($\hat{\mathbf{X}}_f$) absolute future locations of the joints with respect to *their global coordinate*, *e.g.*, $\text{MSE}(\mathbf{X}_f, \hat{\mathbf{X}}_f)$. This metric encodes both global body motion and joint movements respect to the body jointly. However, for the loss as offset prediction is easier task, we minimize the error between the offsets values of the predicted joint locations and the ground-truth joint locations, *i.e.*, $\mathcal{L} = \mathbb{E}_j \left(\text{MSE}(\Delta\mathbf{X}_f^j, \Delta\hat{\mathbf{X}}_f^j)\right)$.

In order to encode and decode this temporal data, we apply sequence-to-sequence model. To this end, we use a previous state-of-the-art recurrent model [1] with GRU cells to implement the encoder and decoder modules. In [1], the encoder receives the past pose states (as velocities/offsets of joints movements with respect to the body location) up to time t and the decoder outputs the predicted pose states for the time steps $t+1$ to T (Fig. 2).

Our framework uses exactly the same encoder and decoder architecture, *i.e.* identical number of parameters and layers, as [1]; However as shown in Fig. 3, we feed the encoder for each person with the person's body and pose motion, $\mathbf{X}_o^j$, to generate a person-specific representation of the historical (past) pose and motions, denoted by $h^j \in \mathbb{R}^K$, where K is the dimension of the feature. These embedding representations are shown by colored and textured boxes, *e.g.*, green, orange, and purple, immediately after encoder module in Fig. 3. Then, this representation is concatenated with social and context features before being fed into the decoder, in order to forecast body motion and pose as $\Delta\mathbf{X}_f^j$.

B. Social and context modules

To incorporate social interaction among the people in the scene, we need to encode a collective feature representation from everyone's body motion and pose. We can use their encoded features, h^j. This feature cannot be simply attained by concatenation of individuals encoded features, h^j, as this representation should not be sensitive to the way each person is ordered. Any ordering of the same set of instance input should generate an identical feature representation. To design this module, similar to all set encoding problems, *e.g.*, point clouds [55], [56] and social trajectories [4], [23], we need a symmetrical (permutation invariant) function, $\mathcal{H}(\cdot)$, applied on all h^j embeddings (as a set) to generate a permutation invariant joint representation $S \in \mathbb{R}^K$, *i.e.*,

$$S = \mathcal{H}(\{h^1, \cdots, h^j, \cdots, h^N\}), \tag{1}$$

The family of these symmetric functions is vast and many mathematical functions has this symmetrical property. However, the most popular ones for being incorporated easily into deep learning pipelines are *max*, *sum* and *average* pooling layers used in [4], [23], [55], [56]. We explore the choice of some of these functions in our experiments (see Table II). Then, the pooled social feature, S, denoted by the red boxes in Fig. 3 is concatenated with features of each person, h^j.

To extract the context features, we encode a spatio-temporal representation from the video. To this end, we use a pre-trained I3D model [6], applied on the RGB video, from which the individual poses were extracted. This leads to another embedding vector that goes through two layers of fully connected neural networks (denoted by multi-layer perceptron, MLP, in Fig. 3). The resulting context feature vector, denoted by C (shown with a blue box in Fig. 3), are additional features shared across all persons in the scene.

Finally, the decoder module for each person j receives a socially and contextually aware feature vector, generated by the concatenation of person specific embedding h^j, the shared socially pooled features S, and the shared context features C. The decoder outputs pose, with both global and local movements encoded in it.

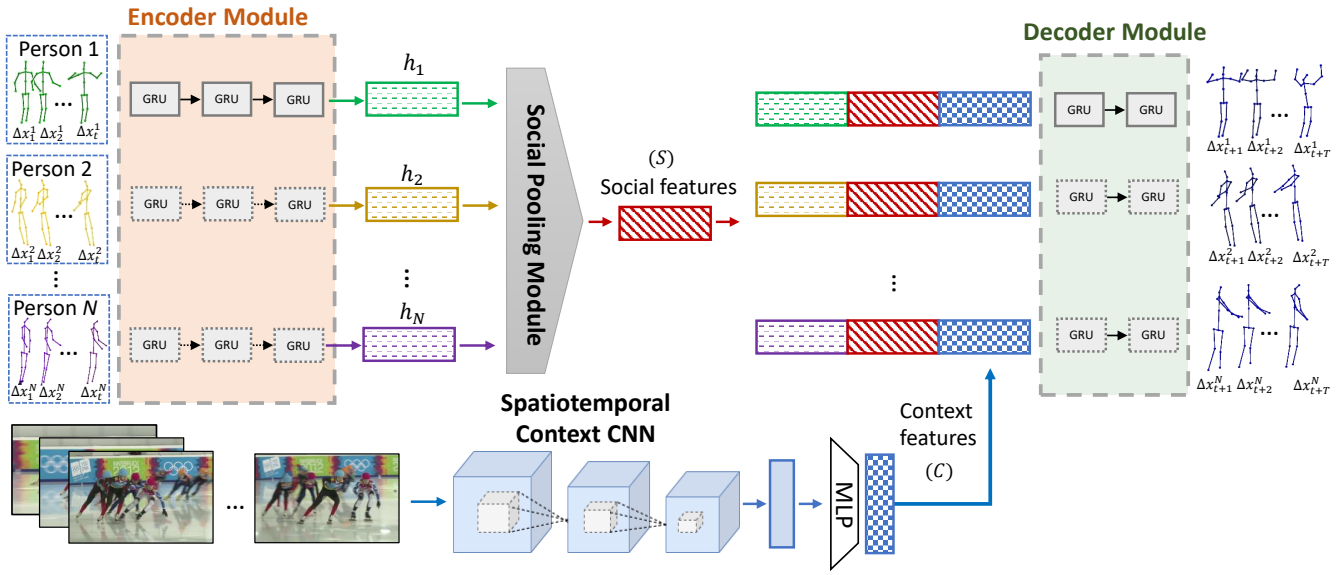

Fig. 3. Overview of the proposed socially and contextually aware human motion and pose forecasting model. Our model consists of four main modules: (1) **Encoder Module** encodes the historical (past) sequence of poses for each individual in the scene; (2) **Social Pooling Module** produces one shared socially aware feature vector by pooling (a permutation invariant operation) features across all embeddings; (3) **Spatiotemporal Context Encoding Module** uses an I3d model to extract spatiotemporal features from the videos, which are passed through two layers of fully connected neural network, MLP, to create the shared context features; (4) **Decoder Module** receives the concatenated person-specific embedding, shared social, and context features to produce the predicted poses for each future time step.

IV. Experiments

In this section we present the experimental results on two challenging datasets: NTU RGB-D 120 [8] and PoseTrack [9]. Different sets of experiments are conducted to assess the impact of our contributions and finally, the results are compared with different baseline methods.

Evaluation metric. As discussed earlier and consistent with the prior work [1], [7], [28], we use MSE metric (ℓ_2 distance between the ground-truth and the predicted poses at each time $i \in \{t+1, \ldots, T\}$) averaged over the number of persons N in that frame. Contrary to the previous work that center the poses (*i.e.*, remove the global body motion from joints motion), our metric evaluates the differences between the predictions and the ground-truth in *the original global space*.

A. Experimental settings

We use a sequence-to-sequence architecture as our encoder-decoder model with GRU modules, exactly the same as [1], as the RNN for both encoder and decoder. We consider a fixed hidden state dimension of 1024 for GRU modules in the encoder and for decoder, it varies depending on the selected dimension for the context feature after embedding to the MLP. We use a two layer dense network for the context features with a dropout probability of 0.7 that reduces its dimension to 256 before feeding it to the forecasting model.

The I3D model [6] pre-trained on Kinetics [57] is used as the context network to extract the spatio-temporal context features from video frames. The I3D 1024-dimensional feature vectors are then used as the context features in our framework. The hyper-parameters are selected through experiments on a validation set. We used a learning rate with initial value of $5e^{-4}$ and a learning rate decay factor of 0.95 for training the model with the Adam optimizer. For all the experiments, the training is performed over all activity types resulting in a single action agnostic model. The best model is selected with early stopping on the validation set.

B. Datasets

There are several commonly used benchmark datasets for human pose forecasting such as Human 3.6M [58] and Penn action dataset [59]. However, these publicly available datasets being used by the previous work [1], [32], [38], [41], [60] focus on single isolated individuals and do not entail social interactions. Moreover, most of these datasets are short-length video sequences, that makes it hard to encode the social behaviour of individuals. To this end, to test the performance of our model for social human pose forecasting, we run extensive experiments on NTU RGB+D 60 and PoseTrack datasets that consist of multiple persons interacting with each other to complete different actions.

PoseTrack: The PoseTrack is a large-scale multi-person dataset based on the MPII Multi-Person benchmark. The dataset covers a diverse variety of interactions including person-person and person-object in dynamic crowded scenarios. 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 occlusions. In each sequence, we select those individuals that are present in all frames. For experimenting on this dataset, we trained our model by observing the past 15 frames as the history of each person and forecast the future 15 frames. We use a set of 14 joints in 2D space as the human pose, 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. Also, we use 60% of sequences in training split for training our model.

TABLE I

ERROR (MSE) ON **POSETRACK** (IN PIXELS) AND **NTU RGB+D** (IN CM) DATASETS FOR DIFFERENT BASELINES AND OUR PROPOSED MODELS. IN EACH COLUMN, THE BEST OBTAINED RESULT IS HIGHLIGHTED WITH BOLDFACE TYPESETTING AND THE SECOND BEST IS UNDERLINED. IN THE FIRST ELEVEN EXPERIMENTS, THE FIRST WORD BEFORE THE HYPHEN NOTATION, IS AN INDICATOR OF METHOD USED FOR BODY SKELETON POSE AND THE SECOND ONE IS FOR METHOD USED AS GLOBAL MOTION (TRAJECTORY). **ZERO**: ZERO VELOCITY, **CONSTANT**: CONSTANT VELOCITY, **LOCALPOSE**: MODEL TRAINED ON THE CENTERED POSES WHETHER BY THE ORIGINAL MODEL [1] OR USING OUR SOCIAL MODULE, **SLSTM**: LEARNING GLOBAL MOTION USING SOCIAL-LSTM [3], **SGAN**: LEARNING GLOBAL MOTION USING SOCIAL-GAN [4].

Method	PoseTrack					NTU RGB + D			
	milliseconds					milliseconds			
	80	160	320	400	560	80	160	320	400
ZeroPose-ZeroMotion	153.3	263.7	432.3	473.9	563.5	14.1	22.1	34.8	40.5
ZeroPose-ConstantMotion	154.0	265.4	436.9	479.6	572.6	14.2	23.0	35.7	44.2
ConstantPose-ConstantMotion	154.4	266.6	440.4	484.3	579.5	15.3	23.6	36.1	46.8
LocalPose [1]-ZeroMotion	72.4	114.7	217.8	253.1	311.5	14.0	23.6	34.9	40.3
LocalPose [1]-ConstantMotion	68.3	109.7	206.9	249.4	306.6	14.1	24.0	34.8	43.1
LocalPose (+ Social pooling)-ZeroMotion	52.2	98.9	186.6	228.3	282.9	14.0	22.0	34.1	40.1
LocalPose (+ Social pooling)-ConstantMotion	47.2	90.2	173.5	207.7	274.6	13.9	21.8	33.8	39.2
LocalPose [1]-SGAN (social) [4]	62.1	105.6	189.9	230.8	282.6	14.9	24.7	36.1	43.9
LocalPose [1]-SLSTM (social) [3]	63.1	107.3	193.4	235.4	285.7	14.2	24.1	35.7	43.5
LocalPose (+ Social pooling)-SGAN (social) [4]	45.6	84.2	167.3	203.9	269.6	14.5	22.4	34.7	40.9
LocalPose (+ Social pooling)-SLSTM (social) [3]	45.9	84.8	168.2	204.1	268.3	14.3	22.3	34.5	40.5
Ours (JointLearning)	**43**	81.7	158.1	197.2	256.3	13.1	20.0	30.5	35.1
Ours (JointLearning+Context)	44.2	82.1	157	194.2	251.5	13.0	20.1	30.4	35.0
Ours (JointLearning+Social)	43.1	81	154.3	191	246.7	**12.8**	**19.6**	29.9	34.3
Ours (JointLearning+Social+Context)	**43**	**80.1**	**152.3**	**188.4**	**243.2**	13.1	19.8	**29.7**	**34.1**

The rest were split equally for validation and test.

NTU RGB+D 60: The NTU dataset contains both single-person actions and mutual actions. We selected mutual sequences for our experiments, which are in 11 different action categories including *punching*, *kicking*, *pushing*, *pat on back*, *point finger*, *hugging*, *giving object*, *touch pocket*, *handshaking*, *walking toward*, and finally *walking apart* actions. In this dataset, each pose is represented by 3D locations of 25 major body joints at each frame. Since some of the joints, such as finger positions, are too fine-grained and are not important in our problem, we use only 13 body joints for our experiments. The pose locations are in camera coordinate and hence the results are reported in centimeter (cm). As suggested in [8], we utilize the standard cross-subject evaluation in which the 40 subjects are split into different groups of training and testing. Also, for validation set we divided the test set into two splits. The model is trained by observing poses in 40 frames and then forecasts for the next 10 frames. Finally, in contrast with previous work [1], [2], [41] that propose action-specific models, we train a single action agnostic model that does not require any prior knowledge about the action categories.

C. Baselines

Since no other work consider the problem of social pose forecasting, there is no previously published work on social datasets. As we are considering two concepts of pose and global motion, to show the effectiveness of modeling them jointly, we conduct some experiments separating these concepts and compare our method against following baselines. Three sets of baseline experiments are conducted (The first three sets in Table I. In the first set of baselines no learning is applied in the method, whereas in the second set, pose information is trained while there is no learning for the global motion. These baselines provide of with a sanity check of our data and show how well can a trivial baseline with no learning perform on our datasets. For the third set of experiments both pose and global motion are learned but separately, using the state-of-the-art methods. Then, in all the baselines, the predicted global motion is added to the predicted pose for estimating metric values in every time step. As the final set of experiments, our method jointly learns and predicts the global motion and pose, involving the social interactions and context.

Zero Pose and Zero Motion velocity, denoted as *ZeroPose-ZeroMotion*, assumes that all the future predictions are identical to the very last seen pose and thus outputs the last observed pose for the future in a same location with no global motion. In other words, pose at time t remains with no change in joint and displacement. The zero velocity baseline for pose has been used by many other methods for comparison and demonstrated to be a high performance baseline that is hard to outperform [1], [41], [61].

Zero Pose and Constant Motion velocity, denoted as *ZeroPose-ConstantMotion*, considers that in future prediction the joints relative position remains fixed and the whole body moves with a constant velocity.

Constant Pose and Constant Motion velocity, denoted as *ConstantPose-ConstantMotion*, assumes that both pose and global trajectory move with a constant velocity model.

Local pose forecasting with zero/constant global trajectory. In order to show that it is fundamental to model both the movements of the joints and the global trajectory jointly, we experiment another baseline that uses the centered human poses as its input, just like almost all previous pose forecasting methods that exclude global body motion. To do so, we center the poses by subtracting the neck position and train a model to predict centered poses. Then during prediction, we add the zero or constant velocity to create the global trajectory. We report the results by training the local pose (i) using the vanila model proposed in [1], or (ii) by incorporating our social pooling module into this model [1].

TABLE II

EFFECT OF DIFFERENT SOCIAL POOLING STRATEGIES IN OUR SOCIAL+CONTEXT MOTION AND POSE FORECASTING MODEL.

Method	milliseconds				
	80	160	320	400	560
Average Pooling	43.9	81.3	153.6	190.5	246.8
Sum Pooling	44.1	81.9	154.1	190.9	247.2
Max Pooling	**43**	**80.1**	**152.3**	**188.4**	**243.2**

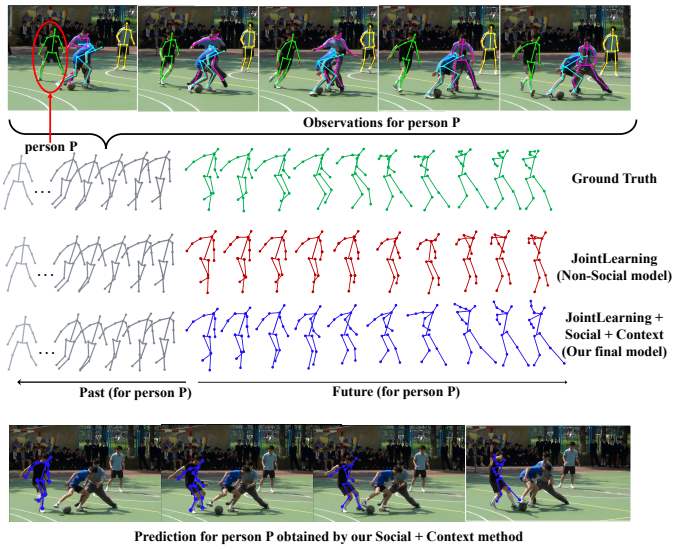

Fig. 4. Visualization of the first 10 pose predictions for a sample sequence from the action playing football in PoseTrack dataset. The first row contains a number of sample observations that the model receives as input including the history of all persons in the scene. The next three rows show historical and future poses for Person P (circled out in the top row). The gray poses are ground-truth data of the history of action. The green, red and blue poses are the future poses of ground-truth, JointLearning Non-social method and our final method (JointLearning+Social+Context), respectively. The last row shows the predicted poses by our social+context model on the frames.

Local pose forecasting with SGAN/SLSTM global trajectory forecasting. In this set of baselines, pose information is trained exactly in the same manner as previous set (centered with vanila and social model), while the global motion is also learned by two state-of-the-art methods in trajectory prediction which are *Social GAN* [4] and *Social LSTM* [3]. These baselines can clearly show the difference between performance of forecasting while both pose and trajectory being trained seperately or jointly.

Our joint learning model (with or without context) baseline uses exactly the same model as [1] followed by our proposed loss for learning to predict body pose and global motion, jointly. For our joint learning model (i.e., non-social model) with context, we also integrate the context feature before decoding the future output. However, these baselines do not contain the social module.

Our joint learning social model without context. In this baseline we intend to investigate the effect of scene context on the proposed framework. The model in this baseline excludes the spatiotemporal context encoding module of our final model and the other modules are remained similarly.

D. Comparison with the baselines

We compare our proposed method against different baselines on both PoseTrack and NTU RGB+D datasets in Table I. It is important to note that in all baselines, the results are reported as the previously described metric in global mode. In the first set of experiments, denoted as *ZeroPose-ZeroMotion*, *ZeroPose-ConstantMotion*, and *ConstantPose-ConstantMotion*, we considered the pose and global trajectory separately and there is no learning involved in these baselines. The reason that the zero or constant baselines are performing more poorly in the PoseTrack dataset than in the NTU is that the PoseTrack contains more complex and realistic scenarios with severe body motions in comparison to the NTU dataset that is recorded in a controlled laboratory settings. Quantitative evaluations on NTU dataset in previous studies [41] confirmed this fact that most existing methods could be outperformed by zero-velocity and constant velocity baselines. However, by considering both body pose and global motion jointly, even without incorporating the social and scene context information, we could outperform these baselines in both datasets. Even the vanila LocalPose model results in better predictions in PoseTrack. However, this superiority for the vanila LocalPose model is marginal in NTU due to the aforementioned limitations in this dataset.

In the second set of experiments, denoted as *LocalPose-(Zero/Constant) Motion*, the pose is centered and the model learns to predict a centred pose [1]. Then, we place the predicted poses in the position of time t (zero velocity on the neck) or we move the whole body with constant velocity. As can be seen, modeling both pose and global motion (Our joint learning results) outperforms each of these methods. This shows that adding a constant velocity model to the centered pose model cannot compensate the global motion term and so they are highly correlated to each other and should be modeled jointly. In the third set of experiments both person centric pose and global motion are learned by state-of-the-art pose forecasting (non-social [1] and social) and trajectory forecasting models (SGAN [4] and SLSTM [3]) separately and the final global pose is obtained by adding the global motion prediction to the predicted local pose. These experiments also demonstrates that pose is better predicted when jointly predicting the global motion and pose. Inspecting precisely, we can observe that methods with constant or zero motion do not considerably differ from trajectory learnable methods in terms of results, in NTU dataset. This is due to small amount of global motion in NTU, which is recorded in constraint laboratory settings. Even in some cases, we see that the complexity of the trajectory model (SGAN/SLSTM) results in worse predictions. Despite this, the results show that the use of joint learning with social and context terms improved the final results in NTU.

As the forth set of experiments, we consider our joint learning (non-social) model, which is an extension of the state-of-the-art method [1] but with our new loss defined on both local pose and global motion jointly. Then, we add the context and social module in order to investigate the effects of each the social and context information in our model. As can be seen, by using the social and context information of the scene and modeling the pose and global motion jointly, our model achieves the highest performance on both NTU and real-world PoseTrack dataset. The results indicate that not only utilizing the social interactions, but also the context information could improve the prediction result.

TABLE III

MSE ACTION LEVEL ERRORS (IN CM) ON THE 11 MUTUAL ACTION CATEGORIES OF NTU RGB+D 60 DATASET. OURS (JOINTLEARNING) IS THE EXTENDED VERSION OF [1] WHICH USES THE LOCAL POSE AND GLOBAL MOTION JOINTLY IT DOES NOT CONSIDER THE SOCIAL TERM.

	Punching				Kicking				Pushing			
milliseconds	80	160	320	400	80	160	320	400	80	160	320	400
Ours (JointLearning)	15.7	24.5	38.3	44.08	16.9	28.1	45.9	53.3	19.7	30.8	52.9	61.7
Ours (JointLearning+Social+Context)	**15.6**	**22.4**	**37.8**	**44.1**	**16.7**	**27.8**	**44.9**	**52.1**	**19.3**	**29.8**	**50.4**	**58.4**
	Pat on back				Point finger				Hugging			
milliseconds	80	160	320	400	80	160	320	400	80	160	320	400
Ours (JointLearning)	**5.4**	**7.9**	**10.7**	**12.2**	**5.9**	**8.2**	**10.2**	**12.0**	25.6	36.9	51.3	56.6
Ours (JointLearning+Social+Context)	5.5	**7.9**	10.8	12.3	**5.9**	8.4	10.7	12.6	**25.5**	**36.7**	**50.1**	**54.9**
	Giving object				Touch pocket				Handshaking			
milliseconds	80	160	320	400	80	160	320	400	80	160	320	400
Ours (JointLearning)	10.2	14.5	22.0	26.3	7.1	10.8	17.1	20.1	**8.0**	**11.1**	**14.5**	**16.0**
Ours (JointLearning+Social+Context)	**10.1**	**14.4**	**21.6**	**25.7**	**7.0**	**10.6**	**16.5**	**19.0**	**8.0**	11.2	14.9	16.5
	Walking towards				Walking apart				Average of all 11			
milliseconds	80	160	320	400	80	160	320	400	80	160	320	400
Ours (JointLearning)	23.6	35.7	56.3	62.9	21.8	36.6	58.7	68.6	13.1	20.0	30.5	35.1
Ours (JointLearning+Social+Context)	**23.0**	**35.0**	**54.9**	**61.3**	**21.4**	**35.7**	**57.2**	**66.7**	**13.1**	**19.8**	**29.7**	**34.1**

This claim is supported by the lower prediction errors of Social + Context model compared to the Social method. We believe the reason that the context make less improvement on NTU is the constraint environmental settings that the videos are recorded in. All videos are obtained in a laboratory setting and therefore concepts of context become impractical.

Finally, we investigated the effect of different social pooling strategies in the proposed Social + Context pose and motion forecasting models (Table II). The best results are achieved using *max* pooling strategy. In Fig. 4, we show a sample qualitative result from the action playing football, obtained by our final (JointLearning+Social+Context) model against the results for our JointLearning (Non-social) model (JointLearning), which is the extended version of [1] with loss jointly on local pose and global motion. As it can be seen in this figure, compared to JointLearning Non-Social model the predictions from our final model (blue poses) are more visually closer to the ground-truth poses (green skeletons). This improvement in predictions increase over time, when the social interaction becomes more meaningful. For example, when a person observes the tackling action by another person, he would have some specific pose or change in direction of motion to avoid it. Besides, observing the scene context as another source of information has a great impact on the final result. Because when the model observes the football pitch, it can better predict the movements of leg joints. Moreover, this claim is also supported by results in Table III, in which we can see our final model (JointLearning+Social+Context) outperforms the Non-social model (JointLearning) in most of the action classes. Precisely investigating the results in this table, we can figure out that in those actions that involve a high amount of social interaction such as "punching" or "hugging", our ultimate model (JointLearning+Social + Context) outperforms the JointLearning Non-social method and only in a few number of actions with minor social interactions the JointLearning Non-social method performs better. On average, our final model results in better predictions in all time steps and this improvement increases over time.

Finally, (Table IV) shows that pose is better predicted when jointly predicting the global motion and pose with a joint loss rather than separating the two information and learning them separately (Human-centric pose and motion). This experiment is conducted on the single-person subset of NTU RGB-D 60 dataset. For the first experiment the pose is centered relative to the neck and the value of neck is concatenated to the input vector as the human motion.

TABLE IV

MSE ERRORS (IN CM) ON THE MODEL TRAINED ON 1) SEPARATED POSE AND MOTION AND 2) A GLOBAL POSE WITH JOINT LOSS.

Metric (MSE)	milliseconds			
	80	160	320	400
Summed over local pose and trajectory	16.7	18.7	20.0	24.0
On global pose with joint loss	**16.4**	**18.4**	**19.7**	**23.8**

V. CONCLUSION

In this paper, we proposed an action agnostic model for simultaneously forecasting global motion (trajectory) and local pose movements. Our model incorporates social and context cues for making the predictions. Contrary to the previous work, we defined a metric and a loss function in the original space of the movements (2D in PoseTrack and 3D in NTU RGB+D) without the global motion removed. Our end-to-end training framework based on encoder-decoder GRU cells outperforms all the appropriate baselines and shows that the use of spatiotemporal context and social considerations both improve the global pose prediction performance. Such modeling can facilitate detailed understanding of the future actions, which can be used for better navigation and human-robot interaction. As a direction for the future work, GAN-based social models can improve the results.

REFERENCES

[1] J. Martinez, M. J. Black, and J. Romero, "On human motion prediction using recurrent neural networks," in *CVPR*, 2017.

[2] A. Jain, A. R. Zamir, S. Savarese, and A. Saxena, "Structural-rnn: Deep learning on spatio-temporal graphs," in *CVPR*, pp. 5308–5317, 2016.

[3] A. Alahi, K. Goel, V. Ramanathan, A. Robicquet, L. Fei-Fei, and S. Savarese, "Social lstm: Human trajectory prediction in crowded spaces," in *CVPR*, pp. 961–971, 2016.

[4] A. Gupta, J. Johnson, L. Fei-Fei, S. Savarese, and A. Alahi, "Social gan: Socially acceptable trajectories with generative adversarial networks," in *CVPR*, pp. 2255–2264, 2018.

[5] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Gated feedback recurrent neural networks," in *International Conference on Machine Learning*, pp. 2067–2075, 2015.

[6] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *CVPR*, pp. 6299–6308, 2017.

[7] B. Wang, E. Adeli, H. kuang Chiu, D.-A. Huang, and J. C. Niebles, "Imitation learning for human pose prediction," in *ICCV*, 2019.

[8] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang, "NTU RGB+D: A large scale dataset for 3d human activity analysis," in *CVPR*, pp. 1010–1019, 2016.

[9] M. Andriluka, U. Iqbal, E. Insafutdinov, L. Pishchulin, A. Milan, J. Gall, and B. Schiele, "Posetrack: A benchmark for human pose estimation and tracking," in *CVPR*, pp. 5167–5176, 2018.

[10] M. Mathieu, C. Couprie, and Y. LeCun, "Deep multi-scale video prediction beyond mean square error," in *ICLR*, 2016.

[11] K.-H. Zeng, W. B. Shen, D.-A. Huang, M. Sun, and J. C. Niebles, "Visual forecasting by imitating dynamics in natural sequences," in *ICCV*, 2017.

[12] J.-T. Hsieh, B. Liu, D.-A. Huang, L. F. Fei-Fei, and J. C. Niebles, "Learning to decompose and disentangle representations for video prediction," in *Advances in Neural Information Processing Systems*, pp. 517–526, 2018.

[13] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, G. Liu, A. Tao, J. Kautz, and B. Catanzaro, "Video-to-video synthesis," *arXiv:1808.06601*, 2018.

[14] C. Vondrick, H. Pirsiavash, and A. Torralba, "Generating videos with scene dynamics," in *NeurIPS*, 2016.

[15] R. Mahjourian, M. Wicke, and A. Angelova, "Geometry-based next frame prediction from monocular video," in *Intelligent Vehicles Symposium (IV), 2017 IEEE*, pp. 1700–1707, IEEE, 2017.

[16] C. Vondrick, H. Pirsiavash, and A. Torralba, "Anticipating visual representations from unlabeled video," in *CVPR*, 2016.

[17] K. M. Kitani, B. D. Ziebart, J. A. Bagnell, and M. Hebert, "Activity forecasting," in *ECCV*, 2012.

[18] C. Sun, A. Shrivastava, C. Vondrick, R. Sukthankar, K. Murphy, and C. Schmid, "Relational action forecasting," in *CVPR*, pp. 273–283, 2019.

[19] N. Rhinehart and K. Kitani, "First-person activity forecasting from video with online inverse reinforcement learning," *IEEE transactions on pattern analysis and machine intelligence*, 2018.

[20] K. Soomro, H. Idrees, and M. Shah, "Online localization and prediction of actions and interactions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018.

[21] M. M. Arzani, M. Fathy, H. Aghajan, A. A. Azirani, K. Raahemifar, and E. Adeli, "Structured prediction with short/long-range dependencies for human activity recognition from depth skeleton data," in *IROS, 2017 IEEE/RSJ International Conference on*, pp. 560–567, IEEE, 2017.

[22] J. Liang, L. Jiang, J. C. Niebles, A. G. Hauptmann, and L. Fei-Fei, "Peeking into the future: Predicting future person activities and locations in videos," in *CVPR*, pp. 5725–5734, 2019.

[23] A. Alahi, V. Ramanathan, and L. Fei-Fei, "Socially-aware large-scale crowd forecasting," in *CVPR*, 2014.

[24] N. Deo and M. M. Trivedi, "Scene induced multi-modal trajectory forecasting via planning," *arXiv preprint arXiv:1905.09949*, 2019.

[25] J. Walker, A. Gupta, and M. Hebert, "Dense optical flow prediction from a static image," in *ICCV*, pp. 2443–2451, IEEE, 2015.

[26] P. Luc, N. Neverova, C. Couprie, J. Verbeek, and Y. LeCun, "Predicting deeper into the future of semantic segmentation," in *ICCV*, 2017.

[27] S. S. Nabavi, M. Rochan, and Y. Wang, "Future semantic segmentation with convolutional lstm," in *BMVC*, 2018.

[28] H.-k. Chiu, E. Adeli, and J. C. Niebles, "Segmenting the future," *arXiv preprint arXiv:1904.10666*, 2019.

[29] G. Rogez, J. Rihan, S. Ramalingam, C. Orrite, and P. H. Torr, "Randomized trees for human pose detection," in *CVPR*, pp. 1–8, 2008.

[30] M. Andriluka, S. Roth, and B. Schiele, "Monocular 3d pose estimation and tracking by detection," in *CVPR*, pp. 623–630, 2010.

[31] D. Mehta, S. Sridhar, O. Sotnychenko, H. Rhodin, M. Shafiei, H.-P. Seidel, W. Xu, D. Casas, and C. Theobalt, "Vnect: Real-time 3d human pose estimation with a single rgb camera," *ACM Transactions on Graphics (TOG)*, vol. 36, no. 4, p. 44, 2017.

[32] H.-k. Chiu, E. Adeli, B. Wang, D.-A. Huang, and J. C. Niebles, "Action-agnostic human pose forecasting," in *WACV*, 2019.

[33] K. Mangalam, E. Adeli, K.-H. Lee, A. Gaidon, and J. C. Niebles, "Disentangling human dynamics for pedestrian locomotion forecasting with noisy supervision," *arXiv preprint arXiv:1911.01138*, 2019.

[34] Y.-W. Chao, J. Yang, B. Price, S. Cohen, and J. Deng, "Forecasting human dynamics from static images," in *CVPR*, 2017.

[35] J. M. Wang, D. J. Fleet, and A. Hertzmann, "Gaussian process dynamical models for human motion," *IEEE transactions on pattern analysis and machine intelligence*, vol. 30, no. 2, pp. 283–298, 2008.

[36] D. Wu and L. Shao, "Leveraging hierarchical parametric networks for skeletal joints based action segmentation and recognition," in *CVPR*, pp. 724–731, 2014.

[37] P. Ghosh, J. Song, E. Aksan, and O. Hilliges, "Learning human motion models for long-term predictions," *arXiv preprint arXiv:1704.02827*, 2017.

[38] J. Walker, K. Marino, A. Gupta, and M. Hebert, "The pose knows: Video forecasting by generating pose futures," in *ICCV*, pp. 3352–3361, IEEE, 2017.

[39] E. Barsoum, J. Kender, and Z. Liu, "Hp-gan: Probabilistic 3d human motion prediction via gan," *arXiv preprint arXiv:1711.09561*, 2017.

[40] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein gan," *arXiv preprint arXiv:1701.07875*, 2017.

[41] K. Fragkiadaki, S. Levine, P. Felsen, and J. Malik, "Recurrent network models for human dynamics," in *ICCV*, pp. 4346–4354, IEEE, 2015.

[42] X. Yan, A. Rastogi, R. Villegas, K. Sunkavalli, E. Shechtman, S. Hadap, E. Yumer, and H. Lee, "Mt-vae: Learning motion transformations to generate multimodal human dynamics," in *ECCV*, pp. 265–281, 2018.

[43] Y. Zhou, Z. Li, S. Xiao, C. He, Z. Huang, and H. Li, "Auto-conditioned recurrent networks for extended complex human motion synthesis," in *ICLR*, 2018.

[44] J. Ho and S. Ermon, "Generative adversarial imitation learning," in *Advances in neural information processing systems*, pp. 4565–4573, 2016.

[45] D. Helbing and P. Molnar, "Social force model for pedestrian dynamics," *Physical review E*, vol. 51, no. 5, p. 4282, 1995.

[46] R. Mehran, A. Oyama, and M. Shah, "Abnormal crowd behavior detection using social force model," in *CVPR*, pp. 935–942, IEEE, 2009.

[47] K. Yamaguchi, A. C. Berg, L. E. Ortiz, and T. L. Berg, "Who are you with and where are you going?," in *CVPR*, pp. 1345–1352, IEEE, 2011.

[48] A. Robicquet, A. Sadeghian, A. Alahi, and S. Savarese, "Learning social etiquette: Human trajectory understanding in crowded scenes," in *ECCV*, pp. 549–565, Springer, 2016.

[49] S. Pellegrini, A. Ess, K. Schindler, and L. Van Gool, "You'll never walk alone: Modeling social behavior for multi-target tracking," in *ICCV*.

[50] N. Lee, W. Choi, P. Vernaza, C. B. Choy, P. H. Torr, and M. Chandraker, "Desire: Distant future prediction in dynamic scenes with interacting agents," in *CVPR*, pp. 336–345, 2017.

[51] T. Fernando, S. Denman, S. Sridharan, and C. Fookes, "Soft+ hardwired attention: An lstm framework for human trajectory prediction and abnormal event detection," *arXiv preprint arXiv:1702.05552*, 2017.

[52] A. Sadeghian, V. Kosaraju, A. Sadeghian, N. Hirose, H. Rezatofighi, and S. Savarese, "SoPhie: An attentive GAN for predicting paths compliant to social and physical constraints," in *CVPR*, 2019.

[53] V. Kosaraju, A. Sadeghian, R. Martín-Martín, I. Reid, H. Rezatofighi, and S. Savarese, "Social-bigat: Multimodal trajectory forecasting using bicycle-gan and graph attention networks," in *Advances in Neural Information Processing Systems*, pp. 137–146, 2019.

[54] H. Joo, T. Simon, M. Cikara, and Y. Sheikh, "Towards social artificial intelligence: Nonverbal social signal prediction in a triadic interaction," in *CVPR*, pp. 10873–10883, 2019.

[55] M. Zaheer, S. Kottur, S. Ravanbakhsh, B. Poczos, R. R. Salakhutdinov, and A. J. Smola, "Deep sets," in *Advances in neural information processing systems*, pp. 3391–3401, 2017.

[56] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "Pointnet: Deep learning on point sets for 3d classification and segmentation," in *CVPR*, pp. 652–660, 2017.

[57] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, *et al.*, "The kinetics human action video dataset," *arXiv preprint arXiv:1705.06950*, 2017.

[58] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu, "Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments," *IEEE transactions on pattern analysis and machine intelligence*, vol. 36, no. 7, pp. 1325–1339, 2013.

[59] W. Zhang, M. Zhu, and K. G. Derpanis, "From actemes to action: A strongly-supervised representation for detailed action understanding," in *ICCV*, pp. 2248–2255, 2013.

[60] Z. Liu, S. Wu, S. Jin, Q. Liu, S. Lu, R. Zimmermann, and L. Cheng, "Towards natural and accurate future motion prediction of humans and animals," in *CVPR*, pp. 10004–10012, 2019.

[61] S. Toyer, A. Cherian, T. Han, and S. Gould, "Human pose forecasting via deep markov models," *arXiv preprint arXiv:1707.09240*, 2017.