AppGAN: Generative Adversarial Networks for Generating Robot Approach Behaviors into Small Groups of People

Fangkai Yang[1], Christopher Peters[1]

***Abstract*— Robots that navigate to approach free-standing conversational groups should do so in a safe and socially acceptable manner. This is challenging since it not only requires the robot to plot trajectories that avoid collisions with members of the group, but also to do so without making those in the group feel uncomfortable, for example, by moving too close to them or approaching them from behind. Previous trajectory prediction models focus primarily on formations of walking pedestrians, and those models that do consider approach behaviours into free-standing conversational groups typically have handcrafted features and are only evaluated via simulation methods, limiting their effectiveness. In this paper, we propose AppGAN, a novel trajectory prediction model capable of generating trajectories into free-standing conversational groups trained on a dataset of safe and socially acceptable paths. We evaluate the performance of our model with state-of-the-art trajectory prediction methods on a semi-synthetic dataset. We show that our model outperforms baselines by taking advantage of the GAN framework and our novel group interaction module.**

I. INTRODUCTION

In interaction situations involving groups of people that are standing and involved in conversation, a vital ability is to join the group by approaching it. When robots approach these *free-standing conversational groups* [1], they should adopt socially acceptable paths in order not to make individuals in the group feel uncomfortable, for example, due to violating their personal boundaries (see Fig. 1). Due to the importance of these approach behaviors for robots that have social roles, including mobile companion robots and delivery robots in social environments [2], [3], [4], recently a number of navigation methods and experiments have been devised in relation to robot approach behaviors [5]. Although these methods and experiments involve safe and socially acceptable paths, collectively they have shortcomings that limit their utility as a full solution for robot navigation into groups: a) most research focuses on robot behaviors that approach an individual human, b) most of the cases involve groups of humans that are totally static within the conversational group i.e. they do not change body orientations and positions at all, while the positions and orientations of the individuals in free-standing groups are not totally static over time, c) the experimental studies typically do not propose computational models, d) existing computational models are only evaluated in simulation with handcrafted features and specific measurements. To overcome the aforementioned difficulties, we present a data-driven method to generate robot approach behaviors into a conversational group. To the best of our knowledge, no other work has previously used data-driven methods for this purpose. The main contribution of the paper is our introduction of a Long-Short Term Memory network (LSTM) based Generative Adversarial Network (GAN) with a novel group interaction module that fuses body position and orientation information of individuals in a group to generate socially acceptable paths for a mobile robot when it approaches a 'quasi-dynamic'[1] conversational group. Our model is trained and evaluated on a semi-synthetic dataset of safe and socially acceptable paths. We show that it outperforms baseline methods via the GAN and our novel group interaction module, leading to efficient and safe trajectories for robots in free-standing conversational groups.

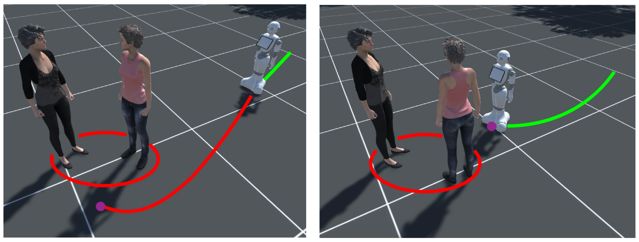


Fig. 1: A sample from our dataset (refer to Fig. 5). Members of the conversational group (red circle) are quasi-dynamic, changing body positions and orientations. The robot initially plans a path (red curve) to approach and join the group without interrupting the conversation while approaching from the front (left). However, the two people in the group change positions as well as body orientations (right) and the robot thus changes its path accordingly (green curve) to continue to approach the group from the front.

II. RELATED WORK

In this section, we present a summary of research on robot approaching behaviors and the related trajectory predictions.

A. Approaching Group Behaviors

There have been many studies concerning mobile robot behaviors in terms of approaching humans recently. They can be classified into two general categories: a) approaching an individual human or b) approaching a group of humans. The first category typically involves proxemics, social distances between humans and robots or virtual agents. Mead and Mataric [6] presented an experiment on how human perceives robots during interactions at different distances. Peters et al.

[1] Department of Computational Science and Technology, School of Electrical Engineering and Computer Science, KTH Royal Institute of Technology, Stockholm, Sweden. {fangkai,chpeters}@kth.se

[1] 'Quasi-dynamic' refers to slight changes in the body positions and orientations of people in a free-standing conversational group.

[7] proposed comfortable proxemics when a virtual robot approaches a human user using Hall's model [8]. Approaching behaviors towards humans in different states, such as sitting, walking, standing, has been studied in [2], [9], [10].

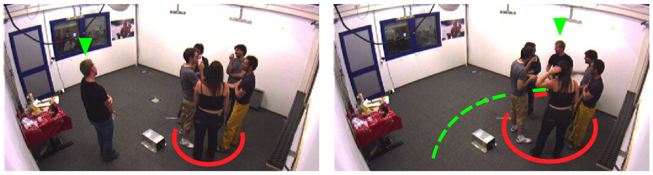


Fig. 2: Images from a cocktail party dataset [11]. A man (green triangle overhead) tries to approach a group (red circle). However, the closest joining position has been occupied, so he goes around the group to find a place to join in without intruding upon the o-space of the group members.

The second category of behaviors, in which an individual approaches a group of humans, has been studied in [12], [13], [14]. In a free-standing conversational group of humans, Kendon [15] proposed the *F-formation* system to define the positions and orientations of individuals within a group. The *O-space*, proposed by Kendon as an exclusive space surrounded by group members, has been used as a basis for models seeking to that prevent robots from intruding into the group. For example, robots outside a group want to approach to join it, they need to calculate a trajectory that does not intersect with the o-space, as shown in Fig. 2. Leveraging the F-formation system, Truong et al. [12] proposed a framework to enable a robot to approach a human group safely and socially. Althaus et al. [16] developed a topological map based model for approaching a group. Gómez et al. [13] extended a fast marching algorithm to navigate a robot for engaging a group of people. Claudio et al. [17] simulate approaching behaviors for virtual characters towards small groups. Other recent works [18], [19] focus on investigating the factors in a conversational human group that may impact robot behaviors, such as human orientations and robot approach distances. Studies show that people prefer to be approached within their field-of-view (FOV) rather than directly behind them [20]. Ball et al. [18] found that seated people feel least comfortable when the approaching robot cannot be seen.

However, most previous works are either experimental studies or computational models implemented and validated in simulation, and all assume that group members remain completely static, i.e. do not change body orientations and positions at all, during conversation. The lack of data-driven models, which can address these limitations, motivates our development of a new model for planning robot approaching behaviors in a human-like and socially acceptable manner.

B. Human Trajectory Prediction

Simulation techniques for human behaviors in crowds have been widely studied at both the *macroscopic* [21], [22], [23] and the *microscopic* [24], [25], [26] perspectives. Our work focuses on the *microscopic* perspective in order to capture human behaviors and interactions of each individual. The Social Force Model [24], [27] is often used in modeling pedestrian future behaviors with attractive forces driving towards the target and repulsive forces encouraging collision avoidance. [13], [28] simulate pedestrian trajectories through modeling cost maps. However, these methods relied on handcrafted features on distances or predefined rules.

With the dawn of deep learning, data-driven methods based on Recurrent Neural Networks (RNNs) have been used to outperform these traditional methods. One of the most popular RNN-based method is the Social LSTM [29] method which represents all pedestrians using Long-Short Term Memory networks (LSTMs) and uses a local pooling method to gather the information from local neighbours. It is further augmented in [30] by a spatio-temporal graph based social attention model and also augmented in [31] through a soft and hardwired attention framework. These methods shown the importance of capturing context information which includes long-short term history of the pedestrian of interest and the neighbours in trajectory prediction.

Generative Adversarial Networks (GANs) [32] based trajectory prediction methods [33], [34], [35] have shown promising results in capturing interactions between pedestrians and the surrounding environment. Social GAN [35] proposed a pooling mechanism to learn social norms in a data-driven approach. SoPhie [33] leveraged social and scene attention-based mechanisms to generate a distribution of predicted trajectories. GD GAN [34] augmented the task of trajectory prediction with group detection to discover social group interactions. GAN models can be used to predict trajectories based on a training set and starting conditions. They can also then be used for generating new trajectories.

However, none of the aforementioned works focus specifically on predicting the trajectory of a human or robot approaching a group. Moreover, only position information is considered in these works since their training datasets have few or no conversational groups, and orientation information is always ignored or defaulted as the moving direction.

III. Method

Our primary objective is to generate a path for a robot to approach a group of people in a socially acceptable manner. To achieve this goal, we have developed a GAN based encoder-decoder architecture with attention mechanisms and trained it on a dataset consists of socially acceptable approaching group trajectories.

A. Problem Definition

We assume that there are N humans $p_1, p_2, \ldots, p_N$ in a conversational group. As shown in [11], [36], humans in a group change body and head orientations as well as positions during conversation. The input position and body orientation of a human i ($i \in [1, N]$) is defined as $\mathbf{X}_{H_i}^t = (x_{H_i}^t, y_{H_i}^t)$ and $\theta_{H_i}^t$ at t time step. Moreover, the input position of a mobile robot at t is defined as $\mathbf{X}_R^t = (x_R^t, y_R^t)$. For the given observed positions and orientations of N humans in a group from time steps $t = 1, 2, \ldots, t_{pred}$ and the robot trajectory from time steps $t = 1, 2, \ldots, t_{obs}$, our model aims at predicting the robot trajectory $\hat{\mathbf{Y}}_R^t$ in the future time period from time steps $t = t_{obs} + 1, \ldots, t_{pred}$.

B. Generative Adversarial Networks

We use LSTM-based Generative Adversarial Networks (GANs) to predict the future path of a mobile robot that approaches a human group. It consists of two neural networks, the generator (G) and the discriminator (D), that are trained to compete with each other [32]. The generator is trained to capture the distribution of the paths to sample a possible future path, while the discriminator is trained to estimate the probability that the generated path is from the data rather than the generator. The generator G takes a latent vector $\mathbf{z}$ to get an output vector $G(\mathbf{z})$. The discriminator D takes a sample $\mathbf{y}$ from the dataset as input and output $D(\mathbf{y})$ which represents its probability of it being real.

In this paper, we augment the generic GAN to be a conditional model [34], [37] by providing the generator and discriminator with the current group content vector $\mathbf{G}_\star^t$ (see Section III-D), in order to generate a better path which is guided by the human group. The training process is a min-max game with the objective function shown as below:

$$\begin{aligned} &\underset{G}{min}\ \underset{D}{max} V(G, D) = \mathbb{E}_{\mathbf{y}^t, \mathbf{G}_\star^t \sim p_{data}}[logD(\mathbf{G}_\star^t, \mathbf{y}^t)] + \\ &\mathbb{E}_{\mathbf{G}_\star^t \sim p_{data}, \mathbf{z}^t \sim p(\mathbf{z})}[1 - logD(\mathbf{G}_\star^t, G(\mathbf{G}_\star^t, \mathbf{z}^t))] \end{aligned} \tag{1}$$

C. Approaching Group GAN

Our model consists of Generator (G), Group Interaction Module (GIM) and Discriminator (D). G is an encoder-decoder framework where the hidden states of the humans in a group are linked with the hidden states of the robot via GIM. G outputs predicted trajectory $\hat{\mathbf{Y}}_R$ and D inputs the entire robot trajectory $[\mathbf{X}_R, \hat{\mathbf{Y}}_R]$ or $[\mathbf{X}_R, \mathbf{Y}_R]$ to classifies them as fake or real as shown in Fig. 3.

1) Generator: As pointed out in [33], [35], in the pooling or attention modules, the relative positions related to the agent of interest (i.e. the mobile robot in this paper) help to capture the importance with respect to other agents. We thus embed the relative positions similar to the spatial edges in [30]. The relative position $\mathbf{X}_{H_i}^{\prime t} = (x_{H_i}^t - x_R^t, y_{H_i}^t - y_R^t)$ of each group member is embedded in order to get a fixed length vector e_{Hi}^t in a high dimensional feature space through a single layer Multilayer Perceptron (MLP). This embedding is used as an input to the LSTM cell at time t as follows:

$$\begin{aligned} e_{H_i}^t &= \phi(\mathbf{X}_{H_i}^{\prime t}; \mathbf{W}_{hp}) \\ h_{H_i}^t &= LSTM(h_{H_i}^{t-1}, e_{H_i}^t; \mathbf{W}_{hpencoder}) \end{aligned} \tag{2}$$

where $\phi(\cdot)$ is an embedding function with ReLU nonlinearity, $\mathbf{W}_{hp}$ is the embedding weight. $h_{H_i}^t$ is the hidden state of the LSTM at time t and $\mathbf{W}_{hpencoder}$ are the LSTM weights which are shared among all group members. The position of the mobile robot is embedded in a similar way but with another LSTM as follows:

$$\begin{aligned} e_R^t &= \phi(\mathbf{X}_R^t; \mathbf{W}_r) \\ h_R^t &= LSTM(h_R^{t-1}, e_R^t; \mathbf{W}_{rencoder}) \end{aligned} \tag{3}$$

The human orientations in the group affect the trajectory of the mobile robot (see Fig. 1). The orientations are thus embedded and fed into an orientation LSTM as follows:

$$\begin{aligned} r_{H_i}^t &= \phi(\theta_{H_i}^t; \mathbf{W}_{ho}) \\ o_{H_i}^t &= LSTM(o_{H_i}^{t-1}, r_{H_i}^t; \mathbf{W}_{hoencoder}) \end{aligned} \tag{4}$$

As shown in [29], [33], [34], [35], naive use of one LSTM for the robot fails to capture the information of the surrounding environment. We have developed an attention module to fuse hidden states of all group members that the robot plans to approach to get a combined content vector $\mathbf{G}_\star^t$. Unlike the work of [33], [34], which takes a white noise vector $\mathbf{z}$ as an input, we use a similar method to that proposed by Gupta et al. [35] and initialize the hidden state of the decoder by conditioning on the group content vector $\mathbf{G}_\star^t$ as:

$$\begin{aligned} g^t &= \gamma(\mathbf{G}_\star^t, h_R^t; \mathbf{W}_g) \\ h_d^t &= [g^t, \mathbf{z}] \end{aligned} \tag{5}$$

where $\gamma(\cdot)$ is an MLP with ReLU nonlinearity and $\mathbf{W}_g$ is the embedding weight.

After initializing the decoder, the decoder states are used to predict trajectories:

$$\begin{aligned} e_R^{t+1} &= \phi(\mathbf{X}_R^t; \mathbf{W}_{rd}) \\ h_d^{t+1} &= LSTM(\gamma(\mathbf{G}_\star^{t+1}, h_d^t), e_R^{t+1}; \mathbf{W}_{rdecoder}) \\ (\hat{x}_R^{t+1}, \hat{y}_R^{t+1}) &= \gamma(h_d^{t+1}) \end{aligned} \tag{6}$$

where $\phi(\cdot)$ is an embedding function with ReLU nonlinearity, $\mathbf{W}_{rd}$ is the embedding weight, $\mathbf{W}_{rdecoder}$ are the LSTM decoder weights and $\gamma(\cdot)$ is an MLP. Unlike prior works [29], [30] that used hidden states to predict parameters of a bivariate Gaussian distribution, we predict the position of the mobile robot directly, which avoids difficulties in training through nondifferentiable backpropagation [35]. We use e_R^{t+1} in (6) to get $\mathbf{G}_\star^{t+1}$ in prediction.

2) Discriminator: The discriminator consists of another LSTM as an encoder, which distinguishes a trajectory sample that is from either the ground truth dataset or the predicted future paths, i.e. $T_{real} = [\mathbf{X}_R, \mathbf{Y}_R]$, $T_{fake} = [\mathbf{X}_R, \hat{\mathbf{Y}}_R]$.

3) Losses: To train AppGAN, we augment the adversarial loss with $L2$ loss on the predicted trajectory which measures its deviation from the ground truth as follows:

$$L_2(\mathbf{Y}_R, \hat{\mathbf{Y}}_R) = ||\mathbf{Y}_R - \hat{\mathbf{Y}}_R||_2^2 \tag{7}$$

D. Group Interaction Module

Similar to the manner in which a human pays attention to the group which he/she is approaching, the mobile robot should focus on the relevant group members in order to predict a better trajectory. Most current trajectory prediction methods use either pooling modules [29], [35] or attention modules [30], [33], [34] in order to represent social interactions between the agent of interest with respect to other agents, either within a fixed neighbourhood or in the whole scene. However, these methods only consider human position data since the models were trained with datasets only contain position information. And the orientation information was

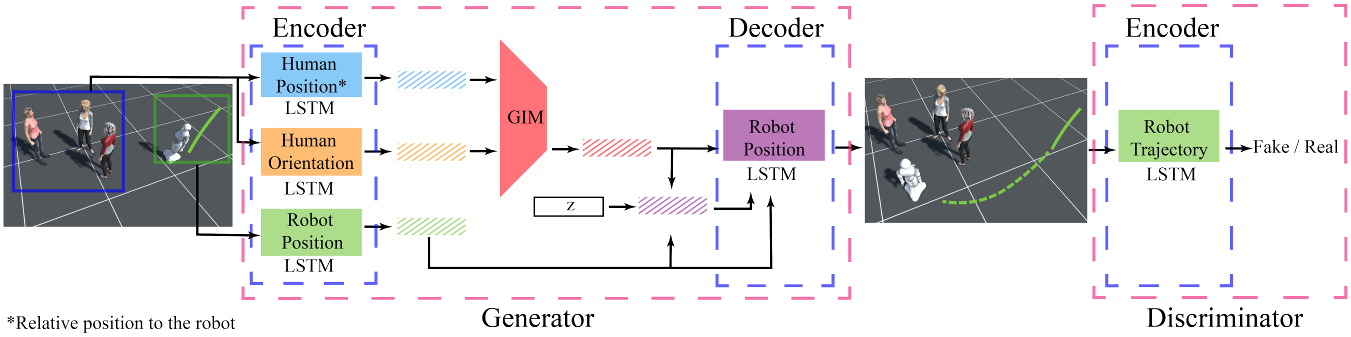


Fig. 3: The overview of *AppGAN*. It consists of Generator (G), Group Interaction Module (GIM) and Discriminator (D). G takes the observed trajectories of humans and the robot, and encodes the relative positions and orientations of humans separately as well as the robot positions. GIM is a multimodal attention-based fusion module which inputs the hidden states of humans and outputs a group content vector for trajectory generation. The decoder generates the future trajectories from the group content vector and the hidden states of the robot. D takes input T_{real} or T_{false} to classify them as real or not.

not taken into consideration in the these works which default the orientation to be the same as the walking direction.

As aforementioned in Section II-A, humans in a free-standing conversational group change their body orientations when talking to different group members, and the body positions in the group also change, ranging from minor movements (e.g. body pose changes) to large (e.g. a human joins or leaves the group). To consider the orientation and position of the group members, we use attention modules similar to [38]. Two separate attention modules are used for position and orientation respectively, as shown in Fig. 4.

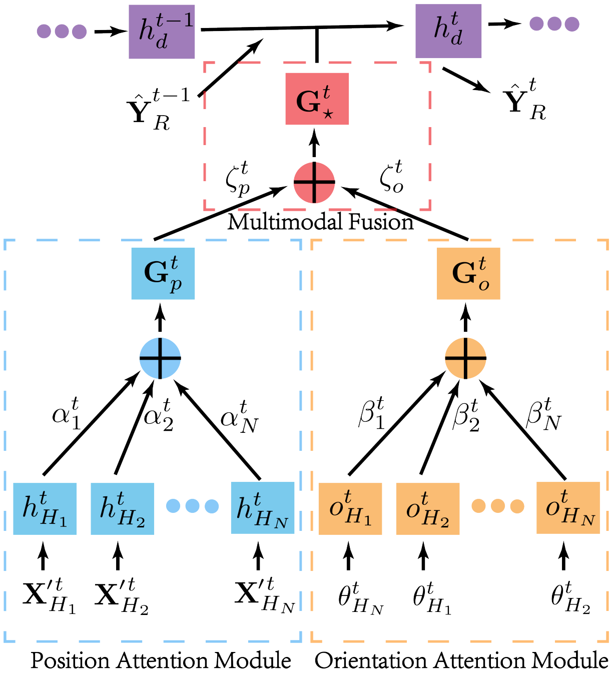


Fig. 4: Our group interaction module (GIM) consists of a position attention module (blue dashed rectangle), an orientation attention module (orange dashed rectangle) and a multimodal fusion module (red dashed rectangle).

For each group member, the position attention module computes a soft attention over the position hidden states $h^t_{H_i}$ at each time step t to get the group position content $\mathbf{G}_p^t$:

$$\mathbf{G}_p^t = \sum_{i=1}^{N} \alpha_i^t h^t_{H_i} \tag{8}$$

where the weight α_i^t is computed as a softmax by:

$$\alpha_i^t = \frac{exp(s_i^t)}{\sum_{l=1}^{N} exp(s_l^t)} \tag{9}$$

where $s_i^t = s(h^t_{H_i}, h^t_R)$ is a score function which measures the interactions between the position of human i and the mobile robot. We use a feed forward neural network with $tanh$ activation function as the score function $s(\cdot)$. Similarly, the group orientation content $\mathbf{G}_o^t$ can be modeled as:

$$\begin{aligned} \mathbf{G}_o^t &= \sum_{i=1}^{N} \beta_i^t o^t_{H_i} \\ \beta_i^t &= \frac{exp(s'^t_i)}{\sum_{l=1}^{N} exp(s'^t_l)} \end{aligned} \tag{10}$$

in which the score function $s'^t_i = s'(o^t_{H_i}, h^t_R)$ is another feed forward network measuring the interactions from the orientation of human i to the position of the robot.

Once we have the group position and orientation content vectors, we need to fuse them to obtain a combined content vector $\mathbf{G}_\star^t$. Rather than using a simple feature fusion method as described in [39], where content vectors are combined with weight matrices, we use a multimodal attention-based feature fusion method proposed by Hori et al. [40] in order to exploit both the position and orientation features effectively. The multimodal attention mechanism enables the decoder to selectively attend to position or orientation content when predicting the robot's position. It is similar to (10) as follows:

$$\begin{aligned} \mathbf{G}_\star^t &= \zeta_p^t \mathbf{G}_p^t + \zeta_o^t \mathbf{G}_o^t \\ \zeta_p^t &= \frac{exp(f(\mathbf{G}_p, h_d^{t-1}))}{exp(f(\mathbf{G}_p, h_d^{t-1})) + exp(f(\mathbf{G}_o, h_d^{t-1}))} \\ \zeta_o^t &= \frac{exp(f(\mathbf{G}_o, h_d^{t-1}))}{exp(f(\mathbf{G}_p, h_d^{t-1})) + exp(f(\mathbf{G}_o, h_d^{t-1}))} \end{aligned} \tag{11}$$

where the score function $f(\cdot)$ is a feed forward network. For initialization in (5), we use a simple fusion method [39].

IV. Evaluations

1) Dataset: To the best of our knowledge, there is no large dataset consisting of individuals approaching groups to join them. Existing datasets that contain conversational groups, such as the CocktailParty dataset [11] and the SALSA dataset [41], have approaching group behaviors, but the number of trajectories is not large enough for training. To overcome this difficulty, we develop a semi-synthetic dataset[2]. It contains 4836 trajectories of individuals approaching groups as well as the group member positions and orientations, with the number of people in each group varying from two to six. Since people in a conversational group are quasi-dynamic, with changing in body positions and orientations, we extract the conversational group information from the CocktailParty dataset [11] including positions and orientations. Graph-Cuts for F formation (GCFF) [36] is used to determine the group centers. Each group member and group center are modeled as a 2D Gaussian function [12], [42] indicating a social-aware space. A speed map is derived from the social-aware space, and the fast marching method [13], [43] is used to generate a path that approaches a group. Fig. 5 shows a sample from our dataset and it is identical to Fig. 1. We evaluated the generating process to be realistic and socially acceptable in the SALSA dataset [41] and the VEIIG dataset [44] concerning trajectories through crowds with groups.

2) Implementation Details: We iteratively train the generator and discriminator with the Adam optimizer [45] using a mini-batch size of 32 and an initial learning rate of 0.001. The model is trained for 200 epochs. The encoder encodes trajectories through an MLP with an embedding dimension of 16. The dimensions of the hidden state is 32 for the encoder and 64 for the decoder. The attention dimension for both position and orientation modules is set to 32. In the group interaction module, all group members are considered to have interactions with the robot. The AppGAN model was trained on a GTX 1070 Ti GPU with a PyTorch implementation.

A. Baselines

We compare our trajectory prediction model to 5 state-of-the-art baselines. We use the Social Force (SF) model introduced in [27], where the group members do not react to the approaching robot. The robot is driven by the attractive force from the group center while the repulsive forces from group members encourage the robot to avoid collisions. We also use the Social Attention (SA) [30] model which has been shown to be able to capture the relative importance of each group member. We also compare Social LSTM (S-LSTM) [29] with a social pooling layer. In relation to GAN methods, we compare our AppGAN with Social GAN (S-GAN), specifically S-GAN with the pooling module and variety loss $k = 10$ [35] and SoPhie [33]. Since there are no physical constraints in the dataset, physical attention is not implemented in SoPhie. For the purposes of a fair comparison, we used the same embedding dimension as well as encoder and decoder dimensions for these GAN methods. The force parameters of SF and pooling parameters of S-LSTM, S-GAN are taken from related original papers.

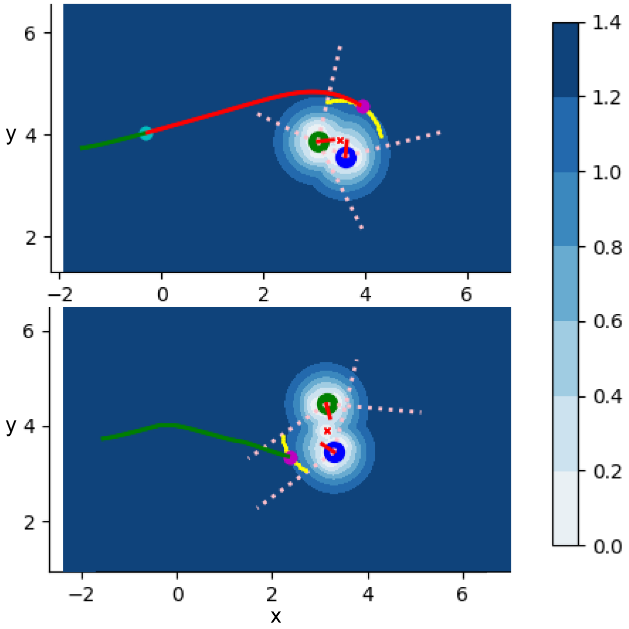


Fig. 5: Top-down view of a trajectory sample from our dataset (identical to Fig. 1). Images depict speed maps generated from the social-aware space of the conversational group. The lighter color contour indicates a lower speed for the robot to walk through. The conversational group contains two people (green and blue circles with red lines indicating body orientations) and a group center (red cross). Potential joining positions (yellow curve) shall be within the FOVs of all group members at a comfortable distance. The changes in positions and orientations deviate the actual trajectory (green curve) from the planned trajectory (red curve).

1) Ablation Study: We also do an ablation study of AppGAN with different control settings: *AppGAN-P* AppGAN only with the position module, *AppGAN-O* AppGAN only with the orientation module, *AppGAN-L* AppGAN using a simple feature fusion.

2) Evaluation Metrics: Similar to other GAN works [33], [35], we use two error metrics: Average Displacement Error (ADE) and Final Displacement Error (FDE).

3) Evaluation Methodology: We observe the trajectory for 4 time steps (1.0 second) and show the prediction for 5 time steps (1.25 seconds).

B. Quantitative Evaluations

We compare our model with different control settings to various baselines (see Table I). We observe poor performance of the SF model due to handcrafted force features and a lack of motion history. AppGAN-O, which consists only of the orientation module, performs as poorly as the SF model due to the lack of position information of group members. The trajectory distribution in AppGAN-O is very diverse (see Fig. 8(b)), which means the robot does not know when and where it should steer. S-LSTM and SA highly improve the prediction performance utilizing the LSTMs of the robot

[2] https://kth.box.com/s/iq7612vurbbnk6rygx9b6m2kil1urczb

motion history and the group position. SA performs better than S-LSTM as it builds a spatial-temporal graph which includes all group members and utilizes an attention model. The GAN based models, S-GAN and SoPhie, increase the performance of LSTM based models via a generative approach. SoPhie does not outperform S-GAN due to the lack of physical attention. AppGAN-P, which solely has the position module, performs as well as SoPhie. They both use a soft attention module, but AppGAN-P uses relative position information in the position attention module, and that could possibly result in a slightly better performance. Our AppGAN outperforms the other methods and AppGAN variations. As expected, the orientation information helps AppGAN in better prediction, but it is not able to do a good prediction alone without the position information. AppGAN-L performs as well as AppGAN which suggests the multimodal attention mechanism is not very necessary.

C. *Qualitative Evaluations*

Currently existing trajectory prediction methods have focused on the position information from walking agents, such as pedestrians. It is imperative to show how orientation information can affect the trajectory prediction when individuals approach small groups. We consider two scenarios from our dataset. Scenario A (see Fig. 6(a)) contains three people in a group while the robot is approaching from the left. We want to test the steering behaviors in this scenario. Scenario B (see Fig. 6(b)) contains two people in a group. Unlike Scenario A where the body orientations are mostly static, the two people in Scenario B are quasi-dynamic (see Fig. 7).

As shown in Fig. 6, SF performs worst and disrespects the group members by approaching from behind and intruding the social-aware space of the people. It is interesting that S-LSTM predictions (gray lines) overreacts to collision avoidance to the group which results in the trajectories deviated from the group rather than approaching it. Although the body orientation information is not used in SA, S-GAN and SoPhie, the predictions from these three methods show a trend to steer towards the group. It suggests that they learn the robot should always approach people.

AppGAN performs especially well in Scenario B (see Fig. 6(b)). Rather than following the observed motion pattern, AppGAN steers the robot in order to approach people from the front within their FOVs and respects the social-aware space in a socially acceptable manner. It is unexpected that AppGAN makes a smooth turn rather than an abrupt one in the ground truth (since the synthetic ground truth trajectory plans only on current frame which results in abrupt turns). However, the other baseline methods follow the observed trajectory and predict trajectories that move downwards and approach from behind without making turns (see Fig. 6(b)).

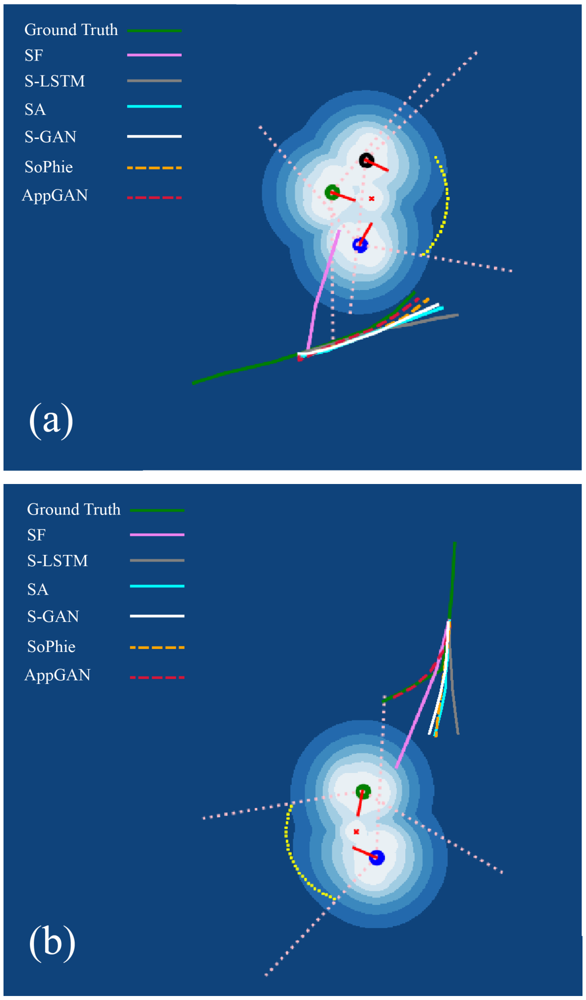


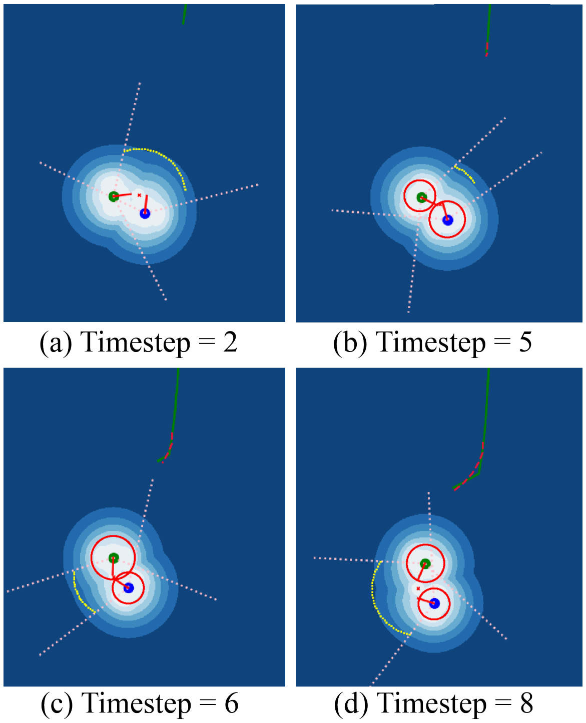
Fig. 6: Top-down view of two sample scenarios where we compare AppGAN with the baseline methods. (a) is a static group where the positions and orientations of group members are mostly static. (b) is a quasi-dynamic group where both positions and orientations change in timesteps (see Fig. 7).

We also compare how the group interaction module performs in Scenario B (see Fig. 7(b-d)). At time step 5, i.e. the first time step in prediction, the attention weights of the two people are close with a slightly higher attention in blue. At time step 6, the green people rotates and leaves his/her back to the robot, and the attention in green is higher than blue. At time step 8, the difference between the attentions from blue and green becomes smaller. It suggests that the robot pays more attention to the people who lost sight of it, which matches the expectation that people prefer to be approached

TABLE I: Quantitative results of all methods in the semi-synthetic dataset. Two evaluation metrics Average Displacement Error (ADE) and Final Displacement Error (FDE) are reported in meters. AppGAN outperforms the baselines and AppGAN variations with only one of the two attention modules (lower is better).

	Baselines					**AppGAN**			
Metric	SF	S-LSTM	SA	S-GAN	SoPhie	AppGAN-P	AppGAN-O	AppGAN-L	AppGAN
ADE	1.01	0.59	0.42	0.36	0.32	0.30	0.92	0.18	**0.15**
FDE	1.75	0.93	0.75	0.64	0.58	0.57	1.53	0.29	**0.27**

within their FOVs. The group member closer to the robot has a higher attention weight regarding collision avoidance.


(a) Timestep = 2

(b) Timestep = 5

(c) Timestep = 6

(d) Timestep = 8

Fig. 7: Scenario B at four time steps. The people (blue and green) initially face towards the robot and leave the joining positions (yellow dots) close to the robot (see (a)(b)). However, they change body orientations and leave their backs to the robot (see (c)(d)). The weights of $\mathbf{G}_{\star}^{t}$ from the two people are marked as red circles (larger radius means higher weight, and the radius have been normalized).

The position attention module and orientation attention module in AppGAN are evaluated in Scenario A (see Fig. 8). AppGAN-P predictions lead steering towards the group. AppGAN-O predictions make a diverse steering and even collide with the group members which leads to unsafe paths and discomfort. The lack of position information makes AppGAN-O blind in both steering directions and timing. AppGAN predictions match the ground truth trajectories with a more packed distribution.

V. DISCUSSION

The dataset we used for training is semi-synthetic which combines recorded group information with generated trajectories. The method to generate these trajectories was evaluated to be realistic in crowds with dynamic and static groups, and the trajectories are safe and socially acceptable. We showed that AppGAN is able to capture both the position and orientation information from the dataset. Thus, our future work will be recording a dataset specifically for approaching group behaviors from humans, including both position and orientation information, and the AppGAN model will be tested in real world settings. Moreover, the robot orientation is defaulted as the moving direction in trajectory generation, the group orientation content vector is thus fused with the robot position which also contains its orientation information. Our AppGAN model could be further developed to explicitly generate the robot orientation.

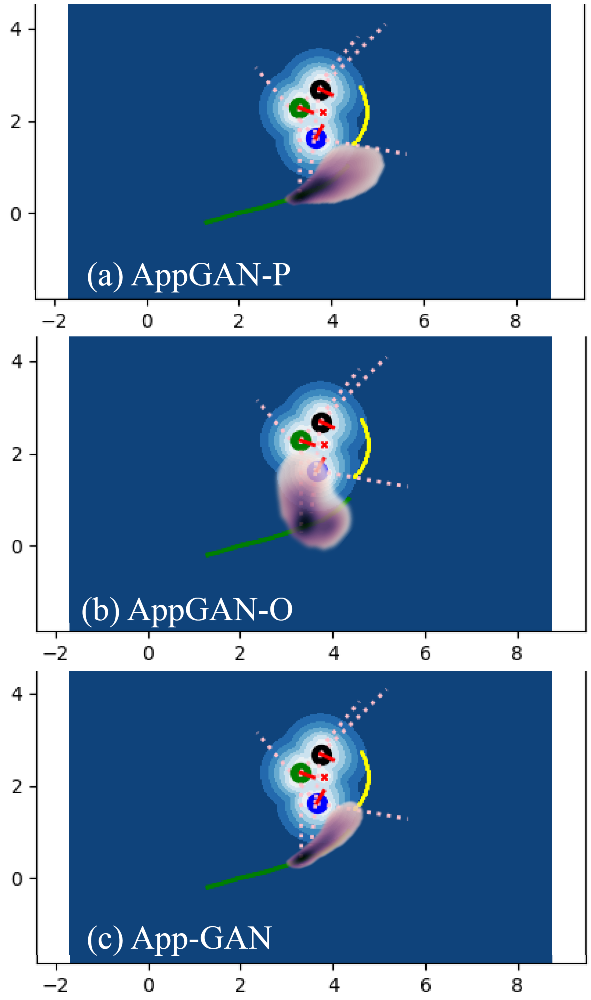


Fig. 8: Comparison between AppGAN and AppGAN variations with only one attention module. 250 samples are drawn to make an approximate distribution of trajectories.

State-of-the-art methods are mostly used to predict trajectories in moving scenarios that involve few or no free-standing conversational groups. Orientation information is thus not taken into consideration, but defaults to the walking direction. This might also be the reason that current public datasets only contain position information. We hope our work highlights the importance of orientation information when using data-driven methods in trajectory prediction.

VI. CONCLUSIONS

In this paper, we generate robot trajectories for approaching small groups. We propose a novel GAN based model that outperforms state-of-the-art methods on the problem of predicting approaching group trajectories. We also propose a novel group interaction module which captures both position and orientation attention information from the people in a quasi-dynamic group. Our model outperforms the baselines by learning safe and socially acceptable trajectories from the data for the robot to approach the group.

References

[1] X. Alameda-Pineda, Y. Yan, E. Ricci, O. Lanz, and N. Sebe, "Analyzing free-standing conversational groups: A multimodal approach," in *Proceedings of the 23rd ACM International Conference on Multimedia*, ser. MM '15. New York, NY, USA: ACM, 2015, pp. 5–14.

[2] K. L. Koay, D. S. Syrdal, M. Ashgari-Oskoei, M. L. Walters, and K. Dautenhahn, "Social roles and baseline proxemic preferences for a domestic service robot," *International Journal of Social Robotics*, vol. 6, no. 4, pp. 469–488, 2014.

[3] R. Triebel, K. Arras, R. Alami, L. Beyer, S. Breuers, R. Chatila, M. Chetouani, D. Cremers, V. Evers, M. Fiore, *et al.*, "Spencer: A socially aware service robot for passenger guidance and help in busy airports," in *Field and service robotics*. Springer, 2016, pp. 607–622.

[4] T. Fong, I. Nourbakhsh, and K. Dautenhahn, "A survey of socially interactive robots," *Robotics and autonomous systems*, vol. 42, no. 3-4, pp. 143–166, 2003.

[5] T. Kruse, A. K. Pandey, R. Alami, and A. Kirsch, "Human-aware robot navigation: A survey," *Robotics and Autonomous Systems*, vol. 61, no. 12, pp. 1726–1743, 2013.

[6] R. Mead and M. J. Matarić, "Proxemics and performance: Subjective human evaluations of autonomous sociable robot distance and social signal understanding," in *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2015, pp. 5984–5991.

[7] C. Peters, F. Yang, H. Saikia, C. Li, and G. Skantze, "Towards the use of mixed reality for hri design via virtual robots," in *Proceedings of the 1st International Workshop on Virtual, Augmented, and Mixed Reality for HRI (VAM-HRI)*, 2018.

[8] E. T. Hall, *The hidden dimension*. Garden City, NY: Doubleday, 1910, vol. 609.

[9] E. Torta, R. H. Cuijpers, and J. F. Juola, "Design of a parametric model of personal space for robotic social navigation," *International Journal of Social Robotics*, vol. 5, no. 3, pp. 357–365, 2013.

[10] D. Carton, A. Turnwald, D. Wollherr, and M. Buss, "Proactively approaching pedestrians with an autonomous mobile robot in urban environments," in *Experimental Robotics*. Springer, 2013.

[11] E. Ricci, J. Varadarajan, R. Subramanian, S. Rota Bulo, N. Ahuja, and O. Lanz, "Uncovering interactions and interactors: Joint estimation of head, body orientation and f-formations from surveillance videos," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 4660–4668.

[12] X.-T. Truong and T.-D. Ngo, "to approach humans?: A unified framework for approaching pose prediction and socially aware robot navigation," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 10, no. 3, pp. 557–572, 2018.

[13] J. V. Gómez, N. Mavridis, and S. Garrido, "Fast marching solution for the social path planning problem," in *2014 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2014, pp. 1871–1876.

[14] V. K. Narayanan, A. Spalanzani, F. Pasteau, and M. Babel, "On equitably approaching and joining a group of interacting humans," in *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2015, pp. 4071–4077.

[15] A. Kendon, *Conducting interaction: Patterns of behavior in focused encounters*. CUP Archive, 1990, vol. 7.

[16] P. Althaus, H. Ishiguro, T. Kanda, T. Miyashita, and H. I. Christensen, "Navigation for human-robot interaction tasks," in *IEEE International Conference on Robotics and Automation, 2004. Proceedings. ICRA'04. 2004*, vol. 2. IEEE, 2004, pp. 1894–1900.

[17] C. Pedica and H. H. Vilhjálmsson, "Study of nine people in a hallway: Some simulation challenges," in *Proceedings of the 18th International Conference on Intelligent Virtual Agents, IVA 2018, Sydney, NSW, Australia, November 05-08, 2018*, 2018, pp. 185–190.

[18] A. K. Ball, D. C. Rye, D. Silvera-Tawil, and M. Velonaki, "How should a robot approach two people?" *Journal of Human-Robot Interaction*, vol. 6, no. 3, pp. 71–91, 2017.

[19] M. Vázquez, E. J. Carter, B. McDorman, J. Forlizzi, A. Steinfeld, and S. E. Hudson, "Towards robot autonomy in group conversations: Understanding the effects of body orientation and gaze," in *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*. ACM, 2017, pp. 42–52.

[20] M. L. Walters, K. Dautenhahn, S. N. Woods, and K. L. Koay, "Robotic etiquette: results from user studies involving a fetch and carry task," in *2007 2nd ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 2007, pp. 317–324.

[21] R. L. Hughes, "The flow of human crowds," *Annual review of fluid mechanics*, vol. 35, no. 1, pp. 169–182, 2003.

[22] A. Treuille, S. Cooper, and Z. Popović, "Continuum crowds," *ACM Transactions on Graphics (TOG)*, vol. 25, no. 3, pp. 1160–1168, 2006.

[23] R. Narain, A. Golas, S. Curtis, and M. C. Lin, "Aggregate dynamics for dense crowd simulation," in *ACM transactions on graphics (TOG)*, vol. 28, no. 5. ACM, 2009, p. 122.

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

[25] J. Van den Berg, M. Lin, and D. Manocha, "Reciprocal velocity obstacles for real-time multi-agent navigation," in *Robotics and Automation, 2008. ICRA 2008. IEEE International Conference on*. IEEE, 2008, pp. 1928–1935.

[26] W. Yu, R. Chen, L. Dong, and S. Dai, "Centrifugal force model for pedestrian dynamics," *Physical Review E*, vol. 72, p. 026112, 2005.

[27] C. Pedica and H. Vilhjálmsson, "Social perception and steering for online avatars," in *International Workshop on Intelligent Virtual Agents*. Springer, 2008, pp. 104–116.

[28] S. Yi, H. Li, and X. Wang, "Understanding pedestrian behaviors from stationary crowd groups," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3488–3496.

[29] A. Alahi, K. Goel, V. Ramanathan, A. Robicquet, L. Fei-Fei, and S. Savarese, "Social lstm: Human trajectory prediction in crowded spaces," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 961–971.

[30] A. Vemula, K. Muelling, and J. Oh, "Social attention: Modeling attention in human crowds," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2018, pp. 1–7.

[31] T. Fernando, S. Denman, S. Sridharan, and C. Fookes, "Soft+ hardwired attention: An lstm framework for human trajectory prediction and abnormal event detection," *Neural networks*, vol. 108, pp. 466–478, 2018.

[32] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[33] A. Sadeghian, V. Kosaraju, A. Sadeghian, N. Hirose, and S. Savarese, "Sophie: An attentive gan for predicting paths compliant to social and physical constraints," *arXiv preprint arXiv:1806.01482*, 2018.

[34] T. Fernando, S. Denman, S. Sridharan, and C. Fookes, "Gd-gan: Generative adversarial networks for trajectory prediction and group detection in crowds," *arXiv preprint arXiv:1812.07667*, 2018.

[35] A. Gupta, J. Johnson, L. Fei-Fei, S. Savarese, and A. Alahi, "Social gan: Socially acceptable trajectories with generative adversarial networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2255–2264.

[36] F. Setti, C. Russell, C. Bassetti, and M. Cristani, "F-formation detection: Individuating free-standing conversational groups in images," *PloS one*, vol. 10, no. 5, p. e0123783, 2015.

[37] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *arXiv preprint arXiv:1411.1784*, 2014.

[38] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[39] H. Yu, J. Wang, Z. Huang, Y. Yang, and W. Xu, "Video paragraph captioning using hierarchical recurrent neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 4584–4593.

[40] C. Hori, T. Hori, T.-Y. Lee, Z. Zhang, B. Harsham, J. R. Hershey, T. K. Marks, and K. Sumi, "Attention-based multimodal fusion for video description," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 4193–4202.

[41] X. Alameda-Pineda, Y. Yan, E. Ricci, O. Lanz, and N. Sebe, "Analyzing free-standing conversational groups: A multimodal approach," in *Proceedings of the 23rd ACM international conference on Multimedia*. ACM, 2015, pp. 5–14.

[42] T. Amaoka, H. Laga, S. Saito, and M. Nakajima, "Personal space modeling for human-computer interaction," in *International Conference on Entertainment Computing*. Springer, 2009, pp. 60–72.

[43] S. Osher and J. A. Sethian, "Fronts propagating with curvature-dependent speed: algorithms based on hamilton-jacobi formulations," *Journal of computational physics*, vol. 79, no. 1, pp. 12–49, 1988.

[44] S. Bandini, A. Gorrini, and G. Vizzari, "Towards an integrated approach to crowd analysis and crowd synthesis: A case study and first results," *Pattern Recognition Letters*, vol. 44, pp. 16–29, 2014.

[45] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.