Social GAN: Socially Acceptable Trajectories with Generative Adversarial Networks

Agrim Gupta[1] Justin Johnson[1] Li Fei-Fei[1] Silvio Savarese[1] Alexandre Alahi[1,2]
Stanford University[1] École Polytechnique Fédérate de Lausanne[2]

Abstract

Understanding human motion behavior is critical for autonomous moving platforms (like self-driving cars and social robots) if they are to navigate human-centric environments. This is challenging because human motion is inherently multimodal: given a history of human motion paths, there are many socially plausible ways that people could move in the future. We tackle this problem by combining tools from sequence prediction and generative adversarial networks: a recurrent sequence-to-sequence model observes motion histories and predicts future behavior, using a novel pooling mechanism to aggregate information across people. We predict socially plausible futures by training adversarially against a recurrent discriminator, and encourage diverse predictions with a novel variety loss. Through experiments on several datasets we demonstrate that our approach outperforms prior work in terms of accuracy, variety, collision avoidance, and computational complexity.

1. Introduction

Predicting the motion behavior of pedestrians is essential for autonomous moving platforms like self-driving cars or social robots that will share the same ecosystem as humans. Humans can effectively negotiate complex social interactions, and these machines ought to be able to do the same. One concrete and important task to this end is the following: given observed motion trajectories of pedestrians (coordinates for the past *e.g.* 3.2 seconds), predict *all* possible future trajectories (Figure 1).

Forecasting the behavior of humans is challenging due to the inherent properties of human motion in crowded scenes:

1. **Interpersonal**. Each person's motion depends on the people around them. Humans have the innate ability to read the behavior of others when navigating crowds. Jointly modeling these dependencies is a challenge.

2. **Socially Acceptable**. Some trajectories are physically possible but socially unacceptable. Pedestrians are governed by social norms like yielding right-of-way or respecting personal space. Formalizing them is not trivial.

3. **Multimodal**. Given a partial history, there is no single correct future prediction. Multiple trajectories are plausible and socially-acceptable.

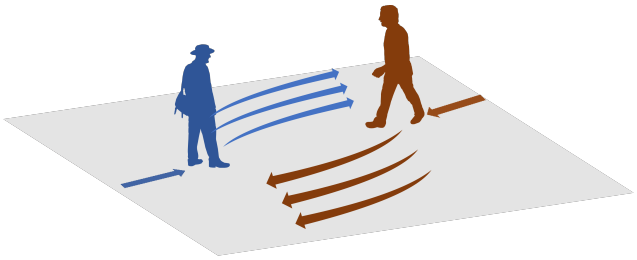

Figure 1: Illustration of a scenario where two pedestrians want to avoid each other. There are many possible ways that they can avoid a potential collision. We present a method that given the same observed past, predicts multiple socially acceptable outputs in crowded scenes.

Pioneering work in trajectory prediction has tackled some of the above challenges. The interpersonal aspect has been exhaustively addressed by traditional methods based on hand-crafted features [2, 17, 41, 46]. Social acceptability has been recently revisited with data-driven techniques based on Recurrent Neural Networks (RNNs) [1, 28, 12, 4]. Finally, the multimodal aspect of the problem has been studied in the context of route choices given a static scene (*e.g.*, which streets to take at an intersection [28, 24]). Robicquet *et al.* [38] have shown that pedestrians have multiple navigation styles in crowded scenes given a mild or aggressive style of navigation. Therefore, the forecasting task entails outputting different possible outcomes.

While existing methods have made great progress in addressing specific challenges, they suffer from two limitations. First, they model a local neighborhood around each person when making the prediction. Hence, they do not have the capacity to model interactions between all people in a scene in a computationally efficient fashion. Second, they tend to learn the "average behavior" because of the

commonly used loss function that minimizes the euclidean distance between the ground truth and forecasted outputs. In contrast, we aim in learning multiple "good behaviors", *i.e.*, multiple socially acceptable trajectories.

To address the limitations of previous works, we propose to leverage the recent progress in generative models. Generative Adversarial Networks (GANs) have been recently developed to overcome the difficulties in approximating intractable probabilistic computation and behavioral inference [14]. While they have been used to produce photo-realistic signals such as images [34], we propose to use them to generate multiple socially-acceptable trajectories given an observed past. One network (the generator) generates candidates and the other (the discriminator) evaluates them. The adversarial loss enables our forecasting model to go beyond the limitation of L2 loss and potentially learn the distribution of "good behaviors" that can fool the discriminator. In our work, these behaviors are referred to as *socially-accepted* motion trajectories in crowded scenes.

Our proposed GAN is a RNN Encoder-Decoder generator and a RNN based encoder discriminator with the following two novelties: (i) we introduce a variety loss which encourages the generative network of our GAN to spread its distribution and cover the space of possible paths while being consistent with the observed inputs. (ii) We propose a new pooling mechanism that learns a "global" pooling vector which encodes the subtle cues for all people involved in a scene. We refer to our model as "Social GAN". Through experiments on several publicly available real-world crowd datasets, we show state-of-the-art accuracy, speed and demonstrate that our model has the capacity to generate a variety of socially-acceptable trajectories.

2. Related Work

Research in forecasting human behavior can be grouped as learning to predict human-space interactions or human-human interactions. The former learns scene-specific motion patterns [3, 9, 18, 21, 24, 33, 49]. The latter models the dynamic content of the scenes, *i.e*. how pedestrians interact with each other. The focus of our work is the latter: learning to predict human-human interactions. We discuss existing work on this topic as well as relevant work in RNN for sequence prediction and Generative models.

Human-Human Interaction. Human behavior has been studied from a crowd perspective in *macroscopic models* or from a individual perspective in *microscopic models* (the focus of our work). One example of microscopic model is the Social Forces by Helbing and Molnar [17] which models pedestrian behavior with attractive forces guiding them towards their goal and repulsive forces encouraging collision avoidance. Over the past decades, this method has been often revisited [5, 6, 25, 26, 30, 31, 36, 46]. Tools popular in economics have also been used such as the Discrete Choice framework by Antonini *et. al.* [2]. Treuille *et. al.* [42] use continuum dynamics, and Wang *et. al.* [44], Tay *et. al.* [41] use Gaussian processes. Such functions have also been used to study stationary groups [35, 47]. However, all these methods use hand crafted energy potentials based on relative distances and specific rules. In contrast, over the past two years, data-driven methods based on RNNs have been used to outperform the above traditional ones.

RNNs for Sequence Prediction. Recurrent Neural Networks are a rich class of dynamic models which extend feedforward networks for sequence generation in diverse domains like speech recognition [7, 8, 15], machine translation [8] and image captioning [20, 43, 45, 39]. However, they lack high-level and spatio-temporal structure [29]. Several attempts have been made to use multiple networks to capture complex interactions [1, 10, 40]. Alahi *et al*. [1] use a social pooling layer that models nearby pedestrians. In the rest of this paper, we show that using a Multi-Layer Perceptron (MLP) followed by max pooling is computationally more efficient and works as well or better than the social pooling method from [1]. Lee *et al.* [28] introduce a RNN Encoder-Decoder framework which uses variational autoencoder (VAE) for trajectory prediction. However, they did not model human-human interactions in crowded scenes.

Generative Modeling. Generative models like variational autoencoders [23] are trained by maximizing the lower bound of training data likelihood. Goodfellow *et al*. [14] propose an alternative approach, Generative Adversarial Networks (GANs), where the training procedure is a minimax game between a generative model and a discriminative model; this overcomes the difficulty of approximating intractable probabilistic computations. Generative models have shown promising results in tasks like super-resolution [27], image to image translation [19], and image synthesis [16, 34, 48] which have multiple possible outputs for a given input. However, their application in sequence generation problems like natural language processing has lagged since sampling from these generated outputs to feed to the discriminator is a non-differentiable operation.

3. Method

Humans possess an intuitive ability to navigate crowds taking into account the people around them. We plan our paths keeping in mind our goal and also simultaneously taking into account the motion of surrounding people like their direction of motion, velocity, etc. However, often in such situations multiple possible options exist. We need models which not only can understand these complex human interactions but can also capture the variety of options. Current approaches have focused on predicting the average future trajectory which minimizes the $L2$ distance from the ground truth future trajectory whereas we want to predict multiple "good" trajectories. In this section, we first present our

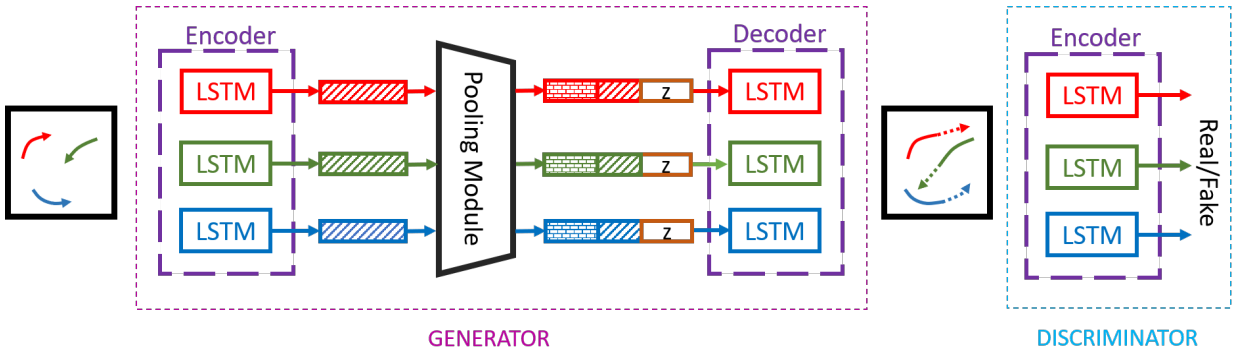

Figure 2: System overview. Our model consists of three key components: Generator (G), Pooling Module, and Discriminator (D). G takes as input past trajectories X_i and encodes the history of the person i as H_i^t. The pooling module takes as input all $H_i^{t_{obs}}$ and outputs a pooled vector P_i for each person. The decoder generates the future trajectory conditioned on $H_i^{t_{obs}}$ and P_i. D takes as input T_{real} or T_{fake} and classifies them as socially acceptable or not (see Figure 3 for PM).

GAN based encoder-decoder architecture to address this issue, we then describe our novel pooling layer which models human-human interactions and finally we introduce our variety loss which encourages the network to produce multiple diverse future trajectories for the same observed sequence.

3.1. Problem Definition

Our goal is to **jointly** reason and predict the future trajectories of **all** the agents involved in a scene. We assume that we receive as input all the trajectories for people in a scene as $\mathbf{X} = X_1, X_2, ..., X_n$ and predict the future trajectories $\hat{\mathbf{Y}} = \hat{Y}_1, \hat{Y}_2, ..., \hat{Y}_n$ of all the people **simultaneously**. The input trajectory of a person i is defined as $X_i = (x_i^t, y_i^t)$ from time steps $t = 1, ..., t_{obs}$ and the future trajectory (ground truth) can be defined similarly as $Y_i = (x_i^t, y_i^t)$ from time steps $t = t_{obs} + 1, ..., t_{pred}$. We denote predictions as $\hat{Y}_i$.

3.2. Generative Adversarial Networks

A Generative Adversarial Network (GAN) consists of two neural networks trained in opposition to each other [14]. The two adversarially trained models are: a generative model G that captures the data distribution, and a discriminative model D that estimates the probability that a sample came from the training data rather than G. The generator G takes a latent variable z as input, and outputs sample $G(z)$. The discriminator D takes a sample x as input and outputs $D(x)$ which represents the probability that it is real. The training procedure is similar to a two-player min-max game with the following objective function:

$$\min_G \max_D V(G, D) = \mathbb{E}_{x \sim p_{data}(x)}[\log D(x)] + \mathbb{E}_{z \sim p_{(z)}}[\log(1 - D(G(z)))]. \quad (1)$$

GANs can used for conditional models by providing both the generator and discriminator with additional input c, yielding $G(z, c)$ and $D(x, c)$ [13, 32].

3.3. Socially-Aware GAN

As discussed in Section 1 trajectory prediction is a multimodal problem. Generative models can be used with time-series data to simulate possible futures. We leverage this insight in designing SGAN which addresses the multimodality of the problem using GANs (see Figure 2). Our model consists of three key components: Generator (G), Pooling Module (PM) and Discriminator (D). G is based on encoder-decoder framework where we link the hidden states of encoder and decoder via PM. G takes as input X_i and outputs predicted trajectory $\hat{Y}_i$. D inputs the entire sequence comprising both input trajectory X_i and future prediction $\hat{Y}_i$ (or Y_i) and classifies them as "real/fake".

Generator. We first embed the location of each person using a single layer MLP to get a fixed length vector e_i^t. These embeddings are used as input to the LSTM cell of the encoder at time t introducing the following recurrence:

$$e_i^t = \phi(x_i^t, y_i^t; W_{ee})$$
$$h_{ei}^t = LSTM(h_{ei}^{t-1}, e_i^t; W_{encoder}) \quad (2)$$

where $\phi(\cdot)$ is an embedding function with ReLU nonlinearity, W_{ee} is the embedding weight. The LSTM weights ($W_{encoder}$) are shared between all people in a scene.

Naïve use of one LSTM per person fails to capture interaction between people. Encoder learns the state of a person and stores their history of motion. However, as shown by Alahi *et al.* [1] we need a compact representation which combines information from different encoders to effectively reason about social interactions. In our method, we model human-human interaction via a Pooling Module (PM). After t_{obs} we pool hidden states of all the people present in the scene to get a pooled tensor P_i for each person. Traditionally, GANs take as input noise and generate samples. Our goal is to produce future scenarios which are consistent with the past. To achieve this we condition the generation of output trajectories by initializing the hidden state of the

decoder as:

$$c_i^t = \gamma(P_i, h_{ei}^t; W_c)$$
$$h_{di}^t = [c_i^t, z] \tag{3}$$

where $\gamma(\cdot)$ is a multi-layer perceptron (MLP) with ReLU non-linearity and W_c is the embedding weight. We deviate from prior work in two important ways regarding trajectory prediction:

- Prior work [1] uses the hidden state to predict parameters of a bivariate Gaussian distribution. However, this introduces difficulty in the training process as backpropagation through sampling process in non-differentiable. We avoid this by directly predicting the coordinates $(\hat{x}_i^t, \hat{y}_i^t)$.
- "Social" context is generally provided as input to the LSTM cell [1, 28] . Instead we provide the pooled context only once as input to the decoder. This also provides us with the ability to choose to pool at specific time steps and results in **16x** speed increase as compared to S-LSTM [1] (see Table 2).

After initializing the decoder states as described above we can obtain predictions as follows:

$$\begin{aligned} e_i^t &= \phi(x_i^{t-1}, y_i^{t-1}; W_{ed}) \\ P_i &= PM(h_{d1}^{t-1}, ..., h_{dn}^t) \\ h_{di}^t &= LSTM(\gamma(P_i, h_{di}^{t-1}), e_i^t; W_{decoder}) \\ (\hat{x}_i^t, \hat{y}_i^t) &= \gamma(h_{di}^t) \end{aligned} \tag{4}$$

where $\phi(\cdot)$ is an embedding function with ReLU non-linearity with W_{ed} as the embedding weights. The LSTM weights are denoted by $W_{decoder}$ and γ is an MLP.

Discriminator. The discriminator consists of a separate encoder. Specifically, it takes as input $T_{real} = [X_i, Y_i]$ or $T_{fake} = [X_i, \hat{Y}_i]$ and classifies them as real/fake. We apply a MLP on the encoder's last hidden state to obtain a classification score. The discriminator will ideally learn subtle social interaction rules and classify trajectories which are not socially acceptable as "fake".

Losses. In addition to adversarial loss, we also apply $L2$ loss on the predicted trajectory which measures how far the generated samples are from the actual ground truth.

3.4. Pooling Module

In order to jointly reason across multiple people we need a mechanism to share information across LSTMs. However, there are several challenges which a method should address:

- Variable and (potentially) large number of people in a scene. We need a compact representation which combines information from all the people.
- Scattered Human-Human Interaction. Local information is not always sufficient. Far-away pedestrians might impact each others. Hence, the network needs to model global configuration.

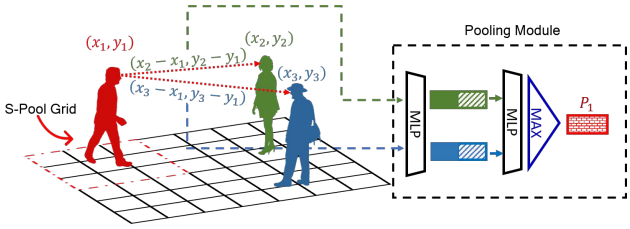

Figure 3: Comparison between our pooling mechanism (red dotted arrows) and Social Pooling [1] (red dashed grid) for the red person. Our method computes relative positions between the red and all other people; these positions are concatenated with each person's hidden state, processed independently by an MLP, then pooled elementwise to compute red person's pooling vector P_1. Social pooling only considers people inside the grid, and cannot model interactions between all pairs of people.

Social Pooling [1] addresses the first issue by proposing a grid based pooling scheme. However, this hand-crafted solution is slow and fails to capture global context. Qi *et al.* [37] show that above properties can be achieved by applying a learned symmetric function on transformed elements of the input set of points. As shown in Figure 2 this can be achieved by passing the input coordinates through a MLP followed by a symmetric function (we use Max-Pooling). The pooled vector P_i needs to summarize all the information a person needs to make a decision. Since, we use relative coordinates for translation invariance we augment the input to the pooling module with relative position of each person with respect to person i.

3.5. Encouraging Diverse Sample Generation

Trajectory prediction is challenging as given limited past history a model has to reason about multiple possible outcomes. The method described so far produces good predictions, but these predictions try to produce the "average" prediction in cases where there can be multiple outputs. Further, we found that outputs were not very sensitive to changes in noise and produced very similar predictions.

We propose a variety loss function that encourages the network to produce diverse samples. For each scene we generate k possible output predictions by randomly sampling z from $\mathcal{N}(0, 1)$ and choosing the "best" prediction in $L2$ sense as our prediction.

$$\mathcal{L}_{variety} = \min_k \|Y_i - \hat{Y}_i^{(k)}\|_2, \tag{5}$$

where k is a hyperparameter.

By considering only the best trajectory, this loss encourages the network to hedge its bets and cover the space of outputs that conform to the past trajectory. The loss is structurally akin to Minimum over N (MoN) loss [11] but to the

Metric	Dataset	Linear	LSTM	S-LSTM [1]	SGAN (Ours) 1V-1	SGAN (Ours) 1V-20	SGAN (Ours) 20V-20	SGAN (Ours) 20VP-20
ADE	**ETH**	0.84 / 1.33	0.70 / 1.09	0.73 / 1.09	0.79 / 1.13	0.75 / 1.03	0.61 / **0.81**	**0.60** / 0.87
	HOTEL	**0.35** / **0.39**	0.55 / 0.86	0.49 / 0.79	0.71 / 1.01	0.63 / 0.90	0.48 / 0.72	0.52 / 0.67
	UNIV	0.56 / 0.82	0.36 / 0.61	0.41 / 0.67	0.37 / 0.60	0.36 / 0.58	**0.36** / **0.60**	0.44 / 0.76
	ZARA1	0.41 / 0.62	0.25 / 0.41	0.27 / 0.47	0.25 / 0.42	0.23 / 0.38	**0.21** / **0.34**	0.22 / 0.35
	ZARA2	0.53 / 0.77	0.31 / 0.52	0.33 / 0.56	0.32 / 0.52	0.29 / 0.47	**0.27** / 0.42	0.29 / **0.42**
AVG		0.54 / 0.79	0.43 / 0.70	0.45 / 0.72	0.49 / 0.74	0.45 / 0.67	**0.39** / **0.58**	0.41 / 0.61
FDE	**ETH**	1.60 / 2.94	1.45 / 2.41	1.48 / 2.35	1.61 / 2.21	1.52 / 2.02	1.22 / **1.52**	**1.19** / 1.62
	HOTEL	**0.60** / **0.72**	1.17 / 1.91	1.01 / 1.76	1.44 / 2.18	1.32 / 1.97	0.95 / 1.61	1.02 / 1.37
	UNIV	1.01 / 1.59	0.77 / 1.31	0.84 / 1.40	0.75 / 1.28	**0.73** / **1.22**	0.75 / 1.26	0.84 / 1.52
	ZARA1	0.74 / 1.21	0.53 / 0.88	0.56 / 1.00	0.53 / 0.91	0.48 / 0.84	**0.42** / 0.69	0.43 / **0.68**
	ZARA2	0.95 / 1.48	0.65 / 1.11	0.70 / 1.17	0.66 / 1.11	0.61 / 1.01	**0.54** / **0.84**	0.58 / 0.84
AVG		0.98 / 1.59	0.91 / 1.52	0.91 / 1.54	1.00 / 1.54	0.93 / 1.41	**0.78** / **1.18**	0.81 / 1.21

Table 1: Quantitative results of all methods across datasets. We report two error metrics Average Displacement Error (ADE) and Final Displacement Error (FDE) for $t_{pred} = 8$ and $t_{pred} = 12$ (8 / 12) in meters. Our method consistently outperforms state-of-the-art S-LSTM method and is especially good for long term predictions (lower is better).

best of our knowledge this has not been used in the context of GANs to encourage diversity of generated samples.

3.6. Implementation Details

We use LSTM as the RNN in our model for both decoder and encoder. The dimensions of the hidden state for encoder is 16 and decoder is 32. We embed the input coordinates as 16 dimensional vectors. We iteratively train the Generator and Discriminator with a batch size of 64 for 200 epochs using Adam [22] with an initial learning rate of 0.001.

4. Experiments

In this section, we evaluate our method on two publicly available datasets: ETH [36] and UCY [25]. These datasets consist of real world human trajectories with rich human-human interaction scenarios. We convert all the data to real world coordinates and interpolate to obtain values at every 0.4 seconds. In total there are 5 sets of data (ETH - 2, UCY-3) with 4 different scenes which consists of 1536 pedestrians in crowded settings with challenging scenarios like group behavior, people crossing each other, collision avoidance and groups forming and dispersing.

Evaluation Metrics. Similar to prior work [1, 28] we use two error metrics:

1. *Average Displacement Error (ADE)*: Average $L2$ distance between ground truth and our prediction over all predicted time steps.
2. *Final Displacement Error (FDE)*: The distance between the predicted final destination and the true final destination at end of the prediction period T_{pred}.

Baselines: We compare against the following baselines:

1. *Linear*: A linear regressor that estimates linear parameters by minimizing the least square error.
2. *LSTM*: A simple LSTM with no pooling mechanism.
3. *S-LSTM*: The method proposed by Alahi *et al*. [1]. Each person is modeled via an LSTM with the hidden states being pooled at each time step using the social pooling layer.

We also do an ablation study of our model with different control settings. We refer our full method in the section as *SGAN-kVP-N* where kV signifies if the model was trained using variety loss ($k = 1$ essentially means no variety loss) and P signifies usage of our proposed pooling module. At test time we sample multiple times from the model and chose the best prediction in $L2$ sense for quantitative evaluation. N refers to the number of time we sample from our model during test time.

Evaluation Methodology. We follow similar evaluation methodology as [1]. We use leave-one-out approach, train on 4 sets and test on the remaining set. We observe the trajectory for 8 times steps (3.2 seconds) and show prediction results for 8 (3.2 seconds) and 12 (4.8 seconds) time steps.

4.1. Quantitative Evaluation

We compare our method on two metrics ADE and FDE against different baselines in Table 1. As expected Linear model is only capable of modeling straight paths and does especially bad in case of longer predictions ($t_{pred} = 12$). Both LSTM and S-LSTM perform much better than the linear baseline as they can model more complex trajectories. However, in our experiments S-LSTM does not outperform LSTM. We tried our best to reproduce the results of the pa-

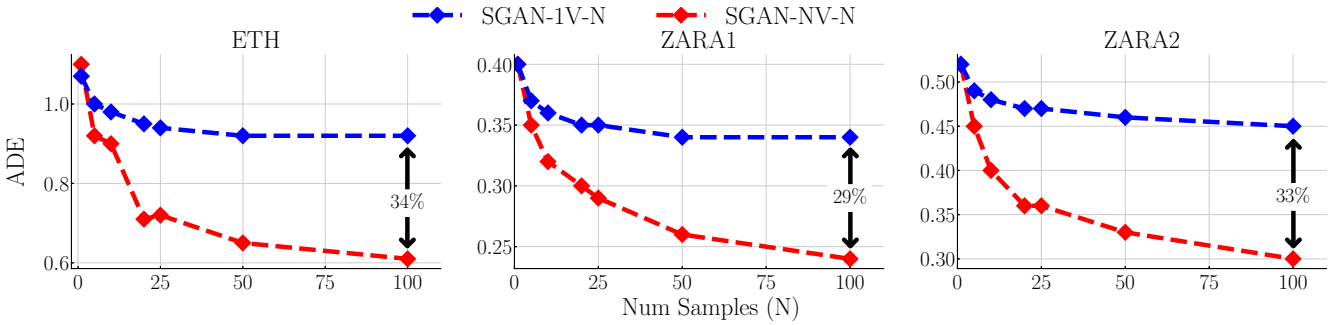

Figure 4: Effect of variety loss. For SGAN-1V-N we train a single model, drawing one sample for each sequence during training and N samples during testing. For SGAN-NV-N we train several models with our variety loss, using N samples during both training and testing. Training with the variety loss significantly improves accuracy.

per. [1] trained the model on synthetic dataset and then fine-tuned on real datasets. We don't use synthetic data to train any of our models which could potentially lead to worse performance.

SGAN-1V-1 performs worse than LSTM as each predicted sample can be any of the multiple possible future trajectories. The conditional output generated by the model represents one of many plausible future predictions which might be different from ground truth prediction. When we consider multiple samples our model outperforms the baseline methods confirming the multi-modal nature of the problem. GANs face mode collapse problem, where the generator resorts to generating a handful of samples which are assigned high probability by the discriminator. We found that samples generated by SGAN-1V-1 didn't capture all possible scenarios. However, SGAN-20V-20 significantly outperforms all other models as the variety loss encourages the network to produce diverse samples. Although our full model with proposed pooling layer performs slightly worse we show in the next section that pooling layer helps the model predict more "socially" plausible paths.

Speed. Speed is crucial for a method to be used in a real-world setting like autonomous vehicles where you need accurate predictions about pedestrian behavior. We compare our method with two baselines LSTM and S-LSTM. A simple LSTM performs the fastest but can't avoid collisions or make accurate multi-modal predictions. Our method is **16x** faster than S-LSTM (see Table 2). Speed improvement is because we don't do pooling at each time step. Also, unlike S-LSTM which requires computing a occupancy grid for each pedestrian our pooling mechanism is a simple MLP followed by max pooling. In real-world applications our model can quickly generate 20 samples in the same time it takes S-LSTM to make 1 prediction.

Evaluating Effect of Diversity. One might wonder what will happen if we simply draw more samples from our model without the variety loss? We compare the performance of SGAN-1V-N with SGAN-NV-N. As a reminder

	LSTM	**S-LSTM**	**SGAN**	**SGAN-P**
8	0.02	1.79	0.04	0.12
12	0.03	2.61	0.05	0.15
Speed-Up	82x	1x	**49x**	**16x**

Table 2: Speed (in seconds) comparison with S-LSTM. We get 16x speedup as compared to S-LSTM allowing us to draw 16 samples in the same time S-LSTM makes a single prediction. Unlike S-LSTM we don't perform pooling at each time step resulting in significant speed bump without suffering on accuracy. All methods are benchmarked on Tesla P100 GPU

SGAN-NV-N refers to a model trained with variety loss with $k = N$ and drawing N samples during testing. As shown in Figure 4 across all datasets simply drawing more samples from the model trained without variety loss does not lead to better accuracy. Instead, we see a significant performance increase as we increase k with models on average performing 33% better with $k = 100$.

4.2. Qualitative Evaluation

In multi-agent (people) scenarios, it is imperative to model how actions of one person can influence the actions of other people. Traditional approaches for activity forecasting and human trajectory prediction have focused on hand crafted energy potentials modeling attractive and repulsive forces to model these complex interactions. We use a purely data driven approach which models human-human interaction via a novel pooling mechanism. Humans walking in the presence of other people plan their path taking into account their personal space, perceived potential for collision, final destination and their own past motion. In this section, we first evaluate the effect of the pooling layer and then analyze the predictions made by our network in three common social interaction scenarios. Even though our model makes **joint** predictions for **all** people in a scene we show predictions for a subset for simplicity. We refer to

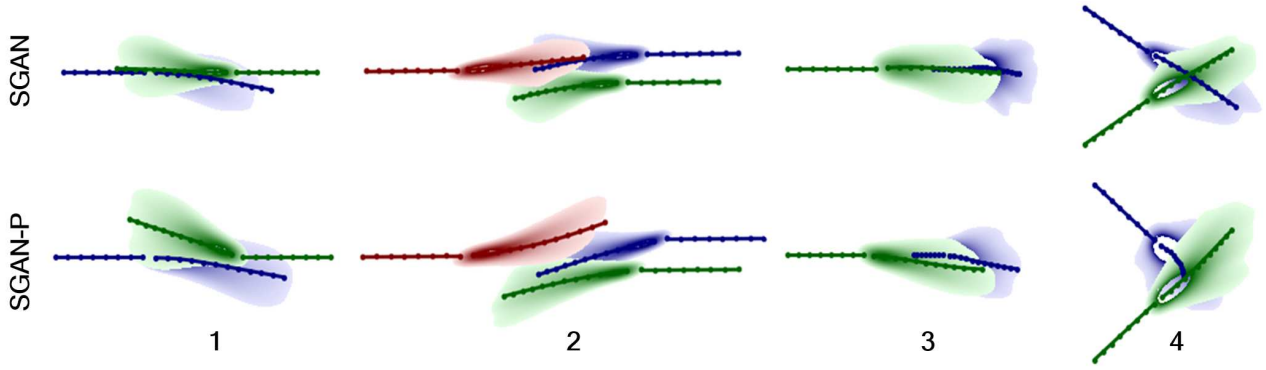

Figure 5: Comparison between our model without pooling (SGAN, top) and with pooling (SGAN-P, bottom) in four collision avoidance scenarios: two people meeting (1), one person meeting a group (2), one person behind another (3), and two people meeting at an angle (4). For each example we draw 300 samples from the model and visualize their density and mean. Due to pooling, SGAN-P predicts socially acceptable trajectories which avoid collisions.

each person in the scene by the first letter of the color in the figure (e.g., Person B (Black), Person R (Red) and so on). Also for simplicity we refer SGAN-20VP-20 as SGAN-P and SGAN-20V-20 as SGAN.

4.2.1 Pooling Vs No-Pooling

On quantitative metrics both methods perform similarly with SGAN slightly outperforming SGAN-P (see Table 1. However, qualitatively we find that pooling enforces a global coherency and conformity to social norms. We compare how SGAN and SGAN-P perform in four common social interaction scenarios (see Figure 5). We would like to highlight that even though these scenarios were created synthetically, we used models trained on real world data. Moreover, these scenarios were created to evaluate the models and nothing in our design makes these scenarios particularly easy or hard. For each setup we draw 300 samples and plot an approximate distribution of trajectories along with average trajectory prediction.

Scenario 1 and 2 depict the collision avoidance capacity of our model by changing direction. In the case of two people heading in the same direction pooling enables the model to predict a socially accepted way of yielding the right of way towards the right. However, SGAN prediction leads to a collision. Similarly, unlike SGAN, SGAN-P is able to model group behavior and predict avoidance while preserving the notion of couple walking together (Scenario 2).

Humans also tend to vary pace to avoid collisions. Scenario 3 is depicts a person G walking behind person B albeit faster. If they both continue to maintain their pace and direction they would collide. Our model predicts person G overtaking from the right. SGAN fails to predict a socially acceptable path. In Scenario 4, we notice that the model predicts person B slowing down and yielding for person G.

4.2.2 Pooling in Action

We consider three real-scenarios where people have to alter their course to avoid collision (see Figure 6).

People Merging. (Row 1) In hallways or in roads it is common for people coming from different directions to merge and walk towards a common destination. People use various ways to avoid colliding while continuing towards their destination. For instance a person might slow down, alter their course slightly or use a combination of both depending on the context and behavior of other surrounding people. Our model is able predict variation in both speed and direction of a person to effectively navigate a situation. For instance model predicts that either person B slows down (col 2) or both person B and R change direction to avoid collision. The last prediction (col 4) is particularly interesting as the model predicts a sudden turn for person R but also predicts that person B significantly slows down in response; thus making a globally consistent prediction.

Group Avoiding. (Row 2) People avoiding each other when moving in opposite direction is another common scenario. This can manifest in various forms like a person avoiding a couple, a couple avoiding a couple etc. To make correct predictions in such cases a person needs to plan ahead and look beyond it's immediate neighborhood. Our model is able to recognize that the people are moving in groups and model group behavior. The model predicts change of direction for either groups as a way of avoiding collision (col 3, 4). In contrast to Figure 5 even though the convention might be to give way to the right in this particular situation that would lead to a collision. Hence, our models makes prediction where couples give way towards the left.

Person Following. (Row 3) Another common scenario is when a person is walking behind someone. One might want to either maintain pace or maybe overtake the person

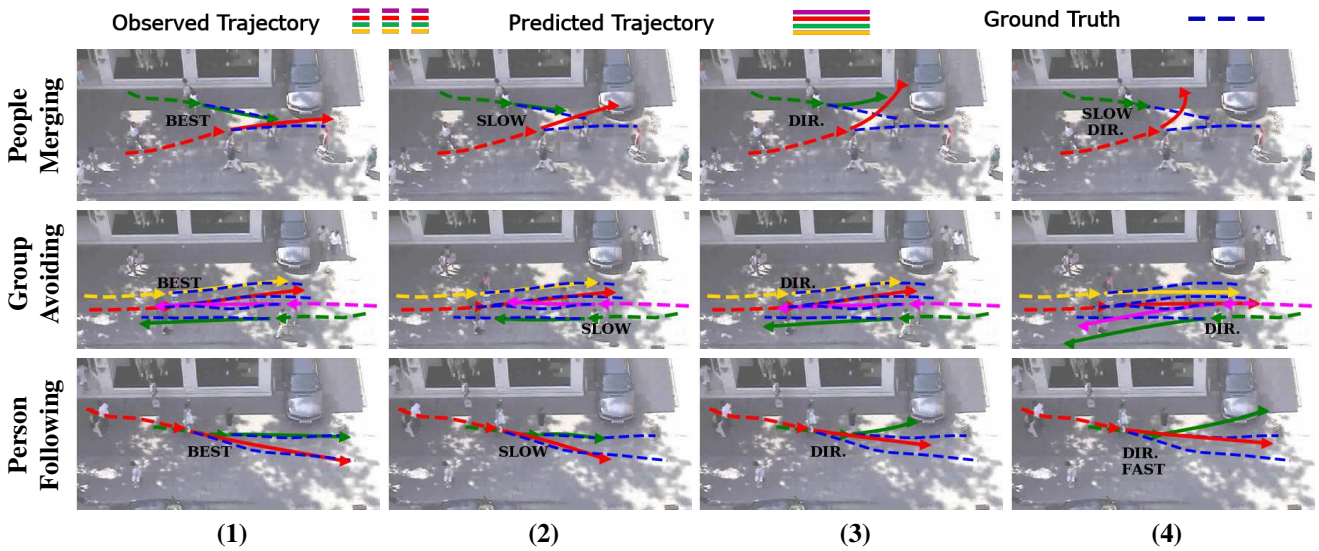

Figure 6: Examples of diverse predictions from our model. Each row shows a different set of observed trajectories; columns show four different samples from our model for each scenario which demonstrate different types of socially acceptable behavior. BEST is the sample closest to the ground-truth; in SLOW and FAST samples, people change speed to avoid collision; in DIR samples people change direction to avoid each other. Our model learns these different avoidance strategies in a data-driven manner, and jointly predicts globally consistent and socially acceptable trajectories for all people in the scene. We also show some failure cases in supplementary material.

in front. We would like to draw attention to a subtle difference between this situation and its real-life counterpart. In reality a person's decision making ability is restricted by their field of view. In contrast, our model has access to ground truth positions of all the people involved in the scene at the time of pooling. This manifests in some interesting cases (see col 3). The model understands that person R is behind person B and is moving faster. Consequently, it predicts that person B gives way by changing their direction and person R maintains their direction and speed. The model is also able to predict overtaking (ground truth).

4.3. Structure in Latent Space

In this experiment we attempt to understand the landscape of the latent space z. Walking on the manifold that is learnt can give us insights about how the model is able to generate diverse samples. Ideally, one can expect that the network imposes some structure in the latent space. We found that certain directions in the latent space were associated with direction and speed (Figure 7).

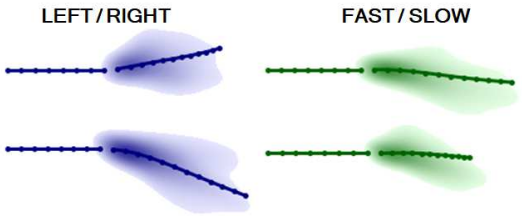

Figure 7: Latent Space Exploration. Certain directions in the latent manifold are associated with direction (left) and speed (right). Observing the same past but varying the input z along different directions causes the model to predict trajectories going either right/left or fast/slow on average.

5. Conclusion

In this work we tackle the problem of modeling human-human interaction and jointly predicting trajectories for all people in a scene. We propose a novel GAN based encoder-decoder framework for trajectory prediction capturing the multi-modality of the future prediction problem. We also propose a novel pooling mechanism enabling the network to learn social norms in a purely data-driven approach. To encourage diversity among predicted samples we propose a simple variety loss which coupled with the pooling layer encourages the network to produce globally coherent, socially compliant diverse samples. We show the efficacy of our method on several complicated real-life scenarios where social norms must be followed.

6. Acknowledgment

This work was supported by Toyota (1186781-31-UDARO), ONR (1165419-10-TDAUZ), Nvidia and MURI (1186514-1-TBCJE). We thank Jayanth Koushik and De-An Huang for their helpful comments and suggestions.

References

[1] 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*, pages 961–971, 2016. 1, 2, 3, 4, 5, 6

[2] G. Antonini, M. Bierlaire, and M. Weber. Discrete choice models of pedestrian walking behavior. *Transportation Research Part B: Methodological*, 40(8):667–687, 2006. 1, 2

[3] L. Ballan, F. Castaldo, A. Alahi, F. Palmieri, and S. Savarese. Knowledge transfer for scene-specific motion prediction. In *European Conference on Computer Vision*, pages 697–713. Springer, 2016. 2

[4] F. Bartoli, G. Lisanti, L. Ballan, and A. Del Bimbo. Context-aware trajectory prediction. *arXiv preprint arXiv:1705.02503*, 2017. 1

[5] W. Choi and S. Savarese. A unified framework for multi-target tracking and collective activity recognition. In *Computer Vision–ECCV 2012*, pages 215–230. Springer, 2012. 2

[6] W. Choi and S. Savarese. Understanding collective activitiesof people from videos. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 36(6):1242–1257, 2014. 2

[7] J. Chorowski, D. Bahdanau, K. Cho, and Y. Bengio. End-to-end continuous speech recognition using attention-based recurrent nn: First results. *arXiv preprint arXiv:1412.1602*, 2014. 2

[8] J. Chung, K. Kastner, L. Dinh, K. Goel, A. C. Courville, and Y. Bengio. A recurrent latent variable model for sequential data. In *Advances in neural information processing systems*, pages 2980–2988, 2015. 2

[9] P. Coscia, F. Castaldo, F. A. Palmieri, L. Ballan, A. Alahi, and S. Savarese. Point-based path prediction from polar histograms. In *Information Fusion (FUSION), 2016 19th International Conference on*, pages 1961–1967. IEEE, 2016. 2

[10] Y. Du, W. Wang, and L. Wang. Hierarchical recurrent neural network for skeleton based action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1110–1118, 2015. 2

[11] H. Fan, H. Su, and L. Guibas. A point set generation network for 3d object reconstruction from a single image. *arXiv preprint arXiv:1612.00603*, 2016. 4

[12] 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. 1

[13] J. Gauthier. Conditional generative adversarial nets for convolutional face generation. *Class Project for Stanford CS231N: Convolutional Neural Networks for Visual Recognition, Winter semester*, 2014(5):2, 2014. 3

[14] 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*, pages 2672–2680, 2014. 2, 3

[15] A. Graves and N. Jaitly. Towards end-to-end speech recognition with recurrent neural networks. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 1764–1772, 2014. 2

[16] K. Gregor, I. Danihelka, A. Graves, D. J. Rezende, and D. Wierstra. Draw: A recurrent neural network for image generation. *arXiv preprint arXiv:1502.04623*, 2015. 2

[17] D. Helbing and P. Molnar. Social force model for pedestrian dynamics. *Physical review E*, 51(5):4282, 1995. 1, 2

[18] W. Hu, D. Xie, Z. Fu, W. Zeng, and S. Maybank. Semantic-based surveillance video retrieval. *Image Processing, IEEE Transactions on*, 16(4):1168–1181, 2007. 2

[19] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. *arXiv preprint arXiv:1611.07004*, 2016. 2

[20] A. Karpathy, A. Joulin, and F. F. F. Li. Deep fragment embeddings for bidirectional image sentence mapping. In *Advances in neural information processing systems*, pages 1889–1897, 2014. 2

[21] K. Kim, D. Lee, and I. Essa. Gaussian process regression flow for analysis of motion trajectories. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 1164–1171. IEEE, 2011. 2

[22] D. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[23] D. P. Kingma and M. Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 2

[24] K. M. Kitani, B. D. Ziebart, J. A. Bagnell, and M. Hebert. Activity forecasting. In *Computer Vision–ECCV 2012*, pages 201–214. Springer, 2012. 1, 2

[25] L. Leal-Taixé, M. Fenzi, A. Kuznetsova, B. Rosenhahn, and S. Savarese. Learning an image-based motion context for multiple people tracking. In *CVPR*, pages 3542–3549. IEEE, 2014. 2, 5

[26] L. Leal-Taixe, G. Pons-Moll, and B. Rosenhahn. Everybody needs somebody: Modeling social and grouping behavior on a linear programming multiple people tracker. In *ICCV Workshops*, 2011. 2

[27] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. *arXiv preprint arXiv:1609.04802*, 2016. 2

[28] 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. *arXiv preprint arXiv:1704.04394*, 2017. 1, 2, 4, 5

[29] J. Liu, A. Shahroudy, D. Xu, and G. Wang. Spatio-temporal lstm with trust gates for 3d human action recognition. In *European Conference on Computer Vision*, pages 816–833. Springer, 2016. 2

[30] M. Luber, J. A. Stork, G. D. Tipaldi, and K. O. Arras. People tracking with human motion predictions from social forces. In *Robotics and Automation (ICRA), 2010 IEEE International Conference on*, pages 464–469. IEEE, 2010. 2

[31] R. Mehran, A. Oyama, and M. Shah. Abnormal crowd behavior detection using social force model. In *Computer Vi-*

sion and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on, pages 935–942. IEEE, 2009. 2

[32] M. Mirza and S. Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014. 3

[33] B. T. Morris and M. M. Trivedi. Trajectory learning for activity understanding: Unsupervised, multilevel, and long-term adaptive approach. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 33(11):2287–2301, 2011. 2

[34] A. Odena, C. Olah, and J. Shlens. Conditional image synthesis with auxiliary classifier gans. *arXiv preprint arXiv:1610.09585*, 2016. 2

[35] H. S. Park and J. Shi. Social saliency prediction. 2

[36] S. Pellegrini, A. Ess, and L. Van Gool. Improving data association by joint modeling of pedestrian trajectories and groupings. In *Computer Vision–ECCV 2010*, pages 452–465. Springer, 2010. 2, 5

[37] C. R. Qi, H. Su, K. Mo, and L. J. Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. *arXiv preprint arXiv:1612.00593*, 2016. 4

[38] A. Robicquet, A. Sadeghian, A. Alahi, and S. Savarese. Learning social etiquette: Human trajectory understanding in crowded scenes. In *European conference on computer vision*, pages 549–565. Springer, 2016. 1

[39] T. Shu, S. Todorovic, and S.-C. Zhu. Cern: confidence-energy recurrent network for group activity recognition. *Proc. of CVPR, Honolulu, Hawaii*, 2017. 2

[40] N. Srivastava, E. Mansimov, and R. Salakhudinov. Unsupervised learning of video representations using lstms. In *International Conference on Machine Learning*, pages 843–852, 2015. 2

[41] M. K. C. Tay and C. Laugier. Modelling smooth paths using gaussian processes. In *Field and Service Robotics*, pages 381–390. Springer, 2008. 1, 2

[42] A. Treuille, S. Cooper, and Z. Popović. Continuum crowds. In *ACM Transactions on Graphics (TOG)*, volume 25, pages 1160–1168. ACM, 2006. 2

[43] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164, 2015. 2

[44] J. M. Wang, D. J. Fleet, and A. Hertzmann. Gaussian process dynamical models for human motion. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 30(2):283–298, 2008. 2

[45] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on Machine Learning*, pages 2048–2057, 2015. 2

[46] K. Yamaguchi, A. C. Berg, L. E. Ortiz, and T. L. Berg. Who are you with and where are you going? In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 1345–1352. IEEE, 2011. 1, 2

[47] 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*, pages 3488–3496, 2015. 2

[48] H. Zhang, T. Xu, H. Li, S. Zhang, X. Huang, X. Wang, and D. Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. *arXiv preprint arXiv:1612.03242*, 2016. 2

[49] B. Zhou, X. Wang, and X. Tang. Random field topic model for semantic region analysis in crowded scenes from tracklets. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 3441–3448. IEEE, 2011. 2