SCAN: A Spatial Context Attentive Network for Joint Multi-Agent Intent Prediction

Jasmine Sekhon[1], **Cody Fleming**[2]
[1] University of Virginia, [2] Iowa State University
js3cn@virginia.edu, flemingc@iastate.edu

Abstract

Safe navigation of autonomous agents in human centric environments requires the ability to understand and predict motion of neighboring pedestrians. However, predicting pedestrian intent is a complex problem. Pedestrian motion is governed by complex social navigation norms, is dependent on neighbors' trajectories, and is multimodal in nature. In this work, we propose **SCAN**, a **S**patial **C**ontext **A**ttentive **N**etwork that can jointly predict socially-acceptable multiple future trajectories for all pedestrians in a scene. SCAN encodes the influence of spatially close neighbors using a novel spatial attention mechanism in a manner that relies on fewer assumptions, is parameter efficient, and is more interpretable compared to state-of-the-art spatial attention approaches. Through experiments on several datasets we demonstrate that our approach can also quantitatively outperform state of the art trajectory prediction methods in terms of accuracy of predicted intent.

Introduction

Modes of autonomous navigation are increasingly being adopted in land, marine and airborne vehicles. In all these domains, these autonomous vehicles are often expected to operate in human-centric environments (e.g. social robots, self-driving cars, etc.). When humans are navigating in crowded environments, they follow certain implicit rules of social interaction. As an example, when navigating in crowded spaces like sidewalks, airports, train stations, and others, pedestrians attempt to navigate safely while avoiding collision with other pedestrians, respecting others' personal space, yielding right-of-way, etc. Any autonomous agent attempting to navigate safely in such shared environments must be able to model these social navigation norms and understand neighbors' motion as a function of such complex spatial interactions. In this work, we aim to understand pedestrian interactions and model these towards *jointly* predicting future trajectories for multiple pedestrians navigating in a scene. The contributions of our work are three-fold:

- We introduce a novel spatial attention mechanism to model spatial influence of neighboring pedestrians in a manner that relies on fewer assumptions, is parameter efficient, and interpretable. We encode the spatial influences experienced by a pedestrian at a point of time into a *spatial context* vector.
- We propose **SCAN**, a **S**patial **C**ontext **A**ttentive **N**etwork, that jointly predicts trajectories for all pedestrians in the scene for a future time window by attending to spatial contexts experienced by them individually over an observed time window.
- Since human motion is multimodal, we extend our proposed framework to predicting multiple socially feasible paths for all pedestrians in the scene.

Related Work

Since a key contribution of our work is the ability of our proposed framework to model spatial interactions between neighboring pedestrians in a novel manner, we briefly discuss how existing trajectory forecasting methods encode spatial influences while predicting pedestrian intent.

Traditional methods have relied on hand-crafted functions and features to model spatial interactions. For instance, the Social Forces model [7] models pedestrian behavior with attractive forces encouraging moving towards their goal and repulsive forces discouraging collision with other pedestrians. Similarly, [2] and [26] proposed trajectory forecasting approaches that rely on features extracted from human trajectories or human attributes. Such methods are limited by the need to hand craft features and attributes and their simplistic models and lack generalizability to complex crowded settings. Further, they only model immediate collision-avoidance behavior and do not consider interactions that may occur in the more distant future.

More recently, deep learning based frameworks are being used to model spatial interactions between pedestrians. LSTM-based (Long short-term memory) approaches are well-suited to predict pedestrian trajectories owing to the sequential nature of the data. Consequently, several LSTM-based approaches have been proposed and successfully applied to predict pedestrian intent in the past. Alahi *et. al.* proposed Social LSTM [1] that uses a social pooling layer to encode spatial influences from neighboring pedestrians within an assumed spatial grid. More recently, Gupta *et. al.* proposed Social GAN [6], which goes beyond modeling only local interactions within a fixed spatial grid, and considers influence of every other pedestrian in the scene on the pedestrian of interest. However, they use maxpooling, which causes all neighboring agents to have an identical representation towards predicting intent for a pedestrian of interest. Therefore, their method treats the influence of all agents on each other uniformly. SophieGAN [17] eliminates this problem by using a sorting mechanism based on distance to create a feature representation to encode spatial influences of neighbors. This causes each neighbor to have its unique feature representation, and hence, all neighbors have different spatial influences on a pedestrian. However, two neighbors at the same distance from a pedestrian may have different spatial influences. For instance, a neighbor at a certain distance from the pedestrian of interest, but not in line-of-sight, may have negligible influence on it, in comparison to another neighbor at the same distance but approaching it head-on. Such factors, like orientation, are therefore, imperative towards encoding spatial influence.

Graph Attention Networks, proposed by Velickovic *et. al.* [22], allow for application of self-attention over any type of structured data that can be represented as a graph. Pedestrian interactions can be naturally represented as graphs, where nodes are pedestrians and edges are spatial interactions. Several attention-based graph approaches [10, 13, 3, 23] are used for modeling spatial interactions. At a very high level, graph attention networks compute weights for edges by using scoring mechanisms (e.g. dot product of the hidden states of the nodes connected by the edge). Such a scoring mechanism does not consider the effect of features such as distances, relative orientations, etc. on the spatial influence of a neighbor. In [23], Vemula *et. al.* proposed Social Attention that takes into account the effect of this relative orientation towards spatial influence by encoding this information in spatial edges of a spatio-temporal graph. Similarly, Social Ways [3] computes spatial influence of a neighbor as the scalar product of the hidden state of the neighbor and a feature vector that contains orientation features. A key disadvantage of such approaches is that the number of trainable parameters towards computing spatial interactions are proportional to the number of nodes in the graph. As we explain later, our proposed spatial interaction mechanism is able to model spatial influence such that the number of trainable parameters are independent of the number of nodes/pedestrians in the graph. Our proposed approach models spatial influence in a manner that is parameter efficient and more interpretable compared to existing approaches.

Proposed Approach

Given N pedestrians present in a given frame at the start of an observation time window, from t_0 to T_{obs}, our goal is to *jointly predict* socially plausible trajectories for each of the N pedestrians in the scene over a time window in the future, from $T_{obs}+1$ to T_{pred}. The trajectory of a pedestrian p at time t is denoted by (x^p_t, y^p_t).

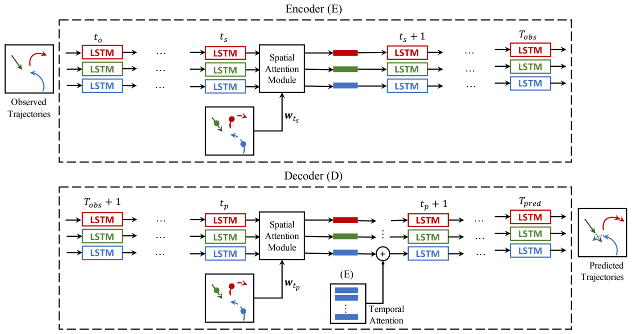

Figure 1: **SCAN** Architecture. $\mathbf{w}_{t_s}$ represents the vector of spatial weights computed for each neighbor with respect to each pedestrian using Equation 2 for $t_s \in [t_0, T_{obs}]$, similarly $\mathbf{w}_{t_p}$ for $t_p \in [T_{obs}+1, T_{pred}]$. In the decoding stage, temporal attention is interleaved with the spatial attention mechanism to enable the model to attend to observed spatial contexts.

Model Architecture. At a high level, **SCAN** is an LSTM-based encoder-decoder framework. The encoder encodes each pedestrian's observed trajectory into a fixed-length vector, and the decoder uses this fixed-length vector to predict each pedestrian's predicted trajectory. Our proposed model architecture is shown in Figure 1. We denote the number of pedestrians in the scene as N, observation time steps as $t_s \in [t_0, T_{obs}]$ and the prediction timesteps as $t_p \in [T_{obs}+1, T_{pred}]$. At a certain timestep t, we denote the trajectory of a pedestrian p, $p \in [1, N]$, by $\mathbf{x}^p_t$. Conventionally, the hidden state of an LSTM associated with modeling the trajectory of p is updated using its hidden state at previous time step $t-1$, h^p_{t-1} and $\mathbf{x}^p_t$. However, this update mechanism does not account for the spatial influences of other pedestrians on p's trajectory.

To take this spatial interaction into account, we incorporate a spatial attention mechanism, which will be explained in detail momentarily. Using this attention mechanism, the LSTM is able to incorporate spatial context experienced by p by computing a spatially weighted hidden state, $\tilde{h}^p_{t_s-1}$. The LSTM then uses this spatially-weighted hidden state to compute the next hidden state for pedestrian p using the conventional update mechanism:

$$h^p_t = \mathbf{LSTM}(\mathbf{x}^p_{t-1}, \tilde{h}^p_{t-1}) \tag{1}$$

This update mechanism is followed by both the LSTM encoder and LSTM decoder in our framework. By doing so, our framework is not only able to account for spatial influences that were experienced by p in the observed trajectory, but also *anticipate* the spatial influence of neighboring pedestrians on the trajectory of p in the future. Using spatial attention in the prediction time window is similar to a pedestrian altering their path if they anticipate collision with another pedestrian at a future time step.

While navigating through crowds, the spatial influence of neighbors causes pedestrians to temporarily digress from their intended trajectory to evade collision, respect personal space, etc. Therefore, while predicting intent for these pedestrians, some observed timesteps would be more reflective of their intent than others based on the spatial context associated with each observed timestep, t_s. In typical attention-based LSTM encoder-decoder frameworks, temporal attention is incorporated to enable the decoder to variably attend to the encoded hidden states. In our approach, we attempt to adopt temporal attention to enable our framework to attend to encoded *spatial contexts*.

At every $t_p \in [T_{obs+1}, T_{pred}]$, for a pedestrian p, the decoder attends to every spatially weighted hidden state, $\tilde{h}^p_{t_s}$, where $t_s \in [t_0, T_{obs}]$. To do so, the decoder compares the current spatially weighted hidden state for p, $\tilde{h}^p_{t_p}$ with all $\tilde{h}^p_{t_s}$, $t_s \in [t_0, T_{obs}]$ and assigns a score of similarity to each. The model then *attends* more to the spatially weighted hidden states that have been assigned a higher score than others. This mechanism of attending variably to different time steps from the observation window is called temporal attention or soft attention [12]. In our model, we use the dot product as the scoring mechanism for temporal attention. Therefore, the score assigned to a $\tilde{h}^p_{t_s}$ would be maximum when $\tilde{h}^p_{t_s} = \tilde{h}^p_{t_p}$, which would mean that the spatial context at t_p is similar to an observed spatial context at t_s. Therefore, in our framework, **SCAN**, the decoder possesses a novel *interleaved* spatially and temporally attentive architecture, that not only accounts for previous spatial interactions, but also accounts for the anticipated spatial interactions in the future, their influence on the pedestrian's intent thereof, and the variable influence of observed spatial contexts on the pedestrian's intent.

Spatial Attention Mechanism. As mentioned earlier, a pedestrian's intent is influenced by other pedestrians' trajectories and their expected intent. However, not all other pedestrians in a scene are of importance towards predicting the intent of a pedestrian. People navigating far off or towards different directions and not in line of sight of the pedestrian would have little to no effect on the pedestrian's intent. Therefore, to be able to understand and model spatial interactions experienced by a pedestrian, it is important to understand what the *neighborhood* of the pedestrian is, i.e., the neighbors that have a spatial influence on the pedestrian. As discussed earlier, prior approaches have either made significant assumptions about this *neighborhood* [1], assumed identical influence of all neighbors within this neighborhood irrespective of their orientations [1, 6] or only used features such as distance from the pedestrian [17]. Others, such as graph-based approaches [13, 10, 23] require learning a 'weight' for all pairs of pedestrians in the scene.

We introduce a concept called pedestrian domain, borrowed from an identical concept in ship navigation [15]. We define the domain of a pedestrian as the boundary of the area around a pedestrian, the intrusion of which by a neighbor causes the neighbor's trajectory to influence the intent of the pedestrian. Any other pedestrian that is beyond this boundary from the pedestrian of interest has no influence on the pedestrian's trajectory. Hereafter, we denote the domain by $\mathbf{S}$. The magnitude of influence of a neighbor, p_2, on that of a pedestrian of interest, p_1 at a certain instant t is largely dependent on three factors: distance between the p_1 and p_2, d^{21}_t, relative bearing of p_2 from p_1 θ^{21}_t, relative heading of p_2 to p_1, ϕ^{21}_t. The influence of p_2 on the intent of p_1 at $t+1$ is then determined by computing its spatial weight or *score* at t:

$$\mathbf{score}(p_1, p_2)_t = w^{21}_t = \mathbf{ReLU}(\mathbf{S}^{\theta^{21}_t, \phi^{21}_t} - d^{21}_t) \qquad (2)$$

where $\mathbf{S}^{\theta^{21}_t, \phi^{21}_t}$ denotes the value of the pedestrian domain $\mathbf{S}$ for θ^{21}_t,ϕ^{21}_t. Imagine if we discretized bearing and heading such that any encounter between agents can be put in a "bin". Let $\mathbf{S} \in R^{m,n}$, where the set $i \in \{1, \ldots, m\}$ (or $j \in \{1, \ldots, n\}$) maps to an interval in the relative bearing $[(i-1)\cdot\alpha, i\cdot\alpha)$ where $\alpha = \frac{360°}{m}$ (similar reasoning for heading). At the risk of overloading notation, we define $\mathbf{S}^{\theta^{21}_t, \phi^{21}_t}$ to be the element $s_{i,j}$ of $\mathbf{S}$ such that the encounter geometry is a kind of indicator function for the appropriate index on i, j. For example, if both bearing and heading are discretized at 30° ($m = n = 12$) increments and an encounter occurs at time $t = 0$ of $\theta^{21}_0 = 5°$ and $\phi^{21}_0 = 185°$ (potentially a collision course, by the way) it will lead to learning of the domain $\mathbf{S}$ in the increment of $\theta^{21}_0 \in [0, 30) \wedge \phi^{21}_0 \in [180, 210)$, or in this case $\mathbf{S}^{\theta^{21}_0, \phi^{21}_0}$ maps to the element $s_{1,7}$ of $\mathbf{S}$.

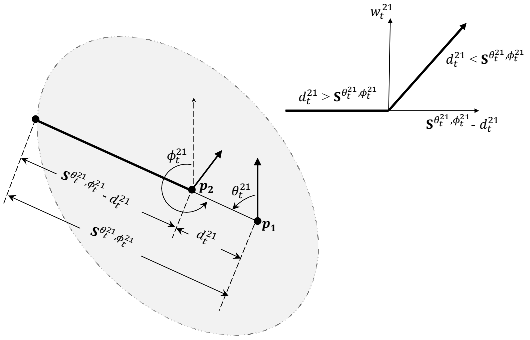
w^{21}_t
$d^{21}_t < \mathbf{S}^{\theta^{21}_t, \phi^{21}_t}$
$d^{21}_t > \mathbf{S}^{\theta^{21}_t, \phi^{21}_t}$
$\mathbf{S}^{\theta^{21}_t, \phi^{21}_t} - d^{21}_t$
ϕ^{21}_t
p_2
θ^{21}_t
p_1
$\mathbf{S}^{\theta^{21}_t, \phi^{21}_t} - d^{21}_t$
d^{21}_t
$\mathbf{S}^{\theta^{21}_t, \phi^{21}_t}$

Figure 2: For p_2 at distance d^{21}_t, relative bearing θ^{21}_t, and relative heading ϕ^{21}_t from p_1, the spatial weight or score of p_2 at t increases with increase in distance from $\mathbf{S}^{\theta^{21}_t, \phi^{21}_t}$. The shaded region corresponds to the domain corresponding to any neighbors with relative heading ϕ^{21}_t from p_1. The elliptical shape of the shaded region is notational and used to indicate that ideally, the neighbors in a larger area in line of sight of the pedestrian would influence its trajectory.

This weighting mechanism directly translates into a pedestrian closer to the self, and hence farther from $\mathbf{S}$ having a larger weight, and hence a larger influence on the self. Similarly, a pedestrian closer to the boundary, $\mathbf{S}$, and hence farther from the self would have a smaller influence on the self. The activation function $\mathbf{ReLU}$ ensures that if a pedestrian p_2 is at a distance $d^{21}_t \geq \mathbf{S}$ from p_1 at t, its influence on the intent of p_1 at $t+1$ is 0. This allows the model to determine the domain as an area beyond which another pedestrian in the scene would not affect the self and vice-versa.

However, using this spatial scoring mechanism, a neighbor at a certain distance and orientation with respect to the pedestrian of interest would always have the same spatial influence on the pedestrian's trajectory, irrespective of crowd densities. However, a certain neighbor p_2 at a (large) distance from p_1, with a small positive value for w^{21}_t might not affect p_1 much in a densely crowded setting but might influence p_1 more in a sparsely crowded environment. Simply put, while navigating in environments that are not too crowded humans often tend to change their trajectories as a response to someone that is relatively far away; however, in crowded settings, the same neighbor at the same orientation and distance does not pose an immediate risk of collision and hence does not influence the pedestrian's trajectory as much. To account for this varying spatial influence as a result of varying crowd densities, we normalize the scores for all neighbors for each pedestrian in the frame,

$$\mathbf{score}(p_1, p_2)_t = \frac{\mathbf{exp}(\mathbf{score}(p_1, p_2)_t)}{\sum_{n \in N'} \mathbf{exp}(\mathbf{score}(p_1, p_n)_t))} \quad (3)$$

where $n \in N'$ are all pedestrians in the frame apart from the pedestrian of interest, p_1. Once the spatial scores have been computed for every pair of pedestrians, we compute a *spatial context vector* for each pedestrian that represents the *spatial context* experienced from the pedestrian's perspective at t. For example, the spatial context vector for p_1 at t is computed as,

$$\tilde{\mathbf{C}}_t^{p_1} = \sum_{n \in N'} \mathbf{score}(p_1, p_n)_t h_t^{p_n} \quad (4)$$

This spatial context vector contains meaningful information about the spatial orientation of other pedestrians in the frame from p_1's perspective at t and hence the amount of knowledge (hidden states) shared with p_1 about its neighbors depending on their orientations. This spatial context is then concatenated with the hidden state of the pedestrian at t before it is fed to the LSTM. For p_1,

$$\tilde{h}_t^{p_1} = \mathbf{concat}(h_t^{p_1}, \tilde{C}_t^{p_1}) \quad (5)$$

This gives the model relevant information of both the pedestrian's own hidden state as well as spatial context from the pedestrian's perspective. Every pedestrian in the frame has a unique spatial context, which is the spatial orientation and influence of neighbors experienced by the pedestrian at t from its own perspective instead of a global perspective.

Multiple Socially Plausible Paths. Given an observed trajectory, there can be more than one *socially plausible* trajectory that a pedestrian can take in the future. A socially plausible trajectory would account for spatial influence of neighboring pedestrians' trajectories and respect social norms. For safe navigation, it is imperative to be able to account for the fuzzy nature of human motion and be able to generate multiple socially plausible future trajectories instead of narrowing down on one average expected behavior. To do so, we leverage the generative modeling abilities of GANs (Generative Adversarial Networks) [5]. Briefly, the training process of GANs is formulated as a two player minmax game between a generator and a discriminator. The generator generates candidate predictions and the discriminator evaluates them and scores them as real/fake. In our case, the goal of the generator is to be able to generate predictions that are consistent with the observed trajectory and are also consistent with the observed and intended spatial contexts, hence *socially plausible*. The discriminator must be able to discern which trajectories are real, and which are generated. GANs have also been previously adopted for pedestrian intent prediction [6, 17, 10, 13].

Generator. The generator of our model is basically the encoder-decoder framework that we described above. The goal of generator is to learn how to generate realistic trajectories that are consistent with the observed trajectories and the observed spatial contexts that are incorporated in the encoded representation of each pedestrian by virtue of the interleaved spatial attention mechanism. We achieve this by initializing the hidden state of the decoder for a pedestrian, p, as

$$h_{T_{obs+1}}^p = [h_{T_{obs}}^p, z] \quad (6)$$

where z is a noise vector, sampled from $\mathcal{N}(0, 1)$ and $h_{T_{obs}}^p$ is the encoded representation for pedestrian, p, or the final hidden state of the LSTM encoder pertaining to p. A difference of our approach in comparison to prior multimodal intent forecasting approaches is that in addition to the pedestrian's encoding, they also condition the generation of output trajectories on social context vectors [6] that summarise the spatial context of the pedestrian, p. In our framework, our interleaved spatial attention mechanism already accounts for spatial context in the encoded representation.

Discriminator. The discriminator contains a separate encoder. This encoder takes as input the N 'ground truth' trajectories over $[t_0, T_{obs}]$ and the N generated trajectories over $[t_0, T_{obs}]$ and classifies them as 'real' or 'fake'. The encoder in the discriminator also uses the spatial attention mechanism at each time step, therefore ideally the goal of the discriminator is to classify the trajectories as real/fake while taking into account social interaction rules. This would imply that trajectories that do not seem to comply with social navigation norms and hence are not socially plausible would be classified as fake.

Experimental Evaluation

Datasets. We evaluate **SCAN** on two publicly available pedestrian-trajectory datasets: ETH[14] and UCY[11]. The datasets contain birds eye-view frames sampled at 2.5 fps and 2D locations of pedestrians walking in crowded scenes. The ETH dataset contains two sub-datasets (annotated ETH and HOTEL) from two scenes, each with 750 pedestrians. The UCY dataset contains two scenes with 786 pedestrians, split into three sub-datasets (ZARA1, ZARA2, UNIV). These datasets contain annotated trajectories of pedestrians interacting in several social situations and include challenging behavior such as collision avoidance, movement in groups, yielding right of way, couples walking together, groups crossing groups, etc.[14].

Baselines. We compare our model against several baselines: (a) *Linear:* A linear regressor with parameters estimated by minimizing least square error; (b) *LSTM:* An LSTM that models only individual pedestrian trajectory without accounting for any spatial interactions; (c) *Social LSTM* [1]: Uses a pooling mechanism to model spatial influence of neighbors within an assumed spatial grid and models each pedestrian's trajectory using an LSTM; (d) *S–GAN* [6]: Models spatial interactions using a grid-based pooling mechanism, and models each pedestrian's trajectory using a GAN-based framework similar to ours; (e) *S–GAN-P* [6]: Similar framework to *S-GAN*, but incorporates their proposed pooling mechanism to model spatial interactions; (f) *SoPhie GAN* [17]: Models agent trajectories using a LSTM-GAN framework with additional modules to incorporate social attention and physical scene context; (g) *Social Attention* [23]: Models pedestrian trajectory prediction as a spatio-temporal graph, also incorporates features like relative orientation and distances in the spatial edges of the graph; (h) *Social Ways* [3]: GAN-based framework that also incorporates relative orientation features as a prior over the attention pooling mechanism; (i) *Social-BiGAT* [10]: Graph-based GAN that uses a graph attention network (GAT) to model spatial interactions and an adversarially trained recurrent encoder-decoder architecture to model trajectories; (j) *Trajectron* [8]: An LSTM-CVAE encoder-decoder which is explicitly constructed to match the spatio-temporal structure of the scene; and (k) *Trajectron++* [18]: Similar to [8], but uses directed edges in the spatio-temporal graph modeling the scene.

Implementation. We follow a leave-one-out evaluation methodology to train and test **SCAN** on each of the five datasets, training on four datasets and testing on the fifth. As with all prior approaches, we observe the trajectory for 8 time steps (2.8 seconds) and predict intent over future 12 time steps (3.2 seconds). Model parameters are iteratively trained using Adam[4] optimizer with a batch size of 32 and learning rate of 0.0005. The model is implemented in PyTorch and trained using a single GPU. In both the encoder and the decoder, the positional information pertaining to each pedestrian in the frame is first embedded into 16 dimensional vectors using a linear layer. The hidden states for both the encoder and the decoder LSTMs are 32 dimensional vectors. In the decoder, a linear layer is used to convert the LSTM output to the (x,y) coordinates predicted for the pedestrians. Relative bearing and relative heading are discretized at 30^o. All the parameters are chosen using grid search based on performance on ZARA1 validation dataset.

Quantitative Comparison. We compare two versions of our model - **SCAN**, the proposed encoder-decoder framework with interleaved spatial and temporal attention, and **vanillaSCAN**, the proposed encoder-decoder architecture sans the temporal attention in the decoder - with the deterministic baselines (*Linear*, *Social LSTM* [1], *Social Attention* [23], deterministic *Trajectron++* [18]) in Table 1. We also compare GAN-based generative framework, **generativeSCAN** with the generative baselines (*S-GAN* [6], *S-GAN-P* [6], *SoPhie GAN* [17], *Social Ways* [3], *Trajectron* [8], generative *Trajectron++* [18]) in Table 2. We report our results using two metrics: *Average Displacement Error (ADE)*, which is the average L2 distance between ground truth trajectories and predicted trajectories over all predicted time steps, and *Final Displacement Error (FDE)*, which is the average displacement error between final predicted destination of all pedestrians at the end of the time window and the true final destination at T_{pred}. In Table 1, while we mention results for *Social Attention* [23], as are reported in their paper, it is not directly comparable to our method because, as mentioned in their paper, they process their dataset differently in comparison to the other baselines (and our method). While *Trajectron++* [18] has an average lower ADE, **SCAN** has a lower final displacement error, implying that its ability to anticipate spatial interactions into the future enable it to predict a more accurate final destination. Both **vanillaSCAN** and **SCAN** are largely able to outperform the other deterministic baselines on the five datasets. Interleaving temporal attention with spatial attention in **SCAN** also enables the model to capture long-term or high-level intent more accurately, which is reflected in its lower FDE values compared to **vanillaSCAN**. In Table 2, we compare **generativeSCAN** with other baselines that account for multimodal pedestrian behavior. *Sophie GAN* [17] takes into account physical scene information while making trajectory predictions. Despite our model being agnostic to such information, it is able to achieve lower ADE and FDE than both *Sophie GAN* and *S-GAN* [6]. Our model is also able to outperform *Social-Ways* on both the Zara datasets. *Social-BiGAT* [10], which uses a graph attention network [22] to model spatial influences, is able to slightly outperform our model on an average. As we explain later, our spatial attention mechanism in fact outperforms a graph-based attention mechanism for modeling spatial influences, hence *Social-BiGAT*'s performance may be attributed to its ability to also include scene information while making its predictions. *Trajectron++* is largely able to outperform **generativeSCAN** across all five datasets. While it simply uses a directed spatiotemporal graph to model agent interactions, *Trajectron++* [8] incorporates a conditional variational autoencoder (CVAE) [20] to sample multimodal trajectories conditioned on future behavior, as opposed to **generativeSCAN** and other baselines that are GAN-based.

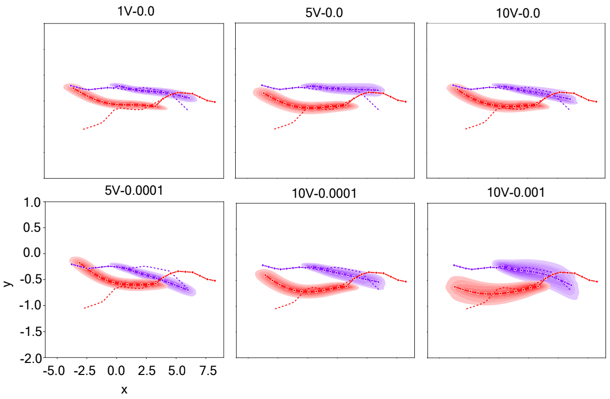

Figure 3: Effect of increasing λ and k on the diversity of generated trajectories on a scenario from Zara1 test dataset. For each sample shown, we generate 300 trajectories and visualize their density and mean. Each plot is titled kV-λ. The ground truth trajectory is denoted by dotted line.

Variety loss and diversity loss. While accounting for multimodal pedestrian behavior, it is important to ensure that the generated predictions are diverse and not simply multiple 'close to average' predictions. We train **generativeSCAN** using adversarial loss and L2 loss. However, while the trained model is able to generate multiple socially plausible trajectories, these are largely very similar predictions. To encourage diversity in generated trajectories, we adopt *variety loss*, as proposed in [6]. For each scene, the generator generates k possible output predictions by randomly sampling z from $\mathcal{N}(0,1)$ and penalizing the 'best prediction', i.e., the one with the least ADE. However, training the model with a large k value is computationally expensive because it involves k forward passes per batch in the training dataset. Further, we observed that increasing k does not improve the diversity of the generated trajectories substantially. There-

Dataset	**ADE / FDE (m)**						
	Linear	*LSTM*	*Social LSTM* [1]	*Social Attention* [23]	*Trajectron*++ [18]	**vanillaSCAN**	**SCAN**
ETH	1.33 / 2.94	1.09 / 2.41	1.09 / 2.35	0.39 / 3.74	0.71 / 1.66	0.75 / 1.44	0.57 / 0.78
Hotel	0.39 / 0.72	0.86 / 1.91	0.79 / 1.76	0.29 / 2.64	0.22 / 0.46	0.45 / 0.91	0.43 / 0.85
Univ	0.82 / 1.59	0.61 / 1.31	0.67 / 1.40	0.33 / 3.92	0.41 / 1.07	0.62 / 1.31	0.61 / 1.28
Zara1	0.62 / 1.21	0.41 / 0.88	0.47 / 1.00	0.20 / 0.52	0.30 / 0.77	0.39 / 0.89	0.39 / 0.84
Zara2	0.77 / 1.48	0.52 / 1.11	0.56 / 1.17	0.30 / 2.13	0.23 / 0.59	0.34 / 0.75	0.34 / 0.74
Average	0.79 / 1.59	0.70 / 1.52	0.72 / 1.54	0.30 / 2.59	0.37 / 0.91	0.51 / 1.06	0.46 / 0.89

Table 1: Comparison of our models, **vanillaSCAN** and **SCAN** against other deterministic baselines.

Dataset	**ADE / FDE (m)**, Best of 20						
	S-GAN [6]	*Sophie GAN* [17]	*Social Ways* [3]	*Social BiGAT* [10]	*Trajectron* [8]	*Trajectron*++ [18]	**generativeSCAN**
ETH	0.81 / 1.52	0.70 / 1.43	0.39 / 0.64	0.69 / 1.29	0.59 / 1.14	0.39 / 0.83	0.84 / 1.58
Hotel	0.72 / 1.61	0.76 / 1.67	0.39 / 0.66	0.49 / 1.01	0.35 / 0.66	0.12 / 0.19	0.44 / 0.90
Univ	0.60 / 1.26	0.54 / 1.24	0.55 / 1.31	0.55 / 1.32	0.54 / 1.13	0.20 / 0.44	0.63 / 1.33
Zara1	0.34 / 0.69	0.30 / 0.63	0.44 / 0.64	0.30 / 0.62	0.43 / 0.83	0.15 / 0.32	0.31 / 0.85
Zara2	0.42 / 0.84	0.38 / 0.78	0.51 / 0.92	0.36 / 0.75	0.43 / 0.85	0.11 / 0.25	0.37 / 0.76
Average	0.58 / 1.18	0.54 / 1.15	0.46 / 0.83	0.48 / 1.00	0.56 / 1.14	0.19 / 0.41	0.51 / 1.08

Table 2: Comparison of our generative model, **generativeSCAN** with other generative baselines. The results reported for all generative models are 'best of 20', which means the ADE for the trajectory with least ADE out of 20 generated trajectories per sample is reported. The FDE value is reported for the trajectory with the best ADE.

fore, we incorporate another loss function, *diversity loss*, which essentially penalizes the generator for generating similar trajectories. For N pedestrians in the frame,

$$\mathcal{L}_{diversity} = \frac{1}{N} \sum_{i,j \in k} \exp(-d_{ij}) \tag{7}$$

where d_{ij} is the average euclidean distance between trajectories i and j. The generator is then trained using the sum of adversarial loss, variety loss and the diversity loss weighted by parameter λ. In Figure 3, we analyze the effect of increasing k and increasing λ on the diversity in generated trajectories in a crossing scenario. More diverse trajectories can be generated by increasing λ value for a smaller k value.

Dataset	**ADE / FDE (m)**		**Inference Time (s)**	
	Graph based SCAN	**SCAN**	**Graph based SCAN**	**SCAN**
ETH	1.14 / 2.28	0.57 / 0.78	0.0683	0.0764
Hotel	0.53 / 1.19	0.43 / 0.85	0.0712	0.0742
Univ	0.64 / 1.29	0.61 / 1.28	0.0737	0.0773
Zara1	0.43 / 0.92	0.39 / 0.84	0.0736	0.0786
Zara2	0.41 / 0.88	0.34 / 0.74	0.0712	0.0776
Average	0.62 / 1.30	0.46 / 0.89	0.0716	0.0768

Table 3: Quantitative comparison of **SCAN** with **Graph based SCAN**, an ablation that models spatial influence using graph attention networks (GATs). The inference time reported is averaged across ten evaluation runs.

Modeling Spatial Interactions as a Graph. Our spatial attention mechanism has certain similarities to graph attention networks [22], since we initially consider all nodes (pedestrians) to be connected, or influence each other, and then proceed to learn the 'domain' which enables us to learn these influences or edges during training. The key difference is that given N pedestrians, hence N nodes in the graph, graph attention networks learn a $\mathbf{W}^{N \times N}$ weight parameter. **SCAN**, on the other hand, is required to learn $\mathbf{S}^{m \times n}$ as explained earlier, where m and n depend on the chosen relative bearing and heading discretization values. In contrast to graph attention based trajectory forecasting methods (*Social Ways* [3], *Social BiGAT* [10], *Trajectron* [8], *Trajectron*++ [18]), **SCAN**'s learnable parameters do not increase with an increase in pedestrians in the frame. To validate the performance benefits of our proposed spatial attention mechanism, we also evaluate an ablation that uses a graph attention network (GAT) in place of our spatial attention mechanism in **SCAN** with the rest of the framework being the same. The results are reported in Table 3. Computationally, both mechanisms are nearly the same. The slight overhead for our method comes from having to compute distance, bearing, heading for each prediction time step in order to compute spatial attention weights. Since the maximum number of nodes (pedestrians) across the datasets is 57, the number of trainable parameters in the **Graph based SCAN** mechanism is $57 \times 57 = 3249$ parameters. For our proposed spatial attention mechanism, the trainable parameters are 144, which is simply the size of the learnable domain, which depends on our chosen values of relative bearing and relative heading discretization (30^o). Our spatial attention mechanism is therefore not only parameter efficient, but also capable of achieving lower error in comparison to a graph attention network. Further, the learned domain parameter informs interpretability of the model's predictions since it provides information about the neighborhood that influences the pedestrian and its intent.

Collision Analysis. To demonstrate the capability of our spatial attention mechanism to predict safe, *socially acceptable* trajectories, we evaluate the ability of trajectories predicted by our model to avoid "collisions". To do so, we calculate the average percentage of pedestrians *near-collisions* across the five evaluation datasets. As in [17], for a given

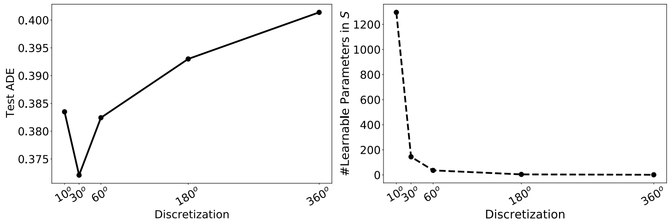

Figure 4: (a) Effect of increasing relative bearing, relative heading discretization on Test ADE, (b) Effect of increasing relative bearing, heading discretization on learnable parameters in the pedestrian domain, **S**.

	Ground Truth	*Linear*	*Social-GAN*	*SoPhie GAN*	**SCAN**
ETH	0.000	3.137	2.509	1.757	**0.793**
Hotel	0.092	1.568	1.752	1.936	**1.126**
Univ	0.124	1.242	0.559	0.621	**0.481**
Zara1	0.000	3.776	1.749	1.027	**0.852**
Zara2	0.732	3.631	2.020	**1.464**	3.109
Average	0.189	2.670	1.717	1.361	**1.272**

Table 4: Average % of colliding pedestrians per frame for each of the five evaluation datasets. A collision is detected if the euclidean distance between two pedestrians is less than 0.10m.

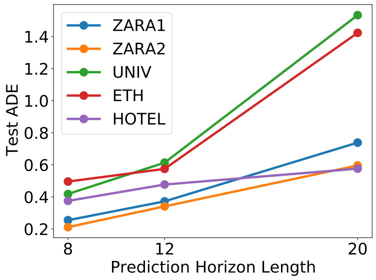

Figure 5: Change in ADE with prediction horizon length.

scene, if the euclidean distance between any two pedestrians drops below 0.10 m, we say that a *near-collision* has occurred. In Table 4, we compare the average percentage of colliding pedestrians for predictions generated by **SCAN** against several other baselines. Our model is able to predict much more socially acceptable trajectories in comparison to other baselines. Further, the average percentage of colliding pedestrians per frame for each dataset as obtained by our model's predictions is much closer to the ground truth as compared to the other baselines. *Social-GAN* [6] uses a pooling mechanism to incorporate spatial influences of neighboring pedestrians, and *Sophie-GAN* uses a sorting mechanism to incorporate distances while taking spatial influences into account. Further, *Sophie-GAN* [17] also incorporates scene context towards making more informed predictions. From Table 4, we can conclude that our proposed spatial attention mechanism is not only able to generate more socially acceptable trajectories, but is also able to capture the social behavior in the ground truth trajectories.

Effect of Different Bearing, Heading. Discretizations. In order to learn the pedestrian domain **S**, we discretize the space of relative bearing and relative heading values such that any encounter between agents can be put in a "bin". In our evaluation, we choose to discretize relative bearing, θ and relative heading, ϕ values into bins of $\Delta\theta = \Delta\phi = 30^o$. Figure 4a. plots the variation in test ADE on ZARA1 dataset with increasing $\Delta\theta = \Delta\phi$. A more fine-grained discretization than 30^o has a higher test ADE. Similarly, more coarse-grained discretizations lead to higher test ADE values. A discretization of 360^o would correspond to a uniform value of **S** irrespective of relative bearing and relative heading values of a neighbor. Figure 4b. also plots the number of learnable parameters in **S** as a function of discretization values. As is true of deep learning based architectures in general, a highly parameterized domain and lower parameterized **S** domains do not generalize well to the test dataset.

Effect of Varying Prediction Horizon Lengths. Figure 5 plots the average displacement error (ADE) for **SCAN** across all five datasets against various prediction horizon lengths for the same observed time window. As expected, as the length of the prediction time window increases, the average displacement error across all the five evaluation datasets increases. For ZARA1, ZARA2 and HOTEL, the increase in ADE as the prediction time window is increased from 12 to 20 timesteps is ≈ 0.2 m. Therefore, using the same observed time window of 8 timesteps, **SCAN** is able to predict longer trajectories fairly accurately.

Conclusion and Future Work

In this work, we propose **SCAN**, a novel trajectory prediction framework for predicting pedestrian intent. A key contribution of this work is the novel spatial attention mechanism, that is able to model spatial influence of neighboring pedestrians in a manner that is parameter efficient, relies on less assumptions and results in more accurate predictions. We also propose **generativeSCAN** that accounts for the multimodal nature of human motion and is able to predict multiple socially plausible trajectories per pedestrian in the scene. Despite being agnostic to scene context and relevant physical scene information, our model is able to match or even outperform existing baselines that use such information. This work can also be extended to predicting trajectories for heterogeneous agents with different trajectory dynamics. The spatial attention mechanism introduced in this work can be used to infer more domain-specific knowledge, such as the influence of different kinds of agents on each other (for example, the effect of a skateboarder on a cyclist's trajectory) and use these to either explain model predictions or inform model predictions.

At a more fundamental level, **SCAN** is a general framework that can be applied to any sequence-to-sequence modeling application where cross-LSTM knowledge can help improve performance. This can include human action recognition [25, 21], modeling human-object interactions [9, 16], video classification [24, 19]. An important advantage of **SCAN** is its ability to infer domain knowledge from the observation dataset and hence yield improved predictions without making significant assumptions about the application domain or the dataset.

Ethical Impact

Deep learning based decision making has ethical implications, especially in safety-critical applications, where failures could possibly lead to fatalities. This especially amplifies in shared settings like our application, where an agent's decisions influence other agents' decisions and so on. Certain features of our model contribute towards ethical decision-making. To begin with, our model is motivated by the need for autonomous agents to practice safety while navigating in human-centric environments. Our proposed framework takes into account the spatial influence of neighbors and implicit social navigation norms such as collision avoiding behavior that pedestrians follow when navigating in crowded environments towards predicting their future behavior. Further, our proposed framework acknowledges the multimodality of human motion and is capable of predicting multiple socially plausible trajectories per pedestrian in the scene. An autonomous agent that may use this framework to inform its navigation decisions would essentially take in to account all these multiple trajectories to negotiate a safe, collision-free path for itself. Often deep learning based models are reflective of inherent biases on the datasets that they are trained on. For instance, in our application, a model trained only on the UNIV dataset may not generalize well to a lower crowd density. However, as is the case with other baselines in our application domain, this is taken care of by using a leave-one-out approach, by training the model on four of five datasets and testing on the fifth. These datasets vary in crowd densities and contain a variety of trajectories of pedestrians interacting in several social situations, hence the training dataset is diverse. Moreover, a predicted trajectory can be mapped to the neighborhood (the learned domain) and hence, the neighbors that influenced the model's decision, hence providing some degree of interpretability to our framework.

However, like all other deep learning models, our proposed framework relies on implicit assumptions that may have ethical consequences. For instance, our model relies on the assumption that the training dataset is reflective of ideal pedestrian behavior in shared environments or general pedestrian dynamics. Further, when deployed in a real-world setting to aid the navigation of an autonomous agent in a human centric environment, our framework's ability to predict intent accurately is largely dependent on the accuracy of input, i.e, the observed trajectory. Our model, by itself, does not account for the presence of adversaries that may provide deceptive input and cause our model to mispredict and cause undesired behavior. Further, in a real world setting, our model is expected to inform safety-critical decision-making of an autonomous agent in human-centric environments. Because deep learning models are black-box in nature, it is difficult to be able to completely ensure safety before deployment. It is therefore also important to incorporate a certain measure of confidence in the model's decisions, based on which its predictions can be followed or overridden.

Acknowledgements

This material is based upon work supported in part by the National Science Foundation under Grant No. CNS: 1650512, conducted in the NSF IUCRC Center of Visual and Decision Informatics, through the guidance of Leidos Corporation.

References

[1] Alahi, A.; Goel, K.; Ramanathan, V.; Robicquet, A.; Fei-Fei, L.; and Savarese, S. 2016. Social LSTM: Human Trajectory Prediction in Crowded Spaces. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*

[2] Alahi, A.; Ramanathan, V.; and Fei-Fei, L. 2014. Socially-Aware Large-Scale Crowd Forecasting. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, 2211–2218.

[3] Amirian, J.; Hayet, J.-B.; and Pettré, J. 2019. Social ways: Learning multi-modal distributions of pedestrian trajectories with GANs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 0–0.

[4] Antonini, G.; Bierlaire, M.; and Weber, M. 2006. Discrete Choice Models for Pedestrian Walking Behavior. *Transportation Research Part B: Methodological* 40: 667–687. doi: 10.1016/j.trb.2005.09.006.

[5] Goodfellow, I. J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative Adversarial Nets. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, 2672–2680. Cambridge, MA, USA: MIT Press.

[6] Gupta, A.; Johnson, J.; Fei-Fei, L.; Savarese, S.; and Alahi, A. 2018. Social GAN: Socially Acceptable Trajectories with Generative Adversarial Networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, CONF.

[7] Helbing, D.; and Molnár, P. 1995. Social force model for pedestrian dynamics. *Physical Review E* 51(5): 4282–4286. ISSN 1095-3787. doi:10.1103/physreve.51.4282. URL http://dx.doi.org/10.1103/PhysRevE.51.4282.

[8] Ivanovic, B.; and Pavone, M. 2018. The Trajectron: Probabilistic Multi-Agent Trajectory Modeling With Dynamic Spatiotemporal Graphs.

[9] Jain, A.; Zamir, A. R.; Savarese, S.; and Saxena, A. 2015. Structural-RNN: Deep Learning on Spatio-Temporal Graphs.

[10] Kosaraju, V.; Sadeghian, A.; Martín-Martín, R.; Reid, I.; Rezatofighi, S. H.; and Savarese, S. 2019. Social-BiGAT: Multimodal Trajectory Forecasting using Bicycle-GAN and Graph Attention Networks.

[11] Lerner, A.; Chrysanthou, Y.; and Lischinski, D. 2007. Crowds by Example. *Comput. Graph. Forum* 26: 655–664.

[12] Luong, M.-T.; Pham, H.; and Manning, C. D. 2015. Effective Approaches to Attention-based Neural Machine Translation.

[13] Mohamed, A.; Qian, K.; Elhoseiny, M.; and Claudel, C. 2020. Social-STGCNN: A Social Spatio-Temporal Graph Convolutional Neural Network for Human Trajectory Prediction.

[14] Pellegrini, S.; Ess, A.; Schindler, K.; and van Gool, L. 2009. You'll never walk alone: Modeling social behavior for multi-target tracking. In *2009 IEEE 12th International Conference on Computer Vision*, 261–268. doi:10.1109/ICCV.2009.5459260.

[15] Pietrzykowski, Z.; and Uriasz, J. 2009. The Ship Domain – A Criterion of Navigational Safety Assessment in an Open Sea Area. *Journal of Navigation* 62(1): 93–108. doi:10.1017/S0373463308005018.

[16] Qi, S.; Wang, W.; Jia, B.; Shen, J.; and Zhu, S.-C. 2018. Learning Human-Object Interactions by Graph Parsing Neural Networks.

[17] Sadeghian, A.; Kosaraju, V.; Sadeghian, A.; Hirose, N.; Rezatofighi, H.; and Savarese, S. 2019. SoPhie: An Attentive GAN for Predicting Paths Compliant to Social and Physical Constraints. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[18] Salzmann, T.; Ivanovic, B.; Chakravarty, P.; and Pavone, M. 2020. Trajectron++: Dynamically-Feasible Trajectory Forecasting With Heterogeneous Data.

[19] Shan, M.; and Atanasov, N. 2017. A spatiotemporal model with visual attention for video classification.

[20] Sohn, K.; Lee, H.; and Yan, X. 2015. Learning Structured Output Representation using Deep Conditional Generative Models. In Cortes, C.; Lawrence, N. D.; Lee, D. D.; Sugiyama, M.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 28*, 3483–3491. Curran Associates, Inc. URL http://papers.nips.cc/paper/5775-learning-structured-output-representation-using-deep-conditional-generative-models.pdf.

[21] Song, S.; Lan, C.; Xing, J.; Zeng, W.; and Liu, J. 2017. An End-to-End Spatio-Temporal Attention Model for Human Action Recognition from Skeleton Data. In *AAAI Conference on Artificial Intelligence*, 4263–4270.

[22] Veličković, P.; Cucurull, G.; Casanova, A.; Romero, A.; Liò, P.; and Bengio, Y. 2017. Graph Attention Networks.

[23] Vemula, A.; Muelling, K.; and Oh, J. 2017. Social Attention: Modeling Attention in Human Crowds.

[24] Wu, Z.; Wang, X.; Jiang, Y.-G.; Ye, H.; and Xue, X. 2015. Modeling Spatial-Temporal Clues in a Hybrid Deep Learning Framework for Video Classification.

[25] Yang, Z.; Li, Y.; Yang, J.; and Luo, J. 2019. Action Recognition With Spatio–Temporal Visual Attention on Skeleton Image Sequences. *IEEE Transactions on Circuits and Systems for Video Technology* 29(8): 2405–2415.

[26] Yi, S.; Li, H.; and Wang, X. 2015. Understanding pedestrian behaviors from stationary crowd groups. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3488–3496.