SoPhie: An Attentive GAN for Predicting Paths Compliant to Social and Physical Constraints

Amir Sadeghian[1,2*] Vineet Kosaraju[1*] Ali Sadeghian[3] Noriaki Hirose[1]
S. Hamid Rezatofighi[1,4] Silvio Savarese[1]

[1]Stanford University [2] Aibee Inc [3] University of Florida [4] University of Adelaide
amirabs@aibee.com

Abstract

This paper addresses the problem of path prediction for multiple interacting agents in a scene, which is a crucial step for many autonomous platforms such as self-driving cars and social robots. We present SoPhie; an interpretable framework based on Generative Adversarial Network (GAN), which leverages two sources of information, the path history of all the agents in a scene, and the scene context information, using images of the scene. To predict a future path for an agent, both physical and social information must be leveraged. Previous work has not been successful to jointly model physical and social interactions. Our approach blends a social attention mechanism with physical attention that helps the model to learn where to look in a large scene and extract the most salient parts of the image relevant to the path. Whereas, the social attention component aggregates information across the different agent interactions and extracts the most important trajectory information from the surrounding neighbors. SoPhie also takes advantage of GAN to generates more realistic samples and to capture the uncertain nature of the future paths by modeling its distribution. All these mechanisms enable our approach to predict socially and physically plausible paths for the agents and to achieve state-of-the-art performance on several different trajectory forecasting benchmarks.

1. Introduction

When people navigate through a park or crowded mall, they follow common sense rules in view of social decorum to adjust their paths. At the same time, they are able to adapt to the physical space and obstacles in their way. Interacting with the physical terrain as well as humans around them is by no means an easy task; because it requires:

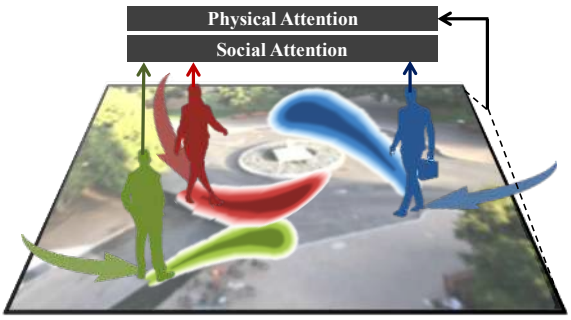

Figure 1. **SoPhie** predicts trajectories that are socially and physically plausible. To perform this, our approach incorporates the influence of all agents in the scene as well as the scene context.

- **Obeying physical constraints of the environment.** In order to be able to walk on a feasible terrain and avoid obstacles or similar physical constraints, we have to process the local and global spatial information of our surroundings and pay attention to important elements around us. For example, when reaching a curved path, we focus more on the curve rather than other constraints in the environment, we call this *physical attention*[26].

- **Anticipating movements and social behavior of other people.** To avoid collisions with other people, disturbing their personal space, or interrupting some social interactions (e.g. a handshake), we must have a good understanding of others' movements and the social norms of an environment and adjust our path accordingly. We should take into account that some agents have more influence in our decision. For example, when walking in a corridor, we pay more attention to people in front of us rather than the ones behind us, we call this *social attention*. Modeling these social interactions is a non-trivial task.

- **Finding more than a single feasible path.** To get to our destination, there often exists more than a single choice for our path, which is the fuzzy nature of human motion.

*indicates equal contribution

Indeed, there is a range for our traversable paths toward our destinations [26, 23, 13, 8, 1].

In this paper, we aim to tackle the problem of future path prediction for a set of agents. The existing approaches follow different strategies to solve this problem. Some methods solely rely on the scene context to predict a feasible path for each agent. For example, the approach in [3] learns a dynamic pattern for all agents from patch-specific descriptors using previously created navigation maps that encode scene-specific observed motion patterns. In [14], the approach learns the scene context from top-view images in order to predict future paths for each agent. [26] applies an attention mechanism to input images in order to highlight the important regions for each agent's future path. However, all above approaches ignore the influence of the other agents' state on predicting the future path for a targeted agent.

Parallel to path prediction using scene context information, several approaches have recently proposed to model interactions between all agents in the scene in order to predict the future trajectory for each targeted agent [5, 6]. Although these methods have shown promising progress in addressing this challenging problem, they still ignore the scene contexts as crucial information. In addition, these methods fall short as instead of treating pedestrian's future movements as a distribution of locations, they only predict a single path, which generally ends up optimizing "average behavior" rather than learning difficult constraints.. To address the second problem, [1, 14, 30] have introduced models that are able to generate multiple feasible paths. However, most of these models only incorporate the influence of few adjacent agents in a very limited search space. Recently, [8] proposed a GAN model that takes into account the influence of all agents in the scene.

In this work, we propose SoPhie an attentive GAN-based approach that can take into account the information from both scene context and social interactions of the agents in order to predict future paths for each agent. Influenced by the recent success of attention networks [29] and also GANs [7] in different real-world problems, our proposed framework simultaneously uses both mechanisms to tackle the challenging problem of trajectory prediction. We use a visual attention model to process the static scene context alongside a novel attentive model that observes the dynamic trajectory of other agents. Then, an LSTM based GAN module is applied to learn a reliable generative model representing a distribution over a sequence of plausible and realistic paths for each agent in future.

To the best of our knowledge, no other work has previously tackled all the above problems together. SoPhie generates multiple socially-sensitive and physically-plausible trajectories and achieves state-of-the-art results on multiple trajectory forecasting benchmarks. To summarize the main contribution of the paper are as follows:

- Our model uses scene context information jointly with social interactions between the agents in order to predict future paths for each agent.
- We propose a more reliable feature extraction strategy to encode the interactions among the agents.
- We introduce two attention mechanisms in conjunction with an LSTM based GAN to generate more accurate and interpretable socially and physically feasible paths.
- State-of-the-art results on multiple trajectory forecasting benchmarks.

2. Related Work

In recent years, there have been many advances in the task of trajectory prediction. Many of the previous studies on trajectory prediction either focus on the effect of physical environment on the agents paths (agent-space interactions) and learn scene-specific features to predict future paths [26], or, focus on the effect of social interactions (dynamic agent-agent phenomena) and model the behavior of agents influenced by other agents' actions [1, 8]. Few works have been trying to combine both trajectory and scene cues [14].

Agent-Space Models. This models mainly take advantage of the scene information, e.g., cars tend to drive between lanes or humans tend to avoid obstacles like benches. Morris et al. [20] cluster the spatial-temporal patterns and use hidden Markov models to model each group. Kitani et al. [13] use hidden variable Markov decision processes to model human-space interactions and infer walkable paths for a pedestrian. Recently, Kim et al. [12], train a separate recurrent network, one for each future time step, to predict the location of nearby cars. Ballan et al. [3] introduce a dynamic Bayesian network to model motion dependencies from previously seen patterns and apply them to unseen scenes by transferring the knowledge between similar settings. In an interesting work, a variational auto-encoders is used by Lee et al. [14] to learn static scene context (and agents in a small neighborhood) and rank the generated trajectories accordingly. Sadeghian et al. [26], also use top-view images and learn to predict trajectories based on the static scene context. Our work is similar to [26] in the sense that we both use attentive recurrent neural networks to predict trajectories considering the physical surrounding; nonetheless, our model is able to take into account other surrounding agents and is able to generate multiple plausible paths using a GAN module.

Agent-Agent Models. Traditional models for modeling and predicting human-human interactions used "social forces" to capture human motion patterns [9, 18, 31, 22, 2, 23, 21, 17]. The main disadvantage of these models is the need to hand-craft rules and features, limiting their ability to efficiently learn beyond abstract level and the domain experts.

Modern socially-aware trajectory prediction work usually use recurrent neural networks [25, 1, 14, 6, 5, 4, 11, 32]. Hug et al. [10] present an experiment-based study the effectiveness of some RNN models in the context socially aware trajectory prediction. These methods are relatively successful, however, most of these methods only take advantage of the local interactions and don't take into account further agents. In a more recent work, Gupta et al. [8] address this issue as well as the fact that agent's trajectories may have multiple plausible futures, by using GANs. Nevertheless, their method treats the influence of all agents on each other uniformly. In contrast, our method uses a novel attention framework to highlight the most important agents for each targeted agent.

Few recent approaches [14, 30, 4, 28], to some extent, incorporate both the scene and social factors into their models. However, these models only consider the interaction among the limited adjacent agents and are only able to generate a single plausible path for each agent. We address all these limitations by applying wiser strategies such as 1- using visual attention component to process the scene context and highlight the most salient features of the scene for each agent, 2- using a social attention component that estimates the amount of contribution from each agent on the future path prediction of a targeted agent, and 3- using GAN to estimate a distribution over feasible paths for each agent. We support our claims by demonstrating state-of-the-art performance on several standard trajectory prediction datasets.

3. SoPhie

Our goal is to develop a model that can successfully predict future trajectories of a set of agents. To this end, the route taken by each agent in future needs to be influenced not only by its own state history, but also the state of other agents and physical terrain around its path. SoPhie takes all these cues into account when predicting each agent's future trajectory.

3.1. Problem Definition

Trajectory prediction can be formally stated as the problem of estimating the state of all agents in future, given the scene information and their past states. In our case, the scene information is fed as an image I^t, *e.g.* a top-view or angle-view image of the scene at time t, into the model. Moreover, the state of each agent i at time t is assumed to be its location, *e.g.* its 2D coordinate $(x_i^t, y_i^t) \in \mathbb{R}^2$ with respect to a reference, *e.g.* the image corner or the top view's world coordinates. Therefore, the past and current states of the N agents are represented by the ordered set of their 2D locations as:

$$X_i^{1:t} = \{(x_i^\tau, y_i^\tau) | \tau = 1, \cdots, t\} \quad \forall i \in [N],$$

where $[N] = \{1, \cdots, N\}$. Throughout the paper, we use the notations $X_{1:N}^{\cdot}$ and $X_{1:N\setminus i}^{\cdot}$ to represent the collection of all N agents' states and all agents' states excluding the target agent i, respectively. We also use the notation Y^τ, to represent the future state in $t+\tau$. Therefore, the future ground truth and the predicted states of the agent i, between frames $t+1$ and $t+T$ for $T > 1$, are denoted by $Y_i^{1:T}$ and $\hat{Y}_i^{1:T}$ respectively, where

$$Y_i^{1:T} = \{(x_i^\tau, y_i^\tau) | \tau = t+1, \cdots, t+T\} \quad \forall i \in [N].$$

Our aim is to learn the parameters of a model W^* in order to predict the future states of each agent between $t+1$ and $t+T$, given the input image at time t and all agents' states up to the current frame t, *i.e.*

$$\hat{Y}_i^{1:T} = f(I^t, X_i^{1:t}, X_{1:N\setminus i}^{1:t}; W^*),$$

where the model parameters W^* is the collection of the weights for all deep neural structures used in our model. We train all the weights end-to-end using back-propagation and stochastic gradient descent by minimizing a loss $\mathcal{L}_{GAN}$ between the predicted and ground truth future states for all agents. We elaborate the details in the following section.

3.2. Overall Model

Our model consists of three key components including: 1- A feature extractor module, 2- An attention module, and 3- An LSTM based GAN module (Fig. 2). First, the feature extractor module extracts proper features from the scene, *i.e.* the image at the current frame I^t, using a convolutional neural network. It also uses an LSTM encoder to encode an index invariant, but temporally dependent, feature between the state of each agent, $X_i^{1:t}$, and the states of all other agents up to the current frame, $X_{1:N\setminus i}^{1:t}$ (Fig. 2(a)). Then, the attention module highlights the most important information of the inputted features for the next module (Fig. 2 (b)). The attention module consists of two attention mechanisms named as *social* and *physical* attention components. The physical attention learns the spatial (physical) constraints in the scene from the training data and concentrates on physically feasible future paths for each agent. Similarly, the social attention module learns the interactions between agents and their influence on each agent's future path. Finally, the LSTM based GAN module (Fig. 2 (c)) takes the highlighted features from the attention module to generate a sequence of plausible and realistic future paths for each agent. In more details, an LSTM decoder is used to predict the temporally dependent state of each agent in future, *i.e.* $\hat{Y}_i^{1:T}$. Similar to GAN, a discriminator is also applied to improve the performance of the generator model by forcing it to produce more realistic samples (trajectories). In the following sections, we elaborate each module in detail.

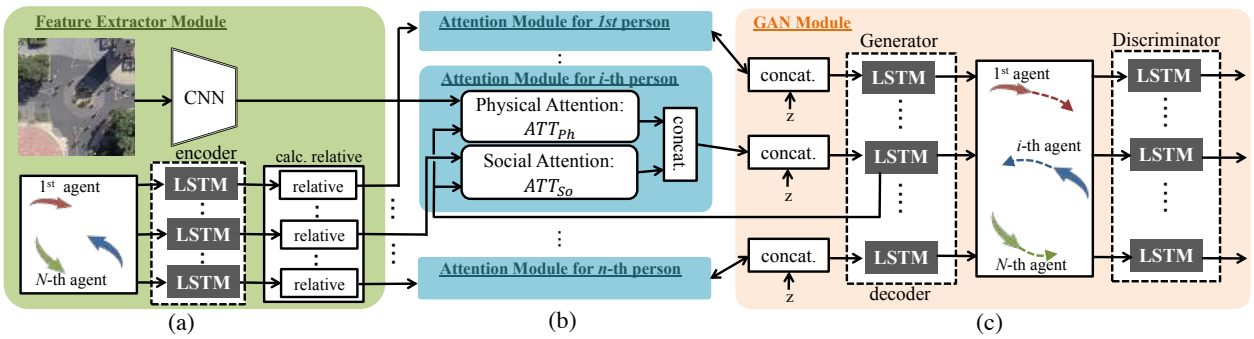

Figure 2. An overview of SoPhie architecture. Sophie consists of three key modules including: (a) A feature extractor module, (b) An attention module, and (c) An LSTM based GAN module.

3.3. Feature extractors

The feature extractor module has two major components, explained below. To extract the visual features V^t_{Ph} from the image I^t, we use a Convolutional Neural Network (CNN).

$$V^t_{Ph} = CNN(I^t; W_{cnn}) \tag{1}$$

In this paper, we use VGGnet-19 [27] as $CNN(\cdot)$, where its weights W_{cnn} is initialized by pre-training on ImageNet [24] and fine-tuning on the task of scene segmentation as described in [16].

To extract joint features from the past trajectory of all agents, we perform the following procedure. Similar to [8], first an LSTM is used to capture the temporal dependency between all states of an agent i and encode them into a high dimensional feature representation for time t, *i.e.*

$$V^t_{en}(i) = LSTM_{en}(X^t_i, h^t_{en}(i); W_{en}), \tag{2}$$

where $h^t_{en}(i)$ represents the hidden state of the encoder LSTM at time t for the agent i. Moreover, to capture the influence of the other agents' state on the prediction of the future trajectory of an agent, we need to extract a joint feature from all agents' encoded features $V^t_{en}(\cdot)$. However, this joint feature cannot be simply created by concatenating them as the order of the agents does matter. To make the joint feature permutation invariant with respect to the index of the agents, the existing approaches use a permutation invariant (symmetric) function such as *max* [8]. Then, this joint global feature is concatenated by each agent's feature $V^t_{en}(i)$ to be fed to the state generator module. However this way, all agents will have an identical joint feature representation. In addition, the permutation invariant functions such as *max* may discard important information of their inputs as they might loose their uniqueness. To address these two limitations, we instead define a consistent ordering structure, where the joint feature for a target agent i is constructed by sorting the other agents' distances from agent i, *i.e.*

$$V^t_{So}(i) = \big(V^t_{en}(\pi_j) - V^t_{en}(i) \big| \forall \pi_j \in [N] \backslash i)\big), \tag{3}$$

where π_j is the index of the other agents sorted according to their distances to the target agent i. In this framework, each agent i has its own unique joint (social) feature vector. We also use *sort* as the permutation invariant function, where the reference for ordering is the euclidean distance between the target agent i and other agents. Note that *sort* function is advantageous in comparison with *max* as it can keep the uniqueness of the input. To deal with variable number of agents, we set a maximum number of agents ($N = N_{max}$) and use a dummy value as features if the corresponding agent does not exist in the current frame.

3.4. Attention Modules

Similar to humans who pays more attention to close obstacles, upcoming turns and people walking towards them, than to the buildings or people behind them, we want the model to focus more on the salient regions of the scene and the more relevant agents in order to predict the future state of each agent. To achieve this, we use two separate soft attention modules similar to [29] for both physical V^t_{Ph} and social $V^t_{So}(i)$ features.

Physical Attention The inputs to this attention module $ATT_{Ph}(\cdot)$ are the hidden states of the decoder LSTM in the GAN module, and the visual features extracted from the image V^t_{Ph}. Note that, the hidden state of the decoder LSTM has the information for predicting the agent's future path. And this module learns the spatial (physical) constraints in the scene from the training data. Therefore, the output would be a context vector C^t_{Ph}, which concentrates on feasible paths for each agent.

$$C^t_{Ph}(i) = ATT_{Ph}(V^t_{Ph}, h^t_{dec}(i); W_{Ph}) \tag{4}$$

Here, W_{Ph} are the parameters of the physical attention module and $h^t_{dec}(i)$ represents the hidden state of the decoder LSTM at time t for the agent i.

Social Attention Similar to the physical attention module, the joint feature vector $V^t_{So}(i)$ together with the hidden state of the decoder LSTM for the i-th agent, are fed to the social attention module $ATT_{So}(\cdot)$ with the parameters W_{So} to obtain a social context vector $C^t_{So}(i)$ for the i-th agent.

This vector highlights which other agents are most important to focus on when predicting the trajectory of the agent i.

$$C_{So}^{t}(i) = ATT_{So}(V_{So}^{t}(i), h_{dec}^{t}(i); W_{So}) \tag{5}$$

We use soft attention similar to [29] for both $ATT_{Ph}(\cdot)$ and $ATT_{So}(\cdot)$, which is differentiable and the whole architecture can be trained end-to-end with back-propagation. Social attention and physical attention aggregate information across all the involved agents and the physical terrain to deal with the complexity of modeling the interactions of all agents in crowded areas while adding interpretability to our predictions. This also suppresses the redundancies of the input data in a helpful fashion, allowing the predictive model to focus on the important features. Our experiments show the contribution of our attention modules in Table 1.

3.5. LSTM based Generative Adversarial Network

In this section, we present our LSTM based Generative Adversarial Network (GAN) module that takes the social and physical context vectors for each agent i, $C_{So}^{t}(i)$ and $C_{Ph}^{t}(i)$, as input and outputs candidate future states which are compliant to social and physical constraints. Most existing trajectory prediction approaches use the L2 norm loss between the ground truth and the predictions to estimate the future states [26]. By using L2 loss, the network only learns to predict one future path for each agent, which is intuitively the average of all feasible future paths for each agent. Instead, in our model, we use GAN to learn and predict a distribution over all the feasible future paths.

GANs consist of two networks, a *generator* and a *discriminator* that compete with each other. The generator is trained to learn the distribution of the paths and to generate a sample of the possible future path for an agent while the discriminator learns to distinguish the feasibility or infeasibility of the generated path. These networks are simultaneously trained in a two player min-max game framework. In this paper similar to [8], we use two LSTMs, a decoder LSTM as the generator and a classifier LSTM as the discriminator, to estimate the temporally dependent future states.

Generator (G) Our generator is a decoder LSTM, $LSTM_{dec}(\cdot)$. Similar to the conditional GAN [19], the input to our generator is a white noise vector z sampled from a multivariate normal distribution while the physical and social context vectors are its conditions. We simply concatenate the noise vector z and these context vectors as the input, *i.e.* $C_{G}^{t}(i) = [C_{So}^{t}(i), C_{Ph}^{t}(i), z]$. Thus, the generated τ^{th} future state's sample for each agent is attained by:

$$\hat{Y}_{i}^{\tau} = LSTM_{dec}\big(C_{G}^{t}(i), h_{dec}^{\tau}(i); W_{dec}\big), \tag{6}$$

Discriminator (D) The discriminator in our case is another LSTM, $LSTM_{dis}(\cdot)$, which its input is a randomly chosen trajectory sample from the either ground truth or predicted future paths for each agent up to τ^{th} future time frame, *i.e.* $T_{i}^{1:\tau} \sim p(\hat{Y}_{i}^{1:\tau}, Y_{i}^{1:\tau})$

$$\hat{L}_{i}^{\tau} = LSTM_{dis}(T_{i}^{\tau}, h_{dis}^{\tau}(i); W_{dis}), \tag{7}$$

where $\hat{L}_{i}^{\tau}$ is the predicted label from the discriminator for the chosen trajectory sample to be a ground truth (real) $Y_{i}^{1:\tau}$ or predicted (fake) $\hat{Y}_{i}^{1:\tau}$ with the truth label $L_{i}^{\tau} = 1$ and $L_{i}^{\tau} = 0$, respectively. The discriminator forces the generator to generate more realistic (plausible) states.

Losses To train SoPhie, we use the following losses:

$$W^{*} = \underset{W}{\operatorname{argmin}} \quad \mathbb{E}_{i,\tau}[\mathcal{L}_{GAN}\big(\hat{L}_{i}^{\tau}, L_{i}^{\tau}\big) + \lambda \mathcal{L}_{L2}(\hat{Y}_{i}^{1:\tau}, Y_{i}^{1:\tau})], \tag{8}$$

where W is the collection of the weights of all networks used in our model and λ is a regularizer between two losses.

The adversarial loss $\mathcal{L}_{GAN}(\cdot,\cdot)$ and L2 loss $\mathcal{L}_{L2}(\cdot,\cdot)$ are shown as follows:

$\mathcal{L}_{GAN}\big(\hat{L}_{i}^{\tau}, L_{i}^{\tau}\big) =$

$$\min_{G}\max_{D} \quad \mathbb{E}_{T_{i}^{1:\tau}\sim p(Y_{i}^{1:\tau})}[L_{i}^{\tau}\log\hat{L}_{i}^{\tau}] + \mathbb{E}_{T_{i}^{1:\tau}\sim p(\hat{Y}_{i}^{1:\tau})}[(1-L_{i}^{\tau})\log(1-\hat{L}_{i}^{\tau})], \tag{9}$$

$$\mathcal{L}_{L2}(\hat{Y}_{i}^{\tau}, Y_{i}^{\tau}) = ||\hat{Y}_{i}^{\tau} - Y_{i}^{\tau}||_{2}^{2}. \tag{10}$$

4. Experiments

In this section, we first evaluate our method on the commonly used datasets such as ETH [22] and UCY [15], and on a recent and larger dataset, *i.e.* Stanford drone dataset [23]. We also compare its performance against the various baselines on these datasets. Next, we present a qualitative analysis of our model on the effectiveness of the attention mechanisms. Finally, we finish the section by demonstrating some qualitative results on how our GAN based approach provides a good indication of path traversability for agents.

Datasets We perform baseline comparisons and ablation experiments on three core datasets. First, we explore the publicly available ETH [22] and UCY [15] datasets, which both contain annotated trajectories of real world pedestrians interacting in a variety of social situations. These datasets include nontrivial movements including pedestrian collisions, collision avoidance behavior, and group movement. Both of the datasets consists of a total of five unique scenes, Zara1, Zara2, and Univ (from UCY), and ETH and Hotel (from ETH). Each scene includes top-view images and 2D locations of each person with respect to the world coordinates. One image is used per scene as the cameras remain static. Each scene occurs in a relatively unconstrained outdoor environment, reducing the impact of physical constraints. We also explore the Stanford Drone Dataset (SDD) [23], a benchmark dataset for trajectory prediction problems. The dataset

	Baselines					SoPhie (Ours)				
Dataset	**Lin**	**LSTM**	**S-LSTM**	**S-GAN**	**S-GAN-P**	$\mathbf{T_A}$	$\mathbf{T_O + I_O}$	$\mathbf{T_O + I_A}$	$\mathbf{T_A + I_O}$	$\mathbf{T_A + I_A}$
ETH	1.33 / 2.94	1.09 / 2.41	1.09 / 2.35	0.81 / 1.52	0.87 / 1.62	0.90 / 1.60	0.86 / 1.65	0.71 / 1.47	0.76 / 1.54	**0.70 / 1.43**
HOTEL	**0.39 / 0.72**	0.86 / 1.91	0.79 / 1.76	0.72 / 1.61	0.67 / 1.37	0.87 / 1.82	0.84 / 1.80	0.80 / 1.78	0.83 / 1.79	0.76 / 1.67
UNIV	0.82 / 1.59	0.61 / 1.31	0.67 / 1.40	0.60 / 1.26	0.76 / 1.52	**0.49 / 1.19**	0.58 / 1.27	0.55 / 1.23	0.55 / 1.25	0.54 / 1.24
ZARA1	0.62 / 1.21	0.41 / 0.88	0.47 / 1.00	0.34 / 0.69	0.35 / 0.68	0.38 / 0.72	0.34 / 0.68	0.35 / 0.67	0.32 / 0.64	**0.30 / 0.63**
ZARA2	0.77 / 1.48	0.52 / 1.11	0.56 / 1.17	0.42 / 0.84	0.42 / 0.84	0.38 / 0.79	0.40 / 0.82	0.43 / 0.87	0.41 / 0.80	**0.38 / 0.78**
AVG	0.79 / 1.59	0.70 / 1.52	0.72 / 1.54	0.58 / 1.18	0.61 / 1.21	0.61 / 1.22	0.61 / 1.24	0.57 / 1.20	0.58 / 1.20	**0.54 / 1.15**

Table 1. Quantitative results of baseline models vs. SoPhie architectures across datasets on the task of predicting 12 future timesteps, given the 8 previous ones. Error metrics reported are ADE / FDE in meters. SoPhie models consistently outperform the baselines, due to the combination of social and physical attention applied in a generative model setting.

	Baselines						SoPhie (Ours)				
Dataset	**Lin**	**SF**	**S-LSTM**	**S-GAN**	**CAR-Net**	**DESIRE**	$\mathbf{T_A}$	$\mathbf{T_O + I_O}$	$\mathbf{T_O + I_A}$	$\mathbf{T_A + I}$	$\mathbf{T_A + I_A}$
SDD	37.11 / 63.51	36.48 / 58.14	31.19 / 56.97	27.246 / 41.440	25.72 / 51.8	19.25 / 34.05	17.76 / 32.14	18.40 / 33.78	16.52 / 29.64	17.57 / 33.31	**16.27 / 29.38**

Table 2. ADE and FDE in pixels of various models on Stanford Drone Dataset. SoPhie's main performance gain comes from the joint introduction of social and physical attention applied in a generative modeling setting.

consists of a bird's-eye view of 20 unique scenes in which pedestrians, bikes, and cars navigate on a university campus. Similar to the previous datasets, images are provided from a top-view angle, but coordinates are provided in pixels. These scenes are outdoors and contain physical landmarks such as buildings and roundabouts that pedestrians avoid.

Implementation details We iteratively trained the generator and discriminator models with the Adam optimizer, using a mini-batch size of 64 and a learning rate of 0.001 for both the generator and the discriminator. Models were trained for 200 epochs. The encoder encodes trajectories using a single layer MLP with an embedding dimension of 16. In the generator this is fed into a LSTM with a hidden dimension of 32; in the discriminator, the same occurs but with a dimension of 64. The decoder of the generator uses a single layer MLP with an embedding dimension of 16 to encoder agent positions and uses a LSTM with a hidden dimension of 32. In the social attention module, attention weights are retrieved by passing the encoder output and decoder context through multiple MLP layers of sizes 64, 128, 64, and 1, with interspersed ReLu activations. The final layer is passed through a Softmax layer. The interactions of the surrounding $N_{max} = 32$ agents are considered; this value was chosen as no scenes in either dataset exceeded this number of total active agents in any given timestep. If there are less than N_{max} agents, the dummy value of 0 is used. The physical attention module takes raw VGG features (512 channels), projects those using a convolutional layer, and embeds those using a single MLP to an embedding dimension of 16. The discriminator does not use the attention modules or the decoder network. When training we assume we have observed eight timesteps of an agent and are attempting to predict the next $T = 12$ timesteps. We weight our loss function by setting $\lambda = 1$. Moreover, the generator/discriminator are trained jointly in a traditional GAN setting.

In addition, to make our model more robust to scene orientation, we augmented the training data by flipping and rotating the scene and also normalization of agents' coordinates. We observed that these augmentations are conducive to make the trained model general enough in order to perform well on the unseen cases in the test examples and different scene geometries such as roundabouts.

Baselines & Evaluation For the first two datasets, a few simple, but strong, baselines are used. These include *Lin*, a linear regressor that estimates linear parameters by minimizing the least square error; *S-LSTM*, a prediction model that combines LSTMs with a social pooling layer, as proposed by Alahi *et. al.* [1]; *S-GAN* and *S-GAN-P*, predictive models that applies generative modeling to social LSTMs [8]. For the drone dataset, we compare to the same linear and Social LSTM baselines, but also explore several other state-of-the-art methods. These include *Social Forces*, an implementation of the same Social Forces model from [31]; *DESIRE*, an inverse optimal control (IOC) model proposed by Lee *et. al.* that utilizes generative modeling; and *CAR-Net*, a physically attentive model from [26]. For all datasets, we also present results of various versions of our SoPhie model in an ablative setting by 1- $\mathrm{T_A}$: Sophie model with social features only and the social attention mechanism, 2- $\mathrm{T_O + I_O}$ Sophie model with both visual and social features without any attention mechanism, 3- $\mathrm{T_O + I_A}$ Sophie model with both visual and social features with only visual attention mechanism, 4- $\mathrm{T_A + I_O}$ Sophie model with both visual and social features with only social attention mechanism, and 5- $\mathrm{T_A + I_A}$ complete Sophie model with all modules.

All models are evaluated using the average displacement error (ADE) metric defined as the average L2 distance between the ground truth and pedestrian trajectories, over all pedestrians and all time steps, as well as the final displacement error metric (FDE). The evaluation task is defined to be performed over 8 seconds, using the past 8 positions consisting of the first 3.2 seconds as input, and predicting the

remaining 12 future positions of the last 4.8 seconds. For the first two datasets, we follow a similar evaluation methodology to [8] by performing a leave-one-out cross-validation policy where we train on four scenes, and test on the remaining one. These two datasets are evaluated in meter space. For the SDD, we utilize the standard split, and for the sake of comparison to baselines we report results in pixel space, after converting from meters.

4.1. Quantitative Results

ETH and UCY We compare our model to various baselines in Table 1, reporting the average displacement error (ADE) in meter space, as well as the final displacement error (FDE). As expected, we see that in general the linear model performs the worst, as it is unable to model the complex social interactions between different humans and the interactions between humans and their physical space. We also notice that S-LSTM provides an improvement over the linear baseline, due to its use of social pooling, and that S-GAN provides an improvement to this LSTM baseline, by approaching the problem from a generative standpoint.

Our first model, $\mathrm{T_A}$, which solely applies social context to pedestrian trajectories, performs slightly better than the S-GAN on average due to better feature extraction strategy and attention module. As expected, although social context helps the model form better predictions, it alone is not enough to truly understand the interactions in a scene. Similarly, while our second model $\mathrm{T_O} + \mathrm{I_O}$ applies both pedestrian trajectories and features from the physical scene (no attention), the lack of any context about these additional features make the model unable to learn which components are most important, giving it a similar accuracy to $\mathrm{T_A}$. Our first major gains in model performance come when exploring the $\mathrm{T_O} + \mathrm{I_A}$ and $\mathrm{T_A} + \mathrm{I_O}$ models. Because the former applies physical context to image features and the latter applies social context to trajectory features, each model is able to learn the important aspects of interactions, allowing them to slightly outperform the previous models. Interestingly, $\mathrm{T_O} + \mathrm{I_A}$ performs slightly better than $\mathrm{T_A} + \mathrm{I_O}$ potentially suggesting that understanding physical context is slightly more helpful in a prediction task. The final SoPhie model, consisting of social attention on trajectories and physical attention on image features ($\mathrm{T_A} + \mathrm{I_A}$) outperformed the previous models, suggesting that combining both forms of attention allows for robust model predictions.

Stanford Drone Dataset We next compare our method to various baselines in Table 2, reporting the ADE and FDE in pixel space. Much like the previous datasets, with SDD we see that the linear baseline performs the worst, with S-LSTM and S-GAN providing an improvement in accuracy. The next major improvement in accuracy is made with CAR-Net, due to the use of physical attention. This is likely due to the nature of SDD, where pedestrian movements based on the curvature of the road can be extrapolated from the birds eye view of the scene. The next major improvement in accuracy is made with the DESIRE framework, which explores trajectory prediction from a generative standpoint, making it the best baseline. Note that the DESIRE results are linearly interpolated from the 4.0s result reported in [14] to 4.8s, as their code is not publicly available. Finally, incorporating social context in $\mathrm{T_A}$, as well as both social and physical context in $\mathrm{T_A} + \mathrm{I_A}$ allow for significant model improvements, suggesting that both attentive models are crucial to tackling the trajectory prediction problem.

Impact of social and physical constraints. Since the goal is to produce socially acceptable paths we also used a different evaluation metrics that reflect the percentage of near-collisions (if two pedestrians get closer than the threshold of $0.10m$). We have calculated the average percentage of pedestrian near collisions across all frames in each of the BIWI/ETH scenes. These results are presented in Table 3. To better understand our model's ability to also produce physically plausible paths, we also split the test set of the Stanford Drone Dataset into two subsets: simple and complex, as previously done in CAR-Net [26] and report results in Table 4. We note that the S-GAN baseline achieves decent performance on simple scenes, but is unable to generalize well to physically complex ones. On the other hand, CAR-Net and SoPhie both achieves a slight performance increase on simple scenes over S-GAN and trajectory only LSTM, as well as nearly halving the error on complex scenes, due to this physical context. This experiment demonstrates that Sophie's use of physical and social attention successfully allows it to predict better physical and socially acceptable paths compared to baseline methods. We also want to note that, unfortunately, the existing benchmarks in trajectory prediction are still naive and were not developed for evaluating the social and physical aspects of trajectories. In this paper we tried to evaluate our method and baselines methods using simple metrics. However, proper benchmarks with specific metrics would be a good future direction.

	GT	LIN	S-GAN	SoPhie
ETH	0.000	3.137	2.509	**1.757**
HOTEL	0.092	**1.568**	1.752	1.936
UNIV	0.124	1.242	**0.559**	0.621
ZARA1	0.000	3.776	1.749	**1.027**
ZARA2	0.732	3.631	2.020	**1.464**
Avg	0.189	2.670	1.717	**1.361**

Table 3. Average % of colliding pedestrians per frame for each of the scenes in BIWI/ETH. A collision is detected if the euclidean distance between two pedestrians is less than 0.10m.

Nexus 6 Little 1 Huang 1

Figure 3. Using the generator to sample trajectories and the discriminator to validate those paths, we present highly accurate traversability maps for SDD scenes. Maps are presented in red, and generated only with 30 starting samples, illustrated as blue crosses.

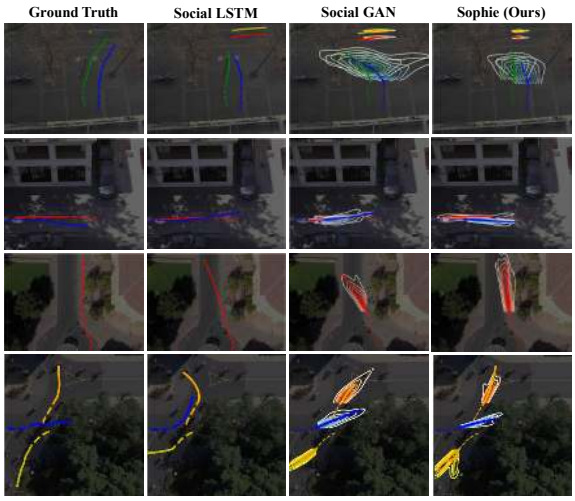

Figure 4. Comparison of Sophie's predictions against the ground truth trajectories and two baselines. Each pedestrian is displayed with a different color, where dashed lines are observed trajectories and solid lines are predicted. Generative models also have a distribution of predicted samples.

Model	Complex	Simple
LSTM	31.31	30.48
CAR-Net	24.32	30.92
S-GAN	29.29	**22.24**
SoPhie	**15.61**	**21.08**

Table 4. Performance of multiple baselines on the Stanford Drone Dataset, split into physically simple and complex scenes. Error is ADE and is reported in pixels.

4.2. Qualitative Results

We further investigate the ability of our architecture to model how social and physical interactions impact future trajectories. Fig. 4 demonstrates the affects that attention can have in correcting erroneous predictions. Here we visualize four unique scenarios, comparing Sophie to two baselines and the ground truth pedestrian movements. In the first two scenarios, the variability of predictions is reduced, allowing pedestrian collisions to be avoided. In the last two scenarios, the physical attention ensures that the pedestrians follow physical constraints, such as staying on sidewalks. As such, the introduction of social and physical attention not only allows for greater model interpretability but also better aligns predictions to scene constraints.

An additional benefit of the generative SoPhie architecture is that it can be used to understand which areas in a scene are traversable. To show the effectiveness of our method, we sampled 30 random agents from the test set (i.e., first 8 seconds of each trajectory) Specifically, given a scene, random trajectories from the test set are sampled at various points in the scene and the generator generated sample trajectories using this starting points. These generated trajectories were then validated using the discriminator. The distribution of these trajectories results in an interpretable traversability map, as in Fig. 3. Each image represents a unique scene from SDD, with the overlayed heatmap showing traversable areas and the blue crosses showing the starting samples. With Nexus 6, the model is able to successfully identify the traversable areas as the central road and the path to the side. With Little 1, the model identifies the main sidewalk that pedestrians walk on while correctly ignoring the road that pedestrians avoid. In Huang 1, the model is able to correctly identify the cross section as well as side paths on the image. We thus observe that the generative network can successfully be used to explore regions of traversability in scenes even with a small number of samples.

5. Conclusion

We propose a trajectory prediction framework that outperforms state-of-the-art methods on multiple benchmark datasets. Our method leverages complete scene context and interactions of all agents, while enabling interpretable predictions, using social and physical attention mechanisms. To capture the uncertain nature of the future paths we generate a distribution over the predicted trajectories using an attentive GAN which can successfully generate multiple physically acceptable paths that respect social constraints of the environment. We showed that by modeling jointly the information about the physical environment and interactions between all agents, our model learns to perform better than when this information is used independently. Our experiments demonstrate that Sophie's use of physical and social attention successfully allows it to predict better physical and socially acceptable paths compared to baseline methods.

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. 2, 3, 6

[2] A. Alahi, V. Ramanathan, and L. Fei-Fei. Socially-aware large-scale crowd forecasting. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, number EPFL-CONF-230284, pages 2211–2218. IEEE, 2014. 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. 3

[5] T. Fernando, S. Denman, A. McFadyen, S. Sridharan, and C. Fookes. Tree memory networks for modelling long-term temporal dependencies. *arXiv preprint arXiv:1703.04706*, 2017. 2, 3

[6] 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. 2, 3

[7] 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

[8] A. Gupta, J. Johnson, L. Fei-Fei, S. Savarese, and A. Alahi. Social gan: Socially acceptable trajectories with generative adversarial networks. *arXiv preprint arXiv:1803.10892*, 2018. 2, 3, 4, 5, 6, 7

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

[10] R. Hug, S. Becker, W. Hübner, and M. Arens. On the reliability of lstm-mdl models for pedestrian trajectory prediction. In *VIIth International Workshop on Representation, analysis and recognition of shape and motion FroM Image data (RFMI 2017)*, 2017. 3

[11] R. Hug, S. Becker, W. Hübner, and M. Arens. Particle-based pedestrian path prediction using lstm-mdl models. *arXiv preprint arXiv:1804.05546*, 2018. 3

[12] B. Kim, C. M. Kang, S. H. Lee, H. Chae, J. Kim, C. C. Chung, and J. W. Choi. Probabilistic vehicle trajectory prediction over occupancy grid map via recurrent neural network. *arXiv preprint arXiv:1704.07049*, 2017. 2

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

[14] 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. 2017. 2, 3, 7

[15] A. Lerner, Y. Chrysanthou, and D. Lischinski. Crowds by example. In *Computer Graphics Forum*, volume 26, pages 655–664. Wiley Online Library, 2007. 5

[16] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015. 4

[17] A. Maksai, X. Wang, F. Fleuret, and P. Fua. Non-markovian globally consistent multi-object tracking. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2563–2573. IEEE, 2017. 2

[18] R. Mehran, A. Oyama, and M. Shah. Abnormal crowd behavior detection using social force model. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 935–942. IEEE, 2009. 2

[19] M. Mirza and S. Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014. 5

[20] B. T. Morris and M. M. Trivedi. Trajectory learning for activity understanding: Unsupervised, multilevel, and long-term adaptive approach. *IEEE transactions on pattern analysis and machine intelligence*, 33(11):2287–2301, 2011. 2

[21] S. Pellegrini, A. Ess, K. Schindler, and L. Van Gool. You'll never walk alone: Modeling social behavior for multi-target tracking. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 261–268. IEEE, 2009. 2

[22] S. Pellegrini, A. Ess, and L. Van Gool. Improving data association by joint modeling of pedestrian trajectories and groupings. In *European conference on computer vision*, pages 452–465. Springer, 2010. 2, 5

[23] 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. 2, 5

[24] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015. 4

[25] A. Sadeghian, A. Alahi, and S. Savarese. Tracking the untrackable: Learning to track multiple cues with long-term dependencies. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 300–311, 2017. 3

[26] A. Sadeghian, F. Legros, M. Voisin, R. Vesel, A. Alahi, and S. Savarese. Car-net: Clairvoyant attentive recurrent network. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 151–167, 2018. 1, 2, 5, 6, 7

[27] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 4

[28] S. Su, J. Pyo Hong, J. Shi, and H. Soo Park. Predicting behaviors of basketball players from first person videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1501–1510, 2017. 3

[29] 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, 4, 5

[30] H. Xue, D. Q. Huynh, and M. Reynolds. Ss-lstm: A hierarchical lstm model for pedestrian trajectory prediction. In *Applications of Computer Vision (WACV), 2018 IEEE Winter Conference on*, pages 1186–1194. IEEE, 2018. 2, 3

[31] 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. 2, 6

[32] S. Zheng, Y. Yue, and J. Hobbs. Generating long-term trajectories using deep hierarchical networks. In *Advances in Neural Information Processing Systems*, pages 1543–1551, 2016. 3