Goal-GAN: Multimodal Trajectory Prediction Based on Goal Position Estimation

Patrick Dendorfer, Aljoša Ošep, and Laura Leal-Taixé

Technical University Munich
{patrick.dendorfer,aljosa.osep,leal.taixe}@tum.de

Abstract. In this paper, we present Goal-GAN, an interpretable and end-to-end trainable model for human trajectory prediction. Inspired by human navigation, we model the task of trajectory prediction as an intuitive two-stage process: (i) goal estimation, which predicts the most likely target positions of the agent, followed by a (ii) routing module which estimates a set of plausible trajectories that route towards the estimated goal. We leverage information about the past trajectory and visual context of the scene to estimate a multi-modal probability distribution over the possible goal positions, which is used to sample a potential goal during the inference. The routing is governed by a recurrent neural network that reacts to physical constraints in the nearby surroundings and generates feasible paths that route towards the sampled goal. Our extensive experimental evaluation shows that our method establishes a new state-of-the-art on several benchmarks while being able to generate a realistic and diverse set of trajectories that conform to physical constraints.

1 Introduction

Modeling human motion is indispensable for autonomous systems that operate in public spaces, such as self-driving cars or social robots. Safe navigation through crowded scenes and collision prevention requires awareness not only of the present position but also of the future path of all moving objects. Human trajectory prediction is particularly challenging since pedestrian trajectories depend primarily on their intention – and the destination of a pedestrian is inherently unknown to the external observer. Consider the example of a pedestrian reaching a crossroad such as the one depicted in Figure 1. Based solely on past observations, we cannot infer the future path of the pedestrian: turning right, left, or going straight, are all equally likely outcomes.

For this reason, a powerful prediction model should be able to capture the *multimodality* of this task, *i.e.*, forecast trajectories that cover the distinctive modes present in the scene. Furthermore, it should produce a *diverse* set of the paths within each mode, reflecting inherent uncertainty in walking style, velocity, and different strategies for obstacle avoidance.

To capture the stochastic nature of trajectory prediction, state-of-the-art methods leverage generative the power of variational autoencoders (VAEs) [1,2,3]

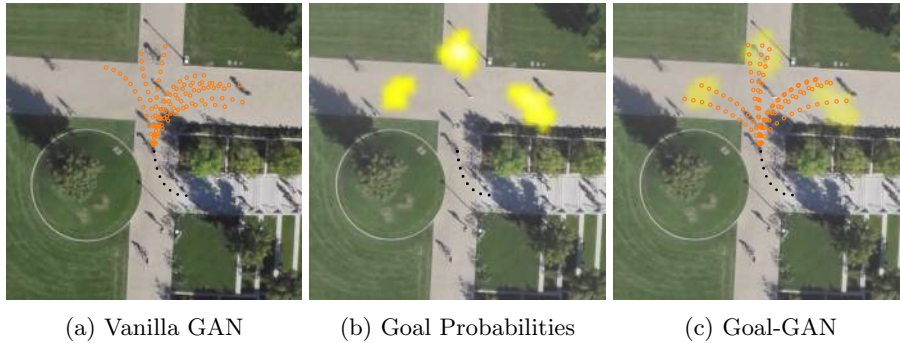| (a) Vanilla GAN | (b) Goal Probabilities | (c) Goal-GAN |

Fig. 1: Visual comparison between predictions of our proposed Goal-GAN and a vanilla GAN. In contrast to the baseline, our proposed model covers all three modes and predicts diverse and feasible trajectories by explicitly estimating realistic goals.

and/or generative adversarial networks (GANs) [4,5,6] to predict a set of trajectories for every observation.

While generative methods are widely used to generate diverse outputs, they are unable to explicitly capture the inherent multimodality of pedestrian trajectories. Often, these methods generate highly diverse trajectories but tend to neglect the physical structure of the environment. The resulting trajectories are not necessarily feasible, and often do not fully cover multiple possible directions that a pedestrian can take (Figure 1a). A more natural way of capturing all feasible directions is to first determine an intermediate goal sampled from a distribution of plausible positions, as shown in Figure 1b. In the second step, the model generates the trajectories reaching the sampled positions (Figure 1c). While social interactions among agents [7,4,5,6] and local scene interaction have been extensively studied, there are almost no methods tackling the challenge of explicitly learning the inherent multimodal distribution of pedestrian trajectories.

In this paper, we aim to bridge this gap and explicitly focus on the underexplored problem of generating diverse multimodal trajectories that conform to the physical constraints. Influenced by recent studies on human navigation [8] we propose an end-to-end trainable method that separates the task of trajectory prediction into two stages. First, we estimate a posterior over possible goals, taking into account the dynamics of the pedestrian and the visual scene context, followed by the prediction of trajectories that route towards these estimated goals. Therefore, trajectories generated by our model take both local scene information and past motion of the agent explicitly into account. While the estimated distribution of possible goal positions reflects the multimodality in the scene, the routing module reacts to local obstacles and generates diverse and feasible

paths. We ensure diversity and realism of the output trajectories by training our network in a generative adversarial setup.

In summary, our main **contribution** is three-fold: (i) we propose Goal-GAN, a two-stage end-to-end trainable trajectory prediction method inspired by human navigation, which separates the prediction task into goal position estimation and routing. (ii) To this end, we design a novel architecture that explicitly estimates an interpretable probability distribution of future goal positions and allows us to sample from it. Using the Gumbel Softmax trick [9] enables us to train the network through the stochastic process. (iii) We establish a new state-of-the-art on several public benchmarks and qualitatively demonstrate that our method predicts realistic end-goal positions together with plausible trajectories that route towards them. The code for Goal-GAN[1] is publicly available.

2   Related Work

Several methods focus on modelling human-human [4,7], human-space interactions [10,2,11], or both [5]. Recent methods leverage generative models to learn a one-to-many mapping, that is used to sample multimodal future trajectories.
Trajectory Prediction. Helbing and Molar introduced the Social Force Model (SFM) [12], a physics-based model, capable of taking agent-agent and agent-space interactions into account. This approach was successfully applied in the domain of multi-object tracking [13,14,15,16]. Since then, data-driven models [17,18,7,19,4] have vastly outperformed physics-based models. Encoder-decoder based methods [2,7] leverage recurrent neural networks (RNNs) [20] to model the temporal evolution of the trajectories with long-short term memory (LSTM) units [21]. These deterministic models cannot capture the stochastic nature of the task, as they were trained to minimize the L_2 distance between the prediction and the ground truth trajectories. This often results in implausible, average-path trajectories.

Recent methods [22,11] focus on human-space interactions using bird-view images [5] and occupancy grids [10,23] to predict trajectories that respect the structural constraints of the scene. Our method similarly leverages bird-eye views. However, we use visual information to explicitly estimate feasible and interpretable goal positions, that can, in turn, be used to explicitly sample end-goals that ease the task of future trajectory estimation.
Generative Models for Trajectory Prediction. Recent works [4,5,6] leverage generative models to sample diverse trajectories rather than just predicting a single deterministic output. The majority of methods either use variational autoencoders (VAEs) [24,3,2,25,26,27,11] or generative adversarial networks (GANs) [28,4,5,6,29]. Social GAN (S-GAN) [4] uses a discriminator to learn the distribution of socially plausible paths. Sadeghian *et al.* [5] extend the model to human-environment interactions by introducing a soft-attention [30] mechanism. GANs have shown promising results for the task of trajectory prediction, but tend to suffer from mode collapse. To encourage the generator to

---

[1] https://github.com/dendorferpatrick/GoalGAN

produce more diverse predictions, [1] uses a best-of-many sampling approach during training while [6] enforces the network to make use of the latent noise vector in combination with BicycleGAN [31] based training. While producing trajectories with high variance, many trajectories are not realistic, and a clear division between different feasible destinations (reflecting inherent multi-modality of the inherent task) is not clear. To account for that, we take inspiration from prior work conditioning the trajectory prediction on specific target destinations.

Goal-conditioned forecasting. In contrast to the aforementioned generative models that directly learn a one-to-many mapping, several methods propose two-stage prediction approaches. Similarly to ours, these methods predict first the final (goal) position, followed by trajectory generation that is conditioned on this position. Early work of [32] models the distribution over possible destinations using a particle filter [33] while other approaches [34] propose a Bayesian framework that estimates both, the destination point together with the trajectory. However, these purely probabilistic approaches are highly unstable during training. The Conditional Generative Neural System (CGNS) [35] uses variational divergence minimization with soft-attention [30] and [36] presents a conditional flow VAE that uses a conditional flow-based prior to effectively structured sequence prediction. These models condition their trajectory generator on initially estimated latent codes but do not explicitly predict a goal distribution nor sample an explicit goal position. Most recently, [37] proposes P2TIRL that uses a maximum entropy inverse reinforcement learning policy to infer goal and trajectory plan over a discrete grid. P2TRL assigns rewards to future goals that are learned by the training policy which is slow and computationally expensive. In contrast, we directly learn the multimodal distribution over possible goals using a binary cross-entropy loss between the (discrete) probability distribution estimate and the ground truth goal position. This makes our work the first method (to the best of our knowledge) that directly predicts an explicit (and discrete) probability distribution for multimodal goals and is efficiently end-to-end trainable.

3   Problem Definition

We tackle the task of predicting the future positions of pedestrians, parametrized via x and y coordinates in the 2D ground plane. As input, we are given their past trajectory and visual information of the scene, captured from a bird-view.

We observe the trajectories $X_i = \{(x_i^t, y_i^t) \in \mathbb{R}^2 | t = 1, \ldots, t_{obs}\}$ of N currently visible pedestrians and a top-down image I of the scene, observed at the timestep t_{obs}. Our goal is to predict the future positions $Y_i = \{(x_i^t, y_i^t) \in \mathbb{R}^2 | t = t_{obs} + 1, \ldots, t_{pred}\}$.

In the dataset, we are only given one future path for t_{obs} – in particular, the one that was observed in practice. We note that multiple distinctive trajectories could be realist for this observed input trajectory. Our goal is, given the input past trajectory X_i, to generate $k \in \{1, \ldots, K\}$ multiple future samples $\hat{Y}_i^k$ for

all pedestrians $i \in \{1, \ldots, N\}$. These should cover all feasible modes and be compliant with the physical constraints of the scene.
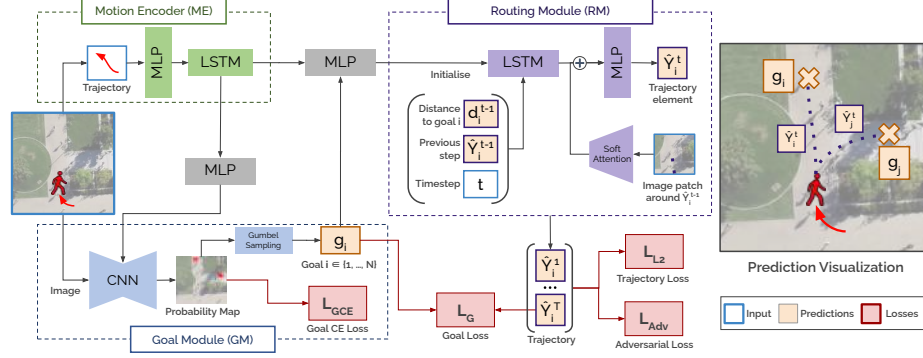

Fig. 2: **Overview of model architecture:** Our model consists of three components: 1) Motion Encoder, 2) Goal Module, and the 3) Routing Module. The Goal Module combines the dynamic features of the Motion Encoder and the scene information to predict a final goal g. The Routing Module generates the future trajectory while taking into account the dynamic features and the estimated goal. During inference, we generate multiple trajectories i by sampling goals from the estimated goal probability map.

4   Goal-GAN

When pedestrians walk through public spaces, they aim to reach a predetermined goal [8], which depends on their intentions and the scene context. Once the goal is set, humans route to their final destination while reacting to obstacles or other pedestrians along their way. This observation motivates us to propose a novel two-stage architecture for trajectory prediction that first estimates the end-goal position and then generates a trajectory towards the estimated goal. Our proposed Goal-GAN consists of three key components, as shown in Figure 2.

- *Motion Encoder (ME):* extracts the pedestrians' dynamic features recursively with a long short-term memory (LSTM) unit capturing the speed and direction of motion of the past trajectory.
- *Goal Module (GM):* combines visual scene information and dynamic pedestrian features to predict the goal position for a given pedestrian. This module estimates the probability distribution over possible goal (target) positions, which is in turn used to sample goal positions.
- *Routing Module (RM):* generates the trajectory to the goal position sampled from the GM. While the goal position of the prediction is determined by the

GM, the RM generates feasible paths to the predetermined goal and reacts to obstacles along the way by using visual attention.

Figure 2 shows an overview of our model. In the following sections, we motivate and describe the different components in detail.

4.1   Motion Encoder (ME)

The past trajectory of a pedestrian is encoded into the Motion Encoder (ME), which serves as a dynamic feature extractor to capture the speed and direction of the pedestrian, similarly to [7,4]. Each trajectory's relative displacement vectors $(\Delta x_i^t, \Delta y_i^t)$ are embedded into a higher dimensional vector e^t with a multi-layer perceptron (MLP). The output is then fed into an LSTM, which is used to encode the trajectories. The hidden state of the LSTM, h_{ME}, is used by the other modules to predict the goal and to decode the trajectory for each pedestrian.
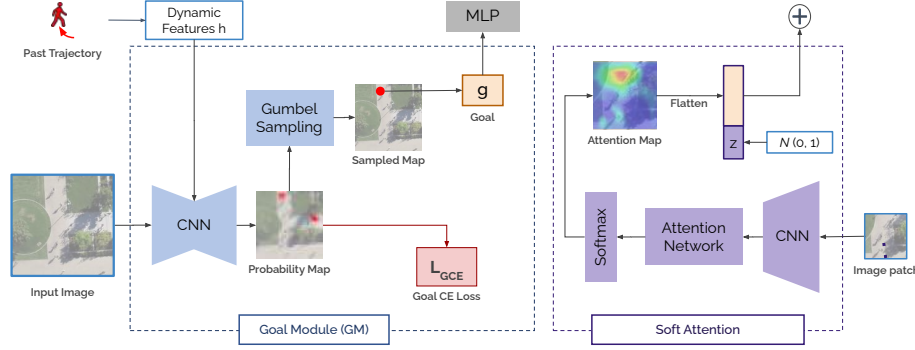
4.2   Goal Module (GM)

Fig. 3: **Goal Module (GM) and Soft Attention (SA).** The Goal Module samples a goal coordinate g while the soft attention assigns attention scores to spatial positions.

In our work, we propose a novel Goal Module (GM). The Goal Module combines visual and dynamic features of the pedestrian to estimate a distribution of possible end goals. As can be seen in Figure 1, the scene dictates the distinctive modes for possible trajectories. Here, the pedestrian can go left, right, or straight. The Goal Module is responsible for capturing all the possible modes and predicting a final goal position, *i.e.*, choosing one of the three options.

Architecture. In order to estimate the goal distribution, the network assesses the visual scene and the dynamics of the pedestrian. The visual scene is represented as an RGB image (or a semantic map) of size $H \times W$, captured from a bird-eye view. This image is input to the goal module.

The scene image is passed through an encoder-decoder CNN network with skip connections similar to [38]. Before the decoder, the scene image features in the bottleneck layer are concatenated with the motion features h_{ME} from the Motion Encoder. Intuitively, the CNN decoder should analyze both the past trajectory and the scene to estimate the future target positions – goals. The module outputs a probability distribution that reflects the multimodal directions for a given input trajectory and scene.

Training through sampling. The CNN decoder outputs a score map $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_n)$ for which each α_i reflects the probability of a particular cell being the end-goal location of the agent.

The discrete probability distribution α is used to sample an end-goal by using the Gumbel-Softmax-Trick [9]. This allows us to sample a discrete distribution over possible goal locations while being able to backpropagate the loss through the stochastic process. The resulting two-dimensional goal position g is sampled randomly from the 2D grid representing the scene,

Goal Sampling vs. Soft Attention. A major novelty of our work is the Goal Module that replaces soft attention [30] to process the scene contextual information [5,11]. Both approaches are illustrated in Figure 3. A soft attention module assigns attention scores to spatially relevant positions based on the visual CNN features. In [5], the attention values are combined with random noise and fed to the trajectory decoder to generate scene-aware multimodal trajectories. However, this often leads to unsatisfying results when the network simply ignores the spatial attention scores or has difficulties combining the attention values with the noise to capture all modes in the scene.

We argue that the attention module is useful when predicting the route towards a goal (as we show in Section 4.3), as it encourages the feasibility of the predicted trajectories. However, the model that solely relies on soft visual attention mechanism tends to generate trajectories that do not capture the multimodal nature of the task, as illustrated in Figure 1. Furthermore, in Section 5, we experimentally confirm that stochasticity of the task is reflected better when sampling from the learned probability distribution, produced by our Goal Module, compared to merely relying on noise injection.

We can directly train the module for the goal position estimation using the Gumbel Softmax trick [9], in combination with the standard cross-entropy loss, which is directly applied to the estimated goal distribution based on the observed (final) position of the ground truth trajectories. We emphasize that we do not use nor need any other data than what is provided in the standard training set.

During the inference, we simply sample the goal from the learned probability distribution and pass it to the decoder. This significantly eases the task for the decoder, as the Goal Module already assesses the visual surroundings and only passes a low dimensional input into the routing module.

4.3   Routing Module (RM)

The Routing Module (see Figure 2) is the third component of our method. It combines the dynamic features and the global goal estimate to generate the final

trajectory prediction. The RM consists of an LSTM network, a visual soft attention network (ATT), and an additional MLP layer that combines the attention map with the output of the LSTM iteratively at each timestep.

First, we forward the goal estimate embedding e_g and the object dynamics embedding h_{ME} (given by the motion encoder, ME) to an MLP to initialise the hidden state h^0_{RM} of the RM.

Then, we recursively estimate predictions for the future time steps. To this end, the LSTM in the RM obtains three inputs: the previous step prediction $\hat{Y}^{t-1}$, the remaining distance to the estimated goal $d_{t-1} = g - \hat{Y}^{t-1}$ and the current scalar timestep value t.

To assess the traversability of the local surroundings, we apply soft attention [30] on the image patch centered around the current position of the pedestrian. As shown in the Figure 3, we combine the output of the LSTM with the attention map F^t to predict the next step $\hat{Y}^t$. The visual attention mechanism allows the RM to react to obstacles or other nearby structures. Finally, we use both the dynamic and visual features to predict the final prediction $\hat{Y}^t$.

4.4   Generative Adversarial Training

In our work, we use a Generative Adversarial Network (GAN) to train our trajectory generator to output realistic and physically feasible trajectories. The GAN consists of a Generator and Discriminator network competing in a two-player min-max game. While the generator aims at producing feasible trajectories, the discriminator learns to differentiate between real and fake samples, i.e., feasible and unfeasible trajectories. Adversarial training is necessary because, in contrast to prediction accuracy, it is not possible to formulate a differential loss in a closed mathematical form that captures the concept of feasibility and realism of the generated trajectories.

The discriminator network consists of an LSTM network that encodes the observed trajectory X. This encoding is used to initialize the second LSTM that processes the predicted trajectory Y together with visual features (obtained from the CNN network, that encodes the image patch centered around the current position) at each time step. Finally, the last hidden state of the $LSTM_{pred}$ is used for the final output of the discriminator.

4.5   Losses

For training our Goal-GAN we use multiple losses addressing the different modules of our model. To encourage the generator to predict trajectories, that are closely resembling the ground truth trajectories, we use a best-of-many [1] distance loss $\mathcal{L}_{L2} = \min_k \|Y - \hat{Y}^{(k)}\|_2$ between our predictions $\hat{Y}$ and the ground truth Y. As an adversarial loss, we employ the *lsgan* [39] loss:

$$\mathcal{L}_{Adv} = \frac{1}{2}\mathbb{E}\left[(D(X,Y) - 1)^2\right] + \frac{1}{2}\mathbb{E}\left[D(X,\hat{Y})^2\right], \tag{1}$$

due to the fact, the original formulation [28] using a classifier with sigmoid cross-entropy function potentially leads to the vanishing gradient problem.

To encourage the network to take into account the estimated goal positions for the prediction, we propose a goal achievement losses $\mathcal{L}_G$ that measures the L_2 distance between the goal prediction g and the actual output $\hat{Y}^{t_{pred}}$,

$$\mathcal{L}_G = \|g - \hat{Y}^{t_{pred}}\|_2. \tag{2}$$

In addition, we use a cross-entropy loss

$$\mathcal{L}_{GCE} = -\log{(p_i)}, \tag{3}$$

where p_i is the probability that is predicted from the Goal Module for the grid cell i corresponding to the final ground-truth position. The overall loss is the combination of the partial losses weighted by λ:

$$\mathcal{L} = \lambda_{Adv}\,\mathcal{L}_{Adv} + \mathcal{L}_{L2} + \lambda_G\,\mathcal{L}_G + \lambda_{GCE}\,\mathcal{L}_{GCE}. \tag{4}$$

5   Experimental Evaluation

In this section, we evaluate our proposed Goal-GAN on three standard datasets used to assess the performance of pedestrian trajectory prediction models: ETH [40], UCY [41] and Stanford Drone Dataset (SDD) [19]. To assess how well our prediction model can cover different possible modes (splitting future paths), we introduce a new, synthetically generated scene.

We compare our method with several state-of-the-art methods for pedestrian trajectory prediction and we qualitatively demonstrate that our method produces multi-modal, diverse, feasible, and interpretable results.

Evaluation measures. We follow the standard evaluation protocol and report the prediction accuracy using Average Displacement Error (ADE) and Final Displacement Error (FDE). Both measures are computed using the L_2 distance between the prediction and ground truth trajectories. The generative models are tested on these metrics with a $N - K$ variety loss [1,4,5]. As in the previous work [19,7], we observe 8 time steps (3.2 seconds) and predict the future 12 time steps (4.8 seconds) simultaneously for all pedestrians in the scene.

Visual input and coordinates. As in [5], we use a single static image to predict trajectories in a given scene. We transform all images into a top-down view using the homography transformation provided by the respective datasets. This allows us to perform all predictions in real-world coordinates.

5.1   Benchmark Results

In this section, we compare and discuss our method's performance against state-of-the-art on ETH [40], UCY [41] and SDD [19] datasets.

Datasets. ETH [40] and UCY datasets [41] contain 5 sequences (ETH:2, UCY: 3), recorded in 4 different scenarios. All pedestrian trajectories are converted

Table 1: Quantitative results for ETH [40] and UCY [41] of Goal-GAN and baseline models predicting 12 future timesteps. We report ADE and FDE in meters.

Dataset	Baseline						Ours
	S-LSTM [7]	S-GAN [4]	S-GAN-P [4]	SoPhie [5]	S-BiGAT [6]	CGNS [35]	Goal GAN
K	1	20	20	20	20	20	20
ETH	1.09/2.35	0.81/1.52	0.87/1.62	0.70/1.43	0.69/1.29	0.62/1.40	**0.59**/**1.18**
HOTEL	0.79/1.76	0.72/1.61	0.67/1.37	0.76/1.67	0.49/1.01	0.70/0.93	**0.19**/**0.35**
UNIV	0.67/1.40	0.60/1.26	0.76/1.52	0.54/1.24	0.55/1.32	**0.48**/1.22	0.60/**1.19**
ZARA1	0.47/1.00	0.34/0.69	0.35/0.68	**0.30**/0.63	**0.30**/0.62	0.32/**0.59**	0.43/0.87
ZARA2	0.56/1.17	0.42/0.84	0.42/0.84	0.38/0.78	0.36/0.75	0.35/0.71	**0.32**/**0.65**
AVG	0.72/1.54	0.58/1.18	0.61/1.21	0.54/1.15	0.48/1.00	0.49/0.97	**0.43**/**0.85**

into real-world coordinates and interpolated to obtain positions every 0.4 seconds. For training and testing, we follow the standard leave-one-out approach, where we train on 4 datasets and test on the remaining one. The Stanford Drone Dataset (SDD) [19] consists of 20 unique video sequences captured at the Stanford University campus. The scenes have various landmarks such as roundabouts, crossroads, streets, and sidewalks, which influence the paths of pedestrians. In our experiments, we follow the train-test-split of [42] and focus on pedestrians.
Baselines. We compare our model to several published methods. S-LSTM [7] uses a LSTM encoder-decoder network with social pooling. S-GAN [4] leverages a GAN framework and S-GAN-P [4] uses max-pooling to model social interactions. SoPhie [5] extends the S-GAN model with a visual and social attention module, and Social-BiGAT [6] uses a BicycleGAN [43] based training. DESIRE [2] is an inverse optimal control based model, that utilizes generative modeling. CARNet [11] is a physically attentive model. The Conditional Generative Neural System (CGNS) [35] uses conditional latent space learning with variational divergence minimization to learn feasible regions to produce trajectories. CF-VAE [36] leverages a conditional normalizing flow-based VAE and P2TIRL [37] uses a grid-based policy learned with maximum entropy inverse reinforcement learning policy. As none of the aforementioned provide publicly available implementation, we outline the results reported in the respective publications.
ETH and UCY. We observe a clear trend – the generative models improve the performance of the deterministic approaches, as they are capable of sampling a diverse set of trajectories. Compared to other generative models, Goal-GAN achieves state-of-the-art performance with an overall decrease of the error of nearly 15% compared to *S-BiGAT* and *CGNS*. While *SoPhie* and *S-BiGAT* also use visual input, these models are unable to effectively leverage this information to discover the dominant modes for the trajectory prediction task, thus yielding a higher prediction error. It has to be pointed out that Goal-GAN decreases the

average FDE by $0.12m$ compared to the current state-of-the-art method. We explain the drastic increase in performance with our new Goal Module as we can cover the distribution of all plausible modes and are therefore able to generate trajectories lying close to the ground truth.

Stanford Drone Dataset. We compare our model against other baseline methods on the SDD and report ADE and FDE in pixel space. As it can be seen in Table 2, Goal-GAN achieves state-of-the-art results on both metrics, ADE and FDE. Comparing Goal-GAN against the best non-goal-conditioned method, *So-Phie*, Goal-GAN decreases the error by 25%. This result shows clearly the merit of having a two-stage process of predicting a goal estimate over standard generator methods using only soft attention modules but does not explicitly condition their model on a future goal. Further, it can be understood that multimodal trajectory predictions play a major role in the scenes of the SDD. Also, Goal-GAN exceeds all other goal-conditioned methods and is on par with P2TIRL (which was not yet published during the preparation of this work).

Table 2: Quantitative results for the Stanford Drone Dataset (SDD) [19] of Goal-GAN and baseline models predicting 12 future timesteps. We report ADE and FDE in pixels.

	Baseline								Ours
	S-LSTM [7]	S-GAN [4]	CAR-NET [11]	DESIRE [2]	SoPhie [5]	CGNS [35]	CF-VAE [36]	P2TIRL [37]	Goal GAN
K	1	20	20	20	20	20	20	20	20
ADE	57.0	27.3	25.7	19.3	16.3	15.6	12.6	12.6	**12.2**
FDE	31.2	41.4	51.8	34.1	29.4	28.2	22.3	**22.1**	**22.1**

5.2 Assessing Multimodality of Predictions on Synthetic Dataset

In this section, we conduct an additional experiment using synthetically generated scenarios to study the multimodality of the predictions. We compare the performance of Goal-GAN against two vanilla GAN baselines, with and without visual soft attention [30]. The synthetic dataset allows us to explicitly control multimodality and feasibility of the (generated) ground truth trajectories, as the other datasets do not provide that information.

Dataset. We generate trajectories using the Social Force Model [12] in the *hyang 4* scene of the SDD dataset [19]. To ensure the feasibility of the generated trajectories, we use a two-class (manually labeled) semantic map, that distinguishes between feasible (walking paths) from unfeasible (grass) areas. We simulate 250 trajectories approaching and passing the two crossroads in the scene.

Additional Evaluation Measures. In addition to ADE and FDE, we follow [26,44] to measure the multimodality of the distribution of generated trajec-

Table 3: Quantitative results on our synthetic dataset. We show results obtained by a GAN baseline [4] and different versions of our Goal-GAN model, predicting 12 future time steps. We report ADE, FDE, F (feasibility) and MC (mode coverage) for $k = 10$ sampled trajectories for each scene. We also report the negative log-likelihood (NLL) of the ground truth trajectory computed with the KDE (Kernel Density Estimate), following [26].

Model	Loss	ADE ↓	FDE ↓	F ↑	MC ↑	NLL ↓
GAN w/o visual	$\mathcal{L}_{L2} + \mathcal{L}_{Adv}$	0.70	1.49	59.94	78.51	4.54
GAN w visual	$\mathcal{L}_{L2} + \mathcal{L}_{Adv}$	0.68	1.27	66.51	85.12	4.47
Goal-GAN	$\mathcal{L}_{GCE} + \mathcal{L}_G$	2.09	1.27	76.78	88.22	**3.76**
Goal-GAN	$\mathcal{L}_{L2} + \mathcal{L}_{GCE} + \mathcal{L}_G$	0.62	1.20	85.05	89.27	3.90
Goal-GAN w/o GST	$\mathcal{L}_{L2} + \mathcal{L}_{Adv} + \mathcal{L}_{GCE} + \mathcal{L}_G$	0.84	1.45	76.84	86.27	4.18
Goal-GAN (full model)	$\mathcal{L}_{L2} + \mathcal{L}_{Adv} + \mathcal{L}_{GCE} + \mathcal{L}_G$	**0.55**	**1.01**	**89.47**	**92.48**	3.88

tories. Here we evaluate the negative log-likelihood (NLL) of the ground truth trajectories using a Kernel Density Estimate (KDE) from the sampled trajectories at each prediction timestep. In addition, we define a new mode coverage (MC) metric. For each scene, MC assesses if at least one of the k generated trajectories $\hat{y}$ reaches the final position of the ground truth final up to a distance of $2m$:

$$\text{MC} = \frac{1}{n} \sum_{i}^{n} S(\hat{\mathbf{y}}_i) \ \text{ with } S(\hat{\mathbf{y}}) = \begin{cases} 1 & \text{if } \exists k, \ \|\hat{y}^k - y\|_2 < 2m \\ 0 & \text{else.} \end{cases} \tag{5}$$

To evaluate the feasibility of the trajectories, we report the ratio of trajectories lying inside the feasible area $\mathcal{F}$, *i.e.*, predictions staying on the path:

$$\text{F} = \frac{1}{n} \sum_{i,k}^{n} f(\hat{y}_i^k) \ \text{ with } f(\hat{y}) = \begin{cases} 1 & \text{if } \hat{y} \in \mathcal{F} \\ 0 & \text{else.} \end{cases} \tag{6}$$

Results. As can be seen in Table 3, the vanilla GAN baseline [4] that is not given access to the visual information, yields ADE/FDE of 0.70/1.49, respectively. Adding visual information yields a performance boost (0.68/1.27), however, it is still not able to generate multimodal and feasible paths. When we add our proposed goal module (Goal-GAN) and train it using our full loss, we observe a large boost of performance w.r.t. multimodality (7.36 increase in terms of MC) and feasibility (10.26 increase in terms of F). To ablate our model, we train our network using different loss components, incentivizing the network to train different modules of the network. A variant of our model, trained only with goal achievement loss $\mathcal{L}_G$ and adversarial loss $\mathcal{L}_{Adv}$ can already learn to produce multimodal trajectories (MC of 88.22), however, yields a high ADE error of 2.09. The addition of L_2 loss $\mathcal{L}_{L2}$ significantly increases the accuracy of the predictions (1.47 reduction in ADE), and at the same time, increases

the quality and feasibility (8.26 increase in F), of the predictions. This confirms that our proposed goal module, which explicitly models the distribution over the future goals, is vital for accurate and realistic predictions. Furthermore, we note that the performance drastically drops if we train the full model without the Gumbel-Softmax Trick (GST) (see Section 4.2) which seems to be crucial for stable training, enabling the loss back-propagation through the stochastic sampling process in the Goal Module.
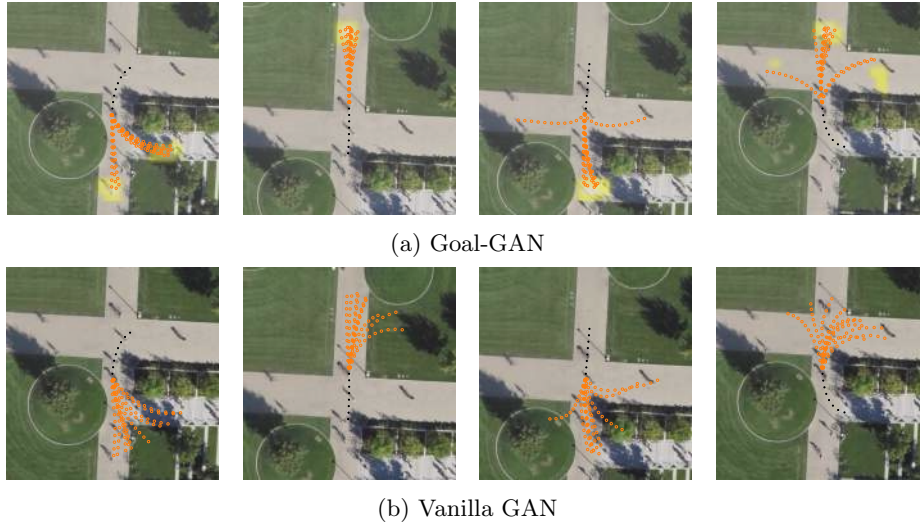

(a) Goal-GAN

(b) Vanilla GAN

Fig. 4: Visualisation of multiple generated trajectories (orange) for past trajectory (black) on the synthetic dataset. We compare the output of our Goal-GAN against the performance of the vanilla GAN using visual attention for $t_{pred} = 12$. For Goal-GAN, the yellow heatmap corresponds to the goal probability map.

5.3   Qualitative Evaluation

In this section, we visually inspect trajectories, generated by our model, and assess the quality of the predictions.

Synthetic Dataset: In Figure 4 we visualize trajectories of the synthetic dataset for our proposed Goal-GAN (top) and the vanilla GAN baseline [4] (bottom). Next to the predicted trajectories (orange circles), we display the probability distribution (yellow heatmap) of goal positions, estimated by the Goal Module. As shown in Figure 4, Goal-GAN predicts a diverse set of trajectories routing to specific estimated modes. Here, we observe that Goal-GAN outputs an interpretable probability distribution that allows us to understand where the model "sees" the dominant modes in the scene. Comparing the quality of the predictions, we can demonstrate that Goal-GAN produces distinct modes while the

GAN baseline tends to instead span its trajectory over a wider range leading to unfeasible paths.

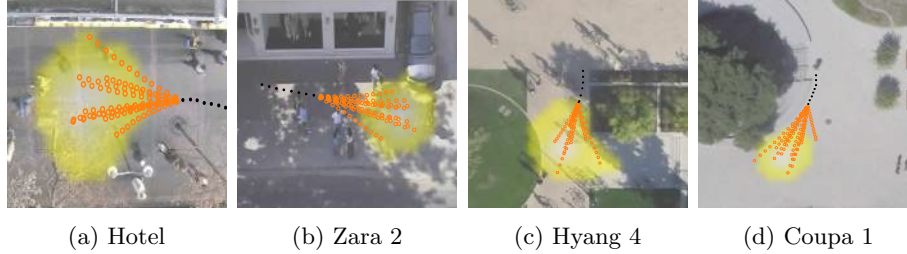| (a) Hotel | (b) Zara 2 | (c) Hyang 4 | (d) Coupa 1 |

Fig. 5: Visualisation of generated trajectories (orange circles) and estimated global goal probabilities (yellow heatmap). The figures show that the model interacts with the visual context of the scene and ensures feasibility predictions.

Real Data: Furthermore, we present qualitative results of the datasets ETH/UCY and SDD in Figure 5. The two figures show predictions on the *Hotel* (Figure 5a) and *Zara 2* (Figure 5b) sequences. Our model assigns high probability to a large area in the scene as in *Hotel* sequence, as several positions could be plausible goals. The broad distribution ensures that we generate diverse trajectories when there are no physical obstacles. Note that the generated trajectories do not only vary in direction but also in terms of speed. In *Zara 2*, the model recognizes the feasible area on the sidewalk and predicts no probability mass on the street or in the areas covered by the parked cars. In the scene *Hyang 4* SDD dataset, we observe that the model successfully identifies that the pedestrian is walking on the path, assigning a very low goal probability to the areas, overgrown by the tree. This scenario is also presented successfully with synthetic data which shows that we can compare the results of the synthetic dataset to the behavior of real data. The trajectories shown for *Coupa 1* demonstrate that the model generates solely paths onto concrete but avoids predictions leading towards the area of the tree.

6   Conclusion

In this work, we present Goal-GAN, a novel two-stage network for the task of pedestrian trajectory prediction. With the increasing interest in the interpretability of data-driven models, Goal-GAN allows us to comprehend the different stages iduring the prediction process. This is an alternative to the current generative models, which use a latent noise vector to encourage multimodality and diversity of the trajectory predictions. Our model achieves state-of-the-art results on the ETH, UCY, and SDD datasets while being able to generate multimodal, diverse, and feasible trajectories, as we experimentally demonstrate.

References

1. Bhattacharyya, A., Schiele, B., Fritz, M.: Accurate and diverse sampling of sequences based on a "best of many" sample objective. In: Conference on Computer Vision and Pattern Recognition. (2018)
2. Lee, N., Choi, W., Vernaza, P., Choy, C.B., Torr, P.H., Chandraker, M.: Desire: Distant future prediction in dynamic scenes with interacting agents. In: Conference on Computer Vision and Pattern Recognition. (2017)
3. Felsen, P., Lucey, P., Ganguly, S.: Where will they go? predicting fine-grained adversarial multi-agent motion using conditional variational autoencoders. In: European Conference on Computer Vision. (2018)
4. Gupta, A., Johnson, J., Fei-Fei, L., Savarese, S., Alahi, A.: Social GAN: Socially acceptable trajectories with generative adversarial networks. In: Conference on Computer Vision and Pattern Recognition. (2018)
5. Sadeghian, A., Kosaraju, V., Sadeghian, A., Hirose, N., Rezatofighi, H., Savarese, S.: Sophie: An attentive GAN for predicting paths compliant to social and physical constraints. In: Conference on Computer Vision and Pattern Recognition. (2019)
6. Kosaraju, V., Sadeghian, A., Martín-Martín, R., Reid, I., Rezatofighi, H., Savarese, S.: Social-BiGAT: Multimodal trajectory forecasting using bicycle-gan and graph attention networks. In: Neural Information Processing Systems. (2019)
7. Alahi, A., Goel, K., Ramanathan, V., Robicquet, A., Fei-Fei, L., Savarese, S.: Social lstm: Human trajectory prediction in crowded spaces. In: Conference on Computer Vision and Pattern Recognition. (2016)
8. Bellmund, J.L.S., Gärdenfors, P., Moser, E.I., Doeller, C.F.: Navigating cognition: Spatial codes for human thinking. Science (2018)
9. Jang, E., Gu, S., Poole, B.: Categorical Reparameterization with Gumbel-Softmax. arXiv e-prints (2016) arXiv:1611.01144
10. Ridel, D.A., Deo, N., Wolf, D.F., Trivedi, M.M.: Scene compliant trajectory forecast with agent-centric spatio-temporal grids. IEEE Robotics Autom. Lett. (2020)
11. Sadeghian, A., Legros, F., Voisin, M., Vesel, R., Alahi, A., Savarese, S.: CAR-Net: Clairvoyant attentive recurrent network. In: European Conference on Computer Vision. (2018)
12. Helbing, D., Molnár, P.: Social force model for pedestrian dynamics. Physical Review E (1995)
13. Scovanner, P., Tappen, M.: Learning pedestrian dynamics from the real world. In: International Conference on Computer Vision. (2009)
14. Pellegrini, S., Ess, A., Schindler, K., van Gool, L.: You'll never walk alone: Modeling social behavior for multi-target tracking. In: International Conference on Computer Vision. (2009)
15. Yamaguchi, K., Berg, A., Ortiz, L., Berg, T.: Who are you with and where are you going? In: Conference on Computer Vision and Pattern Recognition. (2011)
16. Leal-Taixé, L., Pons-Moll, G., Rosenhahn, B.: Everybody needs somebody: Modeling social and grouping behavior on a linear programming multiple people tracker. In: International Conference on Computer Vision Workshop. (2011)
17. Leal-Taixé, L., Fenzi, M., Kuznetsova, A., Rosenhahn, B., Savarese, S.: Learning an image-based motion context for multiple people tracking. In: Conference on Computer Vision and Pattern Recognition. (2014)
18. Milan, A., Rezatofighi, S.H., Dick, A., Reid, I., Schindler, K.: Online multi-target tracking using recurrent neural networks. In: International Conference on Computer Vision. (2017)

19. Robicquet, A., Sadeghian, A., Alahi, A., Savarese, S.: Learning social etiquette: Human trajectory understanding in crowded scenes. In: European Conference on Computer Vision. (2016)
20. E. Rumelhart, D., E. Hinton, G., J. Williams, R.: Learning representations by back propagating errors. Nature (1986)
21. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation (1997)
22. Hiroaki Minoura, Tsubasa Hirakawa, T.Y., Fujiyoshi, H.: Path predictions using object attributes and semantic environment. In: International Conference on Computer Vision Theory and Applications. (2019)
23. Hong, J., Sapp, B., Philbin, J.: Rules of the road: Predicting driving behavior with a convolutional model of semantic interactions. In: Conference on Computer Vision and Pattern Recognition. (2019)
24. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. In: International Conference on Learning Representations. (2014)
25. Deo, N., Trivedi, M.M.: Multi-modal trajectory prediction of surrounding vehicles with maneuver based lstms. Intelligent Vehicles Symposium (2018)
26. Ivanovic, B., Pavone, M.: The trajectron: Probabilistic multi-agent trajectory modeling with dynamic spatiotemporal graphs. In: International Conference on Computer Vision. (2019)
27. Rhinehart, N., McAllister, R., Kitani, K., Levine, S.: Precog: Prediction conditioned on goals in visual multi-agent settings. In: International Conference on Computer Vision. (2019)
28. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Neural Information Processing Systems. (2014)
29. Amirian, J., Hayet, J.B., Pettré, J.: Social ways: Learning multi-modal distributions of pedestrian trajectories with gans. In: Conference on Computer Vision and Pattern Recognition Workshop. (2019)
30. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., Zemel, R., Bengio, Y.: Show, attend and tell: Neural image caption generation with visual attention. In: International Conference on Machine Learning. (2015)
31. Zhu, J.Y., Zhang, R., Pathak, D., Darrell, T., Efros, A.A., Wang, O., Shechtman, E.: Toward multimodal image-to-image translation. In: Neural Information Processing Systems. (2017)
32. Rehder, E., Kloeden, H.: Goal-directed pedestrian prediction. In: International Conference on Computer Vision Workshop. (2015)
33. Thrun, S., Burgard, W., Fox, D.: Probabilistic Robotics (Intelligent Robotics and Autonomous Agents). The MIT Press (2005)
34. Best, G., Fitch, R.: Bayesian intention inference for trajectory prediction with an unknown goal destination. In: International Conference on Intelligent Robots and Systems. (2015)
35. Li, J., Ma, H., Tomizuka, M.: Conditional generative neural system for probabilistic trajectory prediction. In: International Conference on Intelligent Robots and Systems. (2019)
36. Bhattacharyya, A., Hanselmann, M., Fritz, M., Schiele, B., Straehle, C.N.: Conditional flow variational autoencoders for structured sequence prediction. In: Neural Information Processing Systems. (2019)
37. Deo, N., Trivedi, M.M.: Trajectory forecasts in unknown environments conditioned on grid-based plans. arXiv e-prints (2020) arXiv:2001.00735

38. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer Assisted Intervention. (2015)
39. Mao, X., Li, Q., Xie, H., Lau, R.Y.K., Wang, Z., Smolley, S.P.: Least squares generative adversarial networks. In: International Conference on Computer Vision. (2016)
40. Pellegrini, S., Ess, A., Gool, L.V.: Improving data association by joint modeling of pedestrian trajectories and groupings. In: European Conference on Computer Vision. (2010)
41. Lerner, A., Chrysanthou, Y., Lischinski, D.: Crowds by example. Comput. Graph. Forum (2007)
42. Sadeghian, A., Kosaraju, V., Gupta, A., Savarese, S., Alahi, A.: Trajnet: Towards a benchmark for human trajectory prediction. arXiv preprint (2018)
43. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: Conference on Computer Vision and Pattern Recognition. (2016)
44. Thiede, L.A., Brahma, P.P.: Analyzing the variety loss in the context of probabilistic trajectory prediction. In: International Conference on Computer Vision. (2019)