From Goals, Waypoints & Paths To Long Term Human Trajectory Forecasting

Karttikeya Mangalam[†*] Yang An[§*] Harshayu Girase[†] Jitendra Malik[†]

[†] UC Berkeley [§] Technical University of Munich

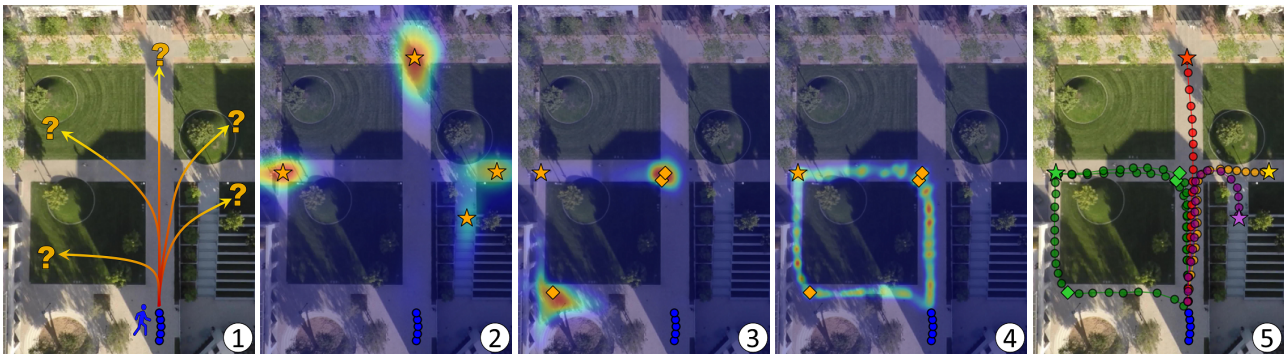

Figure 1: We tackle the problem of long term human trajectory forecasting. Given the past motion of an agent (blue) on a scene over the last five seconds, we aim to predict the multimodal future motion over the next minute ①. To achieve this, we propose factorizing overall multimodality into its *epistemic* and *aleatoric* factors. The *epistemic* factor is modeled with an estimated distribution over the long term goals ② while the *aleatoric* factor is modeled as a distribution over the intermediate waypoints ③ and trajectory ④ for each goal separately. This is repeated for multiple goals and waypoints for scene-compliant multimodal human trajectory forecasting ⑤. Each color indicates predicted trajectories for a different sampled goal.

Abstract

Human trajectory forecasting is an inherently multi-modal problem. Uncertainty in future trajectories stems from two sources: (a) sources that are known to the agent but unknown to the model, such as long term goals and (b) sources that are unknown to both the agent and the model, such as the intent of other agents and irreducible randomness in decisions. We propose to factorize this uncertainty into its epistemic and aleatoric sources. We model the epistemic uncertainty through multimodality in long term goals and the aleatoric uncertainty through multimodality in waypoints and paths. To exemplify this dichotomy, we also propose a novel long term trajectory forecasting setting, with prediction horizons up to a minute, up to an order of magnitude longer than prior works. Finally, we present Y-net, a scene compliant trajectory forecasting network that exploits the proposed epistemic and aleatoric structure for diverse trajectory predictions across long prediction horizons. Y-net significantly improves previous state-of-the-art performance on both (a) The short prediction horizon setting on the Stanford Drone (31.7% in FDE) and ETH/UCY datasets (7.4% in FDE) and (b) The proposed long horizon setting on the re-purposed Stanford Drone and Intersection Drone datasets.

1. Introduction

Sequence prediction is a fundamental problem in several engineering disciplines such as signal processing, pattern recognition, control engineering, and in virtually any domain concerned with temporal measurements. From the seminal work of A. A. Markov [29] on predicting the next syllable in the poem *Eugene Onegin* with Markov chains, to modern day autoregressive descendants like GPT-3 [6], next element prediction in a sequence has a long standing history. Time series forecasting is a key instantiation of the sequence prediction problem in the setting where the sequence is formed by elements sampled in time. Several classic techniques such as Autoregressive Moving Average Models (ARMA) [43] have been incorporated in deep learning architectures [41, 16] in modern day state-of-the-art time series forecasting methods [37].

However, humans are not inanimate Newtonian entities, slaves to predetermined physical laws and forces. Predicting the future motion of a billiard ball smoothly rolling on a pool table under friction and physical constraints is a problem of different nature from forecasting human motion and

∗ indicates equal contribution.

positions. Humans are goal conditioned agents that, unlike the ball, exert their will through actions to achieve a desired outcome [40]. Anticipating human motion is of fundamental importance to dynamic agents such as other humans, autonomous robots [3] and self-driving vehicles [39]. Human motion is inherently goal directed and is put in place by the agent to bring about a desired effect.

Nevertheless, even conditioned on the agent's past motion and overarching long term goals, is the future trajectory deterministic? Consider yourself standing at a crossing on a busy street, waiting for the pedestrian light to turn green. While you have every intention of crossing the street, the exact future trajectory remains stochastic as you might swerve to avoid other pedestrians, speed up your pace if the light is about to turn red, or pause abruptly if an unruly cyclist dashes by. Hence, even conditioned on the past observed motion and scene semantics, future human motion is inherently stochastic [14] owing to both *epistemic* uncertainty caused by latent decision variables like long term goals and *aleatoric* variability [10] stemming from random decision variables such as environmental factors. This dichotomy is even sharper in long term forecasting since due to the increased uncertainty in the future, the aleatoric randomness influences the trajectory much more strongly in long rather than short temporal horizons.

This motivates a factorized multimodal approach for human dynamics modeling where both factors of stochasticity are modeled hierarchically rather than lumped jointly. We hypothesize that the long term latent goals of the agent represent the *epistemic* uncertainty within motion prediction. While the agent has a goal in mind while planning and executing their trajectory, this is unknown to the prediction system. In physical terms, this is akin to the question of *where* the agent wants to go. Similarly, the *aleatoric* uncertainty is expressed in the stochasticity of the path leading to the goal, which encompasses factors like environment variables such as other agents, partial scene information available to the agent and most importantly, the unconscious randomness in human decisions [18]. In physical terms, this is akin to the question of *how* the agent reaches the goal.

Hence, we propose to model the *epistemic* uncertainty first and then model the *aleatoric* stochasticity conditioned on the obtained estimate. Concretely, with the RGB scene and the past motion history, we first estimate an explicit probability distribution over the agent's long term goals. This represents the *epistemic* uncertainty in the prediction system. We also estimate distributions over a few chosen future waypoint positions which along with the sampled goal points are used to obtain explicit probability maps over all the remaining intermediate trajectory positions. This represents the aleatoric uncertainty in the prediction system. Together the samples from the *epistemic* goal distribution and the *aleatoric* waypoint and trajectory distribution form the predicted future trajectory.

In summary, our contribution is threefold. **First**, we propose a novel long term prediction setting that extends up to a minute in the future which is about an order of magnitude longer than previous literature. **Second**, we propose Y-net, a scene-compliant long term trajectory prediction network that explicitly models both the *goal* and *path* multi-modalities while making effective use of the scene semantics. **Third**, we show that the factorized multimodality modeling enables Y-net to improve the state-of-the-art both on the proposed long term settings and the well-studied short term prediction settings. We benchmark Y-net's performance on the Stanford Drone [32] and the ETH [31]/UCY[23] benchmark in the short term setting. It outperforms previous approaches by significant margins of 13.0% in ADE and 31.7% in FDE metric on SDD, and on-par in ADE and by 7.4% in FDE on ETH/UCY. Further, we also study Y-net's performance in the proposed long term prediction setting on the Stanford Drone and the Intersection Drone Dataset [5] where it substantially improves the performance of state-of-the-art short term methods by over 50.7% and 39.7% respectively, on ADE and 77.1% and 56.0% respectively, on FDE metric. The preprocessed data, model, and code can be found here for future work: https://karttikeya.github.io/publication/ynet/

2. Related Works

Several recent studies have investigated human trajectory prediction in different settings. Broadly, these approaches can be grouped based on the proposed formulation for multimodality in forecasting, inputs signals available to the prediction model and the nature and form of prediction results furnished by the model. Several diverse input signals such as agent's past motion history [15], human pose [27], RGB scene image [13, 35, 8, 22, 26], scene semantic cues [8], location [36, 24, 4] and gaze of other pedestrian [27, 46] in the scene, moving vehicles such as cars [36] and also latent inferred signals such as agent's goals [28] have been used. The form of prediction results produced are also diverse with multimodality [26] and scene-compliant forecasting being central to the prior works.

Unimodal Forecasting: Early trajectory forecasting work focused on unimodal predictions of the future. Social Forces [15] proposes modeling interactions as attractive and repulsive forces and future trajectory as a deterministic path evolving under these forces. Social LSTM [1] focuses on other agents in the scene and models their effects through a novel pooling module. [46] forecasts motion in ego-centric views and exploits body pose and gaze along with camera wearer's ego-motion for other agent's future location prediction. [42] proposes to use attention to model target agent's interaction with other agent's. [27] predicts trajec-

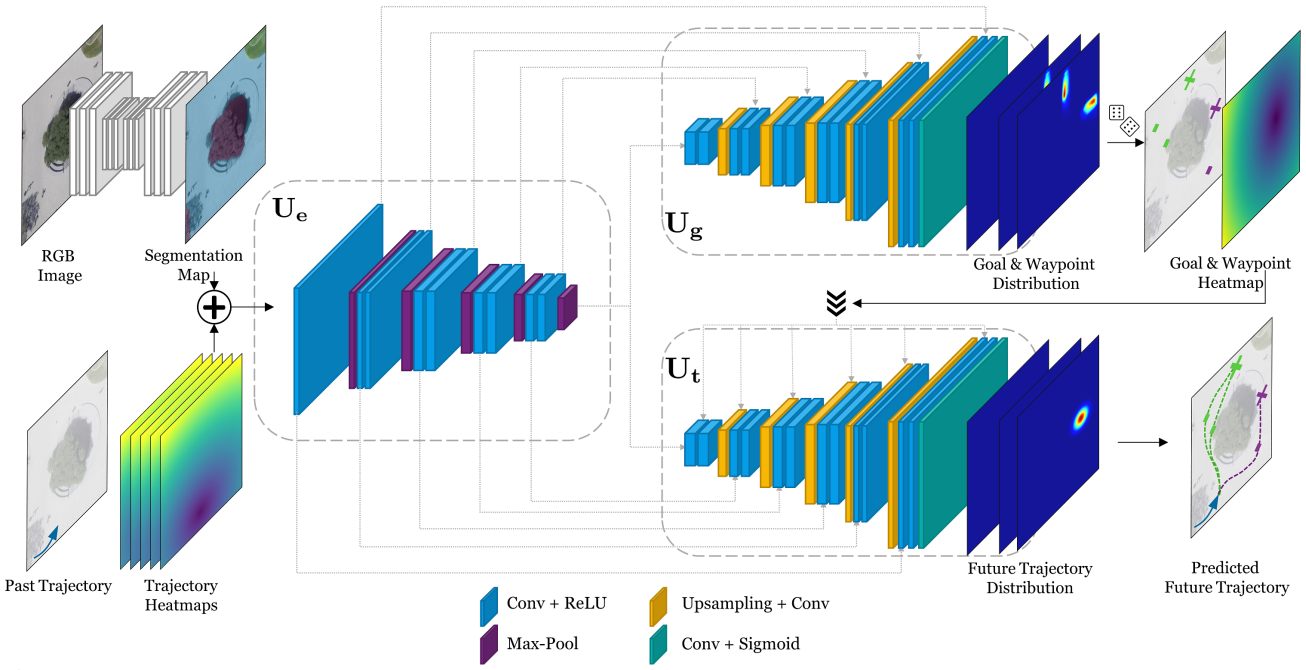

Figure 2: **Model Architecture**: Y-net comprises of three sub-networks $\mathbf{U}_e$, $\mathbf{U}_g$ and $\mathbf{U}_t$ modeled after the U-net architecture [34] (Section 3.1). Y-net adopts a factorized approach to multimodality, expressing the stochasticity in goals and waypoints through estimated distributions furnished by $\mathbf{U}_g$. And multimodality in paths is achieved through estimated probability distributions obtained by $\mathbf{U}_t$ conditioned on samples from $\mathbf{U}_g$ for predicting diverse multimodal scene-compliant futures.

tory as the 'global' branch for pose prediction and proposes to condition downstream tasks such as pose prediction on predicted unimodal trajectories.

Multimodality through Generative Modeling: Another line of work aims to model the stochasticity inherent in future prediction through a latent variable with a defined prior distribution through approaches such as conditional variational auto-encoders [20]. DESIRE [22] is an inverse reinforcement learning based approach that uses multimodality in sampling of a latent variable that is ranked and optimized with a refinement module. [27] introduces the use of a CVAE for capturing multimodality in the final position of the pedestrians conditioned on the past motion history. Trajectron++ [36] represents agent's trajectories in a graph structured recurrent network for scene compliant trajectory forecasting, taking into account the interaction with a diverse set of agents. LB-EBM [30] learns an energy-based model in the latent space and a policy generator to map the latent vector into a trajectory. The attention based method AgentFormer [47] jointly models the time dimension and social interactions using a sequence representation while preserving each agent's identity. Introvert [38] uses a 3D visual attention mechanism conditioned on the observed trajectory to extract scene and social information from videos. A future displacement distribution is predicted and multiple sequences can be sampled.

A different line of work includes Social GAN [13] which uses adversarial losses [12] for incorporating multimodality in predictions. While such generative approaches do produce diverse trajectories, overall coverage of critical modes cannot be guaranteed and little control is afforded over the properties of predicted trajectories such as direction, number of samples, *etc*. In contrast, our method, Y-net, estimates explicit probability maps which allow easily incorporating spatial constraints for a downstream task.

Multimodality through spatial probability estimates: Another line of work obtains multimodality via estimated probability maps. Activity Forecasting from Kitani *et al*. [21] proposes to use a hidden Markov Decision process for modeling the future paths. However, in contrast to our work, the future predictions in [21] are conditioned on activity labels such as 'approach car', 'depart car', *etc*. More recently, some works have used a grid based scene representation to estimate probabilities for future time steps [25, 26, 9]. Relatedly, some prior works such as [27, 48, 8] propose a goal-conditioned trajectory forecasting method. However, no prior works have proposed factorized modeling of *epistemic* uncertainty or goals and *aleatoric* uncertainty or paths as Y-net uses.

3. Proposed Method

The problem of multimodal trajectory prediction can be formulated formally as follows. Given a RGB scene image $\mathcal{I}$ and past positions of a pedestrian in the scene $\mathcal{I}$ denoted by $\{\mathbf{u}_n\}_{n=1}^{n_p}$ for the past $t_p = n_p/\texttt{FPS}$ seconds sampled at the frame rate $\texttt{FPS}$, the model aims to predict the posi-

tion of the pedestrian for the next t_f seconds in the future, denoted by $\{\mathbf{u}_n^i\}_{n=n_p+1}^{n_p+n_f}$ where $t_f = n_f/\texttt{FPS}$. Since the future is stochastic, multiple predictions for the future trajectories are produced. In this work, we factorize the overall stochasticity into two modes. First are the modes relating to *epistemic* uncertainty *i.e*. multimodality in the final destination for which the module produces K_e goals. Second are the modes relating to the *aleatoric* uncertainty *i.e*. multimodality in the path taken to the destination stemming from uncontrolled randomness given the goal, for which the module produces K_a predictions for each estimated goal. In the short temporal horizon limit, since the overall path length is small, the options for paths to a given goal are limited and similar to each other. This is naturally modeled by constraining $K_a = 1$ and so the total number of paths predicted (K in prior works) is the same as K_e in the short horizon setting. However, for longer temporal horizons, there are several paths to the same goal and hence $K_a > 1$. Next, we describe in detail the working of our model, Y-net and its three sub-networks $\mathbf{U}_e$, $\mathbf{U}_g$ and $\mathbf{U}_t$ followed by details of the non-parametric sampling process (Section 3.2) and loss functions used.

3.1. Y-net Sub-Networks

To effectively use scene information in semantic space (image-like) with trajectory information (coordinates), pixel-wise alignment needs to be created between the different modalities. Some prior works [35] achieve this by encoding the RGB image $\mathcal{I}$ as a hidden state vector extracted from a pretrained CNN network. While this provides the network with scene information, any meaningful spatial signal gets highly conflated when flattened into a vector and pixel alignment is destroyed. This is highlighted in [28] establishing previous state-of-the-art without any scene information, underlining the misuse of image information in prior works. In this work, we adopt a trajectory-on-scene heatmap representation that solves the alignment issue by representing the trajectory in the same space as image $\mathcal{I}$.

3.1.1 Trajectory-on-Scene Heatmap Representation

The RGB image $\mathcal{I}$ is first processed with a semantic segmentation network such as a U-net [34] that produces segmentation map $\mathbf{S}$ of $\mathcal{I}$ comprising of C classes determined according to the affordance provided by the surface to an agent for actions such as walking, standing, running etc. In a parallel branch, the past motion history $\{\mathbf{u}_n\}_{n=1}^{n_p}$ is converted to a trajectory heatmap $\mathbf{H}$ of spatial sizes of $\mathcal{I}$ and n_p channels with one channel for each timestep. Mathematically,

$$\mathbf{H}(n,i,j) = 2\frac{\|(i,j)-\mathbf{u}_n\|}{\max\limits_{(x,y)\in\mathcal{I}}\|(x,y)-\mathbf{u}_n\|}$$

The heatmap trajectory representation is then concatenated with the semantic map $\mathcal{S}$ along the channel dimension producing the trajectory-on-scene heatmap tensor $\mathbf{H}_\mathcal{S}$ a $H \times W \times (C + n_p)$ dimensional input tensor which is passed to the encoder network $\mathbf{U}_e$.

3.1.2 Trajectory-on-Scene Heatmap Encoder $\mathbf{U}_e$

The tensor $\mathbf{H}_\mathcal{S}$ is processed with the encoder $\mathbf{U}_e$ designed as a U-net encoder [34] (Fig. 2). The encoder $\mathbf{U}_e$ consists of M blocks, where the spatial dimensions are reduced from $H \times W$ to $H_M \times W_M$ halving after every block using max pooling (stride 2) and the channel depth is increased sequentially from $C + n_p$ to C_M doubling after a certain number of blocks using convolutional layers with ReLU. The final spatially compact and deep representation after block M along with the $M - 1$ intermediate tensors $\mathbf{H}_m$ with $1 \leq m \leq M$ are passed onto the goal decoder $\mathbf{U}_g$ and the trajectory decoder $\mathbf{U}_t$ as discussed below.

3.1.3 Goal and Waypoint Heatmap Decoder $\mathbf{U}_g$

The processed trajectory-on-scene tensors $\mathbf{H}_m$ at various spatial resolutions are passed onto the goal and waypoint heatmap decoder $\mathbf{U}_g$ which is modeled after the expansion arm in the U-net architecture [34]. A center block consisting of two convolutional layers with ReLU first takes in the final and spatially compact feature tensor $\mathbf{H}_M$. Then the expansion arm spatially doubles the resolution at the beginning of every block using bilinear up-sampling and convolution (together forming Deconvolution [34]). After every Deconvolution, the corresponding intermediate representation $\mathbf{H}_m$ from $\mathbf{U}_e$ is fused using skip connections and the features are processed with two convolutional layers with ReLU non-linearity. Merging intermediate high-resolution feature maps from $\mathbf{U}_e$ is necessary since just using the final feature $\mathbf{H}_M$ would severely limit the final resolution of the goal heatmap, thus missing fine spatial details that are preserved in the intermediate feature maps. The U-net block starts with a deconvolution operation followed by feature merging and two convolutional layers, all of which is repeated sequentially M times to form $\mathbf{U}_g$. The output layer consists of a convolutional layer followed by a pixelwise sigmoid that for each N^w chosen waypoint $\mathbf{u}_{w_i}$ and the goal $\mathbf{u}_{n_p+n_f}$ produces an explicit, non-parametric probability distribution, $P(\mathbf{u}_{w_i})$ and $P(\mathbf{u}_{n_p+n_f})$ after normalization. The overall output shape of $\mathbf{U}_g$ is $H \times W \times (N^w + 1)$. Thus, for each N^w waypoint plus the goal, this submodule predicts a $H \times W$ matrix, where the (i,j)th element of the matrix represents the estimate probability value of the agent being at location (i,j) at the selected timestep.

	S-GAN	CF-VAE	P2TIRL	SimAug	PECNet	LB-EBM	Y-net (Ours)	DESIRE	TNT	PECNet	Y-net (Ours)
	$K = 20$							$K = 5$			
ADE	27.23	12.60	12.58	10.27	9.96	8.87	**7.85**	19.25	12.23	12.79	**11.49**
FDE	41.44	22.30	22.07	19.71	15.88	15.61	**11.85**	34.05	21.16	29.58	**20.23**

Table 1: **Short temporal horizon forecasting results on SDD**: Our method significantly outperforms previous state-of-the-art methods on the Stanford Drone Dataset [33] on both the ADE and FDE metrics for both settings of K, where K represents the number of multimodal samples . Reported errors are in pixels with $t_p = 3.2$ sec, $t_f = 4.8$ sec, $n_p = 8, n_f = 12$ and lower is better.

	ETH	HOTEL	UNIV	ZARA1	ZARA2	AVG
S-GAN	0.81/1.52	0.72/1.61	0.60/1.26	0.34/0.69	0.42/0.84	0.58/1.18
PECNet	0.54/0.87	0.18/0.24	0.35/0.60	0.22/0.39	0.17/0.30	0.29/0.48
LB-EBM	0.30/0.52	0.13/0.20	0.27/0.52	0.20/0.37	0.15/0.29	0.21/0.38
Introvert	0.42/0.70	0.11/0.17	0.20/0.32	0.16/0.27	0.16/0.25	0.21/0.34
Trajectron++	0.39/0.83	0.12/0.21	**0.20**/0.44	**0.15**/0.33	**0.11**/0.25	0.19/0.41
AgentFormer	**0.26**/0.39	0.11/0.14	0.26/0.46	**0.15/0.23**	0.14/0.24	**0.18**/0.29
Y-net (Ours)	0.28/**0.33**	**0.10/0.14**	0.24/**0.41**	0.17/0.27	0.13/**0.22**	**0.18/0.27**

Table 2: **Short term forecasting results on ETH/UCY benchmark**: Our proposed method establishes new state-of-the-art results on both the ADE/FDE metrics on the popular ETH-UCY benchmark with standard short-horizon settings (same as SDD). Reported errors are in meters and lower is better.

3.1.4 Trajectory Heatmap Decoder $\mathbf{U}_t$

$\mathbf{U}_t$ comprises of M decoder blocks which proceed in a similar fashion as $\mathbf{U}_g$ (Section 3.1.3). However, in contrast to $\mathbf{U}_g$, $\mathbf{U}_t$ is conditioned on the sampled goal and waypoints in addition to the scene $\mathbf{S}$ and past trajectory $\{\mathbf{u}_n\}_{n=1}^{n_p}$. The probability distributions estimated by $\mathbf{U}_g$ are used to sample potential goal and waypoints sets. The sampling process is described in Section 3.2 and further details are in Supplementary Section 1. In total, K_e goals are sampled and for each goal K_a waypoint sets are sampled representing K_a paths to the same goal. The obtained coordinate sample sets $\hat{\mathbf{u}}_{n_p+n_f}$ for the goal and $\{\hat{\mathbf{u}}_{w_i}\}_{i=1}^{N^w}$ for the intermediate waypoints are converted to a heatmap representation $\mathbf{H}_g$ similar to the past trajectory as described in Section 3.1.1. Finally, the obtained goal and waypoint conditioning tensor $\mathbf{H}_g$ is downsampled to fit the spatial size of each corresponding block and along with the corresponding $\mathbf{H}_m$ is concatenated to the output of the previous $\mathbf{U}_t$ block and passed into the next block. For each future timestep, it predicts a separate probability distribution, resulting in an output of shape $H \times W \times n_f$, with each channel corresponding to the location distribution in each timestep.

3.2. Non-parametric Distribution Sampling

Given a distribution P of the future frame position as a matrix of probabilities X, we aim to sample a two-dimensional point as our estimate for the position of the agent. This is difficult to achieve reliably in practice since the estimated distribution P is noisy during the initial training stages. Hence, taking a naive `argmax` is not robust. Instead, we propose to use the `softargmax` operation [11],

$$\texttt{softargmax}(X) = \left(\sum_i i \frac{\sum_j e^{X_{ij}}}{\sum_{i,j} e^{X_{ij}}}, \sum_j j \frac{\sum_i e^{X_{ij}}}{\sum_{i,j} e^{X_{ij}}} \right)$$

to approximate the most likely position in a robust fashion.

Further details on the sampling process including Test-Time Sampling Trick and Conditional Waypoint Sampling can be found in Supplementary Section 1.1 and 1.2.

3.3. Loss Function

Since the predictions are explicit probability distributions for each timestep, we impose losses directly on the estimated distribution $\hat{P}$ rather than on the drawn coordinate samples. The ground truth future is represented as a Gaussian heatmap P centered at the observed points with a predetermined variance σ_H. All three networks, $\mathbf{U}_e$, $\mathbf{U}_g$ and $\mathbf{U}_t$ are trained end-to-end jointly using a weighted combination of binary cross entropy losses on the predicted goal, waypoint and trajectory distributions.

$$\mathcal{L}_{\text{goal}} = \text{BCE}(P(\mathbf{u}_{n_p+n_f}), \hat{P}(\mathbf{u}_{n_p+n_f}))$$

$$\mathcal{L}_{\text{waypoint}} = \sum_{i=1}^{N^w} \text{BCE}(P(\mathbf{u}_{w_i}), \hat{P}(\mathbf{u}_{w_i}))$$

$$\mathcal{L}_{\text{trajectory}} = \sum_{i=n_p+1}^{n_p+n_f} \text{BCE}(P(\mathbf{u}_i), \hat{P}(\mathbf{u}_i))$$

$$\mathcal{L} = \mathcal{L}_{\text{goal}} + \lambda_1 \mathcal{L}_{\text{waypoint}} + \lambda_2 \mathcal{L}_{\text{trajectory}}$$

4. Results

We use a total of three datasets to study Y-net's performance – the Stanford Drone Dataset (SDD) [33], the Intersection Drone Dataset (InD) [5], and the ETH [31] / UCY [23] forecasting benchmark.

Stanford Drone Dataset: We benchmark our proposed model on the popular Stanford drone dataset [33] where several recently proposed methods have improved state-of-the-art performance significantly in the past few years [45]. The dataset is comprised of more than $11,000$ unique pedestrians across 20 top-down scenes captured on the Stanford university campus in bird's eye view using a flying

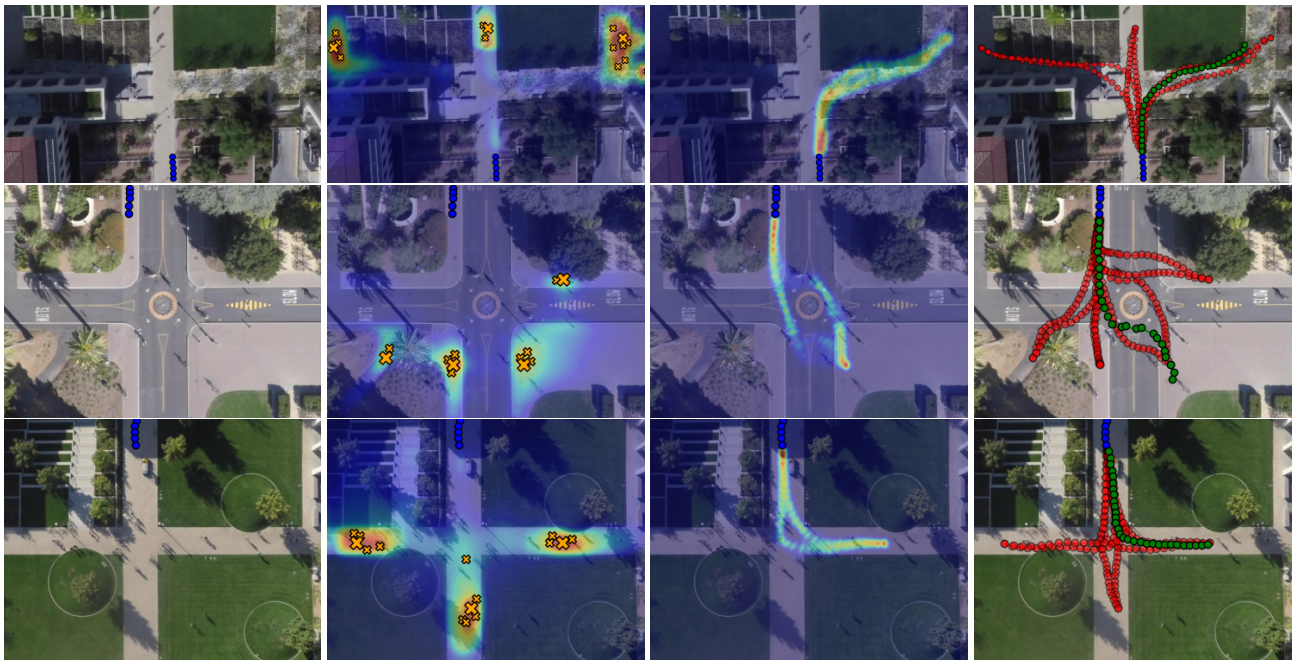

Figure 3: **Qualitative Long Term Trajectory Forecasting Results**: We show various heatmaps and visualizations for three different scenes (rows) in SDD testset. The first column shows the past observed trajectory for last $t_p = 5$ seconds in blue. The second column shows the heatmap from $\mathbf{U}_g$ for $t_f = 30$ seconds in the future (goal multimodality) and some sampled goals from the estimated distribution. The third column shows trajectory heatmaps from $\mathbf{U}_t$ conditioned on a sampled goal from column two (path multimodality). The last column shows the predicted trajectories, green indicating the ground-truth trajectories and red our multimodal predictions.

drone. For short term prediction, we follow the [35, 28] standard setup and dataset split, sampling at $\texttt{FPS} = 2.5$ yielding a input sequence of length $n_p = 8$ and output of length $n_f = 12$, *i.e.* $t_p = 3.2$ sec, $t_f = 4.8$ sec.

In our proposed long term setting, we sample at $\texttt{FPS} = 1$ thus yielding a $n_p = 5$ for $t_f = 5$ seconds in the past and predicting up to one minute into the future. Further, we label the scenes with semantic segmentation maps consisting of $C = 5$ "stuff" classes, namely [7] pavement, terrain, structure, tree and road, depending on the walking affordability of the surface. We split the dataset's scenes in the same way as the short term setup, to evaluate the performance on unseen scenes during training.

Intersection Drone Dataset: We propose to use the Intersection drone dataset [5] for benchmarking long term trajectory forecasting. The dataset comprises over 10 hours of measurements over 4 distinct intersection in an urban environment. The dataset is recorded in $\texttt{FPS} = 25$. We downsample the trajectories to $\texttt{FPS} = 1$ to match our SDD long term setting, filter out non-pedestrian and short trajectories and use a sliding window approach without overlap to split long trajectories. After the preprocessing steps, inD contains 1,396 long term trajectories with $n_p = 5$ and $n_f = 30$. To evaluate performance on unseen environments, we are using location ID 4 only during testing time. The scene is labeled with the same $C = 5$ classes as in SDD. We convert the coordinates from world coordinates (meters) into pixel coordinates using the provided scale factors from the authors and evaluate metrics in pixels.

ETH & UCY datasets: The ETH/UCY benchmarks have been widely used for benchmarking trajectory forecasting models in the short horizon setting in recent years [44]. Forecasting performance has improved by over $\sim 64\%$ on average, within the last two years itself [13]. It comprises of five different scenes all of which report position in world coordinates (in meters). We follow the leave one out validation strategy as outlines in prior work [35, 13, 9]. For all ETH & UCY datasets, since the classes of affordances furnished by the surfaces present is small, we use $C = 2$, identifying each pixel as either belonging to class 'road' or 'not road'. Similar to short term SDD, the frames are sampled at $\texttt{FPS} = 2.5$ predicting $n_f = 12$ frames, $t_f = 4.8$ seconds into the future given the last $n_p = 8$ frames comprising of $t_p = 3.2$ seconds of motion history.

Implementation Details: We train the entire network end to end with Adam optimizer [19] with a learning rate of 1×10^{-4} and batch size of 8. A pre-trained segmentation model is used that is finetuned on the specific dataset. Further details are mentioned in the supplementary materials.

Metrics: We use the established Average Displacement Error (ADE) and Final Displacement Error (FDE) metrics for measuring performance of future predictions. ADE is calculated as the ℓ_2 error between the predicted future and the ground truth averaged over the entire trajectory while FDE

	Stanford Drone Dataset						Intersection Drone Dataset					
	S-GAN	PECNet	R-PECNet	Y-net (Ours)			S-GAN	PECNet	R-PECNet	Y-net (Ours)		
K_a	1	1	1	1	2	5	1	1	1	1	2	5
ADE	155.32	72.22	261.27	47.94	44.94	39.49	38.57	20.25	341.80	14.99	14.02	12.67
FDE	307.88	118.13	750.42	66.71	66.71	66.71	84.61	32.95	1702.64	21.13	21.13	21.13

Table 3: **Long term trajectory forecasting results**: We benchmark performance on our proposed long horizon forecasting setting predicting $t_f = 30$ second into the future given $t_p = 5$ second past motion history. All reported errors are in pixels (lower is better) for $K_e = 20$ with additional results for varying K_a with a fixed K_e.

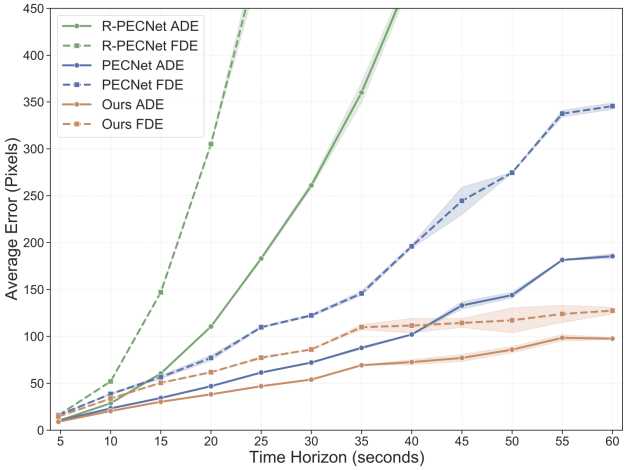

Figure 4: **Benchmarking Performance against Time Horizons**: On prediction horizons up to a minute, we observe a consistently growing difference in ADE between Y-net and PECNet, highlighting the importance of factorized goal and path modeling in long term forecasting.

is the ℓ_2 error between the predicted future and ground truth for the final predicted point [2]. Following prior works [13], in the case of multiple future predictions, the final error is reported as the `min` error over all predicted futures.

ADE and FDE are well suited metrics for deterministic performance evaluation. However, they use samples instead of the predicted distribution for error estimation. Hence, we report the kernel density estimate-based negative log-likelihood (KDE-NLL) metric in the same fashion as [17, 36]. A standardized KDE is used to estimate a probability distribution function for each predicted future timestep, and the NLL of the ground-truth trajectory is calculated using it. Note, that Y-net predicts explicit probability maps. To be consistent with previous literature and enable fair comparison with baselines we also apply the KDE.

4.1. Short Term Forecasting Results

Stanford Drone Results: Table 1 presents results for the SDD in the short term setting. We report results with $K_e = 5$ and 20. Since there is limited aleatoric multimodality in short term settings, we use $K_a = 1$ thus being comparable to prior works using 20 trajectory samples for evaluation. Table 1 shows our proposed model achieving an ADE of 7.85 and FDE of 11.85 at $K_e = 20$ which outperforms the previous state-of-the-art performance of LB-EBM [30] by 13.0% on ADE and 31.7% on FDE. Further, at $K = 5$ it achieves an ADE of 11.49 and FDE of 20.23 outperforming previous state-of-the-art performance of TNT [48].

ETH/UCY Results: We report results on the ETH/UCY benchmark in Table 2. Similar to SDD, we set $K_e = 20, K_a = 1$. We observe that Y-net improves the state-of-the-art performance from AgentFormer [47] in FDE by 7.4% to 0.27 and performs on par in ADE with 0.18.

4.2. Long Term Forecasting Results

To study the effect of *epistemic* and *aleatoric* uncertainty factorization, we propose a long term trajectory forecasting setting with a prediction horizon up to 10 times longer than prior works (up to a minute). To benchmark, we retrain PECNet [28], the previous state-of-the-art method from short term forecasting and Social GAN [13] for each prediction horizon setting separately. We also train a recurrent short term baseline based on PECNet (R-PECNet) where the model is trained only for $t_f = 5$ seconds and is fed its own predictions recurrently for predicting longer horizons.

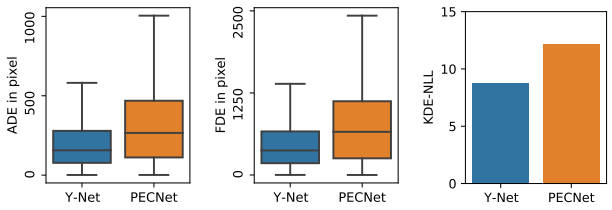

Figure 5: **ADE & FDE boxplot and KDE-NLL**: Left and middle: Boxplots of ADE and FDE, respectively. Right: Results for the KDE-NLL metric. All metrics are estimated for the long term setting on SDD with 100 samples.

Forecasting Results: Table 3 reports the baseline and our results on SDD and InD for a time horizon of $t_f = 30$ seconds in the future given the past $t_p = 5$ second input. All reported results are with $K_e = 20$ for Y-net conditioned on $N^w = 1$ intermediate waypoint at $w_1 = 20$, *i.e.* temporally midway between the observed inputs and the estimated goal. All reported baseline results are at $K = 20$ for fair comparisons with our $K_e = 20, K_a = 1$ setting. On

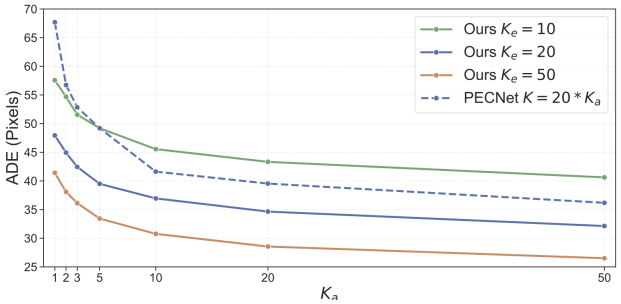

Figure 6: **Benchmarking performance against aleatoric uncertainty** (K_a): Fixing the goal multimodality (K_e) we vary K_a to observe the effect of path multimodality. Also, we benchmark against PECNet by allowing it 20 times more samples for each K_a for a fair compare against the $K_e = 20$ Y-net curve.

SDD, we observe that our proposed model outperforms the state-of-the-art short term baseline on the long horizon setting, achieving an ADE of 47.94 and FDE 66.72 improving upon PECNet's performance by over 50%. Similarly, Y-net outperforms PECNet on InD improving ADE performance from 20.25 to 14.99 and FDE from 32.95 to 21.13.

To gain a more complete assessment of the performance, boxplots of PECNet and Y-net are shown in Fig. 5. These display the median performance and variability of quartiles within the long-term predictions on SDD. Y-net has about half the median error and is much more consistent with less spread.

Further, Y-net achieves a KDE-NLL [17] score of 8.75, significantly better than PECNet's score of 12.15 on the same long-term setting on SDD (Fig. 5). These additional metrics confirm our observations from the ADE and FDE metrics.

Varying Prediction Horizon: We compare Y-net with PECnet and R-PECNet for varying prediction horizons. In Fig. 4 we observe that the difference in performance between Y-net and PECNet grows as prediction horizon increases from 5 to 60 seconds. This shows Y-net's adaptability for long prediction horizons owing to factorized multimodality modeling. We also observe that for PECNet, training a separate model for different time horizons is significantly better than using a short temporal horizon model recurrently (R-PECNet). This motivates our proposal for studying long term forecasting since short term models behave very poorly when applied out of the box recurrently to longer term settings.

Varying $\mathbf{K_a}$: We also report results with $K_a = 2$ *and* 5 for studying the improvement in performance from aleatoric multimodality in Table 3. We observe a consistent improvement in ADE on both datasets, thus indicating the diversity in predicted paths given the same estimated final goal $\mathbf{u}_{n_p+n_f}$. We also report extensive results for varying the path multimodality K_a with a fixed K_e for various choice of K_e and K_a in Figure 6. Additionally for baselining, we benchmark against PECNet [27] evaluated with K_e times more samples than the corresponding Y-net model while varying K_a. We show consistent ADE improvements for various K_e when increasing K_a, indicating effective use of multimodality. Further, even with $K_e = 20$ times more additional samples, PECNet's performance is significantly worse than Y-net at $K_e = 20$ for all K_a highlighting the importance of factorizing goal and path multimodality for diverse and accurate future trajectory modeling.

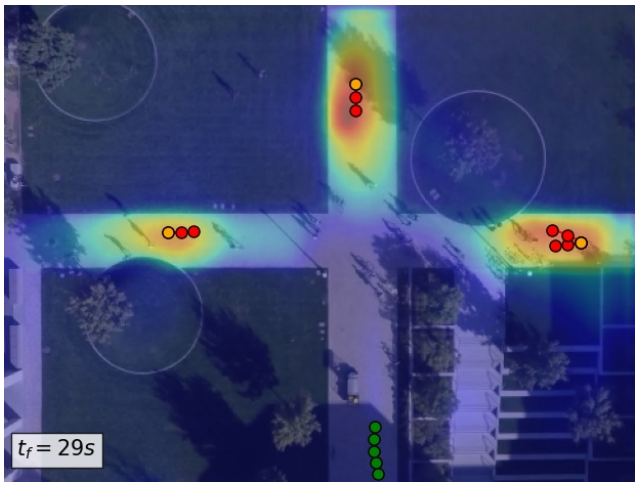

Figure 7: **GIF Visualization**: Demonstrating the goal, waypoint and path multimodality for long term human trajectory prediction (30 seconds). Given the past 5 seconds input history (green) we predict diverse future trajectories (current location in orange, past in red). *Due to restrictions, we can only show a snapshot. Please refer to the supplementary file or ArXiv version for the animation.*

Qualitative Results: We show some qualitative results for long term trajectory prediction ($t_f = 30$) on SDD in Figure 3 and through a GIF temporally in Figure 7. We observe that Y-net predicts diverse scene-complaint trajectories, with both future goals and paths modalities.

5. Conclusion

In summary, we present Y-net, a scene-compliant trajectory forecasting network with factorized goal and path multimodalities. Y-net uses the U-net structure [34] for explicitly modeling probability heatmaps for epistemic and aleatoric uncertainties. Overall, Y-net decrease the error of previous state-of-the-art performance by up to 31.7% on the SDD and by up to 7.4% on ETH/UCY benchmarks in the short term setting. We also propose a new long term trajectory forecasting setting with a prediction horizon of up to a minute for exemplifying the epistemic and aleatoric uncertainty dichotomy. In this setting, we benchmark on the Stanford Drone and Intersection Drone dataset where Y-net exceeds previous state-of-the-art by over 77.1% and 56.0% respectively thereby highlighting the importance of modeling factorized stochasticity.

References

[1] Alexandre Alahi, Kratarth Goel, Vignesh Ramanathan, Alexandre Robicquet, Li Fei-Fei, and Silvio 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

[2] Alexandre Alahi, Vignesh Ramanathan, and Li Fei-Fei. Socially-aware large-scale crowd forecasting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2203–2210, 2014. 7

[3] Maren Bennewitz, Wolfram Burgard, and Sebastian Thrun. Learning motion patterns of persons for mobile service robots. In *Proceedings 2002 IEEE International Conference on Robotics and Automation (Cat. No. 02CH37292)*, volume 4, pages 3601–3606. IEEE, 2002. 2

[4] Apratim Bhattacharyya, Michael Hanselmann, Mario Fritz, Bernt Schiele, and Christoph-Nikolas Straehle. Conditional flow variational autoencoders for structured sequence prediction. *arXiv preprint arXiv:1908.09008*, 2019. 2

[5] Julian Bock, Robert Krajewski, Tobias Moers, Steffen Runde, Lennart Vater, and Lutz Eckstein. The ind dataset: A drone dataset of naturalistic road user trajectories at german intersections. 2019. 2, 5, 6

[6] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020. 1

[7] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Cocostuff: Thing and stuff classes in context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1209–1218, 2018. 6

[8] Zhe Cao, Hang Gao, Karttikeya Mangalam, Qizhi Cai, Minh Vo, and Jitendra Malik. Long-term human motion prediction with scene context. In *Eur. Conf. Comput. Vis.* 2020. 2, 3

[9] Nachiket Deo and Mohan M Trivedi. Trajectory forecasts in unknown environments conditioned on grid-based plans. *arXiv preprint arXiv:2001.00735*, 2020. 3, 6

[10] Armen Der Kiureghian and Ove Ditlevsen. Aleatory or epistemic? does it matter? *Structural safety*, 31(2):105–112, 2009. 2

[11] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 6.2. 2.3 softmax units for multinoulli output distributions. In *Deep Learning.*, pages 180–184. MIT Press, 2016. 5

[12] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014. 3

[13] Agrim Gupta, Justin Johnson, Li Fei-Fei, Silvio Savarese, and Alexandre Alahi. Social gan: Socially acceptable trajectories with generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2255–2264, 2018. 2, 3, 6, 7

[14] Dirk Helbing. *Stochastische Methoden, nichtlineare Dynamik und quantitative Modelle sozialer Prozesse*. Shaker, 1993. 2

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

[16] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 1

[17] Boris Ivanovic and Marco Pavone. The trajectron: Probabilistic multi-agent trajectory modeling with dynamic spatiotemporal graphs. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2375–2384, 2019. 7, 8

[18] Daniel Kahneman. *Thinking, fast and slow*. Macmillan, 2011. 2

[19] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[20] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 3

[21] Kris M Kitani, Brian D Ziebart, James Andrew Bagnell, and Martial Hebert. Activity forecasting. In *European Conference on Computer Vision*, pages 201–214. Springer, 2012. 3

[22] Namhoon Lee, Wongun Choi, Paul Vernaza, Christopher B Choy, Philip HS Torr, and Manmohan Chandraker. Desire: Distant future prediction in dynamic scenes with interacting agents. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 336–345, 2017. 2, 3

[23] Alon Lerner, Yiorgos Chrysanthou, and Dani Lischinski. Crowds by example. In *Computer graphics forum*, volume 26, pages 655–664. Wiley Online Library, 2007. 2, 5

[24] Jiachen Li, Hengbo Ma, and Masayoshi Tomizuka. Conditional generative neural system for probabilistic trajectory prediction. *arXiv preprint arXiv:1905.01631*, 2019. 2

[25] Junwei Liang, Lu Jiang, and Alexander Hauptmann. Simaug: Learning robust representations from 3d simulation for pedestrian trajectory prediction in unseen cameras. 2020. 3

[26] Junwei Liang, Lu Jiang, Kevin Murphy, Ting Yu, and Alexander Hauptmann. The garden of forking paths: Towards multi-future trajectory prediction, 2020. 2, 3

[27] Karttikeya Mangalam, Ehsan Adeli, Kuan-Hui Lee, Adrien Gaidon, and Juan Carlos Niebles. Disentangling human dynamics for pedestrian locomotion forecasting with noisy supervision, 2020. 2, 3, 8

[28] Karttikeya Mangalam, Harshayu Girase, Shreyas Agarwal, Kuan-Hui Lee, Ehsan Adeli, Jitendra Malik, and Adrien Gaidon. It is not the journey but the destination: Endpoint conditioned trajectory prediction. *arXiv preprint arXiv:2004.02025*, 2020. 2, 4, 6, 7

[29] A. A. Markov. An example of statistical investigation of the text eugene onegin concerning the connection of samples in chains. *Lecture at the physical-mathematical faculty, Royal Academy of Sciences, St. Petersburg*, 23 January 1913. 1

[30] Bo Pang, Tianyang Zhao, Xu Xie, and Ying Nian Wu. Trajectory prediction with latent belief energy-based model. In

Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 11814–11824, 2021. 3, 7

[31] Stefano Pellegrini, Andreas Ess, Konrad Schindler, and Luc Van Gool. You'll never walk alone: Modeling social behavior for multi-target tracking. In *2009 IEEE 12th International Conference on Computer Vision*, pages 261–268. IEEE, 2009. 2, 5

[32] A Robicquet, A Sadeghian, A Alahi, and S Savarese. Learning social etiquette: Human trajectory prediction in crowded scenes. In *European Conference on Computer Vision (ECCV)*. 2

[33] Alexandre Robicquet, Amir Sadeghian, Alexandre Alahi, and Silvio Savarese. Learning social etiquette: Human trajectory understanding in crowded scenes. In *European conference on computer vision*, pages 549–565. Springer, 2016. 5

[34] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241, Cham, 2015. Springer International Publishing. 3, 4, 8

[35] Amir Sadeghian, Vineet Kosaraju, Ali Sadeghian, Noriaki Hirose, Hamid Rezatofighi, and Silvio Savarese. Sophie: An attentive gan for predicting paths compliant to social and physical constraints. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1349–1358, 2019. 2, 4, 6

[36] Tim Salzmann, Boris Ivanovic, Punarjay Chakravarty, and Marco Pavone. Trajectron++: Dynamically-feasible trajectory forecasting with heterogeneous data. *arXiv preprint arXiv:2001.03093*, 2020. 2, 3, 7

[37] Rajat Sen, Hsiang-Fu Yu, and Inderjit S Dhillon. Think globally, act locally: A deep neural network approach to high-dimensional time series forecasting. In *Advances in Neural Information Processing Systems*, pages 4837–4846, 2019. 1

[38] Nasim Shafiee, Taskin Padir, and Ehsan Elhamifar. Introvert: Human trajectory prediction via conditional 3d attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16815–16825, 2021. 3

[39] Sebastian Thrun. Probabilistic robotics. *Communications of the ACM*, 45(3):52–57, 2002. 2

[40] Michael Tomasello, Malinda Carpenter, Josep Call, Tanya Behne, and Henrike Moll. Understanding and sharing intentions: The origins of cultural cognition. *Behavioral and brain sciences*, 28(5):675–691, 2005. 2

[41] Pavan Vasishta, Dominique Vaufreydaz, and Anne Spalanzani. Natural vision based method for predicting pedestrian behaviour in urban environments. In *2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC)*, pages 1–6. IEEE, 2017. 1

[42] Anirudh Vemula, Katharina Muelling, and Jean Oh. Social attention: Modeling attention in human crowds, 2018. 2

[43] Peter Whittle. *Hypothesis testing in time series analysis*, volume 4. Almqvist & Wiksells boktr., 1951. 1

[44] Paper with code. Eth/ucy trajectory prediction benchmark, https://paperswithcode.com/sota/trajectory-prediction-on-ethucy, 2020. 6

[45] Paper with code. Stanford drone trajectory prediction benchmark, https://paperswithcode.com/sota/trajectory-prediction-on-stanford-drone, 2020. 5

[46] Takuma Yagi, Karttikeya Mangalam, Ryo Yonetani, and Yoichi Sato. Future person localization in first-person videos, 2018. 2

[47] Ye Yuan, Xinshuo Weng, Yanglan Ou, and Kris Kitani. Agentformer: Agent-aware transformers for socio-temporal multi-agent forecasting. *arXiv preprint arXiv:2103.14023*, 2021. 3, 7

[48] Hang Zhao, Jiyang Gao, Tian Lan, Chen Sun, Benjamin Sapp, Balakrishnan Varadarajan, Yue Shen, Yi Shen, Yuning Chai, Cordelia Schmid, et al. Tnt: Target-driven trajectory prediction. *arXiv preprint arXiv:2008.08294*, 2020. 3, 7