CovarianceNet: Conditional Generative Model for Correct Covariance Prediction in Human Motion Prediction

Aleksey Postnikov[1,2] Aleksander Gamayunov[1,2] Gonzalo Ferrer[2]

***Abstract*— The correct characterization of uncertainty when predicting human motion is equally important as the accuracy of this prediction. We present a new method to correctly predict the uncertainty associated with the predicted distribution of future trajectories. Our approach, CovariaceNet, is based on a Conditional Generative Model with Gaussian latent variables in order to predict the parameters of a bi-variate Gaussian distribution. The combination of CovarianceNet with a motion prediction model results in a hybrid approach that outputs a uni-modal distribution. We will show how some state of the art methods in motion prediction become overconfident when predicting uncertainty, according to our proposed metric and validated in the ETH data-set [1]. CovarianceNet correctly predicts uncertainty, which makes our method suitable for applications that use predicted distributions, e.g., planning or decision making.**

I. Introduction

Human Motion Prediction, during the last years, has received the attention of the research community from different fields: intelligent vehicles, pattern recognition, graphics, robotics, etc. The motivation to understand and predict human motion is immense and it has a deep impact in related topics, such as, decision making, path planning, autonomous navigation, surveillance, tracking, scene understanding, anomaly detection, etc.

The problem of forecasting where pedestrians will be in the near future is, however, ill-posed by nature: human beings tend to be unpredictable on their decisions and motion is neither exempt of it. These random nature in motion brings an open challenge to prediction algorithms, where algorithms are desired to be accurate and correctly grasp the uncertainty associated with their predictions.

To this end, multiple benchmarks have been created and released [1]–[3], providing common grounds to test and evaluate. Most modern motion prediction algorithms focus on accurate prediction of agent position errors on these benchmarks. Nonetheless, the precision due to this inherent uncertainty is equally important, and this paper is an effort to research on this direction. Prediction algorithms should address this issue as well: it provides a high degree of interpretability by estimating the associated uncertainty and it might be of some use for consequent algorithms making use of prediction information, e.g., planning. For example, Fig. 1 shows two predictions, of similar predicted error, but different uncertainty estimation.

[1] The authors are with the Sberbank Robotics Laboratory, Moscow, Russia. {postnikov.a.l,gamayunov.a.r}@sberbank.ru.

[2]Skolkovo Institute of Science and Technology, Moscow, Russia. g.ferrer@skoltech.ru.

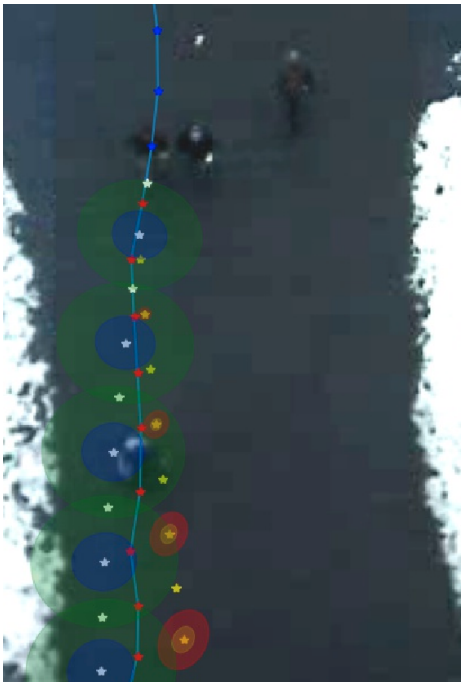

Fig. 1: Comparison of two predicted trajectories in the ETH dataset [1]. CovarianceNet in cyan stars, blue and green ellipses, - mean, 1 sigma, 3 sigma, respectively. Trajectron++ [4] in yellow stars, yellow and red ellipses, red ellipses - mean, 1 sigma, 3 sigma, respectively. The environment is challenging, since there is a pedestrian in the way.

In this paper, we focus on predicting uncertainty and human motion prediction. We propose a hybrid approach shown on Fig. 2, consisting of deep model for Goals Prediction and model-based trajectory prediction, complemented by a modern neural-net approach to predict motion as a uni-modal Gaussian distribution. Several works provide a multi-modal distribution for motion prediction such as [4], achieving excellent results in benchmarks. However, combining a mixture distribution into robot planning approaches (for instance) requires careful considerations. Sampling based techniques require extra attention when considering multi-modality since low-probability prediction outcomes might result in dangerous outcomes, when evaluated under a risk perspective [5].

The contributions of this work are:

- A Conditional Generative Model to correctly predict covariances;
- Hybrid approach combining a model-based motion pre-

diction from the Social Force Model(SFM) [6] and our implementation of a learning-based Goal Prediction;
- Metric to measure the correctness of the uncertainty prediction by counting and averaging the number of times that the prediction lies in the iso-contours of a bi-variate Gaussian.

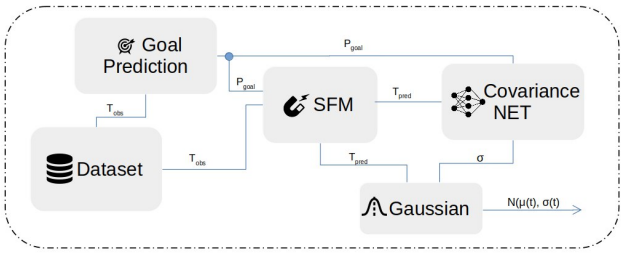

Fig. 2: Our Hybrid approach for Human Motion Prediction and Uncertainty Estimation. We combine customized deep goal prediction model [7] with SFM [6], for mean pedestrian pose prediction and Covariance Net uses Conditional Variational Autoencoder [8] deep generative model for uncertainty prediction.

II. Related Work

Motion prediction has been studied by the robotics community mainly motivated by its direct relation to robot navigation in social environments, embedded at its deepest core. Examples of robot navigation include Human-aware approaches [9], [10], considering prediction into planning [11]–[14] or the effect of planning to prediction [15]–[17].

Other fields have studied human motion prediction, where multiple previous works study the problem of how to forecast pedestrian trajectories and predict their future behaviors. Broadly, human motion prediction can be divided in two classes: model-based that creates an empirical model of these transition functions such as the SFM [6] and its variants [18], [19] or models proposed from the graphics community [20], [21]. And Learning-based approaches [4], [7], [22]–[25] that are becoming the dominant paradigm in human motion prediction, as well as in other topics due to its unrivaled results. Both of these approaches have the same input and output data and predict the next state of the pedestrians in dt time.

Helbing and Molnár [6] have proposed a model based approach for modelling pedestrians' behaviour. Authors have shown that pedestrian motion can be described as a sum of social forces that depend on agents destination point, other pedestrians, static object (borders of buildings, walls, streets, obstacles).

Alahi et al. [22] propose a learning-based method with social pooling, where the Long Short-term Memory(LSTM) [26] states of neighboring agents were pooled based on their locations in a 2-D grid to form social tensor.

Messaoud et al. [23] apply a multihead attention mechanism to the social tensor to directly relate distant vehicles and extract a context representation.

Recently, generative approaches have emerged as state-of-the-art trajectory forecasting methods due to recent advancements in deep generative models. They have caused a shift from focusing on predicting the single best trajectory to producing a distribution of potential future trajectories. Most works in this category use a deep recurrent backbone architecture with a latent variable model, such as a Conditional Variational Autoencoder (CVAE) [8] or a Generative Adversarial Network (GAN) [27].

Trajectron++ [4], [28] - a multi-agent behavior prediction model that accounts for the dynamics of the agents, produces predictions possibly conditioned on potential future robot trajectories which can effectively use heterogeneous data about the surrounding environment.

Jein et al. [29] proposes a discrete residual flow to recursively updates the predicted distribution over a pedestrian's future position in the form of occupancy maps.

Mahgalam et al. [30] propose to address human trajectory prediction by modeling intermediate stochastic goals proposing a socially compliant, endpoint conditioned variational auto-encoder with a novel self-attention based social pooling layer.

In our previous work [31], we propose a new method for motion prediction - HSFM-ΣNN that combines two different approaches: a feed-forward network whose layers are model-based transition functions using the Headed Social Force Model(HSFM) and a Neural Network for covariance prediction. CovarianceNet is the next step in this work, where we have improved the approach substantially.

Gal et al., [32] proposed a simple yet effective method for probabilistic interpretation of dropout which allows to obtain model uncertainty out of existing deep learning models.

Guo et al. [33] introduced a temperature scaling, confidence calibration method, that can effectively correct the miscalibration in modern deep neural networks.

III. Method

A. Problem formulation

The position of a generic agent i at time t is represented by $\boldsymbol{x}_i^t = (x_i, y_i)_t$, where $(x_i, y_i)_t$ are the coordinates of agents in the global reference system at the instant of time t. The agent's trajectory is defined as $X_i^{1:T} = \{\boldsymbol{x}_i^1, ..., \boldsymbol{x}_i^T\}$ from timestamp 1 to T. We aim to generate plausible trajectory distributions for a time-varying number of interacting agents. Every trajectory is split into observed and future: given certain number T_{obs} of observed positions, we seek a distribution over all agents' future states for the next T_{pred} time steps which is denoted as $p(X_i^{T_{obs}+1:T_{pred}} | X_i^{1:T_{obs}})$.

Our approach is visualized in Fig. 2. The first block is the model-based SFM [6], a method for motion prediction the mean positions of the agents' future states.

$$sfm_{1:N}^{T_{obs}+1:T_{pred}} = SFM(X_{1:N}^{1:T_{obs}}, \hat{X}_{goal_{1:N}}^{T}) \quad (1)$$

At a higher level the SFM needs to infer future possible destinations of pedestrians [34]. In particular, we have customized the learning-based approach by [7] and goals are

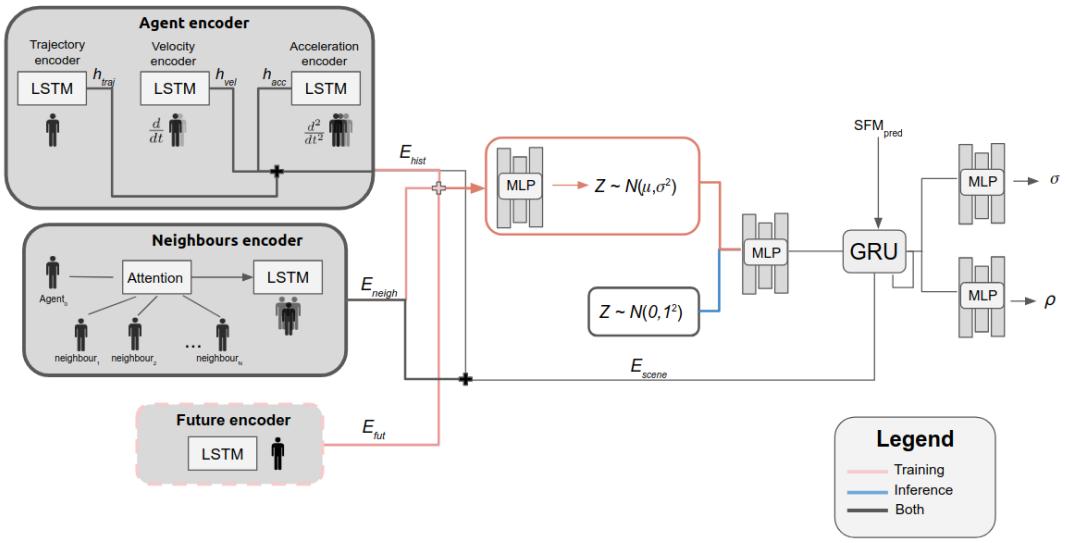

Fig. 3: CovarianceNET architecture. CovarianceNET is a part of our proposed hybrid approach for Human Motion Prediction and Uncertainty Estimation. The input consists of the spatial coordinates over the past T_{obs} seconds of all agents in the scene. We use LSTM-based encoder for the agents' history. Neighbours' impact on the predicted motion is encoded by adding an attention module. We use the CVAE latent variable framework for diverse but realistic uncertainty prediction [8].

predicted as:

$$\hat{X}^{T}_{goal_{1:N}} = \phi(X^{1:T_{obs}}_{1:N}), \tag{2}$$

where ϕ - deep multihead-attention based model that trained to predict the position of pedestrian at timestamp T_{pred}, $X^{1:T_{obs}}_{1:N}$ - the trajectory (scene history) which contains the states of all N agents at timestamp t, $X^{t}_{1:N} = \{X^{t}_{1}, .., X^{t}_{N}\}$, $P_{goal_{1:N}}$ - the position of the pedestrian at timestamp T_{pred}, sfm^{t}_{n} - predicted position of n agent at timestamp t. Here we emphasize that in this work our main goal is to show correct covariance prediction, thus we are not showing implementation details of Goal Prediction model and SFM.

The CovarianceNet network is one of the sub-blocks of our approach, depicted in Fig. 2 and it is shown in detail in Fig. 3. CovarianceNet is designed to complement the predictions by the SFM with accurate x, y variances and its correlation coefficient, such that the bi-variate Gaussian constructed from the mean and covariance matrix accurately assesses the potential trajectory distribution of pedestrians, and statistically ensure the properties of a bi-variate Gaussian distribution.

B. Agent encoder

For each pedestrian in the scene, the current position and all observed previous positions are known. In our work (see Fig. 3), we are using the *Agent* encoder to encode information about agent position and the *Neighbours* encoder for encoding information about the influence of the agent and all his neighbors to each other. Knowing the history of the agent's movement, one can calculate velocities and accelerations as a additional source of data.

The observed trajectory of each agent $X^{1:T_{obs}}_{1:N}$ is encoded using an LSTM encoder

$$h^{t}_{traj_i} = \mathit{LSTM}(X^{1:T_{obs}}_{1:N}, h^{t-1}_{traj_i}, W_{t_{enc}}) \tag{3}$$

Here, $h^{t}_{traj_i}$ is the hidden state vector of the i^{th} agent at time t. All LSTM encoders share the same weights $W_{t_{enc}}$.

Additionally, we use velocities and accelerations encoded in the same fashion as in (3) and sum them into a single encoded vector:

$$h^{t}_{acc_i} = \mathit{LSTM}(\frac{\partial}{\partial t} X^{1:T_{obs}}_{1:N}, h^{t-1}_{i_{vel}}, W_{v_{enc}}) \tag{4}$$

$$h^{t}_{vel_i} = \mathit{LSTM}(\frac{\partial^2}{\partial t^2} X^{1:T_{obs}}_{1:N}, h^{t-1}_{i_{vel}}, h^{t-1}_{i_{acc}}, W_{a_{enc}}) \tag{5}$$

$$E^{t}_{hist_i} = h^{t}_{traj_i} + h^{t}_{vel_i} + h^{t}_{acc_i}; \tag{6}$$

where $h^{t}_{i_{traj}}$ - encoded trajectory, $h^{t}_{i_{vel}}$ - encoded velocities, $h^{t}_{i_{acc}}$ - encoded accelerations and E^{t}_{hist} - full encoded agent's history.

During the training phase, as it is shown in Fig.3, to produce a latent distribution additionally leveraged future trajectory encoding $E^{t}_{fut_i}$ for time indexes $T_{obs} + 1...T_{pred}$, in the same manner as in Eq.(3).

C. Neighbours encoding

To model neighboring agents' influence on the modeled agent, all agents' neighbours in the scene are processed with the additive attention module [35] and are encoded by the LSTM cell to produce a single neighboring agents' influence vector $E^{t}_{neigh_i}$.

$$C^{t}_{i} = \mathit{Attention}(X^{1:T_{obs}}_{neigh_i}, X^{1:T_{obs}}_{i}) \tag{7}$$

$$E^{t}_{neigh_i} = \mathit{LSTM}(C^{t}_{i}, E^{t-1}_{neigh_i}, W_{neigh}) \tag{8}$$

$$E^{t}_{scene_i} = E^{t}_{neigh_i} + E^{t}_{hist_i} \tag{9}$$

where C_i^t - context vectors, $E_{neigh_i}^t$ - encoded neighbours influence on the modeled agent, $E_{scene_i}^t$ is a full encoded scene history.

D. Conditional Variational Autoencoder

The network backbone of our approach (Fig. 3) is realized as a version of the Conditional Variational Autoencoder [8] (CVAE).

The CVAE architecture can be divided in two parts: the prior and the posterior networks. Both prior and posterior distributions are assumed to be Normal distributions. The parameters of the prior are computed by the prior network which only takes the encoded history as input. The posterior parameters are determined from both the encoded history and the target trajectory. The prior distribution is a Normal distribution, denoted as $p_{\phi_i}(z|X_i^{1:T_{obs}}) = \mathcal{N}(\mu_{prior_i}(x), \sigma^2_{prior_i}(x))$.

$$p_{\phi_i}(z|X_i^{1:T_{obs}}) = \mathcal{N}(\mu_{prior_i}, \sigma^2_{prior_i}) \tag{10}$$

$$\mu_{prior_i} = \textit{MLP}(E_{scene_i}) \tag{11}$$

$$\sigma_{prior_i} = \textit{MLP}(E_{scene_i}) \tag{12}$$

During training, the latent variable z will be sampled from the posterior distribution. Specifically, it takes both the encoded past and the encoded future trajectories information X_i passed through an MLP to obtain a latent mean μ_{latent} and a latent sigma σ_{latent} to output a latent distribution $q_{\phi_i}(z|X_i^{T_{obs}+1:T_{pred}}, X_i^{1:T_{obs}})$.

$$q_{\phi_i}(z|X_i^{T_{obs}+1:T_{pred}}, X_i^{1:T_{obs}}) = \mathcal{N}(\mu_{latent_i}, \sigma^2_{latent_i}) \tag{13}$$

$$\mu_{latent_i} = \textit{MLP}(E_{scene_i}, E_{fut_i}) \tag{14}$$

$$\sigma_{latenti} = \textit{MLP}(E_{scene_i}, E_{fut_i}) \tag{15}$$

In our work, we use two fully connected layers with dropout and sigmoid activation function between them as base multi layer perceptron (MLP).

E. Decoder

The decoder models the probability of a target trajectory $X^{T_{obs}+1..T_{pred}}$ given the latent variable z sampled from the latent distribution, encoded node history $E_{scene_i}^t$ and SFM trajectory predictions $\textit{sfm}_i^t$.

We make an assumption that the target distribution $p(\boldsymbol{x}^t|X^{1:T_{obs}})$ is a bi-variate Gaussian:

$$p_i(\boldsymbol{x}_i^t|X_i^{1:T_{obs}}) = N(\mu_i^t, \Sigma_i^t) \tag{16}$$

$$\mu_i^t = \begin{bmatrix} \mu_{x_i}^t \\ \mu_{y_i}^t \end{bmatrix}; \Sigma_i^t = \begin{bmatrix} {\sigma_{x_i}^t}^2 & \rho\sigma_{x_i}^t\sigma_{y_i}^t \\ \rho\sigma_{x_i}^t\sigma_{y_i}^t & {\sigma_{y_i}^t}^2 \end{bmatrix}. \tag{17}$$

For each timestamp of prediction, we use encoded information about current agent state, passed through the MLP as an input to the GRU layer with E_{scene} as the initial state of the GRU cell, witch recurrently outputs the new hidden states for each agent.

$$h_{gru_i}^t = \textit{GRU}([\textit{sfm}_i^{t-1}, z], h_{gru_i}^{t-1}) \tag{18}$$

GRU hidden states are then used to predict the parameters of a bi-variate Gaussian distribution $\mathcal{N}(\mu, \sigma, \rho)$ standard deviations σ and correlation coefficient ρ, while the mean values μ are taken from the SFM prediction $\mu_i = \textit{sfm}_i^t$.

$$\rho_i^t = \textit{MLP}(h_{gru_i}^t) \tag{19}$$

$$\sigma_i^t = \textit{MLP}(h_{gru_i}^t) \tag{20}$$

The entire model is trained by maximising the sequential evidence lower-bound (ELBO):

$$L_{nll}(W) = -\sum_{i=1}^{N}\sum_{t=T_{obs}+1}^{T_{pred}} log(p_i^t(\boldsymbol{x}_i^t|\mu_i^t, \sigma_i^t, \rho_i^t)) \tag{21}$$

$$L_{kl}^n = \sum_{i=1}^{N} D_{kl}(q_{\phi_i}(z|X_i^{T_{obs}+1:T_{pred}}, X_i^{1:T_{obs}})||N(0,1)) \tag{22}$$

$$\textit{Loss} = \alpha L_{nll}^n(W) + L_{kl}^n \tag{23}$$

It is important to mention that in our case, when we utilize Goal Predictor and SFM [6] as model for mean poses prediction we do not optimize μ_i^t parameter from (21).

IV. Evaluation

In our work, we use the combination of goal predictor and SFM [6] modules as a base predictor for mean future poses of agents and combine it with the CovarianceNet in order to predict accurate uncertainties. We compare results of our method with other three popular approaches for evaluating uncertainty on publicly-available ETH [1] pedestrian dataset. It consists of real world human trajectories with rich multi-human interaction scenarios. This dataset is a standard benchmark in the field as it contain challenging behaviors such as couples walking together, groups crossing each other, and groups forming and dispersing. Leave-one-out strategy was used for evaluation, where the model is trained on four datasets and evaluated on the held-out fifth. An observation length of 8 timesteps (3.2s) and a prediction horizon of 12 timesteps (4.8s) is used for evaluation. In this section we will compare results of our method with the following methods:

A. Evaluated Methods

1) Covariance Forward-Propagation (FP): In our work we have used SFM for prediction forces acting on all agents in scene and Human locomotion model which integrates predicted forces to the state variables of a pedestrians 2D poses at timestamp $t+1$, in the following generalized way:

$$x_{t+1} = T(x_t) \tag{24}$$

The transition function $T()$ defined by the SFM (24) is a non-linear differentiable function (by construction). The simplest method for covariance estimation is using the first-order Taylor expansion:

$$x_{t+1} = T(\mu_t) + G_t(x_t - \mu_t) \tag{25}$$

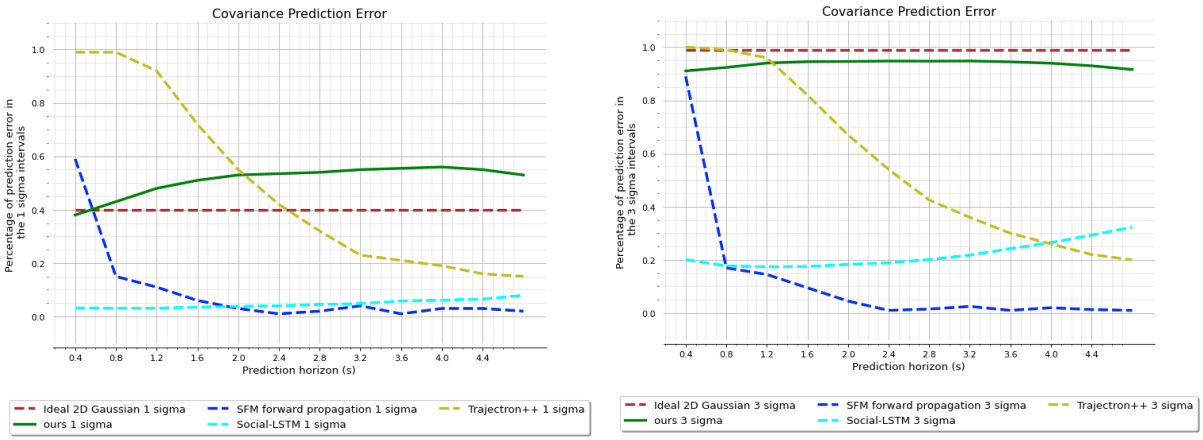

Fig. 4: Evaluation of *PPEI*. *Left:* results for 1σ. *Right:* results for 3σ.

where μ_t is the current state estimate and G_t is the Jacobian matrix $\partial T/\partial x$ evaluated at μ_t. From here, we apply *Covariance Propagation* of a Gaussian random variable ($x_t \sim \mathcal{N}(\mu_t, \Sigma_t)$) over a linear function:

$$x_{t+1} \sim \mathcal{N}\big(T(\mu_t), G_t \cdot \Sigma_t \cdot G_t^\top\big) \tag{26}$$

2) Social-LSTM [22]: One of the most influential works in direction of Human Motion Distribution Prediction based on Neural Networks, where each agent is modeled with an LSTM and nearby agents' hidden states are pooled at each timestep using a proposed social pooling operation.

3) Trajectron++ [4]: One of the latest works in this direction is named *Trajectron++* [4] has four configurations of predictions, of which we used Most Likely, which gives the best results of ADE and FDE. Trajectron++ calculates the result of predictions as a Gaussian Mixture Model (GMM) witch contains 25 Gaussian distributions.

B. Motion Prediction Evaluation

We use Euclidean distance errors, Average Displacement Error (ADE) and Final Displacement Error (FDE) to evaluate the accuracy of then mentioned approaches in a motion prediction task. Metrics are formulated as:

$$ADE^t = \frac{\sum_{j=1}^{N} ||\boldsymbol{x}_j^t - \mu_j^t||_2}{N} \tag{27}$$

$$FDE = \frac{\sum_{j=1}^{N} ||\boldsymbol{x}_j^{T_{pred}} - \mu_j^{T_{pred}}||_2}{N} \tag{28}$$

where N - number of processed pedestrians, $\boldsymbol{x}_j^t$ - ground truth position of j^{th} pedestrian at timestamp t, T_{pred} - prediction horizon, μ - predicted mean position.

The results in Table I shows that modern deep approaches are superior to model-based approaches in terms of ADE and FDE. The Trajectron's++ [4] approach takes into consideration a multimodal distribution, which results in the best Euclidean errors. Still, the contributions of this paper are on the covariance prediction, and combination of SFM and Goal Predictor could be substituted by any modern prediction method.

C. Covariance Estimation Evaluations

We propose to evaluate the accuracy of the covariance prediction methods by the following metric, the Part of Prediction Errors Inside:

$$PPEI_\alpha = E\{\mathbf{1}(||\boldsymbol{x} - \mu||_\Sigma < \alpha)\}, \quad \alpha = \{1, 3\}. \tag{29}$$

where $\mathbf{1}()$ is the indicator function and we simply average the number of times that our prediction lies inside the 1σ and 3σ ellipsoids from the ground truth position by using the Mahalanobis distance.

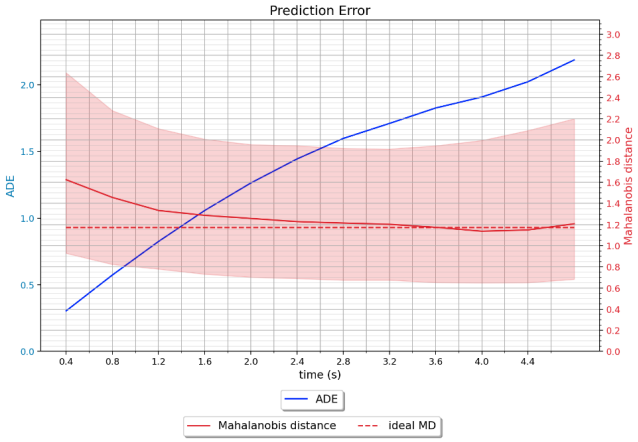

Fig. 5: Mahalanobis error (in red) distances and ADE(in blue) for our proposed method. Mahalanobis solid line is median values, and colored intervals are .25 and .75 percentiles.

Figure 4 shows the $PPEI_1$ and $PPEI_3$ for each of the methods described above. Our proposed method achieves a superior performance (Fig 4) compared to the previously published methods in terms of performance of predicted uncertainties.

It can be seen in Fig. 4 and in Table I that modern deep approaches output accurate ADE, FDE but become overconfident when predicting covariances of distributions, while proposed method with sequentially trained Goal Predictor

TABLE I: Comparison of our method against previously published methods on the ETH dataset [1]. Both ADE and FDE are reported in meters and presented here for reference(*is not valid since ground truth last pedestrian pose is used as input). Our main contribution is accurately predicted uncertainties, which can be measured as part of errors inside 1σ 3σ interval and its deviation from the ideal 2D Gaussian.

Method	ADE	FDE	% errors inside 1σ (Δ from expected)	% errors inside 3σ (Δ from expected)	median Mahalanobis Distance
Social-LSTM [22]	1.09	2.35	4.7±11(35)	22 ±11(76.7)	3.4±5.66
Trajectron++ (Distribution) [4]	**0.71**	**1.66**	**50**±32(**10.3**)	56±30(42.4)	–
SFM + FP	0.98*	0.2*	8.2±15(31.5)	11.6±23(87.2)	–
CovarianceNet	1.39	2.18	51.2±**0.3**(11.5)	**93.7±0.01(5)**	1.21±2.16

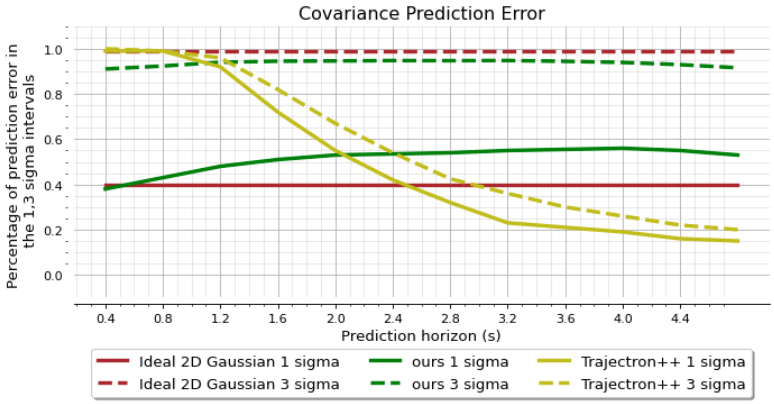

Fig. 6: *PPEI* Evaluation of CovarianceNet method and Trajectron++(Distribution configuration) qualities of predicted uncertainties peaking Trajectron++ Gaussian with the highest probability (out of GMM 25 Gaussians component).

and CovarianceNet models are capable to predict statistically correct covariances. Despite the fact that Trajectron's++ average $PPEI_1$ over all horizon time is close to ideal, it clearly can be observed a high deviation, ranging from a high $PPEI_1$ (under confident) at initial horizon times to low values at the final horizon time (over confident), which is shown with high variance, $PPEI_1$=50 ±32 at Table I, while our proposed method produces more consistent predictions $PPEI_1$=51.2 ±0.3.

The forward propagation (FP) method collapses and provides poor results due to vanishing gradients over multiple iterations. Trajectron shows a decreasing value of *PPEI*, becoming clearly overconfident. Social-LSTM also provides a overconfident covariance estimations, with mean 1σ PPEI=4.7% and 3σ PPEI=22%.

In Fig. 5 we show the median, 25 and 75 percentiles of the Mahalanobis distances for the predicted distributions of CovarianceNet and the Average Displacement Error(ADE) over prediction horizon time for predicted mean trajectories, with dashed red line we show median Mahalanobis Distance (MD) for ideal bi-variate Gaussian. Our projected median MD for the entire prediction horizon is 1.21, which is close to the bi-variate Normal distribution, with a median MD of 1.17.

We use Trajectron++ [4] in Distribution configuration to estimate probability distribution. We used only one Gaussian distribution from GMM for the metric at each prediction step, which had the highest probability. An example of such a prediction and that interpretations can be seen in Fig. 1. This evaluation interpretations have similar to previous interpretations results, with overconfident distribution for large prediction horizons and under confident for smaller time horizons.

Fig 6 shows that individual Gaussians, composing GMM, is far from ideal, which potentially can lead to unexpected results when a person is outside the probabilistic estimate of the pedestrian position.

V. Conclusions

We have presented a new hybrid method, CovarianceNet, that combines model-based Human Motion Prediction with a neural network approach for covariance prediction. Our approach brings an efficient calculation of motion prediction, by using the SFM recursively and calculates the uncertainty associated with this prediction by using a conditional deep generative model CVAE with Gaussian latent variables.

We have evaluated our results in the ETH dataset and compared with state-of-the art approaches. While Social-LSTM and Tajectron methods show better accuracy when evaluating the prediction error, they are overconfident on their predicted distributions, according to our proposed metric to measure uncertainty prediction. Our method, CovarianceNet, is able to predict correctly its uncertainty, thereby, our approach is better suited for applications requiring accurate prediction of distributions.

REFERENCES

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

[2] L. Leal-Taixé, M. Fenzi, A. Kuznetsova, B. Rosenhahn, and S. Savarese, "Learning an image-based motion context for multiple people tracking," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3542–3549, 2014.

[3] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "nuscenes: A multimodal dataset for autonomous driving," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11621–11631, 2020.

[4] T. Salzmann, B. Ivanovic, P. Chakravarty, and M. Pavone, "Trajectron++: Multi-agent generative trajectory forecasting with heterogeneous data for control," *arXiv preprint arXiv:2001.03093*, 2020.

[5] D. Mehta, G. Ferrer, and E. Olson, "Backprop-mpdm: Faster risk-aware policy evaluation through efficient gradient optimization," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1740–1746, IEEE, 2018.

[6] D. Helbing and P. Molnár, "Social force model for pedestrian dynamics," *Physical review E*, vol. 51, no. 5, p. 4282, 1995.

[7] F. Giuliari, I. Hasan, M. Cristani, and F. Galasso, "Transformer networks for trajectory forecasting," *arXiv preprint arXiv:2003.08111*, 2020.

[8] K. Sohn, H. Lee, and X. Yan, "Learning structured output representation using deep conditional generative models," in *Advances in neural information processing systems*, pp. 3483–3491, 2015.

[9] E. A. Sisbot, L. F. Marin-Urias, R. Alami, and T. Simeon, "A human aware mobile robot motion planner," *IEEE Transactions on Robotics*, vol. 23, no. 5, pp. 874–883, 2007.

[10] G. Ferrer, A. G. Zulueta, F. H. Cotarelo, and A. Sanfeliu, "Robot social-aware navigation framework to accompany people walking side-by-side," *Autonomous robots*, vol. 41, no. 4, pp. 775–793, 2017.

[11] B. D. Ziebart, N. Ratliff, G. Gallagher, C. Mertz, K. Peterson, J. A. Bagnell, M. Hebert, A. K. Dey, and S. Srinivasa, "Planning-based prediction for pedestrians," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 3931–3936, IEEE, 2009.

[12] C. Fulgenzi, A. Spalanzani, and C. Laugier, "Probabilistic motion planning among moving obstacles following typical motion patterns," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 4027–4033, IEEE, 2009.

[13] M. Kuderer, H. Kretzschmar, C. Sprunk, and W. Burgard, "Feature-based prediction of trajectories for socially compliant navigation.," in *Robotics: science and systems*, 2012.

[14] P. Trautman, J. Ma, R. M. Murray, and A. Krause, "Robot navigation in dense human crowds: Statistical models and experimental studies of human–robot cooperation," *The International Journal of Robotics Research*, vol. 34, no. 3, pp. 335–356, 2015.

[15] G. Ferrer and A. Sanfeliu, "Proactive kinodynamic planning using the extended social force model and human motion prediction in urban environments," in *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 1730–1735, IEEE, 2014.

[16] G. Ferrer and A. Sanfeliu, "Anticipative kinodynamic planning: multi-objective robot navigation in urban and dynamic environments," *Autonomous Robots*, vol. 43, no. 6, pp. 1473–1488, 2019.

[17] D. Sadigh, S. Sastry, S. A. Seshia, and A. D. Dragan, "Planning for autonomous cars that leverage effects on human actions.," in *Robotics: Science and Systems*, vol. 2, Ann Arbor, MI, USA, 2016.

[18] F. Zanlungo, T. Ikeda, and T. Kanda, "Social force model with explicit collision prediction," *EPL (Europhysics Letters)*, vol. 93, no. 6, p. 68005, 2011.

[19] F. Farina, D. Fontanelli, A. Garulli, A. Giannitrapani, and D. Prattichizzo, "Walking ahead: The headed social force model," *PloS one*, vol. 12, no. 1, p. e0169734, 2017.

[20] J. Van Den Berg, S. J. Guy, M. Lin, and D. Manocha, "Reciprocal n-body collision avoidance," in *Robotics research*, pp. 3–19, Springer, 2011.

[21] S. J. Guy, J. Chhugani, C. Kim, N. Satish, M. Lin, D. Manocha, and P. Dubey, "Clearpath: highly parallel collision avoidance for multi-agent simulation," in *Proceedings of the ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, pp. 177–187, 2009.

[22] A. Alahi, K. Goel, V. Ramanathan, A. Robicquet, L. Fei-Fei, and S. Savarese, "Social LSTM: Human trajectory prediction in crowded spaces," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 961–971, 2016.

[23] K. Messaoud, I. Yahiaoui, A. Verroust-Blondet, and F. Nashashibi, "Non-local social pooling for vehicle trajectory prediction," in *2019 IEEE Intelligent Vehicles Symposium (IV)*, pp. 975–980, IEEE, 2019.

[24] A. Mohamed, K. Qian, M. Elhoseiny, and C. Claudel, "Social-stgcnn: A social spatio-temporal graph convolutional neural network for human trajectory prediction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14424–14432, 2020.

[25] S. Yi, H. Li, and X. Wang, "Pedestrian behavior understanding and prediction with deep neural networks," in *European Conference on Computer Vision*, pp. 263–279, Springer, 2016.

[26] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[27] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, pp. 2672–2680, 2014.

[28] B. Ivanovic and M. Pavone, "The trajectron: Probabilistic multi-agent trajectory modeling with dynamic spatiotemporal graphs," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2375–2384, 2019.

[29] A. Jain, S. Casas, R. Liao, Y. Xiong, S. Feng, S. Segal, and R. Urtasun, "Discrete residual flow for probabilistic pedestrian behavior prediction," in *Conference on Robot Learning*, pp. 407–419, PMLR, 2020.

[30] K. Mangalam, H. Girase, S. Agarwal, K.-H. Lee, E. Adeli, J. Malik, and A. Gaidon, "It is not the journey but the destination: Endpoint conditioned trajectory prediction," *arXiv preprint arXiv:2004.02025*, 2020.

[31] A. Postnikov, A. Gamayunov, and G. Ferrer, "HSFM-SigmaNN: Combining a feedforward motion prediction network and covariance prediction," *arXiv preprint arXiv:2009.04299*, 2020.

[32] Y. Gal and Z. Ghahramani, "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," in *international conference on machine learning*, pp. 1050–1059, 2016.

[33] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *International Conference on Machine Learning*, pp. 1321–1330, PMLR, 2017.

[34] G. Ferrer and A. Sanfeliu, "Bayesian human motion intentionality prediction in urban environments," *Pattern Recognition Letters*, vol. 44, pp. 134–140, 2014.

[35] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.