IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2023, London, UK. Cite as RA-L paper.

Constrained Gaussian Processes with Integrated Kernels for Long-Horizon Prediction of Dense Pedestrian Crowd Flows

Stefan H. Kiss[1], Kavindie Katuwandeniya[1], Alen Alempijevic[1], and Teresa Vidal-Calleja[1]

***Abstract*—In this paper, we present a novel approach for predicting pedestrian crowd dynamics over longer time horizons (30s). In dense environments over long time horizons, the number of pedestrian interactions is high, leading to the degradation of traditional pedestrian trajectory estimation techniques. Alternatively, we consider the macroscopic properties of the crowd as a whole, focusing on the flow of density. This approach benefits from not considering pedestrians individually, and can probabilistically estimate the existence of previously unobserved individuals. We propose a novel approach to imposing a physical constraint on the crowd density flow. Initially, a coarse resolution prediction is generated by a Convolutional Recurrent Neural Network (ConvRNN), and subsequently smoothly interpolated by a Gaussian Process (GP). Using the linearity properties of GPs, a continuous representation of the crowd is produced that complies with both the ConvRNN's prediction and a conservation of density constraint. The approach is trained and analysed on the dense ATC dataset, where we show the advantages of the approach and the improvements from our contributions.**

***Index Terms*—Multi-Modal Perception for HRI, Probabilistic Inference**

I. Introduction

HUMANS tend to anticipate the motion of the people in their surroundings when navigating in crowds. This anticipation is their prediction capability, and is based on a life-long experience of walking in crowds where the physical and social attributes of crowd motions are implicitly learnt. The ability to foresee the future motion of the crowd allows humans to effectively plan paths less invasive of the crowd flow. This is the ideal behaviour we wish a robotic agent to execute when deployed in a dynamic crowded environment. In non-myopic frameworks, the main idea is to predict in some way the motion of the crowd and effectively plan into the predicted future to reach a destination while satisfying the social compliance criteria. In this vein, we focus on developing an effective crowd prediction model by exploiting their macroscopic properties, intended to be utilized for planning robot trajectories through the crowd with minimal social invasiveness.

Crowds can be described using their microscopic or macroscopic properties. Microscopic modelling considers each pedestrian individually, and typically handles interactions in a pairwise manner. Focusing on individual behaviours and interactions has been shown to work well locally [1], and is suitable in small, sparse crowds. Alternatively, macroscopic modelling considers the entire crowd as a whole, focusing on the large-scale dynamics of denser crowds. At higher crowd densities, the microscopic properties of individuals have fewer degrees of freedom [2] and therefore have a small effect on the overall crowd behaviour. This emphasizes the importance of using macroscopic properties at larger scales. In our work, we are interested in modelling large and dense crowds and thus the focus is on macroscopic crowd properties.

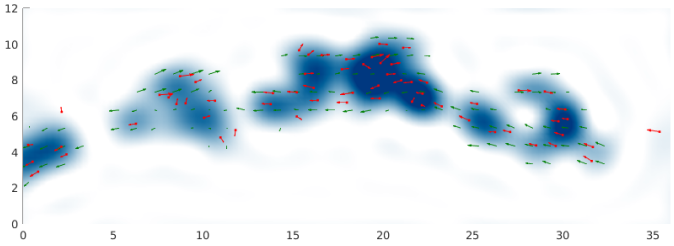


Fig. 1: The output of the presented method, overlaid with the ground-truth. As predicted almost 2.5 seconds into the future, the smooth density field ρ is shown in blue and the velocity field $\mathbf{v}$ is shown by green arrows. The red arrows indicate the true position and velocity of the pedestrians at this future time.

Our previous work on macroscopic crowd modelling [3] was shown to be effective for planning robot trajectories through dense crowds. A Convolutional Recurrent Neural Network (ConvRNN) framework was used to generate a discrete and fixed-resolution model. In this work, we improve upon the predictive model presented by using the coarsely predicted macroscopic crowd features to inform a constrained Gaussian Process (GP), giving a smooth and continuous representation of the crowd into the future. By filling-in the course resolution with a continuous probabilistic estimate, the crowd properties can be more accurately estimated at arbitrary query locations in space-time. More importantly, with the inclusion of a known physical constraint in the proposed framework: the conservation of density, the GP can intelligently interpolate to produce a physically plausible estimate. Additionally, as the framework produces a smooth and continuous representation of the crowd, gradients can be easily extracted from the GP to improve downstream applications such as planning and trajectory optimisation techniques [4].

In summary: the crowd is considered as a whole, and the macroscopic features of density and velocity are captured as averages over discretised space and time. A ConvRNN is used to encode the history of gridded values and predict future

[1]All authors with the Robotics Institute, University of Technology Sydney (UTS:RI), Australia. Corresponding author Stefan H. Kiss.

gridded values. Subsequently, as the main contribution of this paper, a constrained Gaussian process is used to reconstruct a smooth and continuous representation of the pedestrian crowd's macroscopic features, while adhering to physical constraints. This smooth representation of the macroscopic crowd properties produced by the prediction framework is shown in Fig. 1.

II. Related Work and Scope

Robots deployed in crowded environments such as museums, airports, and shopping malls must be socially compliant. The research community has attempted to model socially compliant motion using potential fields [5], velocity obstacles [6], graph-based methods [7], and reinforcement learning methods [8], where social compliance is defined in terms of avoiding collisions [9], [10], executing human-like maneuvers [8], or minimising invasiveness to the crowd's flow [11]. Predicting a pedestrian crowd's motion, also termed "crowd modelling", has been gaining the attention of the research community for various purposes including efficient crowd evacuation [12], traffic management and public safety [13], and planning socially-compliant robot motions [3], [14].

Most of these works consider short-term horizons, however for non-myopic robotic navigation, it is important that a reasonable prediction is made over the entire duration to the destination. We consider this horizon length around $30s$ in the large indoor settings of interest, and find the current research to be lacking at this time span. Additionally, most of the work in the literature considers discrete representations, while a model that produces a smooth and continuous representation of the environment is in fact desirable. An artificially discretised environment representation will induce an optimal (w.r.t. the environment) trajectory conforming to and heavily biased by the discretisation [15]. Furthermore, smooth and differentiable environment maps can be used to quickly converge to locally optimal trajectories via gradient descent [4]. When modelling the crowd spatially it is however important to ensure that densities produced satisfy the *conservation of density*: a prerequisite for probabilistically sound predictions that ensures pedestrians cannot appear, disappear, or teleport.

Table I compares a number of methods and contrasts their differences. State-of-the-art microscopic approaches such as [1], [16] are typically only accurate for predictions of a few seconds as they do not account for the characteristic behaviours within the environment itself. Such methods must also explicitly account for the interaction between pedestrians (and the space, if modelled); macroscopic approaches such as [3], [17], [18] automatically capture the spatial relationships of a crowd. Focusing on the space over longer time frames allows methods such as [18] to predict pedestrian behaviours hours into the future. However, such approaches typically fail to consider recent observations.

Our work aims to bridge and blend the gap between short and long term prediction approaches, leveraging the advantages of macroscopic approaches and generating a smooth and continuous representation that respects the conservation of density.

TABLE I: Comparison of approaches.

Method	Scale	Properties	Discretisation	Horizon	Conservation of Density
[1], [16][1]	Micro-	Position	Discrete per person, 0.4s in time.	4.8 seconds	**Yes**
[17]	Macro-	Density	0.1m in space[2], 0.2s in time.	2.4 seconds	No
[3]	Macro-	**Density, velocity**	1m in space, 0.5s in time.	5 seconds	No
[18]	Macro-	Velocity	**Continuous in space and time.**	24 hours	NA (no density)
Ours	Macro-	**Density, velocity**	**Continuous in space and time.**	**30 seconds**	**Yes**

[1]Representative of most microscopic approaches.
[2]Spatial resolution varies, estimated.

III. Macroscopic Crowd Properties

Typical formulations of pedestrian prediction represent the crowd as a collection of 2D point masses moving in time. In contrast, we consider the crowd to be continuous in time and space, described by a set of *macroscopic properties*. Our model has no concept of individual pedestrians, but is defined over the ambient space. We denote a location in space-time as $\mathbf{x} = [t, x, y]^\mathsf{T} \in \mathbb{R}^3$, and describe the crowd by its density ρ and velocity $\mathbf{v} = [v_x, v_y]^\mathsf{T}$ at every location[3],

$$\begin{aligned} \rho &: \mathbb{R}^3 \to \mathbb{R}_{\geq 0}\,, \\ \mathbf{v} &: \mathbb{R}^3 \to \mathbb{R}^2\,. \end{aligned} \tag{1}$$

This formulation helps to capture the uncertainty of pedestrian positions over time in a probabilistic manner. We propose to model the crowd density and velocity fields as follows.

The density of the crowd ρ can be considered as the intensity of a *point process*: the expected number N of people within an area A at time t is given by the associated integral over that area,

$$\mathbb{E}\left[N_A(t)\right] = \iint_{(x,y)\in A} \rho(t, x, y)\,\mathrm{d}x\,\mathrm{d}y\ . \tag{2}$$

The intensity of a point process is similar in many ways to a probability distribution: it is constrained to be non-negative, however it does not necessarily integrate to unity.

The uncertainty of the velocity field is handled through the explicit inclusion of velocity variance $\sigma^2_{\mathbf{v}}$. Through a probabilistic loss function (described in Section IV-A), this property captures two distinct factors: the uncertainty of the model's prediction, and, due to the discretised nature of the model's output, the variability of the velocities of pedestrians found within each discretisation region. This property is important to understand the coherence (conversely: irregularity) of the crowd flow.

[3]With a slight abuse of notation, we will refer to functions f and their evaluated values $f(\mathbf{x})$ interchangeably.

IV. ConvRNN Model

The problem of dynamic crowd prediction is first tackled by forecasting the macroscopic features at a coarse resolution. The space and time dimensions are discretised into fixed-resolution cells, and a Convolutional Recurrent Neural Network (ConvRNN) is used to learn the average future crowd properties across those cells.

As described in Section III, the macroscopic crowd features of interest are the density ρ and velocity $\mathbf{v}$ of the pedestrian crowd. The neural network is trained to estimate the *average* density $\bar{\rho}$ and velocity $\bar{\mathbf{v}}$ across the discretised cells $\mathfrak{c}$,

$$\begin{aligned} \bar{\rho}(\mathfrak{c}) &= \frac{1}{|\mathfrak{c}|} \iiint_{\mathbf{x}\in\mathfrak{c}} \rho(\mathbf{x})\,\mathrm{d}t\,\mathrm{d}x\,\mathrm{d}y\,, \\ \bar{\mathbf{v}}(\mathfrak{c}) &= \frac{1}{|\mathfrak{c}|} \iiint_{\mathbf{x}\in\mathfrak{c}} \mathbf{v}(\mathbf{x})\,\mathrm{d}t\,\mathrm{d}x\,\mathrm{d}y\,. \end{aligned} \tag{3}$$

We note that while the *true* target density function is not easy to define smoothly (an intuitive definition includes a sum of Dirac δ-functions at each pedestrian), this complication is avoided by considering the spatio-temporal averages.

For our experiments, we consider the network input to be a time-shifted version of the target output; a sequence of past crowd-property images is used to predict a sequence of future crowd-property images. However, we note that the input information could easily be of another type, like colour images of the area of interest. For this work, the input and output are rasterised top-down images with cells describing the average macroscopic crowd properties across them.

The ConvRNN used for the coarse prediction was modelled on that presented in [3]. While there is no drastic change to the neural network architecture in the current work, the model is trained to capture large-scale and long-horizon effects. In [3] the training data is arbitrarily rotated, a form of domain randomisation; in this work a consistent perspective is used to encourage the learning of behavioural patterns of the environment itself. Additionally, the network is trained to a longer horizon of 30 seconds. At this duration, most of the pedestrians seen in the observation period have since left the region of interest, replaced with new, unobserved individuals, which are difficult to predict by training on shorter prediction horizons [17]. Further details of the model are given in [3].

A. Probabilistic Loss Function

Through careful specification of the loss function, the learned model generates a *probabilistic* prediction of the future crowd state. The crowd is modelled as a *marked spatial point process*, with the spatial positions of an uncertain number of pedestrians each associated with an uncertain velocity. The density $\bar{\rho}(\mathfrak{c})$ is considered a Poisson random variable (as a Poisson point process) and the velocity $\bar{\mathbf{v}}(\mathfrak{c})$ a 2-dimensional isotropic Gaussian variable with a mean $\boldsymbol{\mu}_{\mathbf{v}}$ and a covariance $\sigma^2_{\mathbf{v}}\mathbb{I}_2$. Using the forward Kullback-Leibler (KL) divergence from the predictive distribution to the target ground-truth as the model's loss function, the model is effectively optimised by Maximum Likelihood Estimation (MLE). See [3] for more details.

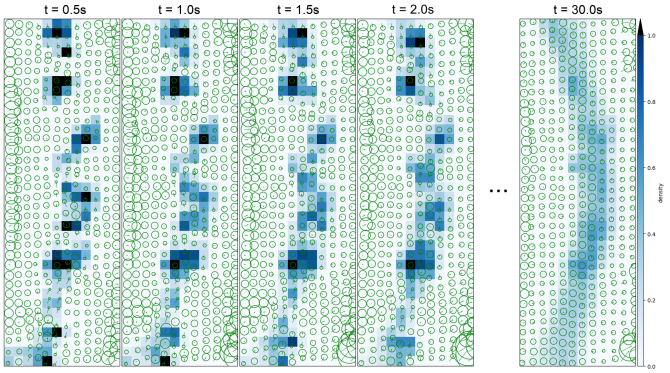


Fig. 2: The short and long time horizon behaviour of the ConvRNN prediction. Note that the final densities are not consistent for all inputs, the model seems to learn the overall density without being explicitly provided with the time of day information. Crowd density ρ is shown in blue, while the green arrows show the expected velocity and a σ-radius circle.

Such a probabilistic model is very beneficial for large-scale, long-horizon prediction of the macroscopic features. When predicting an intrinsically unknowable future, it is crucially important to capture the uncertainty in the modelling process. This allows the learned model to focus on regions it can be more certain about, and convey its uncertainty in regions it is not. The ConvRNN model shows its uncertainty at long time horizons by estimating large velocity variances, and blurring its predicted density image. In our crowd prediction problem, this is particularly useful as the model can seamlessly transition between collision-avoidance-like local behaviours at short time horizons, to a global environment-steady-state-like average behaviour at long time horizons. See Fig. 2 for a pictorial demonstration of this effect.

V. Gaussian Processes

From the course resolution predicted grid of macroscopic crowd features, we wish to obtain a smooth, continuous representation of these values across space and time. We employ a Gaussian Process to smoothly interpolate these values, while constraining the output using known physical constraints.

A GP is a generalisation of a Gaussian random variable to a random function. A GP essentially describes an infinite number of jointly-Gaussian random variables $f(\mathbf{x}_i)$ at locations $\mathbf{x}_i$. It is described by a mean function $\mu(\mathbf{x})$ (commonly zero or constant), and a covariance or kernel function $K(\mathbf{x}, \mathbf{x}')$,

$$\begin{bmatrix} f(\mathbf{x}_1) \\ \vdots \\ f(\mathbf{x}_n) \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu(\mathbf{x}_1) \\ \vdots \\ \mu(\mathbf{x}_n) \end{bmatrix}, \begin{bmatrix} K(\mathbf{x}_1,\mathbf{x}_1) & \cdots & K(\mathbf{x}_1,\mathbf{x}_n) \\ \vdots & \ddots & \vdots \\ K(\mathbf{x}_n,\mathbf{x}_1) & \cdots & K(\mathbf{x}_n,\mathbf{x}_n) \end{bmatrix} \right)$$
$$[f(\mathbf{x}_i)]_i \sim \mathcal{N}([\mu(\mathbf{x}_i)]_i, [K(\mathbf{x}_i,\mathbf{x}_j)]_{i,j})\,,$$
$$f(\mathbf{x}) \sim \mathcal{GP}(\mu(\mathbf{x}), K(\mathbf{x},\mathbf{x}'))\,. \tag{4}$$

We introduce $[\bullet]_i$ and $[\bullet]_{i,j}$ notation to indicate vector and matrix construction (respectively) from function evaluations.

Crucially, as all $f(\mathbf{x})$ variables are jointly-Gaussian, conditioning admits a closed-form solution. Given a set of observed data $\mathbf{y}_d = [f(\mathbf{x}_{d_i})]_i$ and a set of query locations $\mathbf{x}_q$, the

queried values $\mathbf{y}_q = [f(\mathbf{x}_{q_i})]_i$ are jointly-Gaussian, distributed as

$$\begin{aligned}\mathbf{y}_q \mid \mathbf{y}_d \sim \mathcal{N}\Big(\boldsymbol{\mu}_q &+ \boldsymbol{\Sigma}_{q,d}\,\boldsymbol{\Sigma}_{d,d}^{-1}\,(\mathbf{y}_d - \boldsymbol{\mu}_d),\\ &\boldsymbol{\Sigma}_{q,q} - \boldsymbol{\Sigma}_{q,d}\,\boldsymbol{\Sigma}_{d,d}^{-1}\,\boldsymbol{\Sigma}_{d,q}\Big)\,,\end{aligned} \tag{5}$$

where

$$\begin{aligned}\boldsymbol{\mu}_d &= [\mu(\mathbf{x}_{d_i})]_i\,, & \boldsymbol{\mu}_q &= [\mu(\mathbf{x}_{q_i})]_i\,,\\ \boldsymbol{\Sigma}_{d,d} &= [K(\mathbf{x}_{d_i},\mathbf{x}_{d_j})]_{i,j} + \boldsymbol{\Sigma}_{n,n}\,, & \boldsymbol{\Sigma}_{d,q} &= [K(\mathbf{x}_{d_i},\mathbf{x}_{q_j})]_{i,j}\,,\\ \boldsymbol{\Sigma}_{q,d} &= [K(\mathbf{x}_{q_i},\mathbf{x}_{d_j})]_{i,j}\,, \quad \text{and} & \boldsymbol{\Sigma}_{q,q} &= [K(\mathbf{x}_{q_i},\mathbf{x}_{q_j})]_{i,j}\,;\end{aligned}$$

assuming the observed data is additionally corrupted by some (also jointly Gaussian) noise with covariance $\boldsymbol{\Sigma}_{n,n}$.

Note that the definitions given for a scalar function $f(\mathbf{x})$ work equally well for a vector-valued function $\mathbf{f}(\mathbf{x})$. This however requires the definition of a vector-valued mean function $\boldsymbol{\mu}(\mathbf{x})$ and a (square) matrix-valued covariance function $\mathbf{K}(\mathbf{x},\mathbf{x}')$. The resulting data-data and query-data covariance matrices are similarly constructed as block matrices.

A. Linear Transformations of Gaussian Processes

One interesting and useful property of Gaussian processes is their linear transformation rules. Consider $\mathbf{f}_1(\mathbf{x})$, a GP defined by $\boldsymbol{\mu}_1(\mathbf{x})$ and $\mathbf{K}_{1,1}(\mathbf{x},\mathbf{x}')$. Given a linear operator $\mathcal{L}_2$ mapping $\mathbf{f}_1$ to $\mathbf{f}_2 : \mathbf{f}_2(\mathbf{x}) = \mathcal{L}_2\,\mathbf{f}_1(\mathbf{x})$, $\mathbf{f}_2(\mathbf{x})$ is also a Gaussian process, with associated mean and covariance functions (with itself, and with the un-transformed $\mathbf{f}_1$),

$$\begin{aligned}\boldsymbol{\mu}_2(\mathbf{x}) &= \mathcal{L}_2\,\boldsymbol{\mu}_1(\mathbf{x})\,,\\ \mathbf{K}_{2,2}(\mathbf{x},\mathbf{x}') &= \mathcal{L}_2\,\mathbf{K}_{1,1}(\mathbf{x},\mathbf{x}')\,\mathcal{L}_2^{\mathsf{T}}\,,\\ \mathbf{K}_{1,2}(\mathbf{x},\mathbf{x}') &= \quad\ \mathbf{K}_{1,1}(\mathbf{x},\mathbf{x}')\,\mathcal{L}_2^{\mathsf{T}}\,,\\ \mathbf{K}_{2,1}(\mathbf{x},\mathbf{x}') &= \mathcal{L}_2\,\mathbf{K}_{1,1}(\mathbf{x},\mathbf{x}')\,.\end{aligned} \tag{6}$$

The notation of these expressions can be confusing: the base kernel $\mathbf{K}_{1,1}$ is a matrix-valued function of 2 locations; the linear operator $\mathcal{L}_2$ operates from the left on the left location $\mathbf{x}$ and, with a slight abuse of notation, the transposed linear operator $\mathcal{L}_2^{\mathsf{T}}$ is understood to operate on the kernel from the right with respect to the right location $\mathbf{x}'$.

B. Conservation Constraint

Of the possible $\rho(\mathbf{x})$ and $\mathbf{v}(\mathbf{x})$ functions, defined over space and time, not all are physically feasible. The movement of density ρ through time is specified exactly by the flow velocity $\mathbf{v}$; as such these functions must obey the following differential equation to satisfy the conservation of density:

$$\frac{\partial\rho}{\partial t} + \frac{\partial\rho v_x}{\partial x} + \frac{\partial\rho v_y}{\partial y} = 0\,. \tag{7}$$

This equation describes the crowd's momentum vector $\rho\mathbf{v}$. The change in density over time is exactly the negative divergence of the momentum vector field: the density at a position decreases with the net flow of pedestrians away from that position. We can model this property of the flow as a constrained Gaussian process.

By the careful choice of kernel, the GP can be crafted to satisfy this constraint at all locations. Firstly, the constraint is described as a linear operation $\mathcal{L}_c$ (termed the *constraint operator*) on a *state* $\mathbf{s}$,

$$\begin{aligned}\mathbf{s} &= \begin{bmatrix}\rho\\ \rho v_x\\ \rho v_y\end{bmatrix}\,, \qquad \mathcal{L}_c = \begin{bmatrix}\partial_t & \partial_x & \partial_y\end{bmatrix}\,,\\ \mathcal{L}_c\,\mathbf{s} &= \begin{bmatrix}\partial_t & \partial_x & \partial_y\end{bmatrix}\begin{bmatrix}\rho\\ \rho v_x\\ \rho v_y\end{bmatrix} = 0\,,\end{aligned} \tag{8}$$

where $\partial_\bullet$ is the derivative operator $\frac{\partial}{\partial\bullet}$.

Subsequently, our aim is to find a linear operator $\mathcal{L}_s$, we term the *producing operator*, which satisfies

$$\begin{aligned}\mathbf{s} &= \quad \mathcal{L}_s\,\mathbf{g}\,,\\ 0 &= \mathcal{L}_c\,\mathcal{L}_s\,\mathbf{g}\,, \quad \forall\mathbf{g}\,.\end{aligned} \tag{9}$$

For any (vector-valued) continuous function $\mathbf{g}(\mathbf{x})$ defined over space and time, $\mathcal{L}_s\,\mathbf{g}$ should satisfy the constraint. We term $\mathbf{g}$ the *latent state*, and aim to find a valid solution for $\mathcal{L}_s$,

$$0 \underbrace{\xleftarrow{\mathcal{L}_c}}_{\substack{\text{constraint}\\ \text{operator}}} \underbrace{\mathbf{s}(\mathbf{x})}_{\text{state}} \underbrace{\xleftarrow{\mathcal{L}_s}}_{\substack{\text{producing}\\ \text{operator}}} \underbrace{\mathbf{g}(\mathbf{x})}_{\substack{\text{latent}\\ \text{state}}}\,. \tag{10}$$

A 1-dimensional solution for $\mathcal{L}_s$ can be written as a linear combination of scalar basis operators. Higher-dimensional solutions can be found spanning the null space of this linear combination, leading to a more a flexible model still satisfying the constraint. For a complete description of this process, readers are referred to [19]. For our specific problem, we find a 3-dimensional solution of the form:

$$\begin{aligned}\mathcal{L}_s &= \begin{bmatrix}0 & -\partial_y & \partial_x\\ \partial_y & 0 & -\partial_t\\ -\partial_x & \partial_t & 0\end{bmatrix}\,,\\ \mathcal{L}_c\,\mathcal{L}_s &= \begin{bmatrix}\partial_t & \partial_x & \partial_y\end{bmatrix}\begin{bmatrix}0 & -\partial_y & \partial_x\\ \partial_y & 0 & -\partial_t\\ -\partial_x & \partial_t & 0\end{bmatrix}\\ &= \begin{bmatrix}0 & 0 & 0\end{bmatrix}\,.\end{aligned} \tag{11}$$

Note that the composition of linear operators remains a linear operator, namely the *zero operator*, taking any 3D latent state $\mathbf{g}$ and reducing it to (scalar) zero everywhere.

This solution relies on the commutative property of differential operators, $\partial_x\partial_y = \partial_y\partial_x$. The solution can also be seen as the skew-symmetric matrix representation of the *curl* operator, and it is a well known result that the divergence of the curl of a vector field is always zero.

The properties of linear operators on GPs (as seen in Section V-A) allow us to specify a covariance kernel $\mathbf{K}_{g,g}(\mathbf{x},\mathbf{x}')$ defining a GP for the latent state $\mathbf{g}$, and apply the operator $\mathcal{L}_s$. The resulting continuous function $\mathbf{s}$ must, by definition, satisfy the constraint $\mathcal{L}_c$. As the operator $\mathcal{L}_s$ is linear, the function of interest $\mathbf{s}$ remains a GP, with mean and kernel functions defined in terms of $\boldsymbol{\mu}_g$, $\mathbf{K}_{g,g}$, and $\mathcal{L}_s$,

$$\begin{aligned}\mathbf{g}(\mathbf{x}) &\sim \mathcal{GP}\Big(\quad \boldsymbol{\mu}_g(\mathbf{x}),\quad \mathbf{K}_{g,g}(\mathbf{x},\mathbf{x}')\quad\Big)\,,\\ \mathbf{s}(\mathbf{x}) &\sim \mathcal{GP}\Big(\mathcal{L}_s\,\boldsymbol{\mu}_g(\mathbf{x}),\,\mathcal{L}_s\,\mathbf{K}_{g,g}(\mathbf{x},\mathbf{x}')\,\mathcal{L}_s^{\mathsf{T}}\Big)\,.\end{aligned} \tag{12}$$

Note that we define $\boldsymbol{\mu}_s = \mathcal{L}_s \boldsymbol{\mu}_g$ directly, as a constant positive density with zero velocity, with the understanding that the associated $\boldsymbol{\mu}_g$ exists and the constraint $\mathcal{L}_c$ is satisfied.

The base kernel $\mathbf{K}_{g,g}$ is defined with the typical squared-exponential,

$$\mathbf{K}_{g,g} = \begin{bmatrix} K_{g_1} & 0 & 0 \\ 0 & K_{g_2} & 0 \\ 0 & 0 & K_{g_3} \end{bmatrix}, \tag{13}$$

$$K_{g_i} = a_i^2 \exp\left(-\sum_{j \in (t,x,y)} \frac{[\mathbf{x} - \mathbf{x}']_j^2}{2\, l_{i,j}^2} \right), \tag{14}$$

where $[\bullet]_j$ denotes extracting the scalar value from the applicable dimension of the associated vector, such that different length scales $l_{i,j}$ can be used for each dimension.

The diagonal nature of $\mathbf{K}_{g,g}$ establishes the 3 dimensions of the latent state $\mathbf{g}$ as uncorrelated, however we note that the 3 dimensions of the produced state $\mathbf{s}$ *are* correlated as $\mathcal{L}_s$ contains off-diagonal elements.

Three variances a_i^2 and nine length-scales $l_{i,j}$ define hyper-parameters of this kernel. To retain symmetry in the xy-plane, 6 equality constraints are introduced leaving 6 tuneable hyper-parameters, which are selected empirically.

Note that in the constrained kernel $\mathbf{K}_{s,s}$ the density complies with the flow of momentum, satisfying at all locations and times the conservation constraint.

C. Measurement Integral Operator

The output of the ConvRNN outlined in Section IV is a 3-dimensional array of predicted crowd properties, interpreted as the average values across cells in space-time. This naturally gives rise to a piecewise-constant model. A smoother model is produced by using these average values to inform a GP model of the ConvRNN's predictions. We introduced a similar approach for correlating rectangular spatial regions of differing sizes in [20].

The average value over a cellular region is calculated by the triple integral over space and time. This transformation can be described as a linear operator $\mathcal{L}_m$ which we use to define a *measurement state* $\mathbf{m}$,

$$\begin{bmatrix} \bar{\rho} \\ \overline{\rho v_x} \\ \overline{\rho v_y} \end{bmatrix} = \mathbf{m}(\mathfrak{c}) = \mathcal{L}_m\, \mathbf{s}(\mathbf{x}) = \frac{1}{|\mathfrak{c}|} \iiint_{\mathbf{x} \in \mathfrak{c}} \mathbf{s}(\mathbf{x})\, \mathrm{d}x\, \mathrm{d}y\, \mathrm{d}z\ . \tag{15}$$

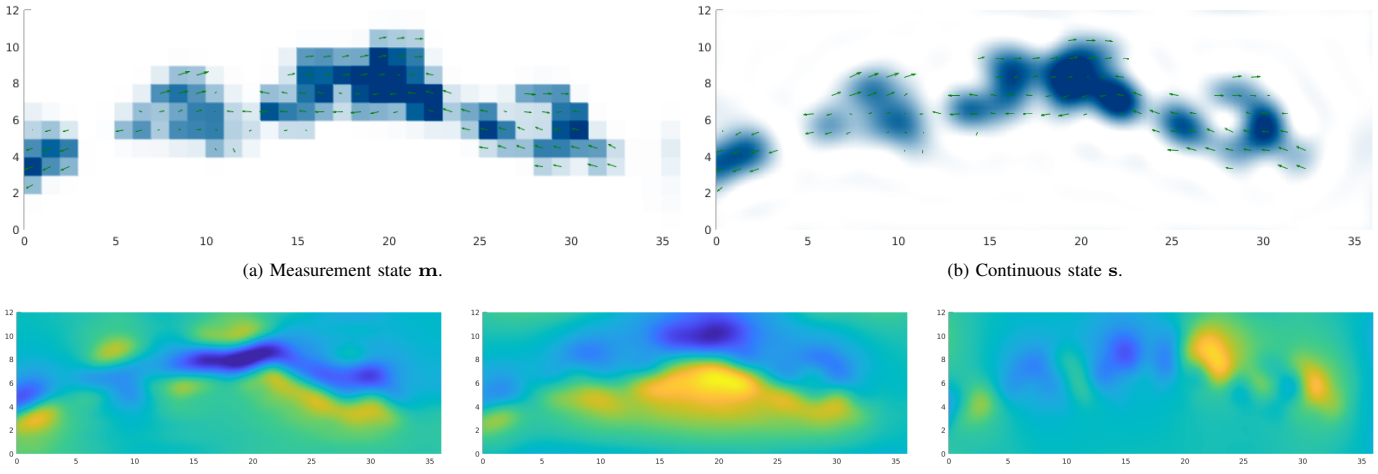
Note the similarities to the average values $\bar{\rho}(\mathfrak{c})$ and $\bar{\mathbf{v}}(\mathfrak{c})$ introduced in Section IV.

As the *measurement operator* $\mathcal{L}_m$ is linear, if the state $\mathbf{s}$ is a GP then the measurement state $\mathbf{m}$ is also a GP with the covariance kernel modified by $\mathcal{L}_m$,

$$\underbrace{\mathbf{m}(\mathfrak{c})}_{\substack{\text{measurement} \\ \text{state}}} \underbrace{\xleftarrow{\quad\mathcal{L}_m\quad}}_{\substack{\text{measurement} \\ \text{operator}}} \underbrace{\mathbf{s}(\mathbf{x})}_{\text{state}}\ . \tag{16}$$

We note that this averaging operation is equivalent to convolving the kernel with an appropriately sized rectangular window filter. This approach can therefore be considered more generally, by convolving the state's kernel with any number of measurement filters, as all convolutions are linear integration operators. Analytic evaluation of the convolved kernel, however, may prove more difficult.

To simplify evaluation, the measurement operator $\mathcal{L}_m$ is split into two parts. We use an approach similar to that of a continuous summed-area table. First, the quantity of interest



(a) Measurement state $\mathbf{m}$.

(b) Continuous state $\mathbf{s}$.

(c) The three components of latent state $\mathbf{g}$: g_1, g_2, and g_3 from left to right.

Fig. 3: The coarse $\mathbf{m}$ grid, the continuous $\mathbf{s}$, and the components of $\mathbf{g}$. All views are taken at a single slice in time, and Fig. 1 additionally shows the true pedestrian locations at this time. In Figs. 3a and 3b, the density field is shown in blue and the flow velocity is shown with green arrows. In Fig. 3c, the latent values are shown in blue-yellow with teal zero. While g_2 and g_3 are harder to interpret, g_1 can be seen as approximately the instantaneous 2-dimensional *stream function*, where the gradient is largely perpendicular to the crowd flow $\rho\mathbf{v}$.

is integrated in space and time with $\int_t \int_x \int_y$, where $\int_\bullet$ is the integration operator

$$\int_\bullet f(\bullet) = \int_{-\infty}^{\bullet} f(\tau)\,\mathrm{d}\tau\ . \tag{17}$$

Secondly, the function values at the corners are added and subtracted as appropriate to form the definite integral of Eq. (15). The data cells $\mathfrak{c}_d$ are specified by the eight (2^3) corners of each observation volume $\mathbf{x}_d$. The average value within each cell can be expressed as a sparse linear combination of the integrated value at the eight corners of the cell. For convenience, this transformation is given by the matrix $\mathbf{M}_{\mathfrak{c},\mathbf{x}}$. The matrix $\mathbf{M}_{\mathfrak{c},\mathbf{x}}$ is a non-square sparse matrix. Due to the surface-volume ratio, as the number of cells increases the aspect of this matrix becomes more square, increasing the efficiency.

Thus, the measurement operator $\mathcal{L}_m$ can be re-written as

$$\mathcal{L}_m = \mathbf{M}_{\mathfrak{c},\mathbf{x}} \int_t \int_x \int_y . \tag{18}$$

The integrations are applied directly to the kernel where appropriate (see Section V-D) and the sparse matrix $\mathbf{M}_{\mathfrak{c},\mathbf{x}}$ is multiplied after calculating the kernel values at the data locations $\mathbf{x}_d$. This efficiently allows the kernel to be evaluated at each corner only once, and used in the calculation of each of the neighboring cell values. As most corners are shared by eight neighboring cells, this reduces the complexity of computing the kernels approximately eight-fold.

D. Kernels

The continuous latent state $\mathbf{g}(\mathbf{x})$ is defined as a GP, and transformed to the discretised cell measurements $\mathbf{m}(\mathfrak{c})$ observed and predicted by the ConvRNN with the linear operators $\mathcal{L}_s$ and $\mathcal{L}_m$ specified in Sections V-B and V-C respectively,

$$\mathbf{m}(\mathfrak{c}) \xleftarrow{\mathcal{L}_m} \mathbf{s}(\mathbf{x}) \xleftarrow{\mathcal{L}_s} \mathbf{g}(\mathbf{x})\,. \tag{19}$$

Visualisations of these different states are shown in Fig. 3.

The kernel functions of interest can be specified following the linear transformation properties of GPs described in Section V-A,

$$\begin{aligned} \mathbf{K}_{m,m}(\mathfrak{c},\mathfrak{c}') &= \mathcal{L}_m\, \mathcal{L}_s\, \mathbf{K}_{g,g}(\mathbf{x},\mathbf{x}')\, \mathcal{L}_s^\mathsf{T}\, \mathcal{L}_m^\mathsf{T}\,, \\ \mathbf{K}_{s,m}(\mathbf{x},\mathfrak{c}') &= \quad\ \mathcal{L}_s\, \mathbf{K}_{g,g}(\mathbf{x},\mathbf{x}')\, \mathcal{L}_s^\mathsf{T}\, \mathcal{L}_m^\mathsf{T}\,. \end{aligned} \tag{20}$$

Finally, the full covariance matrices between the observed data d and itself, and between the observed data and the queried values q, respectively, can be calculated as

$$\begin{aligned} \mathbf{\Sigma}_{d,d} &= \left[\mathbf{K}_{m,m}(\mathfrak{c}_{d_i},\mathfrak{c}_{d_j})\right]_{i,j} + \mathbf{\Sigma}_{n,n}\,, \\ \mathbf{\Sigma}_{q,d} &= \left[\mathbf{K}_{s,m}(\mathbf{x}_{q_i},\mathfrak{c}_{d_j})\right]_{i,j}\,. \end{aligned} \tag{21}$$

The observed data is assumed to be additionally corrupted by measurement noise with covariance $\mathbf{\Sigma}_{n,n}$, the form of which is non-trivial and is described in Section V-E below. The Gaussian conditioning equations can then be used to calculate the posterior distribution as per Eq. (5). However are primarily focused on the mean of the prediction,

$$\mathbb{E}\left[\mathbf{s}_q \mid \mathbf{m}_d\right] = \boldsymbol{\mu}_q + \mathbf{\Sigma}_{q,d}\mathbf{\Sigma}_{d,d}^{-1}\left(\mathbf{m}_d - \boldsymbol{\mu}_d\right)\,. \tag{22}$$

E. Observation Measurement Noise

We consider the predictions from the ConvRNN as measurements to inform the Gaussian process. The estimated values of density $\bar{\rho}$ and velocity $\bar{\mathbf{v}}$ are assumed to be corrupted by noise, which allows the GP to correct unlikely (or impossible, due to the constraint) predictions from the neural network. The noise is assumed to be independent between density $\bar{\rho}$ and velocity $\bar{\mathbf{v}}$ measurements, therefore the noise on the measurement state $\mathbf{m}$, defined in terms of momentum $\overline{\rho\mathbf{v}}$, is in fact correlated.

The neural network is considered to output parameters $(\mu_\rho, \mu_{v_x}, \mu_{v_y}, \sigma^2_\mathbf{v})$, describing the variables with uncertainty (from their Poisson and isotropic Gaussian definitions, as described in Sections III and IV-A),

$$\begin{aligned} \mathbb{E}\left[\bar{\rho}\right] &= \ \mu_\rho\,, \qquad \mathbb{V}[\bar{\rho}] = \ \sigma^2_\rho\,, \\ \mathbb{E}\left[\bar{\mathbf{v}}\right] &= \begin{bmatrix}\mu_{v_x}\\ \mu_{v_y}\end{bmatrix}\,, \quad \mathbb{V}[\bar{\mathbf{v}}] = \begin{bmatrix}\sigma^2_\mathbf{v} & 0\\ 0 & \sigma^2_\mathbf{v}\end{bmatrix}\,. \end{aligned} \tag{23}$$

The density variance σ^2_ρ is assigned a constant, empirically selected value of 0.01.

Based on the definition of covariance of a random variable, and using the independence assumption between ρ and $\mathbf{v}$, the covariance of the observation noise $\mathbf{\Sigma}_{n,n}$ can be written as

$$\begin{aligned} \mathbf{\Sigma}_{n,n} &= \mathbb{V}[\mathbf{m}] = \mathbb{E}\left[\mathbf{m}\mathbf{m}^\mathsf{T}\right] - \mathbb{E}\left[\mathbf{m}\right]\mathbb{E}\left[\mathbf{m}\right]^\mathsf{T} \\ &= \begin{bmatrix} \sigma^2_\rho & & \text{(sym.)} \\ \mu_{v_x}\sigma^2_\rho & \mu^2_\rho\sigma^2_\mathbf{v} + \mu^2_{v_x}\sigma^2_\rho + \sigma^2_\rho\sigma^2_\mathbf{v} & \\ \mu_{v_y}\sigma^2_\rho & \mu_{v_x}\mu_{v_y}\sigma^2_\rho & \mu^2_\rho\sigma^2_\mathbf{v} + \mu^2_{v_y}\sigma^2_\rho + \sigma^2_\rho\sigma^2_\mathbf{v} \end{bmatrix}\,. \end{aligned} \tag{24}$$

VI. Implementation and Evaluation

A. Dataset

The ATC shopping centre dataset [21] was used for both training and evaluation of the proposed framework. The dataset captures pedestrians navigating the ATC shopping centre, two days a week for over two years. We used the Sunday subset of the data as they typically show peak densities over three times greater than the weekday (Wednesday) data. We also focused our attention to the eastern corridor as it is an especially dense region. Note that this subset still contains a wide variety of densities, as the crowd levels fluctuate dramatically according to the time of day. With a total of 41 days, 30 were used for training, 10 for validation, and 1 for evaluation.

Fig. 4 shows an overview of the ATC dataset, showing the discretisation grid used over the eastern corridor.

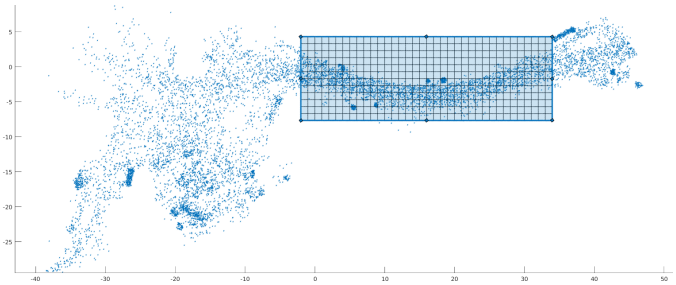


Fig. 4: A visualisation of the dense east hallway section of the ATC dataset used for evaluation. The 12×36 grid used for the discretisation process is shown. For context, 10 000 pedestrian locations are randomly selected from a typical Sunday.

B. Implementation Details

The ConvRNN model described in Section IV was implemented in PyTorch [22]. The input observations and output predictions spatially are 12×36 raster image sequences where each cell is $1m \times 1m$. Temporally, each cell represents a duration of $0.5s$, and the network observes 10 frames for a total of $5s$. The model can then be used to predict any number of frames, however it was trained using gradient descent to predict 60 frames ($30s$).

The output of the ConvRNN is subsequently smoothed to any resolution using the proposed constrained GP method, implemented in MATLAB [23]. The kernel hyper-parameters (described in Section V-B and Eq. (14)) were selected empirically as the following:

$$\begin{aligned} a_1 &= 0.5\,, & a_2 = a_3 &= 0.5\,, \\ l_{1,t} &= 2.0\,, & l_{2,x} = l_{2,y} = l_{3,x} = l_{3,y} &= 1.5\,, \\ l_{1,x} = l_{1,y} &= 1.5\,, & l_{2,t} = l_{3,t} &= 1.0\,. \end{aligned} \tag{25}$$

We note that these hyper-parameters as well as the neural network parameters could be trained using an objective function based on the continuous representation, back-propagated through the GP, however this is left for future work.

After interpolation, generated values are clamped within physically-plausible bounds, to prevent extreme outliers corrupting the results.

C. Evaluation Metrics

Traditional metrics for pedestrian trajectory prediction using microscopic features include Average Displacement Error (ADE) and Final Displacement Error (FDE). These are calculated as the difference in position per person between the prediction and the ground-truth, averaged along the trajectory or at the final endpoint of the trajectory respectively. These metrics are typically extended to $\min_k$ADE and $\min_k$FDE respectively when the predictive model produces a distribution: k trajectory samples (20, as is common practice) are generated and evaluated, with the minimum error selected.

As our work focuses on the macroscopic crowd features, our approach has no concept of individual pedestrians; therefore we need to extend the tradition definitions of these metrics to models in the un-associated, macroscopic feature domain. We achieve this by sampling from the density field, considered as the intensity of a Poisson point process at every slice in time. A Poisson point process can be sampled by first randomly sampling positions in the region of interest at a rate of the maximum intensity (distributed uniformly), and then accepting samples proportional to the underlying intensity at those locations [24]. Positions are sampled from the density field k times over and the ground-truth pedestrian positions are associated with the closest sample (without overlap, using optimal assignment), achieving a generalisation of $\min_k$ADE and $\min_k$FDE.

Note that as the predicted pedestrian positions are initially *un-associated* with the ground-truth pedestrians, and are assigned optimally, this metric is not directly comparable to the typical *associated* definition as it potentially gives an unfair advantage to un-associated predictions.

Additionally note that the proposed macroscopic approach allows us to jointly estimate location *and number* of pedestrians, and extends the $\min_k$ADE and $\min_k$FDE metrics to this modality, as opposed to traditional microscopic methods which are unable to estimate the existence of pedestrians that have not yet been observed. This is especially important in the modelling of dense areas over longer time horizons, where the majority of pedestrians exit the scene and new individuals enter.

We also analyse the velocity predicted by the model, considering the magnitude of the velocity error at the true locations of the pedestrians. Similarly to displacement error, two metrics are produced: Average Velocity Error (AVE) and Final Velocity Error (FVE).

D. Results

Fig. 2 shows the prediction of the ConvRNN model, showing the short and long time horizon prediction behaviour. Qualitatively, we observe that the initial frames are high-contrast and quite certain, whereas the last frame is blurred uncertain. This demonstrates the power of the macroscopic feature approach: while individuals and small groups can be tracked initially, the model can smoothly blend towards a holistic, probabilistic crowd prediction.

At longer time horizons, the prediction converges towards a steady-state average behaviour, capturing the usual densities and traffic patterns of the space. We note that the densities in the corridor fluctuate dramatically throughout the day. Interestingly, we find the ConvRNN manages to learn this property; while the network is not directly informed of the time of day, it is able to learn average density patterns applicable to the time of day from only a few (10) observations of the current densities.

The model was evaluated on a previously unseen day of the ATC dataset. As we focus our attention to particularly dense scenarios, only sequences with an average density of 30 pedestrians or above are selected. A total of 1323 sequences are analysed; for each sequence, 5 seconds are observed by the framework and used to predict the following 30 seconds.

Social GAN (SGAN) [1] was used as a comparison baseline representative of microscopic methods. As SGAN produces *associated* predictions for each pedestrian, the predictions were un-associated and re-associated using optimal assignment to provide a fair comparison with the naturally un-associated macroscopic prediction methods.

Note that SGAN was designed and trained for a short $4.8s$ prediction horizon, so it is evaluated at this horizon. However, for comparison to the macroscopic models' longer prediction, it is also evaluated at $30s$.

The $\min_k$ADE, $\min_k$FDE, AVE, and FVE metrics are calculated as described above, and displayed in Table II. The errors from the Constrained Gaussian Process (CGP) model are additionally compared to a baseline Piece-Wise Constant (PWC) model, directly using the rasterised predictions

TABLE II: Experimental results.

Model	Horizon	$\min_k$ADE $(m)\downarrow$	$\min_k$FDE $(m)\downarrow$	AVE $(m/s)\downarrow$	FVE $(m/s)\downarrow$
SGAN [1]	$4.8s$	0.8248	1.2656	—	—
	$30s$	2.2223	3.9203	—	—
PWC	$30s$	0.2387	0.2611	**0.5780**	0.7158
CGP	$30s$	**0.2372**	**0.2586**	0.5791	**0.7068**

from the ConvRNN. We note that due to the geometry of the spatial discretisation, the PWC model cannot reduce the $\min_k$ distance error metrics below a theoretical limit above 0.1.

Overall, both macroscopic models perform well under the distance error metrics. While the validity of SGAN's $30s$ predicions are debatable, we note that the macroscopic models outperform SGAN's $4.8s$ predictions at over 6 times the prediction horizon. This long-horizon accuracy is due to the holistic perspective of the macroscopic approach, fusing information from recent observations with spatial information about the environment.

The table shows that the CGP model improves upon the PWC model by a small percentage in the $\min_k$ADE and $\min_k$FDE metrics. Notably, the *final* displacement metric makes a greater improvement than the *average* displacement metric, implying that this method is more suitable at longer time horizon predictions. As the ConvRNN predictions become less certain, they become smoother, which agrees with the Gaussian process's assumption of smoothness.

The velocity error metrics demonstrates this phenomenon more strongly, which Fig. 5 shows in detail. Application of the GP is actually detrimental for velocity prediction for horizons less than 8 seconds, however it provides a consistent improvement from 10 seconds onwards. Again, we attribute this to the smoothness assumption of the Gaussian process kernel. Additional careful tuning of the kernel hyper-parameters or modification of the base kernel $\mathbf{K}_{g,g}$ may render the Gaussian process equally applicable across all prediction horizons.

VII. Conclusion

In this work, we proposed a method of crowd prediction using a continuous representation of the macroscopic features over space and time. The framework builds upon a ConvRNN model that produces a coarsely discretised array of values into the future, representing the average expected value in each space-time cell. Using the properties of linear operators, a GP is used to regress the mean values into a smooth and continuous prediction while adhering to the conservation of density.

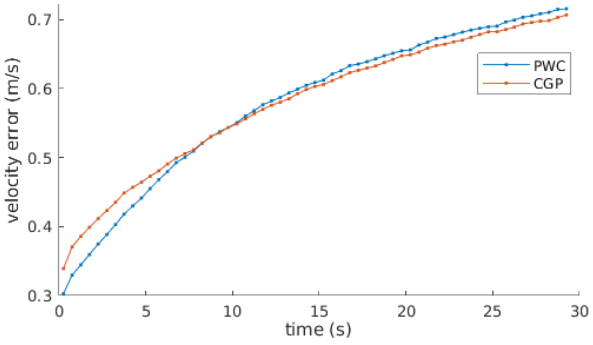


Fig. 5: While the CGP model does not improve upon the PWC prediction at short time horizons, it is more accurate at longer time horizons when the ConvRNN's predictions are smoother.

The framework is validated using the dense ATC dataset and compared to the microscopic approach SGAN. The macroscopic approach is shown to be effective, and the constrained Gaussian process model provides a improvement over the baseline models.

In future work, we intend to make use of the probabilistic forecast to plan non-myopic, non-invasive robot trajectories through dense pedestrian crowds, taking advantage of the smoothness and differentiability of the spatial representation.

References

[1] A. Gupta *et al.*, "Social gan: Socially acceptable trajectories with generative adversarial networks," in *CVPR*, 2018.

[2] R. Narain, A. Golas, S. Curtis, and M. C. Lin, "Aggregate dynamics for dense crowd simulation," in *ACM SIGGRAPH Asia*, 2009.

[3] S. H. Kiss, K. Katuwandeniya, A. Alempijevic, and T. Vidal-Calleja, "Probabilistic dynamic crowd prediction for social navigation," in *ICRA*. IEEE, 2021.

[4] M. Zucker *et al.*, "Chomp: Covariant hamiltonian optimization for motion planning," *IJRR*, vol. 32, no. 9-10, 2013.

[5] S. S. Ge and Y. J. Cui, "Dynamic Motion Planning for Mobile Robots Using Potential Field Method," *Autonomous robots*, vol. 13, no. 3, 2002.

[6] P. Fiorini and Z. Shiller, "Motion Planning in Dynamic Environments Using Velocity Obstacles," *IJRR*, vol. 17, no. 7, 1998.

[7] J. J. H. Lee *et al.*, "Efficient optimal planning in non-fifo time-dependent flow fields," *arXiv:1909.02198*, 2019.

[8] P. Henry, C. Vollmer, B. Ferris, and D. Fox, "Learning to navigate through crowded environments," in *ICRA*. IEEE, 2010.

[9] W. Burgard *et al.*, "The interactive museum tour-guide robot," in *Aaai/iaai*, 1998.

[10] E. Prassler, J. Scholz, and P. Fiorini, "A robotics wheelchair for crowded public environment," *RA Magazine*, vol. 8, no. 1, 2001.

[11] S. H. Kiss, K. Y. C. To, C. Yoo, R. Fitch, and A. Alempijevic, "Minimally Invasive Social Navigation," in *ACRA*. ARAA, 2019.

[12] B. G. Silverman *et al.*, "Human behavior models for agents in simulators and games," *Presence: Teleoperators Virtual Environ.*, vol. 15, 2006.

[13] J. Zhang, Y. Zheng, and D. Qi, "Deep spatio-temporal residual networks for citywide crowd flows prediction," 2017.

[14] A. Rudenko *et al.*, "Joint long-term prediction of human motion using a planning-based social force approach," in *ICRA*. IEEE, 2018.

[15] D. Kularatne *et al.*, "Optimal path planning in time-varying flows using adaptive discretization," *RA-L*, vol. 3, no. 1, 2017.

[16] B. Ivanovic and M. Pavone, "Probabilistic multi-agent trajectory modeling with dynamic spatiotemporal graphs," in *ICCV*. IEEE/CVF, 2019.

[17] H. Minoura *et al.*, "Crowd density forecasting by modeling patch-based dynamics," *RA-L*, vol. 6, no. 2, 2020.

[18] T. Vintr *et al.*, "Time-varying pedestrian flow models for service robots," in *ECMR*. IEEE, 2019, pp. 1–7.

[19] C. Jidling, N. Wahlström, A. Wills, and T. B. Schön, "Linearly constrained gaussian processes," *NIPS*, vol. 30, 2017.

[20] L. Jin *et al.*, "Adaptive-resolution gaussian process mapping for efficient uav-based terrain monitoring," *arXiv preprint arXiv:2109.14257*, 2021.

[21] D. Brščić *et al.*, "Person tracking in large public spaces using 3-d range sensors," *THMS*, vol. 43, no. 6, 2013.

[22] A. Paszke *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *NIPS*, vol. 32, 2019.

[23] *MATLAB version 9.8.0.1873465 (R2020a) Update 8*, The Mathworks, Inc., Natick, Massachusetts, 2020.

[24] R. L. Streit, *Poisson point processes: imaging, tracking, and sensing*. Springer Science & Business Media, 2010.