

# Recent Deep Learning in Crowd Behaviour Analysis: A Brief Review

Jiangbei Yue and He Wang\*

© 2025, Springer Nature

All rights reserved.

January, 2025

**Abstract** Crowd behaviour analysis is essential to numerous real-world applications, such as public safety and urban planning, and therefore has been studied for decades. In the last decade or so, the development of deep learning has significantly propelled the research on crowd behaviours. This chapter reviews recent advances in crowd behaviour analysis using deep learning. We mainly review the research in two core tasks in this field, crowd behaviour prediction and recognition. We broadly cover how different deep neural networks, after first being proposed in machine learning, are applied to analysing crowd behaviours. This includes pure deep neural network models as well as recent development of methodologies combining physics with deep learning. In addition, representative studies are discussed and compared in detail. Finally, we discuss the effectiveness of existing methods and future research directions in this rapidly evolving field. This chapter aims to provide a high-level summary of the ongoing deep learning research in crowd behaviour analysis. It intends to help new researchers who just entered this field to obtain an overall understanding of the ongoing research, as well as to provide a retrospective analysis for existing researchers to identify possible future directions.

## 1 Introduction

The research in crowd behaviour analysis aims to understand, recognise, and predict crowds in different situations, environments, events, *etc.* The behaviour arises from the aggregation of a group of people sharing the same physical space [Wijermans,

---

Jiangbei Yue  
School of Computer Science, University of Leeds, Leeds LS2 9JT, United Kingdom.

He Wang, corresponding author  
UCL Centre for Artificial Intelligence, Department of Computer Science, University College London, London NW1 2AE, United Kingdom  
e-mail: he\_wang@ucl.ac.uk

2011, Swathi et al., 2017]. It is a highly interdisciplinary and cross-disciplinary field, where a wide range of research topics have been investigated, from individual body movements to the overall crowd flow, from the low-level physical constraints to the high-level socio-psychological factors [Zhan et al., 2008, Kok et al., 2016, Murino et al., 2017]. Given its wide impact on a wide range of applications, *e.g.* safety, security, event organisation, and transportation, physicists, mathematicians, computer scientists, and psychologists have jointly explored this area for decades.

In this chapter, we focus on recent research in crowd behaviour analysis in computer science, specifically in deep learning. The emergence of deep learning has advanced the research in Artificial Intelligence (AI) to its ever-greatest today. As a universal computational machinery, deep neural networks have been employed to improve or even replace many research tools in many fields, including crowd behaviour analysis. Given the fast pace of the development of deep learning, it is particularly important to conduct a timely review of the most recent deep learning based research in this field, to summarise the up-to-date achievements, compare different approaches, and discuss the possible future directions of how crowd behaviour analysis can continue to leverage the cutting-edge AI technologies.

The application of deep learning in any field requires two key components: data and task definition. The former is the foundation of AI while the latter enables specific methods to be developed and metrics to be chosen for evaluation. As the history of deep learning in crowd behaviour analysis is relatively short, it covers only a small fraction of the problems that have been widely studied in other fields. This is probably mostly due to the lack of (big) data and probably also the unfamiliarity of the modelling tools to people outside of computer science and engineering. Currently, one of the first and most active research communities for crowd behaviour analysis is computer vision. It is not surprising that computer vision has become an active research field for understanding crowd behaviours. Video data is easy to obtain, and cameras as sensors are commonly accepted. In fact, cameras are not only often used but are the only sensors used in many places, especially in public and communal spaces, *e.g.* CCTV cameras, because of their non-invasiveness. Therefore, large amounts of video data have been obtained for crowds, which establishes the foundation for deep learning for crowd behaviour analysis.

In addition to abundant data, computer vision is also one of the first fields where crowd behaviour analysis tasks are defined for deep learning. So far, there are mainly two tasks: behaviour prediction and recognition. Therefore, in this chapter, we focus on reviewing the most recent research in these two subfields. Specifically, crowd behaviour prediction involves forecasting the future states or actions of a crowd, with or without details on the individuals in it, based on their observed behaviours [Fan et al., 2015, Alahi et al., 2016, Jiang et al., 2021]. This task emphasises modelling the evolving patterns in crowd movements and is essential for applications such as crowd management and autonomous vehicles. Crowd behaviour recognition, on the other hand, aims to identify the type of behaviours exhibited by a given crowd, focusing on analysing the patterns in crowds, which are often regarded as a whole [Khokher et al., 2014]. Such recognition can be applied in domains such as public safety and automatic surveillance [Khokher et al., 2014, Qaraqe et al., 2024].

Besides the tasks, there are still other factors to consider when classifying the current research in a review. An important factor is the type of model. In the past decade, new types of deep neural networks have been proposed one after another at a fast pace. Since there are unique advantages and disadvantages for each type of new network, it is unclear which type of network or what combination of them would be the best for crowd analysis. Therefore, we have seen wave after wave of research applying the latest deep learning model for crowd analysis. Therefore, we will also use the types of neural networks as one major factor to classify recent research in both tasks. Next, given the fast-growing number of papers being published in this field, it is impractical to dive into the details of each one of them. So we only give details of the methodology of some of the representative papers. Finally, since some of the deep learning research in this field is closely related to the traditional data-driven approaches, *i.e.* statistical machine learning, which also has been active in this field, we also briefly review these methods which are from approximately the same period of time.

## 2 Crowd Behaviour Prediction

Crowd behaviours can be studied from many perspectives at different levels [Thida et al., 2013, Yang et al., 2020]. One perspective is microscopic which focuses on individual-level movements, where various levels of details have been modelled, from simplifying individuals into 2D discs to modelling detailed 3D body motions [Van den Berg et al., 2008, Feldmann and Adrian, 2023, Gomez-Nogales et al., 2024]. After modelling the individuals, the collective motions are regarded as crowd behaviours [Yang et al., 2020]. In contrast with the microscopic level, there is also a macroscopic perspective. This type of research often treats the crowd as a whole *e.g.* as a continuum [Hughes, 2002, Golas et al., 2014].

Microscopic analysis is currently trendy in computer vision. One popular modelling perspective is to treat each individual as a 2D disc moving in a 2D environment. The required data are extracted from crowd videos and are normally in the form of trajectories, often in 2D but also occasionally in 3D. These trajectories are either human-labelled or automatically estimated via tracking algorithms [Pellegrini et al., 2009, Robicquet et al., 2016]. Trajectory data has been extensively used to understand individual behaviours and group flows in traditional statistical machine learning [Wang et al., 2016, He et al., 2020], as well as recent deep learning [Sighencea et al., 2021, Gu et al., 2022, Lin et al., 2024]. One core evaluation protocol for these methods is to see whether a model can actually predict the short-horizon and long-horizon future. This is evaluated by several metrics which are based on the difference between the predicted trajectories by the trained model and the ground-truth trajectories. This difference can be measured by trajectory-wise difference, *e.g.* accuracy, or distributions of trajectories, *e.g.* distributional divergences.

## 2.1 Traditional Machine Learning Methods

To predict trajectories, traditional machine learning methods typically rely heavily on feature engineering for a good representation of the crowd/individual state and their dynamics. Popular models include Gaussian processes [MacKay et al., 1998], Markov models [Fosler-Lussier, 1998], Kalman filters [Bishop et al., 2001], *etc.* Ellis et al. [2009] explicitly modelled the probabilistic distribution of the current velocities based on a Gaussian process prior. Given the current positions, instantaneous velocities are then sampled from the estimated distributions to infer the next positions. This process is performed recursively to predict the future trajectories. Kitani et al. [2012] propose a unified model based on hidden variable Markov decision processes. With the prior knowledge of people’s goals, they employed semantic scene labelling and inverse optimal control to model the people-people interaction and the people-environment interaction. BRVO [Kim et al., 2015] follows the reciprocal velocity obstacle [Van den Berg et al., 2008] formulation to build a prediction model, where the ensemble Kalman filter [Evensen, 2003] is exploited to optimise the relevant parameters. Although these methods require data to calibrate model parameters, the amount of data needed is significantly less than what is required for deep learning methods.

Besides individual trajectory prediction, statistical machine learning methods have also been employed to predict trajectories with similar distributions, *i.e.* not hinged on individual motions but collective group motions. However, this requires analysing the flow patterns of trajectories and developing relevant metrics. As labelling flows is too laborious to be practical for crowd trajectories, these methods are often based on unsupervised learning and have been developed to cluster trajectories to identify flows [Wang et al., 2011, 2016, 2017a, Wang and O’Sullivan, 2016, He et al., 2020]. Aiming to find natural clusters of similar behaviours in crowds, these methods automatically discover spatial similarities [Wang et al., 2011, 2016, 2017a], spatial-temporal similarities [Wang and O’Sullivan, 2016], or spatial-temporal-dynamics similarities [He et al., 2020]. Along with such analysis, new metrics have been proposed to measure whether prediction/simulation is similar to real crowds. These metrics are based on rules [Singh et al., 2009], or overall statistical similarities [Guy et al., 2012], scene semantics of flows [Wang et al., 2016, 2017a], *etc.*, so that prediction and simulation can be guided by these metrics [He et al., 2020].

Overall, there is prolific research in using traditional statistical machine learning for crowd behaviour prediction. In comparison with the recent deep learning methods reviewed below, these methods tend to require fewer data samples, employ more white-box models, and have better explainability of the prediction. However, when compared based on prediction accuracy, they are generally inferior to deep learning.

## 2.2 Deep Learning Methods

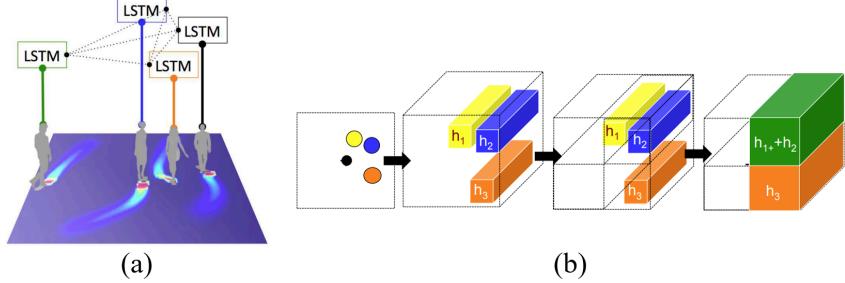
Deep learning methods have recently dominated the field of human trajectory prediction with exceptional prediction accuracy. Since neural networks are universal function approximators [Hornik et al., 1989, Hornik, 1991], they can learn complex and non-linear dynamics and patterns from data, especially when learning from a large amount of data. Despite being almost black-box, it significantly reduces the reliance on handcrafted features, hence less reliance on the expertise and time required for manual feature engineering. At the same time, the techniques for data collection, *e.g.* tracking algorithms in video analysis, have been fast developing, resulting in abundant data. As a result, this has significantly extended the basis of researchers on pedestrian and crowd analysis, especially in the field of computer vision and AI.

Following the development of deep learning itself, a broad set of powerful neural networks, each with its unique characteristics, have been proposed and adapted for human trajectory prediction. These models include Recurrent Neural Networks (RNNs) [Fausett, 1994], Convolutional Neural Networks (CNNs) [LeCun et al., 1989], Graph Neural Networks (GNNs) [Wu et al., 2020], generative models [Goodfellow et al., 2014, Sohn et al., 2015, Sohl-Dickstein et al., 2015], transformers [Vaswani et al., 2017], *etc.* Researchers focus on how to leverage diverse network architectures to capture motion dynamics. Therefore, we can broadly classify existing deep learning methods into RNN-based, CNN-based, GNN-based, generative model-based, and transformer-based methods according to the network architecture. Since this research community has been extremely prolific in the past decade, we will cover only some representative papers.

### 2.2.1 Recurrent Neural Networks

RNNs are a class of neural networks with neurons with an internal state. They can model sequential data with dependence by connecting their output with their input and changing the internal state accordingly, which captures the non-linear dependence between data at different time steps or different positions in a sequence. Specifically, given a sequence, RNNs utilize the same network repeatedly to process each element in the sequence, where the output for the current step is used as a part of the input at the next step. This design enables RNNs to capture temporal dependencies, which is vital for tasks such as human trajectory prediction.

Inspired by the success of RNNs in other sequence prediction tasks *e.g.* speech generation, Alahi et al. [2016] first applied RNNs in trajectory prediction and achieved superior prediction accuracy, which is the pioneering work in trajectory prediction based on deep learning. The authors proposed a new model called Social-LSTM, based on the Long Short Term Memory (LSTM) network [Hochreiter, 1997] (a classic RNN). One technical and modelling novelty in this method is a module which embeds a learnable component in the network to capture the interactions between pedestrians. With each trajectory captured by one LSTM for each pedestrian, a novel social pooling layer is introduced to model the interactions between pedes-



**Fig. 1** Overview of the Social-LSTM (a) and visualization of constructing social tensors (b). The black dot and other dots (yellow, blue and orange) denote the target person and neighbours, respectively, in (b). These two figures are from Alahi et al. [2016].

trians, by connecting the features of the neighbouring pedestrians learned by their LSTMs. The output of the social pooling layer at each step is fed back to the LSTM of the pedestrian in interest as part of the input for the next step. Since Social-LSTM marks the beginning of the new deep learning chapter in human trajectory prediction, a more detailed review of the method is below.

As shown in Fig. 1 (a), one LSTM is employed for each person to model their dynamics, where all LSTMs share the same weights. The LSTM takes as input the current location and interaction features to update its internal states. Specifically, the current location of the  $i$ th person ( $x_t^i, y_t^i$ ) at the time step  $t$  is transformed into the embedding feature  $e_t^i$  by a multi-layer perceptron (MLP). Meanwhile, the social pooling builds a grid near the person in interest, as shown in Fig. 1 (b), and constructs a  $N_o \times N_o \times D$  social tensor  $H_t^i$ :

$$H_t^i(m, n, :) = \sum_{j \in N_i} \mathbf{1}_{mn}[x_t^j - x_t^i, y_t^j - y_t^i] h_{t-1}^j, \quad (1)$$

where  $N_o$  is the neighbourhood size,  $D$  is the dimension of the hidden states  $h_t$  of the LSTM,  $\mathbf{1}_{mn}[x, y]$  is an indicator function showing if  $(x, y)$  is located in the  $(m, n)$  cell of the grid,  $N_i$  denotes the set containing neighbours of the  $i$ th person. Whether a pedestrian is considered a neighbour of the pedestrian in interest is based on their distance at the current time. When the distance between the  $i$ th and the  $j$ th person falls below a pre-defined threshold, the  $j$ th person is considered a neighbour of the  $i$ th person. Note the construction of social tensors preserves the spatial information.

Next, the social tensor  $H_t^i$  is then fed into an MLP to obtain the corresponding embedding feature  $a_t^i$  (Eq. (3)). Subsequently, the LSTM takes as input the last hidden state  $h_{t-1}^i$  and embedded features  $(e_t^i, a_t^i)$  to output the new hidden state (Eq. (4)). Overall, the whole process is formulated as the following recurrence:

$$e_t^i = MLP_e(x_t^i, y_t^i), \quad (2)$$

$$a_t^i = MLP_a(H_t^i), \quad (3)$$

$$h_t^i = LSTM(h_{t-1}^i, e_t^i, a_t^i), \quad (4)$$

where the initial hidden state  $h_0$  is given.  $e_t^i$  is a feature of location  $(x_t^i, y_t^i)$ .

Lastly, the hidden states  $h_t^i$  are used to estimate the distribution of the next position  $(x_{t+1}^i, y_{t+1}^i)$ . Social-LSTM assumes that the next position follows a bivariate Gaussian distribution, which has five parameters: the mean  $\mu_{t+1}^i = (\mu_x, \mu_y)_{t+1}^i$ , the standard deviation  $\sigma_{t+1}^i = (\sigma_x, \sigma_y)_{t+1}^i$ , and the correlation coefficient  $\rho_{t+1}^i$ . The hidden state  $h_t^i$  is fed into an MLP to predict the distribution parameters of the next position:

$$[\mu_{t+1}^i, \sigma_{t+1}^i, \rho_{t+1}^i] = MLP_p(h_t^i). \quad (5)$$

Finally, one can sample  $(\hat{x}_{t+1}^i, \hat{y}_{t+1}^i)$  from the predicted distribution  $\mathcal{N}(\mu_{t+1}^i, \sigma_{t+1}^i, \rho_{t+1}^i)$ .

Since Social-LSTM explicitly parameterises the distribution of the next position as a Gaussian, assuming conditional independence between different time steps and between different pedestrians, one can easily write down the joint likelihood of all observed pedestrian positions as a product of the likelihood of individual positions, given their respective Gaussian parameters. Therefore, Social-LSTM optimises the model parameters by minimising the negative log-likelihood:

$$L = - \sum_i \sum_{t=T_h+1}^{T_h+T_f} \log(P(x_t^i, y_t^i | \mu_t^i, \sigma_t^i, \rho_t^i)), \quad (6)$$

where the first summation involves all pedestrians,  $T_h$  is the length of the history trajectory, and  $T_f$  is the length of the future trajectory. Overall, one of the key novelties in Social-LSTM is the learnable interactions between pedestrians embedded into the design of the RNN. Starting from Social-LSTM, a line of RNN-based methods has been proposed.

Xue et al. [2018] proposed a hierarchical LSTM network named Social-Scene-LSTM based on an encoder-decoder framework, to incorporate the environment factor which is ignored by Social-LSTM. Social-Scene-LSTM uses three LSTM encoders for three scales (person, social, and scene) and an LSTM decoder to estimate future trajectories. The model considers the observed individual trajectories, the positions of the neighbours, and the frames of the video at every time step. These features correspond to the single-person, the social, and the scene scales, respectively. The LSTM encoder outputs the corresponding features at different scales. These features are concatenated and fed into the LSTM decoder to obtain the predicted trajectories. Later, in contrast to the Social-Scene-LSTM which considers the influence of the environment by simply encoding the frames, context-aware LSTM [Bartoli et al., 2018] followed the framework of Social-LSTM and proposed a novel context-aware pooling layer to model the interactions between people and objects in the environment. Zhang and Zheng [2021] used Gated Recurrent Units (GRU) [Chung et al., 2014] (also an RNN) to replace the LSTM in Social-LSTM to avoid overfitting.

They also emphasized that the encoded representation of the trajectory positions is more effective than the raw positions when serving as the input to the GRUs. Given the importance of destinations in trajectory prediction, Tran et al. [2021] proposed a dual-channel neural network, consisting of a goal channel and a trajectory channel, based on GRUs. Specifically, the goal-channel network employs GRUs to extract features from the estimates of multiple possible destinations. The model converts the extracted features into a series of control signals by a flexible attention mechanism. Finally, the trajectory channel network predicts the future trajectories under the guidance of the control signals.

In addition to the methods reviewed above, there are abundant other RNN-based methods with various improvements. There is a series of methods which improve the performance of the social pooling mechanism by considering the motion coherence of groups [Bisagno et al., 2018, 2021], modelling the importance or attention to neighbours [Fernando et al., 2018, Xu et al., 2018b, Shi et al., 2019], adding the repulsion constraint [Xu et al., 2018a], *etc*. Moreover, Zhang et al. [2019] designed a state refinement module based on a message-passing framework to enhance the representation of hidden states of LSTMs and optimize interactions. Zhu et al. [2019] proposed the StartNet with a star topology to model interactions between individuals. Liang et al. [2019] presented a multi-task learning framework, modelling predictions of activities and trajectories jointly.

### 2.2.2 Convolutional Neural Networks

CNN (Convolutional Neural Network) is a type of feed-forward neural network with convolution layers, which excels in capturing the correlations between the features of a location with its neighbours. This type of network has been employed to learn from data with strong temporal and spatial correlations and provide high computational efficiency, compared with *e.g.* multilayer perceptron networks. Therefore, many methods incorporate CNNs to predict pedestrian trajectories.

The key to applying CNNs is to organise the data or features so that the concept of vicinity can be defined. As one of the earliest methods applying CNNs for trajectory prediction, Nikhil and Tran Morris [2018] encoded each position of the past trajectory to obtain the embedding features and organise these features in an image-like form. Then, a CNN is used to apply convolutions on these features and outputs a global feature across all the historical positions. Lastly, the global feature is fed into a fully connected layer to predict future trajectories.

Zamboni et al. [2022] extended the CNN model in Nikhil and Tran Morris [2018] by building different mappings between the global feature and each of the frames in the future trajectory. This is done by dividing the output of the CNN in [Nikhil and Tran Morris, 2018] into N features instead of a single global feature, where N is the number of time steps in the future trajectory. These N features are separately fed into a fully connected layer to predict their corresponding future positions. In addition, the authors explored the effectiveness of 1D convolution, 2D convolution, positional

embeddings, transpose convolutional layers, and residual connections in CNNs for trajectory prediction.

Next, to further improve the modelling of the interactions between pedestrians, Social-IWSTCNN [Zhang et al., 2021] introduces a new social interaction extractor module based on MLPs and an aggregate function to extract features containing social and spatial information from observed trajectories. Subsequently, a temporal convolutional network is proposed to encode the interaction features for refinement by further incorporating the temporal information. The refined features are fed into another CNN to predict distributions of future trajectories.

Zhang et al. [2022] proposed to combine CNNs with RNNs to improve the performance in trajectory prediction. Specifically, they used two LSTMs to encode the input trajectory first to extract the time-related features for each time step. Then, every feature is reshaped and fed as input into a CNN to mine high-dimensional patterns within it. The outputs of the CNN are further enhanced based on an augmented attention mechanism to capture the global information. Then a GRU followed by an MLP takes the enhanced features as the input to predict the next position. The proposed model recursively estimates the future trajectory.

Beyond the papers above, the idea of combining CNNs with RNNs is also utilised by many other people [Ridel et al., 2020, Song et al., 2020, de Brito et al., 2021, Shafiee et al., 2021, Jain et al., 2020].

### 2.2.3 Graph Neural Networks

GNNs are a class of neural networks specifically designed to process data structured as graphs defined by a set of nodes and a set of edges. They have a wide range of applications where the data does not lie in the Euclidean space. Originally, a trajectory is naturally regarded as 2D Euclidean data. However, a group of trajectories from different pedestrians as a whole can be modelled differently. The introduction of GNNs brings an alternative view of them. Trajectories from pedestrians together can be regarded as a graph, *e.g.* if each pedestrian is viewed as a node and the edges between nodes are used to model the interactions between pedestrians. This way, GNNs have the ability to effectively model the spatial, temporal, and relational dynamics in complex graphs [Yan et al., 2018, Wu et al., 2020]. This has motivated various GNN-based methods for trajectory prediction.

Huang et al. [2019] adopted graph attention networks [Veličković et al., 2018] and LSTMs to extract information from spatio-temporal graphs constructed based on past trajectories and predict future motions. Social-STGCNN [Mohamed et al., 2020] constructs spatio-temporal graphs  $G$  consisting of a set of spatial graphs  $G_t = \{V_t, E_t\}$  in all the past time steps to represent the historical trajectories, where nodes in  $V_t$  denote pedestrians, edges in  $E_t$  connect a pedestrian with their neighbours. Each node in  $V_t$  has attributes such as locations and each edge in  $E_t$  is associated with weights describing the interaction intensity *e.g.* whether two pedestrians are close so that they need to steer clear of each other. Subsequently, the proposed spatio-temporal graph convolution neural network takes the spatio-temporal graphs

as input to extract the spatio-temporal features. Finally, time-extrapolator CNNs are used to predict the future trajectories based on the extracted features.

The concept of spatio-temporal graphs is further employed in many other methods [Wang et al., 2021, Zhou et al., 2021, Lian et al., 2023, Sighencea et al., 2023]. DAG-Net [Monti et al., 2021] uses two graph neural networks to estimate the destinations of pedestrians and model the interactions between people. Then a variational RNN [Chung et al., 2015] is employed to incorporate the information of the estimated destinations and the interactions, to obtain the future trajectories by recursively predicting the next position. Shi et al. [2021] argued that previous research typically models redundant interactions between pedestrians. To address the problem, they proposed a sparse graph convolution network model. Specifically, the model introduces a new sparse graph learning based on the self-attention mechanism [Vaswani et al., 2017]. It also uses asymmetric convolution networks to construct sparse spatial graphs with sparse directed interactions, and sparse temporal graphs with motion tendencies, from the input trajectories. Subsequently, graph convolution networks [Kipf and Welling, 2016] are used to extract the features from these sparse graphs. The learned features are then fed into a time convolution network to estimate the future trajectories.

#### 2.2.4 Generative Models

Generative models are machine learning models that can generate similar data to the data on which they are trained. These models normally achieve this by learning a mapping between the unknown data distribution and some simpler distribution where the sampling is easier, *e.g.* a Gaussian or uniform distribution. In deep learning, these methods include Generative Adversarial Networks (GANs) [Goodfellow et al., 2014], Variational Autoencoders (VAEs) [Kingma, 2013], diffusion models [Sohl-Dickstein et al., 2015], *etc.* One advantage of generative models is that they are suitable for modelling randomness in non-deterministic processes, which makes them useful in trajectory prediction. This is because there is intrinsic randomness in pedestrian trajectories due to the influence of various unobserved factors *e.g.* environment and affective states [Luo et al., 2008, Itkina and Kochenderfer, 2023]. Therefore, the mapping from some historical trajectory to the future trajectory can be seen as a stochastic process where generative models can be employed.

GANs consist of a generator network and a discriminator network. The generator is trained to generate synthetic samples resembling real samples, while the discriminator is trained to tell the difference between synthetic and real samples. The adversarial training is for the generator to be able to fool the discriminator, and for the discriminator to be able to distinguish synthetic samples from the real ones, until an equilibrium is reached. Social-GAN [Gupta et al., 2018] is one of the earliest research papers which applies GANs to trajectory prediction. In Social-GAN, the generator is an LSTM-based encoder-decoder network. The discriminator is simply an LSTM-based encoder followed by a classification layer. Given the past trajectories and the noises drawn from a distribution, the generator is trained to predict future

trajectories. Additionally, Social-GAN designs a novel pooling mechanism applied to the generator to model the interactions between individuals. It also employs a new adversarial loss function to obtain diverse predictions which are consistent with the observed trajectories. The initial success of Social-GAN has inspired other similar research. SoPhie [Sadeghian et al., 2019] is a GAN-based method incorporating the restriction of the environment *e.g.* walls, obstacles. The model first leverages CNNs to extract the environment features from the video frames and employs LSTMs to extract the movement features for each person. Then, an attention module is used to capture the physical attention and the social attention, to optimize the environment features and the motion features, respectively. Then both features are concatenated for each person. These concatenated features are finally fed into an LSTM-based GAN to estimate the distributions of the future trajectories. In addition, other methods also introduce different variants of GANs to model crowd dynamics. Amirian et al. [2019] employed Info-GAN [Chen et al., 2016] to avoid mode collapsing and dropping in naive GANs while Kosaraju et al. [2019] utilised Bicycle-GAN [Zhu et al., 2017] to improve the modelling of multi-modal trajectories.

Other than GANs, another type of generative model, VAEs, has also led to a stream of new methods for trajectory prediction. The key idea of VAEs is to assume there is a latent space where the distribution of data is a Gaussian, so that an encode-decoder network can be trained where the encoder maps the data into the latent space and the decoder reconstructs the data from the latent space. After training, samples can be drawn from the Gaussian and then decoded into outputs which are similar to the training data. Often, the encoder and decoder also take as input an additional condition, where it becomes a Conditional VAE [Sohn et al., 2015], or CVAE. PECNet [Mangalam et al., 2020] uses a CVAE to predict the distributions of destinations based on the observed trajectories, which is further used to guide the prediction of future trajectories through a new social pooling mechanism. Here, the condition of the CVAE is the past trajectory. Lee et al. [2022] also employed a CVAE to estimate destinations but another CVAE was used to generate full trajectories. SocialVAE [Xu et al., 2022b] combines a CVAE with a RNN. In this work, the encoder and decoder in the CVAE are implemented as two GRUs to reconstruct the future trajectories directly, while another GRU extracts the features from the historical trajectories to serve as the condition of the CVAE. Yu et al. [2024] modified SocialVAE by using complex gated recurrent units [Xu et al., 2023] as the encoder and decoder in the CVAE and a graph attention network to extract condition features. Zhou et al. [2022] combined CVAEs with GANs to predict trajectories. It adopts the architecture of GANs, with a CVAE-based generator and an RNN-based discriminator.

Recently, diffusion models have gained significant interest in many domains such as image generation, due to their strong generative capabilities. Diffusion models generate data by first adding noises to the data so that they can be seen as samples from a Gaussian. Then the model learns to denoise these noise-polluted samples from the Gaussian to reconstruct the data samples. Once learned, a mapping between a Gaussian and the data distribution is established. Again, this positions diffusion models as natural candidates for modelling stochasticity, which has inspired researchers to use

diffusion models in stochastic trajectory prediction. Gu et al. [2022] is one of the first to do so. In Gu et al. [2022], the denoising process is constructed from all walkable areas to the desired trajectory distributions within some limited areas. Additionally, the past trajectories are encoded into state embeddings serving as conditions of the denoising process. Although the model in Gu et al. [2022] can balance the diversity and the accuracy of prediction by controlling the denoising steps, they typically need a large number of denoising steps, which is time-consuming and hinders its application. To address the issue, Mao et al. [2023] proposed a new diffusion-based model with a leapfrog initialiser. Instead of denoising from all walkable areas, the model denoises from the trajectories initialised by the leapfrog initialiser, which significantly decreases the required number of denoising steps and accelerates the inference process. Yang et al. [2024] built a motion memory bank by clustering trajectories in the training set as prior guidance for the diffusion model. Motion patterns and target distributions in the bank are retrieved based on the proposed addressing mechanism to guide the diffusion model for prediction.

### 2.2.5 Transformers

Transformers utilise attention mechanisms to learn correlations, which can capture global information and long-range dependencies effectively. For sequence data, transformers are often viewed as competitive rivals of more traditional neural network such as RNNs. This drives the application of transformers in human trajectory prediction. Giuliari et al. [2021] proposed new transformer networks for trajectory forecasting, challenging the dominance of LSTM models in the field. The authors explored two variants: the original transformer network [Vaswani et al., 2017] and bidirectional transformers [Devlin, 2018], focusing on predicting individual trajectories without modelling human-human or human-scene interactions. Despite this simplicity, their transformer-based models outperformed more complex techniques, achieving state-of-the-art results on benchmarks such as Trajnet [Sadeghian et al., 2018] and ETH/UCY [Pellegrini et al., 2009, Lerner et al., 2007].

To simultaneously model the temporal and the social dimensions in trajectories, Yuan et al. [2021] introduced a novel transformer model, AgentFormer, with an innovative agent-aware attention mechanism to preserve the personal identity. The AgentFormer enables direct interactions between the agents' states across different time steps, enhancing long-range dependency modelling. Further, the authors presented a stochastic trajectory prediction model based on the AgentFormer following the scheme of CVAEs, fostering socially-aware and diverse trajectory generation. Specifically, the past trajectories are encoded by the AgentFormer encoders to provide conditions, while the AgentFormer decoders map the future trajectories onto the latent space and decode the latent variables to future predictions. Geng et al. [2022] designed a new attention mechanism by introducing a temporal attention module and a spatial attention module to capture motion features.

Li et al. [2022] introduced a graph-based transformer for stochastic trajectory prediction. Multi-scale graphs are first established given the history trajectories and

the scene segmentation maps. Subsequently, an encoder consisting of convolutional LSTMs (ConvLSTMs) [Shi et al., 2015] extracts the features from these graphs to model motion patterns. Finally, a decoder uses the encoded features to predict the future trajectories. During decoding, the graph-based transformer, which models the human-to-human and the human-to-scene spatial interactions, estimates the distributions of the next positions. To ensure temporal consistency, the authors further proposed the Memory Replay module based on a memory graph, smoothening the estimates from the graph-based transformer. In addition, there are other examples where transformers are also combined with graphs such as Yu et al. [2020], Liu et al. [2023, 2024].

TUTR [Shi et al., 2023] unifies social interactions and multimodal trajectory forecasting within a transformer encoder-decoder architecture. Unlike prior approaches [Xu et al., 2022a,b] that rely on computationally expensive post-processing techniques, TUTR eliminates this step by introducing explicit global prediction. This is accomplished by generating common motion modes among pedestrians and using a mode-level transformer encoder to parse diverse motion patterns. The mode-level encoder takes as input the encoded past trajectories which are concatenated with the motion modes, to model the multimodality of the future trajectories. Then, there are neighbour embeddings which are passed through a social-level transformer decoder to capture the social interactions. Finally, the features from the transformer encoder and decoder are fed into a dual prediction network model to estimate multiple future trajectories and their corresponding probabilities. In a similar line of research, Chen et al. [2022] used estimated goals as guidance to capture multimodal trajectories instead of motion modes in Shi et al. [2023] and proposed a goal-conditioned transformer for trajectory prediction.

PPT [Lin et al., 2024] is a novel progressive pretext task learning framework based on transformers. The framework is designed to progressively enhance the model’s ability through three tasks to capture the short-term dynamics and the long-term dependencies in pedestrian trajectory prediction. In Task 1, a transformer encoder is pre-trained to predict the next position for a trajectory of an arbitrary length to understand the short-term motion patterns. Task 2 aims to enhance the long-term dependency modelling of the transformer encoder by predicting diverse trajectory destinations. The pre-trained transformer encoder is then duplicated as two separate models: one dedicated to destination prediction and the other focused on intermediate position prediction in Task 3. The final prediction model involving the two separate models is trained by forecasting the entire future trajectories.

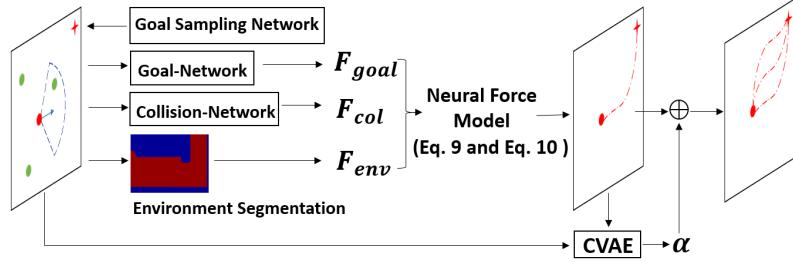
### 2.3 Physics-Inspired Deep Learning Methods

Physics-inspired deep learning methods are a relatively new research trend in this field. The motivation of such research is to combine the advantages of both the traditional empirical non-data-driven methods and the aforementioned deep learning methods. Before statistical machine learning, there was a large body of research re-

lying on empirical modelling. It tends to be empirical or rule-based methods derived via the first-principles approach: summarising observations into rules and deterministic systems based on certain fundamental assumptions on human motions. In such a perspective, social interactions can be modelled as forces in a particle system [Helbing and Molnar, 1995] or an optimization problem [Van den Berg et al., 2008], and individuals can be influenced by affective states [Luo et al., 2008]. Later, data-driven methods were introduced, in which the model behaviour is still dominated by the assumptions on the dynamics, e.g. a linear dynamical system [He et al., 2020]. These methods have good explainability but normally lack prediction accuracy. In contrast, in all the previously mentioned deep learning methods, the prediction accuracy is massively improved. However, due to the black-box nature of deep neural networks, one key modelling effort is to decide the type/architecture of the network to capture as explicitly as possible certain aspects of pedestrian dynamics, e.g. the environmental constraints, social interactions, and intrinsic stochasticity. To this end, since both explainability and accuracy are crucial in applications related to trajectory prediction, e.g. autonomous driving, physics-inspired deep learning methods started to become popular since the year of 2022. One major modelling effort is to embed various physical systems that can best describe a group of pedestrians into deep learning for trajectory prediction. Due to the fact that this field is new, we review some recent research in detail and classify them based on the physics system they employ.

### 2.3.1 Particle Systems with Deep Learning

It is known that relatively sparse crowds can be described by particle systems in the classic work of the Social Force Model (SFM) [Helbing and Molnar, 1995]. To make this kind of physical system fit data better, Yue et al. [2022] designed a new framework, Neural Social Physics (NSP), based on neural differential equations [Chen et al., 2018, Kidger, 2021], combining deep learning with particle systems. In NSP, they proposed a new model NSP-SFM, which embeds the traditional SFM with learnable parameters into a deep neural network. This allows the model to possess explainability (from the physics component) and maintain outstanding prediction performance (from deep learning). NSP achieved state-of-the-art prediction accuracy on public datasets (SDD [Robicquet et al., 2016], ETH [Pellegrini et al., 2009], and UCY [Lerner et al., 2007]), which are the most widely used datasets in computer vision in this area starting from Social-LSTM. Additionally, NSP is capable of providing human-understandable explanations for its predictions, which is significantly different from earlier pure deep learning methods. Following NSP, other methods with similar perspectives have been proposed, such as [Mo et al., 2024] which integrates physics models into neural networks with the support of symbolic regression [Schmidt and Lipson, 2009] for trajectory prediction, or [Yue et al., 2023] which integrates a Bayesian approach into the social force models with deep neural networks. Since NSP is one of the first papers using physics as a prior for trajectory prediction, we provide more details below.



**Fig. 2** Overview of NSP-SFM. This figure is from Yue et al. [2022].

NSP uses a function  $q(t)$  to represent a trajectory, where  $q^t = [p^t, \dot{p}^t]^T$  consisting of the position  $p^t \in \mathbb{R}^2$  and velocity  $\dot{p}^t \in \mathbb{R}^2$ . Then, NSP models the dynamics of pedestrians as follows:

$$\frac{dq}{dt}(t) = f_{\theta, \phi}(t, q(t), \Omega(t), q^{t_h+t_f}, E) + \alpha_\phi(t, q^{t:t-M}), \quad (7)$$

where  $f$  and  $\alpha$  are neural network models with learnable parameters  $\theta$  and  $\phi$ . The  $\theta$  denotes explainable parameters, such as physics-related variables while the  $\phi$  represents unexplainable parameters, such as neural network weights. In NSP,  $f$  governs the dynamics by considering current states  $q(t)$ , neighbours  $\Omega(t)$ , destinations  $q^{t_h+t_f}$ , and the environment  $E$ . To capture stochasticity in trajectories, NSP introduces  $\alpha$  which depends on the brief history  $q^{t:t-M}$ . Using Taylor's expansion on  $q(t)$  gives:

$$q(t + \Delta t) \approx q(t) + \dot{q}(t)\Delta t = \begin{pmatrix} p(t) \\ \dot{p}(t) \end{pmatrix} + \Delta t \begin{pmatrix} \ddot{p}(t) + \alpha \\ \ddot{p}(t) \end{pmatrix}, \quad (8)$$

where  $\Delta t$  represents the time interval, the  $\alpha$  is presumed to solely affect the velocity, and  $p(t)$  is assumed to possess the second-order differentiability. The generality of Eq. 8 enables it to incorporate any physics systems with second-order differentiability. The authors introduced the new model NSP-SFM within NSP by integrating the social force model.

NSP-SFM focuses on modelling  $\ddot{p}(t)$  by considering three factors: goals, neighbours and the environment. As a result, we have:

$$\ddot{p}(t) = F_{goal} + F_{col} + F_{env}, \quad (9)$$

where  $F_{goal}$ ,  $F_{col}$ , and  $F_{env}$  denote the goal attraction, inter-agent repulsion and environmental repulsion, respectively.

We show the overview of NSP-SFM in Fig. 2. The model estimates  $F_{goal}$ ,  $F_{col}$ , and  $F_{env}$  based on neural networks at each time step. The Goal Sampling Network predicts destinations used in the estimation of  $F_{goal}$  given the input trajectories and the environment. Subsequently, NSP-SFM uses Eq. 9 to calculate  $\ddot{p}(t)$  and adopts a semi-implicit scheme to update positions:

$$\dot{p}^{t+1} = \dot{p}^t + \Delta t \ddot{p}^t; \quad p^{t+1} = p^t + \Delta t \dot{p}^{t+1}. \quad (10)$$

Additionally, NSP-SFM introduces stochasticity  $\alpha$ :

$$p^{t+1} = p^t + \Delta t(\dot{p}^{t+1} + \alpha^{t+1}), \quad (11)$$

where a CVAE is designed to predict  $\alpha$ . Conditioned on the historical trajectory, this CVAE takes as input the current intermediate prediction of the position and outputs  $\alpha$  which is a learned random perturbation on the intermediate prediction, which gives the final prediction.

The modelling with forces inherently endows NSP-SFM with explainability. NSP-SFM retains the physical modelling in SFM and uses neural networks to predict relevant parameters which otherwise would need hand-tuning. Specifically, NSP-SFM models the goal attraction as:

$$F_{goal} = \frac{1}{\tau}(v_{des}^t - p^t); \quad \tau = NN(q^t, p^{th+tf}), \quad (12)$$

where  $v_{des}^t = \frac{p^{th+tf} - p^t}{(T-t)\Delta t}$  is the desired velocity and  $NN$  denotes neural networks. Given that the target agent  $i$  has the neighbour set  $\Omega_i$ , the estimation of the inter-agent repulsion is formulated as follows:

$$F_{col}^i = \sum_{j \in \Omega_i} F_{col}^{ij} = \sum_{j \in \Omega_i} -\nabla_{r_{ij}} \mathcal{U}_{ij} (\|r_{ij}\|) = \sum_{j \in \Omega_i} -\nabla_{r_{ij}} r_{col} k_{ij} e^{-\|r_{ij}\|/r_{col}}, \quad (13)$$

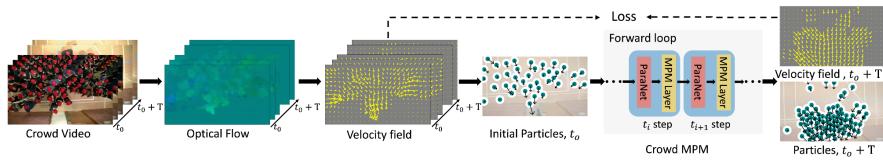
where  $r_{ij} = p_i^t - p_j^t$ ,  $\mathcal{U}_{ij} (\|r_{ij}\|)$  is a repulsive potential field,  $r_{col}$  is a hyper-parameter, and  $k_{ij} = NN(q_i^t, q_j^t)$ . NSP-SFM estimates the environmental repulsion as follows:

$$F_{env} = \frac{k_{env}}{\|p_i^t - p_{obs}\|} \left( \frac{p_i^t - p_{obs}}{\|p_i^t - p_{obs}\|} \right), \quad (14)$$

where  $k_{env}$  is a learnable parameter and  $p_{obs}$  represents the positions of obstacles in the environment.

With an explicit SFM and its key parameters predicted by neural networks, NSP-SFM achieves far more accurate prediction than back then existing methods. Also, since the learn SFM is essentially a simulator, NSP-SFM can simulate more pedestrian behaviours even if the scenario is drastically different from the training data, *e.g.* much higher densities, showing superb generalisability.

The NSP framework is flexible in the sense that the physics component is replaceable depending on what types of crowds, *e.g.* density, are of interest. For instance, Wang et al. [2024] directly estimated the final accelerations through a transformer structure and updated future trajectories based on Newton's laws of motion. Sang et al. [2024] extended Wang et al. [2024] by introducing a diffusion model to smooth predicted trajectories and a probabilistic selection module. Another instance is the particle system in NSP can be alternatively described by the time evolvement of its



**Fig. 3** The overview of the method proposed in He et al. [2025]. This figure is from He et al. [2025].

energy, *e.g.* potential and kinetic, and the system tends to minimise the energy [Xiang et al., 2024].

### 2.3.2 Continuum with Deep Learning

Although particle systems have been applied successfully in modelling relatively sparse crowds, they are not well-suited for capturing the dynamics of dense crowds. This is because the movement of dense crowds shares significantly more similarities with continuum systems such as water, compared with particle systems [van Toll et al., 2020]. Therefore, it is popular to model dense crowds as continuum. Most recently, the combination of continuum dynamics and deep learning has just started to emerge in the research on dense crowds. One example is Zhou et al. [2024] who proposed a hydrodynamic model for crowd simulation based on the similarities between dense crowds and fluids and the unique physical and social characteristics of crowds. In their approach, the governing equation controls the crowd movement. To efficiently solve the governing equation, they designed the hydrodynamics-informed neural network for the next-step prediction.

Instead of seeing dense crowds as passive continuum, He et al. [2025] represented high-density crowds as active matter, a continuum system composed of active particles. They proposed a novel neural differential equation model called CrowdMPM to learn crowd dynamics from in-the-wild videos. The overview of the system is shown in Fig. 3. Initially, optical flows are extracted from the given video and are utilised to generate velocity fields. They can also obtain the initial individual positions (initial particles) in crowds. The proposed CrowdMPM then estimates the subsequent crowd movement. After training, the model is capable of predicting and simulating dense crowd dynamics. Furthermore, this method mitigates the scarcity of dense crowd data by utilising in-the-wild crowd videos. In these videos, tracking individuals in high-density crowds is inherently challenging. As a result, individual trajectories are not available, so that other information such as density distributions and velocity fields need to be employed. Nevertheless, the active matter modelling enables the method in He et al. [2025] to successfully simulate or predict fine-grained movements of dense crowds, including individual trajectories. Since CrowdMPM He et al. [2025] is the first method of its kind, we would like to give more details about the method.

CrowdMPM follows the scheme of the standard material point method (MPM) [Jiang et al., 2016] while considering the features of high-density crowds, where high-density crowds are regarded as active matters. Therefore, similar to MPM, CrowdMPM is also a hybrid Eulerian-Lagrangian method, modelling dense crowd dynamics by combining strengths from the Eulerian and the Lagrangian views. Following the standard MPM, CrowdMPM begins with two governing equations

$$\frac{D\rho}{Dt} + \rho \nabla \cdot v = 0 \quad \text{Conservation of mass} \quad (15)$$

$$\rho \frac{Dv}{Dt} = \nabla \cdot \sigma^{cm} + \rho b + f^{act} \quad \text{Conservation of momentum} \quad (16)$$

where  $\frac{D}{Dt}$  denotes the material derivative,  $\rho$  represents the density of the continuum,  $v$  denotes the velocity field,  $\sigma^{cm}$  is the proposed crowd material stress,  $\rho b$  is the body force, and  $f^{act}$  is the proposed stochastic active force. Eq. (15) and Eq. (16) denote conservation of mass and conservation of momentum, respectively. CrowdMPM is a neural differential equation model since it solves Eq. (15) and Eq. (16) by employing differential operations based on neural networks to estimate crowd dynamics.

In accordance with the standard MPM, crowdMPM solves the governing equations through a three-phase framework: 1. Particle-to-Grid transfer (P2G), 2. Grid Operations (GO), and 3. Grid-to-Particle transfer (G2P). CrowdMPM models each individual in the crowd as a particle in the Lagrangian perspective, while discretising the space using grids from the Eulerian viewpoint. In the first phase, P2G, they transfer mass and momentum from the particles to the grid to leverage the stability and computational efficiency of the Eulerian grid for solving the governing equations. Specifically, at time step  $n$ , the P2G is computed as follows:

$$m_i^n = \sum_p w_{ip}^n m_p, \quad m_i^n v_i^n = \sum_p w_{ip}^n [m_p v_p^n + m_p C_p^n (x_i - x_p^n)], \quad (17)$$

where  $m_i$  ( $m_p$ ),  $v_i$  ( $v_p$ ),  $x_i$  ( $x_p$ ) denote the mass, velocity and position of the grid node  $i$  (particle  $p$ ),  $w_{ip}$  represents weights, and  $C_p$  is the affine velocity gradient. The time superscript  $n$  of  $m_p$  and  $x_i$  are omitted because the mass of particles does not change over time, and the Eulerian grid is reconstructed at each iteration. The weights are calculated as  $w_{ip}^n = \phi(x_i - x_p^n)$ , where  $\phi$  is a quadratic B-spline function.

Subsequently, the governing equations are solved to update the velocity on nodes in the GO phase:

$$\hat{v}_i^{n+1} = v_i^n + \Delta t \frac{f_i^n}{m_i}, \quad (18)$$

$$v_i^{n+1} = BC(\hat{v}_i^{n+1}) = \hat{v}_i^{n+1} - \gamma N \langle N, \hat{v}_i^{n+1} \rangle, \quad (19)$$

where  $\Delta t$  represents the time interval,  $f_i$  is the force on the node  $i$ ,  $BC$  denotes boundary conditions,  $N$  denotes the surface normal vector of the boundary, and  $\gamma$  is a scalar.  $f_i$  is calculated as:

$$f_i = f_i^{st} + f_i^{act} + f_i^{bd}, \quad (20)$$

where  $f_i^{st}$ ,  $f_i^{act}$ , and  $f_i^{bd}$  are the surface force (stress-based), active force and body force, respectively. Velocities updated via Eq. (18) are then corrected by imposing boundary conditions as in Eq. (19).

Finally, G2P transfers the updated grid results back to the particles, updating their velocities and positions. Specifically, the velocities and positions of the particles for the next step are calculated using the following equations:

$$v_p^{n+1} = \sum_i w_{ip}^n v_i^{n+1}, \quad x_p^{n+1} = x_p^n + \Delta t v_p^{n+1}, \quad (21)$$

Additionally, in this phase, the affine velocity gradient  $C_p$  and the deformation gradient  $F_p$  are updated for the next iteration as follows:

$$C_p^{n+1} = \frac{4}{\Delta x^2} \sum_i w_{ip}^n v_i^{n+1} (x_i - x_p^n)^T, \quad F_p^{n+1} = (I + \Delta t C_p^{n+1}) F_p^n, \quad (22)$$

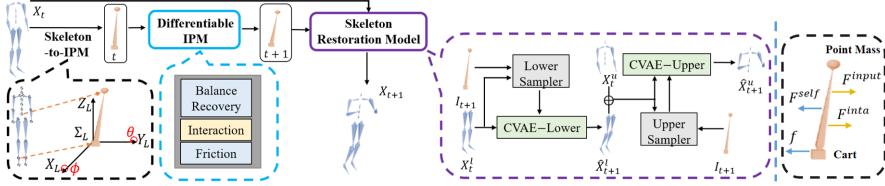
where  $\Delta x$  is the cell size of the grid, and  $I$  is the identity matrix. Note that the deformation gradient  $F_p$  is used in GO to estimate  $f_i^{st}$ .

So far, what has been introduced above is a standard MPM procedure. To modify it to model high density crowds as active matters, CrowdMPM has two key technical novelties. This first technical novelty is a new stress-strain model customised for high density crowds. CrowdMPM designs new  $f_i^{st}$ ,  $f_i^{act}$ , and  $f_i^{bd}$  to capture the characteristics of high-density crowds, in Eq. 20. The surface force  $f_i^{st}$  models three distinctive properties of dense crowds that differ from common materials: elastic asymmetry, exponential resistance and compression dominance.

Another key technical novelty of CrowdMPM is that it further incorporates active forces to represent the self-propelled nature of individuals within crowds. For this purpose, the Toner-Tu equation [Toner and Tu, 1995] is introduced to represent the external force  $f_i^{act}$  in Eq. 20. Since it contains parameters, neural networks are proposed to estimate these parameters during learning. Please refer to the paper for more details.

### 2.3.3 Individual Motions in High-Density Crowds

Very recently, there has been a new line of research, looking into predicting detailed pedestrian full-body motions in high-density crowds. Albeit still aiming at predicting trajectories, the research aims to predict detailed individual joint trajectories for the whole body for a single person, especially under unexpected physical perturbations such as pushing. Compared with all the aforementioned research, this new line of research is significantly more challenging in two folds. First, it stops treating each person as a 2D disc or particle. Instead, it models the full body for each individual. Second, it also considers potential physical interactions between pedestrians, in the presence of push and push propagation within densely packed areas. Despite



**Fig. 4** Overview of LDP. This figure is from Yue et al. [2024].

presenting significant challenges, the modelling of the detailed physical interactions brings significant benefits as it allows people to predict how individuals would react in a highly dense crowd, *e.g.* whether there is likely a fall of an individual, which could cause life-threatening danger.

Yue et al. [2024] is probably the first work in this area to our best knowledge. Yue et al. [2024] proposed a new task: predicting 3D full-body motions under unexpected physical perturbations, to study the physical interactions and interaction propagations in crowds. More precisely, given the initial poses and input force, this task aims to predict the reactive individual motions. To this end, the authors proposed a novel latent differentiable physics (LDP) model. The overview of LDP is shown in Fig. 4. The 3D pose  $X_t$  is first simplified as an inverted pendulum model (IPM)  $I_t$  by the Skeleton-to-IPM module. Then, a differentiable IPM simulates forward to obtain  $I_{t+1}$  based on the 3D IPM motion equation:

$$M(I_t, l_t)\ddot{I}_t + C(I_t, \dot{I}_t, l_t) + G(I_t, l_t) = F_t^{net}, \quad (23)$$

where  $l_t$  is the length of the rod in IPM,  $M \in \mathbb{R}^{4 \times 4}$  represents the inertia matrix,  $C \in \mathbb{R}^{4 \times 1}$  denotes the Centrifugal matrix,  $G \in \mathbb{R}^{4 \times 1}$  represents external forces like gravity, and  $F_t^{net}$  denotes the net force.  $M$ ,  $C$  and  $G$  can be calculated given the IPM state  $I_t$ , its first-order derivative  $\dot{I}_t$  and the rod length  $l_t$ . LDP considers the balance recovery force, interaction force, ground friction, and input force to calculate the net force  $F_t^{net}$  at each step. Finally,  $I_{t+1}$  is mapped onto the 3D pose  $X_{t+1}$  via the Skeleton Restoration Model consisting of CVAEs and MLP samplers.

The core idea of LDP is to model physical interactions and predict motions in the simplified IPM space. Therefore, we introduce the proposed differentiable IPM in detail. The IPM in LDP, illustrated in Fig. 4 right, consists of a cart, a massless rod and a point mass. Together, the rod and point mass form an inverted pendulum, which is mounted to the cart. At time step  $t$ ,  $I_t = [x_t, y_t, \theta_t, \phi_t] \in \mathbb{R}^4$  is used to represent the state of an IPM.  $[x_t, y_t]$  denotes the coordinate of the cart moving in the XY-plane.  $\theta_t$  ( $\phi_t$ ) represents the rod's rotation angle about the  $Y_L$ -axis ( $X_L$ -axis) in the local coordinate system  $\Sigma_L$ . Solving Eq. (23) results in  $\ddot{I}_t$  to update the IPM state through a semi-implicit framework, *i.e.*  $\dot{I}_{t+1} = \dot{I}_t + \Delta t \ddot{I}_t$  and  $I_{t+1} = I_t + \Delta t \dot{I}_{t+1}$ , where  $\Delta t$  is the fixed time interval. Note that the rod length  $l_t$  is changeable to enhance the representation ability of the IPM. LDP updates  $l_t$  at each time step through an MLP:

$$\Delta l_t = MLP([\theta_t, \phi_t, \dot{x}_t, \dot{y}_t, \dot{\theta}_t, \dot{\phi}_t, F_t^{self}, M, l_t]), \quad l_{t+1} = l_t + \Delta l_t, \quad (24)$$

where  $F_t^{self}$  is the balance recovery force and  $M$  denotes the mass of the target person.

The key challenge in solving the motion equation Eq. (23) lies in how to estimate the net force  $F_t^{net}$ . LDP models the net force  $F_t^{net}$  as:

$$F_t^{net} = F_t^{self} + F_t^{inta} + f_t + F_t^{input}, \quad (25)$$

where  $F_t^{inta}$ ,  $f_t$ , and  $F_t^{input}$  denote the interaction force, ground friction and input force, respectively. In the data Yue et al. [2024] used,  $F_t^{input}$  is given. The complexity of modelling is from  $f_t$ ,  $F_t^{self}$ , and  $F_t^{inta}$ , where  $f_t$  is the simplest, and  $F_t^{self}$  and  $F_t^{inta}$  are more complex.

The friction is modelled as a damping force:  $f_t = -\mu[\dot{x}_t, \dot{y}_t, 0, 0]$ , where  $\mu$  is a learnable parameter. So the friction force is learnable, as there is no good way to capture such data.

Next, LDP models the balance recovery force as  $F_t^{self} = F_t^{self-pd} + F_t^{self-nn}$ , where  $F_t^{self-pd}$  is the feed-forward torque based on proportional derivative (PD) control.  $F_t^{self-nn}$  is learned by neural networks which can be viewed as a torque correction.  $F_t^{self-pd}$  drives the IPM to respond to external perturbation and recover balance. LDP adopts a widely used assumption that individuals subjected to perturbation typically try to recover to an upright position with zero linear velocity. So  $F_t^{self-pd}$  is computed using:

$$\begin{aligned} F_t^{self-pd} &= K_p e_t + K_d \dot{e}_t, & e_t &= s_d - s_t, \\ s_d &= [0, 0, 0, 0], & s_t &= [\dot{x}_t, \dot{y}_t, \theta_t, \phi_t], \end{aligned} \quad (26)$$

where  $K_p$  and  $K_d$  are the control hyper-parameters,  $e_t$  is the state error in PD control,  $s_d$  is the desired PD state, and  $s_t$  is the current PD state. Although  $F_t^{self-pd}$  can ensure the balance recovery, the predicted motions only using  $F_t^{self-pd}$  are coarse and inaccurate. Therefore, LDP introduces  $F_t^{self-nn}$  to correct  $F_t^{self-pd}$  and obtain more accurate motion prediction, which is estimated by an LSTM:

$$F_t^{self-nn} = LSTM([\theta_t, \phi_t, \dot{x}_t, \dot{y}_t, \dot{\theta}_t, \dot{\phi}_t, M]). \quad (27)$$

LDP models the interaction force for each pair of individuals and uses the summation of interaction forces  $F_{t,nj}^{inta}$  from all the neighbours in the neighbourhood  $\Omega_{t,n}$  of an individual as the final interaction  $F_{t,n}^{inta}/F_t^{inta}$  for the  $n$ th person:

$$F_{t,n}^{inta} = \sum_{j \in \Omega_{t,n}} F_{t,nj}^{inta} = \sum_{j \in \Omega_{t,n}} F_{t,nj}^{inta-bs} + F_{t,nj}^{inta-nn}, \quad (28)$$

where  $\Omega_{t,n}$  denotes the set of neighbours of the  $n$ th person at time  $t$ . As Eq. (28) shows, each  $F_{t,nj}^{inta}$  consists of the basic interaction force  $F_{t,nj}^{inta-bs}$  and the neural interaction force  $F_{t,nj}^{inta-nn}$ .  $F_{t,nj}^{inta-bs}$  models the primary interaction, while  $F_{t,nj}^{inta-nn}$

**Table 1** The comparison of key techniques in human trajectory prediction.

Technique	Accuracy	Explainability	Data Requirement	Computational Cost
Traditional	Low-Medium	High	Low	Low
RNN	High	Low	High	Medium
CNN	High	Low	High	Medium
GNN	High	Medium	High	High
Generative Model	Very High	Low	High	High
Transformer	Very High	Low	High	High
Physics-inspired Deep Learning	Very High	High	Medium	Medium

serves as a supplementary component. To model  $F_{t,nj}^{inta-bs}$ , LDP adopts different strategies to capture interactions in  $[x, y]$  and  $[\theta, \phi]$ , resulting in forces  $F_{nj}^{bs-xy} \in \mathbb{R}^2$  and  $F_{nj}^{bs-\theta\phi} \in \mathbb{R}^2$ , so that  $F_{t,nj}^{inta-bs} = [F_{nj}^{bs-xy}, F_{nj}^{bs-\theta\phi}]$ , where superscript *inta* and subscript *t* are omitted for simplicity. Further, LDP models  $F_{nj}^{bs-xy}$  by using a repulsive potential energy function  $\mathcal{U}$ :

$$F_{nj}^{bs-xy}(r_{nj}) = -\nabla_{r_{nj}} \mathcal{U}[b(r_{nj})], \quad (29)$$

where  $r_{nj} = r_n - r_j$  is the relative position of carts within the XY-plane between the *n*th target person and their neighbour *j*.  $\mathcal{U}[b] = ue^{-\frac{b}{\sigma}}$  has elliptical equipotential lines, where *u* and  $\sigma$  are hyper-parameters. *b* represents the semi-minor axis of the ellipse and is calculated as:

$$b = \frac{1}{2} \sqrt{(\|r_{nj}\| + \|r_{nj} - \Delta t \dot{r}_{jn}\|)^2 - \|\Delta t \dot{r}_{jn}\|^2}, \quad (30)$$

where  $\dot{r}_{jn} = \dot{r}_j - \dot{r}_n$ .

LDP adopts a different strategy to estimate  $F_{nj}^{bs-\theta\phi}$ . Overall, they are estimated based on the relative orientation of a pair of IPMs. Finally, the neural interaction force  $F_{t,nj}^{inta-nn}$  is learned by an MLP:

$$F_{t,nj}^{inta-nn} = MLP([x_{nj}, y_{nj}, \theta_n, \phi_n, \theta_j, \phi_j, \dot{x}_{nj}, \dot{y}_{nj}, \dot{\theta}_{nj}, \dot{\phi}_{nj}]), \quad (31)$$

where the definition of  $x_{nj}$  and  $y_{nj}$  ( $\dot{x}_{nj}$ ,  $\dot{y}_{nj}$ ,  $\dot{\theta}_{nj}$  and  $\dot{\phi}_{nj}$ ) follows an approach similar to  $r_{nj}$  ( $\dot{r}_{jn}$ ).

## 2.4 Conclusion

The introduction of deep learning has significantly advanced the research in human trajectory prediction. This section first briefly reviewed early traditional machine learning methods and then systematically covered the mainstream deep learning

**Table 2** The comparison of different methods on prediction accuracy. The best results are in bold, and the second-best results are underlined. Methods with \* focus on deterministic trajectory prediction. The others are designed for stochastic trajectory prediction, where 20 future trajectories are generated, and the minimal error is reported. N/A stands for not applicable.

Technique	Method	ETH/UCY		SDD	
		ADE	FDE	ADE	FDE
Traditional	Physics-based Model* [Yamaguchi et al., 2011]	0.55	0.99	36.48	58.14
RNN	Social-LSTM* [Alahi et al., 2016]	0.72	1.54	31.19	56.97
	Group-LSTM* [Zhang et al., 2019]	0.45	0.94	N/A	N/A
CNN	TrajCNN* [Nikhil and Tran Morris, 2018]	0.59	1.22	N/A	N/A
	Ped-CNN* [Zamboni et al., 2022]	0.44	0.91	N/A	N/A
GNN	Social-STGCNN [Mohamed et al., 2020]	0.44	0.75	20.6	33.1
	D-STGCN [Sighencea et al., 2023]	0.42	0.68	15.18	25.50
Generative Model	Social-GAN [Gupta et al., 2018]	0.58	1.18	27.23	41.44
	MLD [Wu and Deng, 2024]	0.17	0.35	6.61	12.65
	IDM [Liu et al., 2025]	0.18	0.27	<b>6.38</b>	11.02
Transformer	AgentFormer [Yuan et al., 2021]	0.18	0.29	N/A	N/A
	PPT [Lin et al., 2024]	0.20	0.31	7.03	10.65
Physics-inspired Deep Learning	NSP-SFM Yue et al. [2022]	<u>0.17</u>	<b>0.24</b>	<u>6.52</u>	<b>10.61</b>
	NDCPM [Wang et al., 2024]	<b>0.15</b>	0.33	N/A	N/A

approaches. To give a high-level comparison of their advantages and drawbacks, we examine them by several indicators, Accuracy, Explainability, Data Requirement, and Computational Cost, shown in Table 1. Note this is a qualitative comparison as the evaluation metrics and experimental settings vary greatly across different publications. Direct quantitative comparison between them across all indicators is itself an interesting, yet open research question.

Traditional machine learning methods typically achieve low to medium prediction accuracy but offer strong explainability. In contrast, deep learning approaches generally excel in prediction accuracy. Although most deep learning methods struggle with explainability, physics-inspired deep learning methods mitigate this limitation and provide substantial explainability. The performance of deep learning models often depends heavily on the volume of available data, resulting in higher data requirements. However, the incorporation of physical priors effectively reduces the data requirement for physics-inspired deep learning methods. Moreover, due to their large number of trainable parameters, deep learning methods typically need more computational resources compared to traditional machine learning methods, particularly when it comes to the methods based on GNNs, generative models or transformers. In conclusion, physics-inspired deep learning methods demonstrate the strongest overall performance, particularly in accuracy and explainability, motivating further exploration of the type of methodologies.

Although it is difficult to directly compare the reviewed methods in the aforementioned four indicators, it is still possible to compare them in their predictive capabilities due to most of the reviewed deep learning methods share common datasets and evaluation metrics for trajectory prediction.

A detailed numerical comparison of prediction accuracy is shown in Table 2 based on the original results reported in the publications. Specifically, we highlight the performance of representative methods which are either pioneering or demonstrating excellent prediction accuracy within each category. This comparison is based on evaluations conducted on two widely used public datasets: ETH/UCY [Pellegrini et al., 2009, Lerner et al., 2007] and SDD [Robicquet et al., 2016]. The primary evaluation metrics used are Average Displacement Error (ADE), which measures the average discrepancy between the predicted trajectory and its ground truth, and Final Displacement Error (FDE), which quantifies the error at the final predicted position. The specific experimental setting is to predict the future trajectory  $\{p^t\}_{t=t_p+1:t_p+t_f}$  for the next  $t_f$  timesteps given the past trajectory  $\{p^t\}_{t=1:t_p}$ , where  $p^t$  denotes the 2D position and each timestep has 0.4 seconds. ADE and FDE are calculated as:

$$ADE = \frac{1}{Nt_f} \sum_{i=1}^N \sum_{t=t_p+1}^{t_p+t_f} \|p_i^t - \hat{p}_i^t\|_2, \quad FDE = \frac{1}{N} \sum_{i=1}^N \|p_i^{t_p+t_f} - \hat{p}_i^{t_p+t_f}\|_2, \quad (32)$$

where  $p_i^t$  is the 2D position for the  $i$ th trajectory,  $\hat{p}_i^t$  is the predicted position, and  $N$  is the total amount of trajectories. Following standard practice in the literature [Gu et al., 2022, Yue et al., 2022, Shi et al., 2023],  $t_p = 8$  and  $t_f = 12$ . The ADE and FDE results are reported in meters for ETH/UCY and pixels for SDD. Notably, employing deep learning techniques such as RNNs, CNNs, and GNNs has substantially improved prediction accuracy, reducing ADE and FDE by approximately 20% and 30%, respectively, on ETH/UCY, and by about 50% for both metrics on SDD, compared with traditional machine learning approaches. The adoption of generative models and transformer architectures has further significantly enhanced the prediction performance across all metrics and datasets. Ultimately, physics-inspired deep learning methods achieve the highest accuracy. In summary, generative models, transformers, and particularly physics-inspired deep learning methods demonstrate superior prediction accuracy for human trajectory forecasting.

### 3 Crowd Behaviour Recognition

Crowd behaviour recognition is the process of analysing and identifying patterns or activities in the movements or behaviours of crowds. It aims to understand how individuals within a crowd interact and respond to their environments, other individuals, and external stimuli. This chapter reviews crowd behaviour recognition research based on machine learning from crowd videos. Similar to trajectory prediction, traditional machine learning played a key role in behaviour recognition, then deep learning has been becoming popular recently.

### 3.1 Traditional Machine Learning Methods

Traditional machine learning methods typically involve manually designing effective features to represent crowd dynamics, followed by training a classifier to recognise crowd behaviours based on these features. These traditional methods generally model crowd movements from two perspectives: holistic and individual-based.

Holistic methods treat the crowd as a single entity. Often this is because the individual movements either cannot be accurately detected *e.g.* due to the view blocked in the camera, the camera being too far away from the crowds, or are regarded as insignificant when it comes to the identification of the global behaviour pattern. Therefore, these approaches focus on the macroscopic behavioural patterns and often treat the crowds as a continuum, which are well-suited for analysing dense crowds where each person occupies merely a small number of pixels hence difficult to detect/track. Solmaz et al. [2012] proposed a framework for identifying crowd behaviours using stability analysis of dynamical systems. Their approach employs a Lagrangian perspective and overlays a grid of particles onto a scene. This way, they defined a dynamical system based on the particle movements, which is indicated by the optical flows of the video. The dynamical system is not fully parameterised but approximated by a Taylor's expansion of the system with the first-order term where the Jacobian matrix dictates the dynamics. By analysing the particle trajectories and employing the eigenvalues of the Jacobian matrix, the method classifies specific behaviours (Lane, Blocking, Bottleneck, Fountainhead, and Arch/Ring) through stability patterns. In comparison, Su et al. [2013] took an Eulerian perspective and proposed a new spatiotemporal viscous fluid field to model crowd dynamics for large-scale crowd behaviour recognition. This approach constructs the novel spatiotemporal features, by incorporating both the appearance variations of the crowds and the interaction forces among pedestrians estimated through the shear stress in the fluid field. Then it employs a latent Dirichlet allocation model [Blei et al., 2003] to identify crowd behaviours such as dispersion and gathering based on these spatiotemporal features. More recently, Matkovic et al. [2022] presented a novel quantum mechanics-inspired method to recognise dominant motion patterns in macroscopic crowd analysis. The method extracts optical flows from the input video and introduces particles similar to Solmaz et al. [2012]. Then, it constructs a wave field based on the optical flows, the particles, and some pre-defined wave functions. The method defines the peak of the wave field as a meta-tracklet, which indicates the most probable particle flow. Eventually, a classifier is employed to recognise the dominant motion patterns such as Inline and Circle, based on the meta-tracklets and the functions of fuzzy predicates.

Individual-based methods regard the crowd as a collection of individuals, enabling detailed analysis of behaviours at the microscopic level. Zhou et al. [2012] used agents with a linear dynamic system to model the motion of each pedestrian. They proposed a novel mixture model of dynamic agents to analyse the collective crowd behaviours. After learning from the data, the mixture model can simulate and classify crowd behaviours. Choi and Savarese [2012] introduced a framework which unifies and integrates multi-target tracking and crowd behaviour recognition. They proposed a

novel hierarchical graphical model to jointly optimise the target tracking and the activity classification. The model links three levels of activities (*i.e.* individual, interaction, and crowd) and leverages the contextual information at each level to improve recognition. The clustering-based methods such as Wang et al. [2009], Zhou et al. [2012] later inspired more fine-grained behavioural analysis on crowd activities. Wang et al. [2016, 2017a] extended [Wang et al., 2009] into more detailed spatial pattern recognition for crowd activities. Wang and O’Sullivan [2016] extended Wang et al. [2017a] and proposed new unsupervised spatio-temporal behaviour recognition. Finally, He et al. [2020] extended [Wang and O’Sullivan, 2016] into unsupervised behaviour analysis based on space, time and dynamics simultaneously.

### 3.2 Deep Learning Methods

Similar to human trajectory prediction, deep learning has made a big impact on the research in crowd behaviour recognition, by overcoming the limitations of traditional approaches, enabling the analysis of complex scenarios with greater accuracy, scalability, and robustness. Deep learning methods for crowd behaviour recognition utilise neural networks to implicitly learn effective features from data, greatly minimising the dependence on manual feature engineering. Additionally, these methods tend to integrate feature extractors with classifiers, providing efficient end-to-end models. With the advancement of deep learning, CNNs, RNNs, GNNs, transformers *etc.* have been explored to identify various crowd behaviours. Similar to the above section, we also classify these methods based on the types of neural networks.

#### 3.2.1 Convolutional Neural Networks

CNNs are widely used in crowd behaviour recognition because they are well-suited for processing image and video data, which is the primary source of information in this domain (*e.g.* CCTV or surveillance videos). Shao et al. [2015] proposed a novel deep model consisting of two CNN branches, with the same architecture, which integrates the appearance and the motion features to jointly learn discerning features for crowd behaviour recognition. One branch of the model takes a frame of the video to capture the static features of the crowd, while the other branch accepts the continuous motion maps derived from the crowd video to capture the dynamic features. Furthermore, inspired by the behavioural features (stability, conflict, and collectiveness) widely employed in prior studies, the authors proposed three continuous motion maps representing various dynamics information. Then the features from the two branches are fused to serve as the input of a classifier, which is a fully connected layer. Extensive experiments demonstrated that the proposed model had superior performance over handcrafted features and state-of-the-art baselines. Moreover, the authors built a large-scale dataset, WWW Crowd Dataset, containing 10,000 crowd videos from 8,257 scenes. As data is the foundation of deep learning,



**Fig. 5** The overview of the proposed deep model in Shao et al. [2015]. Blue, green, orange, and red blocks denote convolutional, pooling, normalization, and fully connected layers, respectively. This figure is from Shao et al. [2015].

the data contribution is considered a distinctive contribution to the field. Shao et al. [2015] is the pioneering work applying deep learning in crowd behaviour recognition, with a new large-scale dataset, which has inspired a stream of research [Shao et al., 2016, Ullah et al., 2019, Deng et al., 2020]. Therefore, we provide more details of the method.

The overview of the proposed model in Shao et al. [2015] is shown in Fig. 5. A single frame and continuous motion maps are extracted from the given video to serve as input on the left. Each branch has a CNN constructed by stacking convolutional, pooling, normalization, and fully connected layers, with a rectified linear unit applied after every convolutional layer as the activation function. The final fully connected layer is the classifier followed by a sigmoid activation function, which predicts the probability of the input data showing a certain behaviour. The calculation of three continuous motion maps is inspired by Zhou et al. [2013], Shao et al. [2014]. Specifically, the authors first detected the tracklets of motions and established a 10-nearest-neighbour graph for all tracklets. Then they used a stability descriptor by averaging over the unchanging neighbours for each node within the graph. Next, they proposed a conflict descriptor based on the graph's velocity correlation of nearby nodes. As for collectiveness, the descriptor from Zhou et al. [2013] was employed. Finally, they could generate a stability/conflict/collectiveness descriptor map for each frame. Averaging the descriptor maps along the temporal dimension produced three motion maps. To enrich the representation of motion information, these motion maps were interpolated to generate continuous motion maps.

Ullah et al. [2019] explored crowd behaviour recognition following the modelling idea in Shao et al. [2015], providing a two-stream model based on CNNs. Both methods use the frames of the crowd video to capture the appearance information, while Ullah et al. [2019] utilised optical flows to capture motion information. Finally, softmax scores from the two streams are fused to obtain classification results. The classic two-stream architecture is also used to explore crowd behaviour recognition in other CNN-based methods [Shuaibu et al., 2017, Wei et al., 2020, Ullah et al., 2021].

Mandal et al. [2018] proposed a novel crowd behaviour recognition model that combines deep residual neural networks with subclass discriminant analysis [Zhu and

Martinez, 2006] to enhance feature extraction and classification accuracy. The model uses a ResNet-50 [He et al., 2016] to extract features from each frame of the input video. To obtain the intra-class variances, the authors introduced spatial partition trees [Wang et al., 2013] which segment each crowd behaviour class into subclasses within the feature maps. Subsequently, the features from these subclasses are further refined using eigenvalue-based regularisation to extract more discriminative features. Next, the model computes a total subclass scatter matrix from the refined features and applies the matrix to extract the final low-dimensional features. Finally, a 1-nearest-neighbour classifier is employed to recognise crowd behaviours based on the low-dimensional features.

PIDLNet [Behera et al., 2021d] combines CNNs with physics-related metrics to characterise crowd videos. Specifically, the model introduces the inflated 3D ConvNets (I3D) [Carreira and Zisserman, 2017], a video classification network, to produce implicit features by capturing and aggregating the appearance and the motion features from the input video. Meanwhile, PIDLNet uses the optical flow information extracted by I3D to estimate physics-based features consisting of entropy and order parameters. Then, the implicit features and the physics-based features are fused. A classifier, a linear layer, takes the fused features as input to recognise structured and unstructured crowds.

Bendali-Braham et al. [2021] integrated existing CNN models by ensemble learning to improve recognition performance. This ensemble includes I3D, two-stream I3D [Carreira and Zisserman, 2017], Convolutional 3D [Tran et al., 2015], and Resnet 3D [Hara et al., 2017].

3D-AIM [Choi et al., 2024] is a new CNN model based on atrous convolutions [Chen et al., 2017] for crowd behaviour recognition. The model integrates two key components: an atrous block with atrous convolutions which captures the spatial features, and an inception block with standard convolutions which extracts the spatiotemporal features. 3D-AIM utilises the atrous convolutions to expand the receptive field without increasing the computational parameters when extracting the spatial features. Furthermore, the authors proposed a novel separation loss to assign greater weighting onto the difficult-to-classify examples and enhance class separation.

### 3.2.2 Recurrent Neural Networks

The importance of temporal information in crowd data motivates the usage of RNNs. RNNs enable the modelling of the dynamics of individual information and how they influence each other when interactions happen. The combination of CNNs with RNNs has further advanced crowd behaviour recognition. Wang et al. [2017b] proposed a hierarchical recurrent modelling approach. Specifically, a person-level LSTM is first used to capture the individual dynamics, which takes the appearance and the motion information of each individual as input at each time step. The method extracts individual information from videos using traditional methods [Choi et al., 2009, Choi and Savarese, 2012] and CNNs. Each person is described by the features based on the outputs of the person-level LSTM. Subsequently, a group-level LSTM

is employed to model the interactions at the group level, where groups are identified according to the spatiotemporal proximity of individuals. Individual features within a group are ordered into a sequence by their spatial coordinates in the frame space and fed into the group-level LSTM. Similarly, each group is described by the group features out of the group-level LSTM. Finally, a scene-level LSTM is employed to capture the whole crowd behaviour.

Vahora and Chauhan [2019] proposed a deep neural network model to recognise crowd behaviours by leveraging contextual information. At each time step, a frame of the input video is fed into a CNN model which extracts scene-level features and outputs activity scores. Meanwhile, the model uses a CNN model to extract the individual features from detected individuals in the frame and employs a pooling operation to aggregate these features. An RNN model (LSTM/GRU) then receives the aggregated features as input and estimates the activity scores computed by softmax. In the end, a probabilistic inference model produces the final activity label based on all the previous softmax scores.

Yan et al. [2019] presented an encoder-decoder model to recognise behaviours in crowd videos and generate corresponding captions. The CNN-based encoder extracts features from input videos. These features are fed into the RNN-based decoder to predict captions.

Behera et al. [2021c] explored the combination of traditional handcrafted features and RNNs. To be specific, the proposed model extracts features incorporating 2D motion histograms, order parameters, and entropy from each frame of the input video. These temporally related features are fed into an LSTM to produce classification results at every time step. Finally, all classification results are aggregated to classify the crowd as structured or unstructured.

Rezaei and Yazdi [2021] proposed the Conv-LSTM-AE model to capture high-level representations of data, where AE stands for the autoencoder [Rumelhart et al., 1986]. Conv-LSTM-AE can be seen as a multi-task learning framework. Its main architecture is an autoencoder. The encoder accepts a sequence of optical flow images obtained from the input video based on [Pérez et al., 2013] as input to extract features. The encoder consists of convolution layers and convolutional LSTMs (ConvLSTMs) [Shi et al., 2015]. Subsequently, the encoded features are fed into two separate branches. One branch is a classification branch which employs an MLP to recognise crowd behaviours. The other branch reconstructs the optical flow images, integrating ConvLSTMs and transposed convolution layers [Zeiler et al., 2011]. The proposed model employs a loss function that accounts for reconstruction and classification errors during training. Prior to Rezaei and Yazdi [2021], ConvLSTMs were also used to model spatio-temporal information for behaviour recognition [Li, 2018]. Later, Chaturvedi et al. [2024] combined ConvLSTMs with the attention mechanism to identify crowd activities.

### 3.2.3 Graph Neural Networks

Similar to trajectory prediction where moving pedestrians can be modelled by a graph, GNNs have unique advantages when used for crowd behavioural analysis. Their abilities to model non-Euclidean data allow individuals and objects in the environment to be seen as graph nodes and the relationship between any two nodes to be modelled by edges. Therefore, GNNs are highly effective at modelling the complex interactions and relationships inherent in crowd dynamics, facilitating the learning of both local and global patterns within a crowd. Moreover, GNNs can handle multimodal data by embedding appearance, motion, and semantic features into the graph structures, allowing a multi-faceted modelling of crowd behaviours. Therefore, a series of GNN-based methods have been proposed in this area.

Behera et al. [2021a] regarded crowd behaviour recognition as a graph classification problem and used a graph convolutional neural network to identify types of graphs which correspond to different behaviours. Given the crowd video, constructing graphs is to introduce a node for each group in the crowd and establish edges based on the orientations of groups. Specifically, the proposed method uses a Langevin equation-based framework [Behera et al., 2021b] to detect groups. Then, seeing crowds as fluids, the method extracts features for each group from the video, including optical flow features and physics-related features. Combining these features gives the node features. The edges of the graph are then built if the orientation similarity between the two groups is higher than a certain pre-defined threshold. Finally, a graph convolutional neural network classifies these graphs.

Liu et al. [2021] introduced a multimodal semantic context-aware graph neural network, designed to capture the visual and the semantic interactions in complex scenes. The proposed model constructs two multimodal visual sub-graphs: an RGB graph to capture the appearance cues and an optical-flow graph for motion patterns. Both sub-graphs build nodes for individuals and are fully connected. Node features are extracted from RGB and optical flow images using I3D [Carreira and Zisserman, 2017]. Subsequently, the model designs the modality-specific and cross-modal aggregation layers to refine graph representations of two sub-graphs. Additionally, a semantic graph is built with nodes corresponding to individual actions and crowd behaviour labels, representing edges by a fixed adjacency matrix. A bi-directional mapping mechanism based on graph convolutional networks is employed to link the multimodal visual graphs to the semantic graph, enriching the visual representations with the semantic contexts. Finally, the multimodal semantic context-aware features are used to recognise the crowd activities. As a follow-up paper, Liu et al. [2022a] extended their work [Liu et al., 2021] by utilizing estimated pose information of individuals in crowds to guide modality-specific and cross-modal aggregation of two sub-graphs.

Longobardi and Riccio [2024] aimed to classify diverse crowd behaviours within the same environment, unlike previous methods [Shao et al., 2014, Su et al., 2016] which rely on pre-segmented scenes with single crowd behaviour. To this end, the authors proposed a novel two-stream graph neural network model. The first stream feeds a set of graphs to a graph convolutional network separately to output a set

of graph representations. They used graphs of two layers, the bottom layer and the top layer, based on a grid/tile discretization of the scene. Bottom-level and top-level graphs are constructed based on detected groups and tiles, respectively. To enhance the classification, the model incorporates motion flow images, capturing both the individual and the collective dynamics. The second stream uses a CNN to extract the motion features from the motion flow images. The model then combines the graph representation with the motion features to conduct node-level classification.

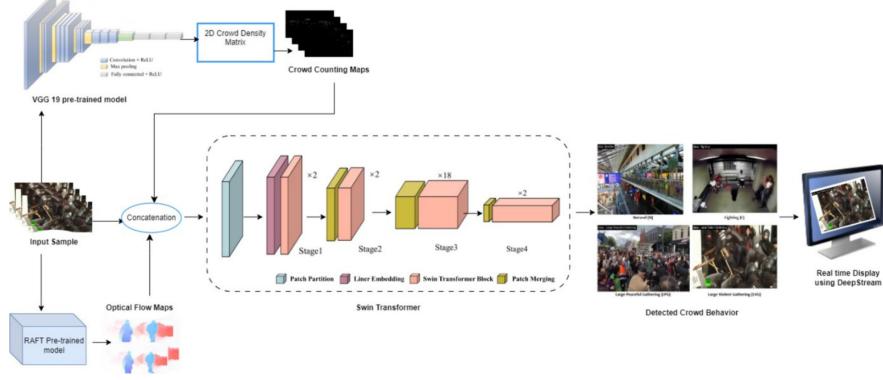
### 3.2.4 Transformers

Given a sequence of input data, transformers use attention mechanisms to learn the correlation between different parts of the data and capture long-range dependencies. In crowd behavioural recognition, this provides an alternative way for a model to capture the spatial and temporal relationships in crowds. Therefore, it is no surprise that transformers have been actively employed in this area.

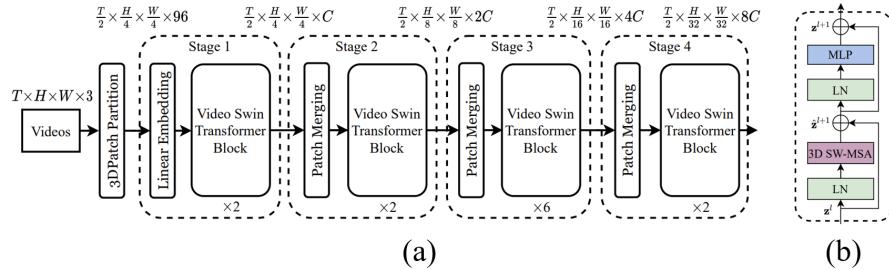
Tamura et al. [2022] proposed a new transformer-based method to recognise crowd activities. The method feeds the input video into I3D to extract a batch of multi-scale feature maps. Each feature map is then passed through an average pooling layer followed by some projection convolution layers to reduce the computational demands of transformers, resulting in altered feature maps. Subsequently, a deformable transformer encoder [Zhu et al., 2021] receives the altered feature maps and a collection of multi-scale position encodings [Zhu et al., 2021] as input to produce refined feature maps. Finally, a set of learnable query embeddings and the refined feature maps are fed into a deformable transformer decoder [Zhu et al., 2021] to obtain effective behaviour representations. Tamura [2024] extended this work by employing two separate deformable transformer decoders, one for individuals and another for groups. Each decoder is equipped with its own learnable query embeddings.

Zuo et al. [2023] proposed another transformer to capture effective features for crowd behaviour recognition. The model processes the input video into a sequence of tokens and maps them onto embedding vectors by a linear layer, following the framework of Visual Transformers [Han et al., 2020]. Then, a transformer accepts these embedding vectors as input to extract features. The multi-head self-attention module in the transformer is modified to contain both spatial and temporal blocks. Finally, the extracted features are fed into an MLP classifier to identify the crowd behaviours.

Qaraqe et al. [2024] provided an in-depth exploration of crowd behaviours varying in size and levels of violence. The authors categorised crowd behaviours into natural, large peaceful gathering, large violent gathering, and fighting. They proposed a swin transformer-based model to classify behaviours. The overall structure of the proposed model is shown in Fig. 6. Specifically, crowd counting and optical flow maps are extracted from a given video. Then, these features are concatenated with the given video to form the input for a swin transformer proposed by Liu et al. [2022b], which is efficient for video recognition. Finally, the transformer produces effective classification features. Comprehensive experiments demonstrated that the



**Fig. 6** Overall structure of the proposed model in Qaraqe et al. [2024]. This figure is from Qaraqe et al. [2024].



**Fig. 7** Architecture of the swin transformer (a) and an illustration of the video swin transformer block (b). These two figures are from Liu et al. [2022b].

proposed method outperformed existing state-of-the-art approaches. Moreover, the authors created a large dataset containing 68 hours of videos from both closed-circuit television and social media. Identifying violent behaviours is crucial for crowd management and public security [Saxena et al., 2008]. The proposed model in Qaraqe et al. [2024] has advanced the research on violence detection in crowds by applying cutting-edge deep learning techniques. Therefore, we would like to provide more details of the model.

Qaraqe et al. [2024] first followed Wan et al. [2021] to generate crowd counting maps as heatmaps. These maps provide the features of crowd distributions to capture the size of crowds. Specifically, a single frame is fed into a pre-trained VGG19 [Simonyan and Zisserman, 2015] to produce a 2D crowd density matrix. Then, a crowd counting map is obtained from the density matrix. The model generates a set of crowd counting maps for all the odd-numbered frames. Subsequently, a pre-trained RAFT model [Teed and Deng, 2020] is employed to produce optical flow maps. The model then uses the optical flow maps to capture motion features. Finally, the video frames, the crowd counting maps and the optical flow maps are concatenated to serve

**Table 3** The comparison of key techniques in crowd behaviour recognition.

Technique	Accuracy	Data Requirement	Computational Cost
Traditional	Low-Medium	Low	Low
CNN	High	High	Medium
RNN	High	High	Medium
GNN	High	High	High
Transformer	High	High	High

as the input for the swin transformer. Then a softmax layer receives the transformer’s output to predict the probabilities of diverse crowd behaviours.

The architecture of the swin transformer is shown in Fig. 7 (a), including four stages. Suppose that the given video has the shape  $T \times H \times W \times 3$ , with the number of frames  $T$ , the frame height  $H$ , the frame width  $W$ , and 3 RGB channels. Subsequently, a 3D patch partition layer segments the video into a set of 3D patches with shape  $2 \times 4 \times 4 \times 3$ . Each 3D patch is flattened to be a token with a 96-dimensional feature. Therefore, there are  $\frac{T}{2} \times \frac{H}{4} \times \frac{W}{4}$  tokens with dimension 96 as input for stage 1. The linear embedding layer converts the feature dimension of every token into an arbitrary dimension  $C$ . The video swin transformer block refines the features further. Next, three stages with the same structure are applied to all tokens to obtain more refined representations. The patch merging layer implements spatial downsampling by concatenating features of  $2 \times 2$  neighbouring tokens and applies a linear layer to reduce the dimension of the merged features by half. Finally, the architecture of the video swin transformer block is shown in Fig. 7 (b), which has two crucial components: 3D shifted window-based multi-head self-attention (3D SW-MSA) and MLP. The 3D SW-MSA is designed to replace the traditional multi-head self-attention for effective video recognition. More details of 3D SW-MSA can be found in Qaraqe et al. [2024], Liu et al. [2022b]. There is a layer normalization (LN) before each component. Additionally, the block introduces the residual connection for 3D SW-MSA and MLP.

### 3.3 Conclusion

In this section, we initially outlined traditional machine learning methods from the holistic and the individual-based perspective. Subsequently, we provided a structured introduction to deep learning approaches, categorised according to their network architectures. To understand their relative strengths and weaknesses, we give a high-level qualitative comparison in Table 3.

First, the introduction of deep learning has substantially improved the recognition accuracy compared with the traditional machine learning approaches. While comparison is qualitative, providing a fair quantitative comparison remains challenging due to the absence of unified experimental settings and datasets. Different methods often utilise diverse datasets, and even when using the same dataset, their experi-

mental conditions vary. For instance, Li [2018] aimed to identify eight distinct crowd behaviour categories using the CUHK Crowd dataset [Shao et al., 2014], whereas Wei et al. [2020] focused on classifying crowd behaviours as either heterogeneous or homogeneous. Furthermore, traditional machine learning methods relying on feature engineering typically require significantly less data and lower computational cost than deep learning approaches. In particular, GNNs and transformers generally demand more computational resources due to their complex architectures. In summary, as data availability and computational power continue to grow, deep learning methods represent both the current standard and the direction for future research in crowd behaviour recognition.

## 4 Discussion

### 4.1 Effectiveness of Crowd Behaviour Analysis Methods

This chapter reviewed two fundamental areas in crowd behaviour analysis: crowd behaviour prediction and recognition. Recent developments in deep learning have significantly advanced research within these domains. As a result, existing methods in crowd behaviour analysis demonstrate varied effectiveness depending on crowd types, environmental contexts, and specific applications.

Based on the diverse behavioural patterns, crowds are often classified into low-density and high-density crowds [Zhao et al., 2018]. Existing crowd analysis methods excel at modelling low-density crowd dynamics. For instance, various human trajectory prediction methods generally focus on low-density crowds, where individual tracking is feasible and trajectory data are available. Current trajectory prediction methods based on deep learning have shown strong performance in capturing low-density crowd dynamics [Yue et al., 2022, Shi et al., 2023], achieving low errors in common evaluation metrics such as ADE and FDE. Particularly, physics-inspired deep learning methods also mitigate the problem of lacking explainability [Yue et al., 2022]. In contrast, modelling high-density crowds dynamics is significantly more challenging because of complex interactions between individuals. Most recently, current research on crowd behaviour analysis has achieved breakthroughs in this area. Specifically, Yue et al. [2024] studied detailed physical interactions at the full-body level through a new human motion prediction task. They proposed an effective interaction model, resulting in accurate motion prediction. He et al. [2025] regarded high-density crowds as continuum active matters and proposed a new crowd material point method. The method can learn from in-the-wild videos, the most common available data for high-density crowds, and then effectively predict and simulate dense crowd movements. Nevertheless, the research on high-density crowds remains comparatively less developed, primarily due to data scarcity and modelling complexity.

Current crowd prediction and recognition research relies heavily on specific datasets [Borja-Borja et al., 2018, Schuetz and Flohr, 2023, Zhang et al., 2025].

Although they contain various scenes, *e.g.* indoor/outdoor, these datasets can not cover all real-world situations. Existing methods have achieved excellent prediction and recognition performance on these benchmark datasets [Yuan et al., 2021, Liu et al., 2022a, Qaraqe et al., 2024], demonstrating their effectiveness in diverse environmental contexts. However, these methods generally require further validation or fine-tuning on more real-world data when deployed in complex real-world situations [Velayutham et al.]. A key future direction is to further enrich the datasets, capturing crowds under diverse events, time, places, etc, and test the generalisability of these methods.

Crowd behaviour analysis methods play an important role in various practical applications such as emergency response, crowd management, and autonomous vehicle systems. For emergency response scenarios, crowd behaviour recognition approaches effectively identify abnormal or potentially dangerous situations, such as stampedes and sudden dispersals. Accurate detection facilitates early warning, improves situational awareness, and supports strategic evacuation planning Wei et al. [2020]. In crowd management applications, trajectory prediction methods can provide accurate future behaviour estimations, enabling proactive interventions to prevent overcrowding and bottlenecks [Tamaru et al., 2024]. Crowd behaviour prediction and recognition are essential components of autonomous vehicle systems, particularly within urban environments where autonomous vehicles must safely navigate around groups of pedestrians. Anticipation of pedestrians' behaviours for the next few seconds and identification of crowd behaviours provide plenty of behavioural cues, allowing the autonomous vehicles to make context-aware decisions [Rasouli and Tsotsos, 2019, Camara et al., 2020].

## 4.2 Future Research

After several decades of research in crowd behaviour analysis, we have observed a wave of new research in the past few years in the field, owing to the fast development of deep learning. Similar to how deep learning has influenced many fields outside of machine learning and AI, it is highly likely that the new research tools based on deep neural networks will become more and more deeply entrenched into the toolbox of crowd analysis. We expect to see the application of deep learning continue to expand more widely into crowd research, way beyond its current scope. There are several key challenges and changes which are likely to emerge in the near future.

The first change is the digital infrastructure development to enable more automated, comprehensive, and multi-modal data capture for crowd research. As mentioned at the beginning of this chapter, the two key elements in applying deep learning are data and tasks. Data, as the foundation of any data learning application, has so far appeared in the form of images and video in crowd research. This is due to the cameras are probably the most commonly used sensors everywhere. Other data based on smartphones, smart watches, GPS, *etc.*, have been actively explored but not at the same scale yet. We expect more types of sensors to be deployed in the

near future providing abundant data, which are not only large in quantity but also of higher quality as well as in more modalities. This also has a strong synergy with the development of the Internet of Things and Smart City.

Along with the wider and deeper data collection on crowds, we expect to see new tasks being defined and solved using deep learning. This might shed light on some of the problems previously deemed to be extremely challenging. One instance is high-density crowds where crushes could happen. Currently, it is rare to see data of crushing and collapsing crowds which can be directly used by deep learning. The insufficient data size and the quality of the data are the main bottlenecks. With new sensors deployed and new data captured, it is possible for deep learning to do real-time analysis and prediction, to prevent crowd crushes. Actually, Yue et al. [2024] have already started in this direction, albeit the data was captured in a laboratory environment, not natural environments. In general, it is possible to re-solve most of the existing problems in crowd analysis using deep learning, when the data is ready and the old problem is reformulated into a form friendly to deep learning.

Another possible future direction is to systematically move from supervised learning tasks to unsupervised ones. Currently, no matter it is trajectory prediction or behaviour recognition, supervision signals need to be obtained either from human labelling or some other algorithms in the pre-processing stage. When data capture becomes more automated, fully automated pre-processing might be hard and human labelling will become prohibitively laborious. We expect to see more unsupervised learning methods being developed in deep learning, following their predecessors in statistical machine learning.

While deep learning will continue to expand its influence in crowd analysis, along with new data collected and new tasks defined, we expect that there are also difficulties in applying deep learning to some research topics in this field. This is because some data are hard to collect and some tasks are difficult to clearly define, in the way that deep learning model can be developed. Crowds are also studied from the perspective of social sciences, such as group dynamics and crowd psychology. While naturalist motion data collection is already challenging to collect, at least they are directly observable. In contrast, some data employed in social sciences are intrinsically hard to collect, *e.g.* personality, affective states, *etc*. Admittedly, it still might be feasible to collect such data for a small number of people. But we expect it is difficult to scale and automate this kind of data collection to meet the demanding requirements of deep neural networks. It will require AI researchers, social scientists, engineers and policymakers to jointly create new solutions, to enable us to leverage deep learning for solving these problems.

## 5 Conclusion

In recent years, deep learning has significantly transformed the research in crowd behaviour analysis, offering powerful tools to predict and recognise crowd actions with remarkable accuracy. This chapter provides a comprehensive review of re-

cent developments in two core tasks: crowd behaviour prediction and recognition. We briefly introduced traditional machine learning approaches and systematically reviewed deep learning methods. In particular, we presented representative deep learning methods in detail. Moreover, this chapter summarises and compares the key techniques in crowd behaviour prediction and recognition to understand their respective strengths and weaknesses. Finally, we discussed the effectiveness of crowd behaviour analysis methods across multiple aspects and provide practical suggestions for future research directions regarding this field. It is our expectation that this review will bring greater visibility to this rapidly developing field and inspires further research in the discussed areas.

## References

- 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.
- Javad Amirian, Jean-Bernard Hayet, and Julien Pettré. Social ways: Learning multi-modal distributions of pedestrian trajectories with gans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019.
- Federico Bartoli, Giuseppe Lisanti, Lamberto Ballan, and Alberto Del Bimbo. Context-aware trajectory prediction. In *2018 24th International Conference on Pattern Recognition (ICPR)*, pages 1941–1946. IEEE, 2018.
- Shreetam Behera, Debi Prosad Dogra, Malay Kumar Bandyopadhyay, and Partha Pratim Roy. Crowd characterization in surveillance videos using deep-graph convolutional neural network. *IEEE Transactions on Cybernetics*, 53(6): 3428–3439, 2021a.
- Shreetam Behera, Debi Prosad Dogra, Malay Kumar Bandyopadhyay, and Partha Pratim Roy. Understanding crowd flow patterns using active-langevin model. *Pattern Recognition*, 119:108037, 2021b.
- Shreetam Behera, Shaily Preetham Kurra, and Debi Prosad Dogra. Characterization of orderly behavior of human crowd in videos using deep learning. In *Intelligence Science III: 4th IFIP TC 12 International Conference, ICIS 2020, Durgapur, India, February 24–27, 2021, Revised Selected Papers 4*, pages 217–226. Springer, 2021c.
- Shreetam Behera, Thakare Kamalakar Vijay, H Manish Kausik, and Debi Prosad Dogra. Pidlnet: A physics-induced deep learning network for characterization of crowd videos. In *2021 17th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–8. IEEE, 2021d.
- Mounir Bendali-Braham, Jonathan Weber, Germain Forestier, Lhassane Idoumghar, and Pierre-Alain Muller. Ensemble classification of video-recorded crowd move-

- ments. In *2021 12th International Symposium on Image and Signal Processing and Analysis (ISPA)*, pages 152–158. IEEE, 2021.
- Niccoló Bisagno, Bo Zhang, and Nicola Conci. Group lstm: Group trajectory prediction in crowded scenarios. In *Proceedings of the European conference on computer vision (ECCV) workshops*, pages 213–225, 2018.
- Niccoló Bisagno, Cristiano Saltori, Bo Zhang, Francesco GB De Natale, and Nicola Conci. Embedding group and obstacle information in lstm networks for human trajectory prediction in crowded scenes. *Computer Vision and Image Understanding*, 203:103126, 2021.
- Gary Bishop, Greg Welch, et al. An introduction to the kalman filter. *Proc of SIGGRAPH Course*, 8(27599-23175):41, 2001.
- David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.
- Luis Felipe Borja-Borja, Marcelo Saval-Calvo, and Jorge Azorin-Lopez. A short review of deep learning methods for understanding group and crowd activities. In *2018 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2018.
- Fanta Camara, Nicola Bellotto, Serhan Cosar, Florian Weber, Dimitris Nathanael, Matthias Althoff, Jingyuan Wu, Johannes Ruenz, André Dietrich, Gustav Markkula, et al. Pedestrian models for autonomous driving part ii: high-level models of human behavior. *IEEE Transactions on Intelligent Transportation Systems*, 22(9):5453–5472, 2020.
- Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017.
- Kunal Chaturvedi, Chhavi Dhiman, and Dinesh Kumar Vishwakarma. Fight detection with spatial and channel wise attention-based convlstm model. *Expert systems*, 41(1):e13474, 2024.
- Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017.
- Ricky TQ Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. *Advances in neural information processing systems*, 31, 2018.
- Weihuang Chen, Zhigang Yang, Lingyang Xue, Jinghai Duan, Hongbin Sun, and Nanning Zheng. Multimodal pedestrian trajectory prediction using probabilistic proposal network. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(6):2877–2891, 2022.
- Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. *Advances in neural information processing systems*, 29, 2016.

- Jong-Hyeok Choi, Jeong-Hun Kim, Aziz Nasridinov, and Yoo-Sung Kim. Three-dimensional atrous inception module for crowd behavior classification. *Scientific Reports*, 14(1):14390, 2024.
- Wongun Choi and Silvio Savarese. A unified framework for multi-target tracking and collective activity recognition. In *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7–13, 2012, Proceedings, Part IV 12*, pages 215–230. Springer, 2012.
- Wongun Choi, Khuram Shahid, and Silvio Savarese. What are they doing?: Collective activity classification using spatio-temporal relationship among people. In *2009 IEEE 12th international conference on computer vision workshops, ICCV Workshops*, pages 1282–1289. IEEE, 2009.
- Junyoung Chung, Caglar Gulcehre, Kyunghyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. In *NIPS 2014 Workshop on Deep Learning, December 2014*, 2014.
- Junyoung Chung, Kyle Kastner, Laurent Dinh, Kratarth Goel, Aaron C Courville, and Yoshua Bengio. A recurrent latent variable model for sequential data. *Advances in neural information processing systems*, 28, 2015.
- Bruno Ferreira de Brito, Hai Zhu, Wei Pan, and Javier Alonso-Mora. Social-vrnn: One-shot multi-modal trajectory prediction for interacting pedestrians. In *Conference on Robot Learning*, pages 862–872. PMLR, 2021.
- Chunhua Deng, Xiaoge Kang, Ziqi Zhu, and Shiqian Wu. Behavior recognition based on category subspace in crowded videos. *IEEE Access*, 8:222599–222610, 2020.
- Jacob Devlin. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.
- David Ellis, Eric Sommerlade, and Ian Reid. Modelling pedestrian trajectory patterns with gaussian processes. In *2009 IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops*, pages 1229–1234. IEEE, 2009.
- Geir Evensen. The ensemble kalman filter: Theoretical formulation and practical implementation. *Ocean dynamics*, 53:343–367, 2003.
- Zipei Fan, Xuan Song, Ryosuke Shibasaki, and Ryutaro Adachi. Citymomentum: an online approach for crowd behavior prediction at a citywide level. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pages 559–569, 2015.
- Laurene Fausett. Fundamentals of neural networks: architectures, algorithms, and applications, 1994.
- Sina Feldmann and Juliane Adrian. Forward propagation of a push through a row of people. *Safety science*, 164:106173, 2023.
- Tharindu Fernando, Simon Denman, Sridha Sridharan, and Clinton Fookes. Soft+ hardwired attention: An lstm framework for human trajectory prediction and abnormal event detection. *Neural networks*, 108:466–478, 2018.
- Eric Fosler-Lussier. Markov models and hidden Markov models: A brief tutorial. *International Computer Science Institute*, 1998.
- Zhiqiang Geng, Te Zhang, and Yongming Han. Spatio-temporal attention transformer model for future trajectory forecast. In *2022 IEEE 11th Data Driven*

- Control and Learning Systems Conference (DDCLS)*, pages 1068–1073. IEEE, 2022.
- Francesco Giuliari, Irtiza Hasan, Marco Cristani, and Fabio Galasso. Transformer networks for trajectory forecasting. In *2020 25th international conference on pattern recognition (ICPR)*, pages 10335–10342. IEEE, 2021.
- Abhinav Golas, Rahul Narain, and Ming C Lin. Continuum modeling of crowd turbulence. *Physical review E*, 90(4):042816, 2014.
- Gonzalo Gomez-Nogales, Melania Prieto-Martin, Cristian Romero, Marc Comino-Trinidad, Pablo Ramon-Prieto, Anne-Hélène Olivier, Ludovic Hoyet, Miguel Otaduy, Julien Pettré, and Dan Casas. Resolving collisions in dense 3d crowd animations. *ACM Transactions on Graphics*, 43(5):1–14, 2024.
- Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.
- Tianpei Gu, Guangyi Chen, Junlong Li, Chunze Lin, Yongming Rao, Jie Zhou, and Jiwen Lu. Stochastic trajectory prediction via motion indeterminacy diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17113–17122, 2022.
- 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.
- Stephen J. Guy, Jur van den Berg, Wenxi Liu, Rynson Lau, Ming C. Lin, and Dinesh Manocha. A statistical similarity measure for aggregate crowd dynamics. *ACM Trans. Graph.*, 31(6), November 2012. ISSN 0730-0301. doi: 10.1145/2366145.2366209. URL <https://doi.org/10.1145/2366145.2366209>.
- Kai Han, Yunhe Wang, Hanting Chen, Xinghao Chen, Jianyuan Guo, Zhenhua Liu, Yehui Tang, An Xiao, Chunjing Xu, Yixing Xu, et al. A survey on visual transformer. *arXiv preprint arXiv:2012.12556*, 2020.
- Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Learning spatio-temporal features with 3d residual networks for action recognition. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 3154–3160, 2017.
- Feixiang He, Yuanhang Xiang, Xi Zhao, and He Wang. Informative scene decomposition for crowd analysis, comparison and simulation guidance. *ACM Transactions on Graphics (TOG)*, 2020.
- Feixiang He, Jiangbei Yue, Jialin Zhu, Armin Seyfried, Dan Casas, Julien Pettré, and He Wang. Learning extremely high density crowds as active matters. *arXiv preprint arXiv:2503.12168*, 2025.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- Dirk Helbing and Peter Molnar. Social force model for pedestrian dynamics. *Physical review E*, 51(5):4282, 1995.
- S Hochreiter. Long short-term memory. *Neural Computation MIT-Press*, 1997.

- Kurt Hornik. Approximation capabilities of multilayer feedforward networks. *Neural networks*, 4(2):251–257, 1991.
- Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5):359–366, 1989.
- Yingfan Huang, Huikun Bi, Zhaoxin Li, Tianlu Mao, and Zhaoqi Wang. Stgat: Modeling spatial-temporal interactions for human trajectory prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6272–6281, 2019.
- Roger L Hughes. A continuum theory for the flow of pedestrians. *Transportation Research Part B: Methodological*, 36(6):507–535, 2002.
- Masha Itkina and Mykel Kochenderfer. Interpretable self-aware neural networks for robust trajectory prediction. In *Conference on Robot Learning*, pages 606–617. PMLR, 2023.
- Ajay Jain, Sergio Casas, Renjie Liao, Yuwen Xiong, Song Feng, Sean Segal, and Raquel Urtasun. Discrete residual flow for probabilistic pedestrian behavior prediction. In *Conference on Robot Learning*, pages 407–419. PMLR, 2020.
- Chenfanfu Jiang, Craig Schroeder, Joseph Teran, Alexey Stomakhin, and Andrew Selle. The material point method for simulating continuum materials. In *Acmsiggraph 2016 courses*, pages 1–52. 2016.
- Renhe Jiang, Zekun Cai, Zhaonan Wang, Chuang Yang, Zipei Fan, Quanjun Chen, Kota Tsubouchi, Xuan Song, and Ryosuke Shibasaki. Deepcrowd: A deep model for large-scale citywide crowd density and flow prediction. *IEEE Transactions on Knowledge and Data Engineering*, 35(1):276–290, 2021.
- Muhammad Rizwan Khokher, Abdesselam Bouzerdoum, and Son Lam Phung. Crowd behavior recognition using dense trajectories. In *2014 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, pages 1–7. IEEE, 2014.
- P Kidger. *On neural differential equations*. PhD thesis, University of Oxford, 2021.
- Sujeong Kim, Stephen J Guy, Wenxi Liu, David Wilkie, Rynson WH Lau, Ming C Lin, and Dinesh Manocha. Brvo: Predicting pedestrian trajectories using velocity-space reasoning. *The International Journal of Robotics Research*, 34(2):201–217, 2015.
- Diederik P Kingma. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.
- Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.
- Kris M Kitani, Brian D Ziebart, James Andrew Bagnell, and Martial Hebert. Activity forecasting. In *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7–13, 2012, Proceedings, Part IV 12*, pages 201–214. Springer, 2012.
- Ven Jyn Kok, Mei Kuan Lim, and Chee Seng Chan. Crowd behavior analysis: A review where physics meets biology. *Neurocomputing*, 177:342–362, 2016.
- Vineet Kosaraju, Amir Sadeghian, Roberto Martín-Martín, Ian Reid, Hamid Rezatofighi, and Silvio Savarese. Social-bigat: Multimodal trajectory forecasting

- using bicycle-gan and graph attention networks. *Advances in neural information processing systems*, 32, 2019.
- Yann LeCun, Bernhard Boser, John Denker, Donnie Henderson, Richard Howard, Wayne Hubbard, and Lawrence Jackel. Handwritten digit recognition with a back-propagation network. *Advances in neural information processing systems*, 2, 1989.
- Mihee Lee, Samuel S Sohn, Seonghyeon Moon, Sejong Yoon, Mubbashir Kapadia, and Vladimir Pavlovic. Muse-vae: Multi-scale vae for environment-aware long term trajectory prediction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2221–2230, 2022.
- Alon Lerner, Yiorgos Chrysanthou, and Dani Lischinski. Crowds by example. In *Computer graphics forum*, volume 26, pages 655–664. Wiley Online Library, 2007.
- Lihuan Li, Maurice Pagnucco, and Yang Song. Graph-based spatial transformer with memory replay for multi-future pedestrian trajectory prediction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2231–2241, 2022.
- Yuke Li. A deep spatiotemporal perspective for understanding crowd behavior. *IEEE Transactions on multimedia*, 20(12):3289–3297, 2018.
- Jing Lian, Weiwei Ren, Linhui Li, Yafu Zhou, and Bin Zhou. Ptp-stgcn: pedestrian trajectory prediction based on a spatio-temporal graph convolutional neural network. *Applied Intelligence*, 53(3):2862–2878, 2023.
- Junwei Liang, Lu Jiang, Juan Carlos Niebles, Alexander G Hauptmann, and Li Fei-Fei. Peeking into the future: Predicting future person activities and locations in videos. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5725–5734, 2019.
- Xiaotong Lin, Tianming Liang, Jianhuang Lai, and Jian-Fang Hu. Progressive pretext task learning for human trajectory prediction. In *European Conference on Computer Vision*, pages 197–214. Springer, 2024.
- Chen Liu, Shibo He, Haoyu Liu, and Jiming Chen. Intention-aware denoising diffusion model for trajectory prediction. *IEEE Transactions on Intelligent Transportation Systems*, 2025.
- Tianshan Liu, Rui Zhao, and Kin-Man Lam. Multimodal-semantic context-aware graph neural network for group activity recognition. In *2021 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2021.
- Tianshan Liu, Rui Zhao, Kin-Man Lam, and Jun Kong. Visual-semantic graph neural network with pose-position attentive learning for group activity recognition. *Neurocomputing*, 491:217–231, 2022a.
- Yao Liu, Binghao Li, Xianzhi Wang, Claude Sammut, and Lina Yao. Attention-aware social graph transformer networks for stochastic trajectory prediction. *IEEE Transactions on Knowledge and Data Engineering*, 2024.
- Yu Liu, Yuexin Zhang, Kunming Li, Yongliang Qiao, Stewart Worrall, You-Fu Li, and He Kong. Knowledge-aware graph transformer for pedestrian trajectory prediction. In *2023 IEEE 26th International Conference on Intelligent Transportation Systems (ITSC)*, pages 4360–4366. IEEE, 2023.

- Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3202–3211, 2022b.
- Francesco Longobardi and Daniel Riccio. Graphic-graph-based representation for analyzing people’s high-level interactions in crowds. In *2024 IEEE International Conference on Image Processing (ICIP)*, pages 868–874. IEEE, 2024.
- Linbo Luo, Suiping Zhou, Wentong Cai, Malcolm Yoke Hean Low, Feng Tian, Yongwei Wang, Xian Xiao, and Dan Chen. Agent-based human behavior modeling for crowd simulation. *Computer Animation and Virtual Worlds*, 19(3-4):271–281, 2008.
- David JC MacKay et al. Introduction to gaussian processes. *NATO ASI series F computer and systems sciences*, 168:133–166, 1998.
- Bappaditya Mandal, Jiri Fajtl, Vasileios Argyriou, Dorothy Monekosso, and Paolo Remagnino. Deep residual network with subclass discriminant analysis for crowd behavior recognition. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 938–942. IEEE, 2018.
- 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. In *European Conference on Computer Vision*, pages 759–776. Springer, 2020.
- Weibo Mao, Chenxin Xu, Qi Zhu, Siheng Chen, and Yanfeng Wang. Leapfrog diffusion model for stochastic trajectory prediction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5517–5526, 2023.
- Franjo Matkovic, Marina Ivasic-Kos, and Slobodan Ribaric. A new approach to dominant motion pattern recognition at the macroscopic crowd level. *Engineering applications of artificial intelligence*, 116:105387, 2022.
- Zhaobin Mo, Yongjie Fu, and Xuan Di. Pi-neugode: Physics-informed graph neural ordinary differential equations for spatiotemporal trajectory prediction. In *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems*, pages 1418–1426, 2024.
- Abduallah Mohamed, Kun Qian, Mohamed Elhoseiny, and Christian Clauzel. 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*, pages 14424–14432, 2020.
- Alessio Monti, Alessia Bertugli, Simone Calderara, and Rita Cucchiara. Dag-net: Double attentive graph neural network for trajectory forecasting. In *2020 25th international conference on pattern recognition (ICPR)*, pages 2551–2558. IEEE, 2021.
- Vittorio Murino, Marco Cristani, Shishir Shah, and Silvio Savarese. The group and crowd analysis interdisciplinary challenge. In *Group and Crowd Behavior for Computer Vision*, pages 1–11. Elsevier, 2017.
- Nishant Nikhil and Brendan Tran Morris. Convolutional neural network for trajectory prediction. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 186–196, 2018.

- 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.
- Javier Sánchez Pérez, Enric Meinhardt-Llopis, and Gabriele Facciolo. G tv-l1 optical flow estimation image process. *On Line*, 2013:137–150, 2013.
- Marwa Qaraqe, Yin David Yang, Elizabeth B Varghese, Emrah Basaran, and Almiqdad Elzein. Crowd behavior detection: leveraging video swin transformer for crowd size and violence level analysis. *Applied Intelligence*, 54(21):10709–10730, 2024.
- Amir Rasouli and John K Tsotsos. Autonomous vehicles that interact with pedestrians: A survey of theory and practice. *IEEE transactions on intelligent transportation systems*, 21(3):900–918, 2019.
- Fariba Rezaei and Mehran Yazdi. Real-time crowd behavior recognition in surveillance videos based on deep learning methods. *Journal of Real-Time Image Processing*, 18(5):1669–1679, 2021.
- Daniela Ridel, Nachiket Deo, Denis Wolf, and Mohan Trivedi. Scene compliant trajectory forecast with agent-centric spatio-temporal grids. *IEEE Robotics and Automation Letters*, 5(2):2816–2823, 2020.
- Alexandre Robicquet, Amir Sadeghian, Alexandre Alahi, and Silvio Savarese. Learning social etiquette: Human trajectory understanding in crowded scenes. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part VIII 14*, pages 549–565. Springer, 2016.
- David E Rumelhart, James L McClelland, PDP Research Group, et al. *Parallel distributed processing, volume 1: Explorations in the microstructure of cognition: Foundations*. The MIT press, 1986.
- Amir Sadeghian, Vineet Kosaraju, Agrim Gupta, Silvio Savarese, and Alexandre Alahi. Trajnet: Towards a benchmark for human trajectory prediction. *arXiv preprint*, 2018.
- 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/CVF Conference on Computer Vision and Pattern Recognition*, pages 1349–1358, 2019.
- Haifeng Sang, Jinyu Wang, Quankai Liu, Wangxing Chen, and Zishan Zhao. Physics constrained pedestrian trajectory prediction with probability quantification. *Expert Systems with Applications*, 255:124743, 2024.
- Shobhit Saxena, François Brémond, Monique Thonnat, and Ruihua Ma. Crowd behavior recognition for video surveillance. In *Advanced Concepts for Intelligent Vision Systems: 10th International Conference, ACIVS 2008, Juan-les-Pins, France, October 20–24, 2008. Proceedings 10*, pages 970–981. Springer, 2008.
- Michael Schmidt and Hod Lipson. Distilling free-form natural laws from experimental data. *science*, 324(5923):81–85, 2009.
- Erik Schuetz and Fabian B Flohr. A review of trajectory prediction methods for the vulnerable road user. *Robotics*, 13(1):1, 2023.

- 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.
- Jing Shao, Chen Change Loy, and Xiaogang Wang. Scene-independent group profiling in crowd. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2219–2226, 2014.
- Jing Shao, Kai Kang, Chen Change Loy, and Xiaogang Wang. Deeply learned attributes for crowded scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4657–4666, 2015.
- Jing Shao, Chen-Change Loy, Kai Kang, and Xiaogang Wang. Slicing convolutional neural network for crowd video understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5620–5628, 2016.
- Liushuai Shi, Le Wang, Chengjiang Long, Sanping Zhou, Mo Zhou, Zhenxing Niu, and Gang Hua. Sgcn: Sparse graph convolution network for pedestrian trajectory prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8994–9003, 2021.
- Liushuai Shi, Le Wang, Sanping Zhou, and Gang Hua. Trajectory unified transformer for pedestrian trajectory prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9675–9684, 2023.
- Xiaodan Shi, Xiaowei Shao, Zhiling Guo, Guangming Wu, Haoran Zhang, and Ryosuke Shibasaki. Pedestrian trajectory prediction in extremely crowded scenarios. *Sensors*, 19(5):1223, 2019.
- Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting. *Advances in neural information processing systems*, 28, 2015.
- Aliyu Nuhu Shuaibu, Aamir Saeed Malik, and Ibrahima Faye. Adaptive feature learning cnn for behavior recognition in crowd scene. In *2017 IEEE International Conference on Signal and Image Processing Applications (ICSIPA)*, pages 357–361. IEEE, 2017.
- Bogdan Ilie Sighencea, Rares Ion Stanciu, and Cătălin Daniel Căleanu. A review of deep learning-based methods for pedestrian trajectory prediction. *Sensors*, 21(22):7543, 2021.
- Bogdan Ilie Sighencea, Ion Rares Stanciu, and Cătălin Daniel Căleanu. D-stgcn: Dynamic pedestrian trajectory prediction using spatio-temporal graph convolutional networks. *Electronics*, 12(3):611, 2023.
- K Simonyan and A Zisserman. Very deep convolutional networks for large-scale image recognition. In *3rd International Conference on Learning Representations (ICLR 2015)*. Computational and Biological Learning Society, 2015.
- Shawn Singh, Mubbasir Kapadia, Petros Faloutsos, and Glenn Reinman. Steerbench: a benchmark suite for evaluating steering behaviors. *Comput. Animat. Virtual Worlds*, 20(5-6):533–548, September 2009. ISSN 1546-4261.
- Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015.

- Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. *Advances in neural information processing systems*, 28, 2015.
- Berkan Solmaz, Brian E Moore, and Mubarak Shah. Identifying behaviors in crowd scenes using stability analysis for dynamical systems. *IEEE transactions on pattern analysis and machine intelligence*, 34(10):2064–2070, 2012.
- Xiao Song, Kai Chen, Xu Li, Jinghan Sun, Baocun Hou, Yong Cui, Baochang Zhang, Gang Xiong, and Zilie Wang. Pedestrian trajectory prediction based on deep convolutional lstm network. *IEEE Transactions on Intelligent Transportation Systems*, 22(6):3285–3302, 2020.
- Hang Su, Hua Yang, Shiba Zheng, Yawen Fan, and Sha Wei. The large-scale crowd behavior perception based on spatio-temporal viscous fluid field. *IEEE Transactions on Information Forensics and security*, 8(10):1575–1589, 2013.
- Hang Su, Yinpeng Dong, Jun Zhu, Haibin Ling, and Bo Zhang. Crowd scene understanding with coherent recurrent neural networks. In *IJCAI*, volume 1, page 2, 2016.
- HY Swathi, G Shivakumar, and HS Mohana. Crowd behavior analysis: A survey. In *2017 international conference on recent advances in electronics and communication technology (ICRAECT)*, pages 169–178. IEEE, 2017.
- Rei Tamaru, Pei Li, and Bin Ran. Enhancing pedestrian trajectory prediction with crowd trip information. *arXiv preprint arXiv:2409.15224*, 2024.
- Masato Tamura. Design and analysis of efficient attention in transformers for social group activity recognition. *International Journal of Computer Vision*, pages 1–20, 2024.
- Masato Tamura, Rahul Vishwakarma, and Ravigopal Venkela. Hunting group clues with transformers for social group activity recognition. In *European Conference on Computer Vision*, pages 19–35. Springer, 2022.
- Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 402–419. Springer, 2020.
- Myo Thida, Yoke Leng Yong, Pau Climent-Pérez, How-lung Eng, and Paolo Re-magnino. A literature review on video analytics of crowded scenes. *Intelligent Multimedia Surveillance: Current Trends and Research*, pages 17–36, 2013.
- John Toner and Yuhai Tu. Long-range order in a two-dimensional dynamical xy model: how birds fly together. *Physical review letters*, 75(23):4326, 1995.
- Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015.
- Hung Tran, Vuong Le, and Truyen Tran. Goal-driven long-term trajectory prediction. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 796–805, 2021.
- Habib Ullah, Sultan Daud Khan, Mohib Ullah, Faouzi Alaya Cheikh, and Muhammad Uzair. Two stream model for crowd video classification. In *2019 8th european workshop on visual information processing (EUVIP)*, pages 93–98. IEEE, 2019.

- Habib Ullah, Ihtesham Ul Islam, Mohib Ullah, Muhammad Afaq, Sultan Daud Khan, and Javed Iqbal. Multi-feature-based crowd video modeling for visual event detection. *Multimedia Systems*, 27:589–597, 2021.
- SA Vahora and NC Chauhan. Deep neural network model for group activity recognition using contextual relationship. *Engineering Science and Technology, an International Journal*, 22(1):47–54, 2019.
- Jur Van den Berg, Ming Lin, and Dinesh Manocha. Reciprocal velocity obstacles for real-time multi-agent navigation. In *2008 IEEE international conference on robotics and automation*, pages 1928–1935. Ieee, 2008.
- Wouter van Toll, Cédric Braga, Barbara Solenthaler, and Julien Pettré. Extreme-density crowd simulation: combining agents with smoothed particle hydrodynamics. In *Proceedings of the 13th ACM SIGGRAPH Conference on Motion, Interaction and Games*, pages 1–10, 2020.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.
- Vivek Velayutham, Sanjay Kumar, Avinash Kumar, Shrinwantu Raha, and Gonesh Chandra Saha. Analysis of deep learning in real-world applications: Challenges and progress. *Tuijin Jishu/Journal of Propulsion Technology*, 44(2):2023.
- Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In *International Conference on Learning Representations*, 2018.
- Jia Wan, Ziquan Liu, and Antoni B Chan. A generalized loss function for crowd counting and localization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1974–1983, 2021.
- Chengxin Wang, Shaofeng Cai, and Gary Tan. Graphtcn: Spatio-temporal interaction modeling for human trajectory prediction. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 3450–3459, 2021.
- He Wang and Carol O’Sullivan. Globally continuous and non-markovian crowd activity analysis from videos. In *The European Conference on Computer Vision (ECCV)*, 2016.
- He Wang, Jan Ondřej, and Carol O’Sullivan. Path patterns: Analyzing and comparing real and simulated crowds. In *ACM SIGGRAPH Symposium on Interactive 3D Graphics and Games (ACM I3D)*, pages 49–57, 2016.
- He Wang, Jan Ondřej, and Carol O’Sullivan. Trending paths: A new semantic-level metric for comparing simulated and real crowd data. *IEEE Transactions on Visualization and Computer Graphics (TVCG)*, 2017a.
- Jingdong Wang, Naiyan Wang, You Jia, Jian Li, Gang Zeng, Hongbin Zha, and Xian-Sheng Hua. Trinary-projection trees for approximate nearest neighbor search. *IEEE transactions on pattern analysis and machine intelligence*, 36(2):388–403, 2013.
- Jinyu Wang, Haifeng Sang, Quankai Liu, Wangxing Chen, and Zishan Zhao. Neural differential constraint-based pedestrian trajectory prediction model in ego-centric

- perspective. *Engineering Applications of Artificial Intelligence*, 133:107993, 2024.
- Minsi Wang, Bingbing Ni, and Xiaokang Yang. Recurrent modeling of interaction context for collective activity recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3048–3056, 2017b.
- Xiaogang Wang, Xiaoxu Ma, and W.E.L. Grimson. Unsupervised activity perception in crowded and complicated scenes using hierarchical bayesian models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(3):539–555, 2009. doi: 10.1109/TPAMI.2008.87.
- Xiaogang Wang, Keng Teck Ma, Gee-Wah Ng, and W Eric L Grimson. Trajectory analysis and semantic region modeling using nonparametric hierarchical bayesian models. *International journal of computer vision*, 95:287–312, 2011.
- Xinlei Wei, Junping Du, Zhe Xue, Meiyu Liang, Yue Geng, Xin Xu, and JangMyung Lee. A very deep two-stream network for crowd type recognition. *Neurocomputing*, 396:522–533, 2020.
- Nanda Wijermans. *Understanding crowd behaviour*. PhD thesis, University of Groningen, 2011.
- Weishang Wu and Xiaoheng Deng. Motion latent diffusion for stochastic trajectory prediction. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6665–6669. IEEE, 2024.
- Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and S Yu Philip. A comprehensive survey on graph neural networks. *IEEE transactions on neural networks and learning systems*, 32(1):4–24, 2020.
- Wei Xiang, YIN Haoteng, He Wang, and Xiaogang Jin. Socialcvae: Predicting pedestrian trajectory via interaction conditioned latents. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 6216–6224, 2024.
- Chenxin Xu, Weibo Mao, Wenjun Zhang, and Siheng Chen. Remember intentions: Retrospective-memory-based trajectory prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6488–6497, 2022a.
- Kaiping Xu, Zheng Qin, Guolong Wang, Kai Huang, Shuxiong Ye, and Huidi Zhang. Collision-free lstm for human trajectory prediction. In *MultiMedia Modeling: 24th International Conference, MMM 2018, Bangkok, Thailand, February 5-7, 2018, Proceedings, Part I* 24, pages 106–116. Springer, 2018a.
- Pei Xu, Jean-Bernard Hayet, and Ioannis Karamouzas. Socialvae: Human trajectory prediction using timewise latents. In *European Conference on Computer Vision*, pages 511–528. Springer, 2022b.
- Yanyu Xu, Zhixin Piao, and Shenghua Gao. Encoding crowd interaction with deep neural network for pedestrian trajectory prediction. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5275–5284, 2018b.
- Zhenwei Xu, Qing Yu, Wushouer Slamu, Yaoyong Zhou, and Zhida Liu. S-cgru: An efficient model for pedestrian trajectory prediction. In *International Conference on Neural Information Processing*, pages 244–259. Springer, 2023.

- Hao Xue, Du Q Huynh, and Mark Reynolds. Ss-lstm: A hierarchical lstm model for pedestrian trajectory prediction. In *2018 IEEE winter conference on applications of computer vision (WACV)*, pages 1186–1194. IEEE, 2018.
- Kota Yamaguchi, Alexander C Berg, Luis E Ortiz, and Tamara L Berg. Who are you with and where are you going? In *CVPR 2011*, pages 1345–1352. IEEE, 2011.
- Liqi Yan, Mingjian Zhu, and Changbin Yu. Crowd video captioning. *arXiv preprint arXiv:1911.05449*, 2019.
- Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.
- Shanwen Yang, Tianrui Li, Xun Gong, Bo Peng, and Jie Hu. A review on crowd simulation and modeling. *Graphical Models*, 111:101081, 2020.
- Yuxin Yang, Pengfei Zhu, Mengshi Qi, and Huadong Ma. Following in the footsteps: Predicting human trajectories using motion pattern memory. In *Proceedings of the 6th ACM International Conference on Multimedia in Asia*, pages 1–7, 2024.
- Cunjun Yu, Xiao Ma, Jiawei Ren, Haiyu Zhao, and Shuai Yi. Spatio-temporal graph transformer networks for pedestrian trajectory prediction. In *Computer Vision-ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16*, pages 507–523. Springer, 2020.
- Qing Yu, Zhenwei Xu, Yaoyong Zhou, Zhida Liu, and Wushouer Silamu. Pedestrian trajectory prediction using spatio-temporal vae. In *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*, pages 297–311. Springer, 2024.
- Ye Yuan, Xinshuo Weng, Yanglan Ou, and Kris M Kitani. Agentformer: Agent-aware transformers for socio-temporal multi-agent forecasting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9813–9823, 2021.
- Jiangbei Yue, Dinesh Manocha, and He Wang. Human trajectory prediction via neural social physics. In *European conference on computer vision*, pages 376–394. Springer, 2022.
- Jiangbei Yue, Dinesh Manocha, and He Wang. Human trajectory forecasting with explainable behavioral uncertainty. *arXiv preprint arXiv:2307.01817*, 2023.
- Jiangbei Yue, Baiyi Li, Julien Pettré, Armin Seyfried, and He Wang. Human motion prediction under unexpected perturbation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1501–1511, 2024.
- Simone Zamboni, Zekarias Tilahun Kefato, Sarunas Girdzijauskas, Christoffer Norén, and Laura Dal Col. Pedestrian trajectory prediction with convolutional neural networks. *Pattern Recognition*, 121:108252, 2022.
- Matthew D Zeiler, Graham W Taylor, and Rob Fergus. Adaptive deconvolutional networks for mid and high level feature learning. In *2011 international conference on computer vision*, pages 2018–2025. IEEE, 2011.
- Beibei Zhan, Dorothy N Monekosso, Paolo Remagnino, Sergio A Velastin, and Li-Qun Xu. Crowd analysis: a survey. *Machine Vision and Applications*, 19:345–357, 2008.
- Chi Zhang, Christian Berger, and Marco Dozza. Social-iwstcnn: A social interaction-weighted spatio-temporal convolutional neural network for pedestrian trajectory

- prediction in urban traffic scenarios. In *2021 IEEE Intelligent Vehicles Symposium (IV)*, pages 1515–1522. IEEE, 2021.
- Ethan Zhang, Neda Masoud, Mahdi Bandegi, Joseph Lull, and Rajesh K Malhan. Step attention: Sequential pedestrian trajectory prediction. *IEEE Sensors Journal*, 22(8):8071–8083, 2022.
- Gang Zhang, Yang Geng, and Zhao G Gong. A comprehensive review of deep learning approaches for group activity analysis. *The Visual Computer*, 41(3):1733–1755, 2025.
- Pu Zhang, Wanli Ouyang, Pengfei Zhang, Jianru Xue, and Nanning Zheng. Sr-lstm: State refinement for lstm towards pedestrian trajectory prediction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12085–12094, 2019.
- Yanbo Zhang and Liying Zheng. Pedestrian trajectory prediction with mlp-social-gru. In *Proceedings of the 2021 13th International Conference on Machine Learning and Computing*, pages 368–372, 2021.
- Mingbi Zhao, Jinghui Zhong, and Wentong Cai. A role-dependent data-driven approach for high-density crowd behavior modeling. *ACM Transactions on Modeling and Computer Simulation (TOMACS)*, 28(4):1–25, 2018.
- Bolei Zhou, Xiaogang Wang, and Xiaou Tang. Understanding collective crowd behaviors: Learning a mixture model of dynamic pedestrian-agents. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2871–2878. IEEE, 2012.
- Bolei Zhou, Xiaou Tang, and Xiaogang Wang. Measuring crowd collectiveness. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3049–3056, 2013.
- Hao Zhou, Dongchun Ren, Huaxia Xia, Mingyu Fan, Xu Yang, and Hai Huang. Ast-gnn: An attention-based spatio-temporal graph neural network for interaction-aware pedestrian trajectory prediction. *Neurocomputing*, 445:298–308, 2021.
- Yanshan Zhou, Pingrui Lai, Jiaqi Yu, Yingjie Xiong, and Hua Yang. Hydrodynamics-informed neural network for simulating dense crowd motion patterns. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 4553–4561, 2024.
- Zhou Zhou, Gang Huang, Zhaoxin Su, Yongfu Li, and Wei Hua. Dynamic attention-based cvae-gan for pedestrian trajectory prediction. *IEEE Robotics and Automation Letters*, 8(2):704–711, 2022.
- Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A Efros, Oliver Wang, and Eli Shechtman. Toward multimodal image-to-image translation. *Advances in neural information processing systems*, 30, 2017.
- Manli Zhu and Aleix M Martinez. Subclass discriminant analysis. *IEEE transactions on pattern analysis and machine intelligence*, 28(8):1274–1286, 2006.
- Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. In *International Conference on Learning Representations*, 2021.
- Yanliang Zhu, Deheng Qian, Dongchun Ren, and Huaxia Xia. Starnet: Pedestrian trajectory prediction using deep neural network in star topology. In *2019 IEEE/RSJ*

*International Conference on Intelligent Robots and Systems (IROS)*, pages 8075–8080. IEEE, 2019.

Yuqi Zuo, Aymen Hamrouni, Hakim Ghazzai, and Yehia Massoud. V3trans-crowd: A video-based visual transformer for crowd management monitoring. In *2023 IEEE International Conference on Smart Mobility (SM)*, pages 154–159. IEEE, 2023.