Dense Crowd Motion Prediction through Density and Trend Maps

Tingting Wang[1], Qiang Fu[†1], Minggang Wang[1], Huikun Bi[2], Qixin Deng[3] and Zhigang Deng[4]

[1]Beijing University of Posts and Telecommunications, China
[2]SenseTime, China
[3]Wabash College, USA
[4]University of Houston, USA

Abstract

In this paper we propose a novel density/trend map based method to predict both group behavior and individual pedestrian motion from video input. Existing motion prediction methods represent pedestrian motion as a set of spatial-temporal trajectories; however, besides such a per-pedestrian representation, a high-level representation for crowd motion is often needed in many crowd applications. Our method leverages density maps and trend maps to represent the spatial-temporal states of dense crowds. Based on such representations, we propose a crowd density map net that extracts a density map from a video clip, and a crowd prediction net that utilizes the historical states of a video clip to predict density maps and trend maps for future frames. Moreover, since the crowd motion consists of the motion of individual pedestrians in a group, we also leverage the predicted crowd motion as a clue to improve the accuracy of traditional trajectory-based motion prediction methods. Through a series of experiments and comparisons with state-of-the-art motion prediction methods, we demonstrate the effectiveness and robustness of our method.

CCS Concepts

• ***Computing methodologies*** → *Neural networks; Tracking;*

1. Introduction

In recent years, significant advances in various fields, including autonomous driving, traffic monitoring, and robotic navigation, have led to a surge in research interest surrounding pedestrian motion prediction within computer graphics, computer vision, and robotics communities. This endeavor aims to forecast the future positions of pedestrians based on their previous movements.

Traditional motion prediction techniques focus on per-pedestrian positions at a specific time represented by coordinate data. Such coordinate-based motion representation serves as the cornerstone for numerous methods designed to capture the spatio-temporal patterns of pedestrian motion [AGR*16, GJFF*18, HBL*19, SICP20, CCLM21]. On the other hand, considering the inherently stochastic nature of human motion in real-world scenarios characterized by intricate and diverse interactions, some existing methods adopt a strategy of predicting multiple coordinate sequences as potential future trajectories for a single pedestrian [GJFF*18, HBL*19, SICP20, XHK22]. Furthermore, some techniques incorporate additional contextual features, such as background images, to better model the complex interactions between pedestrians and their surroundings. Such approaches can enhance the accuracy of predicted pedestrian trajectories, bringing them closer to ground truth [MZC*20, SICP20, BZM*20].

Although generative models can now produce future pedestrian movements based on historical frames, ensuring content consistency remains challenging for these models. As a result, traditional motion prediction methods continue to be the most critical solution for tasks involving pedestrian behavior prediction in surveillance videos and autonomous driving scenarios. However, the demands of run-time pedestrian motion prediction pose challenges in applying these trajectory-based methods effectively across various real-world applications. First, the input frames in trajectory-based methods require accurately tracked or labeled trajectories, leading to tedious data labeling tasks or additional computation. Second, for most trajectory-based methods that focus on per-pedestrian coordinates, the computation overhead and complexity increase with the growth of the number of pedestrians. Last, due to accumulated errors, the accuracy of predictions decreases as the length of predicted frames increases. To tackle the above challenges, a high-level representation for a group of pedestrians that have consistent motion behavior needs to be investigated.

In this paper, we propose a novel representation method for dense crowd motion, and further leverage it for crowd behavior prediction including individual pedestrian trajectory prediction. Specifically, to represent the high-level characteristics of crowd

† Corresponding author: fu.john.qiang@gmail.com

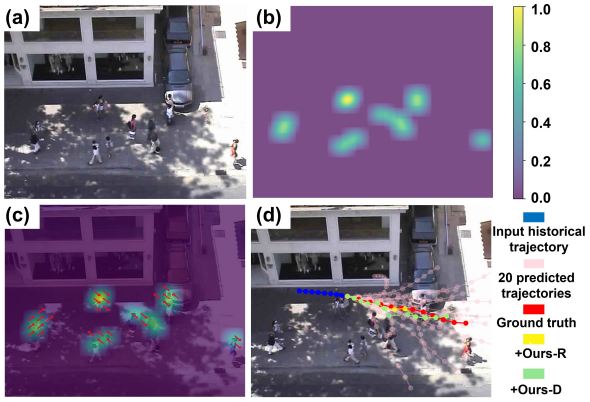

Figure 1: *For a video frame with a dense crowd (a), our method employs a density map (b) as well as the motion trend (red arrows) to represent dense crowd motion (c). The predicted crowd motion can promote the trajectory prediction of individual pedestrians (d).*

motion, we introduce *density maps*, whose pixel values show the number of pedestrians in each pixel (Figure 1-(b)) to depict how a group of pedestrians is distributed. We also introduce *trend maps* (illustrated with red arrows in Figure 1) to represent how the motion trends of pedestrians evolve within crowded scenes, which prioritize capturing the collective spatial-temporal characteristics of a dense crowd over individual pedestrian movements (Figure 1-(c)). To this end, we design novel neural network architectures that use previous frames to predict future pedestrian distributions. Our model excels in predicting dense crowd scenes, even for those cases where annotated pedestrian coordinates are unavailable. Even though such a representation focuses more on the crowd motion of pedestrians in a group, it can promote the trajectory prediction of individual pedestrians as a constraint. For example in Figure 1-(d), for the input historical pedestrian coordinates of an individual pedestrian, we use the trajectory-based method to generate 20 potential trajectories first, and then use the relation vector from the group center to pedestrians (+Ours-R) and the density values of pedestrians (+Ours-D) as constraints to select pedestrian coordinates from the predicted trajectories.

The main contributions of this work include:

- We introduce a novel method for representing pedestrian motion in crowd videos, and a two-stage framework that harnesses such representations to achieve accurate motion prediction of dense crowds in various in-the-wild videos.
- We propose an approach that leverages the predicted crowd motion of a group of pedestrians to refine the trajectory prediction of individual pedestrians in the group.

2. Related Work

In this section, we will first review previous work related to pedestrian trajectory prediction. Then, we delve into existing work that employs density maps and optical flow for representing crowd motion. Finally, we review recent advances in the realm of convolutional LSTM networks, which play a pivotal role in our method.

Pedestrian Trajectory Prediction. Most previous research efforts in pedestrian trajectory prediction, including [AGR*16, GJFF*18, HBL*19, CCLM21, GCL*22, XMZC22, SICP20, XPG18, XHK22], revolve around the accurate prediction of pedestrian position sequences. Broadly, these methods utilize historical pedestrian position sequences as input to forecast future pedestrian positions. To enhance the accuracy and plausibility of these predictions, various strategies have been employed. These strategies include modeling pedestrian interactions [AGR*16], predicting multiple potential pedestrian trajectories [GJFF*18, HBL*19, XHK22], and incorporating contextual scene information [SICP20], among others. Besides, the transformer model has also been effectively applied to motion prediction [SJDS22] and even to multimodal motion of multiple agents [SJDS24]. In computer graphics, particle simulation methods (i.e., [vTCB*21]) can model the movement of large crowds, which can be used to generate realistic crowd motion data.

Recent endeavors have sought to address this limitation by focusing on predicting pedestrian trajectories at the image level. In one category of approaches, images rather than coordinates serve as the input data. Ma et al. [MZC*20], for example, bypass the annotation process and directly derive pedestrian coordinates from input images. Similarly, Su et al. [SPSC19] introduce the concept of potential fields as input data. However, it should be noted that despite these innovations, the output remains in the form of pedestrian coordinates. In another category of approaches, the output aims to represent pedestrian trajectories as images, as demonstrated in [MICB19]. Nevertheless, these methods continue to rely on pedestrian coordinates as their input data. While our method mainly addresses the description and prediction of group movements, it also enhances per-pedestrian trajectory prediction using group motion information. This proves particularly beneficial in densely crowded scenes within target videos, where traditional methods for per-pedestrian trajectory prediction often fall short due to occlusions among pedestrians. As a result, we propose a candidate set recommendation strategy that utilizes density map predictions to improve per-pedestrian trajectory prediction. This endows our method with generalizability, making it applicable to various trajectory prediction models. Additionally, compared to directly using the predicted density map to optimize trajectory positions, candidate sets can reduce the scope of the optimization solution space and minimize the impact of density map prediction errors on trajectory prediction.

Density Maps for Crowd Representation. Utilizing density maps is a highly effective approach for characterizing the density and spatial distribution of objects within images, particularly in scenarios with crowded spaces, such as estimating the presence of pedestrians. In recent years, density maps have gained widespread popularity in crowd-counting tasks, which are designed to estimate the number of individuals present in images. Numerous convolutional neural network (CNN)-based methods have demonstrated remarkable performance in crowd-counting tasks by harnessing the power of density maps, as described below.

Early efforts [FXL*15, WZY*15] introduced CNN-based models for crowd counting, relying solely on the original images without additional feature inputs. Zhang et al. [ZZC*16] pioneered the Multi-column Convolutional Neural Network (MCNN), which employs multi-size filters to extract pertinent features. Jiang et

al. [JZX*20] designed a highly effective CNN-based method that autonomously adjusts the density estimation for each corresponding sub-region through learned scaling factors. In a noteworthy departure from counting specific objects, Ranjan et al. [RSNH21] put forth a novel CNN-based model capable of counting objects across various categories. Additionally, Lin et al. [LMJ*22] introduced an attention mechanism to enhance the performance of their model. All of these methods leverage density maps to estimate the number of people or objects within a given image, showcasing significant advancements in this domain. These notable advancements in leveraging density maps for crowd-related tasks have inspired our approach, wherein we position density maps as a fundamental component of our method.

Optical Flow for Motion Representation. Representing motion information in images or video is a pivotal task in the domains of image processing and computer vision. Optical flow, a widely-used technique, serves as a valuable means to convey the motion of objects across consecutive frames. Optical flow encompasses both sparse and dense variants, with dense optical flow providing pixel-level motion offsets within an image. Farnebck et al. [Far03] were pioneers in leveraging a polynomial expansion transform to compute dense optical flow, marking a significant milestone in this field. Recently, CNN-based methods [DFI*15, IMS*17, ZXZ*19, ZZZ*22] have demonstrated remarkable success in estimating optical flow, yielding impressive results.

In addition to optical flow, alternative techniques have emerged to represent motion information within images. For example, Su et al. [SPSC19] introduced the concept of potential fields as an interpretable and unified representation to model pedestrian motion within crowded scenes, where variations in the potential field signify motion trends. Shi et al. [SWL*21] introduced a sparse directed spatial graph to capture motion tendencies, although it does not directly portray motion within the image. Wu and Yao [WYWL21] proposed transient variations and motion trends to depict object movements in images.

In our work, we use optical flow to generate one of the inputs for our network. Optical flow, in conjunction with historical frames, plays a vital role in predicting crowd motion in future frames, contributing to the effectiveness of our approach.

3. Methodology

3.1. Problem Formulation

As illustrated in Figure 2, for scenes with dense crowds (a), the traditional per-pedestrian trajectory representation struggles to describe the spatial clustering and temporal consistency of crowd motion (b). Therefore, we propose using density maps to represent the spatial positions of crowds (c). Additionally, to better characterize the temporal movement directions of crowd motion, we employ *Trend Maps* which help distinguish between nearby groups in the density maps, instead of using optical flow representation. For visualization purposes, the movement directions (d) can be approximated to four (e) or eight (f) directions. In this paper, we use four directions (NE, NW, SE, and SW) to present trend maps.

Our method consists of two modules: (1) a CNN-based encoder-decoder (*crowd density map net*) to capture the static spatial crowd

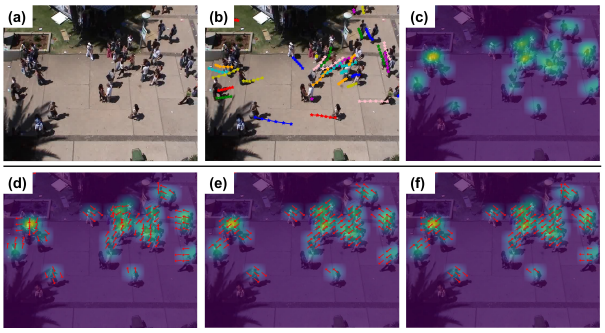

Figure 2: ***Top:*** *the comparison between trajectory-based (b) and density-map-based (c) motion representation for the crowd scene (a).* ***Bottom:*** *the red arrows represent the trend maps with the original motion directions (d) as well as the directions approximated to four (e) and eight (f).*

group distribution from an image frame as a density map (Section 3.2). (2) A ConvLSTM-based predictor (*Crowd Prediction Net*) that leverages optical flow images and the crowd density map net to predict future crowd motion trends (Section 3.3). Building upon the predicted crowd density map, we further use the relation between group and individual pedestrians as constraints, to refine the results of the trajectory-based motion prediction methods.

3.2. Crowd Density Map Net

The primary goal of the first module is to leverage optical flow to generate crowd density maps ($\hat{D}_t, t \in [1,T]$) that represent the spatial distributions of the crowd in the image sequences ($I_t, t \in [1,T]$).

Optical Flow: Optical flow is a well-known technique that reflects changes in pixel values between successive images [BB95]. The motion of the pedestrians in the continuous frames always causes certain pixels occupied by the pedestrians to shift, while the values of these pixels tend to remain approximately the same. This leads to the following equation: $I(x,y,t) = I(x+\delta x, y+\delta y, t+\delta t)$, where $I(x,y,t)$ is the pixel value of the image at position (x,y) at time t, and $\delta x, \delta y$, and δt are the changes in the x-, y-, and time directions, respectively [Far03, LK*81]. The changes in δx and δy are reflected in the optical flow. The optical flow includes sparse optical flow and dense optical flow, depending on the number of pixels used in the calculation. Dense optical flow is more significant for crowd motion because it reflects the movement of all pixels in an image than sparse optical flow [AWS00]. Thus, we employ the dense optical flow method by Farneback [Far03] to obtain corresponding optical flow images $F_1, F_2, \ldots, F_T$ of given $I_1, I_2, \ldots, I_T$.

Our crowd density map net involves a CNN-based encoder-decoder model, illustrated in Figure 3. The model takes in a sequence of original images $I_1, I_2, \ldots, I_T$ and their corresponding optical flow images $F_1, F_2, \ldots, F_T$, while the output is a series of crowd density maps $\{\hat{D}_t\}$. To extract features from the input images and optical flow images, we use an encoder E^{img} for capturing pedestrian distribution features and an encoder E^{of} for capturing spatial distribution variation features between consecutive optical

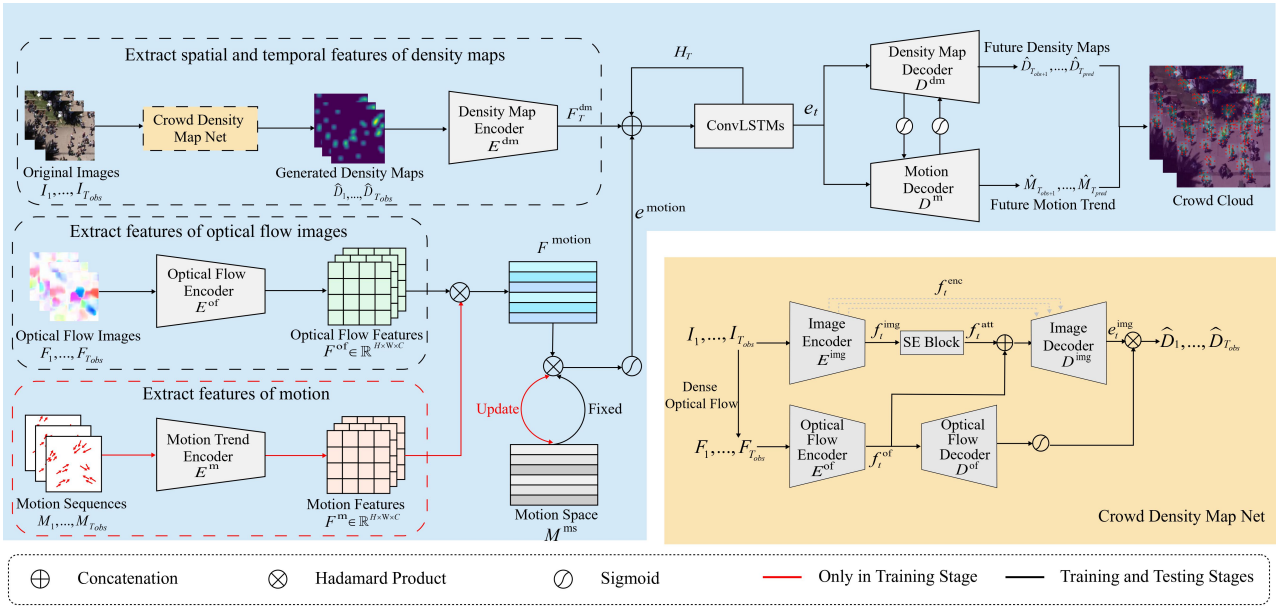

Figure 3: *Our method consists of two stages. In the first stage, we generate density maps from images using Image and Optical Flow Encoders to extract features and decoders to generate the maps. In the second stage, we use a prediction net to update Motion Space with Optical Flow and Motion Trend information during training. We then combine motion space features with the extracted density map features and input them into ConvLSTMs for prediction. Finally, we use decoders to generate future density maps and motion trends.*

flow images. During the training phase, we fine-tune the model using VGG-19 [SZ14] pre-trained on ImageNet [DDS*09] for both encoders. However, convolutional layers may cause the loss of certain local features, such as texture features. To generate stronger crowd density maps by fusing local features with global information in the decoding stage, we keep the features after pooling layers in the encoding stage, denoted as f_t^{enc}.

$$\begin{aligned} f_t^{\text{img}} &= E^{\text{img}}(I_t; W^{\text{img}}) \\ f_t^{\text{of}} &= E^{\text{of}}(F_t; W^{\text{of}}) \end{aligned}, \tag{1}$$

where $E^{\text{img}}(.)$ and $E^{\text{of}}(.)$ have identical structures but different parameters. W^{img} and W^{of} are both learnable parameters. f_t^{img} is the features with static distribution of the crowd and f_t^{of} is the features with motion interactions. I_t represents the image at $t = 1, 2, \ldots, T$ time-step.

Considering pedestrians often have limited space to move in crowded scenarios, we introduce the squeeze and excitation block (SE block) [HSS18] into our model, to better capture the global features of local crowd groups and others. The SE block is an attention model that can select significant portions of feature maps. As shown in Figure 4, the image features $f_t^{\text{img}} \in \mathbb{R}^{H \times W \times C}$ are first passed through a convolution operation that aggregates the feature maps of each channel across spatial dimensions, producing an embedding feature $z \in \mathbb{R}^{1 \times 1 \times C}$. To fully capture all channel-wise dependencies, the SE block employs a simple gating mechanism with a sigmoid activation: $s = \sigma(W_1 \delta(W_2 z))$, where W_1 and W_2 are fully connected layers, and δ is a ReLU layer. Finally, the re-weighted image features f_t^{att}, which emphasize the crowd motion region in the images, are obtained by multiplying f_t^{img} with s.

To decode the features f_t^{att} and f_t^{of}, we use their respective decoders. As shown in Figure 3, we maintain output and input size consistency for the decoder by using the reverse setting on the encoder. To create the input for the image decoder, we concatenate f_t^{att} and f_t^{of}. Furthermore, we concatenate the image features f_t^{enc} saved in the Image Encoder after each pooling layer before each upsampling layer in the Image Decoder. Similarly, we also concatenate the image features f_t^{img} stored in the Image Encoder following each pooling layer before each upsampling layer in the Image Decoder. For the optical flow decoder, we use f_t^{of} as the input. On the other hand, we capture the output features of the optical flow decoder by an activation function and use it to update the output of the Image Decoder.

$$\begin{aligned} e_t^{\text{img}} &= D^{\text{img}}([f_t^{\text{att}}; f_t^{\text{of}}]; f_t^{\text{enc}}; W^{\text{d}}) \\ \hat{D}_t &= e_t^{\text{img}} \sigma(D^{\text{of}}(f_t^{\text{of}}; W^{\text{d}})) \end{aligned}, \tag{2}$$

where D^{img} and D^{of} are the decoders, which share learnable parameters represented as W^{d}, and $\sigma(.)$ denotes the sigmoid function.

In our crowd density map net, we use a pixel-wise L2 loss as follows:

$$L^{\text{dm}} = \frac{1}{N} \sum_{i=0}^{N} \|\hat{D}_t - D_t\|_2^2. \tag{3}$$

Here, $\hat{D}_t$ represents the generated density maps using our crowd

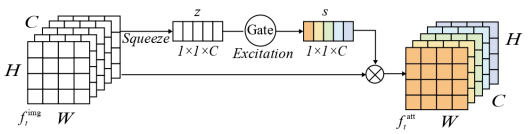

Figure 4: *SE Block is an attention model that updates the feature maps of f_t^{img}. Each feature map is first compressed to obtain z through a convolution operation (squeeze stage). Then, z is activated through a gate mechanism to produce s, which is used to update the original f_t^{img} feature maps (excitation stage).*

density map net, while D_t represents the ground truth calculated by the precise locations of pedestrian trajectories.

3.3. Crowd Prediction Net

In our crowd prediction net, we feed the generated crowd density map sequences $\hat{D}_t (t \in [1,T])$ into a density map encoder E^{dm} to extract additional crowd distribution features. Crowd density maps provide information about the static quantity and distribution of the crowd in the scenario, unlike raw images captured from video. As a result, we do not need to use a complex convolution-based network (such as VGG-19 used in our crowd density map network). E^{dm} is a structure with a basic block unit (Conv2d + Elu). Each density map in a sequence is fed separately into E^{dm} to extract the corresponding distribution features F_t^{dm} as follows:

$$F_t^{\text{dm}} = E^{\text{dm}}(\hat{D}_t; W^{\text{dm}}), \tag{4}$$

where W^{dm} represents the learnable parameters of E^{dm}. We obtain a sequence of features $\{F_1^{\text{dm}}, F_2^{\text{dm}}, \ldots, F_T^{\text{dm}}\}$ that encode the spatial interactions of pedestrians in the aforementioned density maps. The crowd spatial feature sequence derived from the crowd density maps are not sufficient to characterize the movement relationship of the pedestrians' time series. To extract additional crowd temporal motion information, we employ a convolutional LSTM network (ConvLSTM) [SCW*15]. The ConvLSTM retains similar cell states C_t, hidden states H_t, and three gates i_t, f_t, o_t as the LSTM, $t \in [1,T]$. After processed by the ConvLSTMs, H_T will contain information about both the dynamic motion and the static crowd spatial distribution of the density map sequences.

Specifically, at the training stage (Figure 3), the optical flow feature extractor E^{of} processes optical flow images F_t to extract pixel-wise motion information F^{of}, and the motion feature extractor E^{m} is used to extract motion information F^{m} from motion trend maps M_t, representing motion in real-world coordinates. Both extractors share a common unit structure, consisting of a Conv3d layer followed by an Elu activation. We also employ a Motion Space $M^{\text{ms}} \in \mathbb{R}^{HW \times C}$, which is a memory module with its internal parameters continuously updated during the training stage. During each training iteration, specific motion data F^{motion} which is calculated by a Hadamard product operation to merge F^{of} and F^{m}, is stored into M^{ms}. This makes the motion space a dynamic repository of generalized motion patterns used to guide accurate future predictions.

At the runtime stage, the Motion Space, pre-populated with motion patterns from the training phase, does not require updates. The runtime stage solely involves extracting and applying these stored patterns to new data, focusing particularly on motion information derived from image-based optical flows. This retrieval is operationalized through a formula $e^{\text{motion}} = \sigma(\hat{M}^{\text{ms}} \bigotimes F^{\text{of}})$, where the motion patterns in M^{ms} are matched with current optical flow inputs to predict movements. The extracted motion patterns, denoted as e^{motion}, combine with historical density data from the maps (H_T) to predict the features e_t via the ConvLSTM:

$$e_t = ConvLSTM([H_T; e^{\text{motion}}; F_T^{\text{dm}}]; W^{\text{e}}). \tag{5}$$

The features e_t, containing future distribution and motion information, are separately fed into D^{dm} and D^{m} for the prediction of future density maps and motion trends as follows:

$$\hat{D}_t^{i+1} = \begin{cases} D^{\text{dm}}(e_t; W^{\text{dm}}), & i = 0 \\ D^{\text{dm}}([\hat{D}_t^i; Sigmoid(\hat{M}_t^i)]; W^{\text{dm}}), & otherwise \end{cases} \tag{6}$$

$$\hat{M}_t^{i+1} = \begin{cases} D^{\text{m}}(e_t; W^{\text{m}}), & i = 0 \\ D^{\text{m}}([\hat{M}_t^i; Sigmoid(\hat{D}_t^i)]; W^{\text{m}}), & otherwise \end{cases} \tag{7}$$

where $t = T+1, T+2, \ldots, \hat{T}$, and i indicates the i-th layer of the decoder; and W^{dm} and W^{m} are learnable parameters. The output of the last layer of the decoder includes the predicted density map $\hat{D}_t$ and the predicted motion trend $\hat{M}_t$. Note D^{dm} and D^{m} have the same structural (2D DeConv) but do not share weights.

In this step, the total loss L consists of three L_2 losses below:

$$L = L^{\text{dm}} + \omega_1 L^{\text{m}} + \omega_2 L^{\text{f}}, \tag{8}$$

where $L^{\text{dm}} = \frac{1}{N}\sum_{i=0}^{N} \|\hat{D}_i - D_i\|_2^2$ is the density map loss to constrain the generated density maps, $L^{\text{m}} = \frac{1}{N}\sum_{i=0}^{N} \|\hat{M}_i - M_i\|_2^2$ is the motion trend loss to constrain the generated motion trends, and $L^{\text{f}} = \|\hat{D}_T - D_T\|_2^2$ is used to make the generated last frame close to the ground truth. ω_1 and ω_2 are hyperparameters: $\omega_1 = 0.001$ and $\omega_2 = 0.1$.

3.4. Individual Motion Prediction

The pedestrians in the same group always have similar motion speeds and directions, which is caused by crowd behaviors including attraction, repulsion, and alignment. In this way, we can further determine the crowd motion from a pedestrian group to individual pedestrians. For example, in Figure 5, the pedestrian groups (yellow circles) can be extracted from the density map, and individual and group motions (e.g., the blue and red points in Figure 5-Top) typically exhibit coherence within each pedestrian group. Although our method does not directly predict the pedestrians' motion trajectories, the motion coherence between group and individual pedestrian allows us to extend our method to predict trajectories of individual pedestrians. This is achieved by using our predicted crowd motion to enhance existing pedestrian trajectory prediction models, thus ascertaining more accurate pedestrian coordinates. Specifically, we utilize the density values or relation vectors to link individual pedestrians with the group. The former represents the den-

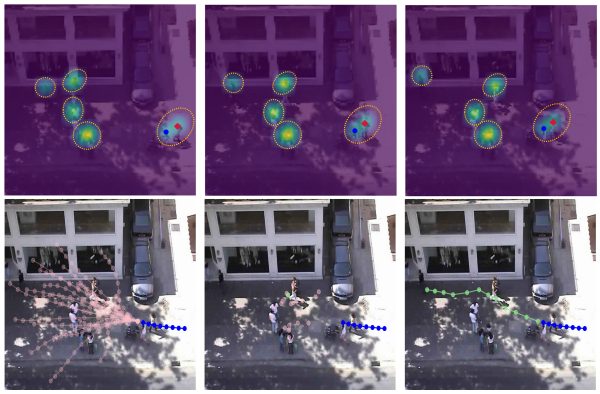

Figure 5: ***Top:*** *Three frame examples overlaid with density maps. We highlight the groups with yellow circles and use the blue and red points to illustrate a pedestrian location and the associated group center, respectively.* ***Bottom:*** *The 20 predicted trajectories (left) lead to 20 candidate pedestrian coordinates in each frame (middle). We pick one from the 20 candidates in each frame to form a new trajectory (right).*

sity value at a pedestrian's location, and the latter is a vector from the group center to the pedestrian's location.

As illustrated in Figure 5-Bottom, for an individual pedestrian, we use the trajectory-based pedestrian motion prediction method to generate N potential pedestrian trajectories. Then, we calculate the average constraint from the historical frames as the constraint (denoted as C), and predict future frames' density maps by using our method. In this manner, in each predicted frame, multiple pedestrian coordinates $\{p_i\}, i = 1, \cdots, N$ are obtained, and we can select pedestrian coordinate p_i that is closest to C in terms of density value or relation vector (green point in Figure 5-Bottom-Middle). A new trajectory can be formed with such selected coordinates across all predicted frames, as the optimized pedestrian motion prediction outcome.

4. Experimental Results

In this section, we first present qualitative evaluations on dense crowd motion representations and predictions, by using both the public dataset and new scenarios. Then, we quantitatively evaluate our method on individual pedestrian trajectory prediction, cooperating with existing trajectory-based motion prediction methods [HBL*19, XHK22, GJFF*18].

4.1. Qualitative Evaluation

Figure 6 shows the prediction results on three scenes of the UCY dataset. In terms of distributions, our predicted distribution is close to the actual pedestrian distribution in the image. In terms of motion trends, our predicted motion trends can reflect the direction of the pedestrians' motion and are consistent with the changes in the distribution of pedestrians. These qualitative results illustrate that we can represent the pedestrian motion at the image level.

Additionally, we also collect new dense crowd videos to show the predicted future dense crowd motion in new scenarios. Figure 7 presents the performance of our crowd density map net on new scenarios. We first use 10% of the data (only 200 density maps per scenario on average) to fine-tune our model. Notably, we did not incorporate crowd motion trends into the model's fine-tuning. The produced results correspond well to the crowd distribution in the raw image. This demonstrates that our method can predict crowd motion without extra trajectory annotation and is easy to transform into new scenarios.

In the individual pedestrian trajectory prediction, as illustrated by the three examples in Figure 8, the top case in each example shows 20 predicted trajectories generated using the trajectory prediction method (SocialVAE [XHK22]) based on the blue input trajectories. The bottom cases display the associated trajectories refined by our method (in green and yellow) compared with the Ground Truth. These examples demonstrate that traditional trajectory prediction methods for pedestrian motion struggle to ensure consistent results, thereby reducing the accuracy of predicted trajectories. By cooperating with our method, more accurate pedestrian coordinates can be extracted from the candidate trajectories generated by the prediction method, resulting in new trajectories that are much closer to the Ground Truth.

4.2. Quantitative Evaluation

To validate our network structure, we conducted ablation experiments to compare our results with those obtained by removing the Crowd Density Map Net (i.e., directly using historical frames as the input for the prediction network) and by removing the fusion of density map features and motion trend features. We selected the prediction frames in T=15 & T=19 and used missing rates to measure the network's performance. The missing rate is the percentage of joint prediction samples that deviate from the ground truth beyond certain thresholds [SHG*22, ECC*21]. As shown in Table 1, the predictions for earlier frames (T=15) are more accurate than those for later frames (T=19). Utilizing the Crowd Density Map Net significantly reduces the missing rates of the predictions, and the fusion operation also enhances prediction accuracy. This indicates that the trend map contributes positively to the prediction of the density map in our network.

The predicted crowd motion can promote the trajectory prediction of individual pedestrian. We perform the state-of-the-art trajectory-based motion prediction methods [HBL*19, XHK22, GJFF*18] on both ETH & UCY datasets which contain real-world pedestrian trajectories from five different scenes. Specifically, we retrained the compared models to align with the output sequences of our approach, which predicts 24 future frames based on 8 historical frames and chooses 12 frames via interval sampling. During testing, we had the trajectory prediction method generate 20 different predicted trajectories for pedestrian movement. This means each frame includes coordinates for 20 candidate pedestrians. Then, we selectively chose pedestrian coordinates to form a new trajectory using the group-pedestrian vector (+Ours-R) and density value (+Ours-D). In tables 2, we respectively compared the optimization effects of our method on previous trajectory prediction methods [HBL*19, XHK22, GJFF*18]. For the results of the trajectory prediction methods, we used the average of 20 trajecto-

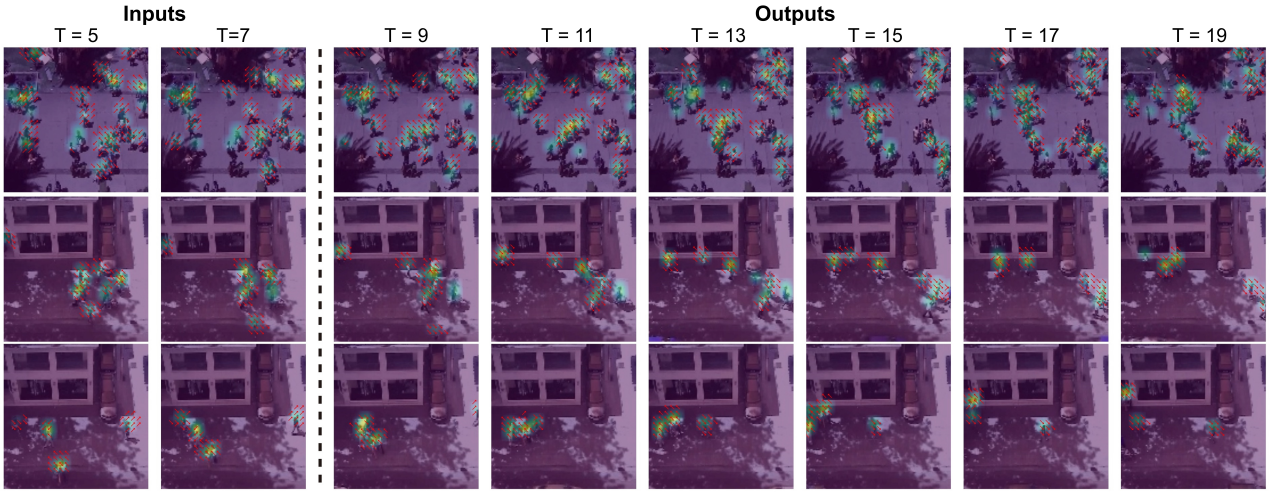

Figure 6: *Prediction results on three scenes of the UCY dataset. We show 2 from 8 input historical frames on the left and 6 from 12 predicted density/trend maps on the right.*

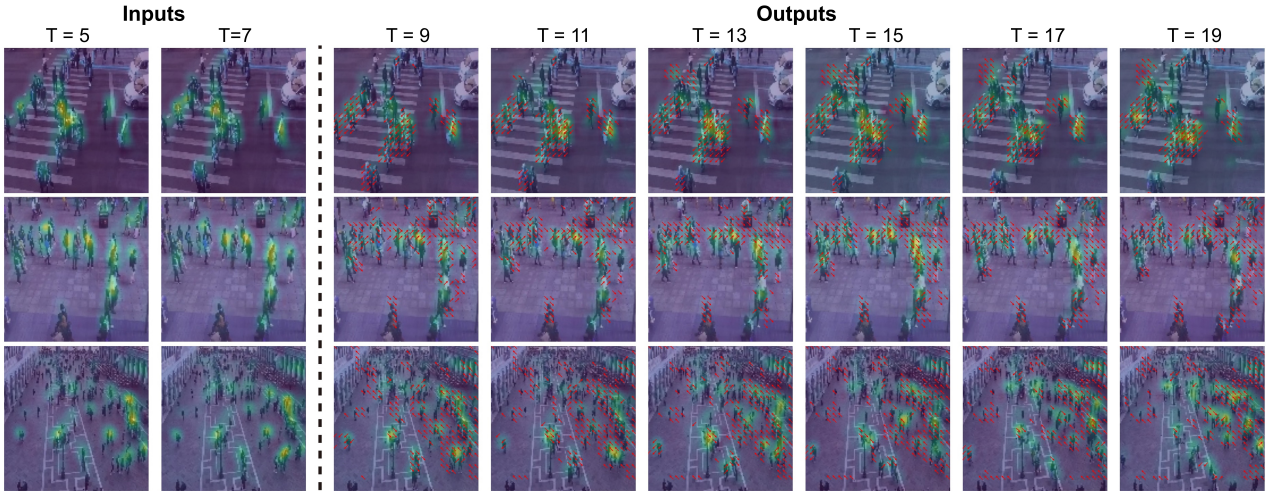

Figure 7: *Prediction results on new data. We show 2 from 8 input historical frames on the left and 6 from 12 predicted density/trend maps on the right.*

ries' ADE and FDE as their performance. It can be seen that, our method is able to improve the accuracy of traditional pedestrian trajectory prediction methods. However, the performance of our method is also dependent on the underlying trajectory prediction algorithm.

The results demonstrate that our method's predicted crowd motion can cooperate with trajectory prediction methods and improve the individual pedestrian's motion prediction accuracy. Besides, we can see that the density value and relation vector constraints have comparable effects on enhancing the accuracy of predicted trajectories. However, in simple scenes with few pedestrians (e.g., Zara1), the performance of such two constraints could be influenced by the adopted trajectory prediction methods. This might be due to the fact that in scenes with fewer pedestrians, groups often contain only one pedestrian, making the relation vector so small that it is easily affected by computational errors. Consequently, our method is more suitable for crowded scenes.

Since the output of our Crowd Density Map Net represents the spatial distribution of the crowd within a frame, we also conducted a quantitative comparison with the crowd-counting method FAMNET [RSNH21]. For the five scenes in the ETH & UCY datasets, FAMNET achieved an average MAE and MSE of 1.67 and 3.68, respectively, whereas our method yielded results of 1.02 and 2.47, respectively. This significant improvement can be attributed to the fact that the FAMNET is a single-image-based method, whereas our Crowd Density Map Net accounts for the temporal continuity of pedestrians across frames.

Table 1: *The ablation experiment results summarized by Missing Rates for $T = 15$ and $T = 19$ on both sides of the vertical bar, respectively. 'w/o CDMN' denotes our method without the Crowd Density Map Net, while 'w/o fusion' denotes our method without the fusion of density map features and motion trends features. The unit of measurement for the data is percentage (%).*

	ETH	HOTEL	UNIV	ZARA1	ZARA2	AVG
Ours(w/o CDMN)	35.06 | 40.25	37.22 | 45.75	24.67 | 31.59	30.14 | 38.05	31.20 | 39.85	31.66 | 39.10
Ours(w/o fusion)	16.32 | 19.40	19.02 | 22.13	9.83 | 14.14	11.25 | 17.56	12.03 | 17.96	13.69 | 18.24
Ours	**15.87** | **19.32**	**17.88** | **21.54**	**9.76** | **13.80**	**10.15** | **16.42**	**11.23** | **16.59**	**12.98** | **17.33**

Table 2: *The pedestrian trajectory prediction accuracies of methods STGAT, SocialVAE, SocialGAN, and the associated results cooperated with our method, respectively. For each cell, we present the ADE on the left and the FDE on the right.*

	Zara1	Zara2	Hotel	Univ	Eth
STGAT	1.43 | 5.59	0.63 | 2.37	0.42 | 1.53	**1.10** | **4.15**	1.26 | 4.98
+Ours-D	1.40 | 5.63	0.53 | 2.04	**0.35** | **1.31**	1.11 | 4.19	0.99 | 3.96
+Ours-R	**1.16** | **4.69**	**0.52** | **1.98**	**0.35** | 1.34	**1.10** | 4.16	**0.66** | **2.87**
SocialVAE	1.95 | 4.59	1.01 | 2.44	1.06 | 2.42	1.94 | 4.29	2.44 | 5.20
+Ours-D	**1.90** | **4.40**	**0.88** | 2.01	**0.54** | **1.31**	**1.80** | **3.85**	2.10 | 4.20
+Ours-R	2.01 | 4.49	0.90 | **2.00**	0.58 | 1.38	1.95 | 4.04	**2.09** | **3.87**
SocialGAN	1.56 | 3.46	1.01 | 2.22	1.49 | 3.18	1.79 | 3.65	0.62 | 1.25
+Ours-D	**1.55** | **3.42**	0.92 | 2.04	1.42 | 3.08	**1.78** | 3.62	0.66 | 1.19
+Ours-R	1.58 | 3.54	**0.91** | **1.95**	**1.38** | **2.99**	**1.78** | **3.61**	**0.50** | **0.99**

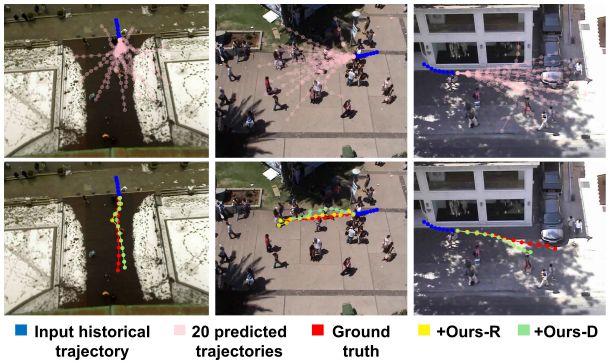

Figure 8: *Three examples of the 20 predicted trajectories (Top), and the associated trajectories suggested by our method compared to the ground truth (Bottom).*

5. Discussion and Conclusion

In this paper, we introduce a new approach to modeling and predicting dense crowd motion in video, including density maps and trend maps, and that captures the collective spatial-temporal characteristics of a dense crowd over individual pedestrian movements. Building upon the representation, we further design highly effective algorithms to predict the motion of dense crowds in video. Our experimental results show that our method can not only predict the crowd motion of pedestrian groups, but also improve the accuracy of trajectory for individual pedestrian cooperated with the trajectory-based motion prediction methods.

Our work has some drawbacks that need to be addressed in future work. One of them is that our method is more suitable for dense crowds rather than scenes with sparse pedestrians. Another challenge is that our method may confuse background objects that look like pedestrians, such as fire hydrants and postboxes, with actual pedestrians, or miss pedestrians that are occluded by other objects. This may result in inaccurate density maps and poor future motion prediction. Despite these limitations, we believe that our method is a practical and novel approach for representing and predicting dense crowd motion from video frames.

Acknowledgements

This work was partially supported by grants from the State Administration of Science, Technology and Industry for National Defense (HTKJ2023KL502004). Zhigang Deng was in part supported by NSF IIS-2005430.

References

[AGR*16] ALAHI A., GOEL K., RAMANATHAN V., ROBICQUET A., FEI-FEI L., SAVARESE S.: Social lstm: Human trajectory prediction in crowded spaces. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2016), pp. 961–971. 1, 2

[AWS00] ALVAREZ L., WEICKERT J., SÁNCHEZ J.: Reliable estimation of dense optical flow fields with large displacements. *International Journal of Computer Vision 39* (2000), 41–56. 3

[BB95] BEAUCHEMIN S. S., BARRON J. L.: The computation of optical flow. *ACM computing surveys (CSUR) 27*, 3 (1995), 433–466. 3

[BZM*20] BI H., ZHANG R., MAO T., DENG Z., WANG Z.: How can i see my future? fvtraj: Using first-person view for pedestrian trajectory prediction. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VII 16* (2020), Springer, pp. 576–593. 1

[CCLM21] CHOI C., CHOI J. H., LI J., MALLA S.: Shared cross-modal trajectory prediction for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2021), pp. 244–253. 1, 2

[DDS*09] DENG J., DONG W., SOCHER R., LI L.-J., LI K., FEI-FEI L.: Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition* (2009), Ieee, pp. 248–255. 4

[DFI*15] DOSOVITSKIY A., FISCHER P., ILG E., HAUSSER P., HAZIRBAS C., GOLKOV V., VAN DER SMAGT P., CREMERS D., BROX T.: Flownet: Learning optical flow with convolutional networks. In *Proceedings of the IEEE international conference on computer vision* (2015), pp. 2758–2766. 3

[ECC*21] ETTINGER S., CHENG S., CAINE B., LIU C., ZHAO H., PRADHAN S., CHAI Y., SAPP B., QI C. R., ZHOU Y., ET AL.: Large scale interactive motion forecasting for autonomous driving: The waymo open motion dataset. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2021), pp. 9710–9719. 6

[Far03] FARNEBÄCK G.: Two-frame motion estimation based on polynomial expansion. In *Scandinavian conference on Image analysis* (2003), Springer, pp. 363–370. 3

[FXL*15] FU M., XU P., LI X., LIU Q., YE M., ZHU C.: Fast crowd density estimation with convolutional neural networks. *Engineering Applications of Artificial Intelligence 43* (2015), 81–88. 2

[GCL*22] GU T., CHEN G., LI J., LIN C., RAO Y., ZHOU J., LU J.: Stochastic trajectory prediction via motion indeterminacy diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022), pp. 17113–17122. 2

[GJFF*18] GUPTA A., JOHNSON J., FEI-FEI L., SAVARESE S., ALAHI A.: Social gan: Socially acceptable trajectories with generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2018), pp. 2255–2264. 1, 2, 6

[HBL*19] HUANG Y., BI H., LI Z., MAO T., WANG Z.: Stgat: Modeling spatial-temporal interactions for human trajectory prediction. In *Proceedings of the IEEE/CVF international conference on computer vision* (2019), pp. 6272–6281. 1, 2, 6

[HSS18] HU J., SHEN L., SUN G.: Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2018), pp. 7132–7141. 4

[IMS*17] ILG E., MAYER N., SAIKIA T., KEUPER M., DOSOVITSKIY A., BROX T.: Flownet 2.0: Evolution of optical flow estimation with deep networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2017), pp. 2462–2470. 3

[JZX*20] JIANG X., ZHANG L., XU M., ZHANG T., LV P., ZHOU B., YANG X., PANG Y.: Attention scaling for crowd counting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2020), pp. 4706–4715. 3

[LK*81] LUCAS B. D., KANADE T., ET AL.: *An iterative image registration technique with an application to stereo vision*, vol. 81. Vancouver, 1981. 3

[LMJ*22] LIN H., MA Z., JI R., WANG Y., HONG X.: Boosting crowd counting via multifaceted attention. *arXiv preprint arXiv:2203.02636* (2022). 3

[MICB19] MAKANSI O., ILG E., CICEK O., BROX T.: Overcoming limitations of mixture density networks: A sampling and fitting framework for multimodal future prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2019), pp. 7144–7153. 2

[MZC*20] MA Y., ZHU X., CHENG X., YANG R., LIU J., MANOCHA D.: Autotrajectory: Label-free trajectory extraction and prediction from videos using dynamic points. In *European Conference on Computer Vision* (2020), Springer, pp. 646–662. 1, 2

[RSNH21] RANJAN V., SHARMA U., NGUYEN T., HOAI M.: Learning to count everything. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2021), pp. 3394–3403. 3, 7

[SCW*15] SHI X., CHEN Z., WANG H., YEUNG D.-Y., WONG W.-K., WOO W.-C.: Convolutional lstm network: A machine learning approach for precipitation nowcasting. *Advances in neural information processing systems 28* (2015). 5

[SHG*22] SUN Q., HUANG X., GU J., WILLIAMS B. C., ZHAO H.: M2i: From factored marginal trajectory prediction to interactive prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022), pp. 6543–6552. 6

[SICP20] SALZMANN T., IVANOVIC B., CHAKRAVARTY P., PAVONE M.: Trajectron++: Dynamically-feasible trajectory forecasting with heterogeneous data. In *European Conference on Computer Vision* (2020), Springer, pp. 683–700. 1, 2

[SJDS22] SHI S., JIANG L., DAI D., SCHIELE B.: Motion transformer with global intention localization and local movement refinement. *Advances in Neural Information Processing Systems 35* (2022), 6531–6543. 2

[SJDS24] SHI S., JIANG L., DAI D., SCHIELE B.: Mtr++: Multi-agent motion prediction with symmetric scene modeling and guided intention querying. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2024). 2

[SPSC19] SU S., PENG C., SHI J., CHOI C.: Potential field: Interpretable and unified representation for trajectory prediction. *arXiv preprint arXiv:1911.07414* (2019). 2, 3

[SWL*21] SHI L., WANG L., LONG C., ZHOU S., ZHOU M., NIU Z., HUA G.: Sgcn: Sparse graph convolution network for pedestrian trajectory prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2021), pp. 8994–9003. 3

[SZ14] SIMONYAN K., ZISSERMAN A.: Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014). 4

[vTCB*21] VAN TOLL W., CHATAGNON T., BRAGA C., SOLENTHALER B., PETTRÉ J.: Sph crowds: Agent-based crowd simulation up to extreme densities using fluid dynamics. *Computers & Graphics 98* (2021), 306–321. 2

[WYWL21] WU H., YAO Z., WANG J., LONG M.: Motionrnn: A flexible model for video prediction with spacetime-varying motions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2021), pp. 15435–15444. 3

[WZY*15] WANG C., ZHANG H., YANG L., LIU S., CAO X.: Deep people counting in extremely dense crowds. In *Proceedings of the 23rd ACM international conference on Multimedia* (2015), pp. 1299–1302. 2

[XHK22] XU P., HAYET J.-B., KARAMOUZAS I.: Socialvae: Human trajectory prediction using timewise latents. In *European Conference on Computer Vision* (2022), Springer, pp. 511–528. doi:10.1007/978-3-031-19772-7_30. 1, 2, 6

[XMZC22] XU C., MAO W., ZHANG W., CHEN S.: Remember intentions: Retrospective-memory-based trajectory prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022), pp. 6488–6497. 2

[XPG18] XU Y., PIAO Z., GAO S.: Encoding crowd interaction with deep neural network for pedestrian trajectory prediction. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2018), pp. 5275–5284. 2

[ZXZ*19] ZHAI M., XIANG X., ZHANG R., LV N., EL SADDIK A.: Learning optical flow using deep dilated residual networks. *IEEE Access 7* (2019), 22566–22578. 3

[ZZC*16] ZHANG Y., ZHOU D., CHEN S., GAO S., MA Y.: Single-image crowd counting via multi-column convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2016), pp. 589–597. 2

[ZZZ*22] ZHAO S., ZHAO L., ZHANG Z., ZHOU E., METAXAS D.: Global matching with overlapping attention for optical flow estimation. *arXiv preprint arXiv:2203.11335* (2022). 3