Appendix

A Personal Space Definition

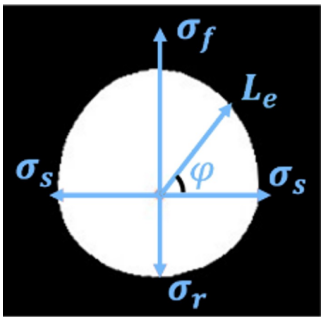

Figure 6: A sample personal space

An example of the personal space is shown in Fig. 6. Each personal space is first constructed by identifying the variances along the four principle axes to the agent's front, sides and rear, defined respectively as:

$$\begin{aligned} \sigma_f^i &= \max(2v^i, 0.5), \\ \sigma_s^i &= 2\sigma_f^i/3 \\ \sigma_r^i &= \sigma_f^i/2 \end{aligned} \quad . \tag{15}$$

Based on the four principle axis, the personal space for agent i is represented as a set of boundary points.

$$\mathcal{P}^i(q^i) = \{b^i(\phi), \phi \in [0, 2\pi)\}, \tag{16}$$

where

$$\begin{aligned} b^i(\phi) &= s^i + L_e(\phi) \begin{pmatrix} \cos(\theta^i + \phi) \\ \sin(\theta^i + \phi) \end{pmatrix}, \\ L_e(\phi) &= \sqrt{\frac{C}{\cos^2 \gamma/(2\sigma_1) + \sin^2 \gamma/(2\sigma_2)}}, \\ \gamma &= \text{mod}(\phi, \pi/2), \\ (\sigma_1, \sigma_2) &= \begin{cases} (\sigma_f^i, \sigma_s^i), & \text{if } 0 \le \phi < \pi/2 \\ (\sigma_s^i, \sigma_r^i), & \text{if } \pi/2 \le \phi < \pi \\ (\sigma_r^i, \sigma_s^i), & \text{if } \pi \le \phi < 3\pi/2 \\ (\sigma_s^i, \sigma_f^i), & \text{if } 3\pi/2 \le \phi < 2\pi \end{cases} \end{aligned} \quad . \tag{17}$$

C is a constant used to adjust the scale of the personal space.

B Partial Input Handling

Note that in a dynamic pedestrian scene, we will have frequent occurrences of partial inputs for individual agents or groups due to new agents entering the scene or new groups being formed respectively. Therefore, our prediction model must be able to handle cases in which the input is complete up to a past window $t_{\hat{h}}$ with $t_{\hat{h}} = t - \hat{h}$, $\hat{h} < h$, i.e., $\boldsymbol{\mathcal{G}}_{t_{\hat{h}}:t}$. To handle these cases, for time $t_h < \tau < t_{\hat{h}}$, we compute $\mathcal{G}_\tau^j$ by making the following membership assumptions:

- For any agent $i \in G_t^j$ such that $i \notin G_\tau^j$ and for whom we have the complete state history $s_{t_h:t}^i$, we set $g_\tau^i = j$. In other words, the prior group membership of any recent members of group j is set to j (despite agent i possibly being a member of another group j' at those instances).

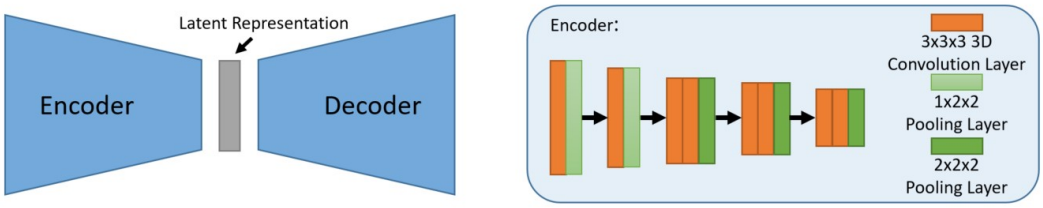

Figure 7: Our simple encoder-decoder network's architecture. The decoder's deconvolution layers mirror the layout of the encoder.

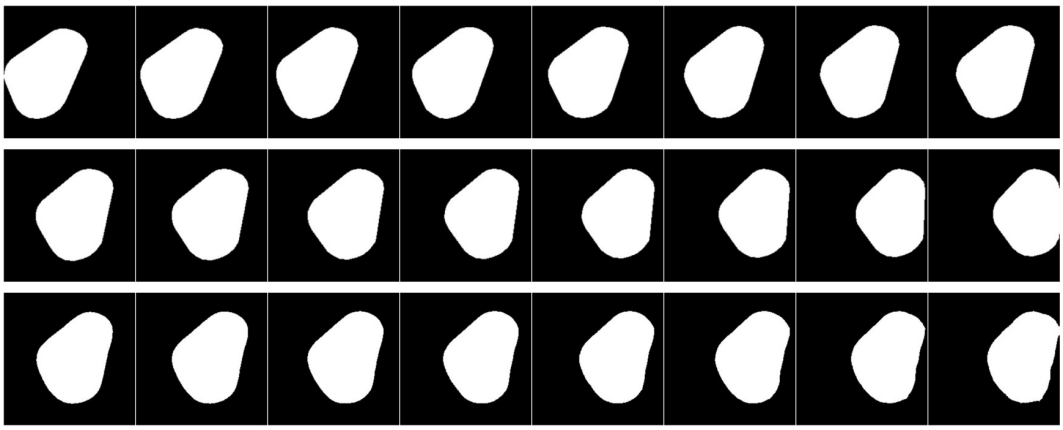

Figure 8: Top: An example group space input sequence for our encoder-decoder network. Mid: The ground truth future sequence of the group. Bottom: The predictioned future sequence of the group as outputted by our encoder-decoder network.

- For any agent $i \in G_t^j$ such that $i \notin G_\tau^j$ and for whom we only have *partial* state history $s^i_{t_{\hat{h}}:t}$, we take the agent's last known state s_t^i and velocity u_t^i and back propagate it as $s^i_{\tau-1} = s^i_\tau - u^i_\tau dt$.

Given a small h, these assumptions should reflect a close approximation of the group's complete history state, because pedestrian group switching process is gradual and pedestrian movements are smooth and predictive.

C Prediction Model Details

Our encoder-decoder network largely leverages [50]'s C3D network. As shown in Fig. 7, the encoder architecture contains the following layers (beginning from the input layer): one $3\times3\times3$ convolution layer with 64 channels, one $1 \times 2 \times 2$ maxpool layer, one $3 \times 3 \times 3$ convolution layer with 128 channels, another $1 \times 2 \times 2$ maxpool layer, another $3 \times 3 \times 3$ convolution layer with 128 channels, one $3 \times 3 \times 3$ convolution layer with 256 channels, one $2 \times 2 \times 2$ maxpool layer, another $3 \times 3 \times 3$ convolution layer with 256 channels, one $3 \times 3 \times 3$ convolution layer with 512 channels, another $2 \times 2 \times 2$ maxpool layer, two $3 \times 3 \times 3$ convolution layers with 512 channels and another $2 \times 2 \times 2$ maxpool layer.

We used an initial learning rate of $1e-5$, batch size of 1 and trained for 200 epochs. We used Adam optimizer with default PyTorch settings. The data samples are generated by sampling a random segment during the evolution of a group for all groups in all the datasets. The data samples are normalized in scale and positions such that the entire group space sequence fits inside the 224×224 image sequences and the geometric center of the group in the last input sequence frame is at the center of the image. After obtaining the predictions from the model, we filter out pixel predictions with confidence level less than 0.5. An example is shown in Fig. 8. For evaluation on a particular dataset, including both evaluation of the encoder-decoder network's performance and the policies in the navigation setting, we use the model that was trained on the other four datasets.

In simulated laser scan settings, we do not retrain the group shape prediction models. Instead, we transfer the learned group shape prediction models on perfect perception settings directly into this new setting. We use a nearest neighbor approach based on geometric centers to identify the history sequence of a group in order to predict the group's future states. If the nearest neighbor of a group in the previous frame is more than $0.25m$ away, then we say no prior history of this group is available and use the technique in section B to linearly back-propagate the group's history.

To integrate this encoder-decoder network into our G-MPC framework, we performed the following additional processing steps (a detailed explanation of footnote 1). Because the encoder-decoder network model takes image-based inputs, we first convert the group space convex hulls into images. We use the homography matrix provided by the dataset to convert the coordinates of the convexhull vertices from meters to pixels. Then we draw these convexhulls on empty canvases to obtain the images. We preprocessed these images to normalize the convexhulls' sizes and positions such that these convexhulls fit inside the images throughout the input sequence and are approximately at the center of the image for the last input frame. Once we obtain the output sequence, we take the coordinates of the vertices at the edge of the output shape blobs and convert them back from pixels to meters using the inverse of the homography matrix.

D Parameter Details

For the parameters of eq. (1), we picked $\epsilon_s, \epsilon_\theta, \epsilon_v$ such that the grouping outcomes match our qualitative inspection of human grouping in the datasets similarly to our prior work [13]. For ETH, HOTEL, ZARA1 and ZARA2 we set $\epsilon_s = 2.0m, \epsilon_\theta = 30^\circ, \epsilon_v = 1.0m/s$. Because UNIV is more crowded than the other four datasets, group formations are tighter and we set $\epsilon_s = 1.5m, \epsilon_\theta = 15^\circ, \epsilon_v = 0.5m/s$.

For the parameter of eq. (17), we selected C under the assumption that closely-interacting pedestrians walk around the boundaries of each other's personal space. For ETH, HOTEL, ZARA1 and ZARA2, we set $C = 0.35$. Again because UNIV has denser crowds, we set $C = 0.25$. If at any given time the robot enters a social group space, we incrementally reduce C by 0.1 with a minimum value of 0.05 until the robot is outside of the group space.

For the time horizon parameter f and the history window parameter h from section 4.2, we set $f = 8$ and $h = 8$ to ensure our MPC formulation's compatibility with the SGAN models.

For the weight parameter λ in the cost function in equation (10), we perform a full parameter sweep to tune λ. We test λ with values from 0.1 to 0.9 with increments of 0.05 on randomly sampled 100 test cases. We then select λ that results in high success rates (at least 90%) for both agent based and group based policies without predictions and that the success rates of the two policies are the closest to each other. For trials with non-reactive agents, we set $\lambda = 0.65$. For trials with reactive agents, we set $\lambda = 0.3$. Note that we want the weight parameter to be the same for both pedestrian-based and group-based policies because the distance from the pedestrians to the boundaries of the social space are the same in both settings. Keeping the same weight allows fair evaluations of these two types of policies.

For the number of control rollouts R in equation (14), we set $R = 12$.

E Numeric Results of Fig. 4 and Fig. 5

Tab. 3 and Tab. 4 are the numerical results of Fig. 4 and Fig. 5. $\mathcal{S}$ is the success rate. $\mathcal{C}$ is percentage of trials in which the robot does not enter any group space (collisions also count as group intrusions). $\mathcal{D}$ is the average minimum distance to pedestrians. $\mathcal{L}$ is the average path length.

Table 3: Performance per scene under the *Offline* condition.

Scene		ETH		HOTEL		ZARA1		ZARA2		UNIV	
Task	Metric	**Flow**	**Cross**	**Flow**	**Cross**	**Flow**	**Cross**	**Flow**	**Cross**	**Flow**	**Cross**
ped-nopred	$\mathcal{S}(\%)$	91.38	75.86	95.35	95.45	96.00	85.71	97.64	96.90	83.96	83.33
	$\mathcal{C}(\%)$	56.9	24.14	81.4	63.64	72.0	46.43	82.68	70.54	68.87	64.91
	$\mathcal{D}(m)$	1.29	0.97	1.35	1.14	1.31	1.03	1.39	1.27	0.95	0.98
	$\mathcal{L}(m)$	27.16	16.93	19.31	10.55	20.75	8.80	20.87	8.92	22.96	17.38
ped-linear	$\mathcal{S}(\%)$	94.83	96.55	97.67	100	96	96.43	94.49	100	86.79	99.12
	$\mathcal{C}(\%)$	75.86	74.14	86.05	88.64	84.0	89.29	88.19	95.35	79.25	90.35
	$\mathcal{D}(m)$	1.42	1.28	1.28	1.25	1.53	1.36	1.53	1.46	1.01	1.09
	$\mathcal{L}(m)$	27.31	18.08	19.28	10.72	20.29	9.48	20.84	9.16	22.93	17.06
ped-sgan	$\mathcal{S}(\%)$	93.1	94.83	95.35	100	96	100	96.06	100	86.79	100
	$\mathcal{C}(\%)$	79.31	70.69	81.4	79.55	80.0	92.86	89.76	95.35	73.58	86.84
	$\mathcal{D}(m)$	1.45	1.27	1.34	1.23	1.46	1.35	1.50	1.47	0.98	1.08
	$\mathcal{L}(m)$	27.05	17.99	19.20	10.10	20.51	9.66	20.83	9.21	23.05	17.22
group-nopred	$\mathcal{S}(\%)$	94.83	87.93	100	88.64	92	75	96.06	92.25	93.4	93.86
	$\mathcal{C}(\%)$	82.76	79.31	97.67	86.36	88.0	71.43	92.91	89.92	90.57	88.6
	$\mathcal{D}(m)$	1.70	1.61	1.8	1.61	1.83	1.59	1.68	1.59	1.18	1.22
	$\mathcal{L}(m)$	29.52	23.32	21.98	14.38	22.86	13.02	23.47	11.17	28.57	20.61
group-pred	$\mathcal{S}(\%)$	96.55	86.21	100	90.91	96	82.14	96.85	94.57	89.62	93.86
	$\mathcal{C}(\%)$	82.76	81.03	100.0	88.64	92.0	82.14	92.91	93.02	86.79	92.98
	$\mathcal{D}(m)$	1.67	1.90	1.83	1.65	1.87	1.75	1.77	1.67	1.19	1.32
	$\mathcal{L}(m)$	29.51	23.17	21.88	13.63	23.01	11.95	23.45	11.13	27.82	20.06
laser-group-pred	$\mathcal{S}(\%)$	94.83	84.48	97.67	95.54	96	75	93.7	86.05	83.02	89.47
	$\mathcal{C}(\%)$	81.03	82.76	97.67	90.91	92.0	75.0	92.13	86.05	82.08	85.09
	$\mathcal{D}(m)$	1.75	2.21	1.86	1.70	1.92	1.81	1.8	1.76	1.25	1.40
	$\mathcal{L}(m)$	29.57	23.99	22.42	15.45	23.50	12.26	23.63	11.58	29.49	22.36

Table 4: Performance per scene under the *Online* condition (simulated pedestrians powered by ORCA[16]).

Scene		ETH		HOTEL		ZARA1		ZARA2		UNIV	
Task	Metric	**Flow**	**Cross**	**Flow**	**Cross**	**Flow**	**Cross**	**Flow**	**Cross**	**Flow**	**Cross**
ped-nopred	$\mathcal{S}(\%)$	96.55	98.28	100	97.73	100	100	97.64	100	88.68	99.12
	$\mathcal{C}(\%)$	50.0	63.79	79.07	77.27	60.0	64.29	65.35	79.84	60.38	78.95
	$\mathcal{D}(m)$	0.93	1.05	0.94	1.04	0.92	0.97	0.98	0.99	0.89	0.98
	$\mathcal{L}(m)$	24.02	13.73	15.75	8.50	16.71	6.13	17.16	7.32	17.20	13.82
ped-linear	$\mathcal{S}(\%)$	94.83	98.28	100	97.73	100	100	98.43	100	89.62	99.12
	$\mathcal{C}(\%)$	43.1	70.69	83.72	90.91	64.0	85.71	74.8	80.62	61.32	88.6
	$\mathcal{D}(m)$	0.94	1.04	0.97	1.03	0.94	0.99	0.98	0.99	0.90	0.98
	$\mathcal{L}(m)$	23.83	13.25	15.43	8.32	16.54	6.22	17.03	6.74	16.87	13.53
ped-sgan	$\mathcal{S}(\%)$	96.55	98.28	100	97.73	100	100	98.43	100	89.62	99.12
	$\mathcal{C}(\%)$	39.66	63.79	88.37	88.64	64.0	82.14	77.95	82.95	62.26	86.84
	$\mathcal{D}(m)$	0.93	1.04	0.95	1.04	0.94	0.99	0.98	0.99	0.90	0.98
	$\mathcal{L}(m)$	23.85	13.20	15.63	8.14	16.54	6.18	17.06	6.72	16.90	13.53
group-nopred	$\mathcal{S}(\%)$	93.1	98.28	95.35	100	96	100	97.64	98.45	89.62	100
	$\mathcal{C}(\%)$	75.86	79.31	81.4	88.64	88.0	71.43	90.55	85.27	68.87	88.6
	$\mathcal{D}(m)$	1.15	1.15	1.17	1.16	1.11	1.03	1.21	1.11	0.94	1.04
	$\mathcal{L}(m)$	25.36	15.37	16.62	9.72	18.16	7.50	18.36	8.56	18.98	15.15
group-pred	$\mathcal{S}(\%)$	94.83	98.28	97.67	97.73	92	96.43	99.21	99.22	93.4	98.25
	$\mathcal{C}(\%)$	74.14	87.93	81.4	95.45	88.0	89.29	91.34	93.8	80.19	91.23
	$\mathcal{D}(m)$	1.13	1.24	1.16	1.21	1.17	1.13	1.18	1.11	0.93	1.07
	$\mathcal{L}(m)$	25.19	14.99	16.45	9.14	17.93	7.62	18.22	7.88	18.71	14.74
laser-group-pred	$\mathcal{S}(\%)$	91.38	98.28	95.35	97.73	96	89.29	97.64	97.67	84.91	92.11
	$\mathcal{C}(\%)$	70.69	84.48	88.37	93.18	88.0	85.71	88.98	94.57	60.38	78.07
	$\mathcal{D}(m)$	1.18	1.33	1.26	1.28	1.18	1.12	1.23	1.14	0.93	1.07
	$\mathcal{L}(m)$	25.40	16.27	16.81	9.82	18.72	8.54	19.07	8.57	21.39	16.46