View Birdification in the Crowd: Ground-Plane Localization from Perceived Movements (Supplementary Material)

Mai Nishimura[12]
mai.nishimura@sinicx.com
Shohei Nobuhara[1]
nob@i.kyoto-u.ac.jp
Ko Nishino[1]
kon@i.kyoto-u.ac.jp

[1] Kyoto University
Kyoto, Japan
[2] OMRON SINIC X
Tokyo, Japan

1 Supplementary Video

Please see the video with audio submitted alongside this pdf. In the video, we show results on the GTAV dataset corresponding to Fig.3 in the paper as sequences. We also compare with static keypoints-based SLAM [5]. These results show that our method successfully birdifies views in very crowded scenes. In contrast, static keypoints-based SLAM fails due to severe occlusions induced by pedestrians on the background which demonstrates the brittleness of the static background assumption in typical view birdification scenarios.

2 Energy Function

2.1 Pedestrian Interaction Models

In Section 4, we formulated view birdification as an iterative energy minimization problem that consists of a pedestrian interaction model $p(\boldsymbol{x}_k^t|\mathcal{X}_k^{t-\tau:t-1})$ and a likelihood $p(\boldsymbol{z}_k^t|\boldsymbol{x}_k^t,\Delta\boldsymbol{x}_0^t)$ defined by the geometric observation model with ambiguities arising from human height estimates (Eq. (5)). Our framework is not limited to a specific pedestrian interaction model, and any type of model that explains pedestrian interactions in a crowd can be incorporated. In the following, we consider two example models with a temporal window of $\tau = 2$.

Constant Velocity. ConstVel [6] is a simple yet effective model of pedestrian interactions in a crowd which simply linearly extrapolates future trajectories from the last two frames

$$p(\boldsymbol{x}_k^t|\mathcal{X}_k^{t-2:t-1}) \sim \exp\left[-\|\boldsymbol{x}_k^t - 2\boldsymbol{x}_k^{t-1} + \boldsymbol{x}_k^{t-2}\|^2\right]. \tag{1}$$

The model is independent of other pedestrians and the overall pedestrian interaction model can be factorized as $p(\mathcal{X}_{1:K}^t|\mathcal{X}_{1:K}^{t-2:t-1}) = \prod_{k=1}^K p(\boldsymbol{x}_k^t|\mathcal{X}_k^{t-2:t-1})$. The energy model $\mathcal{E}_p$ is

rewritten as

$$\mathcal{E}_p = \sum_{k=1}^{K} -\ln p(\boldsymbol{x}_k^t | \mathcal{X}_k^{t-2:t-1}) + \sum_{k=1}^{K} -\ln p(\boldsymbol{z}_k^t | \boldsymbol{x}_k^t, \Delta\boldsymbol{x}_0^t)\,. \tag{2}$$

Social Force. The Social Force Model [7] is a well-known physics-based model that simulates multi-agent interactions with reciprocal forces, which is widely used in crowd analysis and prediction studies [4, 8]. Each pedestrian k with a mass m_k follows the velocity $d\boldsymbol{x}/dt^2$

$$m_k \frac{d^2\boldsymbol{x}_k}{dt^2} = \boldsymbol{F}_k = \boldsymbol{F}_p(\boldsymbol{x}_k) + \boldsymbol{F}_r(\mathcal{X}_\mathcal{C})\,, \tag{3}$$

where $\boldsymbol{F}_k$ is the force on $\boldsymbol{x}_k$ consisting of the personal desired force $\boldsymbol{F}_p$ and the reciprocal force $\boldsymbol{F}_r$. The personal desired force is proportional to the discrepancy between the current velocity and that desired

$$\boldsymbol{F}_p(\boldsymbol{x}_k) = \frac{1}{\eta}\left(\boldsymbol{w}_k - \frac{d\boldsymbol{x}_k}{dt}\right), \tag{4}$$

where $\boldsymbol{w}_k$ denotes the desired velocity which can be empirically approximated as the average velocity of neighboring pedestrians $i \in \mathcal{N}(\boldsymbol{x}_k)$ [4].

The form of reciprocal force $\boldsymbol{F}_r$ can be determined by the set of interactions between pedestrian nodes $\boldsymbol{x}_i \in \mathcal{X}_\mathcal{C}$. To reduce the complexity of optimization, we approximate multi-human interaction $\boldsymbol{F}_r(\mathcal{X}_\mathcal{C})$ with a collection of pairwise interactions $\boldsymbol{F}_r(\boldsymbol{x}_i, \boldsymbol{x}_k)$. We assume a standard Gaussian potential to simulate the reciprocal force between two pedestrians

$$\boldsymbol{F}_r(\boldsymbol{x}_i, \boldsymbol{x}_k) = -\nabla\left(\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{\|\boldsymbol{x}_i - \boldsymbol{x}_k\|^2}{2\sigma^2}\right]\right). \tag{5}$$

Without loss of generality, we omit m_k as $m_k = 1$, assuming that the mass of pedestrians in a crowd is almost consistent. Taking the last two frames as inputs, the complete pedestrian interaction model becomes

$$\begin{aligned} &p(\mathcal{X}_{1:K}^t | \mathcal{X}_{1:K}^{t-2:t-1}) \\ &\sim \prod_k \exp\left[-\left\|\boldsymbol{F}_p(\boldsymbol{x}_k^t) - \frac{d^2\boldsymbol{x}_k^t}{dt^2}\right\|\right] \prod_{(i,k)\in\mathcal{X}_\mathcal{C}} \exp\left[-\left\|\boldsymbol{F}_r(\boldsymbol{x}_i^t, \boldsymbol{x}_k^t)\right\|\right]. \end{aligned} \tag{6}$$

Taking negative log probabilities, the overall energy model in Eq. (6) becomes

$$\mathcal{E}_p = \sum_k D_k(\boldsymbol{x}_k^t; \mathcal{X}_k^{t-2:t-1}) + \sum_{(i,k)\in\mathcal{X}_\mathcal{C}} V_{ik}(\boldsymbol{x}_i^t, \boldsymbol{x}_k^t)\,, \tag{7}$$

where the unary term and pairwise terms are

$$D_k(\boldsymbol{x}_k^t) = \left\|\boldsymbol{F}_p(\boldsymbol{x}_k^t) - \frac{d^2\boldsymbol{x}_k^t}{dt}\right\| - \ln p(\boldsymbol{z}_k^t | \boldsymbol{x}_k^t, \Delta\boldsymbol{x}_0^t)\,, \tag{8}$$

$$V_{ik}(\boldsymbol{x}_i^t, \boldsymbol{x}_k^t) = \boldsymbol{F}_r(\boldsymbol{x}_i^t, \boldsymbol{x}_k^t)\,, \tag{9}$$

respectively.

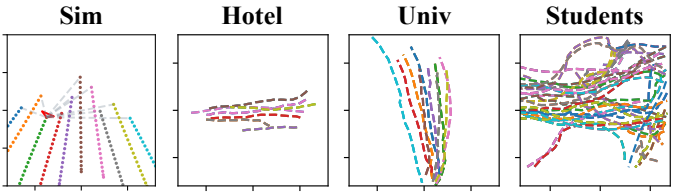

Figure 1: **Typical example trajectories.** Typical example trajectories from the datasets Sim, Hotel, Univ, and Students. In the Sim Example, the red triangle is the virtual camera that observes projected pedestrians on the image plane, where dashed gray lines denote the projection.

2.2 Implementation Details

We use the validation split of each crowd dataset [3] to find the optimal hyperparameters of the pedestrian interaction models. We set the weight parameter of the desired force $\boldsymbol{F}_p$ to $\eta = 0.5$, and the variance of the Gaussian potential to $\sigma^2 = 1.0$ for the social force model. For each dataset of simulated and real trajectories, the size of the ground field, where pedestrians are walking from starting points to their destinations, are scaled to $[-8.0, 8.0]$ m. We also assume that the initial positions of pedestrians $\boldsymbol{x}_k^{\tau_1}$ and $\boldsymbol{x}_k^{\tau_1+1}$ for time $t = \tau_1, \tau_1+1$ are given a priori, and the positions at the next timesteps $\mathcal{X}_k^{\tau_1+2:\tau_2} = \{\boldsymbol{x}_k^{\tau_1+2}, \ldots, \boldsymbol{x}_k^{\tau_2}\}$ are sequentially estimated based on our approach.

3 Dataset Details

3.1 Example Trajectories

Fig. 1 visualizes typical example sequences from the synthetic dataset referred to as **Sim** and from the real trajectory dataset referred to as **Hotel**, **Univ**, and **Students**. In all of these datasets, a virtual observation camera is assigned to one of the trajectories and the observer captures the rest of the pedestrians in the sequence. Fig. 2 shows example trajectories of the GTAV dataset. The size of the ground field, where pedestrians are walking from starting points to their destinations, is configured to be 20m $\times$ 40m. We spawned 50 pedestrians starting from one of the four corners of the field, $[-10,-10], [10,10], [10,-20], [10,20]$, and set the opposite side of the field as their destinations. Both the starting points and destinations were randomized with a uniform distribution. In the **GTAV** dataset, an observation camera is mounted on one of the pedestrians walking in the crowd flow and we can obtain pairs of ground-truth trajectories and ego-centric videos with 90° field -of-view via Script Hook V APIs [7].

3.2 Statistics of the Dataset

In this paper, we constructed several datasets consisting of synthetic pedestrian trajectories (**Sim**), real pedestrian trajectories (**Hotel**, **Univ**, **Students**), and photorealistic crowd simulation (**GTAV**). These datasets are designed differently in several aspects (*i.e.*, densities of a crowd, synthetic view or not, synthetic or real interaction models) for evaluation studies of

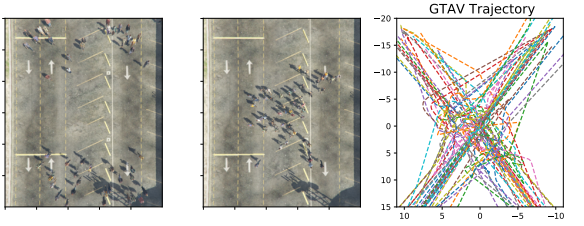

Figure 2: **Example Trajectories from the GTAV dataset.** (Left) Pedestrians are spawned at one of the four corners of the field. (Center) Pedestrians walking towards their destinations while avoiding collisions. (Right) Trajectories of each pedestrian in one sequence.

Table 1: **Overview of birdification dataset.** For real trajectories, we selected scenes of Hotel, Univ, and Students by taking into account the number of people in the crowd. "Seq." corresponds to all the frames captured by a moving observer. "Len." denotes the number of frames included in one sequence.

Dataset	Seq.	Len.	People in Crowd			Int.	observer	input	height	occluded
	Total	Avg	Min	Avg	Max	model	view	bboxes	variances	pedestrians
Sim	500	20.0	10	—	50	synthetic	synthetic	given	✔	
Hotel	340	15.0	3	6.31	15	real	synthetic	given	✔	
Univ	346	14.4	3	9.29	26	real	synthetic	given	✔	
Students	849	45.8	13	44.2	75	real	synthetic	given	✔	
GTAV	—	400	3	6	12	synthetic	photorealistic	MOT [▯]		✔

our proposed view birdification method. Table 1 summarizes the statistics and taxonomy of these datasets.

3.3 Quantitative Results on GTAV dataset

In Section 5.2 of the main text, we omitted quantitative results on the GTAV dataset due to space limitations. Table 2 shows quantitative results on the GTAV dataset with metrics introduced in Sec.5.2 in the manuscript. As introduced in the paper, we prepared two versions of inputs, one manually annotated with centerlines of the people and their heights and the other with those automatically extracted from a multi-object tracker (MOT). We compared view birdification results using these two different inputs, which are referred to as *Birdify-CLine* and *Birdify-MOT*. The results show *Birdify-CLine* and *Birdify-MOT* achieve comparative performance in terms of rotation and translation errors, $\Delta\boldsymbol{r}, \Delta\boldsymbol{t}$ since the localization of the observer is insensitive to pedestrian detection errors. On the other hand, in terms of pedestrian localization errors, $\Delta\tilde{\boldsymbol{x}}$ and $\Delta\boldsymbol{x}$, *Birdify-MOT* results show inferior performance to manually annotated inputs. This is mainly due to the fact that we currently estimate the initial position of a pedestrian $\boldsymbol{x}_k^0$ relative to the observer position $\boldsymbol{x}_0^t$ by Eq. (1) in the main text, whenever a new pedestrian appears in a frame. The accuracy of this initial estimate can be improved by fine-tuning the multi-object tracker or by using the pose of the person [▯, ▯▯]. We will explore these in future work.

Table 2: **Birdification results of real trajectories.** The relative and absolute localization errors of pedestrians, $\Delta\tilde{\boldsymbol{x}}$ and $\Delta\boldsymbol{x}$, respectively, and the errors of camera ego-motion estimation, Δr , and $\Delta\boldsymbol{t}$, computed for each frame whose mean values are shown.

Input	$\Delta\boldsymbol{r}$ [rad]	$\Delta\boldsymbol{t}$ [m]	$\Delta\tilde{\boldsymbol{x}}$ [m]	$\Delta\boldsymbol{x}$ [m]
cline(manual)	0.015	0.097	0.441	0.491
MOT [9]	0.016	0.101	0.491	0.530

4 Failure Cases

We also analyze failure cases of our view birdification to understand the limitations of the method. For this, we picked sequences from **Univ** data that showed a high error rate in terms of camera localization.

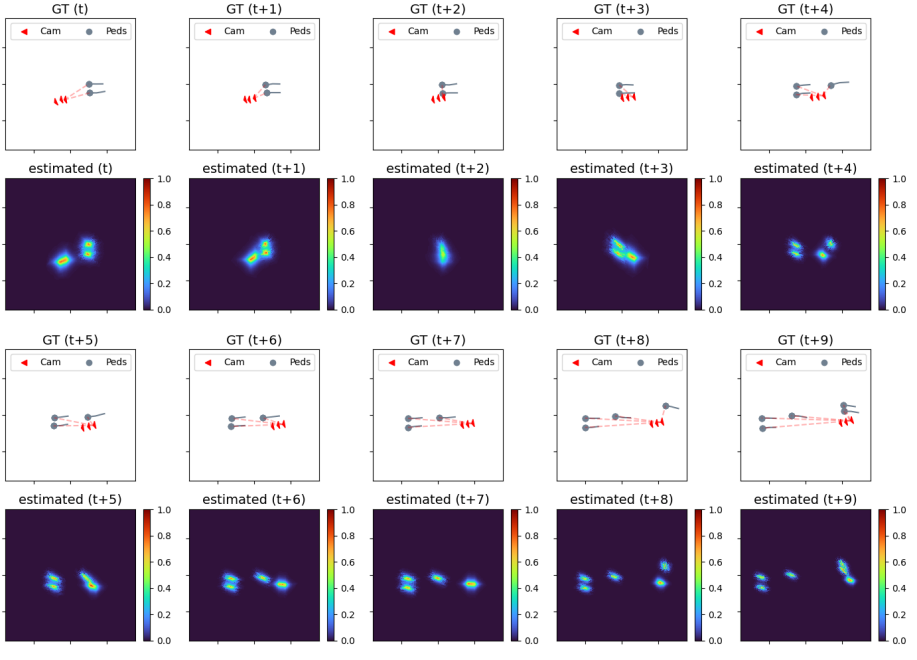

Figure 3: **Visualization of posterior distributions of the Univ dataset.** (First and third rows) Ground truth trajectories of the camera and its surrounding pedestrians. (Second and fourth rows) Visualization of posterior distributions of the location of the observer $\boldsymbol{x}_0^t$ and surrounding pedestrians $\boldsymbol{x}_k^t$. The heatmaps correspond to low (blue) to high (red) probabilities.

4.1 Visualization of Posterior Distributions

Fig. 3 visualizes posterior distributions of the observer location $p(\boldsymbol{x}_0^t|\mathcal{Z}_{1:K}^t, \mathcal{X}_{0:K}^{t-1})$ and surrounding pedestrians $\int_{\boldsymbol{x}_0^t\in\mathcal{X}_s} p(\mathcal{X}_{1:K}^t|\mathcal{Z}_{1:K}^t, \boldsymbol{x}_0^t)p(\boldsymbol{x}_0^t)d\boldsymbol{x}_0^t$ by sampling $\boldsymbol{x}_0^t \in \mathcal{X}_s$ in Eq. (4) and

Eq. (5) in the manuscript, respectively. The first and third rows depict the ground truth trajectories of the camera and pedestrians from t to $t+9$. The number of pedestrians changes from $K=3$ to $K=5$. The second and fourth rows visualize the posterior distributions for each of those two rows. As can be observed in the posteriors shown in the second row, the estimated observer location becomes a heavy-tailed distribution when the number of pedestrians in the crowd is small ($K=3$). In contrast, as shown in the fourth row, the posterior distribution becomes sharper when the crowd is denser ($K=5$). The ambiguity of localization increases when pedestrians walk almost parallel to the observer (*e.g.*, timesteps $t=t+2$ and $t+3$). In contrast, the posterior distribution becomes sharp again when the camera observes more pedestrians walking in diverse directions. Moreover, when the camera observes a large number of pedestrians that conforms to a known crowd motion model, whether or not the camera motion is consistent with dominant crowd flow, the camera ego-motion estimates highly depend on the observed crowd movements and are less sensitive to assumed ego-motion model. That is, as long as the camera observes a sufficient number of pedestrians walking in diverse directions, our method can successfully birdify its views.

References

[1] Lorenzo Bertoni, Sven Kreiss, and Alexandre Alahi. Monoloco: Monocular 3d pedestrian localization and uncertainty estimation. In *Proc. ICCV*, October 2019.

[2] Dirk Helbing and Peter Molnar. Social force model for pedestrian dynamics. *Physical review E*, 51(5):4282, 1995.

[3] Boris Ivanovic and Marco Pavone. The trajectron: Probabilistic multi-agent trajectory modeling with dynamic spatiotemporal graphs. In *Proc. ICCV*, pages 2375–2384, 2019.

[4] Ramin Mehran, Alexis Oyama, and Mubarak Shah. Abnormal crowd behavior detection using social force model. In *Proc. CVPR*, pages 935–942. IEEE, 2009.

[5] Raúl Mur-Artal and Juan D. Tardós. ORB-SLAM2: an open-source SLAM system for monocular, stereo and RGB-D cameras. *IEEE Transactions on Robotics*, 33(5): 1255–1262, 2017. doi: 10.1109/TRO.2017.2705103.

[6] Christoph Schöller, Vincent Aravantinos, Florian Lay, and Alois Knoll. What the constant velocity model can teach us about pedestrian motion prediction. *IEEE Robotics and Automation Letters*, 5(2):1696–1703, 2020.

[7] Script Hook V. Script Hook V. http://www.dev-c.com/gtav/.

[8] Jur Van Den Berg, Stephen J Guy, Ming Lin, and Dinesh Manocha. Reciprocal n-body collision avoidance. In *Robotics research*, pages 3–19. Springer, 2011.

[9] Zhongdao Wang, Liang Zheng, Yixuan Liu, and Shengjin Wang. Towards real-time multi-object tracking. In *Proc. ECCV*, 2020.

[10] Yuliang Xiu, Jiefeng Li, Haoyu Wang, Yinghong Fang, and Cewu Lu. Pose Flow: Efficient online pose tracking. In *Proc. BMVC*, 2018.