CrowdMAC: Masked Crowd Density Completion for Robust Crowd Density Forecasting Supplementary Material

Ryo Fujii[1] Ryo Hachiuma[2] Hideo Saito[1]
[1]Keio University [2]NVIDIA
{ryo.fujii0112, hs}@keio.jp, rhachiuma@nvidia.com

A. Datasets

In this section, we delve into the datasets used in our study. We validate our method using trajectory prediction datasets: SDD [11], ETH-UCY [5, 10], inD [2], and JRDB [8], as well as crowd density analysis datasets: FDST [3], croHD [13], and VSCrowd [6]. Tab. 1 presents the average count of people appearing per frame, highlighting the diversity in density across the datasets.

Stanford Drone Dataset (SDD). SDD [11] consists of 20 scenes captured on the Stanford University campus in a bird's eye view using a flying drone. Following the previous trajectory prediction methods [7], we use the standard setup and train-test split.

ETH-UCY. ETH [10] and UCY [5] are widely used for human trajectory forecasting benchmarks. They consist of five different scenes ETH & HOTEL (from ETH), UNIV, ZARA1, and ZARA2 (from UCY). The leave-one-out validation strategy is employed, followed by prior work [4].

Intersection Drone Dataset (inD). inD [2] acquired with a static drone, comprises 32 recordings collected at 4 distinct intersections. We focus only on pedestrian trajectories and consider the evaluation protocol proposed in [1], where all scenes are split into the train, validation, and test sets according to a 70-10-20 rule.

JackRabbot Dataset (JRDB). JRDB [8] is a real-world dataset that provides a diverse set of pedestrian trajectories and 2D bounding boxes, allowing for a comprehensive evaluation of our models in both indoor and outdoor scenarios. We use the stationary scenes for training and testing. Specifically, we use 'gates-ailab,' 'packard-poster-session,' and 'tressider' for testing and the other scenarios for training.

Fudan-ShanghaiTech dataset (FDST). FDST [3] is curated for video crowd counting tasks, comprising 100 videos capturing crowds in 15 distinct locations, each with unique camera poses and positions, along with annotations for individual heads. We follow the official train-test split.

Crowd of Heads Dataset (croHD). The croHD [13] provides tracking annotation of pedestrian heads in densely populated video sequences. It consists of 9 sequences of 11,463 frames with over $2,276,838$ heads and $5,230$ tracks annotated in diverse scenes. We follow the official train-test split.

Table 1. Comparison of datasets with respect to the average count of people appearing per frame.

Datasets	SDD	ETH-UCY	inD	JRDB	VSCrowd	FDST	croHD
AVG Count	11	12	3	8	30	26	110

Video Crowd dataset (VSCrowd). VSCrowd [6] is a dataset developed for crowd localization. It consists of 634 videos captured in various scenes (e.g., malls, streets, scenic spots) and head annotations. We follow the official train-test split.

B. Evaluation Metrics

Followed by prior work [9], we use Jensen-Shannon (JS) divergence to measure the performance of the forecasting:

$$\mathcal{D}_{JS}(g_t||c_t) = \frac{1}{2}(\mathcal{D}_{KL}(\bar{g}_t||\bar{c}_t) + \mathcal{D}_{KL}(\bar{c}_t||\bar{g}_t)), \quad (1)$$

where $\bar{g}_t = g_t / \sum_{i,j} g_t(i,j)$, $\bar{c}_t = c_t / \sum_{i,j} c_t(i,j)$ are the predicted and ground truth normalized density maps, i, j are the indices of pixel position, and $\mathcal{D}_{KL}$ is Kullback-Leibler (KL) divergence:

$$\mathcal{D}_{KL}(g_t||c_t) = \frac{1}{WH} \sum_{i,j} \bar{g}_t(i,j) \log(\frac{\bar{g}_t(i,j)}{\bar{c}_t(i,j)}). \quad (2)$$

We report the Average JS divergence (AD_{JS}) and the Final JS divergence (FD_{JS}). AD_{JS} is the divergence between the predicted and the ground truth map averaged over all the future time steps, while FD_{JS} is the divergence between the predicted and ground truth map at the final time step.

Table 2. Comparison of the crowd density forecasting and trajectory prediction approaches using ground truth pedestrian positions (see 4.5) on ETH-UCY. The lower metrics (AD_{JS}, FD_{JS}) are better.

Dataset	Trajectory Prediction				Crowd Density Forecasting			
	Y-Net [7]		Social-Trans. [12]		PDFN-ST [9]		Ours	
	AD_{JS}	FD_{JS}	AD_{JS}	FD_{JS}	AD_{JS}	FD_{JS}	AD_{JS}	FD_{JS}
ETH	0.565	0.682	0.587	0.782	0.512	0.702	**0.258**	**0.377**
HOTEL	0.413	0.528	0.383	0.459	0.542	0.764	**0.249**	**0.375**
UNIV	0.178	0.253	0.155	0.217	0.199	0.369	**0.108**	**0.163**
ZARA1	0.345	0.506	0.270	0.396	0.623	0.987	**0.181**	**0.310**
ZARA2	0.242	0.352	0.197	0.275	0.344	0.529	**0.144**	**0.241**
AVG	0.346	0.464	0.318	0.426	0.444	0.670	**0.188**	**0.250**

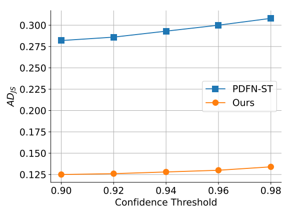

Figure 1. We compare the robustness of the models with realistic miss-detections on the VSCrowd.

C. Comparison Models

We employ the following methods for comparison:

PDFN-ST (RA-L'21) [9]: PDFN-ST is a pioneering work that tackles the crowd density forecasting task by using 3D CNNs to learn local crowd density dynamics in 3D receptive fields, regarded as spatiotemporal patches.

Y-Net (ICCV'21) [7]: Y-Net is a heatmap-based model that predicts future human trajectories by estimating distributions over long-term goals and intermediate waypoints.

Social-Transmotion (ICLR'24) [12]: Social-Transmotion is a Transformer-based model for human trajectory prediction, leveraging diverse visual cues. The model is designed to predict human behavior by capturing spatiotemporal interactions between agents.

D. Additional Results

Forecasting Accuracy Comparison on each ETH-UCY subset with Ground Truth Input Protocol. We compare our model with crowd density forecasting and trajectory prediction models using the ground truth input evaluation protocol on ETH-UCY. As shown in Tab. 2, Our CrowdMAC consistently outperforms both trajectory prediction methods and crowd density forecasting methods across all subsets.

Robustness against Realistic Miss-Detection on VSCrowd. In Fig. 1, we examine the robustness to miss-detections using data preprocessed by the pedestrian detection module on the VSCrowd (as described in Sec. 4.5) . Our proposed method shows a smaller performance drop compared to PDFN-ST, demonstrating greater robustness to realistic miss-detections.

Long Term Forecasting Results. Tab. 3 presents the comparison in a long-term setting, observing 2 seconds in the past and predicting 6 seconds into the future. We observe that our proposed model outperforms both the crowd density forecasting and trajectory prediction methods across multiple datasets.

Table 3. Comparison of crowd density forecasting and trajectory prediction approaches in a long-term setting using ground truth pedestrian positions.

Dataset	Trajectory Prediction		Crowd Density Forecasting			
	Social-Trans. [12]		PDFN-ST [9]		Ours	
	AD_{JS}	FD_{JS}	AD_{JS}	FD_{JS}	AD_{JS}	FD_{JS}
SDD [11]	0.103	**0.139**	0.102	0.197	**0.089**	0.189
JRDB [8]	0.124	0.148	0.091	0.132	**0.090**	**0.125**
VSCrowd [6]	0.372	0.398	0.138	0.153	**0.101**	**0.117**
FDST [3]	-	-	0.073	0.122	**0.060**	**0.104**
croHD [13]	-	-	0.045	**0.052**	**0.042**	**0.052**

Qualitative Comparison. We show some qualitative results on SDD in Fig. 2 and on the FDST in Fig. 3. Our method produces more precise predictions than the state-of-the-art method PDFN-ST at every time step. The performance gap between our method and PDFN-ST becomes more evident as the time step advances.

References

[1] Alessia Bertugli, Simone Calderara, Pasquale Coscia, Lamberto Ballan, and Rita Cucchiara. AC-VRNN: Attentive Conditional-VRNN for multi-future trajectory prediction. CVIU, 2021. 1

[2] Julian Bock, Robert Krajewski, Tobias Moers, Steffen Runde, Lennart Vater, and Lutz Eckstein. The inD Dataset: A Drone Dataset of Naturalistic Road User Trajectories at German Intersections. In IV, 2020. 1

[3] Yanyan Fang, Bi-Sheng Zhan, Wandi Cai, Shenghua Gao, and Bo Hu. Locality-Constrained Spatial Transformer Network for Video Crowd Counting. ICME, 2019. 1, 2

[4] Agrim Gupta, Justin Johnson, Li Fei-Fei, Silvio Savarese, and Alexandre Alahi. Social GAN: Socially Acceptable Trajectories With Generative Adversarial Networks. In CVPR, 2018. 1

[5] Laura Leal-Taixé, Michele Fenzi, Alina Kuznetsova, Bodo Rosenhahn, and Silvio Savarese. Learning an Image-Based Motion Context for Multiple People Tracking. In CVPR, 2014. 1

[6] Haopeng Li, Lingbo Liu, Kunlin Yang, Shinan Liu, Junyu Gao, Bin Zhao, Rui Zhang, and Jun Hou. Video crowd localization with multifocus gaussian neighborhood attention and a large-scale benchmark. TIP, 2022. 1, 2

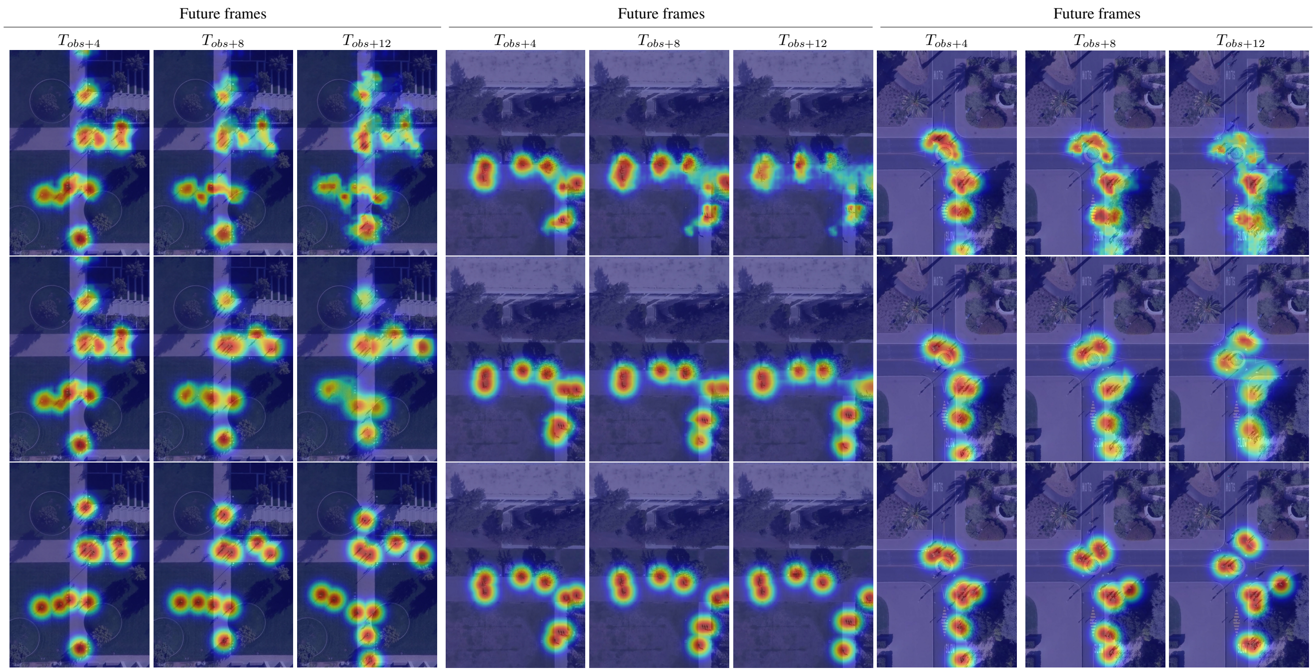

Figure 2. Qualitative results on SDD. The state-of-the-art method (PDFN-ST) prediction (first row), CrowdMAE (second row), and ground truth (third row) are shown. We overlay the crowd density map onto the original image.

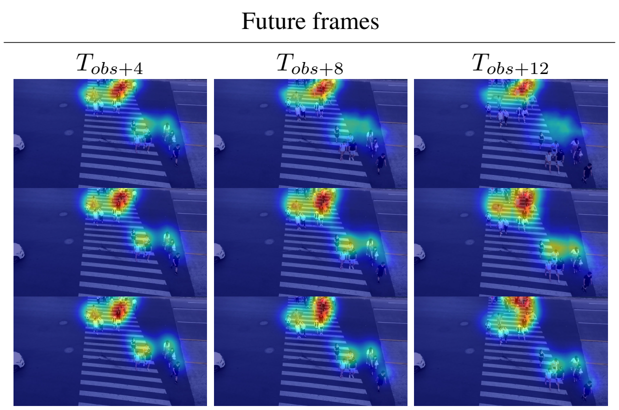

Figure 3. Qualitative results on FDST. The state-of-the-art method (PDFN-ST) prediction (first row), CrowdMAE (second row), and ground truth (third row) are shown. We overlay the crowd density map onto the original image.

[7] Karttikeya Mangalam, Yang An, Harshayu Girase, and Jitendra Malik. From Goals, Waypoints & Paths to Long Term Human Trajectory Forecasting. In ICCV, 2021. 1, 2

[8] Roberto Martin-Martin, Mihir Patel, Hamid Rezatofighi, Abhijeet Shenoi, JunYoung Gwak, Eric Frankel, Amir Sadeghian, and Silvio Savarese. Jrdb: A dataset and benchmark of egocentric robot visual perception of humans in built environments. IEEE TPAMI, 2021. 1, 2

[9] Hiroaki Minoura, Ryo Yonetani, Mai Nishimura, and Yoshitaka Ushiku. Crowd Density Forecasting by Modeling Patch-Based Dynamics. RA-L, 2021. 1, 2

[10] Stefano Pellegrini, Andreas Ess, and Luc Van Gool. Improving Data Association by Joint Modeling of Pedestrian Trajectories and Groupings. In ECCV, 2010. 1

[11] Alexandre Robicquet, Amir Sadeghian, Alexandre Alahi, and Silvio Savarese. Learning Social Etiquette: Human Trajectory Understanding In Crowded Scenes. In ECCV, 2016. 1, 2

[12] Saeed Saadatnejad, Yang Gao, Kaouther Messaoud, and Alexandre Alahi. Social-transmotion: Promptable human trajectory prediction. In ICLR, 2024. 2

[13] Ramana Sundararaman, Cedric De Almeida Braga, Eric Marchand, and Julien Pettre. Tracking Pedestrian Heads in Dense Crowd. In CVPR, 2021. 1, 2