OinkTrack: An Ultra-Long-Term Dataset for Multi-Object Tracking and Re-Identification of Group-Housed Pigs

Feng-Kai Huang
National Taiwan University
Taipei, Taiwan
leonelhuang@cmlab.csie.ntu.edu.tw

Hong-Wei Xu
National Yang Ming Chiao Tung University
Hsinchu, Taiwan
hw413504009.ee13@nycu.edu.tw

Chu-Chuan Lee
Chung Yuan Christian University
Taoyuan, Taiwan
g11378032@cycu.edu.tw

Hong-Yi Tu
National Yang Ming Chiao Tung University
Hsinchu, Taiwan
hongyi.ee11@nycu.edu.tw

Hong-Han Shuai*
National Yang Ming Chiao Tung University
Hsinchu, Taiwan
hhshuai@nycu.edu.tw

Wen-Huang Cheng
National Taiwan University
Taipei, Taiwan
wenhuang@csie.ntu.edu.tw

A Appendix

A.1 Details of Each Scene

Statistics. Detailed statistics for each of the 16 distinct sequences in *OinkTrack* are presented in Tab. A1. This table underscores the dataset's rich diversity, particularly with regard to illumination conditions, as it includes sequences that deliberately capture daytime, nighttime, and critical day-to-night as well as night-to-day transitions. Crucially, the video segments cover a comprehensive 24-hour cycle, spanning early morning (*e.g.*, 05:00-08:00), midday and afternoon (*e.g.*, 11:00-17:00), through to the evening and deep night (*e.g.*, 17:00-00:00). This extensive temporal coverage provides an invaluable resource for analyzing the diurnal behavioral patterns of group-housed pigs and for rigorously evaluating the robustness of MOT algorithms under evolving natural and artificial lighting. The sequences themselves range in duration from approximately one minute to an entire hour, and each features a consistently high number of tracked individuals (averaging 35.88 pigs) and a substantial volume of bounding box annotations. This detailed breakdown reaffirms *OinkTrack*'s unique position as a benchmark for studying challenging, long-term tracking in realistic and densely populated livestock environments.

Table A1: Statistics of each scene in *OinkTrack*

Scene	Time	Lighting	Len. (min)	Tracks	Boxes
C1D-1	08:00-08:01	Day	1	34	2000
C1D-2	11:00-11:10	Day	10	34	18278
C1D-3	07:00-07:30	Day	30	36	60254
C1DN-1	16:00-17:00	Day-night	60	36	116668
C1N-1	18:00-18:01	Night	1	32	1834
C1N-2	17:50-18:00	Night	10	33	18723
C1N-3	23:30-00:00	Night	30	36	60180
C1ND-1	05:00-05:30	Night-day	30	36	58502
C2D-1	15:00-15:01	Day	1	34	1666
C2D-2	14:15-14:30	Day	15	39	25097
C2DN-1	16:30-17:00	Day-night	30	40	51893
C2N-1	17:00-17:01	Night	1	34	1746
C2N-2	04:50-05:00	Night	10	36	18985
C2N-3	17:15-17:30	Night	15	36	29090
C2N-4	04:30-05:00	Night	30	38	52750
C2ND-1	05:15-05:45	Night-day	30	40	56130

Visualization. Fig. A1 showcases more representative annotated frames from diverse sequences within *OinkTrack* and illustrates the variety of conditions and challenges present. These visualizations reveal pigs that engage in a wide array of natural behaviors, including resting, sleeping, playing, agonistic interactions (fighting), and feeding. The extended duration of our collected videos ensures that individual pigs often exhibit multiple distinct behaviors throughout a single sequence, which potentially leads to significant variations in their visual appearance and posture over time. This behavioral diversity, coupled with the extreme length of the recordings, presents a substantial challenge for MOT algorithms, as a tracker must maintain identity consistency despite these appearance shifts. Furthermore, the figure highlights the distinct illumination characteristics between daytime and nighttime (IR) scenes, a key factor that contributes to tracking difficulty. Instances where other pigs or pen structures occlude individuals and subsequently reappear are also evident; this is a common occurrence in our long-duration, crowded footage that poses a significant test for a tracker's ability to perform stable, continuous tracking and re-identification.

A.2 Implementation Details.

We benchmark *OinkTrack* using 11 distinct MOT algorithms. All models are trained on a system equipped with four NVIDIA V100 GPUs.

For the tracking-by-detection methods (SORT [1], DeepSORT [8], MOTDT [3], ByteTrack [11], OC-SORT [2], StrongSORT and its enhanced variant StrongSORT++ [4], and Hybrid-SORT [9]), we consistently employ YOLOX-X [7] as the object detector. The input size for YOLOX-X is set to 1280×736. To ensure a fair comparison, we first pre-train a single YOLOX-X model on the *OinkTrack* training set for 80 epochs, following the default configuration specified

*Corresponding author, hhshuai@nycu.edu.tw

MM '25, Dublin, Ireland
2025. ACM ISBN 979-8-4007-2035-2/2025/10
https://doi.org/10.1145/3746027.3758189

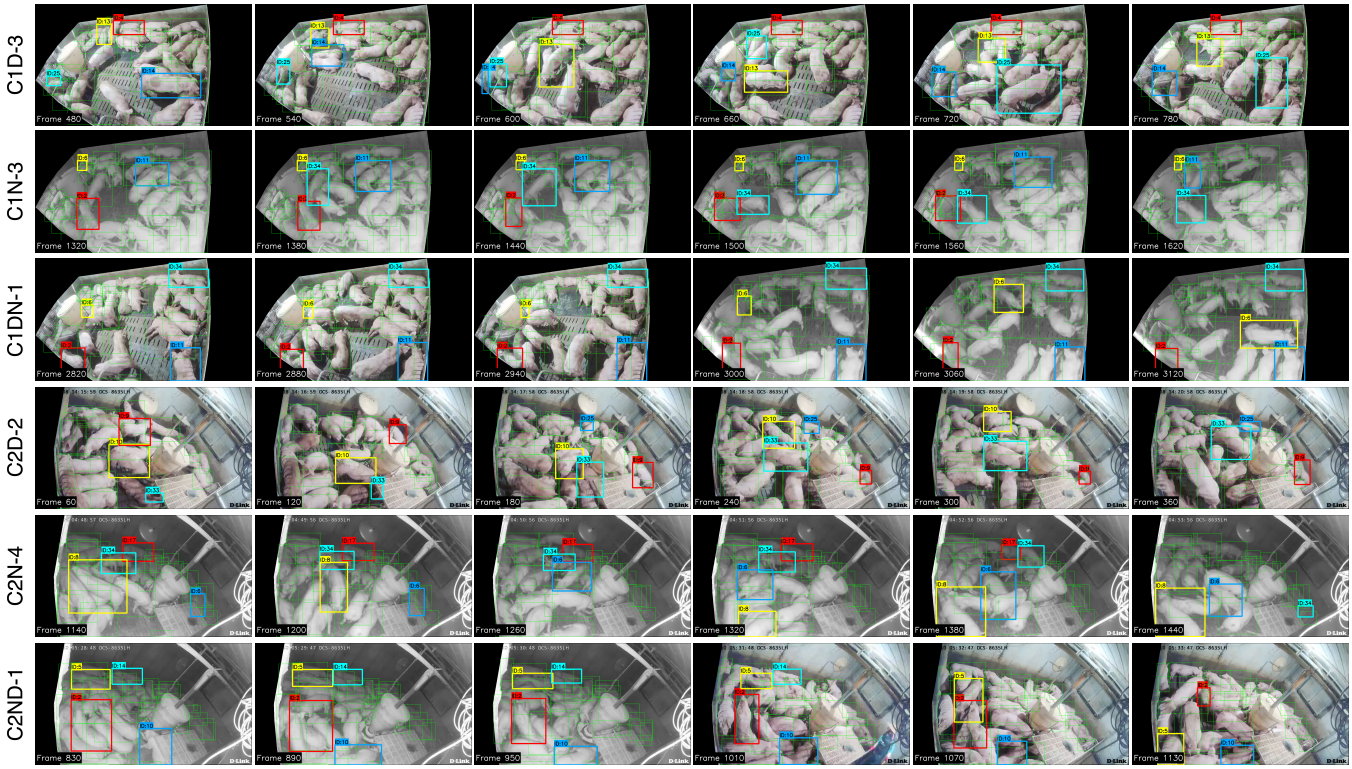

Figure A1: Visualization of annotated frames in different scenes.

in ByteTrack. This single pre-trained detector then provides object detection for all models within this category.

For the transformer-based models (MOTR [10], MeMOTR [6], and MOTIP [5]), we adhere to the training configurations detailed in their respective original publications. Specifically, MOTR and MeMOTR are trained for 20 epochs, while MOTIP is trained for 10 epochs.

A.3 More Qualitative Results.

Fig. A2 presents a qualitative assessment of tracking performance by comparing ground truth (GT) trajectories with predictions from selected state-of-the-art trackers—SORT [1], ByteTrack [11], MOTR [10], MeMOTR [6], Hybrid-SORT [9], and MOTIP [5]. The comparison uses representative *OinkTrack* instances across diverse temporal stages and illumination conditions (*e.g.*, C2N-4, C2N-1, C2D-1, C1DN-1). In shorter temporal segments, visual inspection reveals that transformer-based architectures like MeMOTR and MOTIP generally achieve superior tracking fidelity. They maintain more consistent identities and accurate localization compared to the tracking-by-detection method, ByteTrack, which exhibits a more noticeable proneness to missed detections. This initial robustness of transformer models is attributable to their sophisticated spatio-temporal reasoning capabilities.

However, the critical challenge of extremely long-term tracking, a core feature of *OinkTrack*, becomes evident over extended durations. In these scenarios, even leading approaches, including MeMOTR and MOTIP, begin to falter. We observe instances of identity loss, where individuals are no longer tracked, and significant path deviations, which indicate accumulated localization errors. These qualitative failures become more pronounced in the later stages of long sequences and underscore the severe test that *OinkTrack* poses for sustained identity preservation and continuous localization. The visual evidence in Fig. A2 strongly corroborates our quantitative results and confirms that achieving robust, uninterrupted, and accurate end-to-end tracking in *OinkTrack*'s demanding agricultural environments remains a significant open problem and a fertile ground for future research.

A.4 Full Results of Scene Analysis

Analysis of Illumination Condition. We provide a full breakdown of performance across different illumination conditions in Tab. A2. All evaluated methods achieve their best results in daytime scenes, significantly outperforming their performance under other conditions. An interesting observation arises when comparing day-night transition scenarios against purely nighttime ones: while detection accuracy (DetA) is often higher during transitions than in consistent nighttime, the overall tracking accuracy (HOTA) tends to be worse. A closer examination reveals a dramatic decrease in association accuracy (AssA) during these transitions. This suggests that while objects may be more readily detectable during the shifting light of transitions, the rapid and drastic changes in individual appearance due to varying illumination severely impair the models' ability to maintain consistent associations.

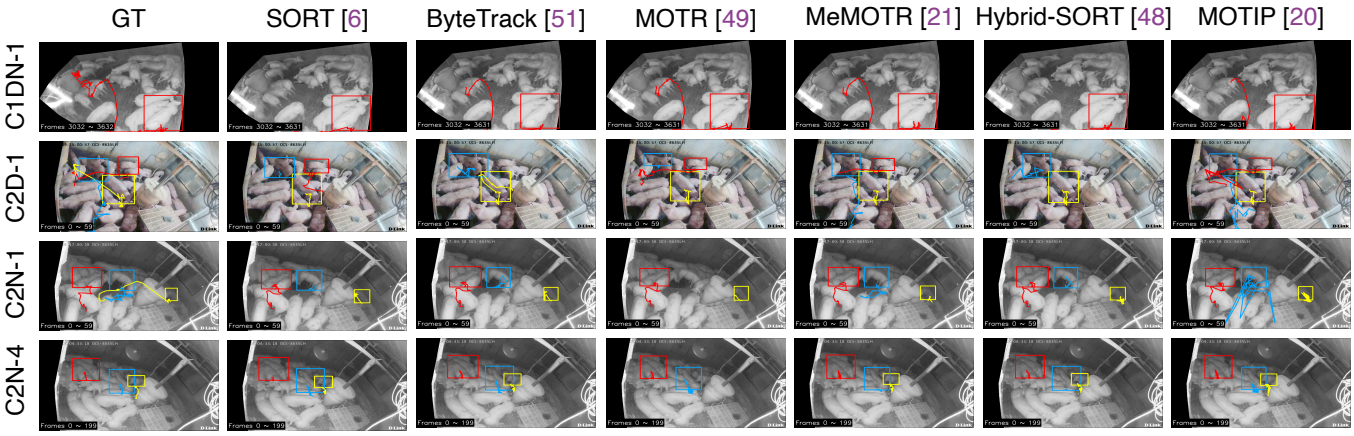

Figure A2: Qualitative results of ground-truth and the predictions of SORT [1], ByteTrack [11], MOTR [10], MeMOTR [6], Hybrid-SORT [9], and MOTIP [5] on different scenes.

Table A2: Overall evaluation results of different algorithms on *OinkTrack* test set.

Method	Day						Night						Cross					
	HOTA↑	MOTA↑	DetA↑	AssA↑	IDF1↑	IDsw↓	HOTA↑	MOTA↑	DetA↑	AssA↑	IDF1↑	IDsw↓	HOTA↑	MOTA↑	DetA↑	AssA↑	IDF1↑	IDsw↓
SORT [1]	50.9	69.0	55.3	48.1	60.8	207	35.5	40.0	40.2	32.3	39.2	512	23.8	58.1	48.2	11.9	24.8	2033
DeepSORT [8]	49.6	65.9	54.4	46.4	58.1	216	34.5	39.1	40.5	30.7	37.4	567	22.8	57.5	48.4	10.9	23.0	2506
MOTDT [3]	42.3	66.1	55.0	34.0	48.1	546	31.5	39.0	39.9	25.9	34.1	1213	19.5	55.2	47.8	8.0	19.1	6622
ByteTrack [11]	48.6	67.4	54.7	44.5	59.2	158	35.0	40.9	41.4	30.8	38.9	397	24.2	59.7	48.4	12.2	26.3	1771
MOTR [10]	55.9	71.6	58.7	54.2	65.9	189	41.9	62.1	55.5	31.9	43.9	532	26.8	64.4	56.0	12.9	25.1	1715
OC-SORT [2]	49.9	67.9	54.5	46.9	59.6	164	35.2	39.6	38.4	32.8	39.0	377	23.3	56.0	46.7	11.6	25.1	1683
StrongSORT [4]	46.8	65.3	53.6	42.5	53.5	215	33.2	39.6	39.8	28.9	37.0	604	22.1	55.8	46.8	10.6	22.6	3127
StrongSORT++ [4]	46.3	61.8	52.3	42.7	53.1	186	33.8	38.0	40.4	29.6	37.1	484	21.9	49.5	45.3	10.8	22.1	2635
MeMOTR [6]	**59.5**	79.9	64.5	**55.7**	**71.7**	**134**	48.5	**68.0**	**59.3**	40.2	53.3	**352**	35.8	75.3	62.4	**20.7**	37.4	**1349**
Hybrid-SORT [9]	50.2	68.7	55.2	46.9	59.9	192	36.2	38.4	41.5	32.6	38.8	619	24.2	57.9	49.3	12.0	24.8	2122
MOTIP [5]	58.4	**82.5**	**66.5**	52.3	70.9	420	**53.6**	64.7	59.2	**49.0**	**61.7**	1387	**35.9**	**78.2**	**64.9**	19.9	**38.8**	6047

Table A3: Evaluation results of different video lengths on *OinkTrack* test set.

Method	≤ 1 min						30 min						60 min					
	HOTA↑	MOTA↑	DetA↑	AssA↑	IDF1↑	IDsw↓	HOTA↑	MOTA↑	DetA↑	AssA↑	IDF1↑	IDsw↓	HOTA↑	MOTA↑	DetA↑	AssA↑	IDF1↑	IDsw↓
SORT [1]	50.1	63.5	51.9	49.8	60.1	327	27.0	39.4	38.6	19.2	30.2	970	23.2	63.5	51.8	10.5	23.0	1455
DeepSORT [8]	48.1	61.5	51.6	46.6	56.0	381	26.9	38.8	39.0	18.9	29.6	1082	21.9	62.8	51.8	9.4	20.9	1826
MOTDT [3]	43.2	61.2	51.7	37.8	49.5	918	23.0	37.6	37.9	14.2	25.1	2452	19.1	60.6	51.6	7.1	17.6	5011
ByteTrack [11]	48.5	63.0	51.9	46.8	59.4	254	27.2	40.9	39.7	19.0	30.7	791	23.4	64.9	51.7	10.7	24.5	1281
MOTR [10]	54.6	70.0	59.0	51.1	62.3	350	33.7	59.9	53.1	21.6	34.4	875	24.9	66.2	57.5	10.8	22.6	1211
OC-SORT [2]	49.2	61.7	50.5	49.0	58.6	255	26.5	38.2	36.3	19.5	30.3	734	22.9	61.6	50.9	10.3	23.5	1235
StrongSORT [4]	45.7	60.8	50.5	42.8	53.3	372	25.3	38.6	38.0	17.1	28.3	1243	22.1	60.9	50.2	9.8	21.2	2331
StrongSORT++ [4]	45.6	58.5	50.0	43.2	53.2	306	25.8	35.9	38.7	17.7	28.1	1082	21.7	53.5	47.8	10.0	20.7	1917
MeMOTR [6]	59.1	78.1	64.3	55.0	70.3	**247**	40.9	**65.5**	56.8	29.7	44.4	**656**	**35.3**	79.3	65.3	**19.1**	**35.7**	**932**
Hybrid-SORT [9]	49.5	63.0	52.1	48.2	58.6	324	28.4	38.4	40.4	20.2	30.5	1089	23.5	63.2	52.5	10.5	23.0	1520
MOTIP [5]	**60.3**	**80.3**	**66.2**	**55.7**	**72.8**	828	**45.8**	64.5	**58.6**	**36.0**	**52.7**	2594	33.5	**82.1**	**66.9**	16.8	34.4	4432

Furthermore, models based on transformer architectures exhibit relatively more stable performance across all lighting conditions, indicating stronger adaptability to illumination changes compared to traditional tracking-by-detection methods. Approaches like SORT and DeepSORT, for instance, particularly struggle when faced with the combined challenges of high-density or occluded environments under nighttime or transitional lighting. The abrupt shifts in illumination during transitions induce substantial changes in target appearance, making it difficult for these models to maintain continuous and accurate tracks. This fine-grained analysis underscores the significant challenge posed by abrupt illumination changes and highlights the critical role of robust appearance modeling and temporal context integration for stable long-term tracking.

Analysis of Video Length. We further analyze model performance across video sequences of varying durations, with detailed results presented in Tab. A3. Our experimental observations indicate that while object detection performance (DetA) remains relatively stable irrespective of sequence length, overall tracking performance (HOTA) and identity preservation (IDF1) deteriorate notably in longer sequences. This decline is primarily attributable to an increased accumulation of errors, leading to more frequent identity switches and trajectory disruptions over extended

periods. In general, traditional tracking-by-detection pipelines exhibit greater instability as sequence duration increases, reflected in more pronounced declines in IDF1 and HOTA scores. By contrast, transformer-based models such as MOTIP and MeMOTR, which leverage query-based attention mechanisms or explicit long-term memory components, demonstrate higher consistency and robustness throughout prolonged sequences. This analysis affirms the intrinsic challenges of long-duration tracking and underscores the significance of *OinkTrack* for evaluating and advancing temporal continuity and identity persistence in extended video scenarios.

References

[1] Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft. 2016. Simple online and realtime tracking. In *IEEE International Conference on Image Processing*. 3464–3468.

[2] Jinkun Cao, Jiangmiao Pang, Xinshuo Weng, Rawal Khirodkar, and Kris Kitani. 2023. Observation-centric sort: Rethinking sort for robust multi-object tracking. In *IEEE Conference on Computer Vision and Pattern Recognition*. 9686–9696.

[3] Long Chen, Haizhou Ai, Zijie Zhuang, and Chong Shang. 2018. Real-time multiple people tracking with deeply learned candidate selection and person re-identification. In *International Conference on Multimedia and Expo*. IEEE, 1–6.

[4] Yunhao Du, Zhicheng Zhao, Yang Song, Yanyun Zhao, Fei Su, Tao Gong, and Hongying Meng. 2023. Strongsort: Make deepsort great again. *IEEE Transactions on Multimedia* 25 (2023), 8725–8737.

[5] Ruopeng Gao, Ji Qi, and Limin Wang. 2025. Multiple object tracking as id prediction. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[6] Ruopeng Gao and Limin Wang. 2023. MeMOTR: Long-term memory-augmented transformer for multi-object tracking. In *International Conference on Computer Vision*. 9901–9910.

[7] Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. 2021. Yolox: Exceeding yolo series in 2021. *arXiv:2107.08430* (2021).

[8] Nicolai Wojke, Alex Bewley, and Dietrich Paulus. 2017. Simple online and realtime tracking with a deep association metric. In *IEEE International Conference on Image Processing*. 3645–3649.

[9] Mingzhan Yang, Guangxin Han, Bin Yan, Wenhua Zhang, Jinqing Qi, Huchuan Lu, and Dong Wang. 2024. Hybrid-sort: Weak cues matter for online multi-object tracking. In *Association for the Advancement of Artificial Intelligence*, Vol. 38. 6504–6512.

[10] Fangao Zeng, Bin Dong, Yuang Zhang, Tiancai Wang, Xiangyu Zhang, and Yichen Wei. 2022. Motr: End-to-end multiple-object tracking with transformer. In *European Conference on Computer Vision*. 659–675.

[11] Yifu Zhang, Peize Sun, Yi Jiang, Dongdong Yu, Fucheng Weng, Zehuan Yuan, Ping Luo, Wenyu Liu, and Xinggang Wang. 2022. Bytetrack: Multi-object tracking by associating every detection box. In *European Conference on Computer Vision*. 1–21.