

Supplementary Materials: DenseTrack: Drone-based Crowd Tracking via Density-aware Motion-appearance Synergy

Anonymous Author(s)

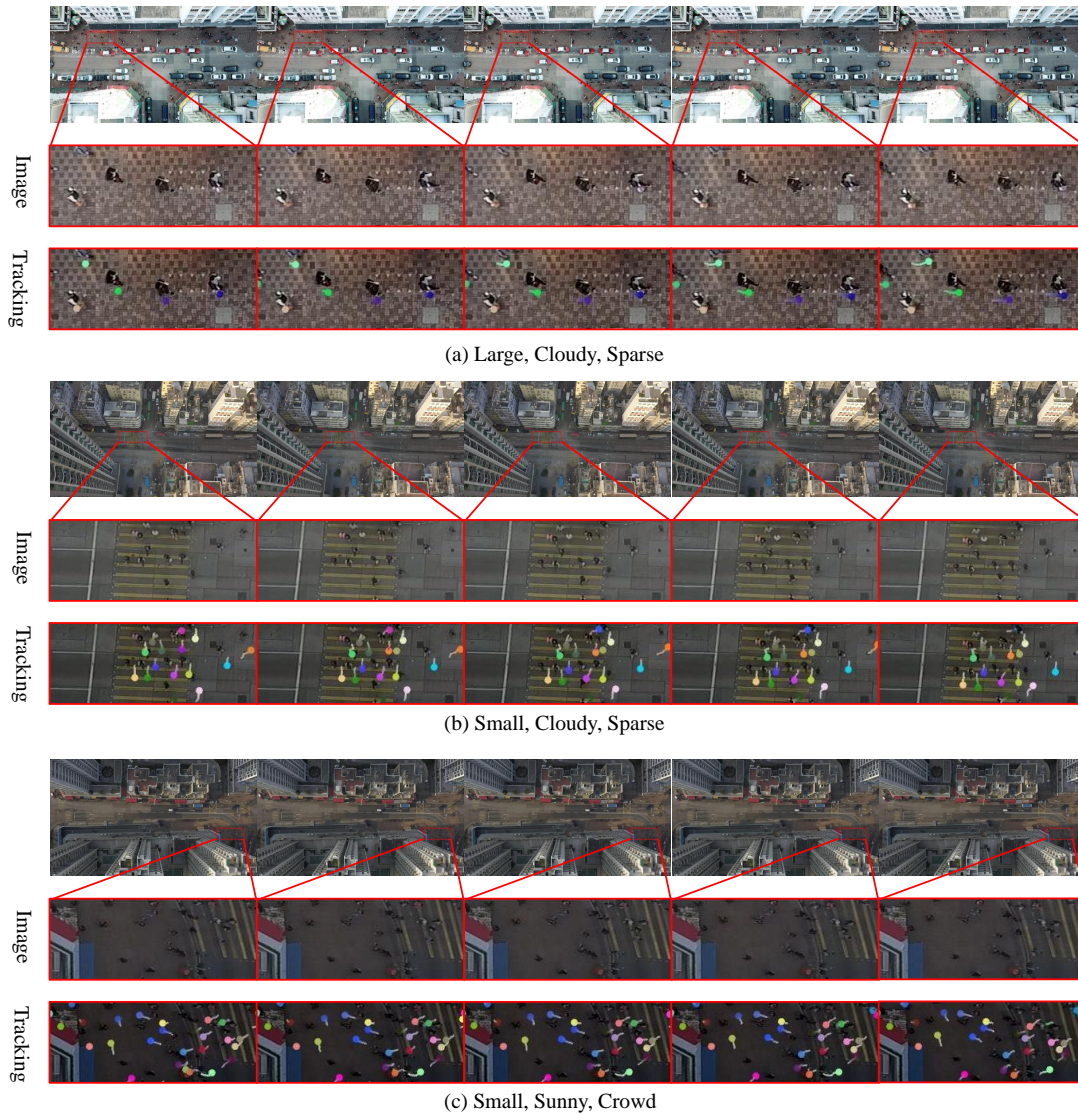


Figure 1: Illustration of tracking under different scenarios. (a) Sparse large objects in cloudy weather scenarios. (b) Sparse small objects in cloudy weather scenarios. (c) Crowd small in sunny weather scenarios.

1 VISUALIZATION OF TRACKING RESULTS IN VARIOUS SCENARIOS

We present visualizations of tracking results in different scenarios as shown in Fig. 1. In multi-object tracking tasks, Sparse large objects are easier to track due to their ease of extracting appearance features

and predicting motion offsets. Conversely, Crowd small objects pose a greater tracking challenge due to the difficulty in extracting appearance features and the insufficient discriminative power of motion offsets, making tracking more difficult. Therefore, these

Table 1: Localization performances on DRONECROWD; average L-mAP, and L-AP at each threshold (L-AP₁₀, L-AP₁₅, and L-AP₂₀). MOT and DCT stands for Multi Object Tracking and Drone-based Crowd Tracking, respectively. The best results are highlighted in bold.

| Method | MOT | DCT | L-mAP | L-AP ₁₀ | L-AP ₁₅ | L-AP ₂₀ |
|-------------------|-----|-----|--------------|--------------------|--------------------|--------------------|
| MCNN [13] | ○ | ● | 9.05 | 9.81 | 11.81 | 12.83 |
| CAN [7] | ○ | ● | 11.12 | 8.94 | 15.22 | 18.27 |
| CSRNet [4] | ○ | ● | 14.40 | 15.13 | 19.17 | 21.16 |
| DM-Count [10] | ○ | ● | 18.17 | 17.90 | 25.32 | 27.59 |
| STNNet [12] | ● | ● | 40.45 | 42.75 | 50.98 | 55.77 |
| DenseTrack (Ours) | ● | ● | 43.52 | 47.75 | 52.21 | 54.71 |

Table 2: Tracking performances on DRONECROWD; average T-mAP, and T-AP at each threshold (T-AP_{0.10}, T-AP_{0.15}, and T-AP_{0.20}).

| Method | T-mAP | T-AP _{0.10} | T-AP _{0.15} | T-AP _{0.20} |
|-------------------|--------------|----------------------|----------------------|----------------------|
| StrongSORT [3] | 8.98 | 10.63 | 8.96 | 7.34 |
| BoT-SORT [1] | 13.60 | 14.60 | 13.63 | 12.58 |
| Deep-OC-SROT [8] | 28.39 | 30.84 | 28.52 | 25.81 |
| OC-SORT [2] | 34.26 | 38.30 | 34.25 | 30.22 |
| DenseTrack (Ours) | 39.44 | 47.48 | 39.88 | 30.95 |

results indicate that our method demonstrates excellent tracking performance in both easy-to-track and challenging scenarios.

2 COMPARISON WITH THE STATE-OF-ARTS IN CROWD LOCALIZATION PERFORMANCE

Tab. 1 provides a competitive analysis of the localization performance of various methods on DRONECROWD. While the success of our localization is influenced by [12] and is not the focus of our research, localization remains a crucial task in object tracking, determining the accuracy of tracking results. Therefore, we still conducted relevant experiments to demonstrate the effectiveness of using density maps for object detection. Following the paper that introduced DRONECROWD, we evaluate the localization performance of crowds using the L-AP score. L-mAP represents the average of L-AP over different distance thresholds (1, 2, ..., 25 pixels). A smaller L-AP distance threshold implies a stricter requirement for precision in localization, while a higher L-AP value indicates better performance.

It is noteworthy that our method exhibits a more significant improvement in L-AP under stricter thresholds (compared to STNNet [12], L-AP₁₀ increases from 42.75% to 47.75%). This suggests that our method is more precise in detecting objects, which is highly beneficial for enhancing tracking performance. Furthermore, our method demonstrates improvements in overall metrics as well.

3 COMPARISON WITH MORE MOT METHOD IN TRACKING PERFORMANCE

Tab. 2 provides a comparative analysis of the tracking performance of more multi-object tracking methods on DRONECROWD. Due to the scarcity of methods explicitly tailored for tracking small objects from the aerial perspective of drones, we replicated recent multi-object methods on the DroneCrowd dataset. Moreover, to ensure fairness, we uniformly input the results of object detection based on

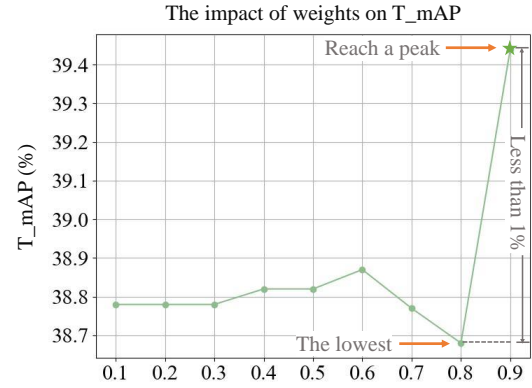


Figure 2: Tracking performances under different weights on DRONECROWD.

density map localization for these methods to avoid errors caused by localization.

OC-SORT [2] proposes a method to recover the tracking of objects lost due to occlusion within a short time window by associating the last observed value of the object with new observations, significantly enhancing its tracking performance. However, OC-SORT solely relies on motion cues for tracking, which may result in erroneous identifications of nearby individuals. Although Deep-OC-SORT [8] represents an improvement over OC-SORT, it faces performance degradation due to the challenge of implementing appearance feature extraction on small objects. In contrast, DenseTrack exhibits excellent tracking capabilities by combining robust appearance feature extraction with motion cues for tracking.

4 COMPARISON OF MOTION WEIGHT

In the inter-frame association stage, the cost matrix $A_{i,i+1}^C$ is obtained by the weighted sum of the similarity matrix $A_{i,i+1}^S$ and the normalized distance matrix $\hat{A}_{i,i+1}^D$, as shown in Eq. (1).

$$A_{i,i+1}^C = (-\lambda) \hat{A}_{i,i+1}^D + (1 - \lambda) A_{i,i+1}^S. \quad (1)$$

Fig. 2 illustrates the impact of different values of λ on tracking performance. In DenseTrack, the choice of λ does not significantly affect tracking performance. For instance, regarding T-mAP, it ranges from a maximum of 39.44% to a minimum of 38.68%, with a difference of less than 1%.

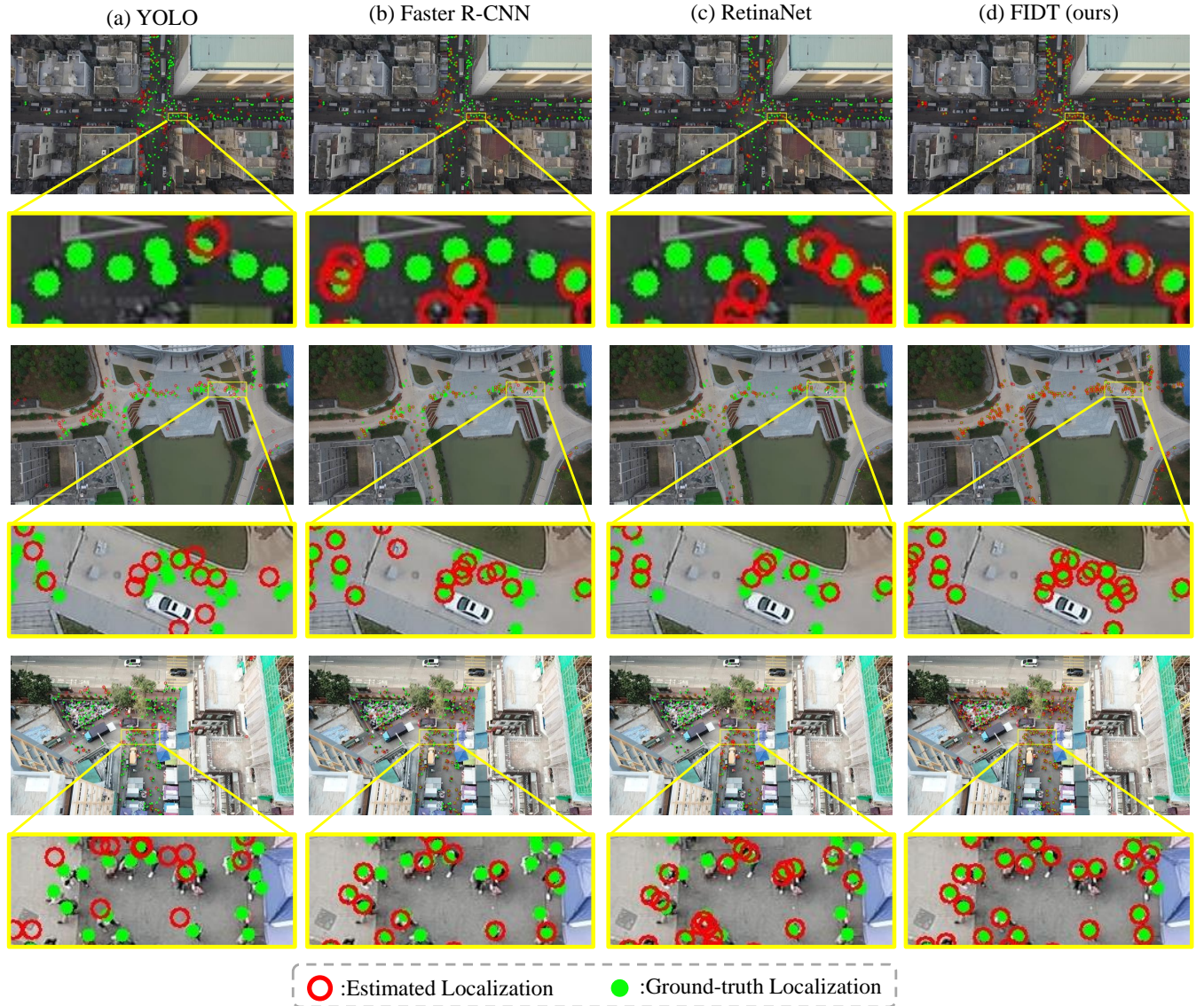


Figure 3: Illustration of localization under different detector. (a) Use YOLOv8 to detect objects. (b) Use Faster R-CNN to detect objects. (c) Use RetinaNet to detect objects. (d) Use density map (FIDT) to detect objects.

Table 3: Localization performances of different detector on DRONECROWD; average L-mAP, and L-AP at each threshold (L-AP₁₀, L-AP₁₅, and L-AP₂₀);

| Method | L-mAP | L-AP ₁₀ | T-AP ₁₅ | T-AP ₂₀ |
|------------------|--------------|--------------------|--------------------|--------------------|
| YOLOv8 [11] | 6.62 | 1.37 | 7.16 | 14.26 |
| Faster R-CNN [9] | 22.39 | 24.35 | 26.68 | 28.27 |
| RetinaNet [6] | 22.63 | 24.28 | 28.37 | 30.78 |
| FIDT (ours) | 43.55 | 47.77 | 52.24 | 54.77 |

5 COMPARISON WITH OTHER DETECTOR IN CROWD LOCALIZATION PERFORMANCE

Tab. 3 displays the localization performance of different detectors on DRONECROWD. Since DroneCrowd is captured from a high-altitude overhead perspective by drones, the distinction between objects and background is not very pronounced, which also affects the performance of commonly used detectors in terms of localization.

Inspired by the work [5], DenseTrack utilizes density maps originally designed for crowd counting in the object localization stage. The results indicate that this approach effectively enhances the accuracy of object detection, facilitating subsequent tracking tasks.

6 VISUALIZATION OF VARIOUS DETECTOR

While localization is not the primary focus of DenseTrack, the accuracy of object localization directly impacts the quality of appearance feature extraction, thereby influencing tracking results. Additionally, the tracking outcomes of the Detection-based tracking paradigm heavily rely on the accuracy of object detection. Therefore, crowd localization is a crucial step in tracking.

Fig. 3 further demonstrates the differences in localization performance among YOLOv8 [11], Faster R-CNN [9], RetinaNet [6], and the FIDT [5] we employ. The visualization results indicate that other detectors may experience missed detections when detecting dense crowds, whereas using density maps for detection can reduce the occurrence of such cases. Additionally, employing density maps for detection can enhance localization accuracy, thereby facilitating subsequent appearance extraction tasks.

REFERENCES

- [1] Nir Aharon, Roy Orfaig, and Ben-Zion Bobrovsky. 2022. BoT-SORT: Robust Associations Multi-Pedestrian Tracking. *arXiv:2206.14651* (2022).
- [2] Jinkun Cao, Jiangmiao Pang, Xinshuo Weng, Rawal Khirodkar, and Kris Kitani. 2023. Observation-Centric SORT: Rethinking SORT for Robust Multi-Object Tracking. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.* 9686–9696.
- [3] Yunhao Du, Zhicheng Zhao, Yang Song, Yanyun Zhao, Fei Su, Tao Gong, and Hongying Meng. 2023. StrongSORT: Make DeepSORT Great Again. *IEEE Trans. Multimedia* 25 (2023), 8725–8737.
- [4] Yuhong Li, Xiaofan Zhang, and Deming Chen. 2018. CSRNet: Dilated Convolutional Neural Networks for Understanding the Highly Congested Scenes. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.* 1091–1100.
- [5] Dingkan Liang, Wei Xu, Yingying Zhu, and Yu Zhou. 2023. Focal Inverse Distance Transform Maps for Crowd Localization. *IEEE Trans. Multimedia* 25 (2023), 6040–6052.
- [6] Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár. 2020. Focal Loss for Dense Object Detection. *IEEE Trans. Pattern Anal. Mach. Intell.* 42, 2 (2020), 318–327. <https://doi.org/10.1109/TPAMI.2018.2858826>
- [7] Weizhe Liu, Mathieu Salzmann, and Pascal Fua. 2019. Context-Aware Crowd Counting. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.* 5099–5108.
- [8] Gerard Maggolino, Adnan Ahmad, Jinkun Cao, and Kris Kitani. 2023. Deep OC-Sort: Multi-Pedestrian Tracking by Adaptive Re-identification. In *Proc. IEEE Int. Conf. Image Process.* 3025–3029.
- [9] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. 2015. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7–12, 2015, Montreal, Quebec, Canada*, Corinna Cortes, Neil D. Lawrence, Daniel D. Lee, Masashi Sugiyama, and Roman Garnett (Eds.), 91–99.
- [10] Boyu Wang, Huidong Liu, Dimitris Samaras, and Minh Hoai Nguyen. 2020. Distribution Matching for Crowd Counting. In *Adv. Neural Inf. Process. Syst.*
- [11] Jiayuan Wang, Q. M. Jonathan Wu, and Ning Zhang. 2023. You Only Look at Once for Real-time and Generic Multi-Task. *CoRR* abs/2310.01641 (2023).
- [12] Longyin Wen, Dawei Du, Pengfei Zhu, Qinghua Hu, Qilong Wang, Liefeng Bo, and Siwei Lyu. 2021. Detection, Tracking, and Counting Meets Drones in Crowds: A Benchmark. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.* 7812–7821.
- [13] Yingying Zhang, Desen Zhou, Siqin Chen, Shenghua Gao, and Yi Ma. 2016. Single-Image Crowd Counting via Multi-Column Convolutional Neural Network. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.* 589–597.