Supplementary material for: NT-VOT211: A Large-Scale Benchmark for Night-time Visual Object Tracking

Yu Liu[1][0009-0009-5898-0113], Arif Mahmood[2][0000-0001-5986-9876], and Muhammad Haris Khan[3][0000-0001-9746-276X]

[1] Xinjiang University
750184785ly@gmail.com
[2] Information Technology University
arif.mahmood@itu.edu.pk
[3] Mohamed bin Zayed University of Artificial Intelligence
muhammad.haris@mbzuai.ac.ae

What will be public along with the dataset: In addition to presenting the dataset itself, we are committed to sharing the source code of our annotation tools, which has played a pivotal role in streamlining our annotation workflow. Moreover, we will disclose the source code that is instrumental in computing the computer-labeled attributes, ensuring transparency and reproducibility.

We are also proud to offer a comprehensive toolkit designed to operate seamlessly with our dataset. This toolkit will be fully compatible with popular tracking libraries, Pytracking, thereby facilitating a wide range of applications and analyses.

To ensure your convenience and ease of use, we will not only supply the raw data as depicted in Table 2 but also grant access to our dedicated evaluation server at eval.com. Drawing inspiration from TrackingNet[6], users will be able to upload their raw results onto the server. Upon upload, their scores will be meticulously evaluated, offering users the option to display their scores on the leaderboard publicly, should they wish to do so. This approach not only fosters a collaborative environment but also encourages continuous improvement and benchmarking within the community.

The Distribution of the proposed dataset We conduct a detailed attribute analysis to explore the characteristics of the proposed dataset. In Figure 1, we present the UMAP visualization of these benchmarks. The visualization reveals distinct patterns in the proposed dataset, showing overlap with other benchmarks while also forming some unique clusters in the top-left corner. This unique pattern poses potential challenges to the generalization ability of tracking algorithms when compared to other datasets.

Why challenging? The primary challenges in our proposed benchmark are attributed to the unique distribution of attributes. To elaborate further, we specifically examined failure scenarios of State-of-the-Art (SOTA) trackers and illustrated the results in Figure 2,3,4,5. (**For more details see our demo.mp4**). At the top of each figure, we display the unique combination of attributes, while at the bottom, we describe the content of the video. It's apparent that even in

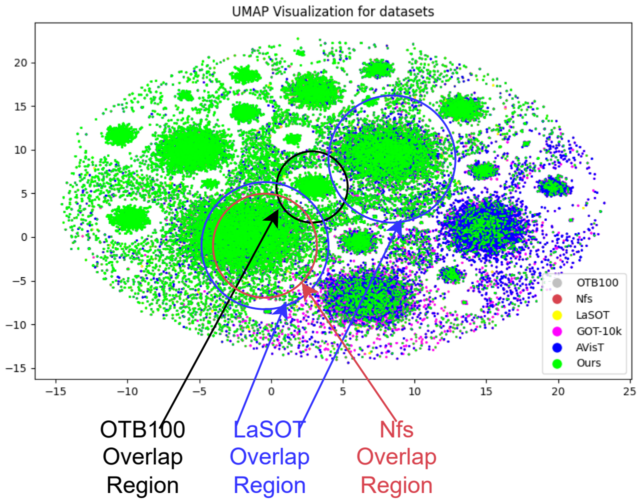
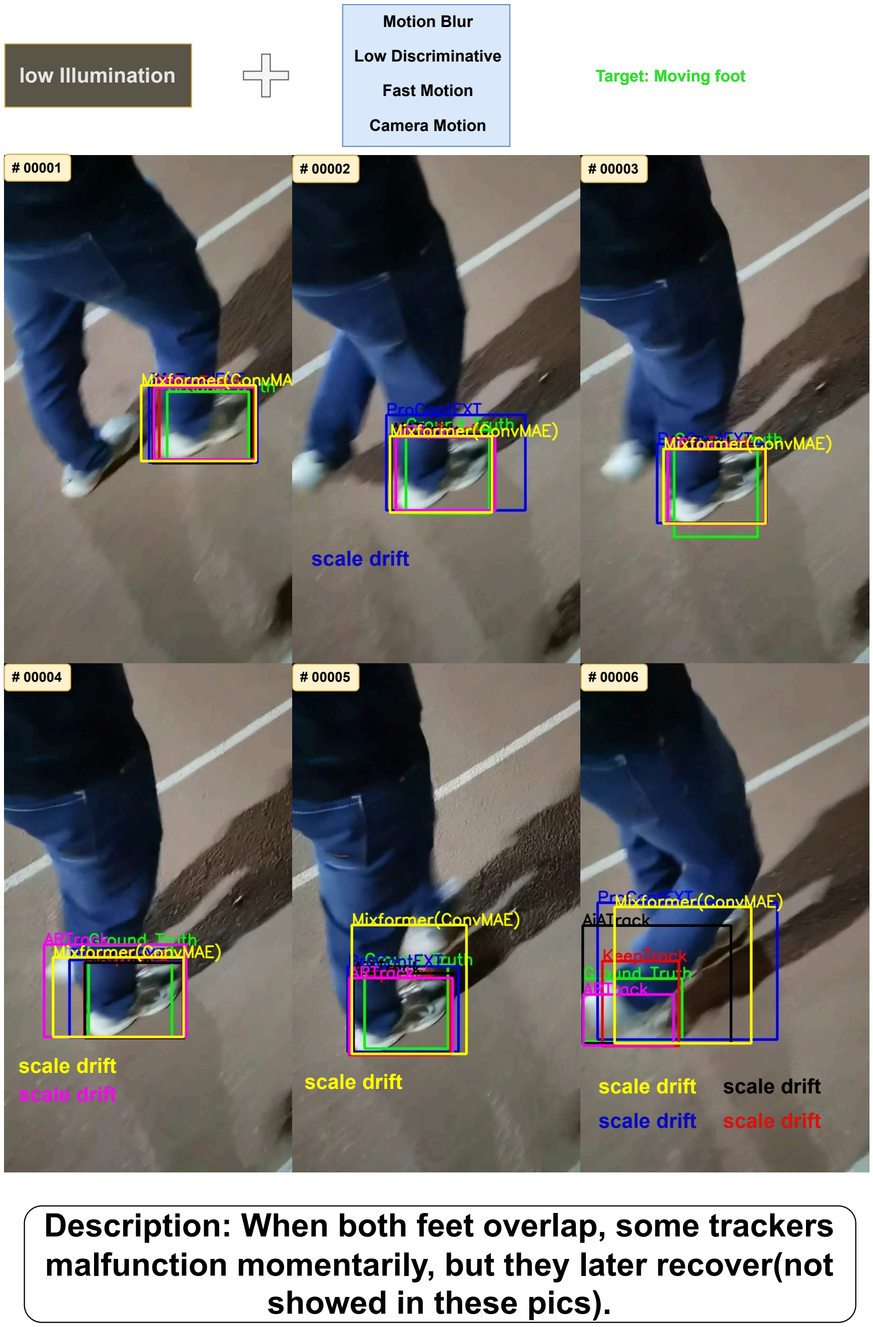


Fig. 1. To analyze the diversity of dataset attributes, each frame is represented by a 6-element feature vector encoding the presence of various computer-labeled attributes. By randomly sampling 50,000 frames per dataset and visualizing the distribution of these attribute features, we can compare dataset characteristics. The distributions for Nfs, OTB100, and LaSOT exhibit significant overlap, concentrated within the circled regions. In contrast, NT-VOT211 shows complementary with AVisT and GOT-10K while maintaining diversity. The breadth of the NT-VOT211 distribution demonstrates its wide variability in attributes, posing generalization challenges for trackers.

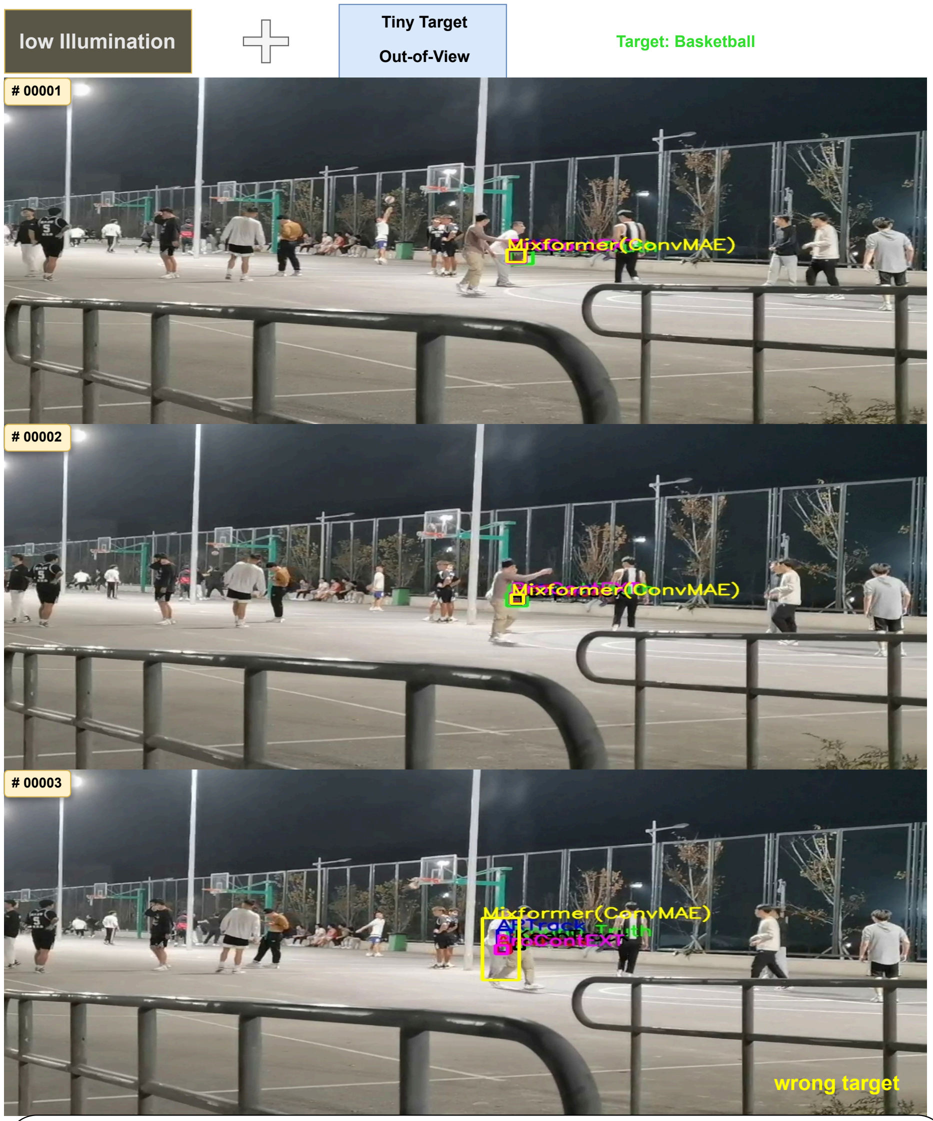
scenes as depicted in Figure 2 and Figure 4, which are relatively less complex, the trackers still failed due to the unique distribution of attributes.

Other Statistics: We also include movement statistics of the target object in our analysis. In this data, we group frames into sets of 60 each, measuring the movement relative to the initial frame within each group. Our statistics include OTB100[8], Nfs[4], LaSOT[2], GOT-10k[3], AVisT[7] and our private dataset. The results are depicted in Figure 6 and Figure 7, where each unit on the x-axis represents movement equivalent to 0.25 times width or height of the target. We noticed that, unlike other datasets such as LaSOT [2], OTB2015 [8], and GOT-10k [3], our dataset presents a distinct challenge. Considering that the most short-distance targets are recorded with handheld devices and are intended to be placed around the center of the frame, these targets tend to move less compared to those in other datasets. However, this characteristic did not prevent our proposed benchmark from being more challenging compared to others. As illustrated in Figure 4, even when the cup and the camera are stationary, two trackers, Mixformer[1] and ProcontEXT[5], still exhibit drifting due to negligible camera shake.



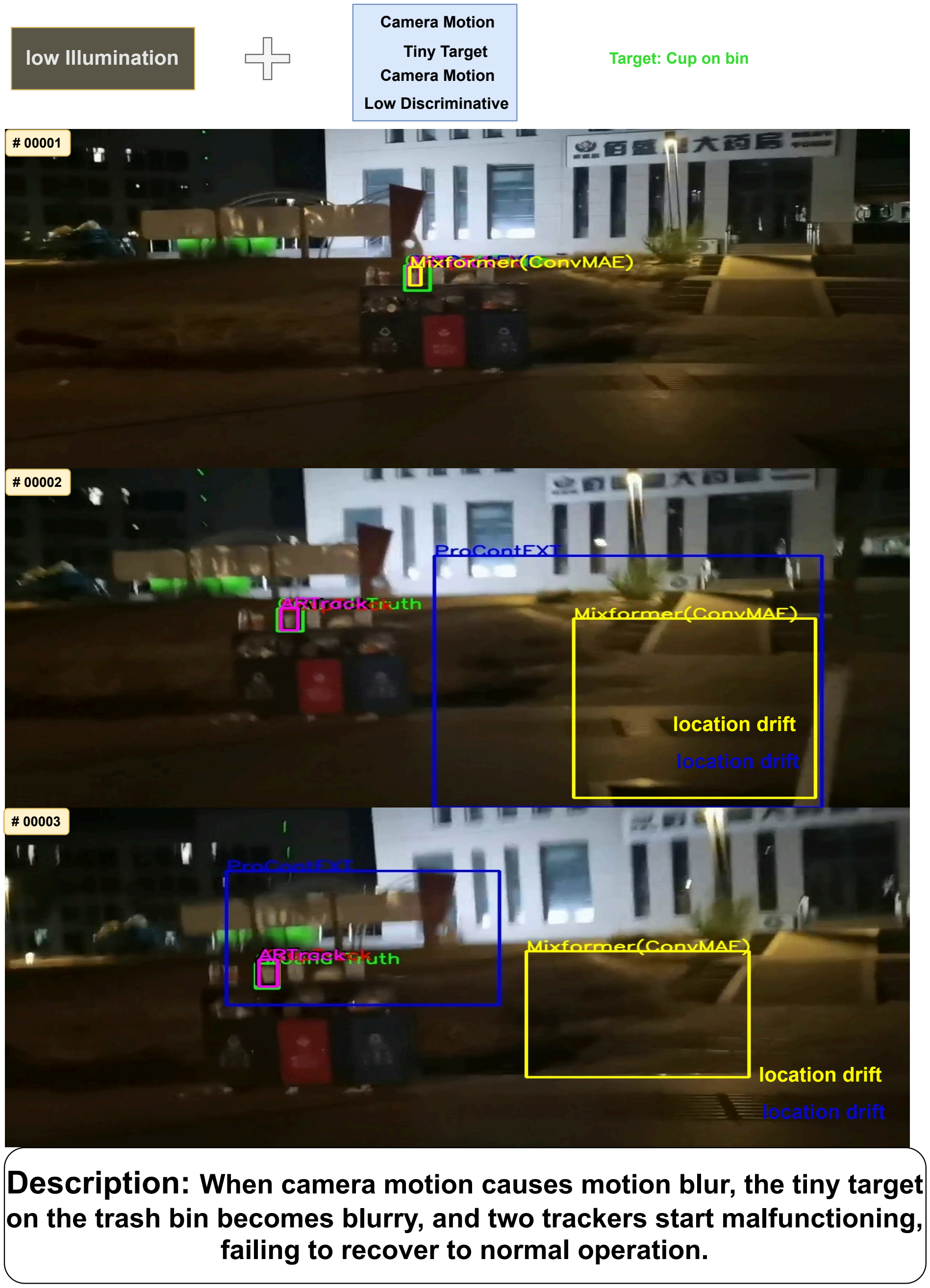
Description: When both feet overlap, some trackers malfunction momentarily, but they later recover(not showed in these pics).

Fig. 2. Failure scenarios of SOTA #1



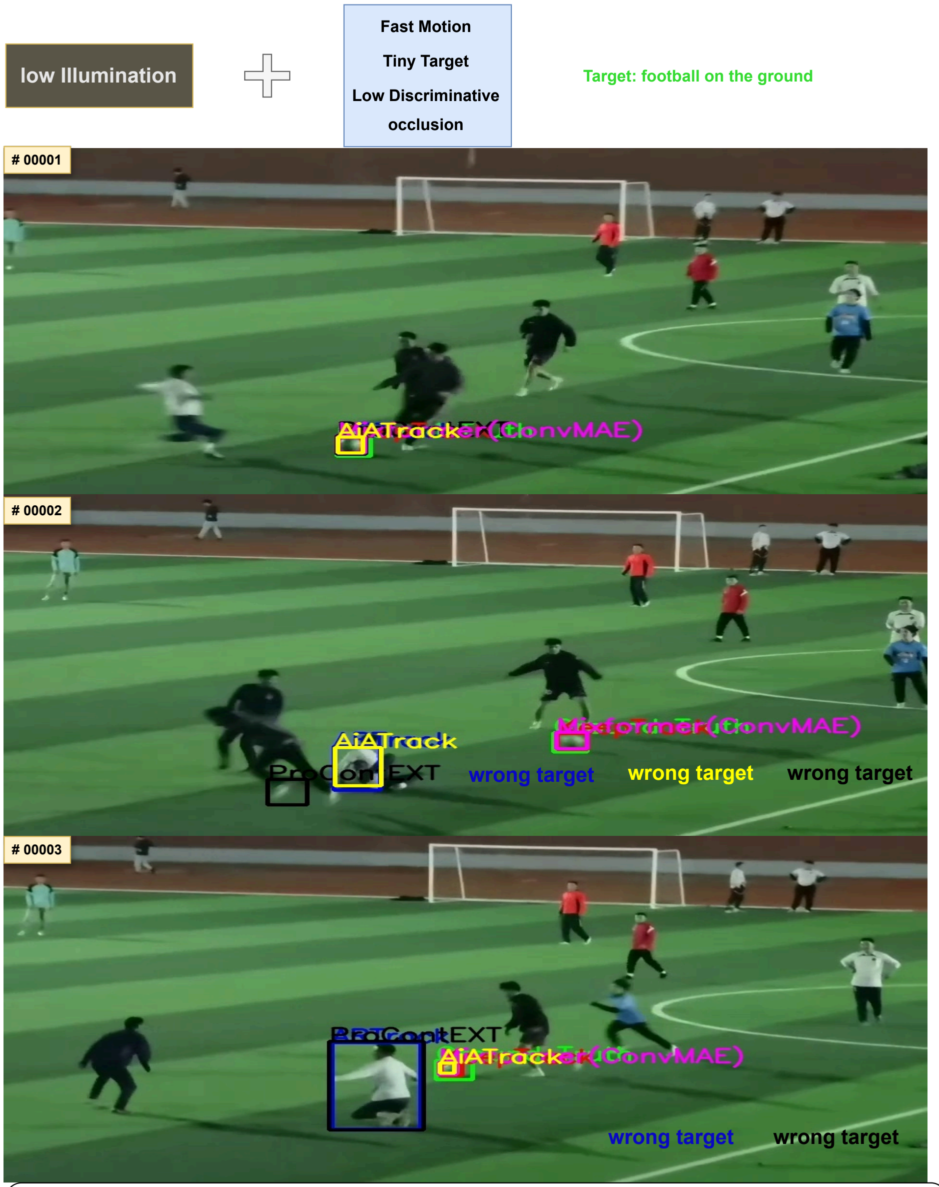
Description: When the basketball goes out of view due to the player's movements, Mixformer starts tracking the entire person instead of the ball. In subsequent footage, Mixformer resumes ball tracking.(not showed in these pics).

Fig. 3. Failure scenarios of SOTA #2



Description: When camera motion causes motion blur, the tiny target on the trash bin becomes blurry, and two trackers start malfunctioning, failing to recover to normal operation.

Fig. 4. Failure scenarios of SOTA #3



Description: A person comes to snatch the football and falls, causing the ball to enter fast motion. The trackers identify the fallen athlete as the target and fail to recover thereafter.

Fig. 5. Failure scenarios of SOTA #4

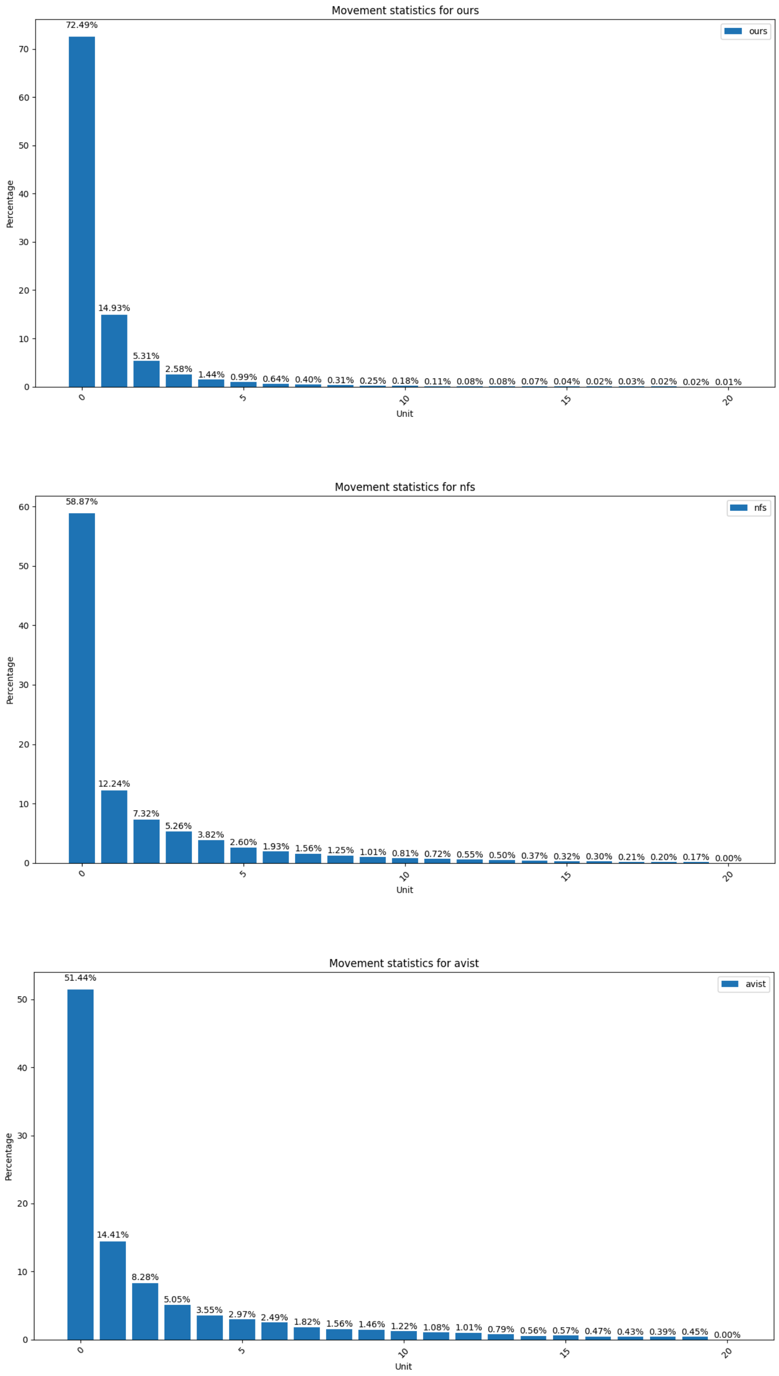


Fig. 6. Statistics on Movement #1

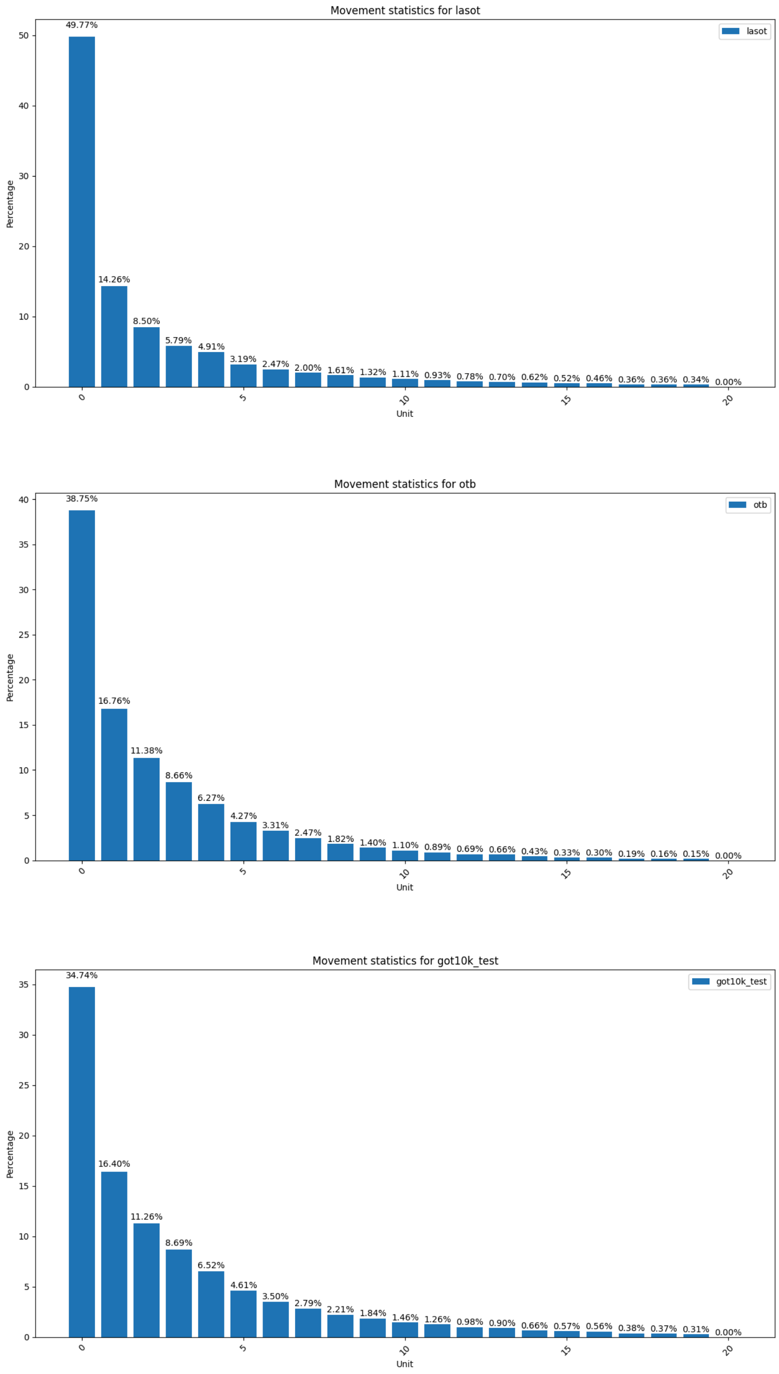


Fig. 7. Statistics on Movement #2

References

1. Cui, Y., Jiang, C., Wang, L., Wu, G.: Mixformer: End-to-end tracking with iterative mixed attention. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13608–13618 (2022)
2. Fan, H., Lin, L., Yang, F., Chu, P., Deng, G., Yu, S., Bai, H., Xu, Y., Liao, C., Ling, H.: Lasot: A high-quality benchmark for large-scale single object tracking. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5374–5383 (2019)
3. Huang, L., Zhao, X., Huang, K.: Got-10k: A large high-diversity benchmark for generic object tracking in the wild. IEEE transactions on pattern analysis and machine intelligence **43**(5), 1562–1577 (2019)
4. Kiani Galoogahi, H., Fagg, A., Huang, C., Ramanan, D., Lucey, S.: Need for speed: A benchmark for higher frame rate object tracking. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1125–1134 (2017)
5. Lan, J.P., Cheng, Z.Q., He, J.Y., Li, C., Luo, B., Bao, X., Xiang, W., Geng, Y., Xie, X.: Procontext: Exploring progressive context transformer for tracking. In: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1–5. IEEE (2023)
6. Muller, M., Bibi, A., Giancola, S., Alsubaihi, S., Ghanem, B.: Trackingnet: A large-scale dataset and benchmark for object tracking in the wild. In: Proceedings of the European conference on computer vision (ECCV). pp. 300–317 (2018)
7. Noman, M., Ghallabi, W.A., Kareem, D., Mayer, C., Dudhane, A., Danelljan, M., Cholakkal, H., Khan, S., Gool, L.V., Khan, F.S.: Avist: A benchmark for visual object tracking in adverse visibility. In: 33rd British Machine Vision Conference 2022, BMVC 2022, London, UK, November 21-24, 2022. p. 817. BMVA Press (2022), https://bmvc2022.mpi-inf.mpg.de/817/
8. Wu, Y., Lim, J., Yang, M.H.: Object tracking benchmark. IEEE Transactions on Pattern Analysis and Machine Intelligence **37**(9), 1834–1848 (2015). https://doi.org/10.1109/TPAMI.2014.2388226