360VOT: A New Benchmark Dataset for Omnidirectional Visual Object Tracking - Supplementary -

Huajian Huang Yinzhe Xu Yingshu Chen Sai-Kit Yeung
The Hong Kong University of Science and Technology
{hhuangbg, yxuck, ychengw}@connect.ust.hk, saikit@ust.hk

Abstract

In this supplementary, we first demonstrate the proposed 360 tracking framework in detail. Then in Sec. 2, we detail the data collection criteria and categorization. In Sec. 3, we provide more information on annotation including the segmentation toolkit and conversion algorithms from masks to bounding boxes. Finally, we show more qualitative and quantitative results, such as performance comparison between tangent BFoV and our extended BFoV, tracking visual results on challenging sequences with exclusive attributes, and more quantitative results among different trackers on 360VOT. In addition, we have a supplementary video[1] to show some sequences with four representation ground truths and display more omnidirectional video object tracking results in challenging scenarios from 360VOT benchmark dataset.

1. 360 Tracking Framework

We use a spherical camera model to depict the relationship between the 3D camera space $[X, Y, Z]$ and the 2D image space $[u, v]$. The projection function $\mathcal{F}$ is formulated as:

$$u = (\frac{lon}{2\pi} + 0.5) * W = arctan(X/Z), \tag{1}$$

$$v = (-\frac{lat}{\pi} + 0.5) * H = arctan(\frac{-Y}{\sqrt{X^2 + Z^2}}), \tag{2}$$

where $-\pi < lon < \pi$ and $-\pi/2 < lat < \pi/2$ denote the longitude and latitude in the spherical coordinate system respectively. W and H are the width and height of the 360° image. As we mention in the main paper, a (r)BFoV is denoted as $[clon, clat, \theta, \phi, \gamma]$, where $clon$ and $clat$ represent the object center in the spherical coordinate system, θ and ϕ are the maximum bounding FoVs of the object, the rotation γ of BFoV is zero. If we use a tangent plane $T \in \mathbb{R}^3$ to model the represented region of (r)BFoV, the corresponding region on 360° is formulated as:

$$I(\text{(r)BFoV} \mid \Omega) = \mathcal{F}(\mathcal{R}_y(clon) \cdot \mathcal{R}_x(clat) \cdot \mathcal{R}_z(\gamma) \cdot \Omega). \tag{3}$$

where,

$$\mathcal{R}_y(clon) = \begin{bmatrix} cos(clon) & 0 & sin(clon) \\ 0 & 1 & 0 \\ -sin(clon) & 0 & cos(clon) \end{bmatrix}, \tag{4}$$

$$\mathcal{R}_x(clat) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & cos(clat) & -sin(clat) \\ 0 & sin(clat) & cos(clat) \end{bmatrix}, \tag{5}$$

$$\mathcal{R}_z(\gamma) = \begin{bmatrix} cos\gamma & -sin\gamma & 0 \\ sin\gamma & cos\gamma & 0 \\ 0 & 0 & 1 \end{bmatrix}, \tag{6}$$

$$\Omega = T = \begin{bmatrix} \mathbf{X} \\ \mathbf{Y} \\ \mathbf{Z} \end{bmatrix} = \begin{bmatrix} -tan(\theta/2) : tan(\theta/2) \\ -tan(\phi/2) : tan(\phi/2) \\ 1 \end{bmatrix}, \tag{7}$$

To handle a large FoV, we extend the represented region of BFoV. When the FoV is larger than the threshold, e.g., 90°, the bounding region of BFoV becomes a surface segment $S \in \mathbb{R}^3$ of the unit sphere:

$$S = \begin{bmatrix} cos(\Phi)sin(\Theta) \\ -sin(\Phi) \\ cos(\Phi)cos(\Theta) \end{bmatrix}, \tag{8}$$

where, $\Phi \in [-\phi/2, \phi/2], \Theta \in [-\theta/2, \theta/2]$. Therefore, the corresponding region of extended (r)BFoV on 360° is formulated as:

$$I(\text{(r)BFoV} \mid \Omega), \quad \Omega = \begin{cases} T(\theta, \phi), & \theta < 90°, \phi < 90° \\ S(\theta, \phi), & otherwise \end{cases}. \tag{9}$$

Based on the I which actually records the corresponding pixel coordinates of 360°, we can remap the 360° image and extract a local search region to perform tracking which generates a BBox or rBBox prediction relative to the local region. After that, we still take advantage of I, converting

[1] https://github.com/HuajianUP/360VOT/

Benchmark	Videos	Total frames	Min frames	Mean frames	Median frames	Max frames	Object classes	Attr.	Annotation	Feature	Year
ALOV300[13]	314	152K	19	483	276	5,975	64	14	sparse BBox	diverse scenes	2013
OTB100[18]	100	81K	71	590	393	3,872	16	11	dense BBox	short-term	2015
NUS-PRO[8]	365	135K	146	371	300	5,040	8	12	dense BBox	occlusion-level	2015
TC128[10]	129	55K	71	429	365	3,872	27	11	dense BBox	color enhanced	2015
UAV123[11]	123	113K	109	915	882	3,085	9	12	dense BBox	UAV	2016
DTB70[9]	70	16K	68	225	202	699	29	11	dense BBox	UAV	2016
NfS[6]	100	383K	169	3,830	2,448	20,665	17	9	dense BBox	high FPS	2017
UAVDT[1]	100	78K	82	778	602	2,969	27	14	sparse BBox	UAV	2017
TrackingNet*[12]	511	226K	96	441	390	2,368	27	15	sparse BBox	large scale	2018
OxUvA[16]	337	1.55M	900	4,260	2,628	37,440	22	6	sparse BBox	long-term	2018
LaSOT*[3]	280	685K	1,000	2,448	2,102	9,999	85	14	dense BBox	category balance	2018
GOT-10k*[5]	420	56K	29	127	100	920	84	6	dense BBox	generic	2019
TOTB[4]	225	86K	126	381	389	500	15	12	dense BBox	transparent	2021
TREK-150[2]	150	97K	161	649	484	4,640	34	17	dense BBox	FPV	2021
VOT[7]	62	20K	41	321	242	1,500	37	9	dense BBox	annual	2022
360VOT	120	113K	251	940	775	2,400	32	20	dense (r)BBox & (r)BFoV	360° images	2023

Table 1: Comparison of current popular benchmarks for visual single object tracking in the literature. * indicates that only the test set of each dataset is reported.

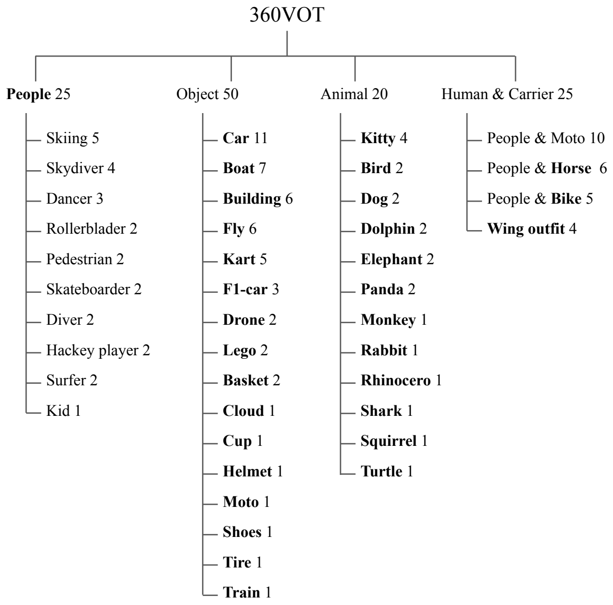


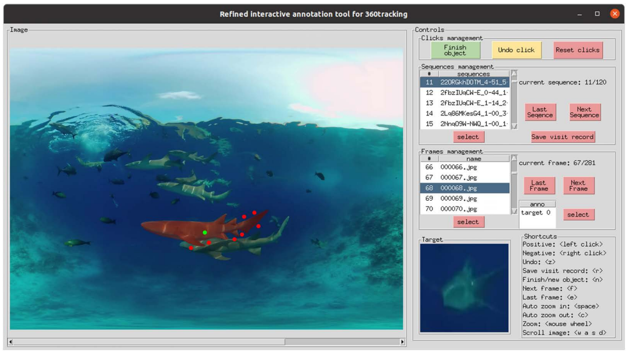
Figure 1: 360VOT contains 120 sequences in diverse scenarios and 32 object categories which are denoted in bold.

the local prediction to obtain a global bounding region. To get the final (r)BBox prediction, we can calculate the minimum area (rotated) rectangle on the 360°. In addition, we can re-project the coordinates of the bounding region on the 360° image to the spherical coordinates system and calculate the maximum bounding FoV for (r)BFoV.

2. Details of 360VOT Collection

We manually collected videos from YouTube and captured some using a 360-degree camera. Four features were recorded for each sequence: *camera motion* (moving and stationary), *target classes* (humans, animals, rigid objects and non-rigid objects), *duration* (18 seconds to 75 minutes) and *environment*. Specifically, the *envonment* varies among indoor-outdoor, illumination (daylight, white light and night) and weather (cloudy, sunny and rainy). We ranked and filtered videos considering four criteria of tracking difficulty scale and some additional challenging cases. First, videos with the more adequate relative motion of the target and camera rank higher. Targets are preferably in a high degree of mobility, appearing in various locations across the frames rather than stationary. Second, videos with higher variability of the environment rank higher. The video background is supposed to be ever-changing across the video, such as with variations in lighting conditions. Third, videos with the target crossing frame boundaries rank higher. The object moving across frame boundaries is a distinct feature in panoramic videos. Finally, videos with a sufficient duration rank higher. A sufficient length of video provides a higher feasibility for the diversity of target movements and deformations, and possible disappearances, increasing tracking difficulties across the video.

Eventually, the 360VOT benchmark dataset contains 120 sequences with about 113K frames in total. The minimum of frames is 251 while the average is 940. The types of targets can be placed in four major categories: *People*, *Object*, *Animal* and *Human & Carrier*. In counting the class number of 360VOT, instead of subdividing the classes of humans, we describe it in a single broad category as *People*. Since Horse and Bike classes in 360VOT always co-occur,



#	sequences
11	22ORGkhDOTM_4-51_5
12	2fbzIUqCW-E_0-44_1
13	2fbzIUqCW-E_1-14_2
14	2Lq86MKesG4_1-00_3
15	2Nnq09W-NWQ_1-00_1

current sequence: 11/120

#	name
66	000066.jpg
67	000067.jpg
68	000068.jpg
69	000069.jpg
70	000070.jpg

current frame: 67/281

anno
target 0

Shortcuts
Positive: <left click>
Negative: <right click>
Undo: <z>
Save visit record: <r>
Finish/new object: <n>
Next frame: <f>
Last frame: <e>
Auto zoom in: <space>
Auto zoom out: <c>
Zoom: <mouse wheel>
Scroll image: <w a s d>

Figure 2: We take advantage of the click-based interactive segmentation model [14] and develop a semi-automatic annotation tool to significantly increase the efficiency and attain high-quality annotation. Annotators refine a segment via green positive and red negative clicks.

we only classify them in *Human & Carrier* but not in *Object* and *Animal*. Finally, we consider it most appropriate to divide all targets into 32 categories, with the details of them shown in Figure 1. The comparison with current popular benchmarks is detailed on Table 1.

3. Annotation

Annotation is usually tedious and labor-intensive. For some existing benchmarks, they mentioned that they hired a large annotation team, more than 10 experts in the tracking domain, to manually label an enormous number of BBoxes over several months. However, such a strategy is not applicable for us to get 4 high-quality types of annotation. In addition, even though we can hire so many annotators with professional backgrounds, it is difficult to guarantee that the subjective annotations are optimal as the ground truth. The work [2] also reports the BBox annotation quality problem in the popular tracking benchmarks [18]. To obtain unbiased ground truth, we decide to segment the per-pixel target instance in each frame and then compute the corresponding (r)BBox and (r)BFoV from the resultant masks.

3.1. Segmentation toolkit

To efficiently obtain fine-grained segmentation, we utilize a state-of-the-art tracker to get the initial positions of the targets with human online revision. The initial positions are then used by a semi-automatic segmentation toolkit to initialize the target object segmentation. The toolkit is based on a click-based interactive segmentation framework [14]. The framework utilizes the HRNet-32 [15, 17] IT-M model trained on the COCO+LVIS dataset which can generate a complete segmentation on the instance with a few clicks. If the initial segmentation does not cover the target completely or contains elements not belonging to the target, positive (green) or negative (red) guiding points are manually added to generate a more accurate refined segmentation as shown in Figure 2.

Algorithm 1 Mask to (r)BBox

```
Input: The mask $M$ and boolean value $needRotation$
  /*Step 1*/
  if $M$ is empty then
      return $None$
  $w_M \leftarrow$ the width of the mask
  Convert Bound $M$ to set of polygons.
  Estimate the largest segment and calculate the centroid $[x_1, y_1]$ in terms of the spherical coordinates, $\theta_1, \phi_1$
  $\Delta x \leftarrow x_1 - w_M/2$
  /*Step 2*/
  Rotate $M$ by $\mathcal{R}_y(\theta_1) \sim$ Eq. 4, giving $M_{R_1}$
  /*Step 3*/
  if $needRotation$ then
      Bound $M_{R_1}$ by the minimum area rotated rectangle $[cx, cy, w, h, \gamma]$
  else
      Bound $M_{R_1}$ by the minimum area rectangle $[cx, cy, w, h]$
      $\gamma \leftarrow 0$
  if $w < w_M - 1$ then
      $cx \leftarrow cx + \Delta x$
  else
      $cx \leftarrow w/2$
  return $(cx, cy, w, h, \gamma)$
```

3.2. Mask to (r)BBox and (r)BFoV

Essentially, the optimal annotation is to minimize the bounding area of the target. We can convert the mask to generate 4 types of unbiased ground truths. Specifically, since the masked target may span the left and right borders of the image, we first estimate the largest segment and then rotate the mask based on the centroid c_1 of the largest segment. To estimate BBox and rBBox, we only need to move the c_1 to the horizontal center of the mask image via Eq. 4 and then calculate the minimum area rectangle and rotate the rectangle respectively. However, for estimating the (r)BFoV, we need to rotate the mask centroid to the image center via Eq. 4 and 5 twice in order to reduce the distortion as much as possible. It is necessary to guarantee the accuracy of the estimations, especially for a large FoV. Next, we can calculate the bounding FoV to get the BFoV. But in terms of rBFoV estimation, we utilize the rotating calipers algorithm to estimate the rotation and then further rotate the mask via Eq. 6 before calculating the bounding FoV. These processes are described in Algo. 1 and 2, and also illustrated in Figure 3.

4. More Results

Tangent BFoV vs extended BFoV. As the FoV increases, the regions extracted by the tangent BFoV suffer extreme distortion, which would impact the tracking performance. To further verify the effectiveness of extended BFoV, we

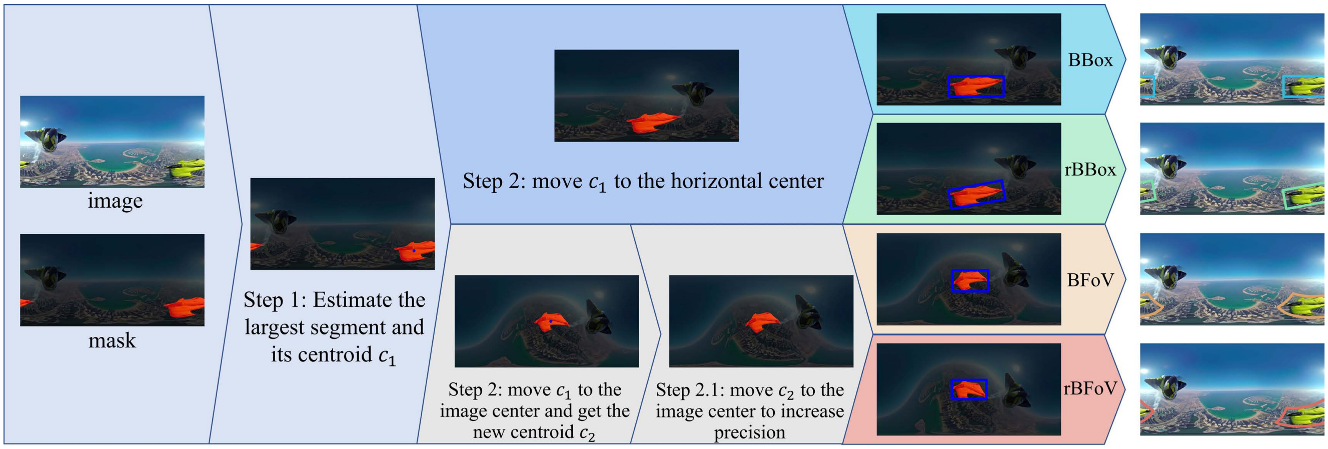


image
mask
Step 1: Estimate the largest segment and its centroid c_1
Step 2: move c_1 to the horizontal center
Step 2: move c_1 to the image center and get the new centroid c_2
Step 2.1: move c_2 to the image center to increase precision
BBox
rBBox
BFoV
rBFoV

Figure 3: The 4 different annotations are generated by minimizing the bounding region of the object according to the segmentation.

Algorithm 2 Mask to (r)BFoV

```
Input: The mask $M$ and boolean value $needRotation$
  /*Step 1*/
  if $M$ is empty then
      return $None$
  Convert Bound $M$ to set of polygons.
  Estimate the largest segment and calculate the centroid in terms of the spherical coordinates, $\theta_1, \phi_1$
  /*Step 2*/
  Rotate $M$ by $R_1 = \mathcal{R}_y(\theta_1)\mathcal{R}_x(\phi_1) \sim$ Eq. 4 and 5, giving $M_{R_1}$
  Calculate and convert the centroid of $M_{R_1}$ to the original $M$ in terms of the spherical coordinates, $\theta_2, \phi_2$
  /*Step 2.1*/
  Rotate $M$ by $R_2 = \mathcal{R}_y(\theta_2)\mathcal{R}_x(\phi_2)$, giving $M_{R_2}$
  Calculate the centroid $c_{R_2}$ , bounding width $w_{R_2}$, height $h_{R_2}$, and rotation $\gamma_{R_2}$ of $M_{R_2}$ by rotating calipers algorithm
  Convert $c_{R_2}$ to the original $M$ and get the centroid in terms of the spherical coordinates, $\theta_3, \phi_3$
  if $needRotation$ then
      if $w_{R_2} > h_{R_2}$ then
          $\gamma \leftarrow \gamma_{R_2}$
      else
          $\gamma \leftarrow \gamma_{R_2} - 90$
  else
      $\gamma \leftarrow 0$
  /*Step 3*/
  Rotate $M$ by $R_3 = \mathcal{R}_y(\theta_3)\mathcal{R}_x(\phi_3)\mathcal{R}_z(\gamma) \sim$ Eq. 4- 6, giving $M_{R_3}$
  Calculate the range of longitude $[lon_{min}, lon_{max}]$ and latitude $[lat_{min}, lat_{max}]$ of $M_{R_3}$
  Convert longitude center $(lon_{max}+lon_{min})/2$ of $M_{R_3}$ to orginal $M$, giving $clon$
  Convert latitude center $(lat_{max}+lat_{min})/2$ of $M_{R_3}$ to orginal $M$, giving $clat$
  $\theta \leftarrow lon_{max} - lon_{min}$
  $\phi \leftarrow lat_{max} - lat_{min}$
  return $(clon, clat, \theta, \phi, \gamma)$
```

conducted extra experiments, tracking based on the unwarped image of tangent BFoV. As reported in the main paper, the new baseline AiATrack-360 achieves 0.534 S$_{dual}$ on 360VOT BBox. However, if we conduct a search based on tangent BFoV, it encounters obvious degradation and only achieves 0.449 S$_{dual}$.

More challenging qualitative results We supplement qualitative results on the sequences with different exclusive attributes of omnidirectional visual tracking in Figure 4 - 10. In the supplementary video, we also compare different representation ground truths and show video object tracking results in challenging scenarios which demonstrate there is big room for improvement.

More quantitative results. The success and precision plots of different trackers on 360VOT are shown in Figure 11. The complete performance results based on BBox for remaining attributes are demonstrated in Figure 12 - 15.

References

[1] Dawei Du, Yuankai Qi, Hongyang Yu, Yifan Yang, Kaiwen Duan, Guorong Li, Weigang Zhang, Qingming Huang, and Qi Tian. The unmanned aerial vehicle benchmark: Object detection and tracking. In *Proceedings of the European conference on computer vision (ECCV)*, pages 370–386, 2018. 2

[2] Matteo Dunnhofer, Antonino Furnari, Giovanni Maria Farinella, and Christian Micheloni. Is first person vision challenging for object tracking? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2698–2710, 2021. 2, 3

[3] Heng Fan, Liting Lin, Fan Yang, Peng Chu, Ge Deng, Sijia Yu, Hexin Bai, Yong Xu, Chunyuan Liao, and Haibin Ling. Lasot: A high-quality benchmark for large-scale single object tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5374–5383, 2019. 2

[4] Heng Fan, Halady Akhilesha Miththanthaya, Siranjiv Ramana Rajan, Xiaoqiong Liu, Zhilin Zou, Yuewei Lin, Haibin Ling, et al. Transparent object tracking benchmark. In

Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 10734–10743, 2021. 2

[5] Lianghua Huang, Xin Zhao, and Kaiqi Huang. 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. 2

[6] Hamed Kiani Galoogahi, Ashton Fagg, Chen Huang, Deva Ramanan, and Simon Lucey. Need for speed: A benchmark for higher frame rate object tracking. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1125–1134, 2017. 2

[7] Matej Kristan, Jiri Matas, Aleš Leonardis, Tomas Vojir, Roman Pflugfelder, Gustavo Fernandez, Georg Nebehay, Fatih Porikli, and Luka Čehovin. A novel performance evaluation methodology for single-target trackers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(11):2137–2155, Nov 2016. 2

[8] Annan Li, Min Lin, Yi Wu, Ming-Hsuan Yang, and Shuicheng Yan. Nus-pro: A new visual tracking challenge. *IEEE transactions on pattern analysis and machine intelligence*, 38(2):335–349, 2015. 2

[9] Siyi Li and Dit-Yan Yeung. Visual object tracking for unmanned aerial vehicles: A benchmark and new motion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017. 2

[10] Pengpeng Liang, Erik Blasch, and Haibin Ling. Encoding color information for visual tracking: Algorithms and benchmark. *IEEE transactions on image processing*, 24(12):5630–5644, 2015. 2

[11] Matthias Mueller, Neil Smith, and Bernard Ghanem. A benchmark and simulator for uav tracking. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pages 445–461. Springer, 2016. 2

[12] Matthias Muller, Adel Bibi, Silvio Giancola, Salman Alsubaihi, and Bernard Ghanem. Trackingnet: A large-scale dataset and benchmark for object tracking in the wild. In *Proceedings of the European conference on computer vision (ECCV)*, pages 300–317, 2018. 2

[13] Arnold WM Smeulders, Dung M Chu, Rita Cucchiara, Simone Calderara, Afshin Dehghan, and Mubarak Shah. Visual tracking: An experimental survey. *IEEE transactions on pattern analysis and machine intelligence*, 36(7):1442–1468, 2013. 2

[14] Konstantin Sofiiuk, Ilya A Petrov, and Anton Konushin. Reviving iterative training with mask guidance for interactive segmentation. In *2022 IEEE International Conference on Image Processing (ICIP)*, pages 3141–3145. IEEE, 2022. 3

[15] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *CVPR*, 2019. 3

[16] Jack Valmadre, Luca Bertinetto, Joao F Henriques, Ran Tao, Andrea Vedaldi, Arnold WM Smeulders, Philip HS Torr, and Efstratios Gavves. Long-term tracking in the wild: A benchmark. In *Proceedings of the European conference on computer vision (ECCV)*, pages 670–685, 2018. 2

[17] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. Deep high-resolution representation learning for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 43(10):3349–3364, 2020. 3

[18] Yi Wu, Jongwoo Lim, and Ming-Hsuan Yang. Object tracking benchmark. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(9):1834–1848, 2015. 2, 3

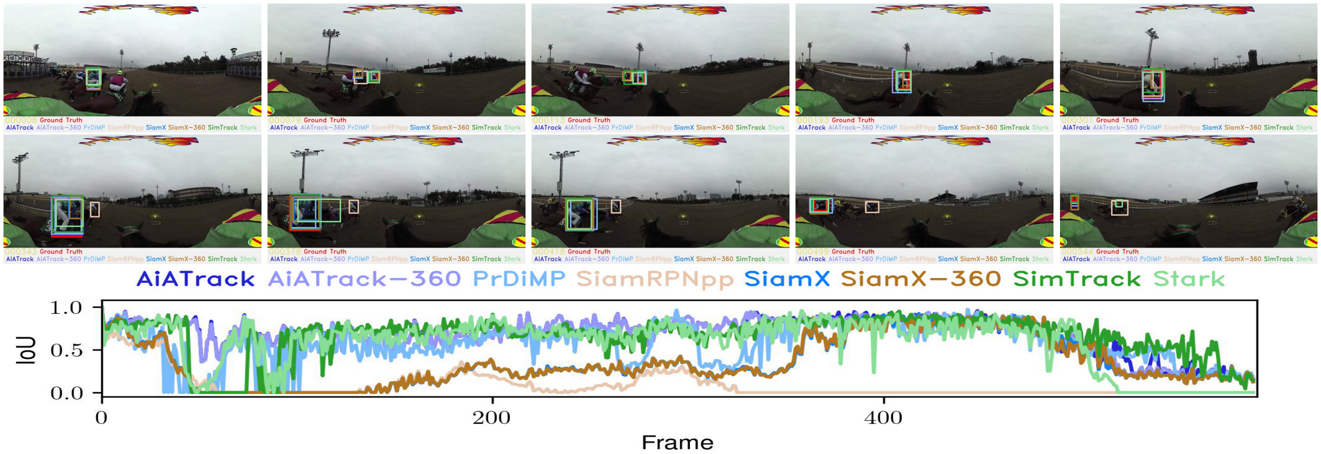


Figure 4: *People*. The qualitative comparison and the IOU plot of 8 top trackers on the scene with stitching artifacts (SA). The target is the rider in blue and green checkered clothes.

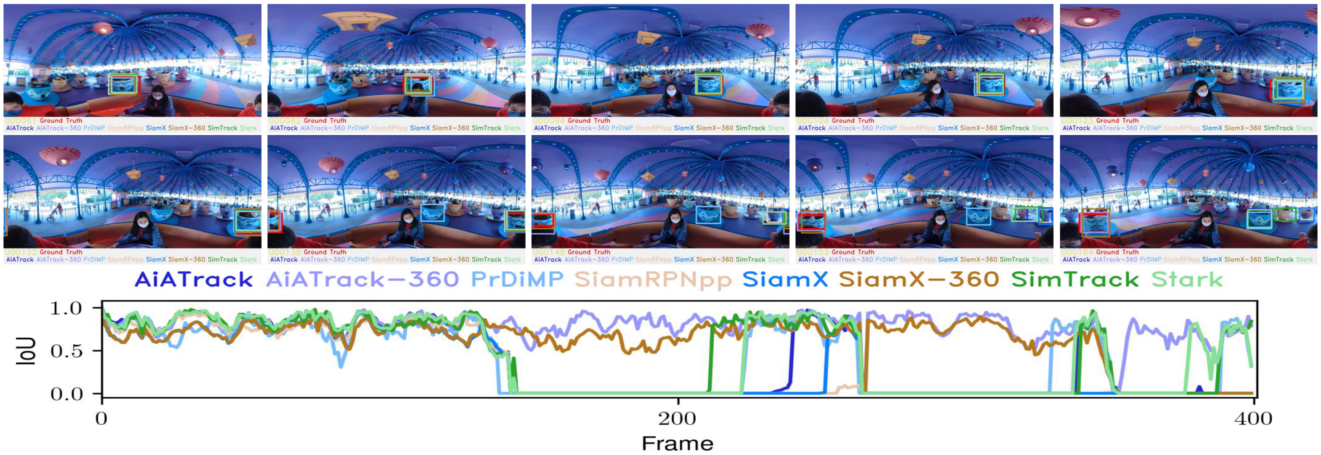


Figure 5: *Cup*. The qualitative comparison and the IoU plot of 8 top trackers on the scene where the target may cross border (CB). The target is the cup-shaped carrier.

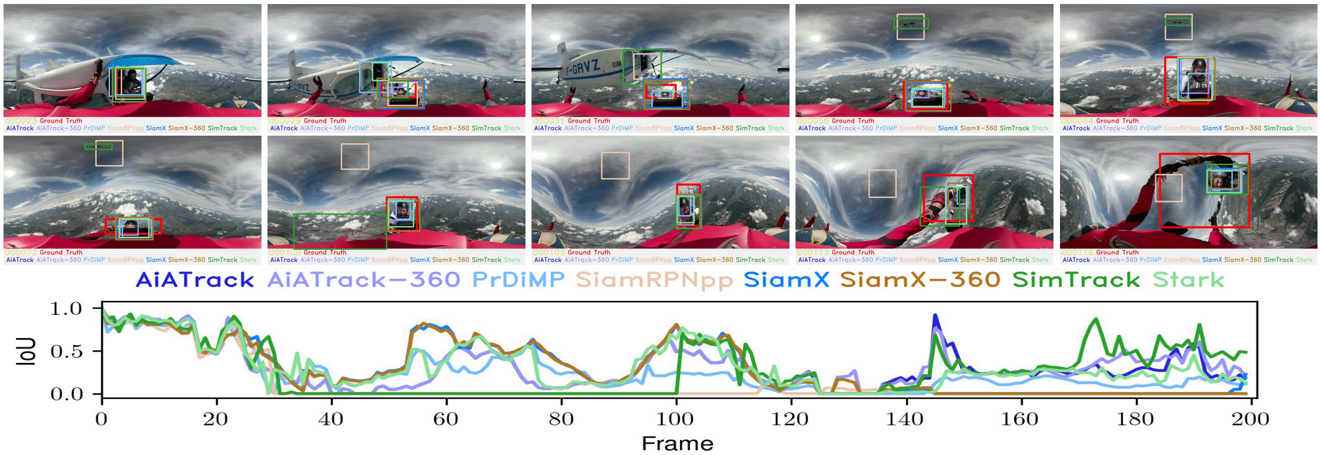


Figure 6: *Skydiver*. The qualitative comparison and the IoU plot of 8 top trackers on the scene where the target fast moves on the sphere (FMS). The target is the skydiver in a black and white suit.

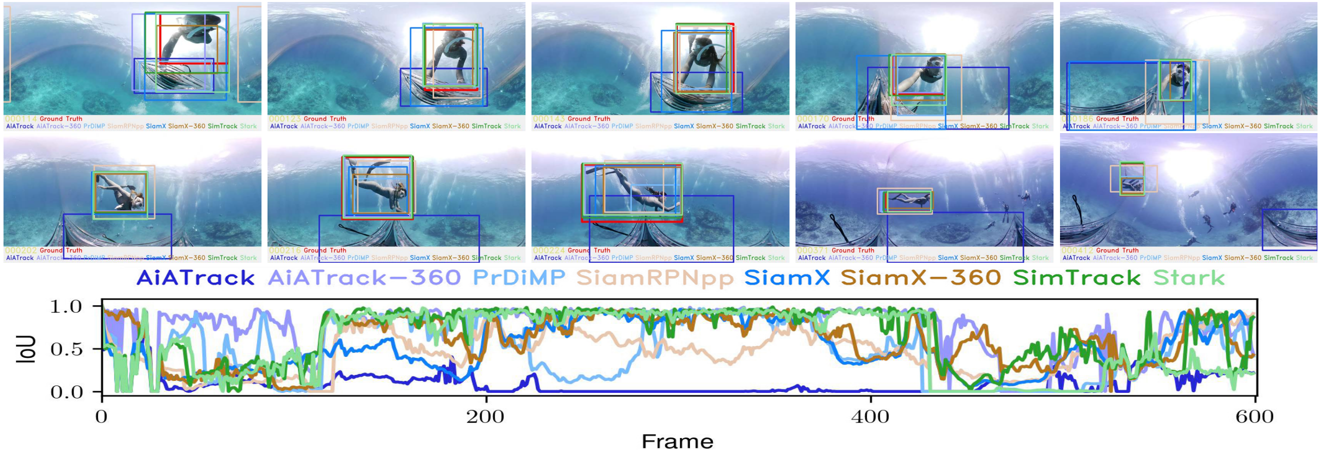


Figure 7: *Diver*. The qualitative comparison and the IoU plot of 8 top trackers on the scene where the target is of large field-of-view (LFoV). The target is the female diver in a gray diving suit.

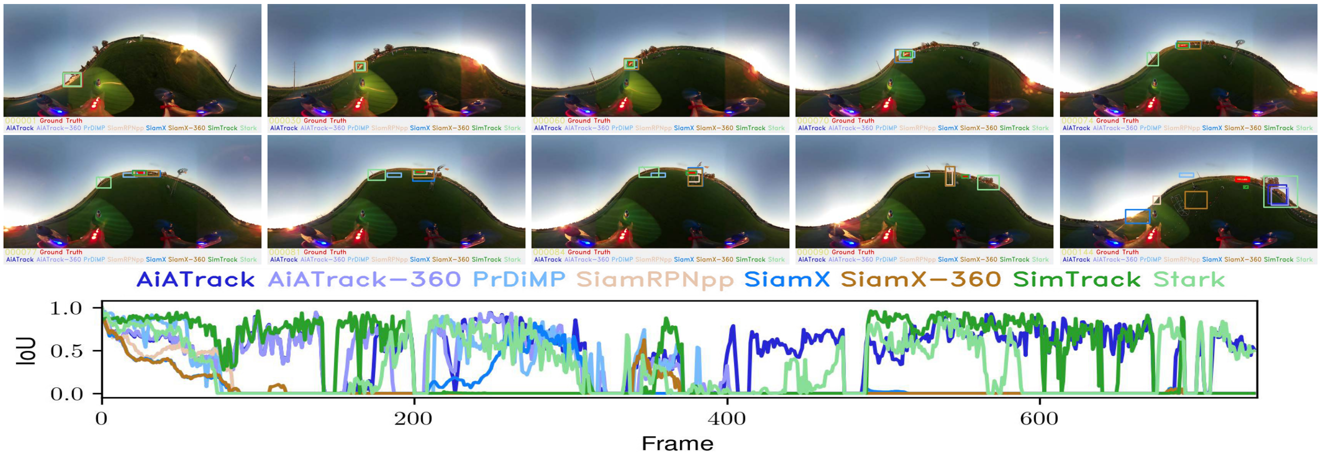


Figure 8: *Building*. The qualitative comparison and the IoU plot of 8 top trackers on the scene with latitude variation (LV). The target is the house with a red roof and white walls.

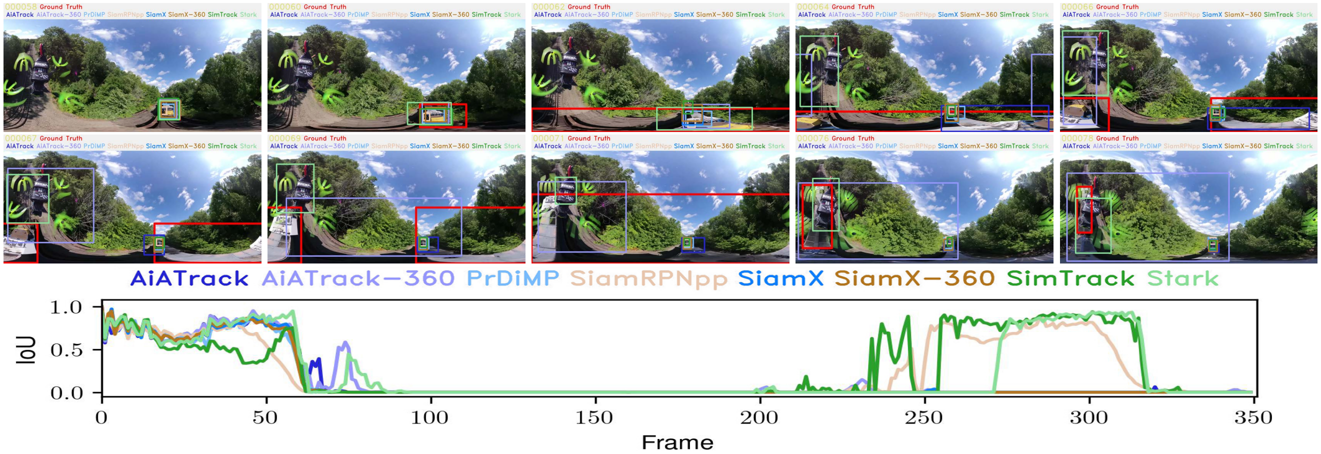


Figure 9: *Train*. The qualitative comparison and the IoU plot of 8 top trackers on the scene where the target is on a high latitude (HL). Best viewed in color and zoom-in. The target is the first carriage of the train.

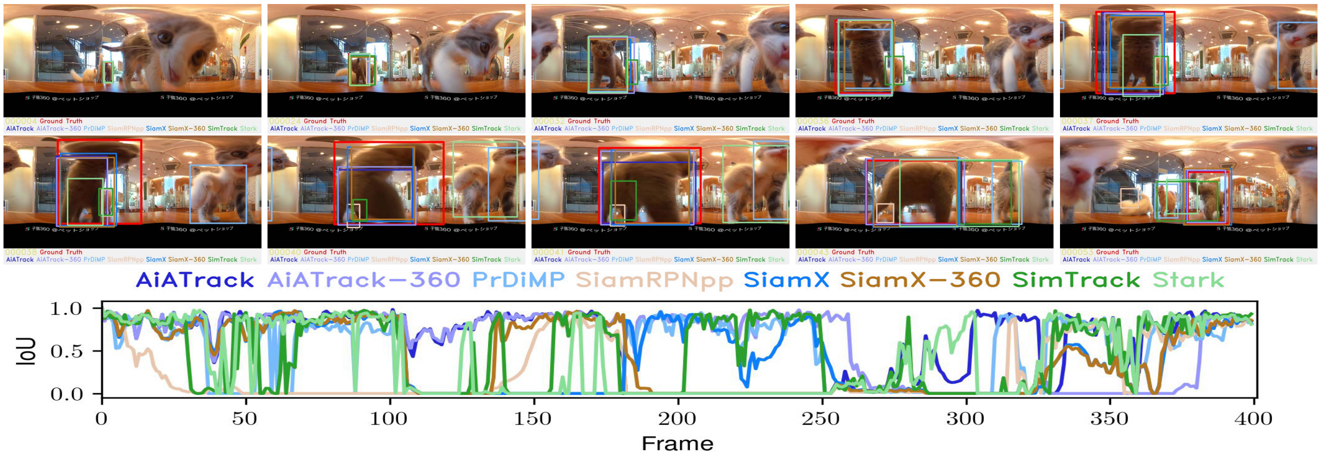


Figure 10: *Kitty*. The qualitative comparison and the IoU plot of 8 top trackers on the scene where the target has large distortion (LD). The target is the grey kitten.

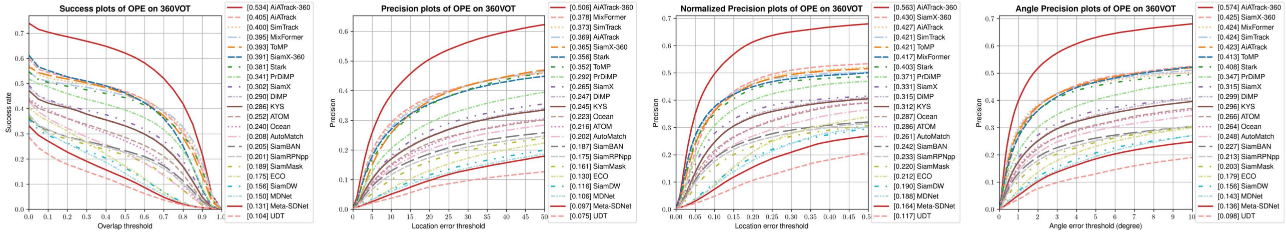


Figure 11: The success and precision plots of different trackers on 360VOT in terms of BBox predictions.

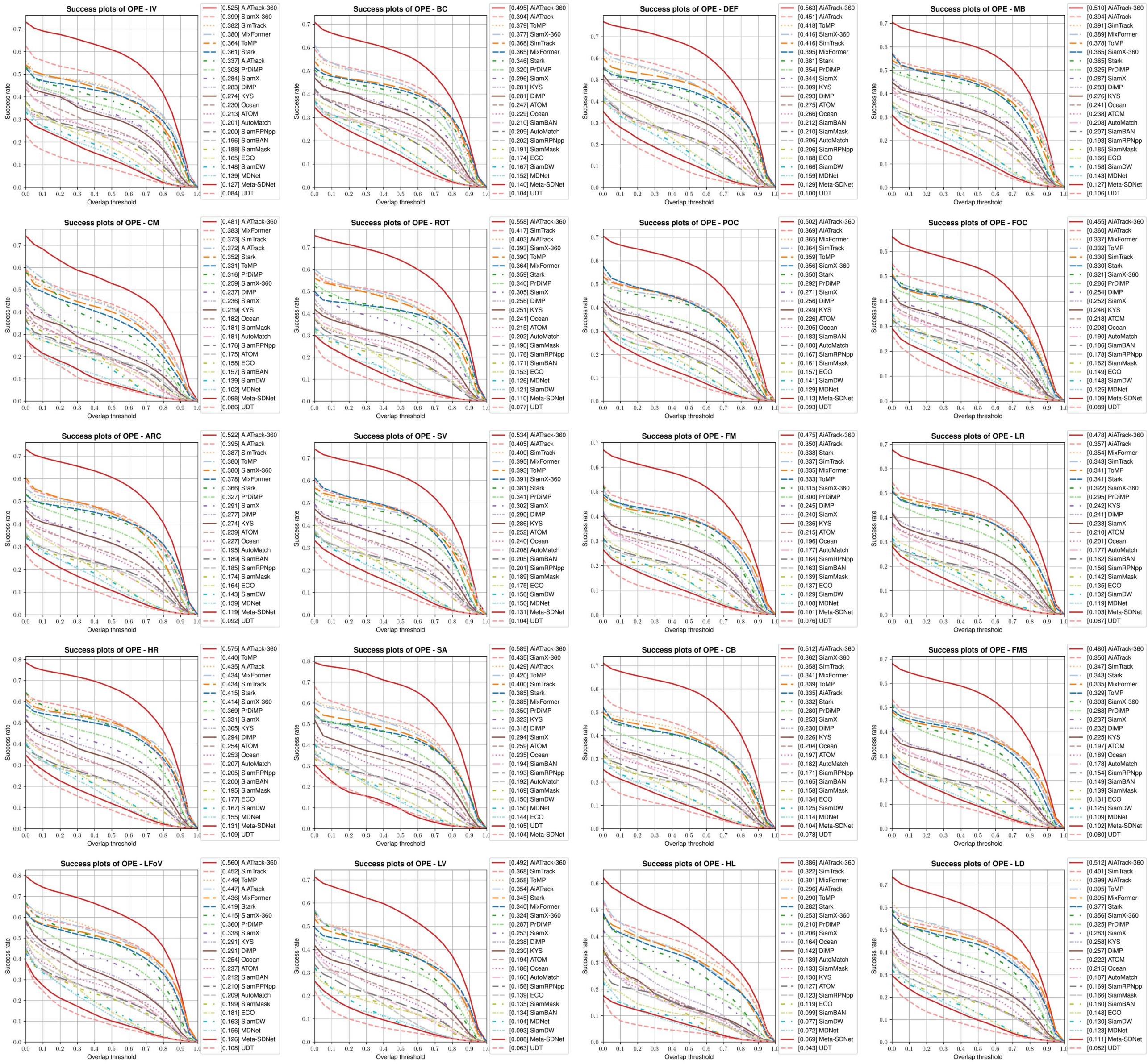


Figure 12: The performance of trackers on each attribute using the BBox dual success metric.

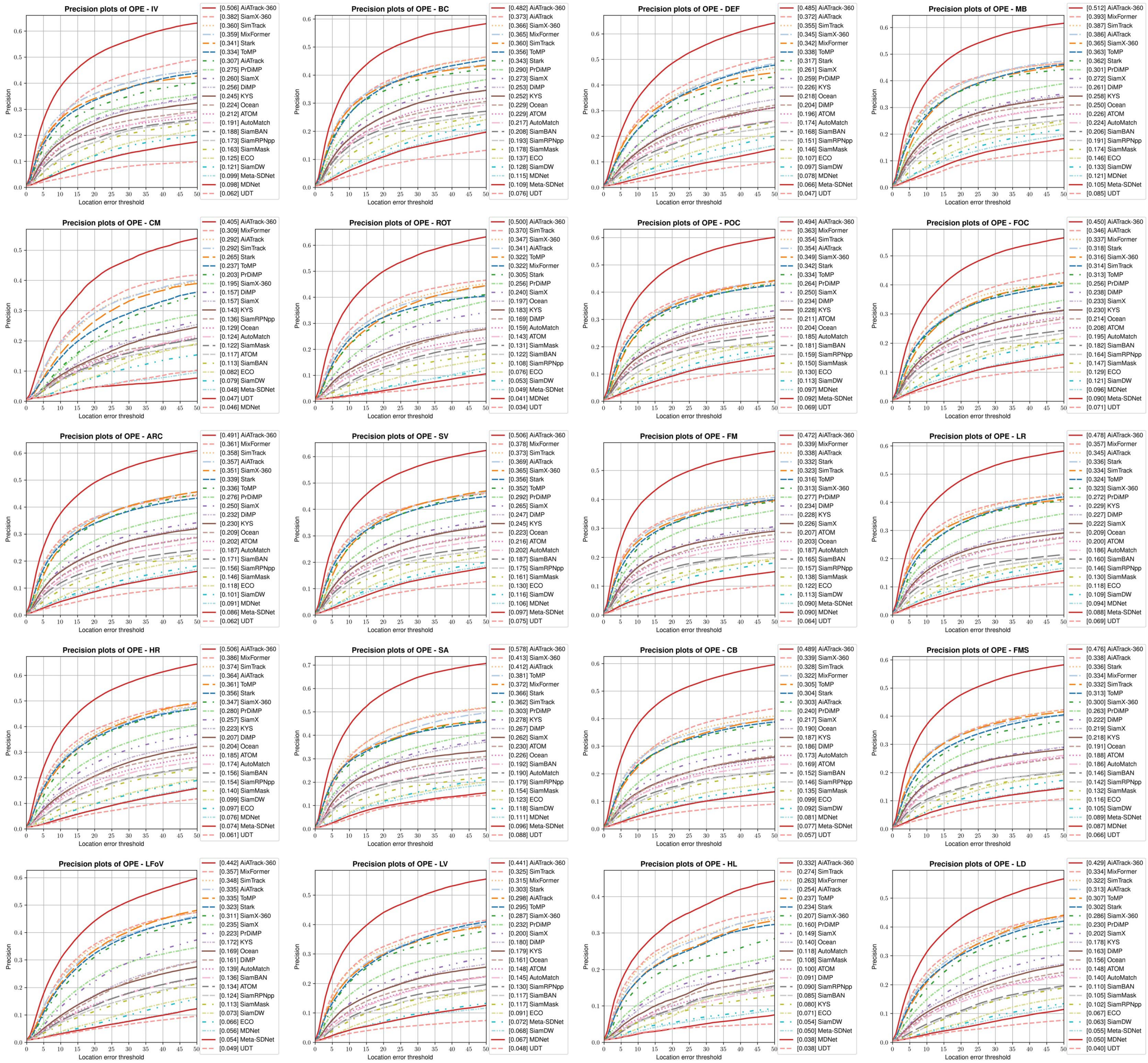


Figure 13: The performance of trackers on each attribute using the BBox dual precision metric.

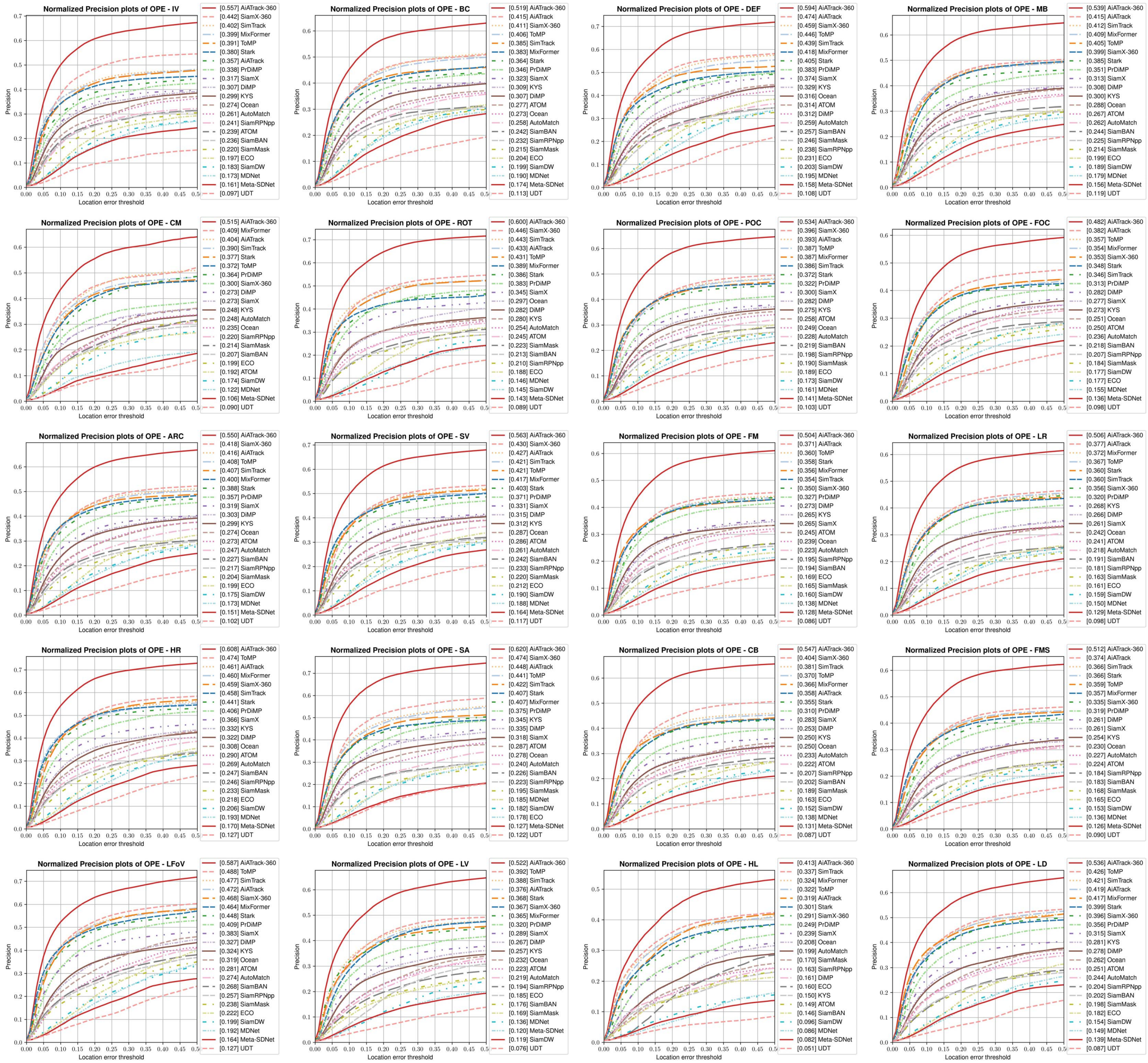


Figure 14: The performance of trackers on each attribute using the BBox normalized dual precision metric.

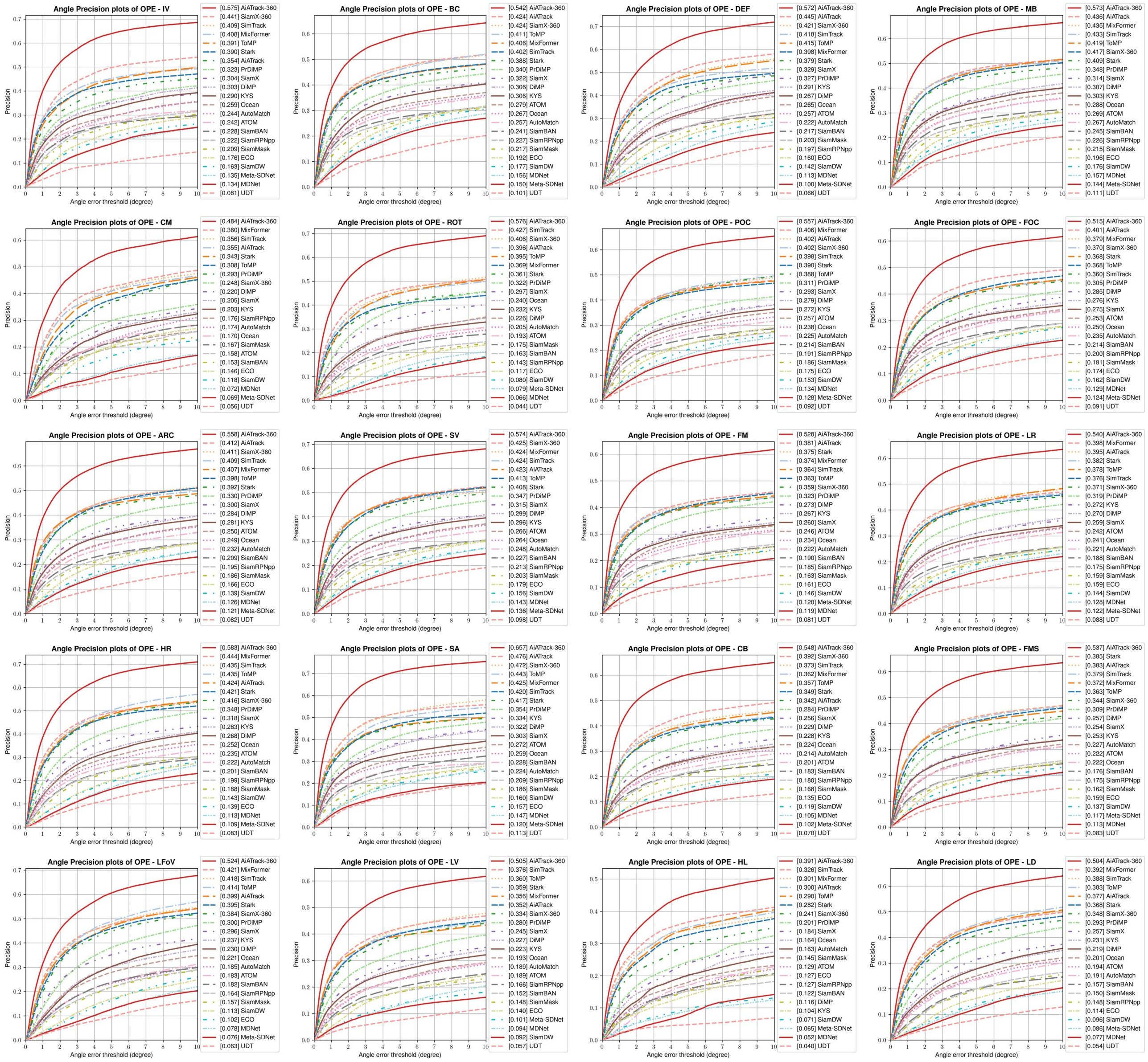


Figure 15: The performance of trackers on each attribute using the BBox angle precision metric.