FISHEYE MULTIPLE OBJECT TRACKING BY LEARNING DISTORTIONS WITHOUT DEWARPING

Ping-Yang Chen[1], *Jun-Wei Hsieh*[2,3], *Ming-Ching Chang*[4],
Munkhjargal Gochoo[5,6], *Fang-Pang Lin*[7], *and Yong-Sheng Chen*[1]

[1]Department of Computer Science and [2]College of AI and Green Energy, National Yang Ming Chiao Tung University, Taiwan
[3]Department of Computer Science and Engineering, National Taiwan Ocean University, Taiwan
[4]Department of Computer Science, University at Albany, State University of New York, USA
[5] College of Information Technology and [6]Emirates Center for Mobility Research, United Arab Emirates University, UAE
[7] National Center for High-Performance Computing, Taiwan

ABSTRACT

We develop a new Multiple Object Tracking (MOT) scheme for fisheye cameras that can directly perform vehicle detection, re-identification, and tracking under fisheye distortions without explicit dewarping. Fisheye cameras provide omnidirectional coverage that is wider than traditional cameras, reducing fewer need of cameras to monitor road intersections. However, the problem of distorted views introduces new challenges for fisheye MOT. In this paper, we propose a Fish-Eye Multiple Object Tracking (FEMOT) approach with two novelties. We develop the Distorted Fisheye Image Augmentation (DFIA) method to improve object detection and re-identification on fisheye cameras, where fisheye model training can be performed on existing datasets of traditional cameras via fisheye data synthesis and augmentation. We also develop the Hybrid Data Association (HDA) method to perform tracking directly on fisheye views, without the need of dewarping. The developed FEMOT framework provides practical design and advancement that enables large-scale use of fisheye cameras in smart city and surveillance applications.

***Index Terms*—** Fisheye distortion, multiple object tracking, object detection, re-identification, data augmentation.

1. INTRODUCTION

Multiple Objects Tracking (MOT) methods have improved greatly with the advancement of object detection and re-identification in deep learning, following the *tracking-by-detection* paradigm. MOT is widely used in surveillance, traffic control, autonomous driving, and many other smart city applications. Camera coverage plays an important role in video analytics, as computer vision (CV) algorithms run solely on visible views of the scene. Fisheye cameras provide a wide field of view and thus are more cost-effective in view coverage, when compared to traditional CCTV cameras. Employing fisheye cameras for traffic monitoring can greatly reduce the number of needed cameras. However, the development of CV algorithms for fisheye cameras is more sophisticated and limited in the literature, as object appearance and movement are no longer linear under distortion.

Fig. 1. Our developed framework provides a concrete step forward for the real-world use of fisheye MOT in surveillance and smart city applications.

We provide two critical contributions in this work toward a practical use of fisheye cameras for MOT traffic analysis (Fig. 1): (1) effective *fisheye object detection and re-identification* that can run directly on the highly distorted fisheye views, without the need of view dewarping, (2) effective *fisheye tracker* that can overcome the nonlinear physical modeling in target/tracklet association on the distorted fisheye views. Our method is superior over most existing MOT algorithms on fisheye cameras, as they rely on the constant velocity assumption of Kalman filtering and thus require fisheye dewarping, which is less effective and prune to error.

Advancement of deep object detection methods [3, 4, 5, 6, 7, 1, 3] have made it feasible to detect and recognize small objects in a distorted camera view. FairMOT [8] can simultaneously detect and track multiple vehicles using appearance features. However, it requires strong supervision of both the bounding box and identity annotations. We point out that the Separate Detection and Embedding (SDE) method is more suitable for fisheye MOT, where only bounding-box annotations are required for detector training. The SORT-based tracking methods [9, 10, 11] are simple and effective in tracking target movements. However, they are not directly applica-

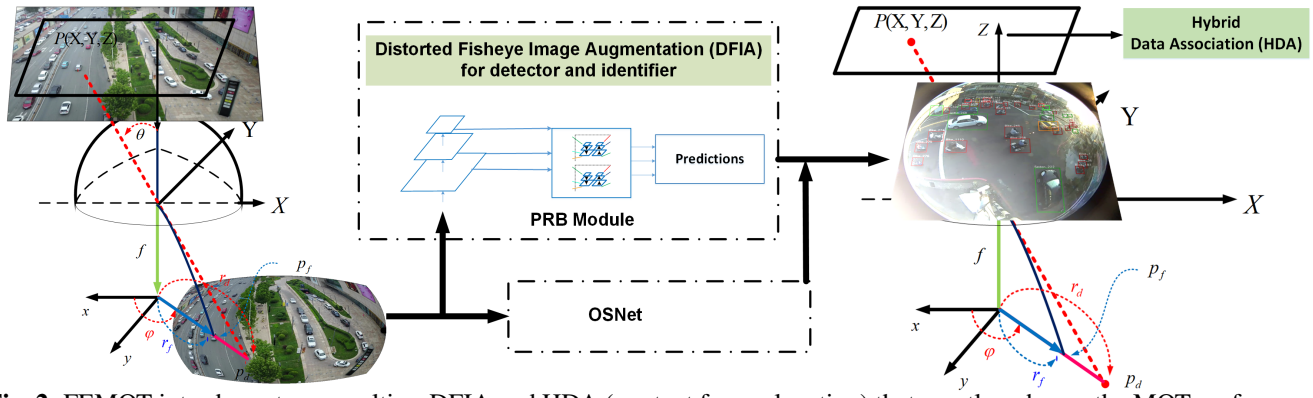

Fig. 2. FEMOT introduces two novelties, DFIA and HDA (see text for explanation) that greatly enhance the MOT performance without the need of explicit fisheye view dewarping. Our implementation of FEMOT combines the strengths of (1) PRBNet [1], which excels at localizing both large and small vehicles in the fisheye view, and (2) OSNet [2] for object re-identification.

ble for fisheye cases, as they rely on the linear physical modeling of target movement on the ground plane. Thus, a *dewarping* step which corrects the distorted fisheye views must be performed beforehand; however, this dewarping operation requires calibration and can potentially introduce errors, new distortions, and increase computational cost.

To address the aforementioned drawbacks in applying the mainstream MOT methods to fisheye cameras, we developed a new **FishEye Multiple Object Tracker (FEMOT)** that can directly learn to address fisheye distortions while performing vehicle detection, re-identification, and tracking without explicit dewarping. FEMOT features a hybrid data association method and a fisheye-specific image augmentation design, which effectively learns the distortion mapping to improve MOT without dewarping. Fig. 2 provides an overview.

The first novelty of FEMOT is the **Distorted Fisheye Image Augmentation (DFIA)** that generates augmented training data from traditional camera views to simulate fisheye perspective distortions. This straightforward design can effectively overcome the performance degradation of vehicle detection and re-identification caused by fisheye distortions. DFIA improves the fisheye detection and re-identification, without the need of new annotation or ground-truthing efforts. It does not rely on any pre-training of models. DIFA can effectively make use of existing MOT datasets made of perspective cameras, and transfer the use to fisheye cameras.

The second novelty of FEMOT is the **Hybrid Data Association (HDA)**, which is a *twin model* that calculates the IoU similarity and predicts vehicle movements under fisheye distortion. The IoU similarity is calculated based on features learned from both fisheye and distorted perspective views. HDA also predicts vehicle movements through a fisheye and distorted perspective Kalman filtering [12], which gracefully resolves the nonlinear modeling of constant movement in fisheye views. The HDA can thus accurately predict the target location directly based its fisheye trajectory for MOT.

We believe that the proposed FEMOT represents a new state-of-the-art for fisheye video analytics. It enables large-scale deployment of fisheye cameras, which takes advantage of the wide-angle fisheye views to improve surveillance, traffic monitoring, and smart city applications.

2. RELATED WORK

Tracking by Detection: Typical Joint Detection and Embedding (JDE) models such as the FairMOT [8] combine the detector and the embedded model in a single shot deep network. FairMOT can simultaneously locate, recognize, and identify vehicles based on the detection results, where the corresponding appearance embeddings are calculated only once. The Separate Detection and Embedding model (SDE) is more suitable for fisheye MOT, as only bounding-box annotations are required for detector training. Representative tracking methods in this category include Kalman Filtering (KF) [12] and SORT [13]. KF is a linear model that predicts the target position from a series of observed measurements over time. SORT [13] is a combination of the KF and a Hungarian algorithm to track multiple objects in real time. SORT-based MOT methods such as [9, 10, 11] can accurately track objects in a general camera setting. However, they do not work well for fisheye cameras due to the large view distortions.

Tracking from Fisheye Cameras: Several works use a single top-view camera for object detection [14, 15] and tracking [16]. CFPN [3] is the first automatic traffic flow estimation system that can run in real-time to detect small objects from fisheye cameras with significant distortions. For fisheye views with large distortion, the State-of-The-Art (SoTA) MOT methods mostly focus on the detector design and performance improvement; they did not address view distortion in the tracker design. To our best knowledge, the proposed FEMOT is the first elegant solution that leverages end-to-end deep learning that extends SoTA methods for fisheye MOT.

3. METHOD

3.1. Fisheye Camera Model

Refer to Fig. 2 on an illustration of the fisheye camera model. Denote the projection of a 3D point $P = (X, Y, Z)^t$ on the

2D undistorted perspective image as $p_d = (x_d, y_d)^t$ and such a 2D projection point on the fisheye image as $p_f = (x_f, y_f)^t$, respectively. Denote the angle between the light ray and the Z-axis as θ, focal length as f, and the distance between p_d and the Z-axis as r_d. The perspective relation between r_d and θ is $r_d = f \cdot \tan(\theta)$. For small θ, we assume that the length r_f between p_f and the Z-axis is approximately $f \cdot \theta$. Denote the center and radius of a circle on the fisheye image as (c_x, c_y) and R, respectively. The r_f is calculated as $r_f = \sqrt{(x_f - c_x)^2 + (y_f - c_y)^2}$. Next, r_d is calculated as: $r_d = R \cdot \tan\left(\frac{r_f}{R}\right)$, where $f = R$. Denote the angle between p_f and the x axis as φ. We have:

$$\varphi = \arctan\left(\frac{y_f - c_y}{x_f - c_x}\right). \tag{1}$$

Next, x_d and y_d can be obtained by $x_d = c_x + r_d \cdot \cos\varphi$ and $y_d = c_y + r_d \cdot \sin\varphi$. Finally, x_d and y_d is calculated as:

$$\begin{aligned} x_d &= c_x + R \cdot \tan\left(\frac{r_f}{f}\right) \cos\left(\arctan\left(\frac{y_f - c_y}{x_f - c_x}\right)\right), \\ y_d &= c_y + R \cdot \tan\left(\frac{r_f}{f}\right) \sin\left(\arctan\left(\frac{y_f - c_y}{x_f - c_x}\right)\right). \end{aligned} \tag{2}$$

3.2. Distorted Fisheye Image Augmentation (DFIA)

The key challenge of building a fisheye object detector and identifier is on how best to effectively deal with distorted views for training a data-driven model end-to-end, given the fact that most object detection and MOT datasets are created using perspective cameras. To this end, we develop DFIA as an effective data augmentation method to transform normal image samples into fisheye samples, as in Fig. 3. Let $I = (A_i, L_i)$ denote the i-th original training sample for MOT, where $A_i \in \mathbb{R}^{W \times H \times C}$ with label L_i. DFIA generates the augmented fisheye image $AF_i \in \mathbb{R}^{W \times H \times C}$ with transformed target labels LF_i, using the equations from § 3.1. DFIA transforms the original perspective sample I into an augmented, distorted fisheye sample $F = (AF_i, LF_i)$, *i.e.*, $F = DFIA(I, \varphi)$, where the parameter φ from Eq. (1) controls the degree of distortion. This way, the newly-generated fisheye samples can be used directly for the training of object detector and re-identifiers for fisheye views.

DCFA is different from the MixUp and Blurring methods [17, 18], which are originally designed for classification rather than object detection. The MixUp operation blurs the *whole* image and results in unclear object boundaries. Such discontinuity of the object boundaries causes problems in localizing the object bounding boxes. In contrast, the DFIA is designed to facilitate the training of the fisheye object detector and identifier directly on the transformed and augmented samples, without the need of explicit view dewarping or model pre-training. Furthermore, we note that the view dewarping MOT approach could be unnecessarily complicated and not extensible to run on different fisheye cameras.

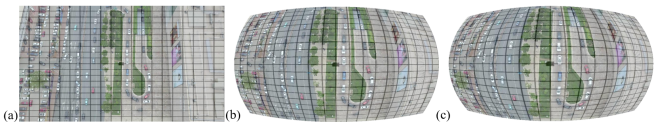

Fig. 3. Examples of fisheye augmentation samples generated using DFIA: (a) the original perspective image, and the generated samples with (b) $\varphi = 0.5$ and (c) $\varphi = 0.9$ in Eq. (1).

3.3. Hybrid Data Association (HDA)

The aim here is to enable effective learning to perform accurate and robust target tracking directly from the distorted fisheye views. Let B_f^V denote the predicted box for a target (vehicle) V with a fish-eye camera. Let (x_f^L, y_f^T) and (x_f^R, y_f^B) denote the positions of the upper-left and bottom-right box of V, B_f^V=$(x_f^L, y_f^T, x_f^R, y_f^B)$. B_f^V can be converted to its new position B_d^V on the distorted perspective image using Eq.(2). HDA creates a hybrid bounded box B^V to represent V both on the fisheye image and the distorted perspective image, $B^V = (B_f^V, B_d^V)$. The movement state S^V of target V is modeled as:

$$S^V = (B_f^V, B_d^V, \dot{B}_f^V, \dot{B}_d^V). \tag{3}$$

Kalman filtering [12] is then adopted on S^V to solve the trajectory prediction problem for target tracking directly on fisheye views. This way, the inter-frame displacement of each target can be effectively predicted via KF. In the case when there is no detection to associate with a target, its positions on the fish eye image and the distorted perspective one are simply predicted using linear velocity terms $\dot{B}_f^V$ and $\dot{B}_d^V$.

4. EXPERIMENTS AND ANALYSIS

We conducted extensive experiments to investigate the effectiveness of the proposed FEMOT as well as ablation studies on the DFIA and HDA modules on multiple MOT datasets.

The proposed FEMOT pipeline is flexible and not limited to use with specific detectors, identifiers, or trackers. Our experiment involves the use of YOLOX [7] detector and multiple identifiers [2, 10] and trackers [19, 13, 10, 11].

4.1. Implementation Details

FisheyeMOT: We collect a new fisheye MOT traffic dataset consisting of 9,288 frames with different splits to evaluate the vehicle detection and MOT tasks. It includes 228.8K bounding boxes and 102.5K identities for eight vehicle classes, namely, *Scooter, Van, Sedan, Truck, Pedestrian, SUV, and Taxi*. Samples are randomly split into Train, Val, and Test sets with a ratio about 50:30:20.

We also select images from the VisDrone Dataset [20] and use DFIA to generate samples for evaluation.

Detector: PRBNet [1] is chosen as our detector in DFIA due to its strength in localizing both large and small objects. PRBNet performs well on detecting small objects in wide

Table 1. Comparisons of FEMOT against SoTA MOT methods on the FisheyeMOT dataset.

Method	HOTA↑	IDF↑	MOTA↑	AssA↑	DetA↑	IDs↓
SORT [13]	22.1	24.1	27.9	20.1	23.3	48,201
FairMOT [8]	37.2	45.8	46.2	32.7	38.7	32,597
ByteTrack [19]	40.8	49.2	50.4	39.6	40.1	25,691
DeepSORT [13]	38.1	47.5	48.8	37.9	40.0	26,984
StrongSORT [11]	40.3	49.8	49.8	40.5	40.3	25,999
BoT-SORT-R [10]	41.2	52.1	50.0	41.2	41.5	19,566
FEMOT	**48.9**	**59.2**	**59.1**	**48.1**	**47.3**	**15,892**

* Higher Order Tracking Accuracy (HOTA), ID F1 (IDF), Multiple Object Tracking Accuracy (MOTA), Association Accuracy (AssA), Detection Accuracy (DetA), ID switch (IDs).

filed of view, which is particularly useful, as many objects appear very small in the fisheye view.

Tracker: We adopted the BoT-SORT-R [10] as our tracker, because it represents the SoTA of MOT and is flexible to run with other detectors and identifiers.

Identifier: OSNet [2] is chosen for its capability for real-time object re-identification.

4.2. Results

We quantitatively evaluate the FEMOT against five SoTA trackers [13, 8, 19, 11, 10] on the FisheyeMOT dataset. Table 1 shows the results using standard MOT evaluation metrics. We highlight two advantages of FEMOT: (1) the DFIA design enables effective vehicle detection from distorted views as well as effective vehicle identification from distorted appearances in the fisheye views. (2) the HDA design for MOT can robustly model the fisheye trajectory prediction effectively with high accuracy. Qualitatively, even if the model does not detect the vehicle in some frames, FEMOT can still re-identify many vehicles, which enables robust traffic flow estimation. Fig. 1 shows a snapshot of FEMOT running on a fisheye camera for real-time traffic flow estimation.

4.3. Ablation studies

We performed ablation studies on the FisheyeMOT dataset to verify the effectiveness of the DFIA and HDA modules and the reliability of FEMOT, with various implementation settings. We analyze the following cases: (1) training a detector with DFIA, (2) training an identifier with DFIA, and (3) analyzing the effectiveness of HDA in building a fisheye MOT system for traffic analysis.

Ablation study of DFIA: We evaluate the effectiveness of DFIA in tackling the perspective distortion issues of the detector and identifier, by bench-marking the mainstream detectors [4, 5, 7, 21, 1] on the FisheyeMOT dataset. We found that the obtained FEMOT scores on fisheye cameras are similar to those scores reported in the standard MOT datasets [22, 23, 20] on normal perspective cameras.

Table 2. Ablation study of DFIA on the FisheyeMOT dataset against two other MOT methods [11, 10].

Method	DET	Re-ID	HOTA
StrongSORT [11] as baseline			40.3
FEMOT [Ours]	✓		42.0
FEMOT [Ours]		✓	44.6
FEMOT [Ours]	✓	✓	**45.0**
BoT-SORT-R [10] as baseline			41.2
FEMOT [Ours]	✓		42.1
FEMOT [Ours]		✓	44.9
FEMOT [Ours]	✓	✓	**46.1**

Table 3. Ablation study of HDA and DFIA on the FisheyeMOT dataset against two other MOT methods [11, 10].

Method	HDA	DFIA	HOTA
StrongSORT [11] as baseline			40.3
FEMOT [Ours]	✓		44.1
FEMOT [Ours]		✓	45.0
FEMOT [Ours]	✓	✓	**48.3**
BoT-SORT-R [10] as baseline			41.2
FEMOT [Ours]	✓		45.6
FEMOT [Ours]		✓	46.1
FEMOT [Ours]	✓	✓	**48.9**

Table 2 compares different data augmentation settings in the detector and the identifier w/wo DFIA data augmentation. Observe that the use of DFIA in re-id results in large improvement of the two trackers [11, 10]. Our explanation is that the deep convolution neural detectors have the ability to handle the fisheye distortion problems, and the key for enabling that is on how best to detect and re-identify targets from highly distorted views. These results show that DFIA is effective in generating distorted augmentation data to perform the required training, without the need for dewarping, nor using additional labeling or model pre-training.

Ablation study of HDA: Table 3 compares the ablation study of FEMOT w/wo HDA or DFIA. Observe that FEMOT with both DFIA and HDA results in the best scores, when compared with other cases.

In summary, the ablation study results show that each of the DFIA and HDA design in FEMOT has its own contribution toward the MOT performance improvement.

5. CONCLUSION

We believe that the proposed FEMOT represents a pioneering effort in pushing the frontiers of deep learning in the fisheye MOT applications. We show that each of the two design of FEMOT has its own contribution. DFIA can effectively generate fisheye training samples from the existing datasets of normal cameras. HDA enables direct tracklet association and tracking in heavily distorted views without dewarping. FEMOT enables large-scale use of fisheye cameras in Intelligent Transportation System (ITS) and smart city applications.

6. REFERENCES

[1] Ping-Yang Chen, Ming-Ching Chang, Jun-Wei Hsieh, and Yong-Sheng Chen, "Parallel residual bi-fusion feature pyramid network for accurate single-shot object detection," *TIP*, vol. 30, pp. 9099–9111, 2021. 1, 2, 3, 4

[2] Kaiyang Zhou, Yongxin Yang, Andrea Cavallaro, and Tao Xiang, "Omni-scale feature learning for person re-identification," in *ICCV*, October 2019. 2, 3, 4

[3] Ping-Yang Chen, Jun-Wei Hsieh, Munkhjargal Gochoo, Chien-Yao Wang, and Hong-Yuan Mark Liao, "Smaller object detection for real-time embedded traffic flow estimation using fish-eye cameras," in *ICIP*, 2019. 1, 2

[4] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao, "YOLOv4: Optimal speed and accuracy of object detection," in *arXiv*, 2020. 1, 4

[5] Glenn Jocher et al., "ultralytics/yolov5: v3.1," Oct. 2020. 1, 4

[6] Xiaozhi Jiang, Qiang Wang, and Zhiqiang Chen, "PP-YOLO: An effective and efficient implementation of object detector," *arXiv*, 2020. 1

[7] Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun, "YOLOX: Exceeding yolo series in 2021," *arXiv*, 2021. 1, 3, 4

[8] Yifu Zhang, Chunyu Wang, Xinggang Wang, Wenjun Zeng, and Wenyu Liu, "Fairmot: On the fairness of detection and re-identification in multiple object tracking," *IJCV*, vol. 129, pp. 3069–3087, 2021. 1, 2, 4

[9] Jinkun Cao, Xinshuo Weng, Rawal Khirodkar, Jiangmiao Pang, and Kris Kitani, "Observation-Centric SORT: Rethinking sort for robust multi-object tracking," *arXiv*, 2022. 1, 2

[10] Nir Aharon, Roy Orfaig, and Ben-Zion Bobrovsky, "BoT-SORT: Robust associations multi-pedestrian tracking," *arXiv*, 2022. 1, 2, 3, 4

[11] Yunhao Du, Yang Song, Bo Yang, and Yanyun Zhao, "StrongSORT: Make DeepSORT great again," *arXiv*, 2022. 1, 2, 3, 4

[12] Tamer Basar and Tamer Başar, *Control theory: twenty-five seminal papers, 1931-1981*, IEEE Press New York, 2001. 2, 3

[13] N. Wojke, A. Bewley, and D. Paulus, "Simple online and realtime tracking with a deep association metric," in *ICIP*, 2017. 2, 3, 4

[14] Han Wang, Dubok Park, David K. Han, and Hanseok Ko, "Top-view people detection based on multiple subarea pose models for smart home system," in *ICCE*, 2016. 2

[15] Xue YUAN, Xue-Ye WEI, and Yong-Duan SONG, "Pedestrian detection for counting applications using a top-view camera," *IEICE Transactions on Information and Systems*, vol. E94.D, no. 6, pp. 1269–1277, 2011. 2

[16] Mamoru Saito, Katsuhisa Kitaguchi, Gun Kimura, and Masafumi Hashimoto, "People detection and tracking from fish-eye image based on probabilistic appearance model," in *SICE Annual Conference*, 2011. 2

[17] Hongyi Zhang et al., "Mixup: Beyond empirical risk minimization," in *ICLR*, 2018. 3

[18] Y. Sangdoo et al., "CutMix: Regularization strategy to train strong classifiers with localizable features," in *ICCV*, 2019. 3

[19] Yifu Zhang, Peize Sun, Yi Jiang, Dongdong Yu, Fucheng Weng, Zehuan Yuan, Ping Luo, Wenyu Liu, and Xinggang Wang, "ByteTrack: Multi-object tracking by associating every detection box," in *ECCV*, 2022. 3, 4

[20] Guanlin Chen, Wenguan Wang, Zhijian He, Lujia Wang, Yixuan Yuan, Dingwen Zhang, Jinglin Zhang, Pengfei Zhu, Luc Van Gool, Junwei Han, Steven Hoi, Qinghua Hu, and Ming Liu, "VisDrone-MOT2021: The vision meets drone multiple object tracking challenge results," in *ICCV Workshops*, 2021. 3, 4

[21] Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao, "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," 2022. 4

[22] Longyin Wen, Dawei Du, Zhaowei Cai, Zhen Lei, Ming-Ching Chang, and Honggang Qi, "Ua-detrac: A new benchmark and protocol for multi-object detection and tracking," *Comput Vis Image Underst*, vol. 198, pp. 103001, 2020. 4

[23] 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 *ECCV*, 2018. 4