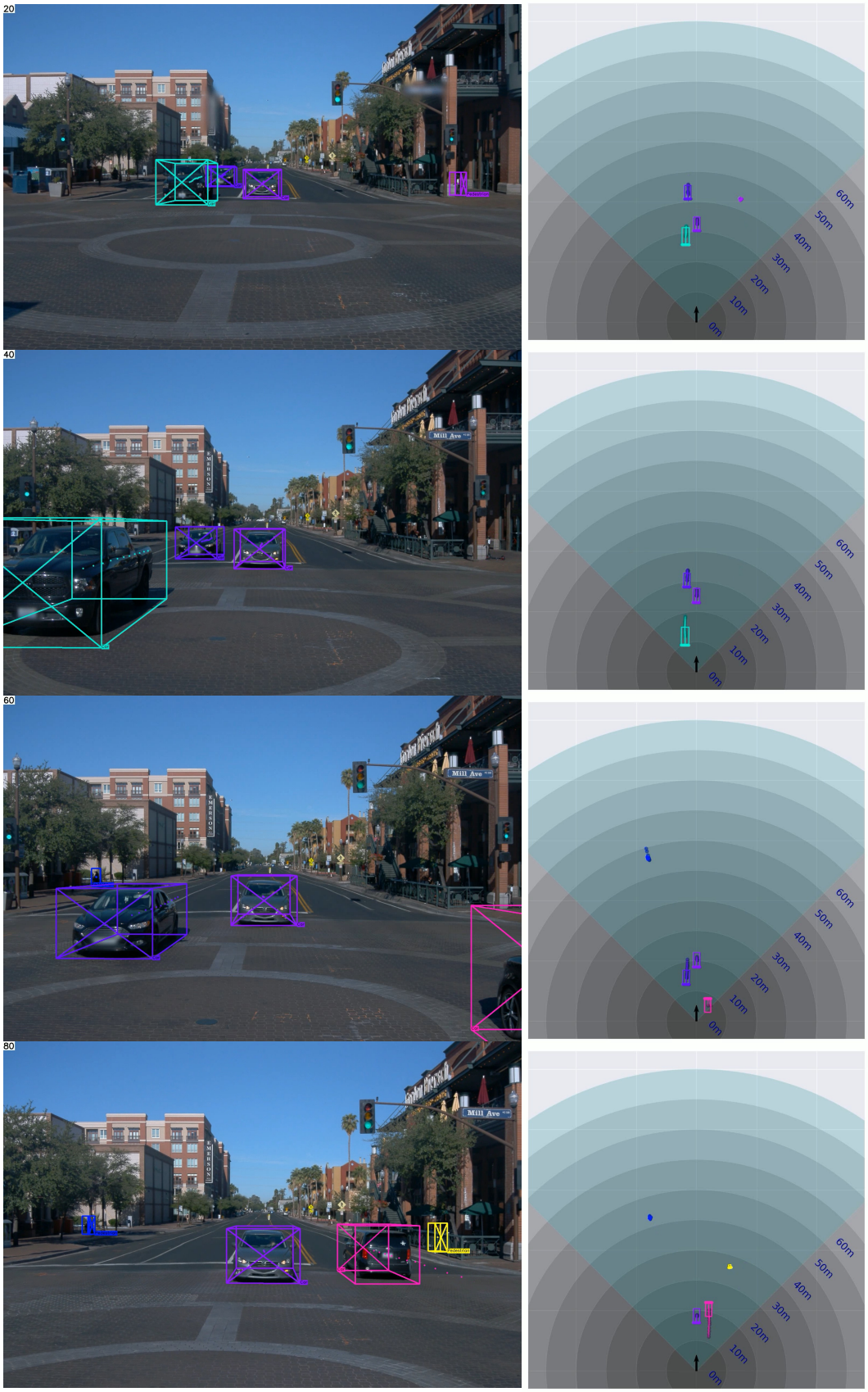

Fig. 14: **Qualitative results on the testing set of Waymo Open dataset.** Our QD-3DT accurately tracks all observed objects and locates them in 3D. We show predicted 3D bounding boxes and trajectories colored with tracking IDs. Better visualization with color.

REFERENCES

[1] L. Wen, D. Du, Z. Cai, Z. Lei, M.-C. Chang, H. Qi, J. Lim, M.-H. Yang, and S. Lyu, "Ua-detrac: A new benchmark and protocol for multi-object detection and tracking," *ArXiv:1511.04136*, 2015.

[2] A. Milan, L. Leal-Taixé, I. Reid, S. Roth, and K. Schindler, "Mot16: A benchmark for multi-object tracking," *ArXiv:1603.00831*, 2016.

[3] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[4] H.-N. Hu, Q.-Z. Cai, D. Wang, J. Lin, M. Sun, P. Krähenbühl, T. Darrell, and F. Yu, "Joint monocular 3d vehicle detection and tracking," in *IEEE International Conference on Computer Vision (ICCV)*, 2019.

[5] J. Pang, L. Qiu, X. Li, H. Chen, Q. Li, T. Darrell, and F. Yu, "Quasi-dense similarity learning for multiple object tracking," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[6] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[7] R. Girshick, "Fast r-cnn," in *IEEE International Conference on Computer Vision (ICCV)*, 2015.

[8] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2015.

[9] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *European Conference on Computer Vision (ECCV)*, 2016.

[10] J. Redmon and A. Farhadi, "Yolo9000: Better, faster, stronger," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[11] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *IEEE International Conference on Computer Vision (ICCV)*, 2017.

[12] X. Zhou, D. Wang, and P. Krähenbühl, "Objects as points," *ArXiv:1904.07850*, 2019.

[13] H. Law and J. Deng, "Cornernet: Detecting objects as paired keypoints," in *European Conference on Computer Vision (ECCV)*, 2018.

[14] Z. Tian, C. Shen, H. Chen, and T. He, "Fcos: Fully convolutional one-stage object detection," in *IEEE International Conference on Computer Vision (ICCV)*, 2019.

[15] X. Chen, K. Kundu, Y. Zhu, A. G. Berneshawi, H. Ma, S. Fidler, and R. Urtasun, "3D object proposals for accurate object class detection," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2015.

[16] X. Chen, K. Kundu, Z. Zhang, H. Ma, S. Fidler, and R. Urtasun, "Monocular 3D object detection for autonomous driving," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[17] A. Mousavian, D. Anguelov, J. Flynn, and J. Košecká, "3D bounding box estimation using deep learning and geometry," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[18] G. Brazil and X. Liu, "M3D-RPN: Monocular 3D region proposal network for object detection," in *IEEE International Conference on Computer Vision (ICCV)*, 2019.

[19] A. Simonelli, S. R. Bulo, L. Porzi, M. López-Antequera, and P. Kontschieder, "Disentangling monocular 3D object detection," in *IEEE International Conference on Computer Vision (ICCV)*, 2019.

[20] Y. Wang, W.-L. Chao, D. Garg, B. Hariharan, M. Campbell, and K. Q. Weinberger, "Pseudo-LiDAR from visual depth estimation: Bridging the gap in 3D object detection for autonomous driving," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[21] X. Ma, Z. Wang, H. Li, P. Zhang, W. Ouyang, and X. Fan, "Accurate monocular 3D object detection via color-embedded 3D reconstruction for autonomous driving," in *IEEE International Conference on Computer Vision (ICCV)*, 2019.

[22] A. Kundu, Y. Li, and J. M. Rehg, "3D-RCNN: Instance-level 3D object reconstruction via render-and-compare," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[23] F. Chabot, M. Chaouch, J. Rabarisoa, C. Teuliere, and T. Chateau, "Deep MANTA: A coarse-to-fine many-task network for joint 2D and 3D vehicle analysis from monocular image," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[24] Y. Zhou and O. Tuzel, "Voxelnet: End-to-end learning for point cloud based 3d object detection," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[25] S. Shi, X. Wang, and H. Li, "PointRCNN: 3D object proposal generation and detection from point cloud," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[26] A. H. Lang, S. Vora, H. Caesar, L. Zhou, J. Yang, and O. Beijbom, "PointPillars: Fast encoders for object detection from point clouds," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[27] X. Chen, H. Ma, J. Wan, B. Li, and T. Xia, "Multi-view 3d object detection network for autonomous driving," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[28] A. Yilmaz, O. Javed, and M. Shah, "Object tracking: A survey," *ACM computing surveys (CSUR)*, 2006.

[29] S. Salti, A. Cavallaro, and L. Di Stefano, "Adaptive appearance modeling for video tracking: Survey and evaluation," *IEEE Transactions on Image Processing (TIP)*, 2012.

[30] A. W. Smeulders, D. M. Chu, R. Cucchiara, S. Calderara, A. Dehghan, and M. Shah, "Visual tracking: An experimental survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2014.

[31] D. S. Bolme, J. R. Beveridge, B. A. Draper, and Y. M. Lui, "Visual object tracking using adaptive correlation filters," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010.

[32] A. Gaidon, Q. Wang, Y. Cabon, and E. Vig, "Virtual worlds as proxy for multi-object tracking analysis," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[33] M. Kristan, J. Matas, A. Leonardis, M. Felsberg, L. Cehovin, G. Fernández, T. Vojir, G. Hager, G. Nebehay, and R. Pflugfelder, "The visual object tracking vot2015 challenge results," in *IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, 2015.

[34] S. Hare, S. Golodetz, A. Saffari, V. Vineet, M.-M. Cheng, S. L. Hicks, and P. H. Torr, "Struck: Structured output tracking with kernels," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2016.

[35] B. Babenko, M.-H. Yang, and S. Belongie, "Visual tracking with online multiple instance learning," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[36] Z. Kalal, K. Mikolajczyk, and J. Matas, "Tracking-learning-detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2012.

[37] W. Luo, J. Xing, A. Milan, X. Zhang, W. Liu, X. Zhao, and T.-K. Kim, "Multiple object tracking: A literature review," *Arxiv:1409.7618*, 2017.

[38] R. Tao, E. Gavves, and A. W. Smeulders, "Siamese instance search for tracking," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[39] L. Bertinetto, J. Valmadre, J. F. Henriques, A. Vedaldi, and P. H. Torr, "Fully-convolutional siamese networks for object tracking," in *European Conference on Computer Vision (ECCV)*, 2016.

[40] C. Feichtenhofer, A. Pinz, and A. Zisserman, "Detect to track and track to detect," in *IEEE International Conference on Computer Vision (ICCV)*, 2017.

[41] P. Bergmann, T. Meinhardt, and L. Leal-Taixé, "Tracking without bells and whistles," in *IEEE International Conference on Computer Vision (ICCV)*, 2019.

[42] D. Mykheievskyi, D. Borysenko, and V. Porokhonskyy, "Learning local feature descriptors for multiple object tracking," in *Asian Conference on Computer Vision (ACCV)*, 2020.

[43] W. Zhang, H. Zhou, S. Sun, Z. Wang, J. Shi, and C. C. Loy, "Robust multi-modality multi-object tracking," in *IEEE International Conference on Computer Vision (ICCV)*, 2019.

[44] Li Zhang, Yuan Li, and R. Nevatia, "Global data association for multi-object tracking using network flows," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008.

[45] W. Choi, "Near-online multi-target tracking with aggregated local flow descriptor," in *IEEE International Conference on Computer Vision (ICCV)*, 2015.

[46] C. Kim, F. Li, A. Ciptadi, and J. M. Rehg, "Multiple hypothesis tracking revisited," in *IEEE International Conference on Computer Vision (ICCV)*, 2015.

[47] A. Ess, B. Leibe, K. Schindler, and L. Van Gool, "A mobile vision system for robust multi-person tracking," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008.

[48] S. Scheidegger, J. Benjaminsson, E. Rosenberg, A. Krishnan, and K. Granstrom, "Mono-camera 3d multi-object tracking using deep learning detections and pmbm filtering," in *IEEE Intelligent Vehicles Symposium (IV)*, 2018.

[49] P. Voigtlaender, M. Krause, A. Osep, J. Luiten, B. B. G. Sekar, A. Geiger, and B. Leibe, "Mots: Multi-object tracking and segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7942–7951.

[50] S. Sharma, J. A. Ansari, J. Krishna Murthy, and K. Madhava Krishna, "Beyond pixels: Leveraging geometry and shape cues for online multi-object tracking," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2018.

[51] J. Luiten, T. Fischer, and B. Leibe, "Track to reconstruct and reconstruct to track," *IEEE Robotics and Automation Letters (RA-L)*, 2020.

[52] P. Li, T. Qin, and a. Shen, "Stereo vision-based semantic 3d object and ego-motion tracking for autonomous driving," in *European Conference on Computer Vision (ECCV)*, 2018.

[53] A. Osep, W. Mehner, M. Mathias, and B. Leibe, "Combined image- and world-space tracking in traffic scenes," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2017.

[54] X. Weng, J. Wang, D. Held, and K. Kitani, "3d multi-object tracking: A baseline and new evaluation metrics," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2020.

[55] Z. Lu, V. Rathod, R. Votel, and J. Huang, "Retinatrack: Online single stage joint detection and tracking," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[56] T. Yin, X. Zhou, and P. Krähenbühl, "Center-based 3d object detection and tracking," *ArXiv:2006.11275*, 2020.

[57] X. Zhou, V. Koltun, and P. Krähenbühl, "Tracking objects as points," in *IEEE International Conference on Computer Vision (ICCV)*, 2020.

[58] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[59] W. Maddern, G. Pascoe, C. Linegar, and P. Newman, "1 year, 1000 km: The oxford robotcar dataset," *International Journal of Robotics Research (IJRR)*, 2017.

[60] F. Yu, H. Chen, X. Wang, W. Xian, Y. Chen, F. Liu, V. Madhavan, and T. Darrell, "Bdd100k: A diverse driving dataset for heterogeneous multitask learning," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[61] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "nuscenes: A multimodal dataset for autonomous driving," *ArXiv:1903.11027*, 2019.

[62] M.-F. Chang, J. Lambert, P. Sangkloy, J. Singh, S. Bak, A. Hartnett, D. Wang, P. Carr, S. Lucey, D. Ramanan, and J. Hays, "Argoverse: 3d tracking and forecasting with rich maps," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[63] P. Sun, H. Kretzschmar, X. Dotiwalla, A. Chouard, V. Patnaik, P. Tsui, J. Guo, Y. Zhou, Y. Chai, B. Caine, V. Vasudevan, W. Han, J. Ngiam, H. Zhao, A. Timofeev, S. Ettinger, M. Krivokon, A. Gao, A. Joshi, Y. Zhang, J. Shlens, Z. Chen, and D. Anguelov, "Scalability in perception for autonomous driving: Waymo open dataset," *ArXiv:1912.04838*, 2019.

[64] G. Ros, L. Sellart, J. Materzynska, D. Vazquez, and A. M. Lopez, "The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[65] S. R. Richter, V. Vineet, S. Roth, and V. Koltun, "Playing for data: Ground truth from computer games," in *European Conference on Computer Vision (ECCV)*, 2016.

[66] S. R. Richter, Z. Hayder, and V. Koltun, "Playing for benchmarks," in *IEEE International Conference on Computer Vision (ICCV)*, 2017.

[67] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun, "Carla: An open urban driving simulator," in *Conference on Robot Learning (CoRL)*, 2017.

[68] P. Krähenbühl, "Free supervision from video games," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[69] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *IEEE International Conference on Computer Vision (ICCV)*, 2017.

[70] P. J. Huber, "Robust estimation of a location parameter," *The Annals of Mathematical Statistics*, 1964.

[71] A. Hermans, L. Beyer, and B. Leibe, "In defense of the triplet loss for person re-identification," *ArXiv:1703.07737*, 2017.

[72] J. Valmadre, L. Bertinetto, J. Henriques, A. Vedaldi, and P. H. S. Torr, "End-to-end representation learning for correlation filter based tracking," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[73] Z. Wu, Y. Xiong, X. Y. Stella, and D. Lin, "Unsupervised feature learning via non-parametric instance discrimination," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[74] A. Kendall and Y. Gal, "What uncertainties do we need in bayesian deep learning for computer vision?" in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[75] G. Brazil, G. Pons-Moll, X. Liu, and B. Schiele, "Kinematic 3d object detection in monocular video," in *European Conference on Computer Vision (ECCV)*, Virtual, 2020.

[76] F. Yu, W. Li, Q. Li, Y. Liu, X. Shi, and J. Yan, "Poi: multiple object tracking with high performance detection and appearance feature," in *European Conference on Computer Vision (ECCV)*, 2016.

[77] H. W. Kuhn, "The hungarian method for the assignment problem," in *Naval Research Logistics Quarterly*, 1955.

[78] N. Wojke, A. Bewley, and D. Paulus, "Simple online and realtime tracking with a deep association metric," in *IEEE International Conference on Image Processing (ICIP)*, 2017.

[79] Y. Xiang, A. Alahi, and S. Savarese, "Learning to track: Online multi-object tracking by decision making," in *IEEE International Conference on Computer Vision (ICCV)*, 2015.

[80] A. Sadeghian, A. Alahi, and S. Savarese, "Tracking the untrackable: Learning to track multiple cues with long-term dependencies," in *IEEE International Conference on Computer Vision (ICCV)*, 2017.

[81] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision (IJCV)*, 2015.

[82] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[83] F. Yu, D. Wang, E. Shelhamer, and T. Darrell, "Deep layer aggregation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[84] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[85] H. Robbins and S. Monro, "A stochastic approximation method," *Ann. Math. Statist.*, vol. 22, no. 3, pp. 400–407, 09 1951.

[86] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European Conference on Computer Vision (ECCV)*, 2014.

[87] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International Journal of Computer Vision (IJCV)*, vol. 88, 2010.

[88] K. Bernardin and R. Stiefelhagen, "Evaluating multiple object tracking performance: the clear mot metrics," *EURASIP Journal on Image and Video Processing (JIVP)*, 2008.

[89] Y. Li, C. Huang, and R. Nevatia, "Learning to associate: Hybridboosted multi-target tracker for crowded scene," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[90] R. E. Kalman, "A new approach to linear filtering and prediction problems," *Journal of basic Engineering*, 1960.

[91] B. Zhu, Z. Jiang, X. Zhou, Z. Li, and G. Yu, "Class-balanced grouping and sampling for point cloud 3d object detection," *ArXiv:1908.09492*, 2019.

[92] Y. Wang, S. Chen, L. Huang, R. Ge, Y. Hu, Z. Ding, and J. Liao, "1st place solutions for waymo open dataset challenges – 2d and 3d tracking," *Arxiv:2006.15506*, 2020.

[93] O. D. Team, "Openpcdet: An open-source toolbox for 3d object detection from point clouds," https://github.com/open-mmlab/OpenPCDet, 2020.

[94] S. Hwang, J. Park, N. Kim, Y. Choi, and I. S. Kweon, "Multispectral pedestrian detection: Benchmark dataset and baseline," *Integrated Computer-Aided Engineering (ICAE)*, 2013.
[95] L. Leal-Taixé, A. Milan, I. Reid, S. Roth, and K. Schindler, "Motchallenge 2015: Towards a benchmark for multi-target tracking," *ArXiv:1504.01942*, 2015.
[96] A. Shenoi, M. Patel, J. Gwak, P. Goebel, A. Sadeghian, H. Rezatofighi, R. Martin-Martin, and S. Savarese, "Jrmot: A real-time 3d multi-object tracker and a new large-scale dataset," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2020.
[97] H. Karunasekera, H. Wang, and H. Zhang, "Multiple object tracking with attention to appearance, structure, motion and size," *IEEE Access*, 2019.
[98] G. Gündüz and T. Acarman, "A lightweight online multiple object vehicle tracking method," in *IEEE Intelligent Vehicles Symposium (IV)*, 2018.
[99] B. Lee, E. Erdenee, S. Jin, M. Y. Nam, Y. G. Jung, and P. Rhee, "Multi-class multi-object tracking using changing point detection," in *European Conference on Computer Vision Workshops (ECCV Workshops)*, 2016.
[100] W. Choi, "Near-online multi-target tracking with aggregated local flow descriptor," in *IEEE International Conference on Computer Vision (ICCV)*, 2015.
[101] D. Frossard and R. Urtasun, "End-to-end learning of multi-sensor 3d tracking by detection," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2018.
[102] J. Hong Yoon, C.-R. Lee, M.-H. Yang, and K.-J. Yoon, "Online multi-object tracking via structural constraint event aggregation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
[103] P. Lenz, A. Geiger, and R. Urtasun, "Followme: Efficient online min-cost flow tracking with bounded memory and computation," in *IEEE International Conference on Computer Vision (ICCV)*, 2015.