TABLE III
PERFORMANCE COMPARISON OF STATE-OF-THE-ART 3D OBJECT DETECTION ON THE KITTI [19] *test* BENCHMARK

Method	Input	Car			Pedestrian			Cyclist			mAP(Mod)	Speed(s)
		Easy	Mod	Hard	Easy	Mod	Hard	Easy	Mod	Hard		
Mono3D [58]	Image	2.53	2.31	2.31	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Deep3DBox [49]	Image	5.84	4.09	3.83	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
OFT-Net [63]	Image	3.28	2.50	2.27	1.06	1.11	1.06	0.43	0.43	0.43	1.35	0.50
3DOP [57]	Stereo Image	6.55	5.07	4.10	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
VeloFCN [68]	LiDAR	15.20	13.66	15.98	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
VoxelNet [69]	LiDAR	77.47	65.11	57.73	39.48	33.69	31.50	61.22	48.36	44.37	49.05	0.23
SECOND [96]	LiDAR	83.13	73.66	66.20	51.07	42.56	37.29	70.51	53.85	46.90	56.69	0.038
PointPillars [112]	LiDAR	79.05	74.99	68.30	52.08	43.53	41.49	75.78	59.07	52.92	59.20	0.016
BirdNet [75]	LiDAR	14.75	13.44	12.04	14.31	11.80	10.55	18.35	12.43	11.88	12.56	0.11
PointRCNN-v1.1 [110]	LiDAR	85.94	75.76	68.32	49.43	41.78	38.63	73.93	59.60	53.59	59.05	0.10
MV3D [8]	Image & LiDAR	71.09	62.35	55.12	N/A	N/A	N/A	N/A	N/A	N/A	N/A	0.36
UberATG-ContFuse [100]	Image & LiDAR	82.54	66.22	64.04	N/A	N/A	N/A	N/A	N/A	N/A	N/A	0.06
RoarNet [99]	Image & LiDAR	84.25	74.29	59.78	N/A	N/A	N/A	N/A	N/A	N/A	N/A	0.10
AVOD [9]	Image & LiDAR	73.59	65.78	58.38	38.28	31.51	26.98	60.11	44.90	38.80	47.40	0.08
AVOD-FPN [9]	Image & LiDAR	81.94	71.88	66.38	50.80	42.81	40.88	64.00	52.18	46.61	55.62	0.10
F-PointNet [65]	Image & LiDAR	81.20	70.39	62.19	51.21	44.89	40.23	71.96	56.77	50.39	57.35	0.17

TABLE IV
PERFORMANCE COMPARISON OF STATE-OF-THE-ART BIRD'S EYE VIEW (BEV) DETECTION ON THE KITTI [19] *test* BENCHMARK

Method	Input	Car			Pedestrian			Cyclist			mAP(Mod)	Speed(s)
		Easy	Mod	Hard	Easy	Mod	Hard	Easy	Mod	Hard		
OFT-Net [63]	Image	9.50	7.99	7.51	1.93	1.55	1.65	0.79	0.43	0.43	3.32	0.50
UberATG-PIXOR [78]	LiDAR	81.70	77.05	72.95	N/A	N/A	N/A	N/A	N/A	N/A	N/A	0.035
VoxelNet [69]	LiDAR	89.35	79.26	77.39	46.13	40.74	38.11	66.70	54.76	50.55	58.25	0.23
SECOND [96]	LiDAR	88.07	79.37	77.95	55.10	46.27	44.76	73.67	56.04	48.78	60.56	0.038
PointPillars [112]	LiDAR	88.35	86.10	79.83	58.66	50.23	47.19	79.14	62.25	56.00	66.19	0.016
PointRCNN-v1.1 [110]	LiDAR	89.47	85.68	79.10	55.92	47.53	44.67	81.52	66.77	60.78	66.66	0.10
BirdNet [75]	LiDAR	75.52	50.81	50.00	26.07	21.35	19.96	38.93	27.18	25.51	33.11	0.11
UberATG-PIXOR++ [82]	LiDAR	89.38	83.70	77.97	N/A	N/A	N/A	N/A	N/A	N/A	N/A	0.035
UberATG-HDNET [82]	LiDAR & Map	89.14	86.57	78.32	N/A	N/A	N/A	N/A	N/A	N/A	N/A	0.05
MV3D [8]	Image & LiDAR	86.02	76.90	68.49	N/A	N/A	N/A	N/A	N/A	N/A	N/A	0.36
UberATG-ContFuse [100]	Image & LiDAR	88.81	85.83	77.33	N/A	N/A	N/A	N/A	N/A	N/A	N/A	0.06
RoarNet [99]	Image & LiDAR	88.19	79.77	69.83	N/A	N/A	N/A	N/A	N/A	N/A	N/A	0.10
AVOD-FPN [9]	Image & LiDAR	88.53	83.79	77.90	58.75	51.05	47.54	68.09	57.48	50.77	64.11	0.10
AVOD [9]	Image & LiDAR	86.80	85.44	77.73	42.51	35.24	33.97	63.66	47.74	46.55	56.14	0.08
F-PointNet [65]	Image & LiDAR	88.70	84.00	75.33	58.09	50.22	47.20	75.38	61.96	54.68	65.39	0.17

TABLE V
PERFORMANCE COMPARISONS OF STATE-OF-THE-ART 3D OBJECT DETECTION FOR 10-CLASSES EVALUATION ON SUN RGB-D [25] DATASET.

Method	Key Input Processing	🛁	🛏	📚	🪑	🗃	🗄	⊓	🛋	⊤	🚽	mAP	Speed
COG [93]	Point Cloud	58.26	63.67	31.80	62.17	45.19	15.47	27.36	51.02	51.29	70.07	47.63	10-30m
LSS [94]	Point Cloud	76.20	73.20	32.90	60.50	34.50	13.50	30.40	60.40	55.40	73.70	51.00	10-30m
2D-driven [66]	Image & Depth	43.45	64.48	31.40	48.27	27.93	25.92	41.92	50.39	37.02	80.40	45.12	4.15s
Rahman *et al.* [67]	Image & Depth	44.10	78.10	12.00	54.40	19.70	33.10	44.50	52.10	37.80	80.90	45.70	0.30s
DSS [7]	Image & Point Cloud	44.20	78.80	11.90	61.20	20.50	6.40	15.40	53.50	50.30	78.90	42.10	19.55s
F-PointNet [65]	Image & Point Cloud	43.30	81.10	33.30	64.20	24.70	32.00	58.10	61.10	51.10	90.90	54.00	0.12s
PointFusion [98]	Image & Point Cloud	37.26	68.57	37.69	55.09	17.16	23.95	32.33	53.83	31.03	83.80	45.38	1.30s
SIFRNet [107]	Image & Point Cloud	64.00	84.40	38.40	57.90	34.10	32.20	67.70	67.30	51.40	86.20	58.40	N/A

TABLE VI
PERFORMANCE COMPARISONS OF STATE-OF-THE-ART 3D OBJECT DETECTION FOR 19-CLASSES EVALUATION ON NYUV2 [20] DATASET.

Method	Key Input Processing	mAP	Speed (s)
Deng *et al.* [40]	Image & Depth	40.9	0.74s
Rahman *et al.* [67]	Image & Depth	43.1	0.30s
DSS [7]	Image & Point Cloud	36.3	19.55s

hope that this survey will serve as a supportive reference and a significant contribution to the research community.

REFERENCES

[1] S. Ren, K. He, R. B. Girshick, and J. Sun, "Faster R-CNN: towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 39, no. 6, pp. 1137–1149, 2017.

[2] J. Redmon, S. K. Divvala, R. B. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 779–788.

[3] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. E. Reed, C. Fu, and A. C. Berg, "SSD: single shot multibox detector," in *Proceedings of the 14th European Conference on Computer Vision ECCV*, 2016, pp. 21–37.

[4] K. He, G. Gkioxari, P. Dollár, and R. B. Girshick, "Mask R-CNN," in

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIP.2019.2955239, IEEE Transactions on Image Processing

Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2980–2988.

[5] A. Ioannidou, E. Chatzilari, S. Nikolopoulos, and I. Kompatsiaris, "Deep learning advances in computer vision with 3d data: A survey," *ACM Comput. Surv.*, vol. 50, no. 2, pp. 20:1–20:38, 2017.

[6] M. Naseer, S. Khan, and F. Porikli, "Indoor scene understanding in 2.5/3d for autonomous agents: A survey," *IEEE Access*, vol. 7, pp. 1859–1887, 2019.

[7] S. Song and J. Xiao, "Deep sliding shapes for amodal 3d object detection in RGB-D images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 808–816.

[8] X. Chen, H. Ma, J. Wan, B. Li, and T. Xia, "Multi-view 3d object detection network for autonomous driving," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 6526–6534.

[9] J. Ku, M. Mozifian, J. Lee, A. Harakeh, and S. L. Waslander, "Joint 3d proposal generation and object detection from view aggregation," in *International Conference on Intelligent Robots and Systems (IROS)*, 2018, pp. 1–8.

[10] R. B. Girshick, "Fast R-CNN," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 1440–1448.

[11] Y. Xiang, W. Choi, Y. Lin, and S. Savarese, "Data-driven 3d voxel patterns for object category recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 1903–1911.

[12] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "Pointnet: Deep learning on point sets for 3d classification and segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 77–85.

[13] S. Gupta, R. B. Girshick, P. A. Arbeláez, and J. Malik, "Learning rich features from RGB-D images for object detection and segmentation," in *Proceedings of the 13th European Conference on Computer Vision (ECCV)*, 2014, pp. 345–360.

[14] H. Badino, U. Franke, and D. Pfeiffer, "The stixel world - A compact medium level representation of the 3d-world," in *Pattern Recognition, 31st DAGM Symposium, Jena, Germany, September 9-11, 2009. Proceedings*, 2009, pp. 51–60.

[15] H. P. Eberhardt, V. Klumpp, and U. D. Hanebeck, "Density trees for efficient nonlinear state estimation," in *13th Conference on Information Fusion, FUSION 2010, Edinburgh, UK, July 26-29, 2010*, 2010, pp. 1–8.

[16] A. Janoch, S. Karayev, Y. Jia, J. T. Barron, M. Fritz, K. Saenko, and T. Darrell, "A category-level 3-d object dataset: Putting the kinect to work," in *IEEE International Conference on Computer Vision Workshops, ICCV*, 2011, pp. 1168–1174.

[17] K. Lai, L. Bo, X. Ren, and D. Fox, "A large-scale hierarchical multi-view RGB-D object dataset," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2011, pp. 1817–1824.

[18] H. S. Koppula, A. Anand, T. Joachims, and A. Saxena, "Semantic labeling of 3d point clouds for indoor scenes," in *Advances in Neural Information Processing Systems*, 2011, pp. 244–252.

[19] 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, pp. 3354–3361.

[20] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from RGBD images," in *Proceedings of the 12th European Conference on Computer Vision (ECCV)*, 2012, pp. 746–760.

[21] J. Xiao, A. Owens, and A. Torralba, "SUN3D: A database of big spaces reconstructed using sfm and object labels," in *IEEE International Conference on Computer Vision, ICCV*, 2013, pp. 1625–1632.

[22] J. J. Lim, H. Pirsiavash, and A. Torralba, "Parsing IKEA objects: Fine pose estimation," in *IEEE International Conference on Computer Vision, ICCV*, 2013, pp. 2992–2999.

[23] M. De Deuge, A. Quadros, C. Hung, and B. Douillard, "Unsupervised feature learning for classification of outdoor 3d scans," in *Australasian Conference on Robitics and Automation*, vol. 2, 2013, p. 1.

[24] Y. Xiang, R. Mottaghi, and S. Savarese, "Beyond PASCAL: A benchmark for 3d object detection in the wild," in *IEEE Winter Conference on Applications of Computer Vision*, 2014, pp. 75–82.

[25] S. Song, S. P. Lichtenberg, and J. Xiao, "SUN RGB-D: A RGB-D scene understanding benchmark suite," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 567–576.

[26] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 3213–3223.

[27] Y. Xiang, W. Kim, W. Chen, J. Ji, C. B. Choy, H. Su, R. Mottaghi, L. J. Guibas, and S. Savarese, "Objectnet3d: A large scale database for 3d object recognition," in *European Conference on Computer Vision - ECCV*, 2016, pp. 160–176.

[28] L. Yi, V. G. Kim, D. Ceylan, I. Shen, M. Yan, H. Su, C. Lu, Q. Huang, A. Sheffer, L. Guibas *et al.*, "A scalable active framework for region annotation in 3d shape collections," *ACM Transactions on Graphics (TOG)*, vol. 35, no. 6, p. 210, 2016.

[29] B. Drost, M. Ulrich, P. Bergmann, P. Härtinger, and C. Steger, "Introducing mvtec ITODD - A dataset for 3d object recognition in industry," in *IEEE International Conference on Computer Vision Workshops, ICCV*, 2017, pp. 2200–2208.

[30] T. Hodan, P. Haluza, S. Obdrzálek, J. Matas, M. I. A. Lourakis, and X. Zabulis, "T-LESS: an RGB-D dataset for 6d pose estimation of texture-less objects," in *IEEE Winter Conference on Applications of Computer Vision, WACV*, 2017, pp. 880–888.

[31] A. Dai, A. X. Chang, M. Savva, M. Halber, T. A. Funkhouser, and M. Nießner, "Scannet: Richly-annotated 3d reconstructions of indoor scenes," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 2432–2443.

[32] I. Armeni, S. Sax, A. R. Zamir, and S. Savarese, "Joint 2d-3d-semantic data for indoor scene understanding," *CoRR*, vol. abs/1702.01105, 2017.

[33] S. Song, F. Yu, A. Zeng, A. X. Chang, M. Savva, and T. A. Funkhouser, "Semantic scene completion from a single depth image," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 190–198.

[34] J. Tremblay, T. To, and S. Birchfield, "Falling things: A synthetic dataset for 3d object detection and pose estimation," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR*, 2018, pp. 2038–2041.

[35] Y. Wu, Y. Wu, G. Gkioxari, and Y. Tian, "Building generalizable agents with a realistic and rich 3d environment," in *International Conference on Learning Representations, ICLR*, 2018.

[36] 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 preprint arXiv:1903.11027*, 2019.

[37] A. Patil, S. Malla, H. Gang, and Y.-T. Chen, "The h3d dataset for full-surround 3d multi-object detection and tracking in crowded urban scenes," in *International Conference on Robotics and Automation*, 2019.

[38] J. Jeong, Y. Cho, Y.-S. Shin, H. Roh, and A. Kim, "Complex urban dataset with multi-level sensors from highly diverse urban environments," in *The International Journal of Robotics Research*, 2019.

[39] S. Gupta, P. Arbelaez, and J. Malik, "Perceptual organization and recognition of indoor scenes from RGB-D images," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013, pp. 564–571.

[40] Z. Deng and L. J. Latecki, "Amodal detection of 3d objects: Inferring 3d bounding boxes from 2d ones in rgb-depth images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 398–406.

[41] E. Barnea and O. Ben-Shahar, "Depth based object detection from partial pose estimation of symmetric objects," in *European Conference on Computer Vision - ECCV*, 2014, pp. 377–390.

[42] M. Everingham, L. J. V. Gool, C. K. I. Williams, J. M. Winn, and A. Zisserman, "The pascal visual object classes (VOC) challenge," *International Journal of Computer Vision*, vol. 88, no. 2, pp. 303–338, 2010.

[43] V. Ramanishka, Y.-T. Chen, T. Misu, and K. Saenko, "Toward driving scene understanding: A dataset for learning driver behavior and causal reasoning," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[44] B. Çalli, A. Singh, A. Walsman, S. S. Srinivasa, P. Abbeel, and A. M. Dollar, "The YCB object and model set: Towards common benchmarks for manipulation research," in *International Conference on Advanced Robotics, ICAR*, 2015, pp. 510–517.

[45] M. Cheng, Z. Zhang, W. Lin, and P. H. S. Torr, "BING: binarized normed gradients for objectness estimation at 300fps," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014, pp. 3286–3293.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIP.2019.2955239, IEEE Transactions on Image Processing

[46] C. L. Zitnick and P. Dollár, "Edge boxes: Locating object proposals from edges," in *European Conference on Computer Vision (ECCV)*, 2014, pp. 391–405.

[47] P. Krähenbühl and V. Koltun, "Learning to propose objects," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 1574–1582.

[48] T. S. H. Lee, S. Fidler, and S. J. Dickinson, "Learning to combine mid-level cues for object proposal generation," in *2015 IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 1680–1688.

[49] A. Mousavian, D. Anguelov, J. Flynn, and J. Kosecka, "3d bounding box estimation using deep learning and geometry," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 5632–5640.

[50] F. Chabot, M. Chaouch, J. Rabarisoa, C. Teulière, 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, pp. 1827–1836.

[51] Y. Xiang, W. Choi, Y. Lin, and S. Savarese, "Subcategory-aware convolutional neural networks for object proposals and detection," in *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2017, pp. 924–933.

[52] B. Pepik, M. Stark, P. V. Gehler, T. Ritschel, and B. Schiele, "3d object class detection in the wild," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 1–10.

[53] M. Aubry, D. Maturana, A. A. Efros, B. C. Russell, and J. Sivic, "Seeing 3d chairs: Exemplar part-based 2d-3d alignment using a large dataset of CAD models," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014, pp. 3762–3769.

[54] J. J. Lim, A. Khosla, and A. Torralba, "FPM: fine pose parts-based model with 3d CAD models," in *European Conference on Computer Vision (ECCV)*, 2014, pp. 478–493.

[55] M. Z. Zia, M. Stark, and K. Schindler, "Explicit occlusion modeling for 3d object class representations," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013, pp. 3326–3333.

[56] Y. Lin, V. I. Morariu, W. H. Hsu, and L. S. Davis, "Jointly optimizing 3d model fitting and fine-grained classification," in *European Conference on Computer Vision (ECCV)*, 2014, pp. 466–480.

[57] 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 *Proceedings of the 28th International Conference on Neural Information Processing Systems (NIPS)*, 2015, pp. 424–432.

[58] 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, pp. 2147–2156.

[59] C. C. Pham and J. W. Jeon, "Robust object proposals re-ranking for object detection in autonomous driving using convolutional neural networks," *Sig. Proc.: Image Comm.*, vol. 53, pp. 110–122, 2017.

[60] V. Lepetit, F. Moreno-Noguer, and P. Fua, "Ep*n*p: An accurate $O(n)$ solution to the p*n*p problem," *International Journal of Computer Vision*, vol. 81, no. 2, pp. 155–166, 2009.

[61] Z. Cai, Q. Fan, R. S. Feris, and N. Vasconcelos, "A unified multi-scale deep convolutional neural network for fast object detection," in *European Conference on Computer Vision - ECCV*, 2016, pp. 354–370.

[62] B. Xu and Z. Chen, "Multi-level fusion based 3d object detection from monocular images," in *IEEE Conference on Computer Vision and Pattern Recognition, (CVPR)*, 2018, pp. 2345–2353.

[63] T. Roddick, A. Kendall, and R. Cipolla, "Orthographic feature transform for monocular 3d object detection," in *Proceedings of the British Machine Vision Conference (BMVC)*, 2019.

[64] Y. Wang, W. 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," 2019.

[65] C. R. Qi, W. Liu, C. Wu, H. Su, and L. J. Guibas, "Frustum pointnets for 3d object detection from RGB-D data," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 918–927.

[66] J. Lahoud and B. Ghanem, "2d-driven 3d object detection in rgb-d images," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 4632–4640.

[67] M. M. Rahman, Y. Tan, J. Xue, L. Shao, and K. Lu, "3D object detection: Learning 3d bounding boxes from scaled down 2d bounding boxes in rgb-d images," *Information Sciences*, vol. 476, pp. 147–158, 2019.

[68] B. Li, T. Zhang, and T. Xia, "Vehicle detection from 3d lidar using fully convolutional network," in *Robotics: Science and Systems*, 2016.

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

[70] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "Pointnet++: Deep hierarchical feature learning on point sets in a metric space," in *Proceedings of the 30th International Conference on Neural Information Processing Systems (NIPS)*, 2017, pp. 5105–5114.

[71] H. Su, S. Maji, E. Kalogerakis, and E. G. Learned-Miller, "Multi-view convolutional neural networks for 3d shape recognition," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 945–953.

[72] B. Wu, A. Wan, X. Yue, and K. Keutzer, "Squeezeseg: Convolutional neural nets with recurrent CRF for real-time road-object segmentation from 3d lidar point cloud," in *IEEE International Conference on Robotics and Automation, ICRA*, 2018, pp. 1887–1893.

[73] K. Minemura, H. Liau, A. Monrroy, and S. Kato, "Lmnet: Real-time multiclass object detection on CPU using 3d lidar," *CoRR*, vol. abs/1805.04902, 2018.

[74] S. Yu, T. Westfechtel, R. Hamada, K. Ohno, and S. Tadokoro, "Vehicle detection and localization on bird's eye view elevation images using convolutional neural network," in *IEEE International Symposium on Safety, Security and Rescue Robotics(SSRR)*, 2017, pp. 102–109.

[75] J. Beltrán, C. Guindel, F. M. Moreno, D. Cruzado, F. García, and A. de la Escalera, "Birdnet: A 3d object detection framework from lidar information," in *International Conference on Intelligent Transportation Systems (ITSC)*, 2018, pp. 3517–3523.

[76] S. Wirges, T. Fischer, C. Stiller, and J. B. Frias, "Object detection and classification in occupancy grid maps using deep convolutional networks," in *International Conference on Intelligent Transportation Systems (ITSC)*, 2018, pp. 3530–3535.

[77] X. Du, M. H. Ang, S. Karaman, and D. Rus, "A general pipeline for 3d detection of vehicles," in *IEEE International Conference on Robotics and Automation, ICRA*, 2018, pp. 3194–3200.

[78] B. Yang, W. Luo, and R. Urtasun, "PIXOR: real-time 3d object detection from point clouds," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 7652–7660.

[79] W. Luo, B. Yang, and R. Urtasun, "Fast and furious: Real time end-to-end 3d detection, tracking and motion forecasting with a single convolutional net," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 3569–3577.

[80] M. Simon, S. Milz, K. Amende, and H. Gross, "Complex-yolo: An euler-region-proposal for real-time 3d object detection on point clouds," in *European Conference on Computer Vision - ECCV*, 2018, pp. 197–209.

[81] W. Ali, S. Abdelkarim, M. Zidan, M. Zahran, and A. E. Sallab, "YOLO3D: end-to-end real-time 3d oriented object bounding box detection from lidar point cloud," in *European Conference on Computer Vision - ECCV*, 2018, pp. 716–728.

[82] B. Yang, M. Liang, and R. Urtasun, "HDNET: exploiting HD maps for 3d object detection," in *Conference on Robot Learning (CoRL)*, 2018, pp. 146–155.

[83] J. Redmon and A. Farhadi, "Yolo9000: Better, faster, stronger," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 6517–6525.

[84] X. Zhao, H. Jia, and Y. Ni, "A novel three-dimensional object detection with the modified you only look once method," *International Journal of Advanced Robotic Systems*, vol. 15, no. 2, p. 1729881418765507, 2018.

[85] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, 2015, pp. 3431–3440.

[86] E. Shelhamer, J. Long, and T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 640–651, 2017.

[87] B. Li, "3d fully convolutional network for vehicle detection in point cloud," in *International Conference on Intelligent Robots and Systems IROS*, 2017, pp. 1513–1518.

[88] S. Song and J. Xiao, "Sliding shapes for 3d object detection in depth images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014, pp. 634–651.

[89] T. Malisiewicz, A. Gupta, and A. A. Efros, "Ensemble of exemplar-svms for object detection and beyond," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2011, pp. 89–96.

[90] D. Z. Wang and I. Posner, "Voting for voting in online point cloud object detection," in *Robotics: Science and Systems*, 2015.

[91] M. Engelcke, D. Rao, D. Z. Wang, C. H. Tong, and I. Posner, "Vote3deep: Fast object detection in 3d point clouds using efficient convolutional neural networks," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2017, pp. 1355–1361.

[92] W. Liu, R. Ji, and S. Li, "Towards 3d object detection with bimodal deep boltzmann machines over RGBD imagery," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 3013–3021.

[93] Z. Ren and E. B. Sudderth, "Three-dimensional object detection and layout prediction using clouds of oriented gradients," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 1525–1533.

[94] ——, "3d object detection with latent support surfaces," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 937–946.

[95] N. Sedaghat, M. Zolfaghari, E. Amiri, and T. Brox, "Orientation-boosted voxel nets for 3d object recognition," in *British Machine Vision Conference (BMVC)*, 2017.

[96] Y. Yan, Y. Mao, and B. Li, "SECOND: sparsely embedded convolutional detection," *Sensors*, vol. 18, no. 10, p. 3337, 2018.

[97] Y. Wang, Y. Sun, Z. Liu, S. E. Sarma, M. M. Bronstein, and J. M. Solomon, "Dynamic graph CNN for learning on point clouds," *CoRR*, vol. abs/1801.07829, 2018.

[98] D. Xu, D. Anguelov, and A. Jain, "Pointfusion: Deep sensor fusion for 3d bounding box estimation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 244–253.

[99] K. Shin, Y. P. Kwon, and M. Tomizuka, "Roarnet: A robust 3d object detection based on region approximation refinement," in *IEEE Intelligent Vehicles Symposium, IV*, 2019.

[100] M. Liang, B. Yang, S. Wang, and R. Urtasun, "Deep continuous fusion for multi-sensor 3d object detection," in *European Conference on Computer Vision - (ECCV)*, 2018, pp. 663–678.

[101] X. Du, M. H. Ang, and D. Rus, "Car detection for autonomous vehicle: LIDAR and vision fusion approach through deep learning framework," in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2017, Vancouver, BC, Canada, September 24-28, 2017*, 2017, pp. 749–754.

[102] D. Matti, H. K. Ekenel, and J. Thiran, "Combining lidar space clustering and convolutional neural networks for pedestrian detection," in *14th IEEE International Conference on Advanced Video and Signal Based Surveillance, AVSS 2017, Lecce, Italy, August 29 - September 1, 2017*, 2017, pp. 1–6.

[103] S. Wang, S. Suo, W. Ma, A. Pokrovsky, and R. Urtasun, "Deep parametric continuous convolutional neural networks," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 2589–2597.

[104] S. Ren, K. He, R. B. Girshick, and J. Sun, "Faster R-CNN: towards real-time object detection with region proposal networks," in *Advances in Neural Information Processing Systems*, 2015, pp. 91–99.

[105] T. Lin, P. Dollár, R. B. Girshick, K. He, B. Hariharan, and S. J. Belongie, "Feature pyramid networks for object detection," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 936–944.

[106] C. Fu, W. Liu, A. Ranga, A. Tyagi, and A. C. Berg, "DSSD : Deconvolutional single shot detector," *CoRR*, vol. abs/1701.06659, 2017.

[107] X. Zhao, Z. Liu, R. Hu, and K. Huang, "3d object detection using scale invariant and feature reweighting networks," in *The Thirty-Third AAAI Conference on Artificial Intelligence*, 2019, pp. 9267–9274.

[108] M. Jiang, Y. Wu, and C. Lu, "Pointsift: A sift-like network module for 3d point cloud semantic segmentation," *CoRR*, vol. abs/1807.00652, 2018.

[109] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 7132–7141.

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

[111] 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, pp. 770–778.

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

[113] M. Everingham, L. J. V. Gool, C. K. I. Williams, J. M. Winn, and A. Zisserman, "The pascal visual object classes (VOC) challenge," *International Journal of Computer Vision*, vol. 88, no. 2, pp. 303–338, 2010.

[114] S. Chadwick, W. Maddern, and P. Newman, "Distant vehicle detection using radar and vision," in *International Conference on Robotics and Automation (ICRA)*, 2019, pp. 8311–8317.

Mohammad Muntasir Rahman received Ph.D. degree in Computer Applied Technology from the School of Engineering Sciences, University of Chinese Academy of Sciences, Beijing, China, in 2019. He also received B.Sc and M.Sc degree in Computer Science and Engineering from the Islamic University, Kushtia, Bangladesh, in 2005 and 2006, respectively. Currently, he is an Associate Professor in the Dept. of Computer Science and Engineering, Islamic University, Kushtia, Bangladesh. His research interest include computer vision, machine learning and pattern recognition.

Yanhao Tan is a Master-Docter combined program graduate student in Computer Applied Technology from the University of Chinese Academy of Sciences, Beijing, China, from 2015. His research interest includes deep learning and computer vision, RGB-D object recognition and detection.

Jian Xue was born in Jiangsu, China, in 1979. He received the Ph.D. degree in Computer Applied Technology from the Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2007. He is currently an Associate Professor with the University of Chinese Academy of Sciences, Beijing. Since 2003, he has long been engaged in the research work about out-of-core medical image analysis and processing, and visualization in scientific computing. His current research interests include image processing, computer graphics and scientific visualization.

Ke Lu was born in Ningxia on March 13th, 1971. He received master degree and Ph.D. degree from the Department of Mathematics and Department of Computer Science at Northwest University in July 1998 and July 2003, respectively. He worked as a postdoctoral fellow in the Institute of Automation Chinese Academy of Sciences from July 2003 to April 2005. Currently he is a professor of the University of the Chinese Academy of Sciences. His current research areas focus on computer vision, 3D image reconstruction and computer graphics.