DOI 10.1007/s10946-023-10113-1
Journal of Russian Laser Research, Volume 44, Number 1, January, 2023

GIoU-CLOCs: IMPORT GENERALIZED INTERSECTION OVER UNION-INVOLVED OBJECT-DETECTION TASKS BASED ON LIDAR AND CAMERA

Tiankai Chen,[1,2] **Yuan Zhou,**[1,2] **Shifeng Wang,**[1,2*] **and Bo Lu**[2]

[1]*School of Optoelectronic Engineering, Changchun University of Science and Technology*
Changchun 130022, China

[2]*Zhongshan Institute of Changchun University of Science and Technology*
Zhongshan 528400, China

*Corresponding author e-mail: sf.wang@cust.edu.cn

Abstract

In recent years, the application of LIDAR has become much more extensive, especially in object detection. While laser researches have encouraging performance in detection, they are typically based on a single modality, being unable to collect information from other modalities. In this paper, we introduce a late fusion way to fuse data from LIDAR and RGB camera. For the disorder of laser, we introduce polynomial functions into the 3D network, which enable the network to take higher-order moments of a given shape into account. Considering the geometric and semantic consistency, we fuse point clouds and images to generate more accurate 3D and 2D detection results. Finally, we address the weaknesses of the intersection over union in the fusion network, employing a generalized version as both a new loss and a new metric. The experimental evaluation of the challenging KITTI object detection benchmark shows significant improvements, especially in the birds' eye view, which shows the feasibility and applicability of our work.

Keywords: LIDAR and camera, point cloud, detection.

1. Introduction

In recent years, object detection has become a widely concerned research topic due to its essential role in autonomous driving and robot navigation, and LIDAR are commonly used in it. The principle of LIDAR is to transmit the detection signal to the object, and then compare the received signal with the transmitted signal. LIDAR points have the advantage of providing accurate depth information, while images preserve much more detailed semantic information. A single sensor may lack RGB information or depth information, so the fusion method may improve the accuracy of detection.

However, the different sensors have different data types. Fusion methods can be divided into three categories: early fusion, deep fusion, and late fusion. Early and deep fusions, being complex computational, can easily be influenced by data error. In order to ensure the fusion efficiency and accuracy, we chose the late fusion. The detectors are divided into one-stage and two-stage methods. Many one-stage detectors like YOLOv7 [1] have high real time, but they have low detection accuracy. Many object detectors are based on the two-stage R-CNN [2–4] framework, and detection are framed as a multitask learning problem combining classification and bounding box regression. Different from traditional object detection, an intersection over union (IoU) [11] threshold is required to define positives and negatives.

Manuscript submitted by the authors in English on November 4, 2022.

However, the commonly used bounding box regression metric is used for single-modality networks. As for the late fusion, if the same metric is used to filtrate the correct candidates, it may cause mismatched results. So we import the generalized intersection over union (GIoU) [5] loss into the late fusion. For 2D and 3D common candidates, GIoU can get outstanding performance.

Existing open-source detectors based on deep learning for LIDAR point processing have obvious shortcomings, which lack more point information. As for the way to process raw point clouds, some scholars have promoted feeding geometric features as input to deep neural networks, and the geometric information of point clouds is classified by the deep neural network [6]. So we use geometric moments for point cloud information expansion.

In this paper, we use a fusion network CLOCs [7] to contact information of the point cloud and camera. To get better 3D candidates, we make improvements for coordinate information. The network can supplement the point cloud coordinate information, using polynomial functions. The proposed network implementation was based on a new version of the SECOND [15] architecture. For the difference between traditional candidates and fusion candidates, we chose the GIoU to achieve classification and bounding box regression. We take 2D and 3D detections, without doing non-maximum suppression (NMS), as some correct detections which may be suppressed because of limited information from a single sensor modality. The network has achieved satisfactory results on the KITTI data set and shows performance improvement on standard object detection benchmarks.

2. Related Works

2.1. Different Fusion Ways

Fusion architectures can be classified according to the fusion method of multimodal data features. Early fusion usually converts multimodal data to one coordinate system before training, and deep fusion combines the intermedia representations of several base networks to serve as inputs for the rest of the network. Vishwanath A. Sindagi et al. proposed a Multimodal VoxelNet [8] to augment LIDAR points with semantic image features and learned to fuse image and LIDAR features at early stages for accurate 3D object detection. To fuse RGB and point cloud information, features are first extracted from the last convolution layer of a 2D detection network. These image features encode semantic information that can be used as prior knowledge to help infer the presence of an object, points or voxels [9]. These objects are projected onto the image and further used in the 3D region proposal network (RPN) to produce 3D bounding boxes.

Xiaozhi Chen et al. [10] proposed a Multi-View 3D object detection network (MV3D) that takes multimodal data as input and predicts the full 3D extent of objects in the 3D space. The main idea of using multimodal information is feature fusion based on region. The Multi-View fusion network extracts regional features by projecting 3D proposals onto multi-view feature maps. They designed a deeply converged approach to achieve information interaction in the middle layers from different views. This network can accurately predict the position, size, and direction of objects in 3D space by directional box regression. Yingwei Li et al. developed an effective multimodal 3D detector [23]; their research shows that when the depth features are well aligned, the late depth feature fusion is more effective.

The crop and resize operations used in the above algorithms to fuse feature vectors from different modalities may destroy the feature structure of each sensor. However, the disadvantages mentioned above will not occur in late fusion. In order to ensure the features are not destroyed, different modality features are entered into different detectors, and the predicted values of the output are spliced together to predict the final result. So we chose the late fusion due to the stability.

2.2. Intersection over Union

The intersection over union (IoU) [11] calculates the ratio of the intersection and union of "predicted anchors" and "real anchors." During training, it is used to distinguish true positives and negatives in a set of predictions. There is an inevitable relationship between the calculation of the non-maximum suppression (NMS) and the score of classification. The maximum score is used as the ground truth, and then the rest is compared with this ground truth by IoU. All performance measures are used to evaluate segmentation, object detection, and tracking rely on this metric. During the evaluation, average precision (AP) calculation needs the threshold of IoU. Therefore, the calculation method and assignment of IoU are crucial for object detection.

2.3. Deep Learning of Geometric Structures

The input of 3D detection is point cloud or voxel in traditional networks, which generally contain coordinates in the point cloud space and the corresponding reflection intensity. There are some methods to raise each point cloud to a high-dimensional space for learning. Qi et al. [21] designed a network to learn the high-dimensional information of the point cloud; they also proved that the network can learn any continuous function. The geometric moment is an important characteristic for judging objects, and every surface can be represented by a finite set of moments [22]. Two surfaces can be considered identical, if their geometric moments are the same. The same theory can be applied to different point clouds. Joseph Rivlin et al. [6] added polynomial functions to the coordinates of the point cloud, leveraging the ability of the network to learn polynomial function information.

3. GIoU-CLOCs

3.1. Polynomial Functions

For the 2D detector, we choose the detection results of Cascad R-CNN [14] as the input, while the 3D detectors use the SECOND [15] to train the LIDAR points. Most 3D detectors are based on Voxelnet [9], and they take the mean of multiple random points within a voxel to represent the voxel information. This will make information of many points incomplete. So we leverage the network ability to operate voxels by adding polynomial functions to their coordinates. Such a design can allow the network to account for higher-order moments and improve classification accuracy to a certain extent. According to the different spatial characteristics of the point cloud, there are many ways to describe point cloud information. The first-order moments represent the non-intrinsic centroid. The second-order moments is used to measure the covariance and can also be viewed as the moments of inertia. Different classes have different applications. Second-order moments of a set of points can be compactly expressed in a voxel, where X_j defines a point given as a vector of its coordinates $X_J = (x_j, y_j, z_j)^T$. Adding these moments to the input helps the network to learn more geometric information on point clouds.

We followed the method of [6], using the relation between point clouds by learning their moments implicitly correlated. Explicitly, the functions $(x^2, y^2, z^2, xy, yz, xz)$ of each point are given to a neural network as input features to obtain better accuracy. In traditional fusion detection networks, those methods focus more on the part of data fusion and ignore the difference between the sparsity of the point cloud and the richness of the image semantic information. Increasing the point cloud geometric moments, we obtain more credible information of our 3D detections, so that they can better integrate with 2D detections.

3.2. GIoU

Intersection over union (IoU) is the ratio of intersection and union for candidates. But if there is no intersection between two candidates: $|A \cap B| = 0$ and IoU $(A, B) = 0$, IoU does not reflect if two shapes are in the vicinity of each other or very far from each other. Most fusion networks currently use IoU as the metric for classification and bounding box regression. If they are ignored, there are fewer candidates on the input to the fusion network compared with the 2D detection network, and if a part of information about intersections is lost, it may affect both convergence rate and training quality. To address this issue, we use a general extension to IoU, named Generalized Intersection over Union (GIoU).

As for classification and bounding box regression [13], GIoU, as a metric for any two arbitrary shapes, can get the gradient in any case. According to [5], the first finding is the smallest convex shape C, which contains A and B: $C \subseteq S \in R^n$. As we can see in Fig. 1, the grey line (the biggest rectangle) is the definition of the smallest convex shape C. Then we calculate a ratio between the volume (area) occupied by C excluding A and B, next use the last step calculated values to divide by the total volume (area) occupied C. Finally, GIoU is defined by the ratio minus the IoU.

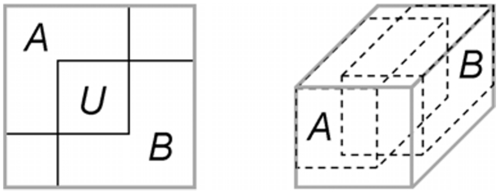

Fig. 1. GIoU defines 2D anchors and 3D anchors.

Choosing two arbitrary convex shapes: $A, B \subseteq S \in R^n$. For A and B, find the smallest enclosing convex object C, where $C \subseteq S \in R^n$,

$$\text{IoU} = \frac{|A \cap B|}{|A \cup B|}\,, \tag{1}$$

$$\text{GIoU} = \frac{|A \cap B|}{|A \cup B|} - \frac{|C - (A \cap B)|}{|C|}\,. \tag{2}$$

In conclusion, GIoU can be a proper substitute for IoU in all performance measures used in 2D and 3D computer vision tasks.

3.3. Network Structure

As we can see in Fig. 2, we complete the 3D detection by adding geometric moments in the VFE layer of SECOND detector. Then import k 2D detections and n 3D detections into the CLOCs network. The camera and LIDAR have been calibrated, so an object correctly detected by both the 2D and 3D detectors will have an identical bounding box in the image plane, while negatives are unlikely to have the same bounding boxes.

In CLOCs network, 2D and 3D detection candidates are transformed into a set of consistent joint detection candidates. Then use a 2D CNN to process the nonempty elements in the sparse tensor. Finally,

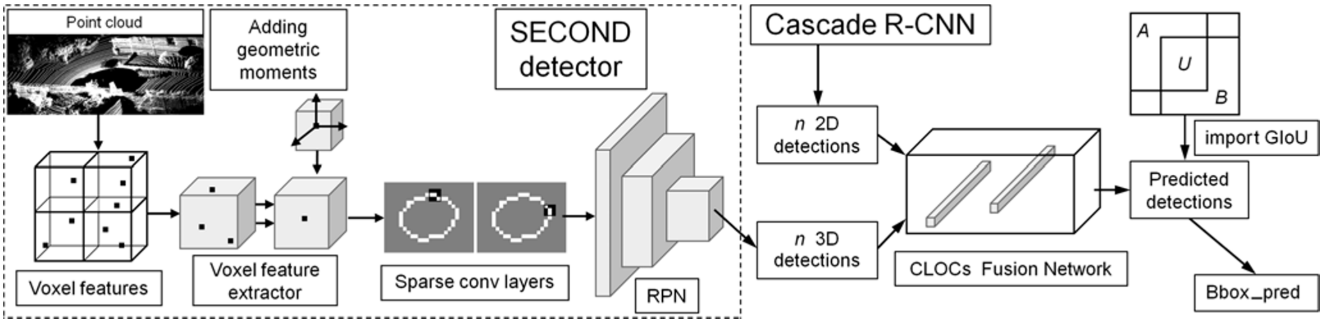

Fig. 2. GIoU-CLOCs Fusion Network Architecture. First, add a polynomial function to the point cloud as input in the SECOND detector. Then put k 2D detections and n 3D detections into the fusion network. Finally, import GIoU as a metric to filtrate candidates.

the probability score map is calculated by max-pooling using the processed tensors, as follows:

$$P_i^{\mathrm{2D}} = \left[|x_{i1}, x_{i2}, y_{i1}, y_{i2}|, s_i^{\mathrm{2D}} \right], \tag{3}$$

$$P_i^{\mathrm{3D}} = \left[|h_i, w_i, l_i, x_i, y_i, z_i, \theta_i|, s_i^{\mathrm{3D}} \right]. \tag{4}$$

According to [7], P_i^{2D} is the set of all k detection candidates in one image, while P_i^{3D} is the set of all n detection candidates in one LIDAR scan. P_i^{2D} and P_i^{3D} are confidence scores generated from 2D and 3D bounding boxes in the fusion network. As for our training and testing data set, using the calibration parameters of the camera and LIDAR, the 3D bounding box in LIDAR coordinates can be accurately projected onto the image plane. After obtaining the candidates through the fusion network, we import GIoU to filter the correct overlapped bounding boxes.

4. Experimental Results

4.1. Data Set

We evaluated the trained model on the KITTI data set [12]. The data set is used to evaluate the performance of computer vision techniques, and the 3D data are collected by the Velodyne HDL-64E Laser scanner. These data are from cities, villages, and streets. It is very similar to the common medical transportation environment. Data of each frame contains pedestrians and vehicles in different scenes, as well as various degrees of occlusion and truncation. There are 7,481 training samples, 7,518 testing samples, and 80,256 tagged objects.

4.2. Analysis of the Detection Results

Average precision (AP) is one of the important indices in the object detection. According to [5], we set the GIOU threshold to 0.5. Calculating the value of the predicted bounding box and ground-truth under the GIOU definition; then, if the value is bigger than 0.5, then regard it as a true positive, otherwise, it is a false positive. The percentage of the number of true positives in all predicted bounding boxes is defined as recall, and the percentage of the number of true positives in all ground truth is defined as precision. Finally, calculating all precision of average in the different recalls we can obtain AP.

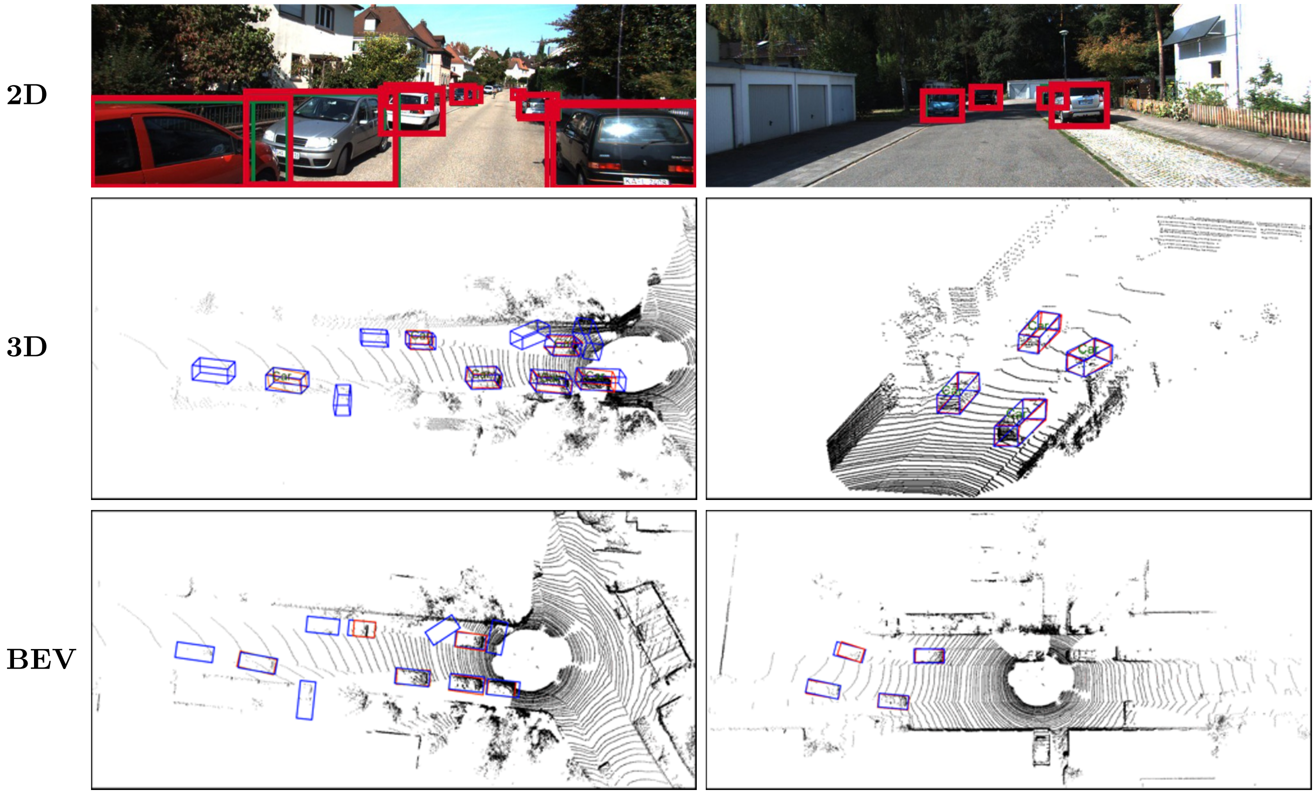

Fig. 3. The results of 2D detection, 3D detection, and BEV detection by GIoU-CLOCs, where green and blue boxes represent detections, and red boxes represent the ground truth.

Table 1. The Results of Most Open-Source and Well-Known Single Sensor Detection Networks Based on KITTI.

Detector	2D AP(%)			3D AP(%)			BEV AP(%)		
	easy	moderate	hard	easy	moderate	hard	easy	moderate	hard
point pillars	94.00	91.19	88.17	84.69	74.04	68.64	90.07	86.56	82.81
Voxel R-CNN	96.49	95.11	92.45	90.90	81.62	77.06	94.85	88.83	86.13
SECOND	96.44	95.79	90.55	87.44	79.46	73.97	92.01	88.98	83.67
GIoU-CLOCs	99.15	95.29	92.64	92.13	82.63	77.53	95.96	91.99	89.04

Object detection experiment: We test our models on KITTI testing samples. Some samples are either visible, semi-occluded, fully occluded, or truncated; according to this, the KITTI authorities divided the samples into three situations: easy, moderate, and hard. Different situations have different average precision (AP). In the KITTI data set, some smaller or less obvious targets may not be labeled by the KITTI authorities. As shown in Fig. 3, our network is able to detect unmarked vehicles. As we can see in Table 1, our network achieves superior performance in 2D, 3D, and BEV compared with the single sensor detection networks [15, 16, 20], our network achieves superior performance on the 2D, 3D, and bird's eye view (BEV) AP in three situations. In the easy situation, it was higher by about 3%

to 8%, while getting some improvements in the moderate and hard situations. Compared with fusion networks [7, 17–19] in Table 2, our network gets the highest accuracy in almost every situation.

Table 2. The Results of Most Open-Source and Well-Known Fusion Networks Based on KITTI.

Detector	2D AP, %			3D AP, %			BEV AP, %		
	easy	moderate	hard	easy	moderate	hard	easy	moderate	hard
Point painting	98.39	92.58	89.71	82.11	71.70	67.08	92.45	88.11	83.36
MV3D	96.47	90.83	78.63	74.97	63.63	54.00	86.62	78.93	69.80
EPNet	96.15	96.44	89.99	89.81	79.28	74.59	94.22	88.47	83.69
CLOCs	96.77	96.07	91.11	89.16	82.28	77.23	92.91	89.48	86.42
GIoU-CLOCs	99.15	95.29	92.64	92.13	82.63	77.53	95.96	91.99	89.04

Ablation study: In order to verify the effect of each module, in this study, we conducted ablation experiments on KITTI. As we can see in Table 3, removing geometric moments has little effect on the 2D detection, but the 3D and BEV detections are significantly improved. Proving this module is necessary to increase geometric moments. When we only imported GIoU, there were some improvements in the three models, but the improvements were not significant. From the ablation experiment, we can see that each module plays an important role and complements each other to achieve the best detection effect of the network.

Table 3. The Results of Ablation Study Based on KITTI.

Import methods applied		2D AP, %	3D AP, %	BEV AP, %
geometric moments	GIoU			
–	–	96.77	89.16	92.91
√	–	96.74	91.89	94.92
–	√	98.94	90.24	93.28
√	√	99.15	92.13	95.96

Different from normal 2D detection, the autonomous driving system has higher requirements for accuracy. It needs not only to quickly identify objects in the surrounding environment, but also make precise positioning of objects in three-dimensional space. So we should pay more attention to 3D and BEV AP. Compared with CLOCs and SECOND detector, as shown in Fig 4, our network reaches the highest AP among the three networks, especially in the BEV detection. However, in 2D and 3D detections, the results of the moderate and hard situations were almost equal to CLOCs. Due to the type differences of the samples, they are either fully occluded or truncated. When using GIoU to filter the correct candidates, the smallest enclosing convex object C may be bigger than the truth value.

As we can see in Fig. 5, the graphs demonstrate BEV AP in three situations. The AP in GIoU-CLOCs falls from 95.96% to 89.04%, while CLOCs is down about 6.5%, and SECOND is down about 8%. Other detectors drop about 7.5% to 16%, the AP of GIoU-CLOCs has high values and tends to be more stable in all situations. This proves that our network has excellent robustness.

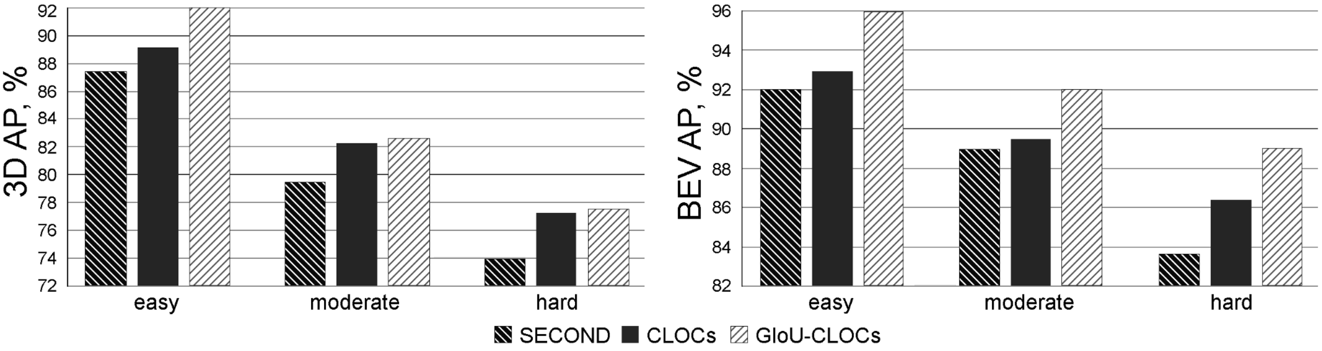

Fig. 4. Histograms of three detection networks in 3D and BEV. We set results of the SECOND as the benchmark to compare with the results of CLOCs and our GIoU-CLOCs.

5. Conclusions

In this paper, we proposed a new fusion network: GIoU-CLOCs, which expanded 3D coordinates by increasing geometric moments, and used GIoU as a metric for bounding box regression. It is an architectural network designed to improve the capability of selecting 2D and 3D candidates and append the informative features to 3D coordinates. The experimental results show the effectiveness and generality of the proposed approach. The AP increased by about 3% compared with CLOCs in BEV. Compared with SECOND, the accuracy of 3D and BEV AP is significantly increased by about 4%. These results prove that the proposed approach can obtain high detection accuracy. For some situations, our network got insufficient improvements. We will consider to use new metrics for bounding box regression in future work. We hope these novel viewpoints to provide insights into future work on the object detection field.

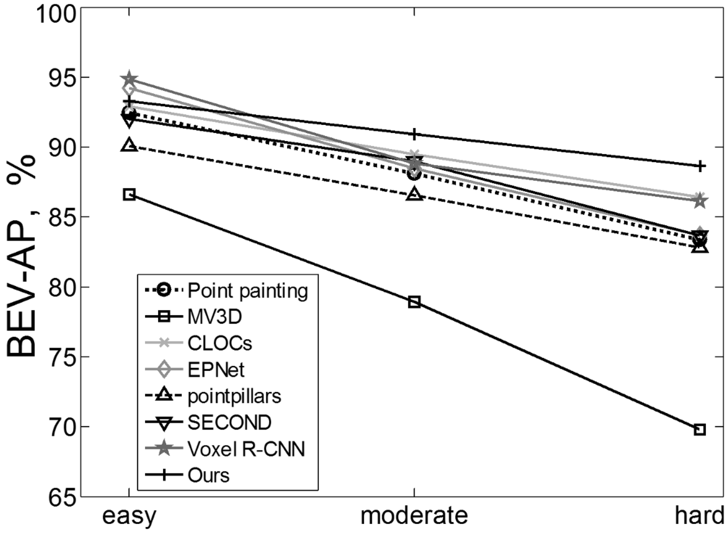

Fig. 5. The results of most open-source and well-known networks in the BEV AP in different situations.

Acknowledgments

This work is funded by the International Cooperation Foundation of Jilin Province under Grant No. 20210402074GH. This work is also supported by the 111 Project of China Nos. D21009 and D17017.

References

1. C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, “YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors,” arXiv:2207.02696 (2022).

2. R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition, IEEE (2014), pp. 580–587; DOI: 10.1109/CVPR.2014.81
3. R. Girshick, "Fast R-CNN," in Proceedings of the IEEE International Conference on Computer Vision, IEEE (2015), pp. 1440–1448; DOI: 10.1109/ICCV.2015.169
4. S. Ren, K. He, R. Girshick, and J. Sun, *IEEE Trans. Pattern Anal. Mach. Intell.*, **39**, 1137 (2017); DOI: 10.1109/TPAMI.2016.2577031
5. H. Rezatofighi, N. Tsoi, J. Y. Gwak, et al., "Generalized intersection over union: A metric and a loss for bounding box regression," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, IEEE (2019), pp. 658–666; DOI: 10.1109/CVPR.2019.00075
6. M. Joseph-Rivlin, A. Zvirin, and R. Kimmel, "Momen(e)t: flavor the moments in learning to classify shapes," in Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, IEEE (2019); DOI: 10.1109/iccvw.2019.00503
7. S. Pang, D. Morris, and H. Radha, "CLOCs: camera-LiDAR object candidates fusion for 3D object detection," IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE (2020), pp. 10386–10393; DOI: 10.1109/IROS45743.2020.9341791
8. V. A. Sindagi, Y. Zhou, and O. Tuzel, "MVX-net: multimodal voxelnet for 3D object detection," 2019 International Conference on Robotics and Automation (ICRA), IEEE (2019), pp. 7276–7282; DOI: 10.1109/ICRA.2019.8794195
9. 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, IEEE (2018), pp. 4490–4499; DOI: 10.1109/cvpr.2018.00472
10. H. Su, S. Maji, E. Kalogerakis, et al., "Multi-view convolutional neural networks for 3D shape recognition," in Proceedings of the IEEE International Conference on Computer Vision, IEEE (2015), pp. 945–953; DOI: 10.1109/ICCV.2015.114
11. B. Jiang, R. Luo, J. Mao, et al., "Acquisition of localization confidence for accurate object detection," in Proceedings of the European Conference on Computer Vision (ECCV), (2018), pp. 784–799; DOI: 10.1007/978-3-030-01264-9_48
12. A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? The KITTI vision benchmark suite," 2012 IEEE Conference on Computer Vision and Pattern Recognition, IEEE (2012), pp. 3354–3361; DOI: 10.1109/cvpr.2012.6248074
13. A. Mousavian, D. Anguelov, J. Flynn, et al., "3D bounding box estimation using deep learning and geometry," in Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, IEEE (2017), pp. 7074–7082; DOI: 10.1109/cvpr.2017.597
14. Z. Cai and N. Vasconcelos, *IEEE Trans. Pattern Anal. Mach. Intell.*, **43**, 1483 (2019); DOI: 10.1109/tpami.2019.2956516
15. Y. Yan, Y. Mao, and B. Li, *Sensors*, **18**, 3337 (2018); DOI: 10.3390/s18103337
16. A. H. Lang, S. Vora, H. Caesar, et al. "Pointpillars: fast encoders for object detection from point clouds," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, IEEE (2019), pp. 12697–12705; DOI: 10.1109/cvpr.2019.01298
17. S. Vora, A. H. Lang, B. Helou, et al., "Pointpainting: sequential fusion for 3D object detection," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, IEEE (2020), pp. 4604–4612; DOI: 10.1109/cvpr42600.2020.00466
18. T. Huang, Z. Liu, X. Chen, et al., "Epnet: enhancing point features with image semantics for 3D object detection," European Conference on Computer Vision, Springer, Cham (2020), pp. 35–52; DOI: 10.1007/978-3-030-58555-6_3
19. X. Chen, H. Ma, J. Wan, et al., "Multi-view 3D object detection network for autonomous driving," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE (2017); pp. 1907–1915; DOI: 10.1109/CVPR.2017.691
20. J. Deng, S. Shi, P. Li, et al., "Voxel R-CNN: towards high performance voxel-based 3D object

detection," in Proceedings of the AAAI Conference on Artificial Intelligence, **35(2)**, 1201 (2021); DOI: 10.1609/aaai.v35i2.16207

21. C. R. Qi, H. Su, K. Mo, et al., "Pointnet: deep learning on point sets for 3D classification and segmentation," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, (2017), pp. 652–660; DOI: 10.1109/CVPR.2017.16
22. A. M. Bronstein, M. M. Bronstein, R. Kimmel, "Rock, paper, and scissors: extrinsic vs. intrinsic similarity of non-rigid shapes," 2007 IEEE 11th International Conference on Computer Vision, IEEE (2007), pp. 1–6; DOI: 10.1109/iccv.2007.4409076
23. Y. Li, A. W. Yu, T. Meng, et al., "Deepfusion: lidar-camera deep fusion for multi-modal 3D object detection," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, IEEE (2022), pp. 17182–17191; DOI: 10.1109/cvpr52688.2022.01667