ASD-Conv: Monocular 3D object detection network based on Asymmetrical Segmentation Depth-aware Convolution

Xingyuan Yu RickyYXY@whu.edu.cn
Neng Du 2019282110109@whu.edu.cn
Ge Gao* gaoge@whu.edu.cn
Fan Wen wenfan@whu.edu.cn
National Engineering Research Center for Multimedia Software
WuHan University
WuHan, China

Editors: Vineeth N Balasubramanian and Ivor Tsang

Abstract

In the field of 3D object recognition, monocular 3D recognition technology is a valuable recognition technology. Compared with binocular technology and lidar technology, its cost is lower. In this paper, based on the existing monocular 3D recognition network, we propose an asymmetrical segmentation depth-aware network: ASD-Conv Network, which is used to better obtain the depth information of monocular images, so as to obtain better recognition results. Compared with other monocular recognition networks, ASD-Conv network performs special segmentation on the image, which can better obtain the depth distribution of the image, and has made a good breakthrough and improvement in the image recognition tasks of 2D, BEV and 3D. The improved algorithm proposed in this paper can improve the detection accuracy while maintaining a certain real-time performance. Experimental results show that compared with the current model, the proposed monocular 3D object detection algorithm based on D-ASDConv has an average improvement rate of 2.82%(AP) in large object detection and the highest average improvement rate of 2.01%(AP) in small object detection on Kitti dataset. The algorithm can effectively learn more advanced features of spatial perception, and the detection results of monocular images are more accurate.

Keywords: Monocular image; 3D object detection, Depth estimation, Asymmetrical segmentation depth-aware convolutio, Self-knowledge distillation

1. Introduction

Object detection technology is currently widely used in many different fields, especially in the field of autonomous driving in recent years. In the field of autonomous driving, 3D object detection (Chu et al. (2018); Guerry et al. (2017); Liang et al. (2018); Qi et al. (2018); Shi et al. (2019); Yang et al. (2018)) is the focus of attention. At present, monocular images (Chen et al. (2016, 2015); Mousavian et al. (2017); Xu and Chen (2018)) have been widely used to estimate 3D detection frames during autonomous driving development (Behl et al. (2017); Chen et al. (2018); Geiger et al. (2012)), because simple 2D detection frames cannot be used for planning and control of 3D space. However, due to the lack of depth information, the monocular image is much more difficult to perform 3D detection relative to the sparse depth data provided by the lidar sensor.

The traditional monocular detection algorithm on the KITTI (Geiger et al. (2012)) data set has a great difference in the accuracy of the detection results of three types of objects (such as cars, pedestrians and cyclists), and the results are not ideal. Therefore, in the face of multiple types of object scenarios, how to improve the detection accuracy of different objects and make the detection accuracy of different types of objects at a higher level is a current challenge.

In order to improve the monocular method, we have proposed a separate end-to-end multi-class 3D object detection network, successfully using 2D and 3D space with shared anchors and classification objects, and established a unified network architecture. First of all, we take into account the rule of the camera imaging perspective, we segment the convolution layer according to special proportions of horizontal stripes, and the accuracy of the result is significantly improved. With this idea, we then further extended to vertical strip segmentation. In order to make the network parameters converge faster and more accurately, we introduced the idea of a distillation mechanism and added a distillation structure to the detection head to solve the problem of consuming more time and equipment resources caused by the huge network structure.

In summary, our contributions are as follows:

- We use the 2D and 3D anchor templates of the object objects and segment the new convolution layer according to a certain proportion of horizontal stripes.
- We extend the vertical strip segmentation based on a certain proportion of horizontal strip segmentation, and conduct experiments of horizontal and vertical fusion detection.
- We optimize the network model with Distillation method, using the output weight model of the trained large model to train the small model with a more concise network structure.

2. Related Work

As mentioned above, simple 2D detection frames cannot be used for planning and control in 3D space; although LiDAR-based methods generally perform well in various 3D tasks, LiDAR-based methods largely depend on the availability of depth information generated from lidar points, so these methods are not suitable for camera applications; while pure image-based 3D detection lacks reliable depth information. One of our goals is to better obtain these depth information, which is also the main purpose of our monocular 3D detection algorithm. Object detection algorithms have emerged endlessly over the years, but the most classic is the regional proposal network (RPN), which is extremely effective. Therefore, we use the RPN-based anchor frame mechanism for monocular multi-class 3D detector tasks.

In terms of extracting image features, we have taken a series of useful measures that we mentioned in the previous paragraph, inspired by the rule of the camera imaging perspective. This paper uses a single network trained with only a 3D box to generate both 2D and 3D object proposal regions.

In the actual training process, our model directly predicts the 3D parameters and optimizes θ. With the above solid network foundation and the inspiration from the results of many experiments, we have further adopted a series of transformation measures, including distillation mechanism, etc., only for better detection effect in the dataset.

3. Methodology

3.1. Pinhole imaging model

Pinhole imaging model is an ideal camera model widely used. The imaging principle is shown in Fig. 1.

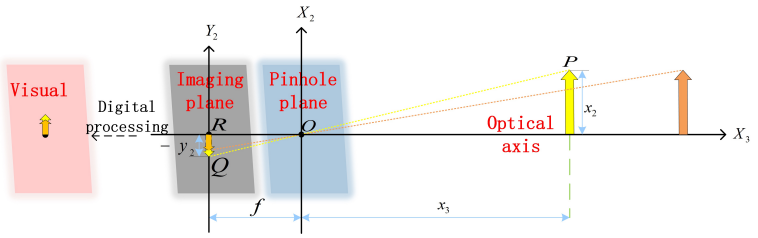

Figure 1: Schematic diagram of pinhole imaging.

In the aperture imaging "camera" shown in the figure above, the conversion relationship between the real world coordinate system and the imaging plane coordinate system can be obtained:

$$-X(imaging) = f * \frac{X}{Z} \tag{1}$$

Where f is the focal length of the camera, X is the height of the object, and Z is the distance from the camera to the object.

In the process of the camera collecting monocular images, the plane view formed by the X-Z axis in the camera coordinate system is shown in Fig. 2, which is the central plane in the imaging principle of the camera in Fig. 1.

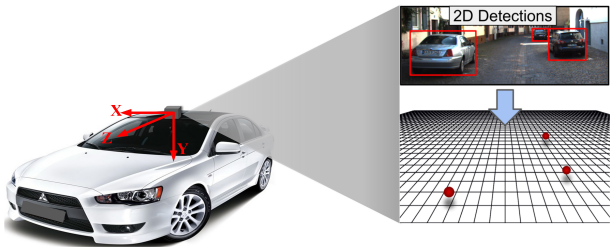

Figure 2: Plan view in camera coordinate system.

Based on the above camera model and on the principle of light propagation along the line, it can be found that in the actual autonomous driving scene, the object distance is much larger than the image distance. At the same time, it can be found that for the same object, if it is closer to the central horizontal line in the monocular image after imaging, it means that the object is farther away from the camera and has deeper depth in the actual scene.

3.2. Asymmetric segmentation depth-aware convolution

The monocular image is essentially a projection result of a stereo scene, and only plane information can be obtained after projection. The earliest monocular depth estimation method is based on machine learning (Saxena et al. (2005, 2008)). Although depth estimation methods based on monocular images are emerging in endlessly (Eigen et al. (2014); Gan et al. (2018)), without exception, these methods require additional sub-networks to assist, resulting in the final overall network model being huge and complex. In the M3D-RPN method proposed by Brazil and Liu (2019), the use of depth-aware convolution is mentioned in the perceptual 3D parameter estimation part, so that the network can learn more spatial high-order features.

Based on the M3D-RPN depth-aware convolutional network, this paper proposes a asymmetrical segmentation depth-aware convolutional network based on the imaging law representation model of 3D space and the principle of small hole imaging. This approach realizes the estimation of the depth information of the monocular image, thereby improving the 3D target detection effect. Fig. 3 shows the distance relationship between each position in the monocular image and the camera. At the same time, in conjunction with the central plan view of Fig. 2, it can be obtained that the depth distribution of the objects in the image is not uniform.

Figure 3: Depth distribution in monocular images.

Therefore, during the detection, if the depth convolution is still performed according to the uniform segmentation mode, the performance of the calculated depth-like information will be degraded. This paper proposes the asymmetrical segmentation deep convolutional network ASD-Conv to try to solve the problems raised above. This technology divides the different positions of the feature map in different proportions, and realizes the efficient extraction of the depth information in the monocular image. Fig. 4 shows the design diagram of the ASD-Conv network in this article.

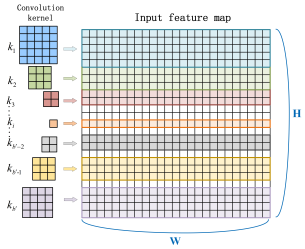

Figure 4: Asymmetric segmentation depth convolution network.

3.3. Converged network structure

The backbone network based on ASD-Conv is DenseNet-121 (Huang et al. (2017)). As the backbone network, DensetNet-121 uses a large number of layer jump connections, which can effectively reduce network parameters and alleviate the problem of gradient disappearance.

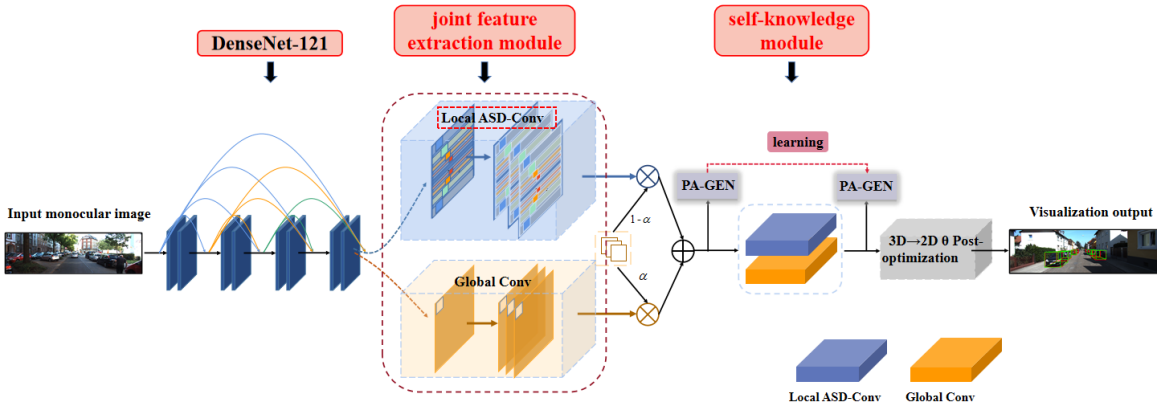

Figure 5: Asymmetric segmentation depth convolution network mode.

Two parallel branches are connected at the end of the backbone network. The following branch is the global convolution branch, and its convolution kernel space is shared. The above branch is a local convolution branch, using asymmetrical segmentation depth-aware convolution ASD-Conv. Here, the two parallel branches are collectively referred to as a joint feature extraction module. The joint feature extraction module includes two stages. The ReLU nonlinear activation function is used for the feature maps that have passed through the first stage of the joint feature extraction module to generate 512 feature maps. Each feature map is connected with 12 outputs:$c, [t_x, t_y, t_w, t_h]_{2D}, [t_x, t_y, t_z]_P, [t_w, t_h, t_l, t_\theta]_{3D}$. The outputs of the second stage of the joint feature extraction module are collectively referred to as O_{globle} and O_{local}. To leverage the depth-aware and spatial-invariant strengths, we fuse each output using a learned attention (after sigmoid) applied for 12 output parameters represented by the index i=1,2,3,...,12 respectively. The weighting calculation formula is as Eqn. 2.

$$O^i = O^i_{global} \cdot \alpha_i + O^i_{local} \cdot (1 - \alpha_i) \tag{2}$$

Since the joint feature extraction module introduces the asymmetrical segmentation depth-aware convolution ASD-Conv, the network can use the connection between the depth information and the position on the monocular image to learn more advanced features of spatial perception. At the same time, the joint effect of the local ASD-Conv branch network and the global branch network can further realize the efficient extraction of monocular image features. At the same time, the joint effect of the local ASD-Conv branch network and the global branch network can further realize the efficient extraction of monocular image features.

3.4. Loss calculation

When calculating the loss function, all the parameters included in the calculation range are shown in Fig. 6. These parameters are used to construct the 2D, 3D, and BEV recognition frame of the object.

Calculate a transition between the anchor box and the real box. The conversion formula under 2D conditions is as follows:

$$\begin{aligned} x_{2D}^{'} &= x_P + t_{x_{2D}} \cdot w_{2D}, \quad w_{2D}^{'} = exp(t_{w_{2D}}) \cdot w_{2D} \\ y_{2D}^{'} &= y_P + t_{y_{2D}} \cdot h_{2D}, \quad h_{2D}^{'} = exp(t_{h_{2D}}) \cdot h_{2D} \end{aligned} \tag{3}$$

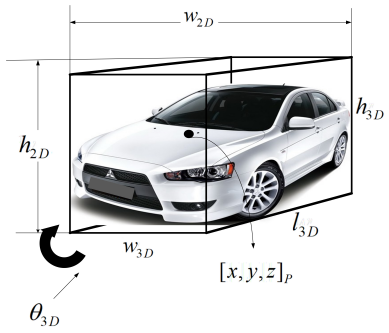

Figure 6: Schematic diagram of parameter definition of box template.

Among them, $[x, y, w, h]_{2D}^{'}$ is used as the output vector of 2D box prediction $b_{2D}^{'}$, and $b_{2D}^{'}$ is the true value. Through the plane coordinate values x and y on the image, the depth information z_P, and the pre-calculated projection matrix P, the three-dimensional coordinate x, y, z of the corresponding point in the three-dimensional space can be obtained. The formula is as follows:

$$\begin{bmatrix} x \cdot z \\ y \cdot z \\ z \end{bmatrix}_P = P \cdot \begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix}_{3D} \tag{4}$$

Then the three-dimensional coordinate information can be obtained. The 3D conversion formula can be analogous to the 2D situation, and the formula is shown in Eqn. 5. This paper uses $[x, y, z, w, h, l, \theta]_{3D}^{'}$ as the detection output vector $b_{3D}^{'}$, where $[x, y, z]_{3D}^{'}$ can be obtained from $[x, y, z]_P^{'}$ using Eqn. 4.

$$\begin{aligned} x_p' &= x_p + t_{x_P} \cdot w_{2D}, \quad w_{3D}' = exp(t_{w_{3D}}) \cdot w_{3D} \\ y_p' &= y_p + t_{y_P} \cdot h_{2D}, \quad h_{3D}' = exp(t_{h_{3D}}) \cdot h_{3D} \\ z_p' &= t_{Z_P} + Z_P, \quad l_{3D}' = exp(t_{l_{3D}}) \cdot l_{3D} \\ \theta_{3D}' &= t_{\theta_{3D}} + \theta_{3D} \end{aligned} \tag{5}$$

The loss value of the calculation model can be calculated by Eqn. 3 and Eqn. 5. In terms of loss function, the classification loss, 2D box loss, and 3D box loss are used as components.

$$L_{pre} = L_c + \lambda_1 L_{b_{2D}} + \lambda_2 L_{b_{3D}} \tag{6}$$

The expressions of L_C, $L_{b_{2D}}$, and $L_{b_{3D}}$ are as follows:

$$L_c = -log(\frac{exp(c_\tau)}{\sum_i^{n_c} exp(c_i)}) \tag{7}$$

$$L_{b_{2D}} = -log\left(IOU\left(b'_{2D}, \hat{b}_{2D}\right)\right) \tag{8}$$

$$L_{b_{3D}} = SmoothL_1\left(b_{3D}, \hat{g}_{3D}\right) \tag{9}$$

3.5. Improved converged network architecture

In the convolutional network, the distillation structure is often used in the backbone network part (Hou et al. (2019)). This paper introduces it to the asymmetrical segmentation convolutional network model.

But different from other traditional distillation structure network, this paper introduces distillation structure in the detection head. Its specific structure is shown in Fig. 7.

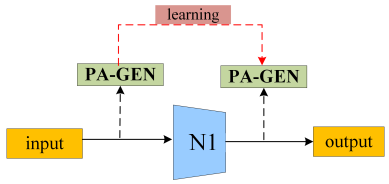

Figure 7: Self-knowledge distillation mechanism SKD module.

The input of the distillation part is the output of ASD-Conv, that is, the output vectors b_{cls}, b_{2D}, and b_{3D} of ASD-Conv. During the entire distillation process, the input and output format does not change, so the output of the distillation network can be directly used as the final prediction result of the network. In the structure of self-knowledge distillation mechanism, this article only uses a teacher-student training pair, namely $N_1 - N_{Input}$. Therefore, the corresponding distillation loss function is:

$$L_{distill} = L_2\left(N_1, N_{Input}\right) \tag{10}$$

Finally, the loss of the distillation network and the previous loss function are combined and superimposed, and the loss function of the ASD-Conv equipped with the distillation structure can be obtained:

$$L_{final} = L_{distill} + L_{pre} \tag{11}$$

The distillation network used in this paper is not a multi-layer network structure, but a single-layer light distillation structure. It will not cause more burden on the overall calculation, and can improve the convergence effect of the detection network, and achieve better detection results in the same number of iterations.

3.6. More details

Similarly, we also get the idea in the horizontal direction and using the distribution of depth information in the horizontal direction. In order to use the depth distribution in the horizontal direction, we need to perform vertical segmentation convolution. The convolution operation is performed on the picture from the horizontal and vertical directions at the same time, and then combined to obtain the final calculation result. Based on the above ideas, we have proposed the Horizontal and Vertical convolutional Network (H plus V Net). The structure of the network is shown in Fig. 8 below.

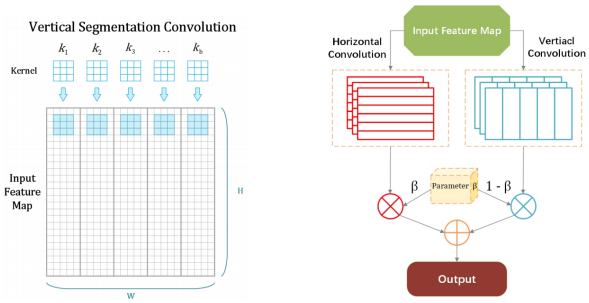

Figure 8: The fusion of Horizontal and Vertical Network.

In H plus V Net, the network input is exactly the same as ASD Network. The situation of each segment is the same as the horizontal direction, and an independent convolution kernel is responsible for the convolution operation. We expect that by fusing horizontal and vertical depth information, a better detection effect can be achieved than simply using horizontally evenly divided convolutions plus global convolutions.

4. Experiments

In this paper, we mainly use data of cars, pedestrians and cyclists in urban scenes (Cai et al. (2016); Cai and Vasconcelos (2018); Liu et al. (2018); Ren et al. (2017); Yang et al. (2016); Zhou and Yuan (2018)). We divided the dataset into three levels of easy, mod, and hard according to the degree of occlusion of samples in the data set for scientificity of the experiment. At the same time, pictures used in this experiment were divided into two parts randomly and the experiment was conducted on the two sub-data sets. To provide fair comparisons between different detectors, we measure the inference times on the same machine with a Titan GPU and an Intel Core i7-5930K CPU.

4.1. Comparison with state-of-the-art methods.

In order to verify the function of indifferent choice segmentation network and add method self-knowledge smart device on this basis, this method is compared with multiple 3D object detection based on deep learning monocular image Mono3D (Chen et al. (2016)), Deep3DBox (Mousavian et al. (2017)), Multi-Fusion (Xu and Chen (2018)), M3D-RPN (Brazil and Liu (2019)), SMOKE (Liu et al. (2020)) and RT-M3D (Li et al. (2020)). The test results of each mainstream method on the KITTI data set are shown in Tab. 1-3. It

can be seen from the final results that the method in this paper achieves the optimal value on the KITTI data set, which proves the superiority of the algorithm in this paper.

Model	$IOU \geq 0.5[car/pedestrian]$			$IOU \geq 0.7[car/pedestrian]$		
	Easy	Mod	Hard	Easy	Mod	Hard
M3D-RPN	48.96/-	39.57/-	33.01/-	20.27/11.42	17.06/11.28	15.21/10.34
+SKD3	41.41/12.68	33.05/13.09	27.21/11.98	20.20/10.26	17.22/9.50	15.47/9.50
+SKD2	44.25/13.53	35.12/13.35	28.35/12.08	21.66/10.72	17.30/10.18	15.73/9.53
+SKD1	47.11/14.50	37.22/14.14	30.93/13.11	22.36/11.46	18.63/11.19	16.02/9.58

Table 1: 3D Box comparison experiment of two types of objects: car and pedestrian

In this article, three types of SKD models are added to M3D-RPN. On the 3D Box task, their experimental comparison results are shown in Tab. 1. The accuracy data in the experimental results are in percentage (%). The above table shows that the performance of the self-knowledge distillation model is not proportional to its scale. The experimental results show that the first-order SKD structure is better than the second-order and third-order SKD structures in both like large object car and like small object pedestrian.

Fig. 9 shows two types of ASD-Conv network structures used in this algorithm, namely ASD-ConvI and ASD-ConvII. In Fig. 9(a), Asymmetric Segmentation Depth convolutional Network with a ratio of 1: 1: 1: 1: 2: 2: 3: 4: 5: 6: 6. Here $b^{'} = 11$, which means there are 11 kernels here. In Fig. 9(b), Asymmetric Segmentation Depth convolutional Network with a ratio of 6: 1: 1: 1: 1: 1: 2: 2: 2: 2: 3: 3: 3: 4. Here $b^{'} = 14$, meaning 14 kernels here.

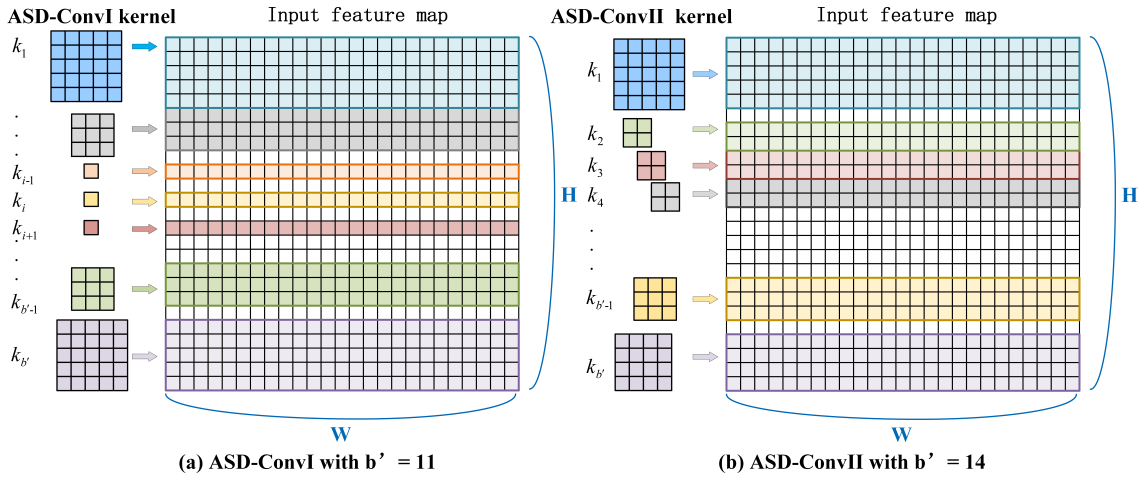

Figure 9: Two types of ASD-Conv structure.

Although the data in Tab. 1 shows that the effect of introducing SKD into M3D-RPN algorithm alone is not significantly improved. However, the combination of Tab. 2 and Tab. 3 experiments proved the effectiveness of ASD-Conv. That is, the monocular feature map is divided into different proportions of strips, and then the feature maps on each strip are subjected to independent convolution operation, which can indeed reflect the relationship between the object depth and its position on the image. The introduction of the SKD

Method	AP_{2D}			AP_{BEV}			AP_{3D}		
	Easy	Mod	Hard	Easy	Mod	Hard	Easy	Mod	Hard
M3D-RPN	51.47	42.77	34.91	2.90	9.09	9.09	2.22	9.09	9.09
ASD-ConvI	51.61	42.90	42.78	10.15	9.09	9.09	10.03	9.09	9.09
D-ASDConvI	50.13	33.91	33.89	2.59	1.38	1.48	2.40	1.38	1.48
ASD-ConvII	60.63	51.82	43.67	2.53	1.30	3.03	0.78	0.91	3.03
D-ASDConvII	59.59	43.50	43.50	3.17	10.38	9.09	2.16	9.09	9.09
H plus V Net	58.89	50.99	43.02	1.88	1.00	1.04	0.80	0.80	0.86

Table 2: Cyclist Detection. Comparison on the cyclist class under the condition of $IOU \geq 0.5$ in the validation set.

structure can amplify and strengthen the effect of ASD-Conv. Moreover, the second ASD-ConvII segmentation network is more suitable for the detection of small objects such as cyclist and large objects such as car than the first ASD-ConvI segmentation network. On the same data set, the AP of the asymmetric segmentation convolution method and its improved method has been greatly improved compared with the original model. This shows that the convolutional network used in this article is effective and scientific in extracting the depth information of monocular images. In the recognition task, the network model with the self-knowledge distillation mechanism is better than the network model without this mechanism, and its performance is more advantageous and the effect is obvious. Although in terms of inference speed, the D-ASDConv-based algorithm proposed in this paper is slightly lower than the baseline algorithm. But in terms of detection accuracy, whether in the BEV task or in the 3D box task, the detection accuracy based on D-ASDConv algorithm is higher than other detection algorithms on the whole, and its accuracy performance is better than the state-of-the-art monocular image detection algorithm RT-M3D. The experimental results show that ASD-Conv is effective. In addition, after introducing SKD mechanism, the performance of D-ASDConv has been further improved. Compared with the baseline model, D-ASDConv has an average improvement rate of 2.82%(AP) in large object detection and the highest average improvement rate of 2.01%(AP) in small object detection on Kitti dataset.

4.2. Comparison of 3D test results.

Monocular images are collected by the front camera of the car. Fig. 10 shows the visual comparison results of 3D detection of some images in Kitti scene between the detection algorithm based on D-ASDConv network and the baseline algorithm. There are eight groups of graphs in Fig. 10. The top of each group is the result graph of baseline algorithm experiment, and the bottom is the result graph of this algorithm. Among them, the red box represents the GT bounding box, the green box represents the car prediction bounding box, the blue box represents the pedestrian prediction bounding box, and the yellow box represents the cyclist prediction bounding box.

It can be seen that the detection results of baseline algorithm are not ideal, and there are many phenomena of missed judgment and wrong classification. In this paper, a 3D object detection algorithm based on asymmetrical segmentation and depth sensing convolution is

Method	AP_{BEV}			AP_{3D}			Time(s)
	Easy	Mod	Hard	Easy	Mod	Hard	
Mono3D	5.22	5.19	4.13	2.53	2.31	2.31	-
Deep3DBox	9.99	7.71	5.30	5.85	4.10	3.84	1.5
Multi-Fusion	19.20	12.17	10.89	7.85	5.39	4.73	-
M3D-RPN	26.86	21.15	17.14	20.40	16.48	13.34	0.16
SMOKE	19.99	15.61	15.28	14.76	12.85	11.50	0.03
RT-M3D	25.56	22.12	20.91	19.47	16.29	15.57	0.055
ASD-ConvI	22.83	18.49	14.48	16.28	13.90	10.75	0.19
D-ASDConvI	27.36	21.62	17.64	17.03	13.78	9.93	0.24
ASD-ConvII	27.29	20.98	17.18	22.55	16.55	14.40	0.19
D-ASDConvII	30.26	23.36	18.90	25.30	18.49	16.00	0.23
H plus V Net	27.79	22.10	18.26	21.12	16.43	14.34	0.14

Table 3: Car Detection.Comparison of several networks on the car class under the condition of $IOU \geq 0.7$ in the validation set

proposed, which achieves good results and can detect most objects in monocular images accurately. In general, the proposed 3D object detection network based on asymmetric segmentation depth sensing convolution has more advantages in Kitti dataset.

Figure 10: Comparison of 3D detection results.

5. Conclusion

In this paper, we adopted our Asymmetry Segmentation Depth convolution network based on the principle of perspective in real world. Introducing proper segmentation proportions gives our network stronger depth acquisition capabilities. In addition to the main work, we introduced vertical strip segmentation and combined horizontal convolution with it for better performance in detection. And we also introduced the structure of the Distillation network at the detection head part of ASD Net in order to improve detection ability and accuracy through accelerating model's convergence. Overall, although the diverse scenes bring many challenges, the monocular 3D object detection algorithm based on asymmetrical segmentation proposed in this paper can still achieve relatively good results in most scenes.

References

Aseem Behl, Omid Hosseini Jafari, Siva Karthik Mustikovela, Hassan Abu Alhaija, Carsten Rother, and Andreas Geiger. Bounding boxes, segmentations and object coordinates: How important is recognition for 3d scene flow estimation in autonomous driving scenarios? In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2574–2583, 2017.

Garrick Brazil and Xiaoming Liu. M3d-rpn: Monocular 3d region proposal network for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9287–9296, 2019.

Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: Delving into high quality object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6154–6162, 2018.

Zhaowei Cai, Quanfu Fan, Rogerio S Feris, and Nuno Vasconcelos. A unified multi-scale deep convolutional neural network for fast object detection. In *European conference on computer vision*, pages 354–370. Springer, 2016.

Xiaozhi Chen, Kaustav Kundu, Yukun Zhu, Andrew G Berneshawi, Huimin Ma, Sanja Fidler, and Raquel Urtasun. 3d object proposals for accurate object class detection. In *Advances in Neural Information Processing Systems*, pages 424–432. Citeseer, 2015.

Xiaozhi Chen, Kaustav Kundu, Ziyu Zhang, Huimin Ma, Sanja Fidler, and Raquel Urtasun. Monocular 3d object detection for autonomous driving. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2147–2156, 2016.

Yiping Chen, Jingkang Wang, Jonathan Li, Cewu Lu, Zhipeng Luo, Han Xue, and Cheng Wang. Lidar-video driving dataset: Learning driving policies effectively. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5870–5878, 2018.

Hang Chu, Wei-Chiu Ma, Kaustav Kundu, Raquel Urtasun, and Sanja Fidler. Surfconv: Bridging 3d and 2d convolution for rgbd images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3002–3011, 2018.

David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. *arXiv preprint arXiv:1406.2283*, 2014.

Yukang Gan, Xiangyu Xu, Wenxiu Sun, and Liang Lin. Monocular depth estimation with affinity, vertical pooling, and label enhancement. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 224–239, 2018.

Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3354–3361. IEEE, 2012.

Joris Guerry, Alexandre Boulch, Bertrand Le Saux, Julien Moras, Aurélien Plyer, and David Filliat. Snapnet-r: Consistent 3d multi-view semantic labeling for robotics. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 669–678, 2017.

Yuenan Hou, Zheng Ma, Chunxiao Liu, and Chen Change Loy. Learning lightweight lane detection cnns by self attention distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1013–1021, 2019.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

Peixuan Li, Huaici Zhao, Pengfei Liu, and Feidao Cao. Rtm3d: Real-time monocular 3d detection from object keypoints for autonomous driving. *arXiv preprint arXiv:2001.03343*, 2, 2020.

Ming Liang, Bin Yang, Shenlong Wang, and Raquel Urtasun. Deep continuous fusion for multi-sensor 3d object detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 641–656, 2018.

Wei Liu, Shengcai Liao, Weidong Hu, Xuezhi Liang, and Xiao Chen. Learning efficient single-stage pedestrian detectors by asymptotic localization fitting. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 618–634, 2018.

Zechen Liu, Zizhang Wu, and Roland Tóth. Smoke: single-stage monocular 3d object detection via keypoint estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 996–997, 2020.

Arsalan Mousavian, Dragomir Anguelov, John Flynn, and Jana Kosecka. 3d bounding box estimation using deep learning and geometry. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7074–7082, 2017.

Charles R Qi, Wei Liu, Chenxia Wu, Hao Su, and Leonidas J Guibas. Frustum pointnets for 3d object detection from rgb-d data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 918–927, 2018.

Jimmy Ren, Xiaohao Chen, Jianbo Liu, Wenxiu Sun, Jiahao Pang, Qiong Yan, Yu-Wing Tai, and Li Xu. Accurate single stage detector using recurrent rolling convolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5420–5428, 2017.

Ashutosh Saxena, Sung H Chung, Andrew Y Ng, et al. Learning depth from single monocular images. In *NIPS*, volume 18, pages 1–8, 2005.

Ashutosh Saxena, Min Sun, and Andrew Y Ng. Make3d: Learning 3d scene structure from a single still image. *IEEE transactions on pattern analysis and machine intelligence*, 31 (5):824–840, 2008.

Shaoshuai Shi, Xiaogang Wang, and Hongsheng Li. Pointrcnn: 3d object proposal generation and detection from point cloud. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–779, 2019.

Bin Xu and Zhenzhong Chen. Multi-level fusion based 3d object detection from monocular images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2345–2353, 2018.

Bin Yang, Wenjie Luo, and Raquel Urtasun. Pixor: Real-time 3d object detection from point clouds. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 7652–7660, 2018.

Fan Yang, Wongun Choi, and Yuanqing Lin. Exploit all the layers: Fast and accurate cnn object detector with scale dependent pooling and cascaded rejection classifiers. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2129–2137, 2016.

Chunluan Zhou and Junsong Yuan. Bi-box regression for pedestrian detection and occlusion estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 135–151, 2018.