Asymmetric Two-Stream Networks for RGB-Disparity Based Object Detection

Ruizhi Lu[1,2,3], Jianhuang Lai[1,2,3(✉)], and Xiaohua Xie[1,2,3]

[1] School of Data and Computer Science, Sun Yat-sen University, Guangzhou, China
lurzh3@mail2.sysu.edu.cn, {stsljh,xiexiaoh6}@mail.sysu.edu.cn
[2] Guangdong Key Laboratory of Information Security Technology, Guangzhou, China
[3] Key Laboratory of Machine Intelligence and Advanced Computing, Ministry of Education, Beijing, China

Abstract. Currently, most methods of object detection are monocular-based. However, due to the sensitivity to color, these methods can not handle many hard samples. With the depth information, disparity maps are helpful to get over this problem. In this paper, we propose the asymmetric two-stream networks for RGB-Disparity based object detection. Our method consists of two networks, Disparity Representations Mining Network (DRMN) and Muti-Modal Detection Network (MMDN), to combine RGB and disparity data for more accurate detection. Unlike normal two-stream networks, our model is asymmetric because of the different capacity of RGB and disparity data. We are the first to propose a deep learning based framework utilizing only binocular information for object detection. The experiment results on KITTI and our proposed BPD dataset demonstrate that our method can achieve a significant increase in performance efficiently and get the state-of-the-art.

Keywords: Object detection · Two-stream networks · RGBD data

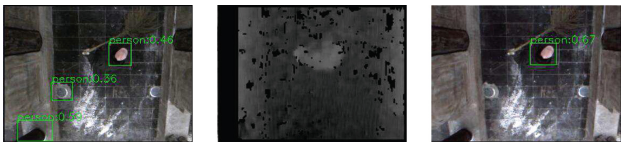
1 Introduction

With the help of deep neural networks, object detection has achieved great progress in recent years. However, in many real-world applications, it is still challenging for object detection to deal with a dramatic variety of illumination, occlusions, viewpoints, and busy backgrounds, etc.

Currently most approaches of object detection are monocular-based [7,8,11,13,15,16,18,19], in other words, they take as input RGB images from single camera. Rich information of color and textures can be extracted from monocular RGB images, and the data only depend on one RGB camera with low cost. Therefore, monocular RGB images are popular in researches on object detection. Nevertheless, monocular-based methods utilizing only RGB data are likely to be mistaken on some hard-negative and hard-positive samples. As illustrated in

J.-H. Lai et al. (Eds.): PRCV 2018, LNCS 11259, pp. 3–15, 2018.
https://doi.org/10.1007/978-3-030-03341-5_1

Fig. 1(a), in RGB image, a lamp has the similar appearance with people's head, so it tends to be mistaken for a pedestrian. In addition, a white hat of the person brings an unusual color of the head, so monocular-based methods recognize him with low confidence due to the sensibility to color. However, if the depth features of objects are known, we can find that lamp's shape is flat, and pedestrian's shape is a paraboloid (Fig. 1(b)), then hard-negative and hard-positive samples can be distinguished, also the confidence of true pedestrian improves, such as Fig. 1(c).



(a) (b) (c)

Fig. 1. An example of pedestrian detection. (a) A lamp is likely to be mistaken for a pedestrian because it looks like a head for the monocular-based methods. In addition, a white hat of the person brings an unusual color of head, so monocular-based methods recognize him with low confidence. But (b) from disparity map we can find that lamp's actual shape is flat, and pedestrian's shape is a paraboloid. According to these, (c) hard-negative and hard-positive samples can be distinguished, also the confidence of true pedestrian improves. Best viewed in color. (Color figure online)

Learning depth features of objects contributes to locating objects accurately. There are two ways to achieve this. On one hand, RGB cameras and extra equipment, such as LIDAR and Microsoft Kinect, can be utilized to get depth information of objects. Some methods took as input LIDAR bird views (3D) [1–3] or depth images (2.5D) [5] with RGB images to locate objects more accurately. However, these approaches need extra expensive equipment, which is unavailable in normal public places. As a result, currently they are too costly to be applied in most public cases, except for some special ones such as autonomous driving, which desperately needs accurate object detection and can afford it.

Besides, from disparity maps we can also get depth information of objects, and these will help learn more discriminative features, such as Fig. 1(b). Furthermore, disparity maps can be got only depending on a pair of normal RGB cameras, so RGB-Disparity based approaches are more feasible and they have much lower cost than the above polymorphic-based approaches. Object detection utilizing only binocular information has not drawn much attention so far. Some approaches regarded binocular information as the correction of monocular detection. For example, Zhang et al. [22] adjusted the detection results of left images with that of right images, but it needed to detect respectively on binocular images so brought much higher computational cost.

In this paper, we propose the asymmetric two-stream networks for RGB-Disparity based object detection. Different from normal two-stream networks [3–5,21,23,24], where both streams were similarly designed as complete network

structures, in our networks one of the streams is based on only part of the whole backbone network with lower computational cost. Our method can significantly improve the performance of basic network. The main contributions of this paper are as follows.

- We propose the asymmetric two-stream networks for RGB-Disparity based object detection. To our knowledge, we are the first to raise a deep learning based framework utilizing only binocular information for object detection.
- Asymmetric two-stream networks are designed to combine RGB and disparity data, so the proposed method can learn discriminative features more easily. Besides, our approach only depends on a pair of normal RGB cameras, so it's low-cost and more feasible in public places.
- The experiment results on KITTI and our proposed BPD dataset have shown that, with less complexity, our method can improve the performance and have comparable effect with some complicated monocular-based methods.

2 Related Work

Object detection has made great progress during these years. For more accurate localization, researchers worldwide have made a lot of efforts, which are twofold: designing stronger networks and utilizing other more reliable equipment.

On the aspect of designing stronger networks, proposal-based methods [7,8,11,18], generating proposals first and then applying high-quality classifiers, have developed for their performances but with higher computational cost. Ren et al. [18] proposed faster R-CNN, which employed RPN and following classification into an end-to-end network. Lin et al. [11] exploited the inherent pyramidal hierarchy of deep convolutional network to construct feature pyramid. On the other hand, [12,13,15,16,19] established regression-based frameworks to locate objects, which removed the step of generating proposals and trained end-to-end detectors directly with higher computational efficiency. Liu et al. [13] proposed SSD, predicting object locations on multi-scale layers thus obtaining desirable performance for objects with different scales. In this paper, for the excellent trade-off between performance and computational cost, our proposed method is based on the SSD network [13].

More complex networks can learn more discriminative features, yet with higher computational cost. On the other hand, some methods utilized other more reliable equipment to locate objects, such as LIDAR and Kinect. [1–3] took as input both LIDAR point clouds and RGB images to predict oriented 3D bounding boxes. Deng et al. [5] stuck to the 2.5D representation framework, taking as input RGB images and Kinect depth maps, to find 3D locations of objects. With the help of other reliable equipment, depth features of objects can be caught, thus localization would be more accurate. However, such expensive devices are not available in most public places. Furthermore, in most public cases 2D localization is enough instead of 3D. As a result, these methods are not feasible in normal public places. In our work, we propose the asymmetric

two-stream networks utilizing RGB and disparity data for 2D object detection, which only require a pair of normal RGB cameras and can achieve a significant increase in performance.

Object detection utilizing only binocular information has not drawn much attention so far. Actually, with the binocular images by a pair of normal RGB cameras, disparity maps can be got and represent the distances between objects and cameras. Utilizing it we can learn more discriminative features of objects. Zhang et al. [22] detected pedestrians, aided by the fusion of binocular information. However, detections were based on traditional sliding-window methods and needed to be processed on binocular images respectively, where disparity maps were used for preprocessing only. To our knowledge, we are the first to propose a deep learning based framework utilizing binocular information for object detection, and it needs to be processed only once per image.

The structure of two-stream networks is popular in dealing with cross-modal data. [21,23,24] designed two-stream networks to fuse RGB and flow features in action recognition. [4] proposed two-stream networks to utilize ensembles of RGB and hypothetical thermal data in pedestrian detection. However, to maintain the symmetry of networks, in existed methods both streams were similarly tended to be designed as complete network structures, so they were multi-parameter and computationally costly. In this paper, we establish a framework of asymmetric two-stream networks, where one of the streams is based on only part of the whole backbone network with lower computational cost. RGB and disparity data go through different networks respectively, and then discriminative features can be learned and fused for better object detection.

3 Two-Stream Networks for Learning and Fusing RGB-Disparity Representations

In this section, the proposed asymmetric two-stream networks for RGB-Disparity based object detection will be described in details. We first present the overview of our approach in Sect. 3.1. Then in Sect. 3.2, the design of asymmetric two-stream networks will be discussed in particular. Finally, specific to the construction of two-stream networks, we will talk about our training strategy in Sect. 3.2.

3.1 Overview

The overview of our asymmetric two-stream networks is illustrated in Fig. 2. Given a pair of binocular RGB images, a disparity map can be got by stereo matching methods, and then two-stream networks get processed. Our proposed two-stream networks consist of two different networks, Disparity Representations Mining Network (DRMN) and Muti-Modal Detection Network (MMDN). DRMN takes as input disparity maps in order to learn discriminative features from them, which will be aided to MMDN later. On the other hand, MMDN learns representations from RGB images and fuses them with the output of DRMN. According to representations from both sides, MMDN processes the

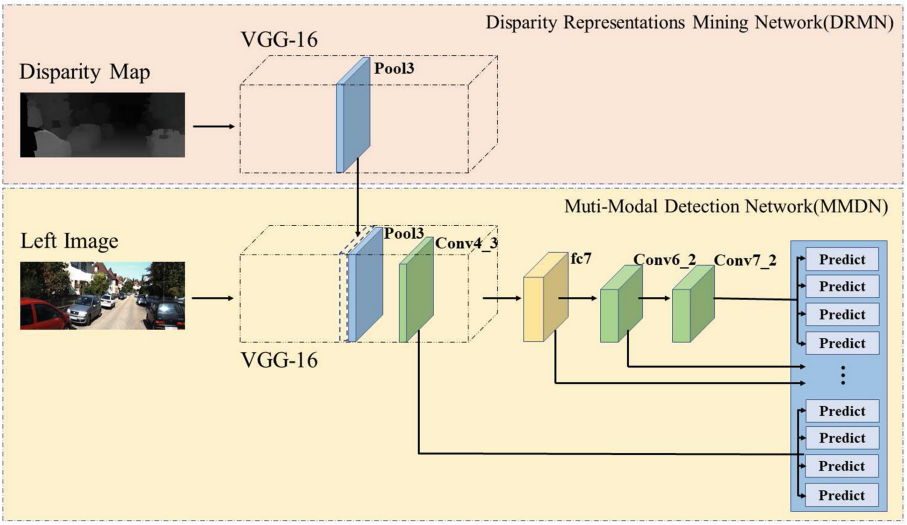


Fig. 2. The overview of our asymmetric two-stream networks.

final detection. With the two-stream networks, representations from RGB and disparity data can be learned and fused to generate more discriminative information, helpful to detection.

3.2 Two-Stream Networks

Given RGB images x_{rgb} and disparity maps x_d, the easy ways to fuse them are addition of them, maximum of them, and concatenation of them. However, features from RGB images and disparity maps, representing information of color and depth respectively, are cross-modal, and it's hard for models to directly learn from them. As a result, We need to learn transform functions $F_1(\cdot)$, $F_2(\cdot)$, projecting them into a common space S, where the fusion of them will be much easier.

$$\tilde{X}_1 = F_1(x_{rgb}), \quad \tilde{X}_2 = F_2(x_d), \quad \tilde{X}_1, \tilde{X}_2 \in S \tag{1}$$

To deal with cross-modal features, referring to [4,24], we establish a framework of asymmetric two-stream networks to model $F_1(\cdot)$ and $F_2(\cdot)$ separately. RGB and disparity data go through different convolution networks respectively, after that semantic information learned can be fused more easily. According to above, our proposed asymmetric two-stream networks consist of two different networks, Disparity Representations Mining Network (DRMN) and Muti-Modal Detection Network (MMDN).

Disparity Representations Mining Network (DRMN). In order to narrow the apparent gap between RGB and disparity data, Disparity Representa-

tions Mining Network (DRMN) is aimed to learn high-level semantic information from disparity data. Actually, it's worth noting that disparity data mainly represent the depths of objects, from which almost only objects' actual shapes can be got, while from RGB images we can extract rich color and texture features, so disparity maps contain less information compared with RGB images. An intermediate experiment shows that, as illustrated in Fig. 3, enough semantic information has been learned on low-level layers, thus it's not necessary for disparity data to go through very deep layers. As a result, different from [4,24], where both streams were designed using complete backbone networks, in our implementation of DRMN we only exploit half of the whole backbone network, which is computationally saving. According to above, we develop DRMN based on VGG-16 network structure [20] except for the first convolution layer. Note that due to the simplicity of disparity data, layers after pool3 are removed, which will be evaluated in Sect. 4 in details.

(a) RGB (b) Disparity (c) Pool3 (d) Pool5 (e) Conv6_1

Fig. 3. Feature maps output from different layers for disparity data. Note that the saliency of a car has already been learned on the layer of pool3 (c), getting approximately the same effects with that from pool5 (d) and conv6_1 (e). Best viewed in color. (Color figure online)

Muti-Modal Detection Network (MMDN). The aim of Muti-Modal Detection Network (MMDN) is to fuse multi-modal data and then process the final detection. There are 3 main steps in MMDN, learning representations from RGB data, fusing heterogeneous features from RGB and disparity, and determining the final detection. With the help of multi-modal data, localization will be more accurate.

For excellent performance on objects with different scales, we develop our MMDN on SSD backbone network [13] to finish the above 3 steps. There are 2 main differences between our method and [13]. Firstly, features learned from RGB data and DRMN will be concatenated on the layer of pool3. At this moment semantic information of them can be fused more easily. Secondly, in [13], feature maps in a layer were responsible for the regression of bounding boxes with a range of size. But if the range was too large, the accuracy of regression will be affected. Referring to [17], we discretize the range of bounding boxes' size and assign 4 regressions for features maps in a layer. However, different from [17], we do not exploit Recurrent Rolling Convolution architecture because of the large computational cost, and evaluations in Sect. 4 will show that our asymmetric two-stream networks have comparable performance with [17], but with less time consumption.

Training Strategy. As discussed above, our proposed two-stream networks consist of 2 different networks. To avoid the problem of slow convergence, the training process includes 3 main phases. Firstly, DRMN is trained on the SSD network [13], initialized using the parameters of VGG-16 [20] pre-trained on ImageNet dataset. After training the layers after pool3 are removed. Secondly, MMDN is trained with the output of DRMN deactivated. The layers of MMDN are initialized similar to DRMM. Finally, the layer of pool3 in MMDN is concatenated with the output of DRMN. The combined two-stream networks are finetuned together, where the coefficients of learning rates of the layers before pool3 drop to 0.1 and others remain 1.

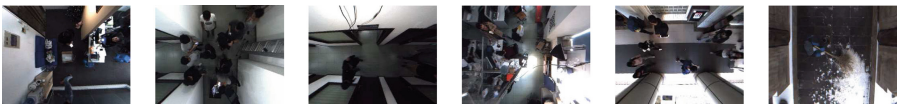
4 Experiment

In this section, details of evaluation will be described. Experiments are performed on KITTI Object Detection Benchmark [6], a publicly available dataset, and our proposed Binocular Pedestrian Detection (BPD) dataset, captured by binocular devices. To evaluate the effectiveness of our method, we conduct 3 experiments in this section. Firstly, we process our method under different stereo matching methods to demonstrate the insensitivity of our algorithm to different stereo matching approaches. Secondly, because DRMN is based on part of the whole backbone network, exploration of different DRMN's depths is performed. Finally, the analysis of performances on both datasets is provided.

4.1 Datasets

The KITTI dataset [6] consists of 7481 images for training and validation, and 7518 images for testing, captured by driving cars with stereo cameras. The groundtruth of the test set is not available, and everyone has only one chance to submit results to a dedicated server for evaluation on the test set. Following [17], We employ an image similarity metric for the training set and validation set separation, which makes our resulting validation set contain 2741 images. In the meanwhile, for a fair comparison, as described in [17], the experiments on KITTI are carried out with only car dataset because the pedestrian data are scarce.

The Binocular Pedestrian Detection (BPD) dataset is captured by ourselves using top-view binocular devices, which covers plenty of indoor and outdoor real-world scenes, such as offices, corridors, laboratory, teaching building, and scenic spots, etc. Besides the BPD dataset is very challenging with many hard samples, and a low image resolution of 320×240, as illustrated in Fig. 4. The BPD dataset consists of 65093 images for training and 12330 images for testing. Specially, all images are captured in the top-view. All left and right images are provided so that we can get a disparity map for each pair of images using stereo matching methods.



(a) (b) (c) (d) (e) (f)

Fig. 4. The BPD dataset. Plenty of indoor and outdoor real-world scenes are covered, such as offices (a), corridors (b, c), laboratory (d), teaching building (e), and scenic spots (f), etc. The dataset is challenging with a low image resolution of 320×240, and a lot of hard samples, for instance, the pedestrians in the dark.

4.2 Experiment Setting

The following settings are used throughout the experiments. In training, we adopt the data augmentation methods described in SSD [13]. Stochastic gradient descent (SGD) is chosen for optimization. Besides, the initial learning rate is set to 0.0005, which will be divided by 10 at iterations of 20000 and 100000. Training is processed for 120000 iterations in total. In the whole evaluation, mAP with IoU of 0.5 is adopted as the criteria, and all experiments on speed are measured with batch size 1 using TITAN X with Intel Xeon E5-2620@2.10 GHz.

4.3 Exploration Under Different Stereo Matching Methods

Our proposed DRMN takes as input disparity map, which can be got using various of stereo matching methods. To demonstrate that our method is insensitive to different stereo matching approaches, we process our method under two typical stereo matching methods, CRL [14] and SGBM [9]. CRL [14] is CNN-based so it's time-consuming but performs very well. SGBM [9] is a time-saving method with the help of classical semi-global matching, but the disparity maps got are coarser, such as Fig. 5. Table 1 shows the performances of the proposed method under these two stereo matching approaches on the KITTI validation set, where all images are resized to 640×192. We can find that although there is a large gap between both of disparity maps, our asymmetric two-stream networks get almost the same accuracy under them (84.1% vs 83.8%). We argue that it's because, on one hand, the semantic information learned by our DRMN can help to reduce the discrepancy on appearances of them. On the other hand, features learned by MMDN can correct the results of DRMN. As a result, it can be confirmed that our method is insensitive to stereo matching approaches, which means we can adopt fast stereo matching (e.g. SGBM (40 ms/frame)) for practical feasibility instead of CNN based ones (e.g. CRL (190 ms/frame)).

4.4 Evaluation of Performances Under Different Depths of DRMN

As mentioned in Sect. 3, DRMN is based on part of the whole backbone network. To explore the depth of DRMN, Table 2 shows the performances under different depths of DRMN on the KITTI validation set, where disparity maps are generated using SGBM [9]. Noting that *Asymmetric Two-Stream Networks (data)*

(a) RGB image (b) CRL [14] (c) SGBM [9]

Fig. 5. Disparity maps from CRL [14] and SGBM [9]. Best viewed in color. (Color figure online)

Table 1. Performances of our method under two different stereo matching methods on KITTI validation set.

	Our mAP (%)	Time (per image)
CRL [14]	84.1	190 ms
SGBM [9]	83.8	40 ms

means the depth of DRMN is zero and raw disparity data are concatenated with RGB data directly. Other *Asymmetric Two-Stream Networks (X)* mean DRMN forwards until the layer of X and then the output are concatenated on the corresponding layer in MMDN. In addition, *One-Stream Network (only RGB)* is also chosen for comparison, where DRMN is removed and MMDN utilizing only RGB data works without the output of DRMN. We can see that because RGB and disparity are cross-modal data, concatenating them directly without two-stream networks even results in a worse accuracy than one-stream network utilizing only RGB data (73.7% vs 80.8%), which strongly confirms the necessity of DRMN. Besides, we demonstrate that performance is not better as DRMN goes deeper, and concatenating features on a low-level layer of pool3 can achieve better performance boost than other high-level layers (i.e. pool5 and conv6_1). As discussed in Sect. 3, the reason is that disparity maps are simpler than RGB images, and enough semantic information has been learned on low-level layers, thus it's not necessary to forward very deep layers. According to above, in our implementation of DRMN, layers after pool3 are removed.

4.5 Results on the KITTI Dataset

To demonstrate the effectiveness of our asymmetric two-stream networks, our method is evaluated on the KITTI validation set. Because our method is regression-based, for fairness, we compare our method only with regression-based methods. Thereinto, our method is compared with the recently published state-of-the-art method, Recurrent Rolling Convolution Detector (RRC) [17], which ranked the first for the hardest category of the car on KITTI testing set by the time this paper was written. Note that our method is not compared with other two-stream or binocular-based methods because of the different applications. Firstly, existed two-stream networks for detection [1–3,5] were mostly designed for 3D localization instead of 2D, and they required data from LIDAR or Kinect, which is not consistent with our case. [4] was for 2D localization but it needed thermal data for training. Secondly, binocular-based approach [22] was based

on traditional sliding-window methods inefficiently and designed for front-view pedestrian detection only. We are the first to propose an end-to-end deep learning based framework for RGB-Disparity based object detection, which only utilizes binocular information.

Table 2. Detection results on KITTI validation set.

Methods	mAP (%)	FPS
One-stream network (only RGB)	80.8	25
Asymmetric two-stream networks (data)	73.7	20
Asymmetric two-stream networks (pool3)	**83.8**	17
Asymmetric two-stream networks (pool5)	82.3	15
Asymmetric two-stream networks (conv6_1)	82.2	14
RRC [17][a]	84.0	11
Asymmetric two-stream RRC	**85.6**	9

[a]The accuracy is lower than that reported in [17], mainly because in [17] an image size of 2560×768 was adopted but here all images are resized to 640×192. Actually, the increase in input size significantly boosts detection accuracy, as pointed out in [10], but it will cause overloaded occupation of GPU memory and be impractical.

Table 2 shows the results on the KITTI validation set. It needs to be pointed out that, in [17], to achieve the best performance, an image size of 2560×768 was adopted, which would cause overloaded occupation of GPU memory and be divorced from reality. In order to develop feasible methods available in most public places, all images are resized to 640×192 and we retrain RRC carefully for evaluation. From the results we can observe that, on one side, our proposed asymmetric two-stream networks exploiting RGB and disparity get the accuracy of 83.8%, outperforming one-stream network exploiting RGB by 3%. On the other side, our method can get comparable accuracy with RRC (83.8% vs 84.0%), while our method runs faster than RRC (17 FPS vs 11 FPS). It can be confirmed that our method exploits disparity data well so that it can learn discriminative features more easily without the large increase in network complexity. Finally, since RRC is based on SSD network too, we employ our asymmetric two-stream networks in it, which achieves the state-of-the-art with an accuracy of 85.6%. All of the above have shown the effectiveness of our asymmetric two-stream networks.

4.6 Results on the BPD Dataset

In this section, we evaluate methods on our proposed BPD dataset, which is captured by ourselves using top-view binocular devices, including lots of hard samples. Following the settings in Sect. 4.5 except that image size of 320×240

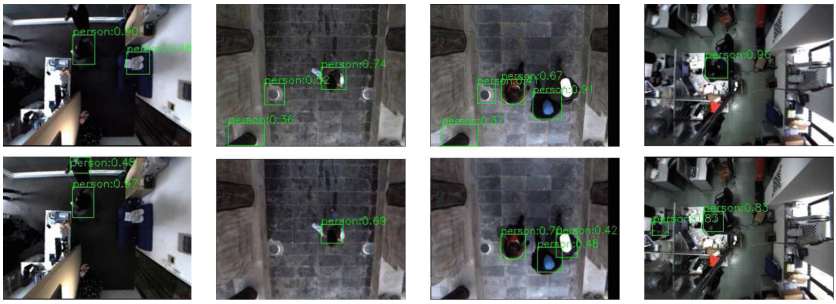


Fig. 6. Examples of detection results on the BPD dataset. The top and bottom rows refer to results of one-stream network and our asymmetric two-stream networks respectively. Compared with the other, Our method can handle more hard samples (e.g. objects on the sofa, lamps, and pedestrians with hats on). Best viewed in color. (Color figure online)

Table 3. Detection results on BPD dataset.

Methods	mAP (%)	FPS
One-stream network (only RGB)	77.8	28
Asymmetric two-stream networks	**83.0**	22
RRC [17]	83.3	14
Asymmetric two-stream RRC	**84.7**	11

is adopted here, results on the BPD dataset are illustrated as Table 3. Similar to results in Sect. 4.5, we can see that our asymmetric two-stream networks bring a significant increase in performance over the one-stream network (83.0% vs 77.8%). Figure 6 shows the qualitative results of one-stream network exploiting only RGB data and our asymmetric two-stream networks respectively, we can see that our method can handle more hard samples. Besides, our method achieves comparable accuracy with RRC (83.0% vs 83.3%), while RRC runs at 14 FPS, slower than our methods (22 FPS). Obviously, our asymmetric two-stream networks utilizing both RGB and disparity information can get comparable performance with lower network complexity. Finally, we employ our asymmetric two-stream networks in RRC, with an accuracy of 84.7%, getting the state-of-the-art. The results shown in this section demonstrate the effectiveness of our asymmetric two-stream networks again.

5 Conclusion

In this paper, we propose the asymmetric two-stream networks for RGB-Disparity based object detection, which exploit both RGB and disparity data to get a higher accuracy of localization. Unlike normal two-stream networks, our model is asymmetric due to the different capacity of RGB and disparity data.

Experiment results show that our asymmetric two-stream networks can learn more discriminative features without the large increase in network complexity, and get the state-of-the-art. In the future, we plan to refine disparity data by detection, to generate disparity better benefitting detection.

Acknowledgement. This project is supported by the Natural Science Foundation of China (61573387, 61672544), and the Tip-top Scientific and Technical Innovative Youth Talents of Guangdong special support program (No. 2016TQ03X263).

References

1. Chen, X., et al.: 3D object proposals for accurate object class detection. In: NIPS, pp. 424–432. MIT Press (2015)
2. Chen, X., Kundu, K., Zhu, Y., Ma, H., Fidler, S., Urtasun, R.: 3D object proposals using stereo imagery for accurate object class detection. PAMI **40**(5), 1259–1272 (2018)
3. Chen, X., Ma, H., Wan, J., Li, B., Xia, T.: Multi-view 3D object detection network for autonomous driving. In: CVPR, pp. 6526–6534. IEEE (2017)
4. Dan, X., Ouyang, W., Ricci, E., Wang, X., Sebe, N., et al.: Learning cross-modal deep representations for robust pedestrian detection. In: CVPR, pp. 4236–4244. IEEE (2017)
5. Deng, Z., Latecki, L.J.: Amodal detection of 3D objects: Inferring 3D bounding boxes from 2D ones in RGB-depth images. In: CVPR, pp. 398–406. IEEE (2017)
6. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? The KITTI vision benchmark suite. In: CVPR, pp. 3354–3361. IEEE (2012)
7. Girshick, R.: Fast R-CNN. In: ICCV, pp. 1440–1448. IEEE (2015)
8. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask R-CNN. In: ICCV, pp. 2980–2988. IEEE (2017)
9. Hirschmuller, H.: Stereo processing by semiglobal matching and mutual information. PAMI **30**(2), 328–341 (2008)
10. Huang, J., et al.: Speed/accuracy trade-offs for modern convolutional object detectors. In: CVPR, pp. 7310–7311. IEEE (2017)
11. Lin, T.Y., Dollar, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: CVPR, pp. 2117–2125. IEEE (2017)
12. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollar, P.: Focal loss for dense object detection. In: CVPR, pp. 2980–2988. IEEE (2017)
13. Liu, W., et al.: SSD: single shot multibox detector. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9905, pp. 21–37. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46448-0_2
14. Pang, J., Sun, W., Ren, J.S., Yang, C., Yan, Q.: Cascade residual learning: a two-stage convolutional neural network for stereo matching. In: ICCV, pp. 887–895. IEEE (2017)
15. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: CVPR, pp. 779–788. IEEE (2016)
16. Redmon, J., Farhadi, A.: Yolo9000: better, faster, stronger. In: CVPR, pp. 6517–6525. IEEE (2017)
17. Ren, J., et al.: Accurate single stage detector using recurrent rolling convolution. In: CVPR, pp. 752–760. IEEE (2017)

18. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. PAMI **39**(6), 1137–1149 (2017)
19. Shen, Z., Liu, Z., Li, J., Jiang, Y.G., Chen, Y., Xue, X.: DSOD: learning deeply supervised object detectors from scratch. In: CVPR, pp. 1919–1927. IEEE (2017)
20. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
21. Sun, S., Kuang, Z., Ouyang, W., Sheng, L., Zhang, W.: Optical flow guided feature: a fast and robust motion representation for video action recognition. arXiv preprint arXiv:1711.11152 (2017)
22. Zhang, Z., Tao, W., Sun, K., Hu, W., Yao, L.: Pedestrian detection aided by fusion of binocular information. Pattern Recognit. **60**, 227–238 (2016)
23. Zhu, J., Zou, W., Zhu, Z.: End-to-end video-level representation learning for action recognition. arXiv preprint arXiv:1711.04161 (2017)
24. Zhu, Y., Lan, Z., Newsam, S., Hauptmann, A.G.: Hidden two-stream convolutional networks for action recognition. arXiv preprint arXiv:1704.00389 (2017)