# ADNet: Asymmetric Dual-mode Network for RGB-D Indoor Scene Parsing*

Mingrong Gong[1,2], Qieshi Zhang[1,2,3], Fusheng Hao[1,3], Shuiming Ouyang[1,2], Jun Cheng[1,2,3], and Yunduan Cui[1(✉)]

[1] Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China
yd.cui@siat.ac.cn
[2] University of Chinese Academy of Sciences, Beijing, China
[3] The Chinese University of Hong Kong, Hong Kong, China

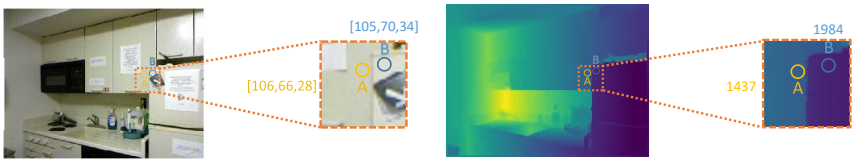
**Abstract.** Indoor scene parsing is a crucial and fundamental task in the field of robotics. In recent years, semantic segmentation based on RGB images and depth images have achieved excellent performance. However, the quality of depth images are not high, and often have noise or information loss. The spatial information that contained in the depth images is not in the same mode as the color information in the RGB images, which cannot be simply overlapped to segment. The limitation can lead to unsatisfactory segmentation results. To overcome this limitation, we propose Asymmetric Dual-mode Network (ADNet) to fuse color information and spatial information more efficiently. The core of ADNet is a shallow network for extracting spatial information in depth images, which can improve the utilization of the network. In addition, Depth Filter (DF) operator is added to the network to optimize the spatial information and reduce the effect of noise on segmentation. We evaluate our ADNet on the common indoor dataset NYUv2 and compare it with the approach of state-of-the-art on this dataset, our proposed model has a competitive performance.

**Keywords:** RGB-D semantic segmentation · Asymmetric Network · Depth filter operator.

## 1 Introduction

lwalIndoor scene parsing is an important task in the field of robotics, The task has a wide range of applications in areas such as autonomous driving [1], robot sensing [2], visual SLAM [3], and so on. The robots with scene understanding are able to perform more advanced tasks and better human-robot interaction. Semantic segmentation is currently well suited for scene understanding because it classifies each pixel and enhances the perceptual capabilities of the robot. In recent years, great progress has been made in semantic segmentation based

* This research is supported the National Natural Science Foundation of China under Grant 62103403



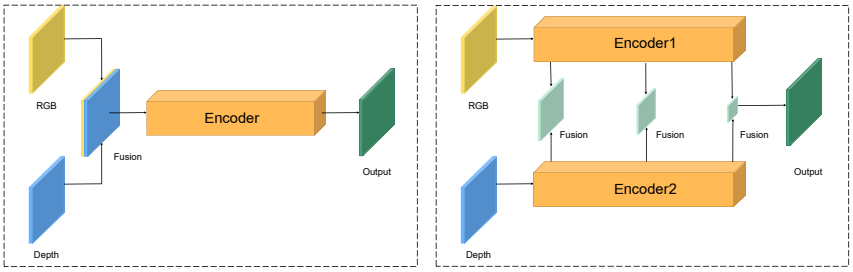

(a) (b)

**Fig. 1.** Illustration spatial information can be provided by depth images to discriminate between objects of similar color. (a) RGB image; (b) Depth image.

RGB [4–9], but performance of semantic segmentation based RGB in complex indoor situations is still a huge challenge. As shown in Fig. 1(a), the point A is the cabinet and the point B is the freezer, it is difficult to distinguish due to the close RGB values. With the availability of commercial RGB-D sensors such as Kinect, depth images are very common in robotics, depth images contain a wealth of spatial information. As shown in Fig. 1(b), the depth values of point A and point B are quite different. Some researchers have introduced depth images containing spatial information to overcome the difficultly to segment objects of close RGB [10–14].

Although adding depth images can complement spatial information to improve the performance of semantic segmentation, RGB and depth are belonging to two modalities. A lot of studies have been devoted to how to fuse color information and spatial information to improve the segmentation effect [15–19]. However, these studies neglect the fact that depth images themselves carry noise, and the noise will inevitably be added to the fusion features. It is often difficult to achieve the desired effect to use a deeper network for depth images. In this paper, we propose an asymmetric structure to improve the network utilization, which is based on two encoders of ResNet [20] to extract color information and spatial information respectively. Before the spatial information is fused with color information, there will be a Depth Filter (DF) operator filtering spatial information. In order to improve the inference speed of the model, we also reduce the number of channels of all fused features to 128. Our contributions can be summarized as follows:

- Rethink the indoor RGB-D scene parsing from a perspective of spatial information contained noisy, for extracting spatial information, an asymmetric network is proposed to effectively reduce the model parameters and improve the utilization of the network.
- A simple and effective DF operator is designed to apply in spatial information without introducing any parameters and computational complexity to make the edge of spatial information smoother and improve the segmentation effect.
- The proposed ADNet outperforms the previous state-of-the-arts on NYUv2 [21] dataset.




(a) (b)

**Fig. 2.** Illustrations of two different fusion methods. (a) Fusion before the encoder module; (b) Fusion in the encoder module.

## 2 Related Work

Convolutional Neural Networks (CNN) has been widely used in the field of semantic segmentation. In general, the mainstream segmentation structure is the encoder-decoder structure nowadays [5, 9]. The encoder module is used for the feature extraction and gradually reduces the resolution of the feature map to capture higher-level semantic information. The decoder module is used for the upsampling to gradually recover the resolution, and both the encoder module and decoder module adopt the convolutional layers as the core building blocks. The semantic segmentation based on RGB-D is generally categorized into two methods, the first category is fusion before the encoder module and the second category is fusion in the encoder module.

### 2.1 Fusion before the Encoder Module

Methods in the first category propose to fuse the color information directly with the spatial information at the input stage before the encoder extracts the features, as shown in Fig. 2(a), then a common encoder with a special operator will be used to extract features. For instance, D-CNN [12] presented to set the weighted convolution according to the size of the object. Malleable 2.5D [22] presented a novel operator called malleable 2.5D convolution. To learn the receptive field along the depth-axis, VCD [16] added deformable convolution [23, 24] to the model to achieve more robust results. but this processing of the two modalities together usually makes it difficult to exploit the spatial information, and some complex operators can make the model inference time much longer. So we choose to replace convolution with pooling to roughly filter out spatial information noise without adding parameters.

### 2.2 Fusion in the Encoder Module

In contrast, methods in the second category target at proposing to feed color information and spatial information to two parallel encoder modules. As shown in Fig. 2(b), and the output features are fused with specific strategies. For example,

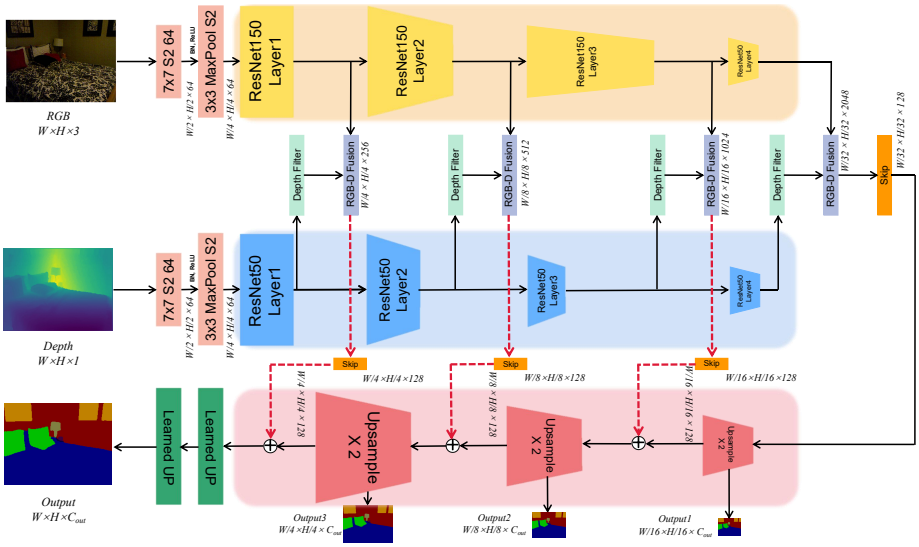



**Fig. 3.** Overview of our proposed ADNet152-50 for RGB-D segmentation. The encoder for extracting color information is in yellow, the encoder for extracting depth information is in blue, and the decoder for upsampling is in pink.

ACNet [15], NANet [25], BCMFP [26], Esanet [18] and CANet [19] presented a gata-fusion method and fuse the features in multi-levels of the backbone stages. However, we used an asymmetric network, which is different from the symmetric network of the above method. Because on the one hand, we can reduce the model parameters to decrease the inference time. More importantly, we consider that a deeper model will not make better use of the spatial information in the process of extracting depth images features. Due to the depth images often containing plenty of noises, which will cannot fully utilize the effect of deep model.

## 3 Asymmetric Dual-mode Network

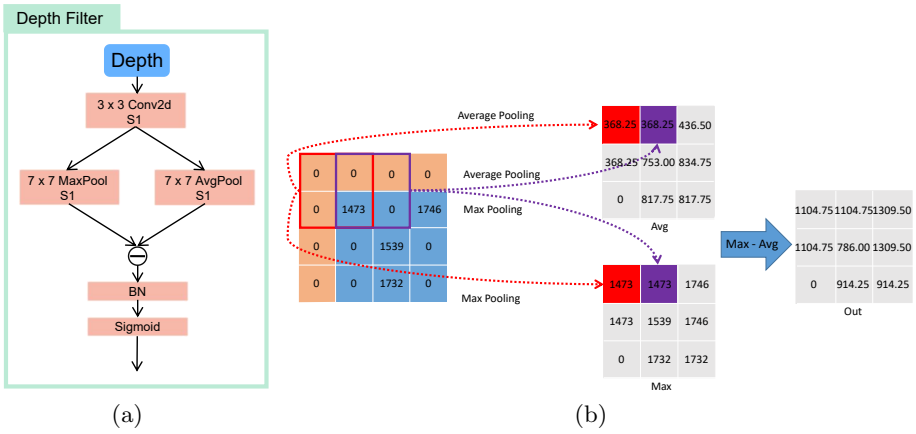
The overall framework of our proposed ADNet is shown in Fig. 3, which mainly consists of the structure of the encoder-decoder. We will briefly describe our encoder in Sec. 3.1, and then in Sec. 3.2 and Sec. 3.3, we explain in detail how to improve the utilization of depth images and how to combine color information and spatial information more effectively, respectively. In Sec. 3.4, we describe the implementation of the decoder.

### 3.1 Encoder

ADNet152-50 has two asymmetric encoder modules, one using the ResNet152 stream for extracting color information from RGB images and the other one using the ResNet50 stream for extracting spatial information from depth images. In

addition, unlike the number of fusions with RedNet [28] and ACNet [15], ADNet drops the maximum resolution fusion feature. As shown in Fig. 3, we skipped fusing the features of size $\frac{W}{2} \times \frac{H}{2} \times 64$ that are fed into the 7×7 convolution layer.

### 3.2 Depth Filter




(a) (b)

**Fig. 4.** Detailed description of DF operator. (a) Implementation of Upsample module; (b) illustration of specific DF operator, the upper result is average operator and the lower one is max operator.

Spatial information is still accompanied by noises before fusing with color information (depth images with missing data when surfaces are too glossy, bright, thin, close, or far from the camera) [29], and we need to filter the noises of spatial information before fusing color information. Inspired by D-CNN [12] and ShapeConv [30], which is done with a special operator for extracting spatial information from depth images, this operator first takes the average of the sum of the convolution kernel, then subtracting the average from the original feature map. But this will add too many parameters to affect the inference time. We propose the a simple and intuitive DF operator (see the Fig. 4(a) for specific implementation), and the pooling is chosen to replace the convolution because the pooling operator does not take additional parameters. The input depth feature map $\mathbf{D} \in R^{h \times w \times c_{in}}$ and the out feature map which feed into DF operator $out \in R^{h \times w \times c_{in}}$ , where $h$ is the height and $w$ is the width, $c_{in}$ is the number of depth feature map channels. After the DF operator, it does not change the resolution and the number of channels. DF operator is consists of a max filter operator and an average filter operator. The max filter operator is

$$out_{max}(h, w, c_j) = \max_{m=0,1,...6,7} \max_{n=0,1,...6,7} \mathbf{D}(h+m, w+n, c_j), \tag{1}$$

the average operator is

$$out_{avg}(h, w, c_i) = \frac{1}{7 \times 7} \sum_{m=0}^{7} \sum_{n=0}^{7} \mathbf{D}(h + m, w + n, c_i), \tag{2}$$

and the $out$ is

$$out(h, w, c_i) = out_{max}(h, w, c_i) - out_{avg}(h, w, c_i), \tag{3}$$

where $c_i$ is the $i$-th channel. As is shown in Fig. 4(b), the $out_{avg}$ is similar to the avg part of Fig. 4(b), the $out_{max}$ is similar to the max part of it.

### 3.3 RGB-D Fusion

The previous subsections are optimizing the noises of spatial information. Usually, there is a misalignment of color information and spatial information. When color information and spatial information are fused, they cannot be added directly. We follow the ACNet's [15] RGB-D fusion strategy, as shown in the Fig. 5, the two features are first fed into the Squeeze-and-Excitation channel attention mechanism [31] and then fused, which allows the model to learn the effective color information and spatial information.

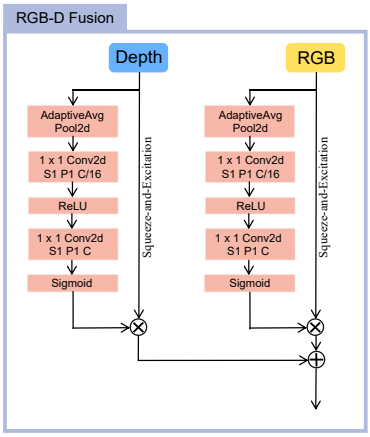



**Fig. 5.** Specific implementation of the RGB-D fusion module.

### 3.4 Decoder

To improve the processing speed of the decoder module, we reduce the number of channels in each layer of RGB-D fusion features to 128 relying on the skip module. As shown in Fig. 6(b), the Upsample layer module processes only 128-channel RGB-D fusion features, and the specific structure is depicted in Fig.

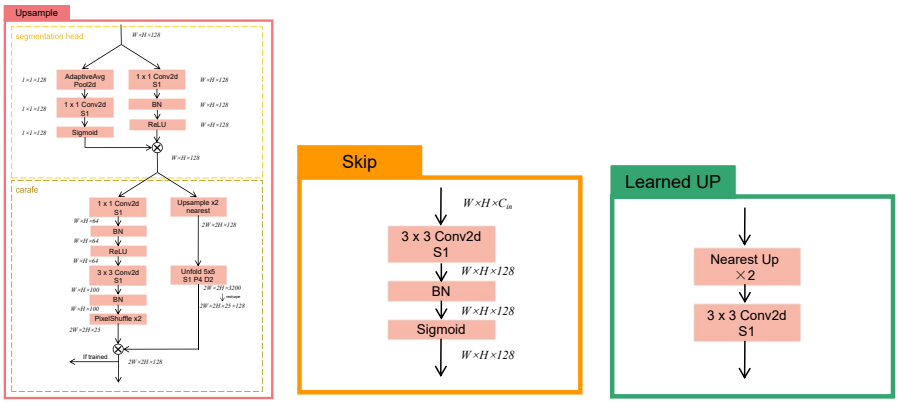



**Fig. 6.** Specific implementation of the key parts of the decoder. (a) Implementation of Upsample moudle; (b) Implementation of Skip module; (c) Implementation of Learned up module.

6(a). The module is composed of a segmentation head and an upsampling, which is combined according to the Lite R-aspp (LR-ASPP) [32] and CARAFE [33]. Replacing the original bilinear interpolation of LR-ASPP upsampling with the lightweight upsampling module CARAFE. The segmentation head is formed by two branches, the left branch does a global average pooling for RGB-D fusion features that skipped, then the number of channels is reduced by half using a $1 \times 1$ convolution layer, and then after a sigmoid activation layer. The right branch does a $1 \times 1$ convolution layer to reduce the number of channels of RGB-D fusion features and then goes through the batch normalization layer and the ReLU activation layer. Lastly, the result of the two branches is multiplied and fed into upsampling.

The Upsample is also composed of two branches, the left branch is an upsampling using pixelshuffle [34] with a multiplier of 2. The right branch takes a larger receptive field, making better use of the surrounding information and extending the number of channels to 3200, and multiply this result with the one on the left branch to get the upsample features. If it is the training process, as shown Fig. 3, output the *output1*, *output2* and *output3* respectively.

Finally, *output3* is fed into the Learned UP module, as shown in Fig. 6(c), which is a learnable upsampling layer that, unlike the atrous convolution, is first upsampled with nearest neighbors, afterwards, a $3 \times 3$ depthwise convolution is applied to combine adjacent features.

The final result *output, output1, output2*, and *output3* together are fed into a softmax layer and cross-entropy function to build the loss function.

# 4 Experiments

## 4.1 Dataset and Implementation Details

To evaluate the performance of the proposed ADNet and DF operator, we conducted comprehensive experiments on a widely used indoor scene RGB-D datasets: NYUv2. NYUv2 contains a total of 1,449 labeled RGB-D images taken by Kinect depth camera with a resolution of 480×640. The dataset has been split into training and test sets, with 795 RGBD images in the training set and 654 images in the test set, which are categorized into 40 groups.

We used the PyTorch to implement the proposed ADNet on a computer equipped with an NVIDIA RTX graphics processor and 24 GB memory. For data augmentation, we applied random scaling with scales from 0.5 to 2.0, cropping, flipping, as well as brightness, contrast, and saturation adjustments. The input RGB-D images were normalized by subtracting the mean and dividing by the variance. The training and testing images in both datasets were cropped to 480 × 640 pixels. Note that augmentation was only applied to the RGB-D images used for training.

We used both Stochastic Gradient Descent (SGD) with momentum of 0.9 and Adam with learning rates of {0.00125, 0.0025, 0.005, 0.01, 0.02, 0.04} and {0.0001, 0.0004} as the optimizer respectively, and a small weight decay of 0.0001. We adapted the learning rate using PyTorch's one-cycle learning rate scheduler. We evaluated the proposed ADNet and existing the best model in terms of mean Intersection over union (mIou).

## 4.2 Ablation Study

**Asymmetric Network.** To verify the functionality of the asymmetric network structure, the following experiments are conducted and the results obtained are listed in Table 1.

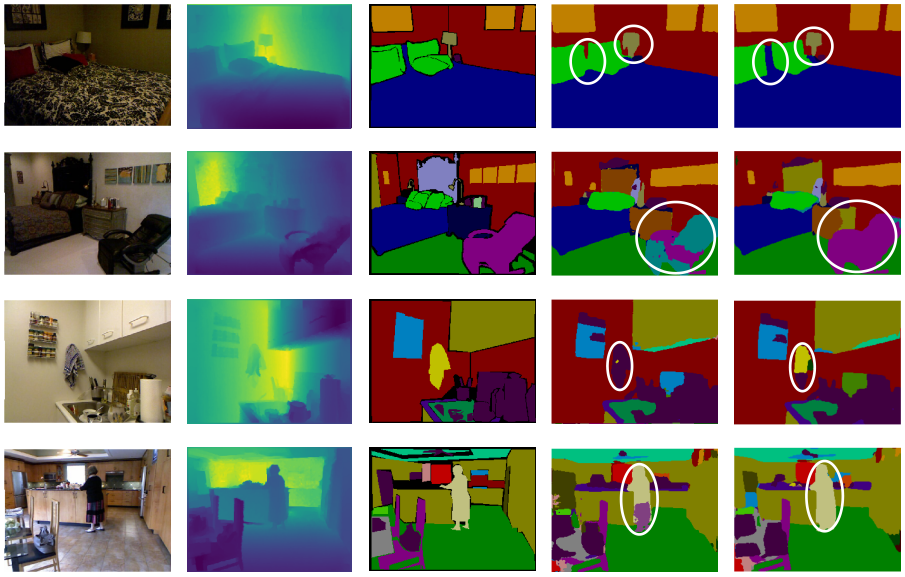
We chose ResNet101, and ResNet50 as the backbone for extracting spatial information, and ResNet152 as the backbone is the baseline. As shown in Table 1, not the deeper network, the better effect. The effect of ADNet152-50 is better than baseline, which mIou improved by 0.46. Apart from that, the inference time dropped from 0.4s to 0.31s and the model parameters dropped by about 20M. It is more efficient to choose a shallow network to extract spatial information.

**Table 1.** mIou, parameters and inference time of different networks for extracting spatial information on NYUv2 dataset.

| Method | NYUv2 | Inference Time | Parameters |
|---|---|---|---|
| Baseline | 52.16 | 0.40s | 222M |
| ADNet152-101 | 52.84 | 0.33s | 207M |
| ADNet152-50 | 52.62 | 0.31s | 188M |

**Table 2.** mIou of the different networks with DF operator compared to the one without DF operator on NYUv2 dataset

| Method | NYUv2 |
|---|---|
| Baseline | 52.16 |
| Baseline+DF | 53.46 |
| ADNet152-101 | 52.84 |
| ADNet152-101+DF | 53.38 |
| ADNet152-50 | 52.62 |
| ADNet152-50+DF | 53.32 |



(a) RGB (b) Depth (c) GT (d) ADNet (e) ADNet+DF

**Fig. 7.** Visualization of segmentation results on the NYUv2 dataset. The segmentation effect of the white circle part of (e) is significantly better than that one of (d).

**Depth Filter.** After verifying the effect of ADNet, we need to go further to verify the effect of the DF operator. We added the DF operator to our asymmetric network and baseline. The results are shown in Table 2, after adding the DF operator, no additional parameters are introduced. From the results, the proposed method is effective in improving the mIou, and we also found that the effect of the baseline after adding the DF operator is also excellent, demonstrating that the DF operator can effectively filter out the noise in the extraction of spatial information. The visualization results are shown in Fig. 7, and the method with the DF operator is superior significantly in segmenting the edges of the object.

### 4.3 Comparison with the State-of-the-art Methods

We compared different methods on the NYUv2 dataset as shown in Table 3, specifically. We compared with the single-modal methods that using and it can be shown that the single-modal models do not perform as well as the dual-modal model. Compared with the dual-modal, Zig-ZagNet [11], SCN-ResNet [37] and ShapeConv [30] are the fusion before the encoder, which is less effective than our proposed method. CEN [17], NANet [25], CANet [19] and our proposed method are the fusion in the decoder. The number of parameters in ResNet152 and ResNet50 is closed to that in double ResNet101, but the mIou is improved by about 1.8, which fully reflects the advantage of our proposed model.

**Table 3.** Table captions should be placed above the tables.

| Method | Modal | Backbone | NYUv2 |
|---|---|---|---|
| RefineNet [14] | RGB | Resnet152 | 47.6 |
| SANet [35] | RGB | Resnet50 | 50.7 |
| LWM [36] | RGB | ResNet152 | 51.51 |
| Zig-Zag Net [11] | RGB-D | ResNet152 | 51.2 |
| CEN [17] | RGB-D | ResNet101 + ResNet101 | 52.5 |
| SCN-ResNet [37] | RGB-D | ResNet152 | 49.6 |
| ShapeConv [30] | RGB-D | ResNet101 | 50.2 |
| NANet [25] | RGB-D | ResNet101 + ResNet101 | 52.3 |
| CANet [19] | RGB-D | ResNet101 + ResNet101 | 51.5 |
| ADNet152-50+DF | RGB-D | ResNet152 + ResNet50 | 53.32 |
| ADNet152-101+DF | RGB-D | ResNet152 + ResNet101 | 53.38 |
| Baseline+DF | RGB-D | ResNet152 + ResNet152 | 53.41 |

## 5 Conclusion

In this paper, we propose an ADNet for extracting spatial information, ADNet reduces the parameters while maintaining accuracy. The DF operator is also proposed, and this operator can effectively filter part of the noise in spatial information and does not introduce additional parameters. Through extensive experiments on the NYUv2 dataset it is demonstrated that our proposed modules are effective. In future work, we will investigate how to adapt ADNet with DF operator to existing pre-trained models. We will also try to extend the application to other popular tasks, such as general or depth estimation, instance segmentation, and fusing data from other modalities.

## References

1. Wang, X., Jiang, Y., Luo, Z., Liu, C.L., Choi, H., Kim, S.: Arbitrary shape scene text detection with adaptive text region representation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6449–6458 (2019)

2. Feng, D., Haase-Schuetz, C., Rosenbaum, L., Hertlein, H., Duffhauss, F., Gläser, C., Wiesbeck, W., Dietmayer, K.C.J.: Deep multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges. IEEE Transactions on Intelligent Transportation Systems (TITS) **22**(3), 1341–1360 (2021)
3. Qin, T., Chen, T., Chen, Y., Su, Q.: Avp-slam: Semantic visual mapping and localization for autonomous vehicles in the parking lot. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 5939–5945 (2020)
4. Shelhamer, E., Long, J., Darrell, T.: Fully convolutional networks for semantic segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) **39**(4), 640–651 (2017)
5. Badrinarayanan, V., Kendall, A., Cipolla, R.: Segnet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) **39**(12), 2481–2495 (2017)
6. Peng, C., Zhang, X., Yu, G., Luo, G., Sun, J.: Large kernel matters — improve semantic segmentation by global convolutional network. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1743–1751 (2017)
7. Lin, G., Milan, A., Shen, C., Reid, I.D.: Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5168–5177 (2017)
8. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6230–6239 (2017)
9. Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: European conference on computer vision (ECCV). pp. 801–818 (2018)
10. Fu, J., Liu, J., Tian, H., Fang, Z., Lu, H.: Dual attention network for scene segmentation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3141–3149 (2019)
11. Lin, D., Huang, H.: Zig-zag network for semantic segmentation of RGB-D images. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) **42**(10), 2642–2655 (2020)
12. Wang, W., Neumann, U.: Depth-aware cnn for RGB-D segmentation. In: European Conference on Computer Vision (ECCV). pp. 135–150 (2018)
13. Zhou, W., Yu, L., Zhou, Y., Qiu, W., Wu, M.W., Luo, T.: Local and global feature learning for blind quality evaluation of screen content and natural scene images. IEEE Transactions on Image Processing (TIP) **27**(5, 2086–2095 (2018)
14. Lin, G., Liu, F., Milan, A., Shen, C., Reid, I.: Refinenet: Multi-path refinement networks for dense prediction. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) **42**(5), 1228–1242 (2020)
15. Hu, X., Yang, K., Fei, L., Wang, K.: Acnet: Attention based network to exploit complementary features for RGB-D semantic segmentation. In: IEEE International Conference on Image Processing (ICIP) pp. 1440–1444 (2019)
16. Xiong, Z., Yuan, Y., Guo, N., Wang, Q.: Variational context-deformable convnets for indoor scene parsing. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 3991–4001 (2020)
17. Wang, Y., Huang, W., Sun, F., Xu, T., Rong, Y., Huang, J.: Deep multimodal fusion by channel exchanging. Advances in Neural Information Processing Systems (NeurIPS) **33**, 4835–4845 (2020)
18. Seichter, D., Köhler, M., Lewandowski, B., Wengefeld, T., Gross, H.M.: Efficient RGB-D semantic segmentation for indoor scene analysis. In: IEEE International Conference on Robotics and Automation (ICRA). pp. 13525–13531 (2021)

19. Zhou, H., Qi, L., Huang, H., Yang, X., Wan, Z., Wen, X.: Canet: Co-attention network for RGB-D semantic segmentation. Pattern Recognition (PR) **124**, 108468 (2022)
20. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 770–778 (2016)
21. Silberman, N., Hoiem, D., Kohli, P., Fergus, R.: Indoor segmentation and support inference from RGB-D images. In: European Conference on Computer Vision (ECCV). pp. 746–760 (2012)
22. Xing, Y., Wang, J., Zeng, G.: Malleable 2.5 d convolution: Learning receptive fields along the depth-axis for RGB-D scene parsing. In: European Conference on Computer Vision (ECCV). pp. 555–571. (2020)
23. Dai, J., Qi, H., Xiong, Y., Li, Y., Zhang, G., Hu, H., Wei, Y.: Deformable convolutional networks. In: IEEE/CVF International Conference on Computer Vision (ICCV). pp. 764-773 (2017)
24. Zhu, X., Hu, H., Lin, S., Dai, J.: Deformable convnets v2: More deformable, better results. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 9308–9316 (2019)
25. Zhang, G., Xue, J.H., Xie, P., Yang, S., Wang, G.: Non-local aggregation for RGB-D semantic segmentation. IEEE Signal Processing Letters (SPL) **28**, 658–662 (2021)
26. Chen, X., Lin, K.Y., Wang, J., Wu, W., Qian, C., Li, H., Zeng, G.: Bi-directional cross-modality feature propagation with separation-and-aggregation gate for RGB-D semantic segmentation. In: European Conference on Computer Vision (ECCV). pp. 561–577 (2020)
27. Su, Y., Yuan, Y., Jiang, Z.: Deep feature selection-and-fusion for RGB-D semantic segmentation. In: International Conference on Multimedia and Expo (ICME). pp. 1–6 (2021)
28. Jiang, J., Zheng, L., Luo, F., Zhang, Z.: Rednet: Residual encoder-decoder network for indoor RGB-D semantic segmentation. arXiv preprint arXiv:1806.01054 (2018)
29. Zhang, Y., Funkhouser, T.: Deep depth completion of a single RGB-D image. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 175–185 (2018)
30. Cao, J., Leng, H., Lischinski, D., Cohen-Or, D., Tu, C., Li, Y.: Shapeconv: Shape-aware convolutional layer for indoor RGB-D semantic segmentation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 7088–7097 (2021)
31. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 7132–7141 (2018)
32. Howard, A., Sandler, M., Chu, G., Chen, L.C., Chen, B., Tan, M., Wang, W., Zhu, Y., Pang, R., Vasudevan, V., et al.: Searching for mobilenetv3. In: IEEE/CVF International Conference on Computer Vision (ICCV). pp. 1314–1324 (2019)
33. Wang, J., Chen, K., Xu, R., Liu, Z., Loy, C.C., Lin, D.: Carafe: Content-aware reassembly of features. In: IEEE/CVF International Conference on Computer Vision (ICCV). pp. 3007–3016 (2019)
34. Shi, W., Caballero, J., Huszár, F., Totz, J., Aitken, A.P., Bishop, R., Rueckert, D., Wang, Z.: Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1874–1883 (2016)
35. Yu, L., Gao, Y., Zhou, J., Zhang, J., Wu, Q.: Multi-layer feature aggregation for deep scene parsing models. arXiv preprint arXiv:2011.02572 (2020)

36. Gu, Z., Niu, L., Zhao, H., Zhang, L.: Hard pixel mining for depth privileged semantic segmentation. IEEE Transactions on Multimedia (TMM) **23**, 3738–3751 (2021)
37. Lin, D., Zhang, R., Ji, Y., Li, P., Huang, H.: Scn: Switchable context network for semantic segmentation of RGB-D images. IEEE Transactions on Cybernetics (TYCB) **50**(3), 1120–1131 (2020)