
# EBANet: Efficient Boundary-Aware Network for RGB-D Semantic Segmentation

Ruiquan Wang(✉), Qingxuan Jia, Yue Shen, Zeyuan Huang, Gang Chen, and Junting Fei

Beijing University of Posts and Telecommunications, Beijing, China
wangruiquan@buptrobot.com

**Abstract.** Semantic segmentation is widely used in robot perception and can be used for various subsequent tasks. Depth information has been proven to be a useful clue in the semantic segmentation of RGB-D images for providing a geometric counterpart to the RGB representation. At the same time, considering the importance of object boundaries in the robot's perception process, it is very necessary to add attention to the boundaries of the objects in the semantic segmentation model.

In this paper, we propose Efficient Boundary-Aware Network (EBANet) which relies on both RGB and depth images as input. We design a boundary attention branch to extract more boundary features of objects in the scene and generate boundary labels for supervision by a Canny edge detector. We also adopt a hybrid loss function fusing Cross-Entropy (CE) and structural similarity (SSIM) loss to guide the network to learn the transformation between the input image and the ground truth at the pixel and patch level. We evaluate our proposed EBANet on the common RGB-D dataset NYUv2 and show that we reach the state-of-the-art performance.

**Keywords:** RGB-D semantic segmentation · Boundary attention · Hybrid loss

## 1 Introduction

For autonomous and intelligent robots, accurate scene perception is necessary. Semantic segmentation is well suited for such an initial step, as it provides precise pixel-wise information that can be used for numerous subsequent tasks.

Besides exploiting various contextual information from the visual cues [1–6], depth data have recently been utilized as supplementary information to RGB data to achieve improved segmentation accuracy [7–14]. Depth data naturally complements RGB signals by providing the 3D geometry to 2D visual information, which is robust to illumination changes and helps better distinguishing various objects. Especially in the environments where robots work, cluttered scenes may impede semantic segmentation. Incorporating depth images can alleviate this effect by providing complementary geometric information, as shown in [15–17].

Furthermore, in robot tasks, the accurate distinction of the boundaries of objects in the scene is more important than the accuracy of the overall cognition of the scene. However, in many studies of semantic segmentation, among the multitude of papers


F. Sun et al. (Eds.): ICCSIP 2021, CCIS 1515, pp. 213–224, 2022.
https://doi.org/10.1007/978-981-16-9247-5_16

contributing to the impressive 86% relative improvement in pixel accurancy (e.g., [18–22]), only a few address mask boundary quality.

Based on these insights we propose our Efficient Boundary-Aware Network (EBANet) which relies on both RGB and depth images as input. We design a boundary attention decoder branch to extract more boundary features of objects in the scene. A Canny edge detector is used to generate boundary-labeled images from the original segmentation labels for supervising the boundary attention branch. We also adopt a hybrid loss function fusing Cross-Entropy (CE) and structural similarity (SSIM) loss inspired by BASNet [23]. The hybrid loss guides the network to learn the transformation between the input image and the ground truth at the pixel and patch level, which helps the optimization to focus on the boundary. Our EBANet shows better performance than the state-of-the-art RGB-D segmentation methods as shown by our experiments (see Fig. 1).

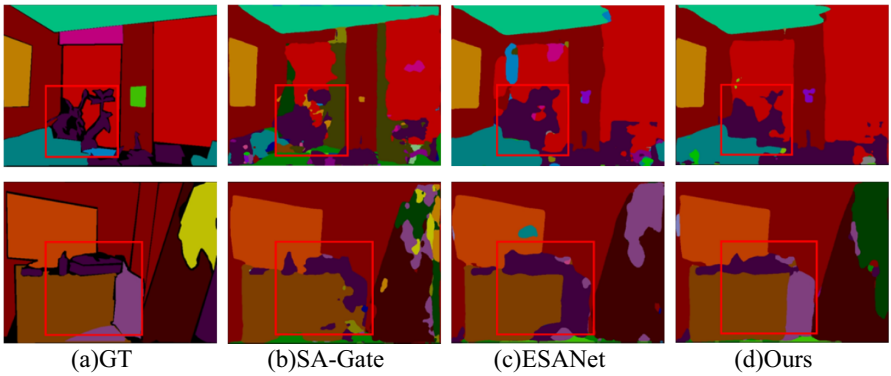


**Fig. 1.** Sample result of our method (EBANet) compared to SA-Gate [24] and ESANet [25]. Column (a) shows the ground truth (GT). (b), (c) and (d) are results of SA-Gate, ESANet and ours.

The main contributions of this paper are:

- A novel boundary-aware RGB-D semantic segmentation network: EBANet, which is a deeply supervised encoder-decoder architecture with dual decoder branches of semantic segmentation branch and boundary attention branch.
- A hybrid loss that fuses CE and SSIM to supervise the training process at the pixel and patch level to make the network focus on the boundary while training.
- A evaluation of the proposed method with a comparison with 2 state-of-the-art methods on the common RGB-D dataset NYUv2. Our method achieves state-of-the-art results in terms of mean Intersection over Union (mIoU).

## 2 Related Work

### 2.1 RGB-D Semantic Segmentation

With the development of depth sensors, recently there is a surge of interest in leveraging depth data for the semantic segmentation task, dubbed as RGB-D semantic segmentation [7–9, 26–28]. Depth images provide complementary geometric information to RGB images and, thus, improve segmentation [15, 16].

Incorporating depth information into RGB segmentation architectures is challenging as depth introduces deviating statistics and characteristics from another modality. The majority of approaches for RGB-D segmentation simply use two branches, one for RGB and one for depth data. Each branch can focus on extracting modality-specific features in this way. For example, the RGB branch extracts color and texture features while the depth branch extracts geometric and illumination-independent features. Then the feature representations are fused in the network. It leads to stronger feature representations to fuse these modality-specific features. [15] shows that if the features are fused at multiple stages, the segmentation performance increases. Typically, the features are fused once at each resolution stage with the last fusion at the end of both encoders. FuseNet [15], RedNet [16] and ESANet [25] fuse the depth features into the RGB encoder, which follows the intuition that the semantically richer RGB features can be further enhanced using complementary depth information. SA-Gate [24] combines RGB and depth features and fuses the recalibrated features back into both encoders. In order to make the two encoders independent of each other, ACNet [17] uses an additional, virtual, third encoder that obtains modality-specific features from the two encoders and processes the combined features. Instead of fusing in the encoder, the modality-specific features can also be used to refine the features in the common decoder via skip connections as in RDFNet [7], SSMA [29] and MMAF-Net [30].

In addition, the introduction of depth data has increased the scale of the network and the requirements for hardware. In [31], depth information is used to project the RGB images into a 3D space. However, it leads to significantly higher computational complexity to process the resulting 3D data. [8, 32–35] design specifically tailored convolutions, taking into account depth information, which often lack optimized implementations for hardware. [25] builds a light-weight network named Efficient Scene Analysis Network (ESANet) by exchanging the basic block in all ResNet layers with more efficient blocks and using a decoder that utilizes a novel learned upsampling. ESANet achieves the state-of-the-art performance with using fewer hardware resources than other methods. Following ESANet, we further improve the performance of the network at object boundaries by the attention mechanism.

### 2.2 Attention Mechanism

Attention mechanisms have been widely utilized in kinds of computer vision tasks, serving as the tools to spotlight the most representative and informative regions of input signals [2, 36–40]. For example, to improve the performance of the image/video classification task, SENet [36] introduces a self recalibrate gating mechanism by model importance among different channels of feature maps. Based on similar spirits, SKNet

[37] designs a channel-wise attention module to select kernel sizes to adaptively adjust its receptive field size based on multiple scales of input information. [38] introduces a non-local operation that explores the similarity of each pair of points in space. For the segmentation task, a well-designed attention module could encourage the network to learn helpful context information effectively. For instance, DFN [39] introduces a channel attention block to select the more discriminative features from multi-level feature maps to get more accurate semantic information. DANet [2] proposes two types of attention modules to model the semantic inter-dependencies in spatial and channel dimensions respectively. BANet [40] introduces a boundary feature mining branch to generate boundary feature maps which help the network focus on boundary area and extract low-level features selectively. However, BANet is an RGB-only network for the task of portrait segmentation and its boundary feature mining branch is not suitable for an RGB-D semantic segmentation network architecture. In our work, we design a boundary attention branch that extracts boundary features from both RGB and depth information as well as make the boundary features are fused with segmentation maps at multiple stages.

## 3 Method

This section starts with the architecture overview of our proposed EBANet. We describe the details of our designed boundary attention branch in Sect. 3.2. The formulation of our hybrid loss is presented in Sect. 3.3.

### 3.1 Overview

The proposed EBANet is a deeply supervised encoder-decoder architecture as shown in Fig. 2.

The RGB and depth encoder both use a ResNet architecture [41] as the backbone. Each $3 \times 3$ convolution is replaced by a $3 \times 1$ and a $1 \times 3$ convolution with a ReLU in-between, called Non-Bottleneck-1D-Block (NBt1D), which is shown that can simultaneously reduce inference time and increases segmentation performance in ESANet [25]. At each of the five resolution stages in the encoders (see Fig. 2), depth features are fused into the RGB encoder. The features from both modalities are first reweighted with a Squeeze and Excitation (SE) module [36] and then summed element-wisely. Using this channel attention mechanism, the model can learn which features of which modality to focus on and which to suppress, depending on the given input. Due to the limited receptive field of ResNet [42], a context module is used to incorporate context information by aggregating features at different scales by several branches.

Our decoder contains two branches with the similar structure, one is used to output the result of semantic segmentation, and the other is used to output the result of boundary detection. Each decoder uses the structure in ESANet which extends the one of SwiftNet [43], but different loss functions are used for supervision (See Sect. 3.3). 512 channels in the first decoder module are used and the number of channels in each $3 \times 3$ convolution is decreased as the resolution increases. Three additional Non-Bottleneck-1D-blocks are incorporated to further increase segmentation performance. Finally, the feature maps are unsampled by a factor of 2. A light-weight learned

upsampling method is used. In addition, skip connections from encoder to decoder stages of the same resolution are used. We will further introduce how the boundary attention branch works in Sect. 3.2.

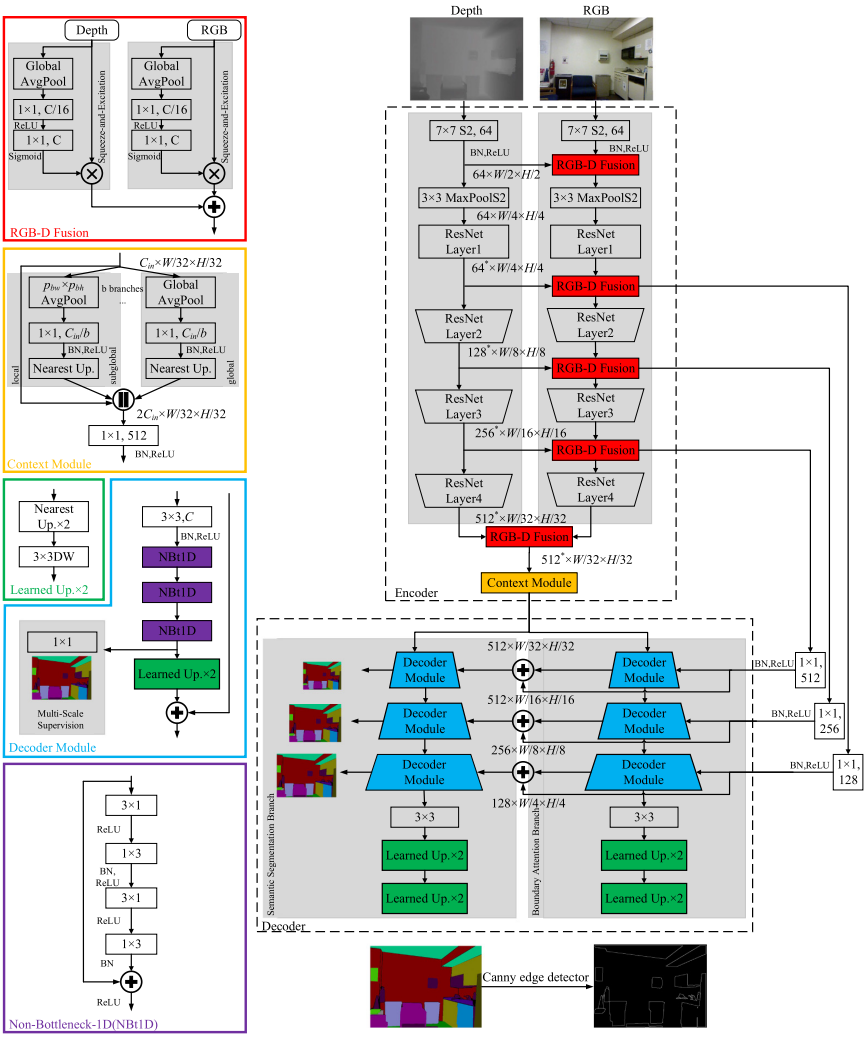


**Fig. 2.** Overview of our proposed EBANet.

## 3.2 Boundary Attention Branch

Our boundary attention branch is designed to introduce an attention mechanism to the boundaries of objects for the network. The boundary attention branch can be regarded as a copy of the original semantic segmentation branch. The difference is that the boundary attention branch only outputs a prediction of a single category, that is, it is judged

whether a certain pixel belongs to the object boundaries or not. It doesn't care what kind of object the pixel belongs to. Thus, after training, the boundary attention branch can further extract boundary features from the RGB-D features that output from the encoder. By adding the output of each decoder module of the boundary attention branch to the feature map of the corresponding size of the semantic segmentation branch, the signal of the boundary feature in the semantic segmentation branch is enhanced.

To train the boundary attention branch, we use a Canny edge detector to generate boundary labels from the original labels. The sigmoid function is used to normalize the network output. We use the Binary Cross-Entropy loss to guide the branch to learn the transformation between the RGB-D features and the boundary labels. It is defined as:

$$l_{boundary} = -\sum_{(r,c)} [G(r,c)\log(S(r,c) + (1 - G(r,c))\log(1 - S(r,c))] \tag{1}$$

where $G(r,c) \in \{0,1\}$ is the boundary label of the pixel $(r,c)$ and $S(r,c)$ is the predicted probability of being boundaries of objects.

### 3.3 Hybrid Loss

To obtain high-quality regional segmentation and clear boundaries, we introduce a hybrid loss into the semantic segmentation branch. The structural similarity index measure (SSIM) is a method for measuring the similarity between two images [44]. It is used to guide the network to learn structural information. SSIM loss is a patch-level measure, which considers a local neighborhood of each pixel. It assigns higher weights to the boundary, which is shown in BASNet [23]. It is defined as:

$$l_{SSIM} = 1 - \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \tag{2}$$

where $x = \{x_j : j = 1, ..., N^2\}$ and $y = \{y_j : j = 1, ..., N^2\}$ are the pixel values of two corresponding patches (size: $N \times N$) cropped from the predicted probability map S and the binary ground truth mask G respectively. $\mu_x, \mu_y$ and $\sigma_x, \sigma_y$ are the mean and standard deviations of $x$ and $y$ respectively, $\sigma_{xy}$ is their covariance, $C_1 = 0.01^2$ and $C_2 = 0.03^2$ are used to avoid dividing by zero.

In order to calculate SSIM loss, the labels need one-hot encoding. As the image size increases, these tensors will take up a lot of memory, which will affect training efficiency. Therefore, we did not use SSIM loss to supervise the final output of the semantic segmentation branch. Instead, we only use the Cross-Entropy (CE) loss [45] at the final output, and use the CE and SSIM loss at each decoder module in semantic segmentation. All these losses are summed to get the loss of the segmentation branch, which is defined as:

$$l_{segmentation} = l_{final} + \sum_{i}^{n} l_i \tag{3}$$

where $l_{final}$ is the loss of the final output and $l_i$ is the loss of the i-th decoder module output. Losses of the decoder module output is defined as:

$$l_{decoder} = l_{CE} + l_{SSIM} \tag{4}$$

where $l_{CE}$ is the CE loss. Our training loss is defined as the summation of all outputs:

$$l_{total} = \alpha \cdot l_{segmentation} + \beta \cdot l_{boundary} \tag{5}$$

where $\alpha, \beta$ are hyperparameters for adjusting the weights of the two branches.

## 4 Experiments

We evaluate our approach on the commonly used RGB-D dataset NYUv2 [46] and present an ablation study to essential parts of our network.

### 4.1 Datasets and Implementation Details

NYUv2 contains 1,449 indoor RGB-D images, of which 795 are used for training and 654 for testing. We used the common 40-class label setting. We used a network input resolution of 640 × 480 and applied median frequency class balancing [47]. As the input to the context module has a resolution of 20 × 15 due to the downsampling of 32, we used $b = 2$ branches, one with global average pooling and one with a pooling size of 4 × 3.

We trained our networks using PyTorch [48] for 500 epochs with batches of size 8. We use 2 RTX 2080ti GPU cards (with 11 GB memory) for both training and testing. For optimization, we used both SGD with momentum of 0:9 and Adam [49] with learning rates of {0.00125; 0.0025; 0.005; 0.01; 0.02; 0.04} and {0.0001; 0.0004}, respectively, and a small weight decay of 0:0001. We adapted the learning rate using PyTorch's onecycle learning rate scheduler. To further increase the number of training samples, we augmented the images using random scaling, cropping, and flipping. For RGB images, we also applied slight color jittering in HSV space. We used 0.9–1.1 times for jittering H and S while [−25, +25] for jittering V randomly. The best models were chosen based on the mIoU. And we get our best model when $\alpha = 1, \beta = 1.2$ in the total loss.

### 4.2 Results on NYUv2

Table 1 lists the results of our RGB-D approach for NYUv2 dataset. We compared the proposed EBANet with the current state-of-the-art methods, SA-Gate and ESANet. Our model achieves leading performance. According to the results, the performance of SA-Gate is far from the results given in the paper (52.4% for mIoU). The success of SA-Gate is due to a larger batch size during training. And this is difficult to reproduce under

our existing hardware conditions. Instead, we only use about a quarter of the hardware resources to achieve similar results. The effectiveness of our proposed method can be proven by the results.

**Table 1.** Mean intersection over union of our ESANet compared to state-of-the-art methods on NYUv2. The results in the table are produced in the environment we use. (*: SA-Gate achieved 51.4% mIoU using 8 NVIDIA TITAN V GPU cards with batches of size 16 according to the paper. For our 2 cards we can only training with batches of size 8.)

| Method | BackBone | mIoU (%) | Pixel Acc. (%) |
|---|---|---|---|
| SA-Gate* | 2 × ResNet101 | 48.94 | 75.80 |
| ESANet | 2 × ResNet50 | 50.00 | 75.93 |
| **EBANet (ours)** | 2 × ResNet50 | **51.51** | **76.82** |

**Table 2.** Ablation study for the proposed parts on NYUv2 test set.

| Model | Boundary Attention | Hybrid Loss | mIoU (%) |
|---|---|---|---|
| Baseline (ESANet) | / | / | 50.00 |
| Hybrid-loss-only | / | √ | 50.48 |
| Boundary-attention-only ( $\alpha$=1, $\beta$=1.2 ) | √ | / | 49.69 |
| EBANet ( $\alpha$=1, $\beta$=1.2 ) | √ | √ | **51.51** |

### 4.3 Ablation Study on NYUv2

Further, we perform ablation studies on the NYUv2 dataset under the same hyperparameters. Table 2 shows the study results for the proposed parts of our network architecture. ESANet is used as the baseline. There is no boundary attention branch in the hybrid-loss-only experiment, and only CE loss at the final output without hybrid loss is used in the boundary-attention-only experiment. According to the results, it can be seen that the improvement of network performance by using only hybrid loss is slight. However, using only boundary attention branch may have harmful effects on network performance. This is caused by hyperparameters $\alpha, \beta$. The hyperparameters that achieve better performance in EBANet may not be suitable for networks that do not use hybrid loss. At present, we have not studied their selection extensively, which will be one of our next research directions. However, the performance of EBANet, which uses both the boundary attention branch and the hybrid loss, has greatly improved. The results can show that boundary attention branch and hybrid loss play a mutually reinforcing role.

## 5 Conclusion

In this paper, we have presented a boundary-aware RGB-D segmentation approach, called EBANet, which is characterized by the boundary attention decoder branch and the hybrid loss function fusing Cross-Entropy (CE) and structural similarity (SSIM) loss to make the network extract more boundary features. On the common RGB-D dataset NYUv2, our EBANet achieves state-of-the-art results in terms of mIoU. Thus, it provides new possibilities for the application of semantic segmentation in robots.

**Acknowledgment.** This work was supported by Major Project of the New Generation of Artificial Intelligence (No. 2018AAA0102900).

## References

1. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, United States, pp. 431–440. IEEE Computer Society (2015)
2. Fu, J., et al.: Dual attention network for scene segmentation. In: 32nd IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, United States, pp. 3141–3149. IEEE Computer Society (2019)
3. He, J., Deng, Z., Qiao, Y.: Dynamic multi-scale filters for semantic segmentation. In: 17th IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Republic of Korea, pp. 3561–3571. Institute of Electrical and Electronics Engineers Inc., United States (2019)
4. Fu, J., et al.: Adaptive context network for scene parsing. In: 17th IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Republic of Korea, pp. 6747–6756. Institute of Electrical and Electronics Engineers Inc., United States (2019)
5. Cheng, B., et al.: SPGNet: semantic prediction guidance for scene parsing. In: 17th IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Republic of Korea, pp. 5217–5227. Institute of Electrical and Electronics Engineers Inc., United States (2019)
6. Zhang, F., et al.: ACFNet: attentional class feature network for semantic segmentation. In: 17th IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Republic of Korea, pp. 6797–6806. Institute of Electrical and Electronics Engineers Inc., United States (2019)
7. Lee, S., Park, S.J., Hong, K.S.: RDFNet: RGB-D multi-level residual feature fusion for indoor semantic segmentation. In: 16th IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, pp. 4990–4999. Institute of Electrical and Electronics Engineers Inc., United States (2017)
8. Wang, W., Neumann, U.: Depth-aware CNN for RGB-D segmentation. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11215, pp. 144–161. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01252-6_9
9. Zhang, Z., Cui, Z., Xu, C., Yan, Y., Sebe, N., Yang, J.: Pattern-affinitive propagation across depth, surface normal and semantic segmentation. In: 32nd IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, United States, pp. 4106–4115. IEEE Computer Society (2019)

10. Chen, Y., Mensink, T., Gavves, E.: 3D neighborhood convolution: learning depthaware features for RGB-D and RGB semantic segmentation. In: 7th International Conference on 3D Vision, 3DV 2019, Quebec, QC, Canada, pp. 173–182. Institute of Electrical and Electronics Engineers Inc., United States (2019)
11. He, Y., Chiu, W.C., Keuper, M., Fritz, M.: STD2P: RGBD semantic segmentation using spatio-temporal data-driven pooling. In: 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, United States, pp. 7158–7167. Institute of Electrical and Electronics Engineers Inc. (2017)
12. Li, Z., Gan, Y., Liang, X., Yu, Y., Cheng, H., Lin, L.: LSTM-CF: unifying context modeling and fusion with LSTMs for RGB-D scene labeling. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9906, pp. 541–557. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46475-6_34
13. Cheng, Y., Cai, R., Li, Z., Zhao, X., Huang, K.: Locality-sensitive deconvolution networks with gated fusion for RGB-D indoor semantic segmentation. In: 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, United States, pp. 1475–1483. Institute of Electrical and Electronics Engineers Inc. (2017)
14. Hung, S.W., Lo, S.Y., Hang, H.M.: Incorporating luminance, depth and color information by a fusion-based network for semantic segmentation. In: 26th IEEE International Conference on Image Processing, ICIP 2019, Taipei, Taiwan, pp. 2374–2378. IEEE Computer Society (2019)
15. Hazirbas, C., Ma, L., Domokos, C., Cremers, D.: FuseNet: incorporating depth into semantic segmentation via fusion-based CNN architecture. In: Lai, S.-H., Lepetit, V., Nishino, K., Sato, Y. (eds.) ACCV 2016. LNCS, vol. 10111, pp. 213–228. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-54181-5_14
16. 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)
17. Hu, X., Yang, K., Fei, L., Wang, K.: ACNet: attention based network to exploit complementary features for RGBD semantic segmentation. In: 26th IEEE International Conference on Image Processing, ICIP 2019, Taipei, Taiwan, pp. 1440–1444. IEEE Computer Society (2019)
18. Zhu, X., Hu, H., Lin, S., Dai, J.: Deformable ConvNets v2: more deformable, better results. In: 32nd IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, United States, pp. 9300–9308. IEEE Computer Society (2019)
19. Cai, Z., Vasconcelos, N.: Cascade R-CNN: delving into high quality object detection. In: 31st Meeting of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, United States, pp. 6154–6162. IEEE Computer Society (2018)
20. Bodla, N., Singh, B., Chellappa, R., Davis, L.S.: Soft-NMS - improving object detection with one line of code. In: 16th IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, pp. 5562–5570. Institute of Electrical and Electronics Engineers Inc., United States (2017)
21. Huang, Z., Huang, L., Gong, Y., Huang, C., Wang, X.: Mask scoring R-CNN. In: 32nd IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, United States, pp. 6402–6411. IEEE Computer Society (2019)
22. Li, Z., Zhuang, Y., Zhang, X., Yu, G., Sun, J.: COCO instance segmentation challenges 2018: winner (2018). http://presentations.cocodataset.org/ECCV18/COCO18-Detect-Megvii.pdf
23. Qin, X., Zhang, Z., Huang, C., Gao, C., Dehghan, M., Jagersand, M.: BASNet: boundary-aware salient object detection. In: 32nd IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, United States, pp. 7471–7481. IEEE Computer Society (2019)

24. Chen, X., et al.: Bi-directional cross-modality feature propagation with separation-and-aggregation gate for RGB-D semantic segmentation. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12356, pp. 561–577. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58621-8_33
25. Seichter, D., Köhler, M., Lewandowski, B., Wengefeld, T., Gross, H.: M Efficient RGB-D semantic segmentation for indoor scene analysis. arXiv preprint arXiv:2011.06961
26. Kong, S., Fowlkes, C.: Recurrent scene parsing with perspective understanding in the loop. In: 31st Meeting of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, United States, pp. 956–965. IEEE Computer Society (2018)
27. Lin, D., Chen, G., Cohen-Or, D., Heng, P.A., Huang, H.: Cascaded feature network for semantic segmentation of RGB-D images. In: 16th IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, pp. 1320–1328. Institute of Electrical and Electronics Engineers Inc., United States (2017)
28. Chen, X., Lin, K., Qian, C., Zeng, G., Li, H.: 3D sketch-aware semantic scene completion via semi-supervised structure prior. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Virtual, Online, United States, pp. 4192–4201. IEEE Computer Society (2020)
29. Valada, A., Mohan, R., Burgard, W.: Self-supervised model adaptation for multimodal semantic segmentation. Int. J. Comput. Vision **128**(5), 1239–1285 (2020)
30. Fooladgar, F., Kasaei, S.: Multi-modal attention-based fusion model for semantic segmentation of RGB-depth images. arXiv preprint arXiv:1912.11691 (2019)
31. Zhong, Y., Dai, Y., Li, H.: 3D geometry-aware semantic labeling of outdoor street scenes. In: 24th International Conference on Pattern Recognition, ICPR 2018, Beijing, China, pp. 2343–2349. Institute of Electrical and Electronics Engineers Inc., United States (2018)
32. Xing, Y., Wang, J., Chen, X., Zeng, G.: 2.5D convolution for RGB-D semantic segmentation. In: 26th IEEE International Conference on Image Processing, ICIP 2019, Taipei, Taiwan, pp. 1410–1414. IEEE Computer Society (2019)
33. Xing, Y., Wang, J., Zeng, G.: Malleable 2.5D convolution: learning receptive fields along the depth-axis for RGB-D scene parsing. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12364, pp. 555–571. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58529-7_33
34. Chen, L., Lin, Z., Wang, Z., Yang, Y.L., Cheng, M.M.: Spatial information guided convolution for real-time RGBD semantic segmentation. IEEE Trans. Image Process. **30** (2021), 2313–2324 (2021)
35. Chen, Y., Mensink, T., Gavves, E.: 3D neighborhood convolution: learning depth-aware features for RGB-D and RGB semantic segmentation. In: 7th International Conference on 3D Vision, 3DV 2019, Quebec, QC, Canada, pp. 173–182. Institute of Electrical and Electronics Engineers Inc., United States (2019)
36. Hu, J., Shen, L., Albanie, S., Sun, G., Wu, E.: Squeeze-and-excitation networks. IEEE Trans. Pattern Anal. Mach. Intell. **42**(8), 2011–2023 (2020)
37. Li, X., Wang, W., Hu, X., Yang, J.: Selective kernel networks. In: 32nd IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, United States, pp. 510–519. IEEE Computer Society (2019)
38. Wang, X., Girshick, R., Gupta, A., He, K.: Non-local neural networks. In: 31st Meeting of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, United States, pp. 7794–7803. IEEE Computer Society (2018)

39. Yu, C., Wang, J., Peng, C., Gao, C., Yu, G., Sang, N.: Learning a discriminative feature network for semantic segmentation. In: 31st Meeting of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, United States, pp. 1857–1866. IEEE Computer Society (2018)
40. Chen, X., Qi, D., Shen, J.: Boundary-aware network for fast and high-accuracy portrait segmentation. arXiv preprint arXiv:1901.03814
41. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: 29th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, United states, pp. 770–778. IEEE Computer Society (2016)
42. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, United states, pp. 6230–6239. Institute of Electrical and Electronics Engineers Inc., United States (2017)
43. Orsic, M., Kreso, I., Bevandic, P., Segvic, S.: In defense of pre-trained imagenet architectures for real-time semantic segmentation of road-driving images. In: 32nd IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, United states, pp. 12599–12608. IEEE Computer Society (2019)
44. Wang, Z., Simoncelli, E.P., Bovik, A.C.: Multiscale structural similarity for image quality assessment. In: The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, vol. 2, no. 2003, pp. 1398–1402 (2003)
45. Murphy, K.P.: Machine Learning: A Probabilistic Perspective. MIT Press, Cambridge (2012)
46. Silberman, N., Hoiem, D., Kohli, P., Fergus, R.: Indoor segmentation and support inference from RGBD images. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012. LNCS, vol. 7576, pp. 746–760. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-33715-4_54
47. Eigen, D., Fergus, R.: Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In: 15th IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, pp. 2650–2658. Institute of Electrical and Electronics Engineers Inc., United States (2015)
48. Paszke, A., et al.: PyTorch: an imperative style, high-performance deep learning library. In: Advances in Neural Information Processing Systems, vol. 32, no. 2019, pp. 8024–8035 (2019)
49. Kingma, D.P., Ba, J.L.: Adam: a method for stochastic optimization. In: 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, United states (2015)