DEPTH REMOVAL DISTILLATION FOR RGB-D SEMANTIC SEGMENTATION

Tiyu Fang[1], Zhen Liang[1], Xiuli Shao[2], Zihao Dong[1], Jinping Li[1*]*

[1]School of Information Science and Engineering, University of Jinan, Jinan 250022, China
[2]College of computer science, Nankai University, Tianjin 300350, China
[*]Corresponding Author: ise_dongzh@ujn.edu.cn, ise_lijp@ujn.edu.cn

ABSTRACT

RGB-D semantic segmentation is attracting wide attention due to its better performance than conventional RGB methods. However, most of RGB-D semantic segmentation methods need to acquire the real depth information for segmenting RGB images effectively. Therefore, it is extremely challenging to take full advantage of RGB-D semantic segmentation methods for segmenting RGB images without the depth input. To address this challenge, a general depth removal distillation method is proposed to remove depth dependence from RGB-D semantic segmentation model by knowledge distillation, which can be employed to any CNN-based segmentation network structure. Specifically, a depth-aware convolution is adopted to construct the teacher network for getting sufficient knowledge from RGB-D images. Then according to the structure consistency between depth-aware convolution and general convolution, the teacher network is used to transfer the learned knowledge to the student network with general convolutions by sharing parameters. Next, the student network makes up for the lack of depth in manner of learning by RGB images. Meantime, a Variable Temperature Cross Entropy (VTCE) loss function is proposed to further increase the accuracy of the student model by soft target distillation. Extensive experiments on NYUv2 and SUN RGB-D datasets demonstrate the superiority of our proposed approach.

Index Terms— RGB-D semantic segmentation, convolutional neural networks, knowledge distillation

1. INTRODUCTION

Semantic segmentation is a key problem in the computer vision tasks, which focuses on achieving accurate pixel-wise segmentation for different categories of objects. With the extensive use of convolutional neural networks (CNNs), significant advances in RGB semantic segmentation algorithms have been witnessed in recent years [1, 2, 3, 4]. However, although some optimized methods were proposed in RGB semantic segmentation from different aspects [5, 6, 7], its accuracy was still hardly improved only by RGB images due to the lack of spatial information. Therefore, RGB-D semantic segmentation receives wide attention since it can utilize depth to get the 3D spatial information [8, 9, 10, 11, 12]. To our regret, at present, most of RGB-D semantic segmentation methods [10, 11, 12] only segment RGB images accurately when the depth information is accessible. As a matter of fact, most installed cameras can not obtain the depth directly, which causes great difficulty for the application of RGB-D semantic segmentation methods. Therefore, it is a necessary and worthy study to make full use of RGB-D semantic segmentation methods for segmenting RGB images without depth information.

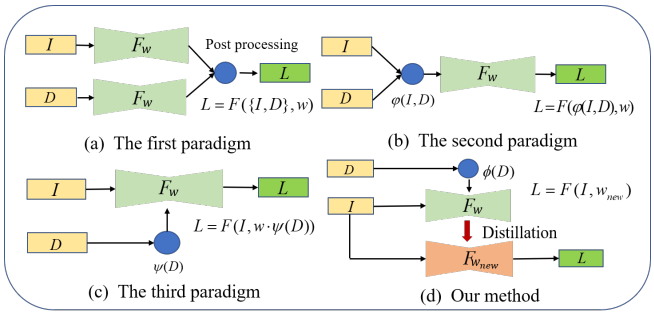

Fig. 1. The paradigms of RGB-D semantic segmentation.

Now, RGB-D semantic segmentation methods are mainly divided into three paradigms, which are shown in Fig. 1. For the convenience of description, let I, D and L denote RGB image, depth image and the obtained label respectively. The operation process of the network is represented by F and the weight in the network is represented by w. For the first paradigm [13, 14, 10], I and D are used as the input of network respectively, the segmentation results are obtained by combining the output of I and D through the network, which is the most common paradigm for RGB-D semantic segmentation. However, D always needs to be the input of the network, so this paradigm has a strong dependence on the depth. In the second paradigm [15, 11], I and D are fused as the input of network by preprocessing operation $\varphi(I, D)$, for example, I and D are encoded as HHA image [15]. Using the image, the network can predict accurate segmentation results. However, in the paradigm, the training and testing process can not be separated from depth information, so the paradigm cannot achieve depth dependence removal. Different from the

above two paradigms, in the third paradigm [16, 12], D becomes an auxiliary factor to optimize the weight w through $\psi(D)$ instead of the input of the network. It is possible to remove depth dependence for RGB-D semantic segmentation based on this paradigm. Inspired by this idea, we propose a general depth removal distillation method for RGB-D semantic segmentation, which aims at the depth dependence removal for RGB-D semantic segmentation.

In the method, the knowledge distillation is fully applied to solve depth dependence removal of model. In the depth image, the pixels occupied by an object often have similar depth values. According to the priori, a depth-aware convolution is constructed to form the teacher network, which is fully trained by RGB-D images for improving the accuracy. Then based on the characteristic that depth-aware convolution and general convolution have the same parameter structure, the teacher network transfers the learned knowledge to the student network with general convolutions by sharing parameters. Next, the student network adopts the learning strategy to make up for the lack of depth by RGB images. In addition, a Variable Temperature Cross Entropy (VTCE) loss function is proposed to further promote the performance of the student network by soft target distillation. Finally, the trained student network is able to segment accurately RGB images without real depth. Since the method mainly improves the basic convolution layer in CNNs, it can be applied to any CNN-based semantic segmentation structure.

2. PROPOSED METHOD

The proposed method is shown in Fig. 2, which is divided into two parts: teacher network and student network. Depth-aware convolution (D-Conv) is adopted to construct teacher network and general convolution (G-Conv) is used to construct student network with the same structure as teacher network.

2.1. Basic Principle

In our proposed method, for the teacher network, depth image D and RGB image I are used to train CNNs model with the network refining strategies ϕ for improving the model accuracy, and after training, the network parameter set w_{new} of w is gotten, which is expressed in the following form:

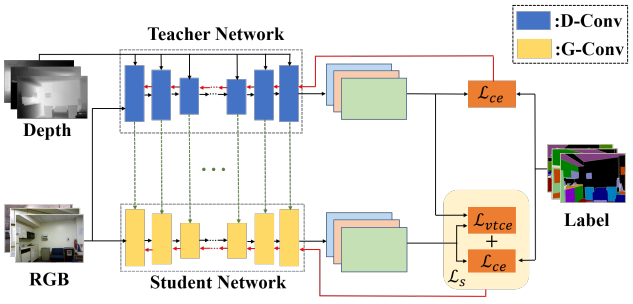

Fig. 2. The overall architecture of our proposed method.

$$w_{new} \leftarrow argmin_w(F(I, \overbrace{w \cdot \phi(D)}^{w_{best}}) - L_{gt}), \quad (1)$$

where L_{gt} is ground truth. For the fully trained teacher network, when the depth information exists, $w \cdot \phi(D)$ can be regarded as the parameter set w_{best}. Assuming that w_{best} is known, the accurate prediction results can be obtained only by using RGB images. However, when removing depth actually, we can only know the parameter set w_{new} of w at this time. Fortunately, from the above formula, it can be seen that there is a clear linear relationship between w_{new} and w_{best}. Based on the characteristic, the network parameter set w_{new} of teacher network is shared to student network and the learning-based strategy is adopted to adjust w_{new} by gradient descent in the student network, as shown below:

$$w_{new} = w_{new} - r\frac{\partial \mathcal{L}_s}{\partial w_{new}}, \quad (2)$$

where r is learning rate and $\mathcal{L}_s$ represents the loss of student network, which will be discussed in Section 2.3. Through iterative training, w_{new} can gradually tend to w_{best}. Using the optimized w_{new}, the final segmentation result is obtained, which is shown as follow:

$$L = F(I, w_{new}). \quad (3)$$

Base on the principle, the network construction strategy and loss function of student and teacher network are designed.

2.2. Network Construction Strategy

For the general convolution operation, it is assumed that input feature map is $X \in R^{(c_x \times h_c \times w_c)}$, output feature map is $Y \in R^{(c_y \times h_c \times w_c)}$, and the convolution kernel is $W \in R^{(l \times l)}$, where c_x is the number of the input feature channels, h_c is the height of the feature map, w_c is the width of the feature map, c_y is the number of the output feature channels and l is the size of the convolution kernel. Since 3-D convolution and 2-D convolution are based on same principle, for simplicity, we choose 2-D convolution to explain the principle. For any point (i_0, j_0) of the output feature map, we use i, j to traverse its corresponding convolution range. According to the above assumption, a general 2-D convolution operation can be expressed as:

$$Y(i_0, j_0) = \sum_{i,j}^{l \times l} W_{(i_0,j_0)}(i,j) X_{(i_0,j_0)}(i,j). \quad (4)$$

From the general convolution, we know W is mainly adjusted by the relationship between RGB images and labels. In fact, depth information also has a certain correspondence with label. Based on the priori, an auxiliary function $H \in R^{(h_c \times w_c)}$ is introduced to the general convolution for forming the depth-aware convolution, which is expressed as:

$$Y(i_0, j_0) = \sum_{i,j}^{l \times l} W_{(i_0,j_0)}(i,j) H_{(i_0,j_0)}(i,j) X_{(i_0,j_0)}(i,j). \quad (5)$$

In the depth D, the pixels occupied by an object often have

similar depth values. By using this peculiarity, H is defined in the form of Eq. 6. Specifically, the smaller depth difference is given a larger weight, and the larger depth difference is given a smaller weight.

$$H_{(i_0,j_0)}(i,j) = e^{-\alpha|D(i_0+i,j_0+j)-D(i_0,j_0)|}, \quad (6)$$

where α are a constant. By using the depth-aware convolution, the network can introduce depth into convolution operation and realize faster and more accurate convergence. Compared with general convolution, depth-aware convolution does not add any additional parameters, so the parameter sharing between them can be easily realized. In addition, after the parameters of depth-aware convolution are shared with general convolution, the accuracy reduction of general convolution due to the lack of depth information can be compensated by gradient descent (GD), as shown in Fig. 3.

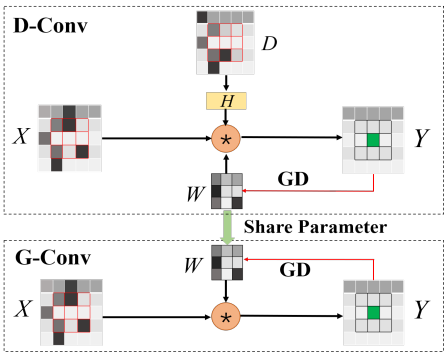

Fig. 3. The interactive process of depth-aware convolution and general convolution.

2.3. Loss Function

The loss of teacher network. In the process of training teacher network, the logit of each class z_i is transformed into the class probability q_i by the softmax layer, which is described as the following form:

$$q_i = \frac{e^{z_i}}{\sum_{j=1}^{n} e^{z_j}}, \quad (7)$$

where n is the number of classes. Based on the class probability q_i and the one-hot vector v_i of ground truth, the cross entropy loss is used to maximize the performance of teacher network, as shown below:

$$\mathcal{L}_{ce} = -\sum_{i=1}^{n} v_i log(q_i). \quad (8)$$

The loss of student network. For making the student network fully learn the knowledge of the teacher network, a Variable Temperature Cross Entropy (VTCE) loss is proposed to transfer the knowledge to the student network in manner of soft target distillation. In the loss, the class probability q_i is optimized in the following form:

$$q_i = \frac{e^{z_i/T}}{\sum_{j=1}^{n} e^{z_j/T}}, \quad (9)$$

where

$$T = e^{-\beta \frac{\lfloor \lambda/N \rfloor^2 - \lfloor \lambda/N \rfloor + 2}{2}}. \quad (10)$$

In T, β is a constant, λ is the current network iteration number and N is the number of training images. T is a variable that varies with λ and is used to control the impact of the teacher network output on the student network. The final VTCE loss is described as

$$\mathcal{L}_{vtce} = -\sum_{i=1}^{n} T q_i^t log(q_i^s), \quad (11)$$

where q_i^s and q_i^t represent the class probability of student network and teacher network respectively. Through the change of T, VTCE loss can simulate the distillation process and transfer knowledge of teacher network from coarse to fine.

In addition to the VTCE loss, the difference between the student network output and ground truth is also optimized by using the cross entropy loss $\mathcal{L}_{ce}$. Based on the above mentioned loss functions, the final loss is described as:

$$\mathcal{L}_s = \mathcal{L}_{ce} + \mathcal{L}_{vtce}. \quad (12)$$

3. EXPERIMENTS

3.1. Implementation Details

We evaluate the proposed method in the both NYUv2 dataset [17] and SUN RGB-D dataset [18]. For NYUv2 dataset, we take 795 pairs of images as training images and 654 images as test images. Similarly, for SUN RGB-D dataset, we use 5285 pairs of images as training images and 5050 images as test images. The encoding-decoding structure of DeepLab-VGG16 [2] is chosen as test model in this paper. Specifically, the convolutions of the structure are replaced by depth-aware convolutions to form the teacher network. And the network structure with general convolutions is used as the student network. The learning rate r is set to 0.00025, batch size is 1, α is set to 8.3 and β is set to 0.9. We use the following two metrics for evaluation: mean pixel accuracy (mPA) and mean Intersection over Union (mIoU).

3.2. Comparative Experiments

In NYUv2 and SUN RGB-D datasets, the accuracy of the proposed method is verified in two cases: the model is trained from scratch ('No Initialization') and the model is initialized by the pre-trained parameters of ImageNet ('Initialization'). We use RGB images as training and testing data to get the baseline results ('BL') in DeepLab-VGG16. Meantime, RGB-D images are applied to train and test the teacher network ('TN'), then the accuracy of student network ('SN') is gotten by training and testing RGB images. The final results are shown in Table 1 and Fig. 4. Compared with the baseline, TN and SN segment images more accurately, which proves the effectiveness of depth-aware convolution and shows that the proposed depth removal distillation can effectively maintain the model accuracy. In addition, we also compare our method with some existing state-of-the-art methods, as shown in Table 2. The final results show that the proposed depth removal distillation method gets better performance by RGB

images in a simple model structure.

Table 1. Segmentation results of experimental methods on NYUv2 and SUN RGB-D datasets.

	NYUv2					
	No Initialization			Initialization		
	BL	TN	SN	BL	TN	SN
Input Data	RGB	RGB-D	RGB	RGB	RGB-D	RGB
mPA(%)	22.8	48.2	**51.0**	35.0	51.6	**51.0**
mIoU(%)	15.2	32.3	**38.1**	24.6	39.1	**38.2**
	SUN RGB-D					
	No Initialization			Initialization		
	BL	TN	SN	BL	TN	SN
Input Data	RGB	RGB-D	RGB	RGB	RGB-D	RGB
mPA(%)	31.6	40.4	**39.3**	39.8	50.8	**48.9**
mIoU(%)	22.9	30.8	**28.5**	31.7	41.0	**39.5**

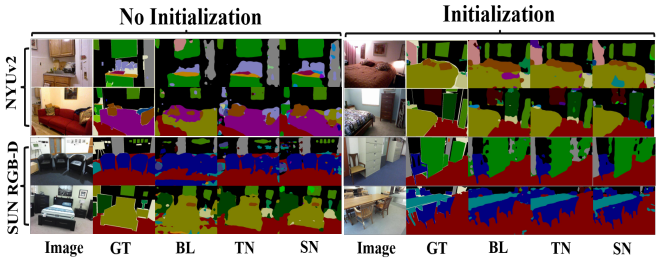

Fig. 4. The results of semantic segmentation on NYUv2 dataset and SUN RGB-D dataset, where GT is ground truth.

Table 2. The segmentation accuracy comparison of some existing methods on NYUv2 test set and SUN RGB-D test set. Network is initialized by the pre-trained parameters.

NYUv2			
Method	**Test Image**	**mPA(%)**	**mIoU(%)**
Deng et al.[19]	RGB-D	-	31.5
Wang et al.[13]	RGB-D	47.3	-
FCN[1]	RGB-D	46.1	34.0
Eigen et al.[20]	RGB	45.1	34.1
FCN[1]	RGB	44.7	31.6
B-SegNet[21]	RGB	45.8	32.4
PSPNet18[22]	RGB	46.9	35.9
TD^4-PSP18[6]	RGB	48.1	37.4
Ours	RGB	**51.0**	**38.2**
SUN RGB-D			
Method	**Test Image**	**mPA(%)**	**mIoU(%)**
Liu et al.[23]	RGB-D	10.0	-
Ren et al.[24]	RGB-D	36.3	-
LSTM-CF[25]	RGB-D	48.1	-
ParseNet [26]	RGB	-	34.7
DeconvNet [27]	RGB	33.3	22.6
FuseNet [14]	RGB	-	37.3
B-SegNet[21]	RGB	45.9	30.7
AdapNet [28]	RGB	-	32.5
DeepLab v3 [29]	RGB	-	35.7
AdapNet++ [7]	RGB	-	38.4
Ours	RGB	**48.9**	**39.5**

3.3. Ablation and Parameter Sensitivity Analysis

The accuracy of student network mainly comes from the proposed parameter sharing mechanism (PS) and VTCE loss. And each component of student network is explored on NYUv2 and SUN RGB-D datasets. The DeepLab-VGG16 with general convolutions is trained and tested from scratch by RGB image for obtaining the baseline results. Each component is added to get the final results, as shown in Table 3. The results show that the parameter sharing mechanism and VTCE loss can all improve the performance of segmentation. In particular, the parameter sharing mechanism greatly improves the segmentation accuracy compared with baseline.

Table 3. The component ablation of experimental method on NYUv2 and SUN RGB-D datasets.

	NYUv2		SUN RGB-D	
	mPA(%)	mIoU(%)	mPA(%)	mIoU(%)
BL	22.8	15.2	31.6	22.9
BL+VTCE	24.1	17.9	32.8	23.5
BL+PS	49.1	37.0	38.1	26.4
BL+PS+VTCE	**51.0**	**38.1**	**39.3**	**28.5**

Meantime, we also explore the sensitivity of parameters including α and β, which is shown in Table 4. Specifically, when α and β are set to different parameters in NYUv2 dataset, the effect of α is studied by training and testing the teacher network. In addition, the performance of β is discussed by training and testing the student network. The results show that when α is set to 8.3 and β is set to 0.9, the proposed method can achieve the best results.

Table 4. The sensitivity analysis of α and β on NYUv2 dataset.

α	0.1	7	8.3	10	15
mPA(%)	42.8	45.4	**48.2**	46.3	42.6
mIoU(%)	28.7	30.0	**32.3**	30.8	28.5
β	1	0.9	0.5	0.3	0.1
mPA(%)	50.5	**51.0**	49.5	49.2	48.7
mIoU(%)	37.5	**38.1**	37.7	37.1	36.5

4. CONCLUSION

In this paper, a depth removal distillation method is proposed to remove depth dependence from RGB-D semantic segmentation model. The knowledge distillation is fully used in the method. Specifically, a depth-aware convolution is constructed to form the teacher network, which is fully trained by RGB-D images. Next, based on the structure consistency between depth-aware convolution and general convolution, the parameter sharing mechanism and the learning strategy based on the proposed VTCE loss are applied for transfering the knowledge of the teacher network to the student network with general convolutions. The basic convolution operation is optimized in the proposed method, therefore, it can be employed to any CNN-based semantic segmentation network. In the future, the method can also be used to solve other multimodal classification problems in the computer vision.

5. REFERENCES

[1] Jonathan Long, Evan Shelhamer, and Trevor Darrell, "Fully convolutional networks for semantic segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 4, pp. 640–651, 2016.

[2] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille, "Semantic image segmentation with deep convolutional nets and fully connected crfs," in *ICLR*, 2015.

[3] Guosheng Lin, Anton Milan, Chunhua Shen, and Ian Reid, "Refinenet: Multi-path refinement networks for high-resolution semantic segmentation," in *CVPR*, 2017, pp. 1925–1934.

[4] Xiong Zhang, Hongmin Xu, Hong Mo, Jianchao Tan, Cheng Yang, Lei Wang, and Wenqi Ren, "Dcnas: Densely connected neural architecture search for semantic image segmentation," in *CVPR*, 2021, pp. 13956–13967.

[5] Guosheng Lin, Chunhua Shen, Anton Van Den Hengel, and Ian Reid, "Efficient piecewise training of deep structured models for semantic segmentation," in *CVPR*, 2016, pp. 3194–3203.

[6] Ping Hu, Fabian Caba, Oliver Wang, Zhe Lin, Stan Sclaroff, and Federico Perazzi, "Temporally distributed networks for fast video semantic segmentation," in *CVPR*, 2020, pp. 8818–8827.

[7] Abhinav Valada, Rohit Mohan, and Wolfram Burgard, "Self-supervised model adaptation for multimodal semantic segmentation," *International Journal of Computer Vision*, pp. 1–47, 2019.

[8] Zhenyu Zhang, Zhen Cui, Chunyan Xu, Zequn Jie, Xiang Li, and Jian Yang, "Joint task-recursive learning for semantic segmentation and depth estimation," in *ECCV*, 2018, pp. 235–251.

[9] Jianbo Jiao, Yunchao Wei, Zequn Jie, Honghui Shi, Rynson WH Lau, and Thomas S Huang, "Geometry-aware distillation for indoor semantic segmentation," in *CVPR*, 2019, pp. 2869–2878.

[10] Hao Zhou, Lu Qi, Zhaoliang Wan, Hai Huang, and Xu Yang, "Rgb-d co-attention network for semantic segmentation," in *ACCV*, 2020.

[11] Xiaokang Chen, Kwan-Yee Lin, Jingbo Wang, Wayne Wu, Chen Qian, Hongsheng Li, and Gang Zeng, "Bi-directional cross-modality feature propagation with separation-and-aggregation gate for rgb-d semantic segmentation," in *ECCV*, 2020, pp. 561–577.

[12] Lin-Zhuo Chen, Zheng Lin, Ziqin Wang, Yong-Liang Yang, and Ming-Ming Cheng, "Spatial information guided convolution for real-time rgbd semantic segmentation," *IEEE Transactions on Image Processing*, vol. 30, pp. 2313–2324, 2021.

[13] Jinghua Wang, Zhenhua Wang, Dacheng Tao, Simon See, and Gang Wang, "Learning common and specific features for rgb-d semantic segmentation with deconvolutional networks," in *ECCV*, 2016, pp. 664–679.

[14] Caner Hazirbas, Lingni Ma, Csaba Domokos, and Daniel Cremers, "Fusenet: Incorporating depth into semantic segmentation via fusion-based cnn architecture," in *ACCV*, 2016, pp. 213–228.

[15] Saurabh Gupta, Ross Girshick, Pablo Arbeláez, and Jitendra Malik, "Learning rich features from rgb-d images for object detection and segmentation," in *ECCV*, 2014, pp. 345–360.

[16] Weiyue Wang and Ulrich Neumann, "Depth-aware cnn for rgb-d segmentation," in *ECCV*, 2018, pp. 135–150.

[17] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus, "Indoor segmentation and support inference from rgbd images," in *ECCV*, 2012, pp. 746–760.

[18] Shuran Song, Samuel P Lichtenberg, and Jianxiong Xiao, "Sun rgb-d: A rgb-d scene understanding benchmark suite," in *CVPR*, 2015, pp. 567–576.

[19] Zhuo Deng, Sinisa Todorovic, and Longin Jan Latecki, "Semantic segmentation of rgbd images with mutex constraints," in *ICCV*, 2015, pp. 1733–1741.

[20] David Eigen and Rob Fergus, "Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture," in *ICCV*, 2015, pp. 2650–2658.

[21] Alex Kendall, Vijay Badrinarayanan, and Roberto Cipolla, "Bayesian segnet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding," in *BMVC*, 2017.

[22] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia, "Pyramid scene parsing network," in *CVPR*, 2017, pp. 2881–2890.

[23] Ce Liu, Jenny Yuen, and Antonio Torralba, "Sift flow: dense correspondence across scenes and its applications.," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 5, pp. 978–994, 2011.

[24] Xiaofeng Ren, Liefeng Bo, and Dieter Fox, "Rgb-(d) scene labeling: Features and algorithms," in *CVPR*, 2012, pp. 2759–2766.

[25] Zhen Li, Yukang Gan, Xiaodan Liang, Yizhou Yu, Hui Cheng, and Liang Lin, "Lstm-cf: Unifying context modeling and fusion with lstms for rgb-d scene labeling," in *ECCV*, 2016, pp. 541–557.

[26] Wei Liu, Andrew Rabinovich, and Alexander C Berg, "Parsenet: Looking wider to see better," *arXiv preprint arXiv:1506.04579*, 2015.

[27] Hyeonwoo Noh, Seunghoon Hong, and Bohyung Han, "Learning deconvolution network for semantic segmentation," in *ICCV*, 2015, pp. 1520–1528.

[28] Abhinav Valada, Johan Vertens, Ankit Dhall, and Wolfram Burgard, "Adapnet: Adaptive semantic segmentation in adverse environmental conditions," in *ICRA*, 2017, pp. 4644–4651.

[29] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *ECCV*, 2018, pp. 801–818.