Non-local Aggregation for RGB-D Semantic Segmentation

Guodong Zhang, Jing-Hao Xue, *Member, IEEE*, Pengwei Xie, Sifan Yang, and Guijin Wang, *Senior Member, IEEE*

***Abstract*—Exploiting both RGB (2D appearance) and Depth (3D geometry) information can improve the performance of semantic segmentation. However, due to the inherent difference between the RGB and Depth information, it remains a challenging problem in how to integrate RGB-D features effectively. In this letter, to address this issue, we propose a Non-local Aggregation Network (NANet), with a well-designed Multi-modality Non-local Aggregation Module (MNAM), to better exploit the non-local context of RGB-D features at multi-stage. Compared with most existing RGB-D semantic segmentation schemes, which only exploit local RGB-D features, the MNAM enables the aggregation of non-local RGB-D information along both spatial and channel dimensions. The proposed NANet achieves comparable performances with state-of-the-art methods on popular RGB-D benchmarks, NYUDv2 and SUN-RGBD.**

***Index Terms*—Convolutional neural network, RGB-D semantic segmentation, Multi-modality feature fusion.**

I. Introduction

IMAGE semantic segmentation, aiming at assigning a semantic category to each pixel in an image, is a fundamental task of computer vision. With remarkable achievements [1]–[3], it has been widely applied in autonomous driving [4], virtual reality [5] and medical diagnosis [6] to name a few. In recent years, the development of commercial RGB-D sensors has leveraged additional 3D geometric information to further improve the performance of semantic segmentation.

The pioneering work of RGB-D semantic segmentation, FuseNet [7], makes use of convolutional neural networks (CNNs) to incorporate complementary depth information into a semantic segmentation framework. Recently, some works achieve great progress by adaptively extracting or fusing RGB-D features. RDFNet [8] proposed to capture multi-scale RGB-D features via multi-modality feature fusion blocks and multi-scale feature refinement blocks. Jiao *et al.* [9] proposed to improve the quality of semantic segmentation by distilling geometry-aware embedding. [10] adaptively fused RGB-D features by replacing identity mappings with idempotent mappings, but the mapping matrix is hard to choose. ACNet [11] proposed an Attention Complementary Module to extract weighted features from RGB and Depth branches, but it lacks long-range cross-modality dependencies. Chen *et al.* [12] proposed an SA-Gate unit to ensure cross-modality features aggregation via channel-wise attention mechanism, but it lacks non-local spatial cross-modality interaction, which is profoundly important for RGB-D semantic segmentation. CANet [13] proposed to take advantages of long-range cross-modality interdependencies via position and channel attention modules, but it only aggregates the non-local cross-modality features at the final stage of the encoder, which cannot exploit multi-scale non-local cross-modality information.

In this letter, we propose a Non-local Aggregation Network (NANet) to better exploit the non-local context of RGB-D features at multi-stage. Particularly, a new Multi-modality Non-local Aggregation Module (MNAM) is designed to capture both spatial and channel-wise long-range dependencies in the NANet. In detail, the MNAM enlarges the receptive fields along the horizontal and vertical spatial dimensions to capture non-local spatial context. Furthermore, we adopt channel-wise global average pooling and multi-layer perceptron (MLP) to capture cross-modality channel-wise dependencies. By repeating this aggregation process several times, comprehensive long-range dependencies can be built over the whole RGB-D image. Extensive experiments on two challenging RGB-D semantic segmentation benchmarks, NYUDv2 [14] and SUN-RGBD [15], confirm the effectiveness of our NANet.

In short, the main contributions of our work are two-fold. 1) We propose a non-local aggregation network called NANet, which achieves comparable performances with state-of-the-art methods on popular RGB-D benchmarks, NYUDv2 and SUN-RGBD. 2) We propose a novel multi-modality non-local aggregation module called MNAM, which can effectively integrate non-local RGB-D features along different dimensions.

II. Method

This section first presents an overview of the proposed NANet, and then explains technical details about how to effectively aggregate non-local RGB-D features via the MNAM.

A. Overview of Non-local Aggregation Network

As shown in Fig. 1, the NANet incorporates several MNAM into a two-stream backbone network, which enables a multi-stage exploitation of non-local RGB-D information. Additionally, we also use the intermediate non-local RGB-D features, from the MNAM at the fourth stage, to generate coarse segmentation results supervised by an auxiliary loss to facilitate the learning.

This work was partially supported by the Beijing Advanced Innovation Center for Future Chip (ICFC). (Corresponding author: Guijin Wang.)

G. Zhang and S. Yang are with the Shenzhen International Graduate School / Department of Electronic Engineering, Tsinghua University, Shenzhen 518055, China (e-mail: zhanggd18@mails.tsinghua.edu.cn; fyang@sz.tsinghua.edu.cn).

J.-H. Xue is with the Department of Statistical Science, University College London, London WC1E 6BT, U.K (e-mail: jinghao.xue@ucl.ac.uk).

P. Xie and G. Wang are with the Department of Electronic Engineering, Beijing National Research Center for Information Science and Technology, Tsinghua University, Beijing 100084, China (e-mail: xpw18@mails.tsinghua.edu.cn; wangguijin@tsinghua.edu.cn).

Differing from [8]–[13], the proposed NANet has two novelties. First, the MNAM, in the NANet, has a strong capability of modeling both spatial and channel-wise long-range dependencies. Secondly, the NANet fully exploits multi-stage RGB-D features for non-local information aggregation.

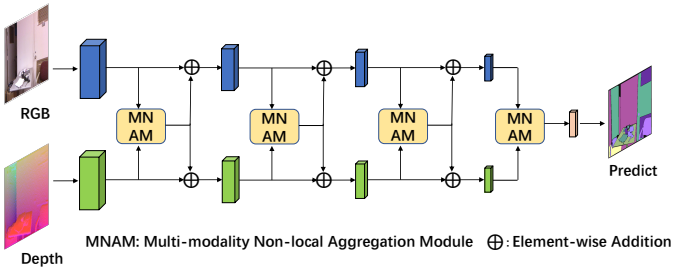

Fig. 1. Non-local Aggregation Network.

B. Multi-modality Non-local Aggregation Module

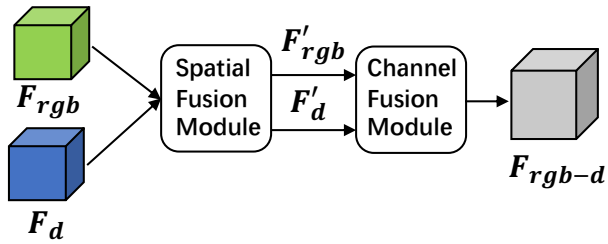

Fig. 2. Schematic diagram of the MNAM.

As illustrated in Fig. 2, the MNAM first models the RGB-D long-range dependencies in the spatial dimension by a Spatial Fusion Module (SFM), and then models the cross-modality context dependencies along the channel dimension by a Channel Fusion Module (CFM).

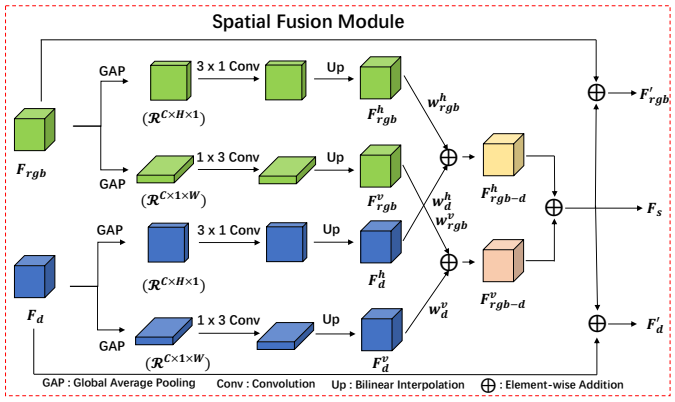

Fig. 3. Schematic diagram of the Spatial Fusion Module.

1) Spatial Fusion Module: As shown in Fig. 3, given raw RGB-D features ($F_{rgb} \in \mathcal{R}^{C\times H\times W}$ and $F_d \in \mathcal{R}^{C\times H\times W}$), where C is the channel number and H and W are the spatial height and width, for each spatial location, the SFM is first used to aggregate long-range contextual information of RGB and Depth features along the horizontal and vertical spatial dimensions, respectively. The non-local information for features ($F \in \mathcal{R}^{C\times H\times W}$) along vertical dimension ($F^v \in \mathcal{R}^{C\times 1\times W}$) and horizontal dimension ($F^h \in \mathcal{R}^{C\times H\times 1}$) can be obtained via global average pooling as

$$F^v(c,1,j) = \frac{1}{H}\sum_{i=0}^{H-1} F(c,i,j), \tag{1}$$

$$F^h(c,i,1) = \frac{1}{W}\sum_{j=0}^{W-1} F(c,i,j). \tag{2}$$

Then, 1×3 and 3×1 convolutions are used to enlarge the receptive fields along vertical and horizontal spatial dimensions, respectively. And these global priors are expanded into the original dimension via bilinear interpolation. In this way, each locations in F^v and F^h can build relationships with the pixels in F that are with the near horizontal or vertical coordinate.

After aggregating the non-local features of raw RGB and Depth features, respectively (*i.e.*, $F_{rgb} \rightarrow {F^v}_{rgb}$, ${F^h}_{rgb}$; and $F_d \rightarrow {F^v}_d$, ${F^h}_d$), we use a softmax function to adaptively get the weight (w^v_{rgb}, w^v_d, w^h_{rgb}, and w^h_d) for ${F^v}_{rgb}$, ${F^v}_d$, ${F^h}_{rgb}$, and ${F^h}_d$ according to their spatial responses (Eq. 3-Eq. 6). Then, the merged non-local features (${F^v}_{rgb\text{-}d}$ and ${F^h}_{rgb\text{-}d}$) can be obtained as the weighted sum (Eq. 7 and Eq. 8), respectively. In this way, the modality (RGB or Depth), having a stronger response, will make more contribution to the ${F^v}_{rgb\text{-}d}$ and ${F^h}_{rgb\text{-}d}$. This mechanism effectively exploits the complementarity of cross-modality features.

$$w^v_{rgb}(c,i,j) = \frac{e^{F^v_{rgb}(c,i,j)}}{e^{F^v_{rgb}(c,i,j)} + e^{F^v_d(c,i,j)}}, \tag{3}$$

$$w^v_d(c,i,j) = 1 - w^v_{rgb}(c,i,j), \tag{4}$$

$$w^h_{rgb}(c,i,j) = \frac{e^{F^h_{rgb}(c,i,j)}}{e^{F^h_{rgb}(c,i,j)} + e^{F^h_d(c,i,j)}}, \tag{5}$$

$$w^h_d(c,i,j) = 1 - w^h_{rgb}(c,i,j). \tag{6}$$

$$\begin{aligned} {F^v}_{rgb-d}(c,i,j) &= w^v_{rgb}(c,i,j)\cdot {F^v}_{rgb}(c,i,j) \\ &+ w^v_d(c,i,j)\cdot {F^v}_d(c,i,j), \end{aligned} \tag{7}$$

$$\begin{aligned} {F^h}_{rgb-d}(c,i,j) &= w^h_{rgb}(c,i,j)\cdot {F^h}_{rgb}(c,i,j) \\ &+ w^h_d(c,i,j)\cdot {F^h}_d(c,i,j), \end{aligned} \tag{8}$$

Finally, the non-local RGB-D spatial features (F_s) can be written as

$$F_s = F^v_{rgb-d} + F^h_{rgb-d}. \tag{9}$$

That is, we eventually embed the non-local RGB-D spatial features (F_s) into the raw RGB-D features (F_{rgb} and F_d) to integrate the local RGB-D information flow and the non-local RGB-D spatial information (Fig. 3).

To facilitate the understanding of the SFM, we illustrate the receptive fields on the SFM in Fig. 4. Each position in the raw RGB-D features (F_{rgb} and F_d) is allowed to build relationships with various positions (in the red bounding box) via the SFM.

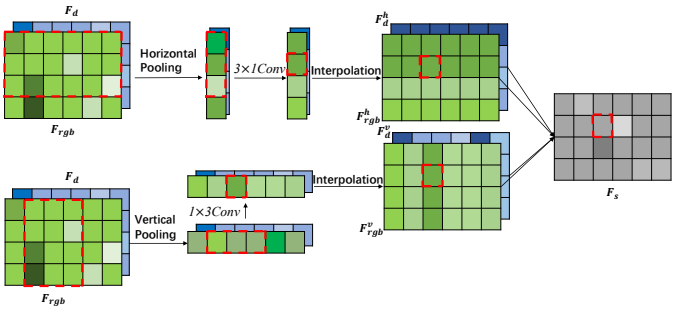

Fig. 4. Illustration of the receptive fields on the SFM.

Repeating this aggregation process several times, during the features extraction stage, will enable us to build long-range dependencies over the whole RGB-D image.

2) Channel Fusion Module: We also propose a CFM to exploit cross-modality channel dependencies.

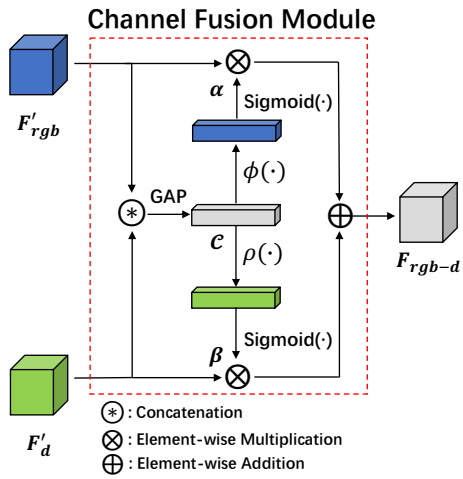

Fig. 5. Schematic diagram of the Channel Fusion Module.

As shown in Fig. 5, the proposed CFM first captures RGB-D channel-wise responses ($\mathcal{C} \in \mathcal{R}^{2C\times1\times1}$):

$$\mathcal{C} = GAP(Cat(F^{'}_{rgb}, F^{'}_{d})), \tag{10}$$

where $Cat(\cdot)$ denotes the concatenation operation along the channel dimension, and $GAP(\cdot)$ denotes the global average pooling along the channel-wise dimension.

Then, the CFM models the dependencies weight ($\alpha \in \mathcal{R}^{C\times1\times1}$ and $\beta \in \mathcal{R}^{C\times1\times1}$) between cross-modality channels. The α and β can be written as

$$\alpha = Sigmoid(\phi(\mathcal{C})), \tag{11}$$

$$\beta = Sigmoid(\rho(\mathcal{C})), \tag{12}$$

where $Sigmoid(\cdot)$ is the activation function, and $\phi(\cdot)$ and $\rho(\cdot)$ are two fully connected layers which can adaptively transform $\mathcal{C}$ to different embeddings ($\mathcal{R}^{2C\times1\times1} \rightarrow \mathcal{R}^{C\times1\times1}$).

Finally, the fused RGB-D features ($F_{rgb-d} \in \mathcal{R}^{C\times H\times W}$) can be written as

$$F_{rgb-d} = \alpha \otimes F^{'}_{rgb} + \beta \otimes F^{'}_{d}, \tag{13}$$

where $\otimes$ is the channel-wise multiplication.

The proposed CFM can learn a nonlinear interaction between cross-modality channels, which models the cross-modality channel-wise dependencies comprehensively.

III. Experimental Results

In this section, we conduct extensive experiments to verify the effectiveness of the proposed method.

1) Datasets: We conduct experiments on two popular benchmark datasets: NYUDv2 [14] and SUN-RGBD [15]. The NYUDv2 dataset contains 1,449 RGB-D images with 40 classes, in which 795 images are used for training, and the rest 654 images are for testing. The SUN-RGBD dataset has 37 categories and contains 10,335 RGB-D images (5,285 images for training and 5,050 for testing).

2) Metrics: To evaluate the performance of different methods, we use the prevailing pixel accuracy (*Pixel Acc.* $= \sum_i \frac{n_{ii}}{s}$) and mean Intersection-over-Union ($mIoU = \frac{1}{n_c}\sum_i \frac{n_{ii}}{s_i - n_{ii} + \sum_j n_{ji}}$), where n_{ij} is the number of pixels with ground-truth class i predicted as class j; n_c is the total number of classes; s_i is the number of pixels with ground truth class i; and s is the total number of all pixels.

3) Implementation Details: We use the PyTorch framework. The Depth images are encoded into HHA [16] images. For a fair comparison, following [10]–[12], [17], we use the DeepLab V3+ [1] as the baseline. All the backbone networks (ResNet-50 [18] and ResNet-101) are pre-trained on ImageNet dataset [19]. For training, we set the initial learning rate as 0.02, weight decay as 0.0005, crop size as 480×480. We use a batch size of 16 and train the proposed model for 400/100 epochs on the NYUDv2 [14]/ SUN-RGBD [15], respectively. Notably, we choose the polynomial learning rate policy with factor $(1 - \frac{iter}{iter_{max}})^{0.9}$. Our loss function is composed by two cross entropy losses. The weight on the final output loss is 1, and the weight on the auxiliary loss (the output layer of the fourth MNAM) is 0.2. For data augmentation, we use random horizontal flipping, scaling with scale $\in \{0.75, 1, 1.25\}$. When compared with state-of-the-art methods, we adopt flipping and multi-scale inference strategies as a test-time augmentation.

A. Comparisons with State-of-the-arts

As shown in Table I, the proposed NANet achieves leading performance. The ResNet-101 based NANet achieves an mIoU score of 52.3% on the NYUDv2 dataset. Moreover, our ResNet-50 based model still outperforms many ResNet-101 or ResNet-152 based models, indicating the effectiveness of aggregating the non-local RGB-D features. In addition, as shown in Table II, our ResNet-101 based NANet also performs well on the SUN-RGBD dataset, achieving an mIoU score of 48.8%.

B. Ablation Studies

To demonstrate the proposed MNAM's effectiveness, we conduct various ablation studies on the NYUDv2 dataset.

In the first ablation study, we gradually embed MNAM behind different stages of ResNet50 to verify the impact of MNAM at different stages. We average the predictions of two

TABLE I
COMPARISONS WITH THE STATE-OF-THE-ARTS ON THE NYUDV2 [14] DATASET. TOP TWO ARE IN BOLD.

Method	Input	Backbone	mIoU(%)	Pixel Acc.(%)
RefineNet [2]	RGB	ResNet-152	46.5	73.6
Jiao *et al.* [9]	RGB	ResNet-50	**59.6**	**84.8**
ACNet [11]	RGB-D	ResNet-50	48.3	-
D-CNN [17]	RGB-D	ResNet-152	48.4	-
RDF-101 [8]	RGB-D	ResNet-101	49.1	75.6
PAD-Net [20]	RGB-D	ResNet-50	50.2	75.2
PAP [21]	RGB-D	ResNet-50	50.4	76.2
Xing *et al.* [10]	RGB-D	ResNet-101	50.6	76.3
CANet [13]	RGB-D	ResNet-101	51.2	76.6
Chen *et al.* [12]	RGB-D	ResNet-50	51.3	-
NANet	RGB-D	ResNet-50	51.4	77.1
NANet	RGB-D	ResNet-101	**52.3**	**77.9**

TABLE II
COMPARISONS WITH THE STATE-OF-THE-ARTS ON THE SUN-RGBD DATASET [15]. TOP TWO ARE IN BOLD.

Method	Input	Backbone	mIoU(%)	Pixel Acc.(%)
RefineNet [2]	RGB	ResNet-152	45.9	80.6
Jiao *et al.* [9]	RGB	ResNet-50	**54.5**	**85.5**
Kong *et al.* [22]	RGB-D	ResNet-50	45.1	80.3
3DGNN [23]	RGB-D	ResNet-101	45.9	-
RDF-152 [8]	RGB-D	ResNet-152	47.7	81.5
RedNet [24]	RGB-D	ResNet-50	47.8	81.3
CFN [25]	RGB-D	ResNet-152	48.1	-
ACNet [11]	RGB-D	ResNet-50	48.1	-
CANet [13]	RGB-D	ResNet-50	48.1	81.6
NANet	RGB-D	ResNet-50	48.0	82.1
NANet	RGB-D	ResNet-101	**48.8**	**82.3**

TABLE III
ABLATION STUDIES ON THE NYUDV2 DATASET FOR MNAM BEHIND DIFFERENT STAGES OF RESNET-50.

Method	Stage1	Stage2	Stage3	Stage4	mIoU(%)
ResNet50					46.3
ResNet50	✓				47.8
ResNet50		✓			48.1
ResNet50			✓		47.7
ResNet50				✓	47.1
ResNet50	✓	✓			48.3
ResNet50	✓	✓	✓		48.8
ResNet50	✓	✓	✓	✓	49.4

parallel ResNet-50 as the final segmentation result. As shown in Table III, when stacking MNAM stage by stage, the ResNet-50 can be boosted continuously, indicating the effectiveness of using multi-scale non-local cross-modality features.

In the second ablation study, we average the predictions of two parallel ResNet-50 based DeepLab V3+ [1] as the baseline. As shown in Table IV, both the CFM and the SFM

TABLE IV
ABLATION STUDIES ON THE NYUDV2 DATASET FOR CFM, SFM AND MNAM.

Method	mIoU(%)	Pixel Acc.(%)
Baseline	47.8	75.2
Baseline + CFM	49.6	76.4
Baseline + SFM	50.5	76.8
Baseline + MNAM	**51.4**	**77.1**

can boost performance via aggregating cross-modality non-local dependencies. Moreover, by embedding the MNAM at multi-stage, the baseline's performance can be boosted by about 3.6% in mIoU (from 47.8% to 51.4%).

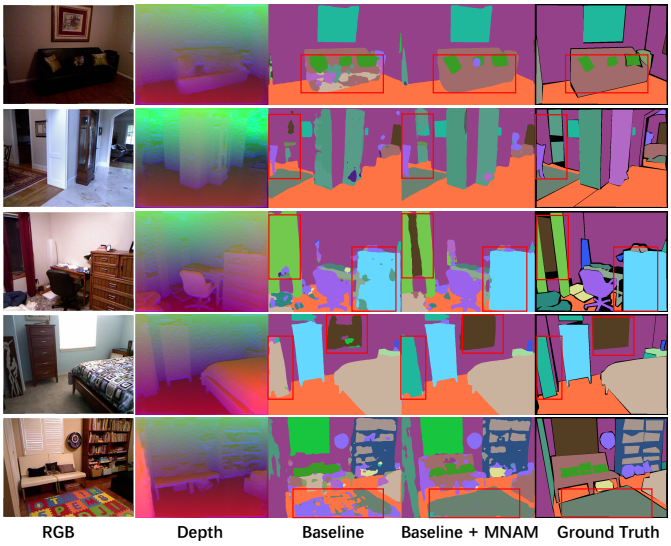

Fig. 6. Visualization of results on the NYUDv2 dataset. The red boxes mark where our method (fourth column) is superior to the baseline method (third column).

Finally, in Fig. 6, we display some qualitative results on the NYUDv2 dataset, which clearly demonstrate that adding the MNAM to the baseline can effectively improve the semantic consistency. For example, the baseline is troubled with local inconsistency on large objects like sofa (first row), wall (second row) and cabinet (third row), window (fourth row), *etc.*, while our method can mitigate this issue.

IV. CONCLUSION

In this letter, we propose a non-local aggregation network (NANet) to better exploit the non-local context of RGB-D features at multi-stage for RGB-D semantic segmentation. Particularly, a new multi-modality non-local aggregation module (MNAM), embedded in multi-stage of the NANet, is well designed to capture both spatial and channel-wise long-range dependencies in RGB-D features. Extensive experimental results confirm the effectiveness of our method on the popular NYUDv2 and SUN-RGBD datasets.

REFERENCES

[1] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 833–851.

[2] G. Lin, A. Milan, C. Shen, and I. Reid, "Refinenet: Multi-path refinement networks for high-resolution semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 5168–5177.

[3] Q. Hou, L. Zhang, M. M. Cheng, and J. Feng, "Strip pooling: Rethinking spatial pooling for scene parsing," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 4002–4011.

[4] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 3213–3223.

[5] X. Hu, G. Wang, J.-S. Hyun, Y. Zhang, H. Yang, and S. Zhang, "Autofocusing method for high-resolution three-dimensional profilometry," *Opt. Lett.*, vol. 45, no. 2, pp. 375–378, 2020.

[6] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention*, 2015, pp. 234–241.

[7] C. Hazirbas, L. Ma, C. Domokos, and D. Cremers, "Fusenet: Incorporating depth into semantic segmentation via fusion-based cnn architecture," in *Proc. Asian Conf. Comput. Vis.*, 2016, pp. 213–228.

[8] S. Lee, S. Park, and K. Hong, "Rdfnet: Rgb-d multi-level residual feature fusion for indoor semantic segmentation," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 4990–4999.

[9] J. Jiao, Y. Wei, Z. Jie, H. Shi, R. Lau, and T. S. Huang, "Geometry-aware distillation for indoor semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 2864–2873.

[10] Y. Xing, J. Wang, X. Chen, and G. Zeng, "Coupling two-stream rgb-d semantic segmentation network by idempotent mappings," in *Proc. IEEE Int. Conf. Image Process.*, 2019, pp. 1850–1854.

[11] X. Hu, K. Yang, L. Fei, and K. Wang, "Acnet: Attention based network to exploit complementary features for rgbd semantic segmentation," in *Proc. IEEE Int. Conf. Image Process.*, 2019, pp. 1440–1444.

[12] X. Chen, K.-Y. Lin, J. Wang, W. Wu, C. Qian, H. Li, and G. Zeng, "Bi-directional cross-modality feature propagation with separation-and-aggregation gate for rgb-d semantic segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2020.

[13] H. Zhou, L. Qi, Z. Wan, H. Huang, and X. Yang, "Rgb-d co-attention network for semantic segmentation," in *Proc. Asian Conf. Comput. Vis.*, 2020.

[14] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from rgbd images," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 746–760.

[15] S. Song, S. P. Lichtenberg, and J. Xiao, "Sun rgb-d: A rgb-d scene understanding benchmark suite," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 567–576.

[16] S. Gupta, R. Girshick, P. Arbeláez, and J. Malik, "Learning rich features from rgb-d images for object detection and segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 345–360.

[17] W. Wang and U. Neumann, "Depth-aware cnn for rgb-d segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 144–161.

[18] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[19] J. Deng, W. Dong, R. Socher, L. Li, Kai Li, and Li Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 248–255.

[20] D. Xu, W. Ouyang, X. Wang, and N. Sebe, "Pad-net: Multi-tasks guided prediction-and-distillation network for simultaneous depth estimation and scene parsing," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 675–684.

[21] Z. Zhang, Z. Cui, C. Xu, Y. Yan, N. Sebe, and J. Yang, "Pattern-affinitive propagation across depth, surface normal and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 4101–4110.

[22] S. Kong and C. Fowlkes, "Recurrent scene parsing with perspective understanding in the loop," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 956–965.

[23] X. Qi, R. Liao, J. Jia, S. Fidler, and R. Urtasun, "3d graph neural networks for rgbd semantic segmentation," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 5209–5218.

[24] J. Jiang, L. Zheng, F. Luo, and Z. Zhang, "Rednet: Residual encoder-decoder network for indoor rgb-d semantic segmentation," 2018, arXiv:1806.01054.

[25] D. Lin, G. Chen, D. Cohen-Or, P. Heng, and H. Huang, "Cascaded feature network for semantic segmentation of rgb-d images," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 1320–1328.