SEMANTICS-GUIDED MULTI-LEVEL RGB-D FEATURE FUSION FOR INDOOR SEMANTIC SEGMENTATION

Yabei Li[1,2], *Junge Zhang*[1,2], *Yanhua Cheng*[1,2], *Kaiqi Huang*[1,2,3], *Tieniu Tan*[1,2,3]

[1]CRIPAC & NLPR, CASIA [2]University of Chinese Academy of Sciences
[3]CAS Center for Excellence in Brain Science and Intelligence Technology

ABSTRACT

Indoor RGB-D semantic segmentation is a new and challenging problem. Traditional methods usually apply two-stream convolutional neural networks (CNNs) to represent RGB and depth images respectively, and fuse the two streams on a specific layer. In this paper, we explore several fusion strategies based on this two-stream-CNN framework and point out such a single-layer fusion method cannot exploit the complementary RGB and depth cues well for semantic segmentation. To address this problem, we propose a novel Semantics-guided Multi-level feature fusion approach, which first learns deep feature representation from bottom to up, and then gradually fuses the RGB and depth features from high level to low level under the guidance of the semantic cues. Experimental results on SUN RGB-D dataset demonstrate the advantages of the proposed method over the state of the arts.

***Index Terms*—** Indoor semantic segmentation, RGB-D, Multimodal fusion

1. INTRODUCTION

With the release of advanced depth sensors, the use of the RGB-D data in computer vision has attracted much attention recently [1][2][3][4][5]. By incorporating the depth information, we can obtain more geometric information which is more invariant to illumination, appearance and occlusion. In this paper, we mainly focus on how to fuse the RGB and depth features for indoor scene semantic segmentation.

For indoor semantic segmentation, remarkable efforts have been put into the RGB-D fusion. Before the wide application of CNN, people usually extract and concatenate the hand crafted RGB and depth features [6][7][8][9]. Recently, the state-of-the-art methods usually employ two-stream Convolutional Neural Networks (CNNs), especially two-stream Fully Convolutional Neural Networks (FCNs) to represent the RGB and depth in each stream and fuse the two streams on a specific layer [10][11][12][13][14].

We summarize above two-stream CNNs based fusion methods into three categories: Early Fusion, Middle Fusion and Late Fusion. Early Fusion (**Fig.1(a)**) directly concatenates the raw RGB and depth images[1]. Middle Fusion (**Fig.1(b)**), concatenates the two modalities in a specific convolutional layer or fully connected layer. Late Fusion (**Fig.1(c)**) sums the output score maps from two streams.

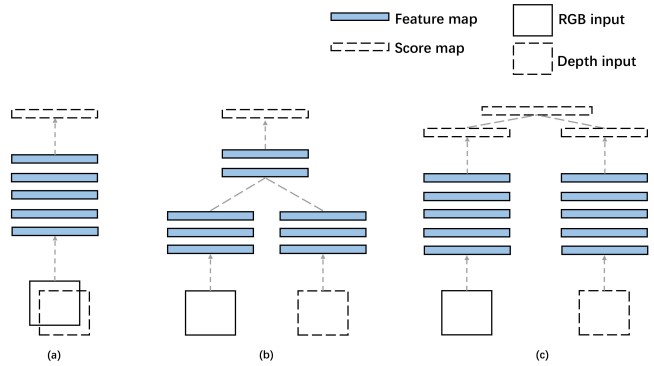

Fig. 1: Different strategies for fusing RGB and depth. (a) Early Fusion; (b) Middle Fusion; (c) Late Fusion

Extensive experimental evaluations on the three fusion approaches under the two-stream FCNs framework are conducted for indoor semantic segmentation in this paper. We find: 1) Early Fusion and Middle Fusion on shallow layers (low-level fusion) suffers from the modality gap. Although the spatial cues are well retained in low-level RGB-D features, the visual information in RGB image and the geometric information in depth image are not calibrated in low levels. For example, in **Fig.2(a)**, the rgb feature highlights the visual boundary and the depth feature highlights the geometric boundary of the whiteboard. After fusing the RGB and depth features, the resulted features are less informative. 2) Late Fusion and Middle fusion on deep layer (high-level fusion) fuses the high-level features that represent semantic information, which are more compatible in different modalities. However, the complementary spatial cues in RGB and depth high-level features have already been weakened after performing max pooling. For example, in **Fig.2(b)** the boundary of the chair legs is hard to see in RGB image due to its similar color to the ground. The legs' boundary can be activated strongly in depth features, yet the segmentation result of high-level fusion cannot

[1]we use HHA [11] (Horizontal Disparity, Height above the ground, Angle of surface norm) to encode the depth images in this paper.

delineate the boundary of the chair legs well.

To bridge the modality gap while utilize the retained complementary spatial cues in low-level features, we propose a novel Semantics-guided Multi-level fusion (S-M Fusion) approach in this paper. The proposed method first learns feature representations in different levels, then learns to gradually fuse the features from high level to low level under the guidance of the semantic cues. When fusing low-level features which have large modality gap, the semantic cues enforce the model to learn to transform, extract and fuse the complementary RGB and depth information. We show the effectiveness of our fusion strategy. Our results on SUN RGB-D dataset achieve the best accuracy compared to the state of the arts.

2. METHODOLOGY

Our Semantics-guided Multi-level Fusion (S-M Fusion) framework for RGB-D indoor semantic segmentation is illustrated in **Fig.3**. The proposed framework first uses two-stream FCNs to learn the RGB and depth features in different levels from bottom to up. Then it gets the fused high-level feature by summing the coarse score maps (i.e. the last layers from the two-stream FCNs) from the two-stream FCNs. To fuse the low-level features, we propose the Semantics-guided Fusion Block (SFB). The final segmentation result can be obtained by cascading the SFBs from top to down.

2.1. Bottom-to-up feature representation

We first train the two-stream FCNs to learn the RGB and depth feature representation in different levels. To get larger receptive field, we use the DeepLab-LargeFoV [15] as our base architecture in each stream. We minimize the softmax loss between the summed score maps from two streams and ground truth when we train the network. The features extracted from the deep layers of the two-stream FCNs (high-level features) contain more semantic information, while the features extracted from the shallow layers (low-level fusion) activate strongly on spatial information. We aim to utilize both the high-level features and low-level features in RGB-D feature fusion in the next two sections.

2.2. Coarse score maps

After learning the RGB and depth features in different levels in section 2.1, we can directly get the high-level fused features by summing the score maps from the two-stream FCNs. The obtained score maps are very coarse as we discussed in section 1. However, with the high-level coarse score maps, the low-level features are able to focus to provide the residual complementary low-level RGB-D information to get the finer segmentation results. In practice, since the number of categories c is always small, the last convolutional layers with c channels are reformulated to convolutional layers with k channels ($k > c$) to capture more information.

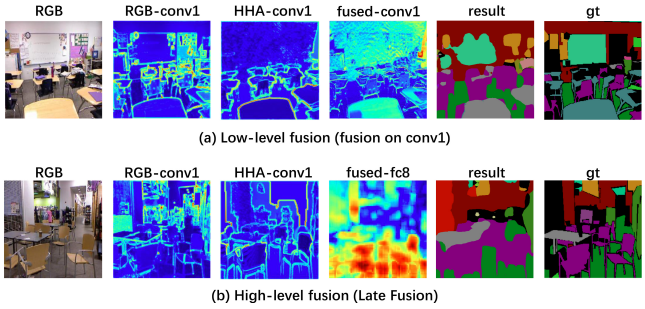

Fig. 2: The weakness of the low-level fusion and high-level fusion. (a) Low-level fusion on the RGB features highlight visual boundary and the depth features highlight geometric boundary of the whiteboard makes the resulted features less informative. (b) Although there are strong complementary boundary cues of chair legs in RGB and depth images, segmentation result from high-level fusion cannot delineate them well.

2.3. Cascaded Semantics-guided Fusion Blocks

The SFB is proposed to fuse the low-level RGB and depth features. As illustrated in **Fig.3**, the i^{th} SFB has three inputs: the score map from the output of the previous SFB (or the coarse score maps in section 2.2) h_{i-1} and two feature maps r_i, d_i extracted from RGB and depth FCNs respectively. The two feature maps r_i and d_i are connected by the convolution layer g_{i1} and g_{i2} respectively before concatenating together with the score map h_{i-1}. The concatenated feature maps then are connected to another convolutional layer w_i with the number of channels equals to c. The output of the i^{th} SFB is h_i.The processes can be represented as:

$$h_i = w_{i1} \circ (g_{i1} \circ r_i) + w_{i2} \circ (g_{i2} \circ d_i) + w_{i3} \circ h_{i-1}, \quad (1)$$

where $\circ$ indicates convolution and + indicates element-wise summation. Each output h_i is associated with a classifier, in which we minimize the softmax loss function $l_i(h_i, t)$. t is the ground truth score map.

In SFB, the g_{i1} and g_{i2} first nonlinearly transform the input low-level features r_i and d_i. The h_{i-1} from the last SFB provides the coarse score maps from the higher-level feature fusion. In order to improve the coarse score maps h_{i-1} to finer score maps h_i with lower-level feature fusion, the transformed r_i and d_i, together with the coarse score maps h_{i-1} are embedded by weighting w_{i1}, w_{i2} and w_{i3}. The r_i and d_i can be regarded to provide residual low-level information between h_{i-1} and h_i. The deep supervision on the h_i is the key to our architecture. The block is called "semantics-guided" for two points: the h_{i-1} gives the higher-level semantics information (coarse score maps) from the deeper layer; the deep supervision on h_i provides semantics guidance to learn to transform, select and fuse the residual complementary low-level RGB-D information by learning g_i and w_i.

In practice, we normalize the feature maps as in [16] before concatenating them to train the network more effectively.

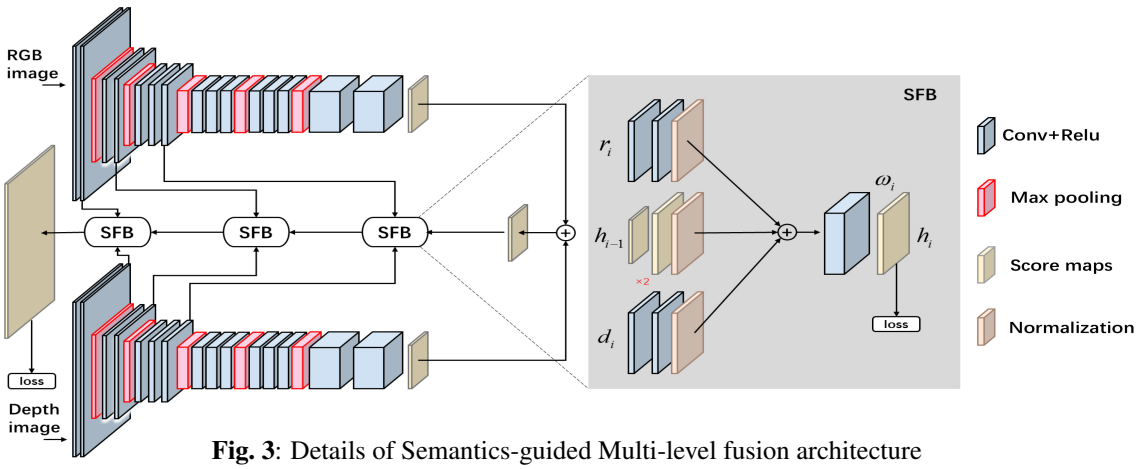

Fig. 3: Details of Semantics-guided Multi-level fusion architecture

Since the number of categories c in application is usually small, in order to transfer more semantic information to the next SFB, for all SFBs except the last one, we reformulate the structure by branching the concatenated features into score maps h_{i1} and h_{i2}. The h_{i1} has the same number of channels as the number of categories c and associates to the classifier. The number of channels of h_{i2} is set to k ($k > c$). h_{i2} then is sent to the next SFB as one of the input.

To get the full resolution, we cascade the SFBs from top to down. To bridge the resolution gaps between different layers, the input score map h_{i-1} can be up-sampled in the SFB. We minimize the objective function of the cascaded SFBs as:

$$L = \sum_i \partial_m l_i(h_i, t), \tag{2}$$

where we set $\partial_m = 1$ in our practice.

3. EXPERIMENTS

3.1. Experimental setup

Dataset: We evaluate our method on the challenging SUN RGB-D [17] dataset for scene segmentation. SUN RGB-D is the largest indoor scene dataset which contains 10355 RGB and depth image pairs captured from different cameras. There are 37 semantic classes and about 0.25% unannotated pixels that do not belong to any of the 37 classes. We divide the 10355 image pairs into 5285 image pairs for training and 5050 image pairs for test, the same setting as [17].

Evaluation Metric: Three widely used metrics in literatures [18] are employed to evaluate the results: the global accuracy, the mean accuracy and the mean IOU. Among the three metrics, the global accuracy is less informative for it doesn't take the class imbalance into account. In [19], it shows that mean IoU prefers region smoothness and does not evaluate boundary accuracy. Since our method aims to improve the results with low-level information in RGB and depth, we mainly focus on the mean accuracy metrics.

Model Training: We train the model via stochastic gradient descent (SGD) based on caffe [20]. Our model is trained in two steps: we first train the bottom-up feature learning stream, and then we train the top-down semantics-guided feature fusion stream. The weights in the convolution network are initialized by using VGG 16-layer net pre-trained on ILSVRC [21] dataset. The convolution weights in SFB are initialized with zero-mean Gaussians. The input RGB and depth images are randomly cropped to 449×449. For optimization, we set initial learning rate to 0.001 for both steps. The learning rate is updated using the "poly" policy [15], i.e. $lr_{iter} = lr_0(1 - \frac{iter}{\max\_iter})^{power}$. Power, weight decay, max iterations and the number of channels k are set to 0.9, 0.0005, 20000 and 128. Our model works in an end to end manner in test.

3.2. Evaluation and discussion

Different Fusion Strategies: The segmentation results of different fusion strategies are shown in **Table 1**. We can see when we fuse features in deeper layer, the classification accuracy increases. However, we argue that high-level fusion still cannot fully exploit the RGB and depth information. Our proposed multi-level fusion model (S-M Fusion) improves the Late Fusion model with a significant margin (5.3%).

Performance Comparison with State-of-the-art Methods: **Table 2** shows the performance of our proposed Semantics-guided Multi-level fusion model (S-M Fusion) compared with other state-of-the-art methods. Our final model fuses the features from the Conv1_2, Conv2_2, Conv3_3 layers. We compare our segmentation results of 37 classes with other state of the arts. It can be seen that our model has achieved best performance among them. Some qualitative segmentation results are shown in **Fig.4**. Compared to Late Fusion model, our model get finer segmentation results. It shows our model can extract and fuse the complementary low-level information in RGB and depth effectively.

3.3. Discussion

Which layer's low-level features are helpful to semantic segmentation? We use SFB to fuse RGB and Depth low-level features in different single layers, from conv1_2 to conv5_3. The results are shown in **Table 3** (see from the 1^{st} to the 6^{th} row). We can see that low-level features from all of these layers are helpful, which also demonstrates that high-

	Early Fusion	Middle Fusion						Late Fusion	Skip Connections	S-M Fusion
		Conv1_2	Conv2_2	Conv3_3	Conv4_3	Conv5_3	fc6			
Global Acc.	76.29%	76.69%	77.05%	75.55%	76.26%	77.07%	77.25%	77.68%	78.09%	**78.07%**
Mean Acc.	44.83%	45.51%	46.45%	47.80%	48.86%	49.86%	48.48%	49.00%	50.75%	**53.93%**
Mean IOU	35.65%	36.36%	37.28%	37.97%	38.48%	39.83%	39.85%	40.04%	40.66%	**40.98%**

Table 1: Evaluations on different fusion strategies. For middle fusion, we evaluate the results of fusing the two-stream FCNs at different layer.

Methods	Global Acc.	Mean Acc.	Mean IOU
SIFT flow [17]	-	10.10%	-
Bayesian SegNet (RGB) [18]	71.20%	45.90%	30.70%
LSTM [22]	-	48.10%	-
FuseNet-SF5 [23]	76.27%	48.30%	37.29%
FuseNet-DF1 [23]	73.37%	50.07%	34.02%
S-M Fusion(ours)	**78.07%**	**53.93%**	**40.98%**

Table 2: Performance comparison with other state of the arts on SUN RGB-D dataset

level feature fusion doesn't fully exploit the RGB and depth information.

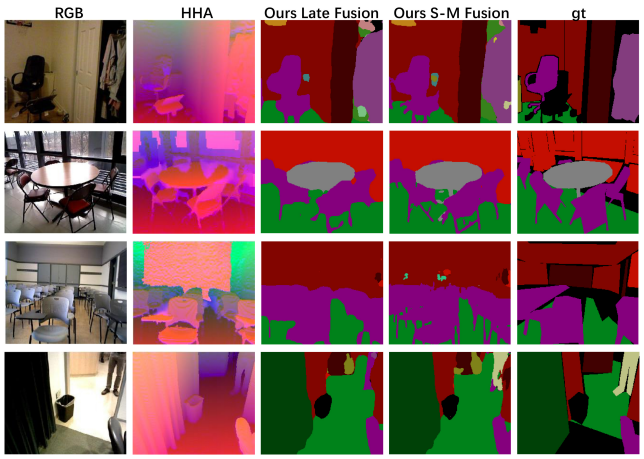

Fig. 4: Qualitative segmentation results on SUN RGB-D dataset. The segmentation results from our model are much finer than the Late Fusion Model.

How many low-level layers to fuse? As we show in **Table 3**, the more layers fused, the higher accuracy is achieved. The improvement decreases when we keep increasing the number of low-level layers to fuse. This is because the information in different layers are not orthogonal.

Is Semantics-guided Fusion necessary? We use skip connections strategy as a comparison, which directly connects low-level features and high-level features from different modalities. For fair comparison, we also use low-level features from the Conv1_2, Conv2_2, Conv3_3 layers. The results are shown in **Table 1** (skip connections). We can see that the skip connections get better results than Late Fusion since it benefits from the multi-level feature fusion, but it's not as effective as our model. This shows that Semantics-guided Fusion does help to bridge the modality gap in low-level fusion.

Num. of SFB	Fusion Layers	Global Acc.	Mean Acc.	Mean IOU
0	None	77.68%	49.00%	40.04%
1	Conv1_2	77.75%	50.52%	40.27%
	Conv2_2	77.88%	51.18%	41.02%
	Conv3_3	78.22%	51.44%	41.04%
	Conv4_3	78.37%	51.25%	41.22%
	Conv5_3	77.99%	50.91%	40.76%
2	Conv2_2, 3_3	78.30%	52.40%	41.08%
3	Conv1_2, 2_2, 3_3	78.07%	53.93%	40.98%

Table 3: Ablation study of our architecture. The Fusion Layers indicates the low-level layers we use to fuse. The architecture using 3 SFBs is our final S-M fusion model.

What if we aggregate each modality's low-level information separately? We also compare the architecture which fuses the low-level features in semantics-guided manner in each stream and then sums the score maps from the two streams. The resulted Mean Acc. and Mean IOU is 50.21% and 40.08% respectively. It demonstrates that although low-level information itself can increase the performance, our model benefits most from the effective fusion of these complementary low-level information in RGB and depth.

4. CONCLUSION

In this paper, different RGB-D fusion approaches are explored for indoor scene semantic segmentation. We point out that single-level fusion cannot fully exploit the complementary information in RGB and depth. We then propose a novel Semantics-guided Multi-level RGB-D fusion strategy, which uses the Semantics-guided Fusion Block (SFB) to guide the lower-level features to fuse across modalities. Experimental results on SUN RGB-D dataset demonstrate the effectiveness of our model.

5. ACKNOWLEDGEMENTS

This work is funded by the National Natural Science Foundation of China (Grant No. 61403387 and Grant No. 61673375) and the National Key Research and Development Program of China (Grant No. 2016YFB1001004 and Grant No. 2016YFB1001005).

6. REFERENCES

[1] Luciano Spinello and Kai O Arras, "People detection in rgb-d data," in *IROS*, 2011, pp. 3838–3843.

[2] Li Liu and Ling Shao, "Learning discriminative representations from rgb-d video data.," in *IJCAI*, 2013, vol. 4, p. 8.

[3] Srinath Sridhar, Antti Oulasvirta, and Christian Theobalt, "Interactive markerless articulated hand motion tracking using rgb and depth data," in *ICCV*, 2013, pp. 2456–2463.

[4] Yang Zhao, Zicheng Liu, Lu Yang, and Hong Cheng, "Combing rgb and depth map features for human activity recognition," in *APSIPA ASC*, 2012, pp. 1–4.

[5] Yanhua Cheng, Rui Cai, Chi Zhang, Zhiwei Li, Xin Zhao, Kaiqi Huang, and Yong Rui, "Query adaptive similarity measure for rgb-d object recognition," in *ICCV*, 2015, pp. 145–153.

[6] Saurabh Gupta, Pablo Arbelaez, and Jitendra Malik, "Perceptual organization and recognition of indoor scenes from rgb-d images," in *CVPR*, 2013, pp. 564–571.

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

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

[9] Nathan Silberman and Rob Fergus, "Indoor scene segmentation using a structured light sensor," in *ICCV Workshops*, 2011, pp. 601–608.

[10] Camille Couprie, Clément Farabet, Laurent Najman, and Yann LeCun, "Indoor semantic segmentation using depth information," *arXiv preprint arXiv:1301.3572*, 2013.

[11] 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.

[12] Andreas Eitel, Jost Tobias Springenberg, Luciano Spinello, Martin Riedmiller, and Wolfram Burgard, "Multimodal deep learning for robust rgb-d object recognition," in *IROS*, 2015, pp. 681–687.

[13] Jonathan Long, Evan Shelhamer, and Trevor Darrell, "Fully convolutional networks for semantic segmentation," in *CVPR*, 2015, pp. 3431–3440.

[14] Yanhua Cheng, Xin Zhao, Rui Cai, Zhiwei Li, Kaiqi Huang, and Yong Rui, "Semi-supervised multimodal deep learning for rgb-d object recognition," in *IJCAI*, 2016, pp. 3345–3351.

[15] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille, "Semantic image segmentation with deep convolutional nets and fully connected crfs," *arXiv preprint arXiv:1412.7062*, 2014.

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

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

[18] Alex Kendall, Vijay Badrinarayanan, and Roberto Cipolla, "Bayesian segnet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding," *arXiv preprint arXiv:1511.02680*, 2015.

[19] Gabriela Csurka, Diane Larlus, Florent Perronnin, and F Meylan, "What is a good evaluation measure for semantic segmentation?," *IEEE PAMI*, vol. 26, pp. 1, 2004.

[20] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell, "Caffe: Convolutional architecture for fast feature embedding," in *Proceedings of the 22nd ACM international conference on Multimedia*, 2014, pp. 675–678.

[21] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *CVPR*, 2009, pp. 248–255.

[22] 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.

[23] Caner Hazirbas, Lingni Ma, Csaba Domokos, and Daniel Cremers, "Fusenet: Incorporating depth into semantic segmentation via fusion-based cnn architecture," in *ACCV*, 2016, vol. 2.