INDOOR SEMANTIC SEGMENTATION FROM RGB-D IMAGES BY INTEGRATING FULLY CONVOLUTIONAL NETWORK WITH HIGHER-ORDER MARKOV RANDOM FIELD

J. Yang [1], Z. Kang [1, *]

[1] Department of Remote Sensing and Geo-Information Engineering, School of Land Science and Technology, China University of Geosciences, Xueyuan Road, Beijing, 100083 CN – jtyang66@126.com, zzkang@cugb.edu.cn

Commission VI, WG VI/5

KEY WORDS: Fully convolutional network, RGB-D images, Higher order potentials, Indoor scenes, Semantic segmentation

ABSTRACT:

Indoor scenes have the characteristics of abundant semantic categories, illumination changes, occlusions and overlaps among objects, which poses great challenges for indoor semantic segmentation. Therefore, we in this paper develop a method based on higher-order Markov random field model for indoor semantic segmentation from RGB-D images. Instead of directly using RGB-D images, we first train and perform RefineNet model only using RGB information for generating the high-level semantic information. Then, the spatial location relationship from depth channel and the spectral information from color channels are integrated as a prior for a marker-controlled watershed algorithm to obtain the robust and accurate visual homogenous regions. Finally, higher-order Markov random field model encodes the short-range context among the adjacent pixels and the long-range context within each visual homogenous region for refining the semantic segmentations. To evaluate the effectiveness and robustness of the proposed method, experiments were conducted on the public SUN RGB-D dataset. Experimental results indicate that compared with using RGB information alone, the proposed method remarkably improves the semantic segmentation results, especially at object boundaries.

1. INTRODUCTION

Semantic segmentation is a fundamental problem in computer vision, which decomposes a scene into meaningful parts and assigns semantic labels to them (Wolf et al., 2015). Compared with outdoor counterpart, indoor scene annotation is a relatively difficult issue since it usually contains illumination variations, occlusions and overlaps among objects, significant appearance variations and imbalanced representations of object categories (Chu et al., 2017). Therefore, semantic segmentation for indoor scene has seen an increased interest.

In recent years, many methods about indoor semantic segmentation have been presented. Most pervious researches primarily rely on hand-crafted features from both color channels and depth channel, as input of the frequently-used classifier for automatic classification. Silberman and Fergus (2011) developed a CRF-based model, combining 3D location prior from depth channel with features captured from both depth channel and color channels, for indoor scene segmentation. Ren et al. (2012) adopted the kernel-based framework for transforming the pixel-level similarity within each super-pixel into the patch descriptor, which were then integrated with contextual information for labeling RGB-D images. Gupta et al. (2013) made effectively use of depth information for optimizing image segmentation and defined the features of super-pixels for automatic classification using random forest classifier and support vector machine. Müller and Behnke (2014) conducted conditional random filed, into which color, depth and 3D scene features were incorporated, for semantic annotation of RGB-D images. Unfortunately, these conventional methods usually consist of segmentation, feature extraction and classification and their final results depend on the results of each stage (Husain et al., 2016).

With the success of convolutional neural network (CNN) in many applications, a large variety of CNN architectures, especially fully CNN, have been developed to extract the high-level semantic features for semantic segmentation in recent years and worked in an end-to-end manner. He et al. (2017) developed a spatio-temporal pooling layer for combining contextual information derived from multi-view images for semantic image segmentation. Chu et al. (2017) integrated learnable constraint layers that encode contextual regularization between the neighboring pixels with a deep convolutional segmentation network for enhancing the semantic segmentation results of indoor scene images. More recently, inexpensive RGB-D sensors are proving to be a rich source of information for indoor scenes and can provide color and depth images in real-time (Khan et al., 2014). To effectively use the depth channel, Höft et al. (2014) presented the histogram of oriented depth descriptor as input of convolutional neural network. Lin et al. (2017a) proposed context-aware receptive field and performed a multiple branches-based network model for segmenting RGB-D images. To sufficiently exploit contextual information, Li et al. (2017) carried out a two-stream FCNs to learn the RGB and depth features respectively and gradually fused these features from high level to low level for indoor scene semantic segmentation. Jiang et al. (2018) developed an encoder-decoder architecture to extract RGB information and depth information separately and fuse the information over several layers for indoor semantic segmentation. By incorporating the depth information, the spatial geometric information, which is more invariant to illumination changes

* Corresponding author.

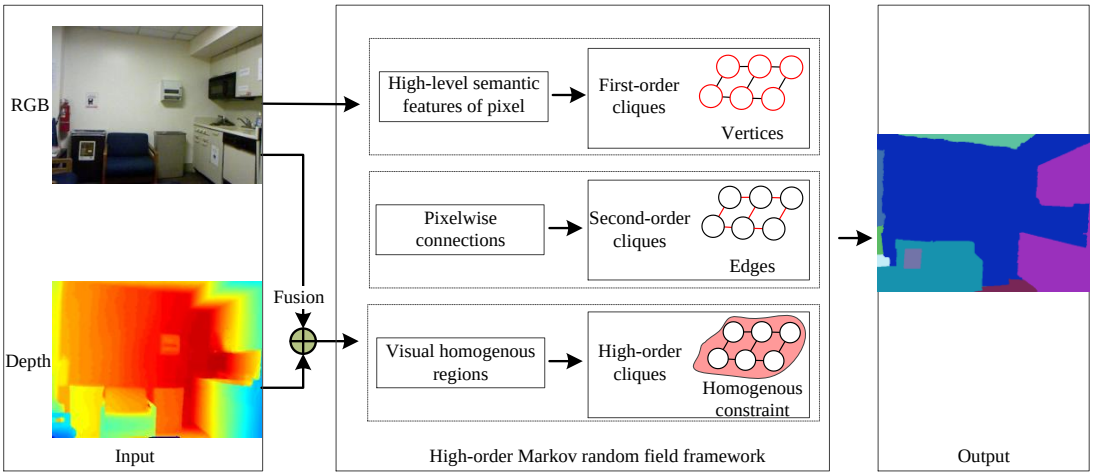


Fig. 1 Workflow of the proposed method

and appearances, can be derived for the improvement of semantic segmentation.

To address the issues raised from the state-of-the-art of the semantic segmentation for indoor scenes, we develop a method based on higher-order Markov random field model for indoor semantic segmentation from RGB-D images. Due to illumination changes, occlusions and overlaps among objects in indoor scenes, the spatial location relationship from depth channel and the spectral information from color channels are integrated as prior information for a marker-controlled watershed algorithm to derive the robust and accurate visual homogenous regions, which will encode the low-level visual features for complementarily reconstructing the detailed boundaries. Moreover, to alleviate the fact that the pooling operations result in the blurry object boundaries, higher-order Markov random field model is adopted to encode the short-range context among the adjacent pixels and the long-range context within each visual homogenous region for refining the semantic segmentations, especially at object boundaries.

The rest of this paper is organized as follows. Section 2 describes the proposed method in detail. Section 3 presents the experimental results and analysis for evaluating the proposed method. This paper concludes with a discussion of future research considerations in Section 4.

2. METHODOLOGY

In this paper, we develop a method based on higher-order Markov random field (MRF) model, which combines the high-level semantic information derived from RefineNet and the low-level visual information captured from a marker-controlled watershed algorithm, for indoor semantic segmentation from RGB-D images. As shown in Fig. 1, the proposed method consists of the following steps: (1) initial semantic segmentation using RefineNet, (2) Visual homogenous regions generated by combining color information and depth information, (3) Region-level label consistency based on higher-order MRF model. As a result, the indoor scenes are interpreted into 38 classes. Key algorithms of the proposed method are given in more detail below.

2.1 Initial semantic segmentation using RefineNet

To date, numerous FCNN architectures have been developed, such as U-Net (Ronneberger et al., 2015), SegNet (Badrinarayanan et al., 2017), PSPNet (Zhao et al., 2017) and DeepLab (Chen et al., 2017), for semantic segmentation. To efficiently exploit all the information available along the down-sampling process for reconstructing the high-resolution prediction, these architectures presented a large variety of strategies, such as atrous convolutions (Chen et al., 2017) and skip connections (Ronneberger et al., 2015; Badrinarayanan et al., 2017). Since RefineNet (Lin et al., 2017b) effectively integrated low-resolution semantic features with fine-grained low-level visual features for generating high-resolution semantic feature maps and adopt residual connections with identify mappings for addressing the problems of vanishing the gradients during the training stage (He et al., 2016), which achieved new state-of-the-art performance on seven public datasets. Thus, we in this paper use the trained RefineNet model to predict the initial semantic segmentation on RGB images alone. An illustration of RefineNet architecture is presented in

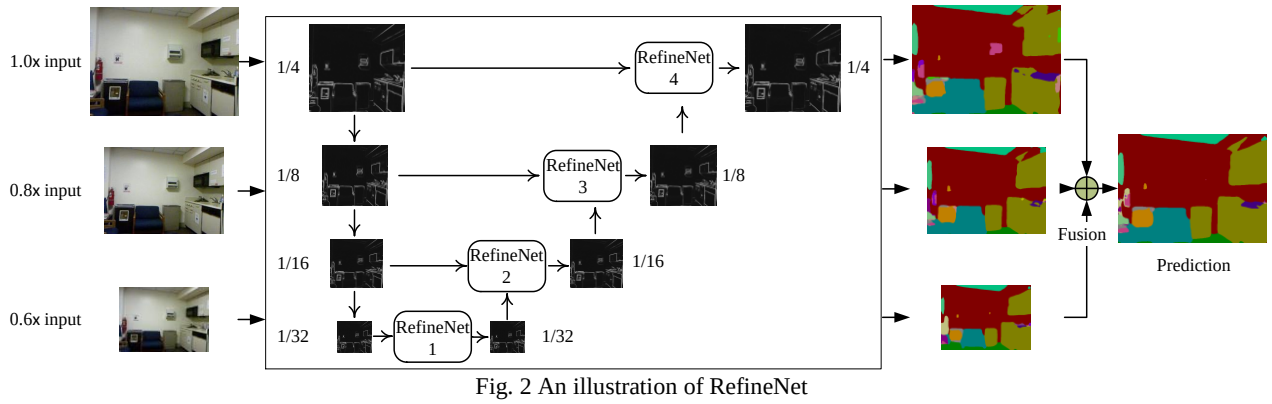


Fig. 2 An illustration of RefineNet

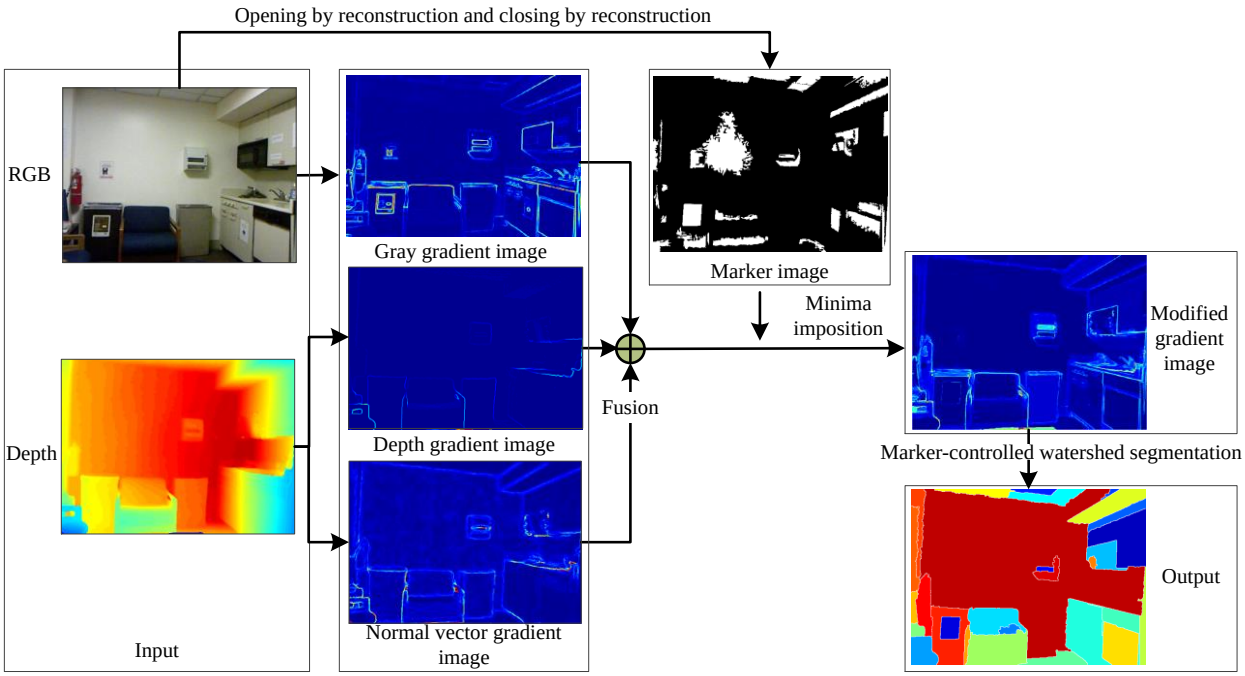


Fig. 3 An example of the marker-controlled watershed algorithm. Different visual homogeneous regions in output results are randomly rendered in different colors.

Fig. 2. For more details about RefineNet architecture, please refer to (Lin et al., 2017b).

2.2 Visual homogenous regions generated by combining color information and depth information

As aforementioned, incorporating depth information for enhancing the performance of semantic segmentation achieved great successes (Höft et al., 2014; Husain et al., 2016; Lin et al., 2017a; Li et al., 2017; Jiang et al., 2018) since depth data can provide the 3D spatial location relationships among objects (Lin et al., 2017a) and be insensitive to illumination changes from which the color channels suffer. As a matter of fact, the data quality of depth sensors, which is a measure of point precision, is limited (Khoshelham, 2012). For example, random error of depth measurement increases drastically with increasing distance from sensors, which inevitably causes the semantic segmentation errors if depth data directly serves as the input of CNN architectures. In our implementation, the depth data is just used for assisting the generation of visual homogenous regions. Furthermore, abundant semantic categories, occlusions and overlaps among objects are common in indoor scenes, which easily results in over-segmentation during the procedure of producing the visual homogenous regions. Since a marker-controlled watershed segmentation algorithm is simple and intuitive and can be parallelized (Xu et al., 2011), effectively avoiding over-segmentation with the marker constraints. Therefore, we use a marker-controlled watershed segmentation algorithm by combining color information with depth information to efficiently and robustly derive a set of visual homogenous regions from RGBD images.

Marker-controlled watershed segmentation is a variant of the conventional watershed segmentation (Vincent and Soille, 1991) for solving the over-segmentation issues from numerous potential but trivial regional minima. Watershed segmentation considers a gray-level image as topographic surface, where the gray value of each pixel is interpreted as its altitude. Suppose a water source is placed in each regional minimum and the entire topography structure is flooded from below. When water from two sources (i.e., regional minima) are about to meet, a dam is constructed to prevent the merging. The flooding and dam construction process continues until only the dams are visible from above. These dames effectively segment the image into regions. Due to noises and quantization error (Parvati et al., 2008), the over-segmentation is an intrinsic problem of watersheds. In our implementation, we constrain the watershed segmentation with marker image that is generated through multiple morphological operations. As a result, each marker is associated with a region in the segmented image. Fig. 3 shows a simple example of the marker-controlled watershed algorithm based on the morphological operations, which consists of the following steps.

(1) Generation of gray gradient image, depth gradient image and normal vector gradient image. The original RGB image and depth image are transformed into the associated gradient images based on Sobel filter (Sobel et al., 1968), respectively. The original depth image is used for producing the 3D point cloud based on the corresponding camera intrinsics and the normal vector of each pixel is estimated for deriving the normal vector gradient image. In such cases, high gradient magnitudes are at object boundaries while low gradient magnitude occurs inside objects. At the subsequent procedures, we perform the watershed segmentation on the derived gradient images instead of the original image; These gradient images associated with both RGB image and depth image are fused for providing redundant and complementary object boundaries from different perspectives.

(2) Because compared with the traditional opening and closing operator, opening by reconstruction and closing by reconstruction are less destructive and can maintain the object shape better (Lewis and Dong, 2012). Thus, the marker image is derived based on the morphological operations, including opening by reconstruction and closing by reconstruction, from the original RGB images.

(3) The combined gradient image is modified based on minima imposition technique (Vincent, 1993), which makes regional

minima occur at marker pixels, using the marker image derived in Step (2);
(4) Marker-controlled watershed segmentation is performed on the modified gradient image.

2.3 Region-level label consistency based on higher-order MRF model

It is noted that the down-sampling operation in CNN architectures, such as pooling layer, causes the burry boundaries in the semantic segmentation results. Recently, higher order potentials were incorporated into MRF model for modeling higher-level contextual information and achieved successes in many applications (Woodford et al, 2009; Ren et al., 2015; Yang et al., 2018). For these models, visual homogenous regions can help to model long-range contextual information, which is particularly useful for obtaining object segmentations with fine boundaries (Kohli and Torr, 2009). Hence, we in this section use higher-order MRF model (Kohli and Torr, 2009) for optimizing the semantic segmentation through encoding the short-range contextual information among the adjacent pixels and the long-range contextual information within each visual homogenous region.

MRF model (Geman and Geman, 1987) is a weighted undirected graph $G = \langle V, E \rangle$, where V denotes a set of vertices, and E represents a set of undirected edges between the neighboring vertices. For the image semantic segmentation, an observed image with V pixels is denoted by a discrete random filed, where each random variable is associated with a pixel. The goal is to infer the labeling of the image $Y = \{y_1, y_2, ..., y_V\}$, where each variable y_i is the label of pixel i and takes a value from the set $C = \{1,2,..., L\}$, L is the number of classes. In the field of computer vision, finding the optimal label configuration Y^* can be naturally formulated into the energy function minimization as the following Eq. (5).

$$\mathrm{En}(Y) = \mathrm{En}_{\mathrm{unary}}(Y) + \lambda_1 \cdot \mathrm{En}_{\mathrm{pairwise}}(Y) + \lambda_2 \cdot \mathrm{En}_{\mathrm{region}}(Y) \quad (5)$$

where first order (or unary) energy term $\mathrm{En}_{\mathrm{unary}}(Y)$ measures the disagreement between Y and the observed data, second order (or pairwise) energy term $\mathrm{En}_{\mathrm{pairwise}}(Y)$ measures the extent to which Y is not piecewise smooth, higher order energy term $\mathrm{En}_{\mathrm{region}}(Y)$ measures the label consistency over visual homogenous regions, λ_1 and λ_2 are the weighted parameters.

The form of unary term $\mathrm{En}_{\mathrm{unary}}(L)$ is typically

$$\mathrm{En}_{\mathrm{unary}}(Y) = \sum_{i \in V} \mathrm{D}_i(y_i), \quad (6)$$

where $\mathrm{D}_i(y_i)$ quantitatively measures the degree of "fit" between the label y_i and the observed data. In this paper, the output of the softmax layer in the learned RefineNet architecture quantitatively measures the disagreement between the label y_i and the observed data. As defined in Eq. (6), the class posterior probability, the smaller the unary term.

To generate locally continuous and globally optimal label configuration, the pairwise energy term $\mathrm{En}_{\mathrm{pairwise}}(Y)$ is generally defined as the following Eq. (8).

$$En_{smooth}(L) = \sum_{\{i,j\} \in E} S_{i,j}(y_i, y_j), \quad (8)$$

where $S_{i,j}(y_i, y_j) = g(i,j) \cdot \delta(y_i, y_j)$, $\delta(y_i, y_j) = \begin{cases} 1, & \text{otherwise} \\ 0, & \text{if } y_i = y_j \end{cases}$, $g(i,j) = \exp(-\|x_i - x_j\|)$, x_i and x_j denote the semantic feature vectors of the pixel i and j respectively derived from the learned RefineNet architectures. As defined in Eq. (8), the smoothness penalty term is zero for the neighboring pixels with the same label. With regards to the adjacent pixels with different labels, the smaller the distance between them is, the larger the smoothness penalty term is. Consequently, the pairwise energy term $\mathrm{En}_{\mathrm{pairwise}}(Y)$ encodes the extent to which the adjacent pixels belong to the same label.

To reconstruct the semantic segmentation objects with refine boundaries, the higher order energy term $\mathrm{En}_{\mathrm{region}}(Y)$ is incorporated into the energy function Eq. (5) for capturing the long-range contextual information within each visual homogenous region derived from Section 2.2. Although the combination of color and geometric information can improve the performance of generating the visual homogeneous regions, some inaccurate segmentations might still exist due to the complexity of the indoor scene. Thus, we use a Robust P^n model (Kohli and Torr, 2009; Yang et al., 2018) (as defined in Eq. (9)) to capture the long-range contextual information, which allows some pixels inside the same segmented object to take different labels and effectively avoids the over-smoothness caused by a rigid consistency.

$$\left.\begin{aligned} \mathrm{En}_{\mathrm{region}}(Y) &= \sum_{c \in S} \psi_c(y_c, x_c) \\ \psi_c(y_c, x_c) &= \begin{cases} W_i(y_c) \dfrac{1}{Q} \lambda_{\max}, & W_i(y_c) < Q \\ \lambda_{\max}, & \text{otherwise} \end{cases} \\ W_i(y_c) &= \min_k \Big(\sum_{j \in c} w_j - \sum_{j \in c} w_j \varsigma(y_j = k) \Big), k = 1,2,..., L \\ \lambda_{\max} &= |c|^{\theta_\alpha} (\theta_p + \theta_v H(c)) \\ H(c) &= \exp\Big(-\theta_\beta \frac{\left\| \sum_{j \in c} (x_j - \mu)^2 \right\|}{|c|}\Big) \end{aligned}\right\} \quad (9)$$

where S denotes the number of visual homogeneous regions, $\psi_c(y_c, x_c)$ denotes the higher order potentials on region c, w_j denotes the class probability for pixel j, $\varsigma(\cdot)$ is the zero-one indicator function, $W_i(y_c)$ measures the inconsistency cost by accumulating the class probability of pixels, Q represents the threshold controlling the rigidity of the higher order potentials, $\lambda_{\max}$ is the homogeneity of each segmented objects, $\|\bullet\|$ denotes the l_2 norm, $\mu = \sum_{j \in c} x_j / |c|$ denotes the mean semantic feature vector, θ_β, θ_β, θ_β and θ_β are parameters.

3. EXPERIMENTATION AND ANALYSIS

To evaluate the effectiveness and robustness of the proposed method, in this section, we performed both qualitative and quantitative analysis on the public SUN RGB-D dataset (Song et al., 2015).

3.1 Experimental data and evaluation criteria

SUN RGB-D dataset is a scene understanding dataset with indoor scene images, 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. Like (Song et al., 2015), the whole dataset was divided into 5285 image pairs for training and 5050 image pairs for test.

In this paper, we use 5 common evaluation criteria, including the global accuracy, the class accuracy, the mean class accuracy, the intersection-over-union (IoU) score (Everingham et al., 2010) and the mean IoU, to measure the segmentation quality: the global accuracy represents the percentage of pixels correctly classified by the division of the total number of pixels of true positive and the total number of pixels of ground true, the class accuracy measures the percentage of pixels correctly classified in a class i, the mean class accuracy represents the mean of the accuracy over all classes by the division of the sum of class accuracy in all classes and the number of classes, IoU is a measure which imposes the penalty of false positive on the class accuracy in a class i, and the mean IoU is the mean of intersection over union in all classes.

$$\text{Globalaccuracy} = \frac{TP}{GT} \tag{10}$$

$$\text{Classaccuracy}_i = \frac{TP_i}{GT_i} \tag{11}$$

$$\text{Mean classaccuracy} = \frac{\sum_{i=1}^{C} \text{Classaccuracy}_i}{C} \tag{12}$$

$$\text{IoU}_i = \frac{TP_i}{GT_i + FP_i} \tag{13}$$

$$\text{Mean IoU} = \frac{\sum_{i=1}^{C} \text{IoU}_i}{C} \tag{14}$$

where TP and GT denote the total number of pixels of true positive and ground true respectively, TP_i, GT_i and FP_i denote the number of pixels of true positive, ground true and false positive in a class i respectively.

3.2 Experimental analysis

As aforementioned, the burry boundaries are common in the semantic segmentation results of the conventional fully convolutional network architectures because of the pooling operations. Furthermore, the depth information can be used for improving the performance of the semantic segmentation and how to use the depth information is still an open area. Thus, we propose a higher-order MRF framework for exploiting the depth data and further optimizing the semantic segmentation, particular over the boundaries among objects, deriving from the existing RefineNet architecture. First, to evaluate the effectiveness of the proposed method, we compared the proposed method with the conventional RefineNet architecture. Table I lists the performance comparisons in the class accuracy, the mean class accuracy, the IoU and the mean IoU between the conventional RefineNet architecture and the proposed method on SUN RGB-D dataset. Fig. 4 demonstrates some typical comparisons between the conventional RefineNet architecture and the proposed method. Experiments suggested that for most indoor objects, the proposed method could further optimize the semantic segmentation results, especially over the object boundaries (as shown in Fig.4), and provide the better performance compared with the conventional RefineNet, with the difference in class accuracy of 2.26% on average and in IoU of 2.33 on average.

Class	Class accuracy (%)		IoU (%)	
	Refine Net	Proposed method	Refine Net	Proposed method
Wall	90.79	**91.73**	78.13	**78.93**
Floor	93.93	**94.20**	83.54	**86.26**
Cabinet	65.42	**67.72**	44.34	**46.66**
Bed	75.94	**77.19**	63.04	**63.84**
Chair	79.97	**82.98**	65.34	**68.43**
Sofa	65.09	**68.15**	53.13	**55.62**
Table	67.98	**69.54**	46.69	**50.58**
Door	58.58	**59.18**	44.84	**45.48**
Window	65.82	**66.87**	50.35	**51.68**
Bookshelf	**44.21**	43.94	35.55	**36.34**
Picture	70.36	**73.60**	54.04	**57.99**
Counter	54.52	**55.75**	43.88	**45.46**
Blinds	**53.57**	53.18	38.20	**39.99**
Desk	22.08	**22.68**	16.15	**16.70**
Shelves	15.97	**16.89**	9.68	**10.43**
Curtain	64.37	**69.00**	55.22	**59.38**
Dresser	33.05	**39.93**	29.71	**33.83**
Pillow	51.63	**54.21**	37.38	**40.47**
Mirror	47.79	**51.29**	38.75	**42.51**
FloorMat	0	0	0	0
Clothes	42.80	**46.51**	29.33	**31.84**
Ceiling	79.77	**81.43**	68.12	**69.47**
Books	56.42	**61.53**	35.95	**39.92**
Fridge	52.02	**56.31**	46.69	**51.14**
TV	74.44	**78.64**	57.76	**59.03**
Paper	41.06	**43.79**	26.22	**29.24**
Towel	33.45	**38.85**	24.34	**29.00**
ShowerCuratain	0	0	0	0
Box	40.15	**43.95**	26.00	**30.62**
Whiteboard	68.30	**69.70**	61.32	**62.84**
Person	76.15	**78.28**	63.77	**67.88**
NightStand	**5.46**	4.07	5.00	**3.41**
Toilet	81.61	**87.17**	74.85	**80.71**
Sink	72.70	**75.86**	57.57	**61.27**
Lamp	**49.08**	48.33	38.34	**39.64**
Bathtub	60.23	**60.99**	**54.71**	**54.51**
Bag	31.28	**36.26**	20.51	**23.92**
Mean	53.68	**55.94**	42.67	**45.00**

Table I Performance comparison between the conventional RefineNet architecture and the proposed method on SUN RGB-D dataset. The best performance is marked with **BOLD** fonts.

Second, to further evaluate the effectiveness of the proposed method, the other existing architectures were used to compare with the proposed method. Table II lists performance comparison between the proposed method and the other existing architectures. For the oexisting architectures in Table II, we copied the best performances in these papers (Chen et al., 2014; Kendall et al., 2015; Badrinarayanan et al., 2017; He et al., 2017; Li et al., 2017). Among all the methods, the proposed method achieved the best performance in global accuracy, mean class accuracy and mean IoU, with difference of 7.69%, 12.17% and 12.89% on average. Experimental comparisons further

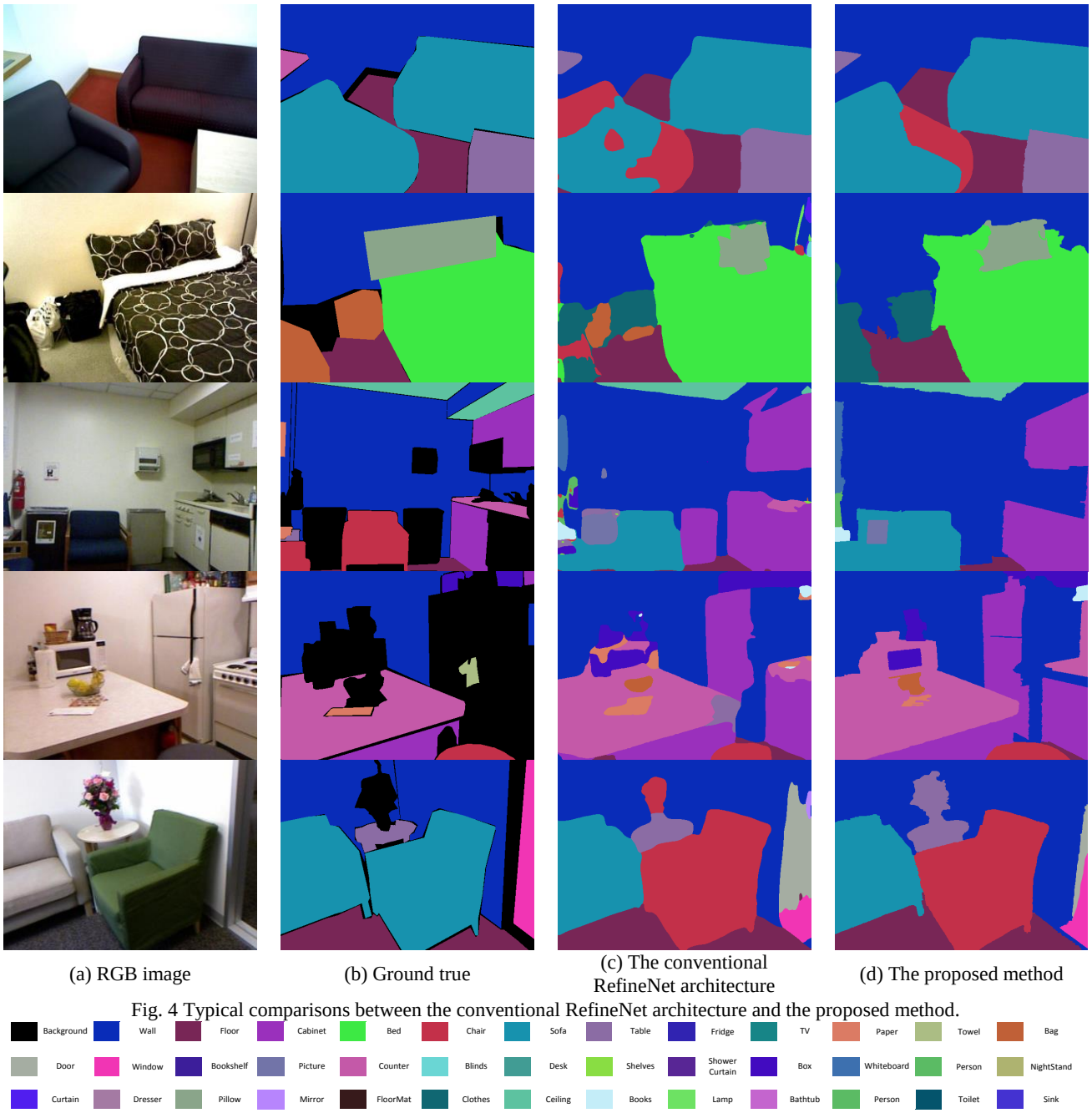
Methods	Data	Global accuracy	Mean class accuracy	Mean IoU
DeepLab+DenseCRF (Chen et al., 2014)	RGB	66.69	33.06	24.13
SegNet (Badrinarayanan et al., 2017)	RGB	72.63	44.76	31.84
Bayesian SegNet (Kendall et al., 2015)	RGB	71.20	45.90	30.70
The superpixel-based multi-view method (He et al., 2017)	RGBD	65.50	41.20	32.90
The semantics-guided multi-level method (Li et al., 2017)	RGBD	78.07	53.93	40.98
The proposed method	RGBD	**78.51**	**55.94**	**45.00**

Table II Performance comparison between the proposed method and the other existing architectures. The best performances in all methods are marked with **BOLD** fonts.

illustrate that the proposed method succeeded in the improvement of semantic segmentations.

4. CONCLUSION

We developed a method based on higher-order Markov random field model for indoor semantic segmentation from RGB-D images. In this paper, we used the depth information for enhancing the performance of the watershed algorithm and combined the high-level semantic information with the long-range contextual information for improving the semantic segmentation results under the higher-order MRF framework. Although experimental results suggested the improvements in the semantic segmentation results to some extent, the final



(a) RGB image (b) Ground true (c) The conventional RefineNet architecture (d) The proposed method

Fig. 4 Typical comparisons between the conventional RefineNet architecture and the proposed method.

results primarily depend on the initial semantic segmentation derived from the fully convolutional network architecture and the robust visual homogeneous region generations. Our future work will focus on further improving the performance of the fully convolutional network architecture itself and enhancing the robustness of producing the visual homogeneous regions.

References

Badrinarayanan, V., Kendall, A., and Cipolla, R., 2017. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Scene Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 99, pp. 1-1.

Chen, L. C., Papandreou, G., Kokkinos, I., Murphy, K., and Yuille, A. L., 2014. Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs. *Computer Science*, no. 4, pp. 357-361.

Chen, L. C., Papandreou, G., Schroff, F., and Adam, H., 2017. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv*:1706.05587.

Chu, J., Xiao, X., Meng, G., Wang, L., and Pan, C., 2017. Learnable contextual regularization for semantic segmentation of indoor scene images. *IEEE International Conference on Image Processing. IEEE*. pp.1267-1271.

Everingham, M., Van, Gool. L., Williams, C. K. I., Winn, J., and Zisserman, A., 2010. The pascal visual object classes (voc) challenge. *International journal of computer vision*, vol. 88, no. 2, pp. 303-338.

Geman, S., and Geman, D., 1987. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *Readings in Computer Vision*. pp. 564-584.

Gupta, S., Arbelaez, P., and Malik, J.,2013. Perceptual organization and recognition of indoor scenes from RGB-D images. *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on. IEEE*, pp. 564-571.

He, K., Zhang, X., Ren, S., and Sun, J., 2016. Deep residual learning for image recognition. *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 770-778.

He, Y., Chiu, W., Keuper, M., and Fritz, M., 2017. STD2P: RGBD Semantic Segmentation Using Spatio-Temporal Data-Driven Pooling. *Computer Vision and Pattern Recognition. IEEE*, pp. 7158-7167.

Höft, N., Schulz, H., and Behnke, S., 2014. Fast Semantic Segmentation of RGB-D Scenes with GPU-Accelerated Deep Neural Networks. *German Conference on Artificial Intelligence*. pp. 80-85.

Husain, F., Schulz, H., Dellen, B., Torras, C., and Behnke, S., 2016. Combining Semantic and Geometric Features for Object Class Segmentation of Indoor Scenes. *IEEE Robotics & Automation Letters*, vol. 2, no. 1, pp. 49-55.

Jiang, J., Zheng, L., Luo, F., and Zhang, Z., 2018. RedNet: Residual Encoder-Decoder Network for indoor RGB-D Semantic Segmentation. *arXiv preprint arXiv*:1806.01054.

Kendall, A., Badrinarayanan, V., and Cipolla, R., 2015. Bayesian SegNet: Model Uncertainty in Deep Convolutional Encoder-Decoder Architectures for Scene Understanding. *Computer Science*.

Khan, S. H., Bennamoun, M., Sohel, F., and Togneri, R., 2014. Geometry driven semantic labeling of indoor scenes. *European Conference on Computer Vision*. Springer, Cham, pp. 679-694.

Khoshelham, K.,2012. Accuracy Analysis of Kinect Depth Data. *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 3812, no. 5, pp. 133-138.

Kohli, P., and Torr, P. H. S., 2009. Robust higher order potentials for enforcing label consistency. *International Journal of Computer Vision*, vol. 82, no. 3, pp. 302-324.

Lewis, S. H., and Dong, A., 2012. Detection of breast tumor candidates using marker-controlled watershed segmentation and morphological analysis. *Image analysis and interpretation (SSIAI)*, 2012 IEEE southwest symposium on. IEEE, pp. 1-4.

Li, Y., Zhang, J., Cheng, Y., Huang, K., and Tan, T., 2017. Semantics-guided multi-level RGB-D feature fusion for indoor semantic segmentation. *Image Processing (ICIP)*, 2017 IEEE International Conference on. IEEE, pp. 1262-1266.

Lin, D., Chen, G., Cohen-Or, D., Heng, P., and Huang, H., 2017a. Cascaded Feature Network for Semantic Segmentation of RGB-D Images. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pp. 1311-1319.

Lin, G., Milan, A., Shen, C., and Reid. I., 2017b. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Müller, A. C., and Behnke, S., 2014. Learning depth-sensitive conditional random fields for semantic segmentation of rgb-d images. *Robotics and Automation (ICRA)*, 2014 IEEE International Conference on. IEEE, pp. 6232-6237.

Parvati, K., Rao, P., Mariya, and Das. M., 2008. Image segmentation using gray-scale morphology and marker-controlled watershed transformation. *Discrete Dynamics in Nature and Society*.

Ren, J., Gong, X., Yu, L., Zhou, W., and Yang, M., 2015. Exploiting global priors for RGB-D saliency detection. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. pp. 25-32.

Ren, X., Bo, L., and Fox, D., 2012. Rgb-(d) scene labeling: Features and algorithms. *Computer Vision and Pattern Recognition (CVPR)*, 2012 IEEE Conference on. IEEE, pp. 2759-2766.

Ronneberger, O., Fischer, P., and Brox, T., 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, Cham, pp. 234-241.

Silberman, N., and Fergus, R.,2011. Indoor scene segmentation using a structured light sensor. *Computer Vision Workshops (ICCV Workshops)*, 2011 IEEE International Conference on. IEEE, pp. 601-608.

Sobel, I., Feldman, G., and Feldman, G., 1968. A 3x3 isotropic gradient operator for image processing. *Die Pharmazie*, vol. 7, no. 8.

Song, S., Lichtenberg, S. P., and Xiao, J., 2015. SUN RGB-D: A RGB-D scene understanding benchmark suite. *Computer Vision and Pattern Recognition*. IEEE, pp. 567-576.

Vincent, L., 1993. Morphological grayscale reconstruction in image analysis: applications and efficient algorithms. *IEEE Trans Image Process*, vol. 2, no. 2, pp. 176-201.

Vincent, L., and Soille, P., 1991. Watersheds in digital spaces: an efficient algorithm based on immersion simulations. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 6, pp. 583-598.

Wolf, D., Prankl, J., and Vincze, M., 2015. Fast semantic segmentation of 3D point clouds using a dense CRF with learned parameters. *IEEE International Conference on Robotics and Automation*. IEEE, pp. 4867-4873.

Woodford, O., Torr, P., Reid, I., and Fitzgibbon, A., 2009. Global stereo reconstruction under second-order smoothness priors. *IEEE transactions on pattern analysis and machine intelligence*, vol. 31, no. 12, pp. 2115-2128.

Xu, S., Liu, H., and Song, E., 2011. Marker-controlled watershed for lesion segmentation in mammograms. *Journal of digital imaging*, vol. 24, no. 5, pp. 754-763.

Yang, J., Jiang, Z., Hao, S., and Zhang, H., 2018. Higher Order Support Vector Random Fields for Hyperspectral Image Classification. *ISPRS International Journal of Geo-Information*, vol. 7, no. 1, pp. 19.

Zhao, H., Shi, J., Qi, X., Wang, X., and Jia, J., 2017. Pyramid scene parsing network. *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. pp. 2881-2890.