SCIENCE CHINA
Information Sciences

• LETTER •

December 2019, Vol. 62 229101:1–229101:3
https://doi.org/10.1007/s11432-018-9828-3

Multi-view based neural network for semantic segmentation on 3D scenes

Yonghua LU[1], Mingmin ZHEN[2*] & Tian FANG[2]

[1]*School of Resource and Environmental Sciences, Wuhan University, Wuhan* 430072*, China;*
[2]*Department of Computer Science and Engineering, The Hong Kong University of Science and Technology, Hong Kong, China*

Received 28 October 2018/Accepted 31 January 2019/Published online 4 September 2019

Citation Lu Y H, Zhen M M, Fang T. Multi-view based neural network for semantic segmentation on 3D scenes. Sci China Inf Sci, 2019, 62(12): 229101, https://doi.org/10.1007/s11432-018-9828-3

Dear editor,

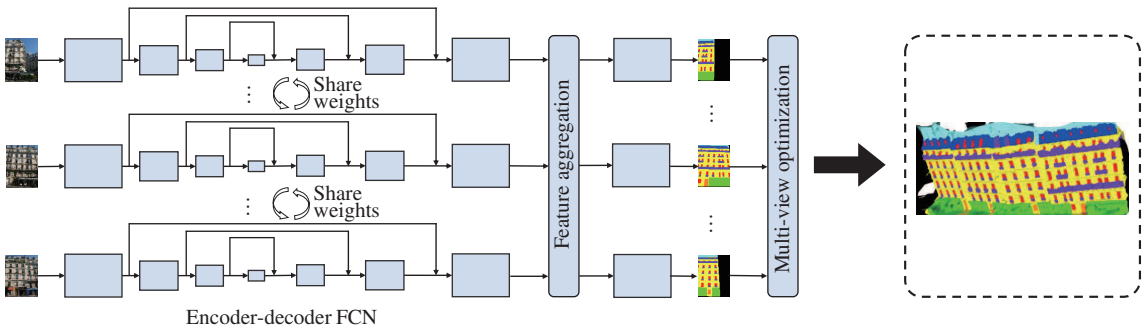
For semantic segmentation tasks, convolutional neural network (CNN) based methods have been prevalent for both 2D image semantic segmentation and 3D semantic segmentation. Though traditional methods often use local features to segment a target (For example, in [1] both 2D local features and 3D local features are used to boost recognition ability), CNN based methods [2,3] exhibit much better performance than traditional methods [4]. In all the CNN based methods on images, fully convolutional networks (FCNs) [2] are firstly proposed for end-to-end training. Basically, all the following methods are the variants of FCNs. For 3D input, some studies leverage 3D convolution to predict dense 3D semantic voxel maps [5]. However, 3D convolution has the limitation of low resolution as the GPU memory constraint. Additionally, RGB information is not well considered though it is very important. As semantic segmentation on images has been very good, we can project the semantic segmentation results of images to 3D mesh based on the geometric relationship. In this study, we mainly exploit the multi-view based neural network for semantic segmentation on 3D scenes.

We bridge the gap between 2D images and 3D scene understanding because each pixel in the images has a corresponding point or face in the reconstructed 3D model. We try to make better use of well-defined 2D neural networks to segment images. After that, the views are projected to a 3D model by the one-to-one correspondence. In order to make the entire segmentation consistent and coherent, conditional random field (CRF) is adopted to take multi-view projective results into account so that the semantic segmentation on the whole 3D scene is optimal.

In summary, we present a novel architecture about a multi-view based neural network for semantic segmentation on 3D scenes. Our contributions are twofold: (1) we propose a pipeline to deal with 3D scene semantic segmentation by using a multi-view based method, which bridges the gap between images and a 3D scene; (2) we use the CRF based method to optimize the final labels of a 3D scene to obtain a coherently consistent results, which obviously improves the results.

Given an input 3D scene and corresponding images, the goal of our method is to label each face in the 3D scene. In order to make use of the images with rich textural information and the 3D geometric structure of a 3D scene, we construct a multi-view encoder-decoder architecture as shown in Figure 1(a). It takes as input a set of images from multiple views which are used to reconstruct the 3D scene and extracts confidence maps through 2D CNN layers. The confidence maps are then combined and projected onto the 3D scene. Finally, we use the CRF based multi-view optimization method to obtain the final semantic segmentation result in order to keep geometric consistency.

* Corresponding author (email: mzhen@connect.ust.hk)



(a)

Method	GCO [8]	GCO+recode [8]	AutoContext [9]	Ours
Mean IoU (%)	37.33	41.92	63.70	67.34

(b)

Figure 1 (Color online) (a) The proposed multi-view based neural network architecture for semantic segmentation on the 3D scenes. The encoder-decoder structure is based on the ResNet101 backbone and multi-stage decoder is used to restore resolution. The feature aggregation module is used to aggregate the features from multi-view feature maps. The multi-view optimization module is used to optimize the semantic segmentation result by conditional random field (CRF). (b) The performance (mean intersection of unison) comparison with other methods on the 3D scenes of the RueMonge dataset.

Encoder-decoder. In the 2D image CNN step, we use encoder-decoder structure, which is mostly adopted in 2D image semantic segmentation tasks [6]. The encoder network extracts features from an image and the decoder network produces the final semantic segmentation labels. The encoder network is combined with layers of convolution, pooling, non-linear activation which generally uses a pretrained model (ResNet [7]) as the backbone. The output of each convolution layer in the encoder network can be interpreted as features with different receptive fields. After the spatial pooling operation, the size of the feature map produced by the encoder network is smaller than the original image. The decoder network will then enlarge the feature map using upsampling and unpooling in order to produce the probility map with the same size as the input.

In our encoder-decoder structure, we use "skip connections" [6] to make better use of previous feature maps. The skip connection directly links an encoder layer to a decoder layer. We use ResNet-101 as the backbone to extract feature maps and then use them in the decoder module. The outputs of the encoder-decoder are of the same size as the input images. We use multiple branches for multiple views and the weights for these branches are shared.

Feature aggregation. In order to encode the geometric information into 2D CNN, we add a feature aggregation layer at the end of 2D CNNs. Because the images from multiple views can be used to reconstruct the 3D scene, the projection between the pixels and the faces of 3D scene is known. In other words, we can obtain the pixels of different images that are mapped to the same face. As these pixels correspond to the same face in a 3D scene, they should have similar features. In order to gain better performance, we aggregate the feature maps of the related pixels from different images. Comprehensively, after obtaining features $f_1, f_2, \ldots, f_K$ (K views), we can use the precomputed Face-ID map, which stores the face id of each corresponding pixel in the image, to find all the pixels $(p_1^i, p_2^i, \ldots, p_n^i)$ in different images which are mapped to the same face id i. As these pixels correspond to the same face in a 3D scene, the feature maps for them should be consistent. Thus, we aggregate these feature maps as

$$F_i = \frac{1}{K}\sum_{j=1}^{K} f_j, \tag{1}$$

where F_i is of the same channels as the original features. The F_i is then used to replace all the features $(f_1, f_2, \ldots, f_K)$. In this way, after feature aggregation the feature maps are consistent for all the pixels corresponding to the same face. We then add a layer for each view. It is composed of convolution, BN and relu operation. The output of 2D CNN is the probability map $C \times H \times W$, where C is the category we want to classify.

Multi-view optimization. In order to keep geometric consistency in the final segmentation results of the 3D scene, we use conditional random field (CRF) to optimize the labeling results from all the views. We can define a graph $\mathcal{G}_{\mathcal{M}} = (\mathcal{F}, \mathcal{E})$ where the nodes represent the triangular faces $\mathcal{F} = \{f_i\}$ of the 3D mesh $\mathcal{M}$, and $\mathcal{E}$ is the set of graph edges, which encode 3D adjacencies between the faces. For all the faces in the set $\mathcal{F}$, the probability vector

set is $\mathcal{D} = \{d_1, d_2, \ldots, d_n\}$. The labeling problem can be defined as $L = (l_1, l_2, \ldots, l_n)$ for n faces. We define CRF on the graph and try to get the maxiumum-a-posteriori (MAP) labeling L^* of 3D mesh where $L^* = (l_1^*, l_2^*, \ldots, l_n^*)$ for n faces and $l_i \in C = \{c_1, c_2, \ldots, c_k\}$ for k categories. The objective function can be

$$L^* = \underset{L \in C^n}{\operatorname{argmin}} E(L|D), \tag{2}$$

and we try to find an optimal multi-label result L^*. The function consists of an unary data term for each face f_i and a pairwise regularity term for each pair of adjacent faces (f_i, f_j).

$$E(L) = \sum_{f_i \in \mathcal{F}} \phi(d_i|l_i) + \lambda \sum_{(f_i, f_j) \in \mathcal{E}} \Psi(l_i, l_j), \tag{3}$$

where $\phi(d_i|l_i)$ is the penalty for assigning label l_i for face f_i. That is

$$\phi(d_i|l_i) = \sum_{c_j \in C} -\log p(c_j|l_i = c_j), \tag{4}$$

where $p(c_j|l_i = c_j)$ is the probability of f_i assigned label c_j.

The pairwise potential $\Psi(l_i, l_j)$ enforces spatially smooth labelling solutions over the mesh faces by penalizing occurrences of adjacent faces f_i and f_j obtaining different labels $(l_i \neq l_j)$. We use a Potts model:

$$\Psi(l_i, l_j) = \begin{cases} 0, & \text{if} \quad l_i = l_j, \\ D(f_i, f_j), & \text{if} \quad l_i \neq l_j, \end{cases} \tag{5}$$

where $D(f_i, f_j)$ is a function used to describe the relationship of the pair of faces (f_i, f_j). We use a Gaussian function here. That is

$$D(f_i, f_j) = \mathrm{e}^{-\mathrm{dist}(f_i, f_j)}, \tag{6}$$

where $\mathrm{dist}(f_i, f_j)$ is the distance of two faces. For each face f_i, we recompute the point representing the face $p(f_i) = \frac{f_i(p_1) + f_i(p_2) + f_i(p_3)}{3}$. Thus, the distance between f_i and f_j is the Euclidean distance of $p(f_i)$ and $p(f_j)$.

The coefficient parameter λ is used to control the balance between the unary term and the pairwise term. Here we use $\lambda = 0.5$.

Implementation and experiments. In order to evaluate our proposed method, we perform experiments on the RueMonge2014 dataset [8]. We implement our work by using PyTorch on a PC with Intel Core i7 3.10 GHz, 32 GB RAM and a 1080Ti GPU. For the training step, we do some data augmentation as commonly done in the literature. The input images are randomly cropped, rotated, scaled (ranging from 0.5 to 1.2) and horizontally flipped. The training input size is 400×400 in all our experiments. In the testing step, we use multi-scale and flipping input. As shown in Figure 1(b) [8, 9], we compare our results with other methods. Compared with other methods, our pipeline shows a better performance.

Conclusion. In this study, we propose a multi-view based neural network architecture for 3D scene segmentation. Our architecture firstly extracts feature maps from different views by using the encoder-decoder structure with ResNet-101 as the backbone. The skip connection is used in the encoder-decoder structure. The feature maps are aggregated through the projection relationship between the images and the 3D scene. Therefore, the feature maps from different views are kept consistent. In order to make the labeling result coherently consistent, we use a CRF based method to optimize multiple view results. The experiments show that the proposed architecture exhibits better performance compared with other methods.

Acknowledgements This work was supported by GRF (Grant No. 16203518), Hong Kong RGC (Grant Nos. 16208614, T22-603/15N), Hong Kong ITC (Grant No. PSKL12EG02), and National Basic Research Program of China (973 Program) (Grant No. 2012CB316300).

References

1 Wang J L, Lu Y H, Liu J B, et al. A robust three-stage approach to large-scale urban scene recognition. Sci China Inf Sci, 2017, 60: 103101
2 Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation. In: Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015. 3431–3440
3 Kalogerakis E, Averkiou M, Maji S, et al. 3D shape segmentation with projective convolutional networks. In: Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017. 6630–6639
4 Kalogerakis E, Hertzmann A, Singh K. Learning 3D mesh segmentation and labeling. ACM Trans Graph, 2010, 29: 102
5 Dai A, Chang A X, Savva M, et al. Scannet: richly-annotated 3D reconstructions of indoor scenes. In: Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017. 2432–2443
6 Noh H, Hong S, Han B. Learning deconvolution network for semantic segmentation. In: Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV 2015), 2015. 1520–1528
7 He K M, Zhang X Y, Ren S Q, et al. Deep residual learning for image recognition. In: Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2016), 2016. 770–778
8 Riemenschneider H, Bodis-Szomoru A, Weissenberg J, et al. Learning where to classify in multi-view semantic segmentation. In: Proceedings of European Conference on Computer Vision (ECCV 2014), 2014. 516–532
9 Gadde R, Jampani V, Marlet R. Efficient 2D and 3D facade segmentation using auto-context. 2016. ArXiv: 1606.06437