3D Semantic Labeling of Photogrammetry Meshes Based on Active Learning

Mengqi Rong[1,2], Shuhan Shen[1,2], Zhanyi Hu[1,2]
[1]National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China
[2]School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100049, China
{mengqi.rong,shshen,huzy}@nlpr.ia.ac.cn

1. Introduction

In recent years, the traditional geometry-based 3D reconstruction has reached a relatively mature stage. Many scholars are not satisfied with just obtaining the structural information of the scene, and then focus on the expression and understanding of 3D scenes. There is no doubt that an urban model with richer information can be better applied to smart city, urban planing, virtual reality, autonomous driving and so on. To deal with this problem, we propose a procedural approach for 3D semantic expression of urban scenes based on active learning. We first start with a small labeled image set to fine-tune a semantic segmentation network and then project its probability map onto a 3D mesh model for fusion, finally outputs a 3D semantic mesh model in which each facet has a semantic label and a heat model showing each facet's confidence. Our key observation is that our algorithm is iterative, in each iteration, we use the output semantic model as a supervision to select several valuable images for annotation to co-participate in the fine-tuning for overall improvement. In this way, we reduce the workload of labeling but not the quality of 3D semantic model.

2. Overview

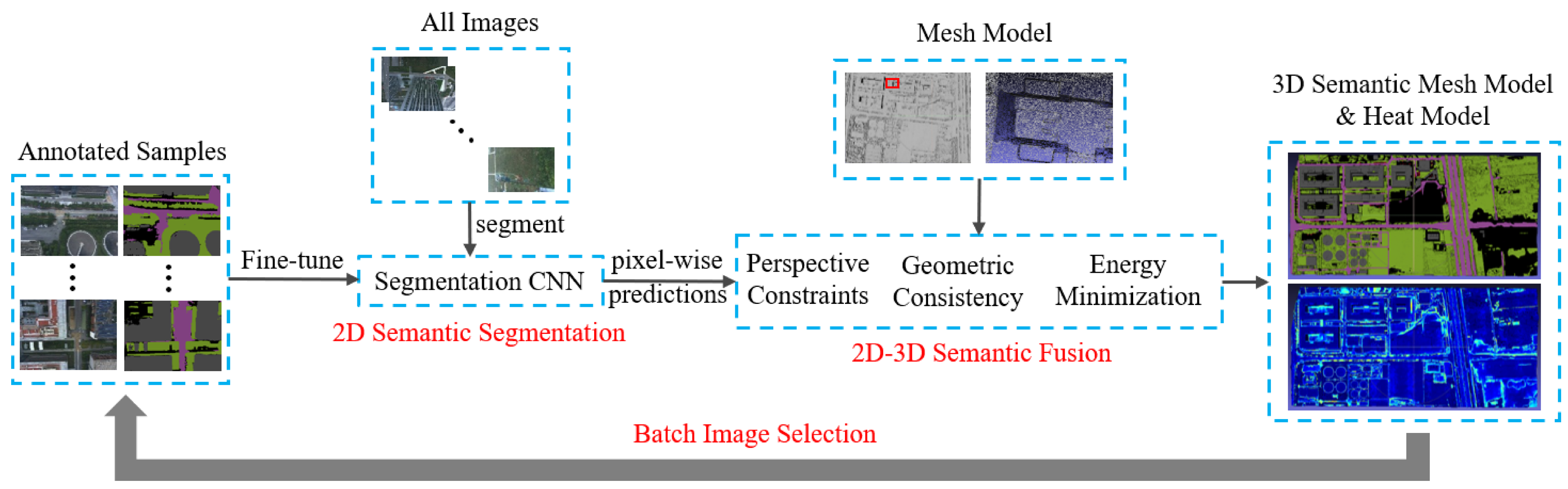


Fig.1 The pipeline consists of three main steps: fine-tuning the 2D semantic segmentation network with an ever-enlarging annotated image set, back-projecting the pixel-wise predictions onto 3D mesh model for semantic fusion based on geometric consistency, selecting a batch of images for annotation and adding them into the training set for the next iteration.

2.1. 2D Semantic Segmentation

At the beginning, all images are unlabeled and we randomly selected several images for manual annotation to fine-tune a segmentation network, here we used DeepLabv3+[1] pre-trained on cityscapes[2]. And we modify the last layer of the network to output the probability that each pixel corresponds to each label.

2.2. 2D-3D Semantic Fusion

A. Back-projection
Given a set of calibrated cameras, the correspondence between the pixels of the images and the facets of the mesh model can be easily calculated by ray intersection. Subsequently, the simplest weighted-average method will be utilized to unify the per-pixel class scores.

B. Geometric Constraint
Besides, we introduced spatial smoothness constraints described in [3] to optimize the label assignment, which means the adjacent facets have a higher probability of being assigned the same label. Given two adjacent facets f_1 and f_2 with the corresponding labels l_{f_1} and l_{f_2}, we define:

$$V_{f_1,f_2} = \begin{cases} 1 & if\, l_{f_1} \neq l_{f_2} \\ \min\left(1, s\left\|W_{f_1} - W_{f_2}\right\|_2\right) & otherwise \end{cases}$$

$$W = \begin{pmatrix} k_{\min} \cdot w_{\min} \\ k_{\max} \cdot w_{\max} \end{pmatrix}$$

where s is a scale factor, and W_{f_1} and W_{f_2} are vectors combining the principal curvature $k_{\min}$, $k_{\max}$ and their principal direction $w_{\min}$, $w_{\max}$.

2.3. Batch Image Selection

The optimized 3D semantic mesh model incorporates both 2D semantic segmentation and 3D geometry information, it could be used as a more reliable supervisor to measure the segmentation quality and help us to determine the next batch of training data for high quality performance.

A. least-Scoring Subset/Uncertainty A straightforward strategy for finding the most valuable annotation areas is to use uncertainty samplings, which is obtained by re-projecting the label and its confidence of each facet onto different 2D images.

$$u_p = \begin{cases} 1 - s_{l_f} & p \cap F = f \\ 0 & p \cap F = \emptyset \end{cases} \qquad U_{I_s} = \sum_{i \in I_s} \sum_{p \in i} u_p$$

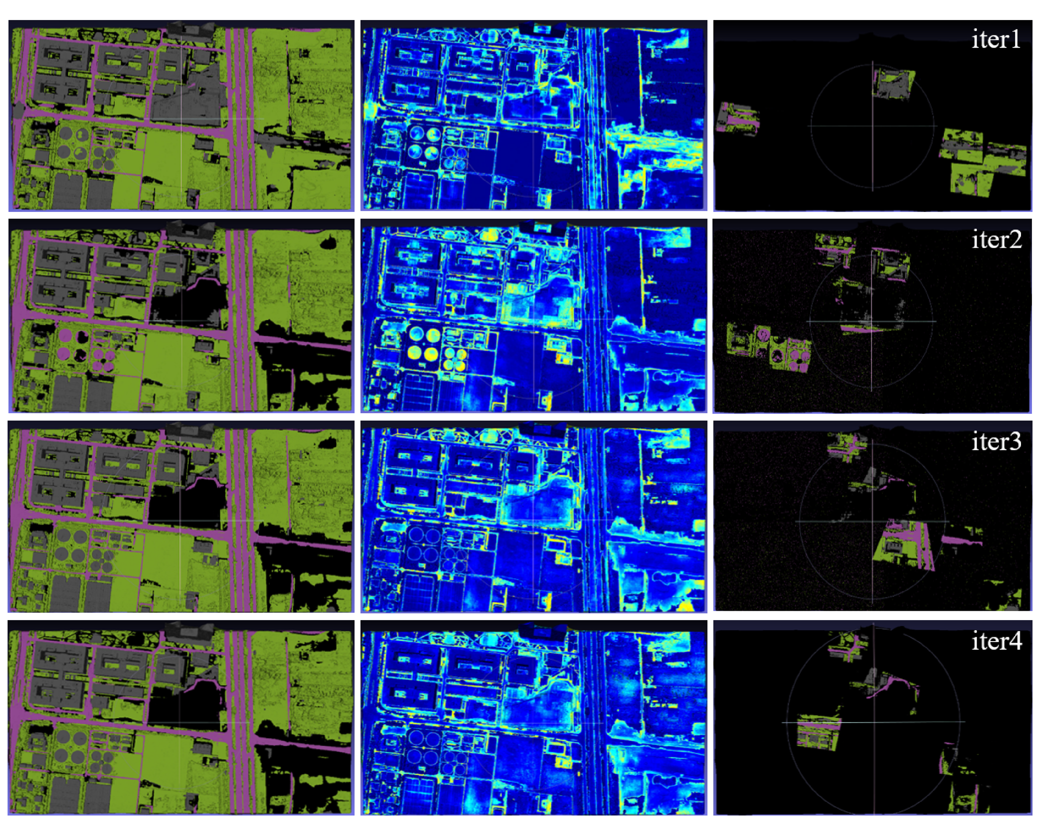
B. largest Coverage Area/Divergence The selected areas are expected to carry as many useful characteristics or features of the unannotated images as possible. We use the coverage area as another measure, that is, the ratio of the intersection of the visible facet by the selected images subset, as:

$$\tilde{C}_{I_s} = \frac{\tilde{F}_\cap}{\tilde{F}_\cup}$$

3. Experimental Results

The regions that were assigned to incorrect semantic labels and accompanied by lower confidences would be selected by our annotation suggestion approach and revised in the next iteration. Four iterations later, the semantic mesh model reached a relatively stable level and the heat model also became smoother.

iterations	2D_Seg_Acc	Number / Percentage	3D_Seg_Acc
iter1	0.8894	——	0.7036
iter2	0.8886	770108 / 0.1569	0.8168
iter3	0.8633	317575 / 0.0647	0.8485
iter4	0.8351	164582 / 0.0335	0.8739



4. Conclusion

We have presented a complete framework of semantic modeling from meshes of large urban scenes. Inspired by active learning, we make our algorithm iterative and propose an annotation suggestion algorithm for selecting the most effective data which would greatly improve the quality of semantic segmentation and then the 3D semantic mesh model. It demands limited human labor but not reduces the labeling quality of 3D models.

5. References

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

[2] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, pages 3213–3223, 2016.

[3] Florent Lafarge, Renaud Keriven, and Mathieu Brédif. Insertion of 3-d-primitives in mesh-based representations: towards compact models preserving the details. *IEEE Transactions on Image Processing*, 19(7):1683–1694, 2010.