POSE-FCN SUPERPIXEL SEGMENTATION FOR BUILDING FACADES BASED ON 2D TEXTURE AND 3D LOCAL POSE-VARIED SEMANTIC FEATURES

Rongchun Zhang [a, c], Yiting He [a], Xuefeng Yi [b, *], Guanming Lu [c], Xiantao Guo [a]

[a] School of Geographic and Biologic Information, Nanjing University of Posts and Telecommunications, 210023 Nanjing, China - (rongchunzhang, 1020173003, xiantaoguo)@njupt.edu.cn
[b] School of Earth Sciences and Engineering, Hohai University, 211100 Nanjing, China - hhuyxf@sina.com
[c] School of Communication and Information Engineering, Nanjing University of Posts and Telecommunications, 210023 Nanjing, China - (rongchunzhang, lugm)@njupt.edu.cn

Commission II, WG II/4

KEY WORDS: Segmentation, Superpixel, Fully Convolution Netural Networks, Multi-modal Features, Building Facades, Integration.

ABSTRACT:

The extraction of building facades based on image sequences has a great contribution to the construction of digital realistic cities. The superpixel segmentation algorithm is a pre-processing tool for segmentation because of its advantages of fast speed, universality and great accuracy. However, the 2D features are less reliable because building facades usually have complex texture and geometric feature. It is difficult to obtain accurate detail information of the façades by clustering the superpixels. Moreover, the process of acquiring building image sequences is easily disturbed by environmental factors, which also leads to the poor results of the superpixel segmentation. In this paper, 3D local pose-varied semantic features of buildings are defined for this problem, which are computed by 3D point clouds generated from multi-view images of buildings based on SfM and PMVS. Then, multi-modal superpixels with integration of 2D texture and 3D pose-varied semantic features are computed by using fully convolutional networks. The new method is compared with traditional superpixel segmentation method by standard superpixel segmentation result evaluation metrics such as achievable segmentation accuracy , boundary recall, and undersegmentation error. The method achieve accurate segmentation results and effectively exclude the influence of complex texture and environmental factors. In summary, The multi-modal superpixels obtained by the integration have better reliability and provide a new idea for the superpixel segmentation of building facades, which has important theoretical and practical significance.

1. INTRODUCTION

The visualization of urban scenes is an important part of digital cities. The reconstruction of building facades in cities values in 3D modeling. Semantic segmentation is an essential techniques that have been widely used with the purpose of segmenting images according to their respective categories by color blocks (Hongshan et al., 2018). Due to the diversity of building facades, it is difficult to obtain accurate information about the details of building facades by relying only on image information when faced with complex textures such as dilapidated building surfaces, advertising painted building surfaces, and irregularly patterned building surfaces. In addition, during the image acquisition process, the obtained images may not entired express the building façade information due to the influence of external environmental factors. Besides 2D features, the building facade also contains rich geometric features that are not affected by external factors, so even the small structure of the surface can be well expressed. However, the building façade segmentation by massive point cloud data is usually inefficient, and the quality of point cloud is difficult to guarantee. Therefore, multi-view images are considered to integrate 2D and 3D features of buildings to obtain images with multi-modal features, in order to improve the accuracy of superpixel segmentation of building facades, and provide better clustering effects for subsequent semantic recognition.

In this study, the multi-view images of the building are used to create a dense point model by structure from motion (SfM). The association model between the images and point cloud is obtained at the same time. Then the normal vector of point cloud is calculated by the method of local neighborhood plane fitting, and the 3D features of building facades are extracted. Using the previously obtained mapping relationship between image and dense point cloud, the 3D features of the building are reprojected onto the 2D image, and the building image with 2D texture features corresponding to 3D pose-varied features is obtained. Next, the images with integrated multi-modal features are used as input to obtain a soft association map through a fully convolutional network. Meanwhile, the superpixel segmentation results of the building façade are obtained based on the prediction results of the soft association map. Finally, the superpixel segmentation results obtained by the new method are evaluated and compared. The superpixel segmentation based on the integration of multi-modal features effectively compensates for the shortcomings of the traditional method and provides higher accuracy and richer information for the subsequent clustering step.

* Corresponding author

The contributions of this paper are:

- We constructed 3D pose-varied semantic features to represent the geometric features of the building façade.
- We obtained the mapping relationship between 2D and 3D features through the process of generating point clouds by images. Meanwhile, we constructed the building images with integrated multi-modal features through the reprojection algorithm.
- By the FCN model, we obtained the results of superpixel segmentation of images based on 2D texture semantics and 3D pose-varied semantics. We demonstrated the reliability of the novel method when targeting various kinds of buildings by visualizing images and quantitative metrics.

The remainder of this paper is organized as follows. Related work is introducted in Section 2, followed the detailed method of Pose-FCN algorithm for superpixel segmentation for building facades based on 2D texture semantics and 3D pose varied semantics in Section 3. Section 4 contains the experiments and the evaluation of the method. The conclusions are in Section 5.

2. RELATED WORK

With the development of digital cities, the requirements for building model accuracy are getting higher and higher. Building façade features, such as doors, windows, walls and other structures are important components of building models. At present, how to extract building façade features with high accuracy and efficiency has become an important research topic. This section reviews researches related to semantic segmentation of building facades.

With the advent of various high-precision sensors, 2D images have become easier to acquire and can truly and reliably represent the rich color and texture information of building facades. Since the 1970s, methods for 2D image segmentation have flourished (Chao et al., 2012, Tyleek et al., 2013). But for building façade, the traditional optical imaging-based system is vulnerable to external conditions during image acquisition and the image quality is unstable. 3D point clouds are independent of the external environment and image distortion, and can reflect the building structure more accurately. With the application of DNN in 3D point cloud segmentation, many segmentation methods based on point cloud alone have also been proposed (Qi et al., 2017, Li et al., 2018, Lin et al., 2020). However, the current algorithms are all applicable to very small 3D point clouds, which are less efficient and do not have good generalization in the face of massive point cloud data and complex scenes.

2D images can provide rich texture semantic information, but 2D image semantic segmentation is not effective when the color and texture of dilapidated building surfaces are scattered or influenced by the environment. Compared with 2D data, 3D point cloud data has richer pose semantic information, but the segmentation accuracy is not high considering only geometric information. Therefore, we consider fusing the two kinds of information to get more accurate segmentation results based on more efficient image semantic segmentation algorithms. To fuse the 2D and 3D features of buildings, the reconstruction of 3D information from the images is a very important part. In recent years, many methods based on single-view or multi-view images to extract 3D information have been proposed, such as recovering depth map from a single image (Eigen et al., 2014), expressing 3D information using voxels based on single or multiple views (Choy et al., 2016), and enerating triangular grid models using a single RGB image (Wang et al., 2018). Currently, the method of generating SfM dense point cloud models based on multi-view images has been widely used in various fields (Riemenschneider et al. 2014, Martinovic et al. 2015). The method can visualize the geometric features of the building facade, and at the same time construct a transformation relationship between images and point clouds to realize the integration of 2D and 3D features.

To reduce data redundancy and computational complexity, superpixel was proposed in 2003 (Ren et al., 2003) and became a key preprocessing technique in image segmentation. In recent years, superpixel segmentation algorithms have achieved rich results. SLIC (Achanta et al., 2012) proposed in 2012 has pushed the superpixel research into a high speed development period. CAS (Xiao et al., 2018) has the ability of adaptively updating feature weights; MBS (Hu et al., 2018) uses the minimum obstacle distance as the metric distance and can segment image contours more accurately; SCALP (Giraud et al., 2018) is a linear path finding method based on color and edge contour features; SNIC (Achanta et al., 2017) has further improved the performance of the superpixel segmentation algorithm.

Compared with traditional methods, deep convolutional neural networks can automatically extract more discriminative high-level features. In 2015, Berkeley's team proposed the FCN method (Long et al. 2015), which extends deep learning-based image segmentation to the pixel level, and FCN replaces the fully connected layer of DCNN used for image classification with a fully convolutional layer and outputs a 2D feature map. Subsquently, deconvolutes to obtain the semantic segmentation result of the original map size, formally applying convolutional neural networks to image semantic segmentation.

In summary, the current method of semantic segmentation using fully convolutional neural networks based on image and point cloud semantic information has become a hot research topic. Considering 2D texture semantics and 3D pose-varied semantics together can significantly improve the accuracy and efficiency of image segmentation. How to achieve more efficient and accurate superpixel semantic segmentation of building facades based on fully convolutional neural networks, integrating 2D color texture and 3D pose-varied features of building facades, is the focus of this paper.

3. METHODOLOGY

In this paper, we proposed a Pose-FCN superpixel segmentation method for building facades based on multi-modal semantic features and FCN. The flowchart of the method is show as Figure 1. It includes three parts as follows :

1) Firstly, we use the multi-view images of the building to generate a 3D dense point cloud model based on SfM and PMVS (Furukawa et al., 2009), and extract 3D pose-

varied semantics (angle, orientation, curvature, etc.) of buildings by point cloud.

2) The mapping relationship between point cloud and images is obtained by a 2D-3D coordinate transformation mathematical model, so the point cloud is able to be reprojected onto the 2D image. In other words, we construct an image that can express the 3D pose-varied semantics of the building.

3) Building images with 2D and 3D semantics are used as inputs to a fully convolutional network, respectively. Through the mapping relationship between the multi-modal features of buildings, an integrated soft associate map Q based on the integration of multi-modal features that considers both 2D and 3D semantics of buildings is established. So far, the superpixel segmentation results of building facades based on multi-modal semantics has been achieved.

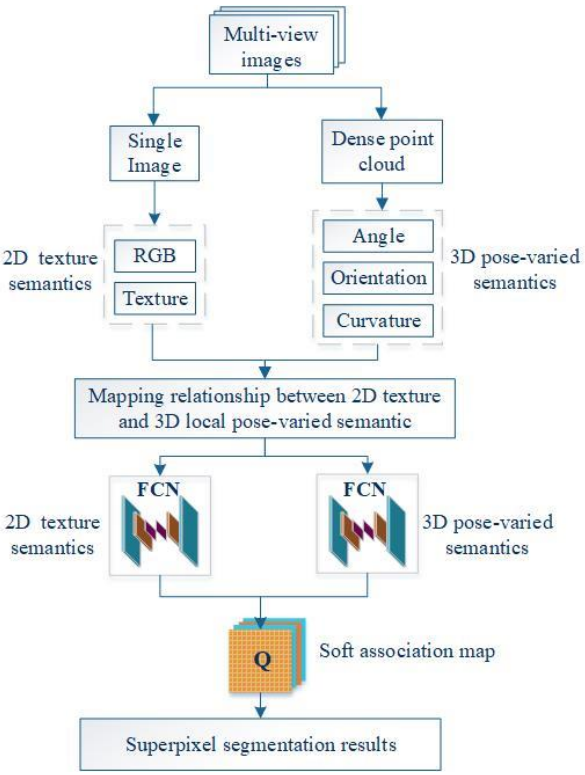


Figure 1. Flowchart of Pose-FCN.

3.1 3D local pose-varied semantic features of buildings

Building fa çades not only has rich color and texture features, but also contains geometric features. When the building fa çades has complex textures, or when the building is affected by external conditions, such as foreground occlusion, shadow coverage, or interlacing of similar buildings, it is not easy to obtain superpixel segmentation accurately by relying on 2D texture features alone. Therefore, in this study, we define 3D pose-varied semantic features, which contain geometric features such as angle, horizontal orientation, and curvature of the building fa çade, as shown in Figure 2.

3D pose-varied semantic features describe the geometric features of the building facade. The spatial pose of each point is calculated one by one based on the 3D coordinates of the points in the dense point cloud. The points in the dense point cloud generated by multi-view images based on SfM and PMVS are fitted with other points in the local neighborhood to form the fitting plane. Angle indicates the angle of the vertical normal vector projection of the fitting plane in the horizontal plane coordinate system, and ranges from 0° to 360°; horizontal orientation indicates the angle between the fitting plane and the horizontal plane, and ranges from 0° to 90°; curvature indicates the curvature of the fitting plane. Angle and horizontal orientation can effectively reflect the orientation of building fa çades, separating the adjacent fa çades. Curvature is suitable for identifying buildings with curved or spherical surfaces.

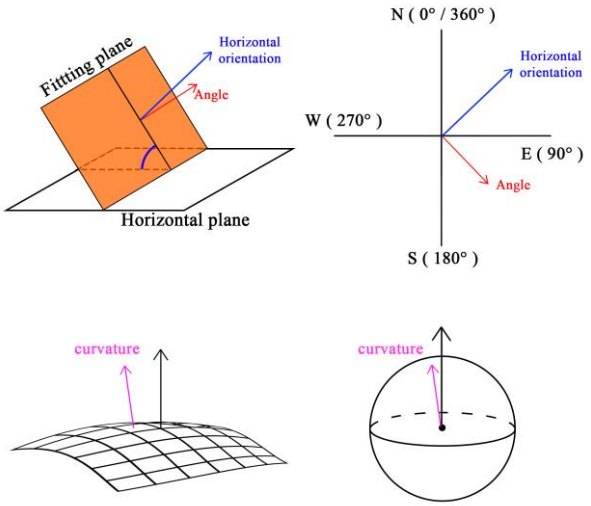


Figure 2. 3D pose-varied semantics of buildings.

3.2 Segmentation with integrated multi-modal features and FCN

The mapping relationship between 2D texture features and 3D pose-varied semantic features is established by the projection relationship between the pixel points on the a single image among multi-view images and the coordinates of SfM dense point. Based on interior and exterior orientation parameters corresponding to each image in the process of generating a dense point cloud through multi-view images, we can calculation the the image coordinates (x, y) of point cloud. Then,the pixel coordinates A of SfM dense point cloud in the image are :

$$A = \begin{bmatrix} x_{col} \\ y_{row} \end{bmatrix} = \begin{bmatrix} (width/2) + x \\ (height/2) - y \end{bmatrix}. \quad (1)$$

Through the equation above, the SfM dense point cloud is reprojected on to the corresponding image of the building. In the meanwhile, mapping relationship between 2D texture semantic feathers from the image and 3D pose-varied semantic features from the dense point cloud has been established.

Subsequently, the images with 2D texture semantic features and 3D pose-varied semantic features of the building are

initialized into a series of regular grid cells, respectively. The Pose-FCN method based on fully convolution neural networks uses a standard encoder-decoder structure. Therefore, the soft association maps Q_1 and Q_2 of the surrounding 3 × 3 neighborhood of pixel-grid cell based on 2D color and texture semantic features and 3D pose-varied semantic features are obtained, respectively. The structure of FCN network is show as Figure 3.

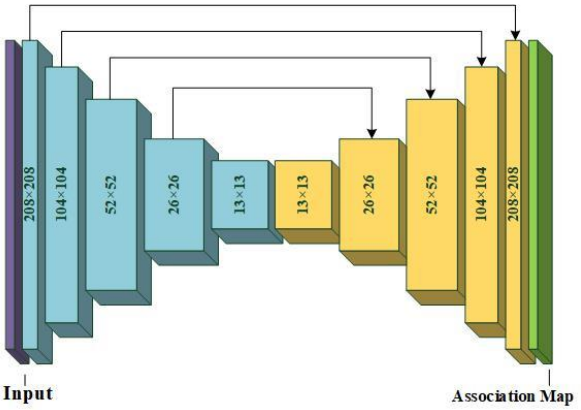


Figure 3. The structure of FCN network for predicted soft association map.

The superpixel s can be obtained by soft association maps Q_1 and Q_2, where m_s represents the texture or pose-varied feature corresponding to superpixel s and n_s represents the location of the center of this superpixel. With these vectors, the pixel points can be reconstructed as:

$$f'(A) = \sum_{s \in N_0} m_s \cdot q_s(A), A' = \sum_{s \in N_0} n_s \cdot q_s(A) \cdot \quad (2)$$

where $f'(A)$ = created property of pixel
A' = created position of pixel
$q_s(A)$ = attribution probability of A
$N_0 = 9$

Finally, the loss function can be constructed from the original pixel points and the created pixel points, as in Equation 3.

$$loss(Q) = \sum_p dist(f(A), f'(A)) + \frac{k}{I}\|A - A'\|_2 \cdot \quad (3)$$

where $dist(\cdot,\cdot)$ = a specific distance metric
I = sampling interval
k = a weight factor

Based on the mapping relationship between the 2D texture semantics and 3D pose-varied semantics that calculated above, together with the predicted soft association maps Q_1 and Q_2 by FCN, a soft association map for multi-modal semantic features of pixels is established as:

$$Q = \omega_1 Q_1 + \omega_2 Q_2 \cdot \quad (4)$$

where $\omega_1 + \omega_2 = 1$
ω_1, ω_2 = weight factors for the two soft association maps

A soft association map considering both 2D texture semantics and 3D pose-varied semantics is obtained. The final results are get by distributing each pixel to the surrounding 9 neighborhood of grid cell with the highest prediction probability. So far, superpixel segmentation results of integrated FCN soft association map based on multi-modal features are achieved.

4. EXPERIMENTS AND RESULTS

In this study, we used the publicly available dataset from LUND University (Enqvist et al., 2012), which contains image sequences of buildings with completely different color textures and geometric features. In this paper, the Skansen Kronan in Gothenburg is used as an example. 131 multi-view images of buildings with overlap of more than 70%, which have a resolution of 1936 × 1296 are included in the Skansen Kronan image sequences. A thumbnail of part of the image sequence is shown in Figure 4.

Figure 4. Mult-view images of the building.

By aligning the image sequences of buildings with high overlap rate, the sparse point cloud of the building can be calculated by applying SfM theory. At the same time, the camera coordinates and interior and exterior orientation parameters of each photograph taken can be obtained. The 3D reconstruction was completed by VisualSFM (Wu, 2013, Wu et al., 2011) in this experiment. SfM extracts image feature points by SIFT operator with scale and rotation invariance, and then uses kd-tree to calculate the distance between two image feature points and perform feature point matching. Then, the base matrix is calculated by RANSAC and the matched point pairs that do not satisfy the base matrix are eliminated. In the next step, the camera parameters and 3D point coordinates are solved by iteration. At this point, the sparse point cloud model is reconstructed.The generated sparse point cloud with 47,208 points and the corresponding camera positions are shown in Figure 5.

Based on the sparse point cloud, the dense point cloud of the building with 4,119,345 points is reconstructed by the feature

points between the images. Then the dense point cloud is cropped in order to remove the influence of external environmental factors such as sky, foreground occlusion and vegetation interference, and keep the point cloud of the building completely. The reconstructed and pre-processed SfM dense point cloud is shown in Figure 6(a) with 1,460,736 points. The reconstructed dense point cloud detail part is shown in Figure 7. The PMVS method is very effective for the geometric structure of details such as eaves, windows, and façade edges. For regions where the point cloud is sparse, interpolation can be used to supplement it.

Figure 5. Sprase point cloud and camera positions generated by multi-view images based on SfM.

(a) dense point cloud (b) angle

(c) horziontal orientation (d) curvature

Figure 6. Dense point cloud and visualized 3D pose-varied semantics of the building.

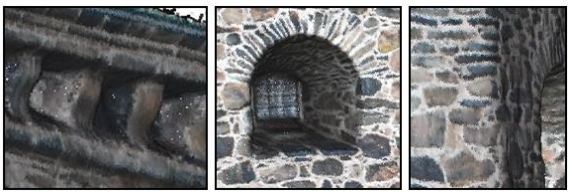


Figure 7. Details of the sprase point cloud.

In the case that the sampled surface of the point cloud is smooth, the local neighborhood of any point is able to be well fitted with a plane. Therefore, we use the local surface fitting method for the normal vector estimation of the SfM dense point cloud. Then, 3D pose-varied semantic features such as angle, horizontal orientation and curvature of the building façade are obtained from the computed normal vectors. Figure 6(b)(c)(d) shows the visualized images of 3D information such as angle, horizontal orientation and curvature of the building façade, respectively. It can be found that the images displayed according to the 3D features ignore the interference of the textural information on the surface of the building and well distinguish the different facades of the building.

Then, selecting a building image with good view, as shown in Figure 8(a), and the interior and exterior orientation parameters of the image are derived to construct a mapping relationship between the 2D and 3D features of the building. Through the reprojection algorithm mentioned in Section 3.1, the geometric features of the building are reprojected onto the selected image to construct a building image with integrated 2D and 3D features, as shown in Figure 8(b). Comparing the two figures, we can find that the reprojection algorithm can correspond two images with 2D features and 3D features, which provides a good basis for the subsequent feature integration and superpixel segmentation.

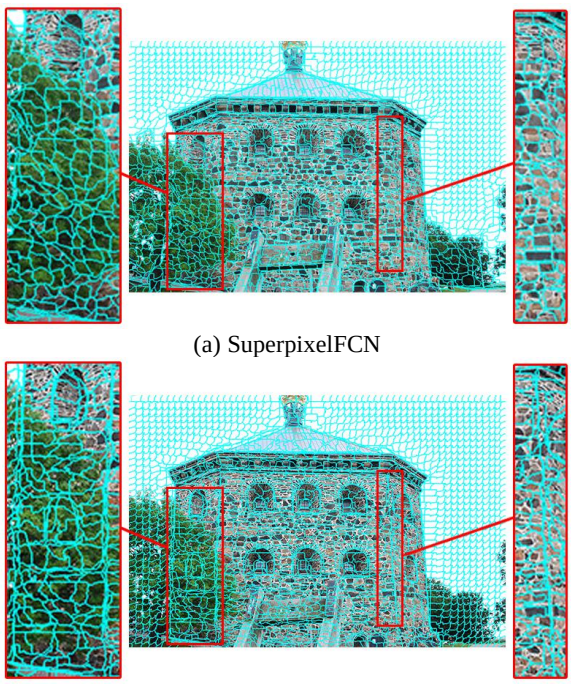
(a) digital image

(b) image with 3D pose-varied semantics by reprojection

Figure 8. Image with multi-modal semantic features.

In our experiments, we applied the training method of SuperpixelFCN (Yang et al., 2020). The network was trained

in PyTorch with k=0.003 using the BSDS500 dataset, which contains 1633 training samples and 1063 test samples. We partitioned the training set into 208 × 208 images as input and set the training units to 16 × 16, with the corresponding number of superpixel outputs varying with the test image size. Based on the FCN superpixel segmentation algorithm mentioned in Section 3.2, the image integrated with 2D texture semantics and 3D pose -varied semantics is segmented into a 16 × 16 regular grid cells as the input to FCN to obtain a predicted soft association map based on multi-modal features. In this experiment, the weight factors ω_1 and ω_2 in Eq.4 are set to 0.2 and 0.8, respectively. Considering that all the images used in the experiment have the problem that the 2D features are more seriously disturbed by the environment, the weight factors of the 3D soft associtation map are considered to be increased. When facing different kinds of building facades, the most suitable integration results can be calculated by adjusting the weight factors. With the highest probability of surrounding neighborhood, we can achieve the segmentation results. Figure 9 compares the results of the traditional image-only based FCN superpixel segmentation method with those of the proposed integrated 2D texture semantics and 3D pose-varied semantics. It can be seen from the images that since the building is a stone façade, in the traditional FCN superpixel segmentation results, the stones of varing colors and shapes are mainly segmented, and the distinction for the façade of the building with different orientations is not obvious. At the same time, the left side of the building is obscured by greenery, and the building outline cannot be recognized. While the method proposed in this paper overcomes the shortcomings of the traditional method, ignores the influence of varing colors of stone textures, identifies the building facades of different orientations and shapes according to the 3D features, and can accurately identify the edges of the facades. Moreover, by cropping the point cloud in the pre-processing process, the influence of foreground occlusion is removed, and the natural building outline is recognized.



(a) SuperpixelFCN

(b) Ours

Figure 9. Comparison of the results for the proposed method and the traditional FCN superpixel segmentation method.

To better demonstrate the reliability of our proposed method for building facade segmentation, we conducted experiments using more buildings with different color textures and geometries and compared them with the traditional FCN superpixel segmentation method, as shown in Figure 10 and Figure 11. We show the whole image and detailed examples of the original image of the building, the segmentation result of the traditional FCN superpixel method, and the segmentation result of the proposed method, respectively. The red dash lines in the figures mark the different facades of the building.

From the building image in Figure 10, it can be seen that the building façade contains rich texture information, and the texture pattern has high contrast and obvious edges. Moreover, the differences between the corrensponding parts of different elevations on the surface of this building are slight, and there are many tiny undulations on the surface. The traditional FCN superpixel segmentation method is greatly influenced by the surface texture pattern, and almost relies on the pattern color to complete the segmentation, which cannot separate the different facades of the building. Our method, on the other hand, uses 3D pose-varied semantic features of the building as the primary segmentation basis, which excludes the interference of texture patterns, and can accurately identify even the tiny structures on the building surface.

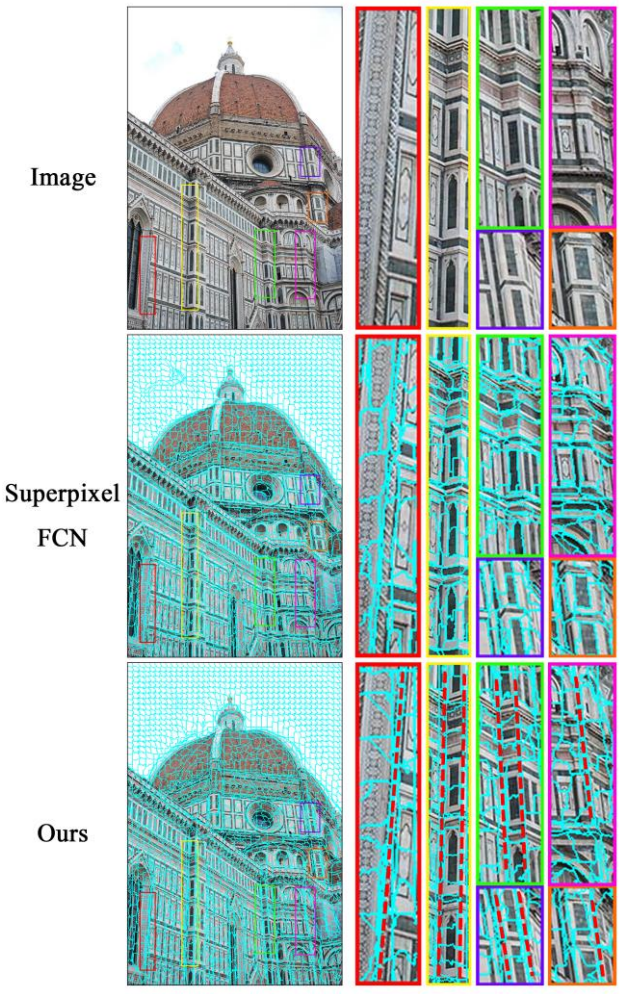


Figure 10. Comparison of the results for superpixel segmentation method on building with flexible texture.

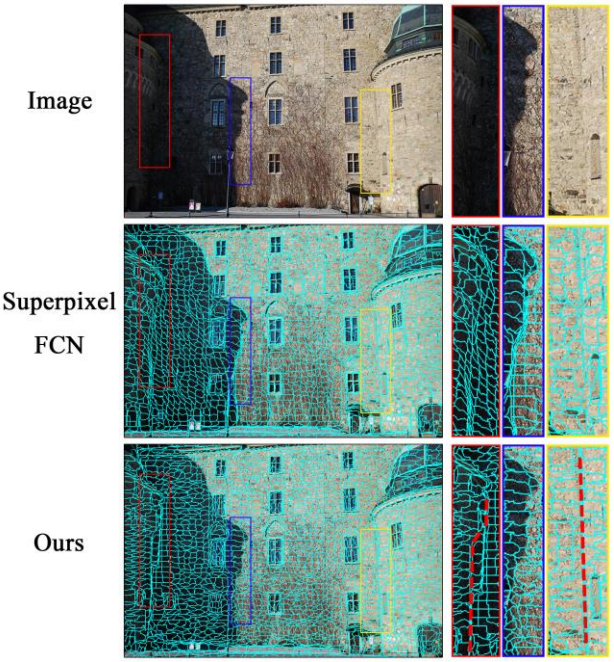
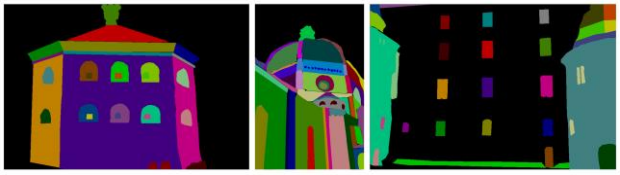


Figure 11. Comparison of the results for superpixel segmentation method on building with shadow coverage.

Figure 11 contains a total of three different buildings that are staggered, and they share similar surface color and texture information. From the original image, it can be seen that the building image is greatly affected by light during the acquisition process. There is a large area of shadow coverage on the building surface. The traditional FCN superpixel method cannot recognize the building structure in the shadowed part. It treats the whole shadowed area as a whole fa çade, with obvious segmentation effect on the shadowed edge part, while the important building features are lost. Moreover, the segmentation results are ineffective as the two buildings with similar color and texture cannot be distinguished in the part of the right side of the image where the two buildings are interspersed. Our proposed method, by extracting the geometric features of different building surfaces, effectively distinguishes the buildings by different normal vector orientations ignoring the influence of shadow occlusion and interlaced distribution of buildings. The segmented building edges are clearly outlined.

To verify the reliability of Pose-FCN by quantitative metrics, this paper evaluates the method by standard superpixel segmentation result evaluation metrics (Stutz et al., 2018) such as ASA (Achievable Segmentation Accuracy), BR (Boundary Recall), and UE (Undersegmentation Error) with the groundtruth data of each image, and compares it with the traditional FCN superpixel segmentation method, as shown in Table 1. In the groundtruth data, different facades of the building are labeled with different properties, while categories such as trees and shadow are not labeled, as show in Figure 12.



(a) Building 1 (b) Building 2 (c) Building 3

Figure 12.The groundtruth data of the buildings.

Images	Method	ASA	BR	UE
Building 1	SuperpixelFCN	0.9671	0.9315	0.0329
	Ours	0.9828	0.9715	0.0173
Building 2	SuperpixelFCN	0.9326	0.9223	0.0674
	Ours	0.9561	0.9704	0.0439
Building 3	SuperpixelFCN	0.9616	0.9764	0.0384
	Ours	0.9698	0.9897	0.0302

Table 1. Comparison of superpixel segmentation results on evaluation metrics.

The results show that the novel method is optimized in all metrics compared to the superpixel segmentation method based on 2D semantics only. The new method achieves better accuracy and higher boundary recall. In particular, the integration of multi-modal semantic features effectively eliminates the interference of texture information. It reduces the false segmentation in the case of complex textures and obvious geometric features of building facades. The novel method is also able to segment different building facades by extracting geometric features when the color textures of the edges of different building facades are similar, and it can also effectively exclude the effects of foreground occlusion, irregular textures and shadows on building facade extraction.

5. CONCLUSION

In this paper, a novel superpixel segmentation method for building facades based on integrated multi-modal semantics was proposed. Firstly, a dense point cloud of buildings is generated by multi-view images based on SfM. Meanwhile, 3D pose-varied semantic features of buildings including angle, horizontal orientation are calculated based on the dense point cloud. Then, the geometric features of the building are reprojected onto the 2D image with the corresponding camera interior and exterior orientation parameters, and the mapping relationship between the 2D and 3D features is constructed. Therefore, the integrated multi-modal features of building are obtained. Following this, a FCN soft association map based on integrated semantics is computed to achieve superpixel segmentation results of building facades. The experimental results demonstrate that the method outperforms the traditional FCN superpixel segmentation method when targeting different kinds of buildings. The method has a reasonable performance in the integrity and accuracy of building facade segmentation, especially when the buildings are affected by irregular patterns, shadow coverings, and foreground occlusions.

Furthermore, the superpixel obtained by the proposed method contains texture and geometric semantics, so the geometric and texture information of the building fa çade can be accurately extracted using multi-level attribute clustering in the next clustering step. For example, geometric planes such as walls and roofs with dilapidated surfaces, irregular patterns and

shadow coverage can be extracted mainly using angles and horizontal orientation accurately; objects such as columns and round roofs can be clustered by adding curvature features; building facade information with a single texture can be clustered mainly using color and texture. Moreover, the method is computationally more efficient as compared to the segmentation of massive point clouds, which are computed on a 2D space.

In summary, the method proposed in this study provides a new idea to integrate the 2D and 3D semantic features of the target, and its further application in other fields will be studied in future work.

ACKNOWLEDGEMENTS

This research was funded by the National Natural Science Foundation of China (Grants Nos. 41901401 and 42101070), the Natural Science Foundation of Jiangsu Province (Grants No. BK20190743), and the China Postdoctoral Science Foundation (Grants No. 2021M691653)

REFERENCES

Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., Süsstrunk, S., 2012. SLIC superpixels compared to state-of-the-art superpixel methods. *IEEE transactions on pattern analysis and machine intelligence*, 34(11), 2274-2282.

Achanta, R., Susstrunk, S., 2017. Superpixels and polygons using simple non-iterative clustering. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4651-4660.

Choy, C. B., Xu, D., Gwak, J., Chen, K., Savarese, S., 2016. 3d-r2n2: A unified approach for single and multi-view 3d object reconstruction. *In European conference on computer vision* , pp. 628-644. Springer, Cham.

Eigen, D., Puhrsch, C., Fergus, R., 2014. Depth map prediction from a single image using a multi-scale deep network. *Advances in neural information processing systems*, 27.

Enqvist, O., Kahl, F., Olsson, C., 2012. Non-sequential structure from motion. *IEEE International Conference on Computer Vision Workshops*. IEEE.

Furukawa, Y., Ponce, J., 2009. Accurate, dense, and robust multiview stereopsis. *IEEE transactions on pattern analysis and machine intelligence*, 32(8), 1362-1376.

Giraud, R., Ta, V. T., Papadakis, N., 2018. Robust superpixels using color and contour features along linear path. *Computer Vision and Image Understanding*, 170, 1-13.

Hongshan, Y., Zhengeng, Y., Lei, T., Yaonan, W., Wei. S., Mingui, S., Yandong, T., 2018. Methods and datasets on semantic segmentation: A review. *Neurocomputing*, 304, 82-103.

Hu, Y., Li, Y., Song, R., Rao, P., Wang, Y., 2018. Minimum barrier superpixel segmentation. *Image and Vision Computing*, 70, 1-10.

Li, Y., Bu, R., Sun, M., Wu, W., Di, X., Chen, B., 2018. Pointcnn: Convolution on x-transformed points. *Advances in neural information processing systems*, 31, 820-830.

Lin Y., Yan Z., Huang H., Du D., Liu L., Cui S., Han X, 2020. FPConv: Learning Local Flattening for Point Convolution. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*.

Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation. *In Proceedings of the IEEE conference on computer vision and pattern recognition.* 3431-3440.

Martinovic, A., Knopp, J., Riemenschneider, H., Van Gool, L., 2015. 3D all the way: Semantic segmentation of urban scenes from start to end in 3d. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4456-4465.

Qi, C. R., Su, H., Mo, K., Guibas, L. J., 2017. PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE.

Ren, X., Malik, J., 2003. Learning a classification model for segmentation. *In Computer Vision, IEEE International Conference on*, Vol. 2, pp. 10-10. IEEE Computer Society.

Riemenschneider, H., Bódis-Szomorú, A., Weissenberg, J., Van Gool, L., 2014. Learning where to classify in multi-view semantic segmentation. *European Conference on Computer Vision*, 516-532. Springer, Cham.

Stutz, D., Hermans, A., Leibe, B., 2018. Superpixels: An evaluation of the state-of-the-art. *Computer Vision and Image Understanding, 166*, 1-27.

Tyleek, R., Radi Sára, 2013. Spatial Pattern Templates for Recognition of Objects with Regular Structure. *Gcpr*.

Wang, N., Zhang, Y., Li, Z., Fu, Y., Liu, W., Jiang, Y. G., 2018. Pixel2mesh: Generating 3d mesh models from single rgb images. *In Proceedings of the European conference on computer vision (ECCV)*, pp. 52-67.

Wu, C., Agarwal, S., Curless, B., Seitz, S. M., 2011. Multicore bundle adjustment. *In CVPR 2011*, pp. 3057-3064. IEEE.

Wu, C., 2013. Towards linear-time incremental structure from motion. *In 2013 International Conference on 3D Vision-3DV 2013*, pp. 127-134. IEEE.

Xiao, X., Zhou, Y., Gong, Y. J., 2018. Content-adaptive superpixel segmentation. *IEEE Transactions on Image Processing*, 27(6), 2883-2896.

Yang, F., Sun, Q., Jin, H., Zhou, Z., 2020. Superpixel segmentation with fully convolutional networks. *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13964-13973.