Facade Segmentation from Oblique UAV Imagery

Xiangyu Zhuo[1], Milena Mönks[2], Thomas Esch[2], and Peter Reinartz[1]

[1]Remote Sensing Technology Institute, German Aerospace Center, Oberpfaffenhofen, Germany
[2]German Remote Sensing Data Center, German Aerospace Center, Oberpfaffenhofen, Germany
{xiangyu.zhuo, milena.moenks, thomas.esch, peter.reinartz}@dlr.de

***Abstract*—**

Building semantic segmentation is a crucial task for building information modeling (BIM). Current research generally exploits terrestrial image data, which provides only limited view of a building. By contrast, oblique imagery acquired by unmanned aerial vehicle (UAV) can provide richer information of both the building and its surroundings at a larger scale. In this paper, we present a novel pipeline for building semantic segmentation from oblique UAV images using a fully convolutional neural network (FCN). To cope with the lack of UAV image annotations at facade level, we leverage existing ground-view facades databases to simulate various aerial-view images based on estimated homography, yielding abundant synthetic aerial image annotations as training data. The FCN is trained end-to-end and tested on full-tile UAV images. Experiments demonstrate that the incorporation of simulated views can significantly boost the prediction accuracy of the network on UAV images and achieve reasonable segmentation performance.

***Index Terms*—Building semantic segmentation, facade parsing, convolutional neural network, FCN**

I. Introduction

Detailed Building Information Modeling (BIM) is demanded in many civil engineering applications such as urban planning and scene understanding. A crucial step for BIM is the semantic parsing of buildings, which aims at the pixel-level interpretation of various structural components, such as wall, door and window. Manual interpretation of building components is quite labor-intensive and cost-expensive, therefore automated building semantic segmentation is of great importance in urban-scale tasks. Despite having been actively studied, this problem is still not adequately solved. The main reason is that many previous approaches define handcrafted grammars based on the geometry and appearance characteristics of facade components, which cannot robustly cope with the significant variation of buildings. Besides, external factors such as occlusions, illumination and perspective differences also pose difficulty to robust segmentation.

Recent advances in deep learning techniques have opened up new possibilities in various computer vision tasks such as object detection and semantic segmentation. A number of deep neural network-based approaches have been proposed for facade parsing tasks and demonstrated superior performance compared to traditional methods. For the sake of generalization, deep learning techniques usually require enormous training data. Nevertheless, the current facade parsing benchmarks such as eTRIMS [1], ECP [2] and CMP [3] are generally collected from the ground and thus have limited views of buildings. In contrast, oblique UAV images are able to offer information of the building and its surroundings at a larger scale, however, there are rarely UAV imagery databases annotated at facade level.

To tackle this problem, we propose in this paper a novel pipeline for semantic facade parsing from UAV oblique imagery, as illustrated in Figure 1. Instead of manually annotating UAV imagery, we exploit existing ground-view facade parsing benchmarks to generate synthetic aerial-view images as training data. Subsequently, a Fully Convolutional Network (FCN) [4] is trained end-to-end. In order to improve label consistency and accuracy at class boundaries, we plug in a Conditional Random Field (CRF) represented as Recurrent Neural Network (CRFasRNN) [5], [6] at the end of the FCN, which combines the strengths of both the CNN and CRF based graphical model in one unified framework. In the end, the trained network is used to predict the UAV imagery at pixel level.

The contributions of this paper lie in two aspects:

1) we proposed simulating aerial views from terrestrial views based on estimated homography in order to generate training data.
2) we firstly investigated into building segmentation on full-tile UAV imagery.

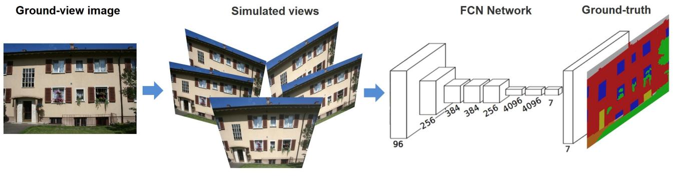


Fig. 1: Pipeline of the proposed method

II. Related Work

In the past decades, an extensive body of research works have been done towards the goal of facade segmentation and can be broadly classified into two categories: *model-based* methods [7]–[10] and *model-free* methods [11]–[13]. With the dramatic development of deep learning techniques, deep neural network-based segmentation methods have gained increasingly

978-1-7281-0009-8/19/$31.00 ©2019 IEEE

attention in facade parsing applications and demonstrated dominant performance compared with traditional approaches. A convolutional neural network based on ConvNet was introduced in [14] for facade parsing (including facade, door, window and clutter) from terrestrial images, the method achieved an F1 score of 82% on the eTRIMS dataset. A FCN was used in [15] for semantic segmentation of facades and vertical edges of buildings, where a 3D tracking system was proposed to ease the annotation of training data. A deep convolutional neural network named "DeepFacade" with a symmetric regularization term was proposed in [16], where a novel loss function was proposed for training. This method has outperformed previous state-of-the-art methods on ECO and eTRIMS datasets. There are limited previous works in the context of facade parsing from aerial views. A pipeline of facade parsing from oblique aerial images was presented in [17], where 2D and 3D features were both extracted to train a random forest (RF) model for pixel-level facade segmentation, then the result was refined by a fully connected conditional random field (CRF). Experiments showed that the incorporation of 3D features can significantly improve segmentation accuracy.

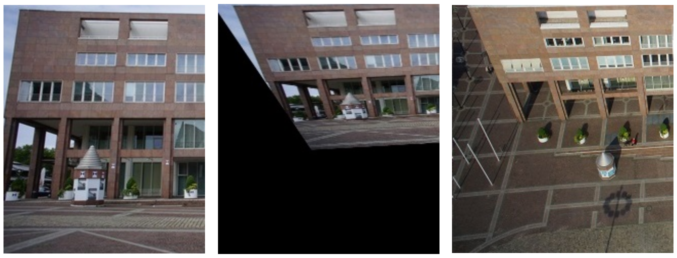
The aforementioned works investigate into facade parsing of homologous imagery, i.e. the images for training and testing are acquired by the same modalities and therefore have similar scale and appearance. In contrast, our paper propose to use existing terrestrial image annotation databases to train CNNs and then apply the networks to parse UAV images. This can dramatically save manual labors for creating training data.

III. Methodology

The performance of deep neural networks usually relies on sufficient training data. Due to the lack of annotated aerial facade dataset, we propose in this paper to exploit the existing ground-view facade datasets to generate training data.

Though it has been demonstrated in [14] that large datasets are not necessary for training the network when transfer learning and data augmentation are employed, our segmentation task still faces the challenge of different appearance of terrestrial imagery and UAV imagery, especially the substantial difference in viewing direction and scale. To tackle this problem, we eliminate the scale difference by image downsampling and simulate aerial views from terrestrial images by perspective transformation.

In view of the fact that building facade is generally planar, the perspective deformation of the facade plane under a camera motion in close-range can be described by a planar homographic transform. Although it is not always possible to retrieve parameters for camera displacement, the deformation can be approximated by estimating the homography matrix. As illustrated in Figure 2, (a) shows a facade from ground-view, (b) shows the simulated aerial view that is computed with given corner points and (c) depicts the real aerial view of the facade in a UAV image. Visually, the simulated aerial view of the facade has high similarity with the real facade in the UAV image, despite the minor difference of illumination and misplacement of some non-coplanar objects.



(a) Terrestrial image (b) Simulated image (c) UAV image

Fig. 2: Comparison between simulated image and real UAV image.

Motivated by this observation, we propose in this paper to simulate aerial images from terrestrial images based on assumptions of the homography. More specifically, the homographic matrix is estimated from a set of arbitrary coplanar points, and the variation of homographies corresponds to different perspectives. Figure 3 illustrates several views simulated from a terrestrial image, including left-view, top-view and right-view. The simulated views can not only enrich the training data, but also narrow the gap between terrestrial image and UAV image.

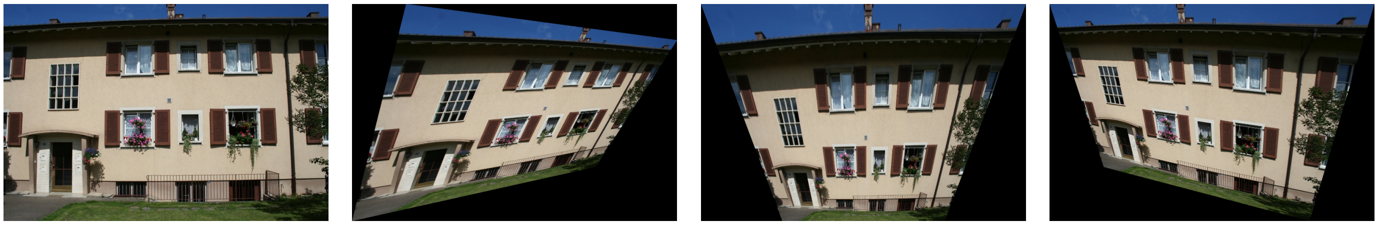
Subsequently, the simulated views are further augmented via rotation and cropping, resulting in around thousands of patches of 300 × 300 pixels. These patches are used as input data for training a fully convolutional neural network [4]. For the sake of efficiency, we fine-tune the pre-trained weights from existing networks on our own dataset. More specifically, we truncate the last layer of the pre-trained network and replace it with a new softmax layer of 5 categories. In view of the fact that the CNN usually yields inaccurate class boundaries due to its large receptive field, we plug in a Conditional Random Field (CRF) represented as Recurrent Neural Network (CR-FasRNN) [5], [6] at the end of the FCN, which can provide more visually appealing results with sharper boundaries.

IV. Data Description

Terrestrial dataset

Instead of manually annotating UAV images for training, we exploit existing facades databases, including the "LabelMe-Facade Image Dataset" [18], [19] and the "eTraining for Interpreting Images of Man-Made Scenes (eTRIMS) Image Database" [1]. The LabelMeFacade Image Dataset consists of 975 street-view images with 9 annotated object classes, however, some of the annotations have low accuracy at class *window* or *door*. Thus, we select a subset of 50 accurate annotations for training. The eTRIMS database has two variants. In this experiment, we use the 8-Class eTRIMS Dataset, including 60 images with 8 annotated object classes showing front-view of buildings.

Images of both databases are taken from ground view and without rectification, presenting buildings of various styles from different perspectives. Although the original image annotations involve various object classes, some classes are



(a) Original image (b) Simulated left-view (c) Simulated top-view (d) Simulated right-view

Fig. 3: Multi-view simulation from terrestrial images.

invisible (e.g. *sky*) or in the minority (e.g. *pavement*) in UAV images. Therefore we merge them into 5 classes, namely *building* (including roof and wall), *window, door, vegetation and ground*. All the rest classes (e.g. *car* and *sky*) are masked during training.

UAV dataset

The UAV images for testing are selected from existing UAV datasets, namely the ISPRS "Zeche Zollern" and "Stadthaus" datasets [20], showing buildings of both modern-style and traditional-style. Besides, we also collected facades images of a detached house in Morschenich, Germany, using a rotary-wing UAV. Different from the ISPRS datasets where the windows are featured by transparent glass, some windows of this house are covered by the blinds and appear to be opaque. In order to evaluate the segmentation accuracy, we labeled the test images by hand as ground-truth data.

V. Experiments

For each input terrestrial image, we simulated three different aerial views from it. Subsequently, the simulated images are further augmented via rotation and cropping, resulting in 3924 patches of 300 × 300 pixels. In order to validate the effect of view simulation, we also set up a control group by augmenting the original images into 3924 patches via only rotating and cropping. The network was fine-tuned in the Caffe [21] framework using the pretrained weights from FCN-8s [4]. The whole training process took around 7 hours and the trained model was applied to predict the UAV images. Since the UAV images for testing have much higher spatial resolution than the terrestrial images for training, we downsampled the UAV images to the same scale. In addition, we manually annotated the test images for accuracy evaluation.

Figure 4 illustrates some of the best segmentation results. It can be seen that the trained network has achieved reasonable accuracy at most classes except for *door*, there can be two possible reasons for that: first, there are only a few samples of doors in the training data so that the network can not well learn the features of doors; second, the doors in UAV images are often severely deformed or hardly visible, which is difficult to recognize even for human eyes.

The impact of view simulation is demonstrated in Table I. Where, the row *Original* lists the pixel accuracy (in percentage) of segmentation using original terrestrial images as training data, while the row *Simulated* gives the pixel accuracy of segmentation using simulated views as training data. It can be seen that the view simulation can significantly improve the segmentation accuracy for building facades.

TABLE I: Pixel accuracy (%) for segmentation results

Method	Building	Window	Door	Veg	Ground
Original	84.92	80.39	37.70	91.83	83.11
Simulated	87.66	84.01	41.87	92.60	84.34

VI. Conclusion

In this paper we presented a novel pipeline for semantic building segmentation from UAV images based on a CNN. By simulating various aerial-view images from terrestrial images based on estimated homography, we achieved abundant training data without manually annotating UAV images. Experiments demonstrated that the trained network can achieve reasonable segmentation performance for UAV images. Besides, the simulated views can also boost the segmentation accuracy significantly.

VII. Acknowledgement

The authors would like to acknowledge the provision of the datasets by ISPRS and EuroSDR, released in conjuction with the ISPRS scientific initiative 2014 and 2015, lead by ISPRS ICWG I/Vb. The authors would also like to thank the European Space Agency (ESA) for funding the project Artificial Intelligence for Smart Cities (ESRIN/Contract No. 4000126100/19/I-EF) that supported parts of the research conducted by DFD-LAX.

References

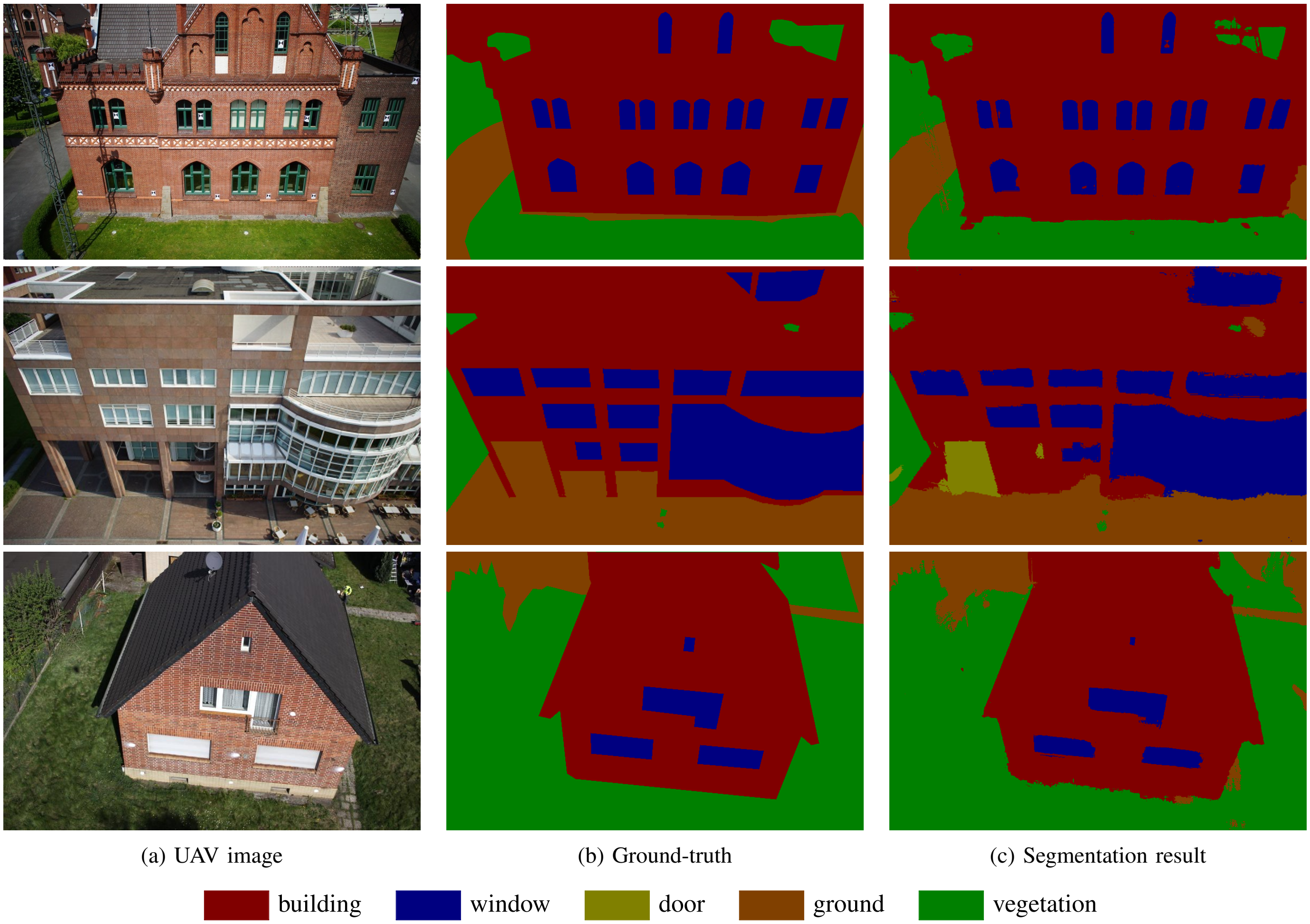
[1] F. Korc and W. Förstner, "eTRIMS Image Database for interpreting images of man-made scenes," *Dept. of Photogrammetry, University of Bonn, Tech. Rep. TR-IGG-P-2009-01*, 2009.

[2] O. Teboul, L. Simon, P. Koutsourakis, and N. Paragios, "Segmentation of building facades using procedural shape priors," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, June 2010, pp. 3105–3112.

[3] R. Tyleček and R. Šára, "Spatial pattern templates for recognition of objects with regular structure," in *German Conference on Pattern Recognition*. Springer, 2013, pp. 364–374.

[4] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[5] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, and P. H. S. Torr, "Conditional random fields as recurrent neural networks," in *International Conference on Computer Vision (ICCV)*, 2015.


(a) UAV image

(b) Ground-truth

(c) Segmentation result

Fig. 4: Segmentation result

[6] A. Arnab, S. Jayasumana, S. Zheng, and P. H. S. Torr, "Higher order conditional random fields in deep neural networks," in *European Conference on Computer Vision (ECCV)*, 2016.

[7] M. Kozinski, R. Gadde, S. Zagoruyko, G. Obozinski, and R. Marlet, "A MRF shape prior for facade parsing with occlusions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 2820–2828.

[8] A. Martinovic and L. Van Gool, "Bayesian grammar learning for inverse procedural modeling," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 201–208.

[9] O. Teboul, I. Kokkinos, L. Simon, P. Koutsourakis, and N. Paragios, "Shape grammar parsing via reinforcement learning," in *CVPR 2011*, June 2011, pp. 2273–2280.

[10] H. Riemenschneider, U. Krispel, W. Thaller, M. Donoser, S. Havemann, D. Fellner, and H. Bischof, "Irregular lattices for complex shape grammar facade parsing," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 1640–1647.

[11] W. Li and M. Y. Yang, "Efficient semantic segmentation of man-made scenes using fully-connected conditional random field," *The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 41, p. 633, 2016.

[12] M. Mathias, A. Martinović, and L. Van Gool, "ATLAS: A three-layered approach to facade parsing," *International Journal of Computer Vision*, vol. 118, no. 1, pp. 22–48, 2016.

[13] A. Cohen, A. G. Schwing, and M. Pollefeys, "Efficient structured parsing of facades using dynamic programming," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 3206–3213.

[14] M. Schmitz and H. Mayer, "A convolutional network for semantic facade segmentation and interpretation." *International Archives of the Photogrammetry, Remote Sensing & Spatial Information Sciences*, vol. 41, 2016.

[15] A. Armagan, M. Hirzer, and V. Lepetit, "Semantic segmentation for 3d localization in urban environments," in *Urban Remote Sensing Event (JURSE), 2017 Joint*. IEEE, 2017, pp. 1–4.

[16] H. Liu, J. Zhang, J. Zhu, and S. C. H. Hoi, "DeepFacade: A deep learning approach to facade parsing," in *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, 2017, pp. 2301–2307. [Online]. Available: https://doi.org/10.24963/ijcai.2017/320

[17] Y. Lin, F. Nex, and M. Yang, "Semantic building façade segmentation from airborne oblique images." *ISPRS Annals of Photogrammetry, Remote Sensing & Spatial Information Sciences*, vol. 4, no. 2, 2018.

[18] B. Fröhlich, E. Rodner, and J. Denzler, "A fast approach for pixelwise labeling of facade images," in *Proceedings of the International Conference on Pattern Recognition (ICPR 2010)*, 2010.

[19] C.-A. Brust, S. Sickert, M. Simon, E. Rodner, and J. Denzler, "Efficient convolutional patch networks for scene understanding," in *CVPR Workshop on Scene Understanding (CVPR-WS)*, 2015.

[20] F. Nex, F. Remondino, M. Gerke, H.-J. Przybilla, M. Bäumker, and A. Zurhorst, "ISPRS benchmark for multi-platform photogrammetry." *ISPRS Annals of Photogrammetry, Remote Sensing & Spatial Information Sciences*, vol. 2, 2015.

[21] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," in *Proceedings of the 22nd ACM international conference on Multimedia*. ACM, 2014, pp. 675–678.