DOI: 10.3217/978-3-85125-663-5-42

Combining Deep Learning and Variational Level Sets for Segmentation of Buildings

Muntaha Sakeena[1] and Matthias Zeppelzauer[2]

***Abstract*— The larger context behind this work is the automated visual assessment of building characteristics (e.g. building age and condition) for the estimation of real estate prices from outdoor pictures. A basic requirement to this end is the automated segmentation of buildings from photos, which is the focus of this work. We propose a combined deep-learned and variational segmentation method for the extraction of the building area from real estate images and our experimental results with dice similarity of approx. 92% demonstrate its capabilities on a novel dataset for building segmentation.**

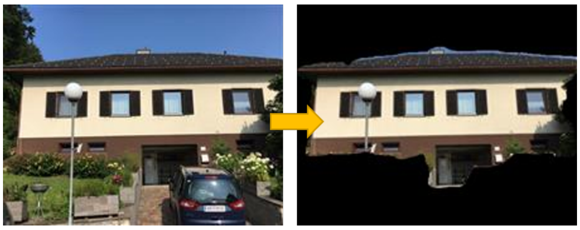

Fig. 1. The task of building segmentation

I. INTRODUCTION

Deep learning-based approaches proved to be effective in many computer vision applications including the image-based appraisal of real estate [6]. The estimation of certain building characteristics, however, requires a different perspective, i.e. the street view perspective. The segmentation of buildings in street view perspective has hardly been performed and is more complex than aerial image segmentation because perspective and appearance varies much stronger. Furthermore, occluding foreground objects, such as trees and plants impede segmentation.

For the robust assessment of building parameters there is a need for a precise segmentation of buildings to remove unnecessary and potentially misleading information (see Fig. 1). Today, segmentation is used for the tasks like autonomous driving [1] and urban scene understanding [3]. In this work, we present a robust method for the segmentation of unconstrained building views to lay the foundations for extended real estate image analysis (REIA) which is an emerging and challenging computer vision problem [4], [7]. To mitigate shortcomings of existing segmentation networks, we propose a combination with Variational Level sets (VLS) [5] to improve the segmentation quality. We evaluate the approach on a set of pixel-wise annotated building images from real estate websites and show (i) that transfer learning (fine tuning) is essential for obtaining satisfactory results and (ii) that VLS improve the boundary, which seems to be difficult to learn for the network, especially when training data is limited.

II. APPROACH

A major challenge is that buildings can be captured from different perspectives and with different scales. To capture this variety, we propose a two-step approach, illustrated in Fig. 2. In the first step, we employ SegNet [1], which is a powerful and flexible encoder-decoder network for semantic segmentation to initially detect the building in the image. Each encoder stores the feature indices extracted from each layer and uses them at the decoder side. The SegNet input is usually an RGB image with a pixel-wise label image as ground truth. Due to the pixel-wise nature of SegNet and in the presence of a limited amount of training data, there is a risk of obtaining noisy, irregular or over-smooth boundaries for the segmented objects as well as outliers (small false positive areas). Therefore, we add a second processing step to refine the initial output mask. To get more exact and accurate boundaries, we integrate Variational Level Sets (VLS) into the approach which are well-suited for the detection of smooth and regular boundaries around objects having different shapes and topology [5].

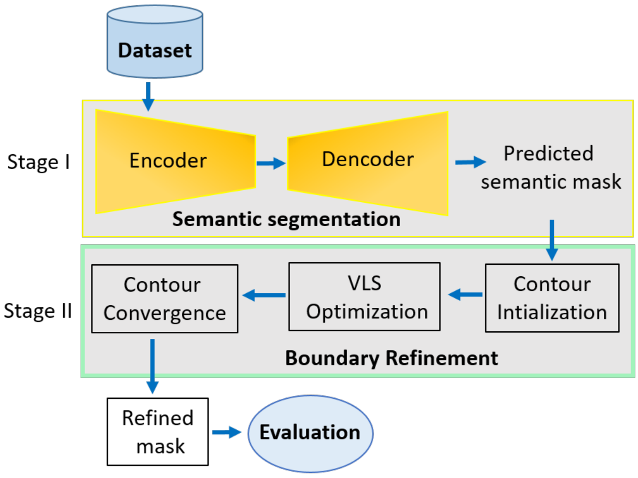

Fig. 2. Proposed overall approach

The idea of VLS is based on the minimization of energy E which integrates the geometric properties of the level set with image features. As input, we consider an image I composed of two regions: "building" and "no-building", which are

*This work was supported by FFG projects 855784, 856333 and 865973

Sankt Pölten University of Applied Sciences, Austria
[1] muntaha.sakeena@fhstp.ac.at
[2] matthias.zeppelzauer@fhstp.ac.at

mutually exclusive and separated by a contour ϕ. Gradient descent is then applied for optimizing and converging ϕ around the boundaries. Contour convergence VLS further includes a regularization term $R(\phi) = \frac{1}{2}\int_I (\nabla\phi - 1)^2 dI$ to limit the contour to a signed distance function. The curve evolves until separate regions can be defined by minimizing E as $(\partial\phi)/(\partial t) = -(\partial E)/(\partial\phi)$ which represents the gradient flow that minimizes the energy E.

III. Experiments & Results

Dataset: We collected the dataset from real estate websites to train and evaluate our approach. The dataset is composed of 975 images in which each building was manually labeled at the pixel level. The dataset is divided into three subsets: training (50%), validation (20%), and testing (30%).

Performance Measures: We employ the Dice similarity coefficient (DSC) to assess the performance of our approach which is specifically designed to evaluate segmentation tasks: $DSC = 2(|S \cap G|)/(|S| + |G|)$, where S is the segmented area in the result image and G is the ground truth mask.

Experimental setup: Initially, we employ SegNet pre-trained on the CamVid dataset [2] which contains a building class (aside from other classes like road and car) to evaluate its generalization ability. In a next step, we fine-tune SegNet for 50 epochs for domain adaption. Input images were resized proportionally to the size of the input layer (360×480). We trained SegNet with a learning rate of $1e^{-3}$, a momentum of 0.9 and a batch size of 4. For VLS, we directly used the output of SegNet to initialize ϕ. This reduces the computation time for contour initialization significantly and mitigates the instabilities resulting from the sensitivity of VLS to initial contour selection. We keep the contour evolving for 110 iterations to assure convergence.

Results: We evaluate our approach on the independent test set. For a fair comparison, we evaluate (i) SegNet pre-trained on the CamVid [2] dataset, (ii) SegNet fine-tuned on our training set and (iii) the latter net with boundary refinement by VLS, see Table I for the results. Our experiments show that the pre-trained SegNet generalized poorly to buildings of different types and from different perspectives than seen before (in the CamVid dataset), see column 2 of Fig. 3. Fine-tuning the network strongly boosts performance from 60% to approx. 89%, showing that domain adaption is of high importance. Still, we observe that the output contours are noisy. By applying VLS refinement to the output of the fine-tuned network, the boundaries further improve which is also reflected in the segmentation results in Table I and column 3 of Fig. 3. Further results are depicted in Fig. 4.

TABLE I

Achieved segmentation performance

Methods	DSC %
Random Baseline	51.17
Pre-trained SegNet	60.23
Fine-tuned SegNet	89.27
Boundary refinement by VLS	**91.70**

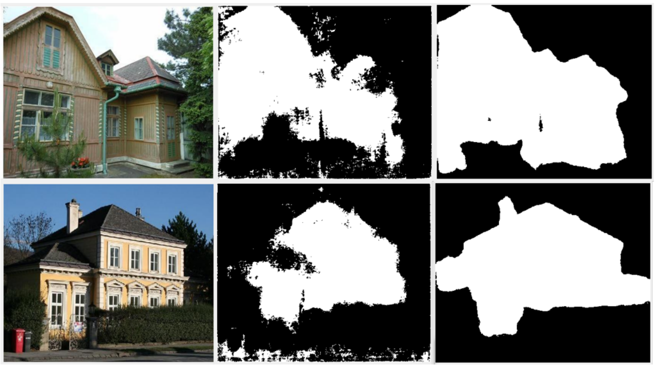

Fig. 3. Test images with predicted contours/masks: Original image (first column), Pre-trained SegNet (second column), predicted mask by proposed method with boundary refinement (third column)

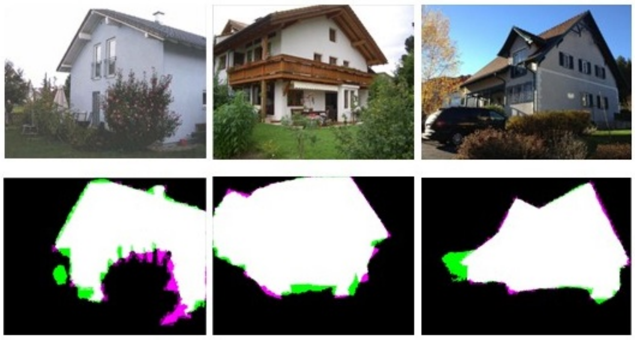

Fig. 4. Test results of the proposed method: original image (first row), predicted masks (second row): true positives (white), false positives (green), true negatives (black), false negatives (pink).

IV. Conclusions

We have presented an approach for the segmentation of buildings by combining a semantic segmentation network with VLS. Results are promising and sufficiently accurate for future visual extraction of higher-level building parameters for real estate appraisal. In future, we plan to compare our method with Conditional random field (CRF) methods.

References

[1] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *CoRR*, vol. abs/1511.00561, 2015.

[2] G. J. Brostow, J. Shotton, J. Fauqueur, and R. Cipolla, "Segmentation and recognition using structure from motion point clouds," in *ECCV (1)*, 2008, pp. 44–57.

[3] 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.

[4] D. Koch, M. Despotovic, M. Sakeena, M. Döller, and M. Zeppelzauer, "Visual estimation of building condition with patch-level convnets," in *Proc. ACM Wsp. on Multimedia for Real Estate Tech*, 2018, pp. 12–17.

[5] C. Li, C. Xu, C. Gui, and M. D. Fox, "Distance regularized level set evolution and its application to image segmentation," *Image Processing, IEEE Transactions on*, vol. 19, no. 12, pp. 3243–3254, 2010.

[6] Q. You, R. Pang, and J. Luo, "Image based appraisal of real estate properties," *CoRR*, vol. abs/1611.09180, 2016.

[7] M. Zeppelzauer, M. Despotovic, M. Sakeena, D. Koch, and M. Döller, "Automatic prediction of building age from photographs," in *Proc. of ACM ICMR*, 2018, pp. 126–134.