Direct Window-to-Wall Ratio Prediction Using Deep Learning Approaches

1st Xiangyu Zhuo
Remote Sensing Technology Institute
German Aerospace Center
Weßling, Germany
xiangyu.zhuo@dlr.de

2nd Jiaojiao Tian
Remote Sensing Technology Institute
German Aerospace Center
Weßling, Germany
jiaojiao.tian@dlr.de

3rd Karl-Heinz Häfele
Institute for Automation and Applied Informatics
Karlsruhe Institute of Technology
Eggenstein-Leopoldshafen, Germany
karl-heinz.haefele@kit.edu

2023 Joint Urban Remote Sensing Event (JURSE) | 978-1-6654-9373-4/23/$31.00 ©2023 IEEE | DOI: 10.1109/JURSE57346.2023.10144162

***Abstract*—A building's window-to-wall ratio (WWR) plays a critical role in estimating heat loss, solar gain and daylighting levels, and is therefore essential for building energy modeling applications. Typically, an accurate WWR estimation corresponds to an accurate window segmentation result, which requires high quality rectified and annotated façade images. In this paper, we propose a novel end-to-end regression model that directly predicts the invisible building attribute, the WWR, from façade imagery. For comparison, we have adopted the latest proposed semantic segmentation of windows from façade images and calculate the WWR based on the result of the semantic segmentation. These two approaches are performed and compared on three public façade benchmarks. The experimental results demonstrate that the direct prediction of invisible building attributes is feasible. Furthermore, the regression-based approach can achieve similar WWR accuracy as the segmentation-based method when they use the same backbone.**

***Index Terms*—Deep learning, semantic segmentation, regression, window-to-wall ratio (WWR), urban building energy modeling (UBEM).**

I. Introduction

Urban Building Energy Modeling (UBEM) considers impacts of neighborhoods and estimates energy demands at an urban level, it is used in urban planning, energetic refurbishment of neighborhoods and in the planning of energy infrastructures [1]. Accurate UBEM requires various input parameters for dynamic thermal simulation of buildings, an important parameter for the thermal simulation is the window-to-wall ratio (WWR). If the models are textured with terrestrial or oblique aerial imagery, these could be used to determine the WWR. A straightforward approach of image based WWR estimation consists of two steps, firstly the façade segmentation and window segmentation need to be done to obtain pixelwise prediction of windows and walls, secondly the WWR is calculated from the pixel number of windows and walls. However, this approach has some drawbacks. First, the accuracy of estimated WWR is influenced by the accuracy of both façade segmentation and window segmentation. Besides, the segmentation model is restricted by the type of façade imagery and window shape, for example, the model Deepfacade [2] is only applicable for rectified façade imagery. Furthermore, the applicability of the method is restricted by the availability of training data, which has to be pixelwise annotations of windows and walls. On the contrary, direct prediction of WWR from images has less requirements on images quality. And in some cases the WWR data can be obtained from other ways, for example the Commercial Building Energy Consumption Survey (CBECS) [3]. Therefore, we are motivated to investigate into the possibility of directly predicting WWR from images using a regression model.

To tackle this problem, we propose in this paper a novel end-to-end regression model that directly predicts the invisible building attribute, the WWR, from façade imagery. More specifically, the network first extracts deep image features through a convolutional neural network (CNN), and then passes the features onto a regression module, producing the required parameter as output.

This paper is organized as follows: we give an overview of related research work in Section II, and explain in depth the two WWR estimation approaches in Sections III. Experimental results are shown and discussed in Section IV. We conclude the paper in Section V.

II. Literature review

A. Segmentation-based WWR estimation

Recent advances in deep neural networks have contributed to the applications of deep learning-based approaches in WWR estimation. But existing WWR estimation methods generally rely on semantic segmentation of windows and the WWR is calculated in a post-processing step. For example, Touzani et al. [4] propose a deep learning-based approach to segment windows from drone images and then compute the WWR. Similarly, Szcześniak et al. [5] propose a pipeline to extract building layouts from street view images and compute the WWR based on the semantic information. Tarkhan et al. [6] utilize a CNN-based workflow to detect window key-points and group them to form discrete window geometries, and then the dominant outer edges of the façade are cropped to form a polygon. The WWR value is calculated based on the polygonal areas of wall and windows. Thus, the WWR estimation accuracy greatly relies on the quality of the semantic segmentation.

B. Direct numerical prediction

In recent years, deep neural networks have been successfully applied in direct numeric prediction. For example, Amirkolaee and Arefi [7] apply a UNet structure for elevation estimation

from single aerial image. Ptak et al. [8] employ DeepLabV3+ [9], UNet and UNet++ for density estimation from drone images. Islam et al. [10] use a 3D UNet architecture with attention module to estimate the survival duration of patients.

Instead of predicting visible image features like window segments, we propose in this paper to employ a deep neural network to directly predict the invisible building attribute, namely the WWR value, from façade imagery. In comparison with the traditional segmentation-based method where the WWR value is estimated as a post-processing step, the regression-based network is trained end-to-end, thus improving the training efficiency.

III. Methods

A. Regression-based WWR prediction

The regression network takes a façade image as input and deep features are extracted from the input image by a UNet [11]. Then the feature maps are passed on to a fully connected layer and a regression layer, resulting in the final WWR prediction.

Figure 1 illustrates the architecture of the proposed regression network for WWR estimation. It can be seen that the main structure is a UNet whose output layer is replaced by a fully connected layer and a regression layer.

In order to train the regression network, we use the mean absolute error (MAE) as the loss function, as it is the default loss to use for regression problems and more robust to outliers than other loss functions. The MAE loss is defined as:

$$L_{(y,\hat{y})} = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y_i}| \tag{1}$$

Where N is the number of images, y denotes the predicted value and $\hat{y_i}$ denotes the ground-truth.

In implementation, the ground-truth WWR is calculated from the image annotations of the façade datasets. We employ a UNet that is pretrained on ImageNet dataset [12] as backbone. The Adam (adaptive moment estimation) [13] optimization algorithm is used with an initial learning rate of 0.0001, the exponential decay rate of the first moment is set to 0.9 and the second moment 0.999. The learning rate is decayed every 20 epochs by a factor of two.

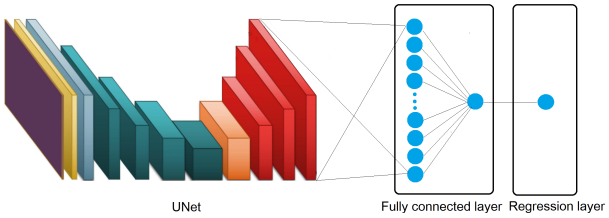

Fig. 1: Architecture of the proposed regression network.

B. Segmentation-based WWR calculation

1) Window segmentation: In our previous work [14] on window vectorization, we proposed a two-fold approach to improve the prediction accuracy of window corners. The workflow is illustrated in Figure 2. First, façade imagery is passed to the U-ResNet101 segmentation network that is integrated with a cross-field and augmented by a self-attention module to improve the segmentation performance, resulting in pixel-wise window segmentation. Second, the segmentation results and the original façade imagery are passed to a regression neural network that is augmented by Squeeze-and-Excitation (SE) attention blocks, resulting in the coordinates of window corners.

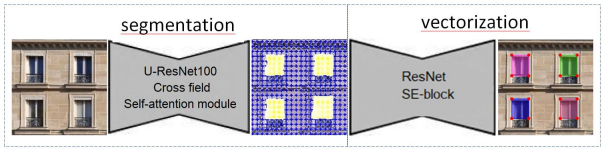

Fig. 2: Workflow of the two-fold method.

2) WWR calculation: Given a rectified façade image C with the size of $H^C \times W^C$, for each rectangular window S_i on the image, the two-fold segmentation and vectorization network outputs the pixel coordinates of the top left corner (x_1^i, y_1^i) and the bottom right corner (x_2^i, y_2^i) of the window. Then the WWR of the façade image C with a total amount of N windows can be calculated by:

$$WWR^C = \frac{\sum_{i=1}^{N}((x_2^i - x_1^i)(y_2^i - y_1^i))}{H^C \times W^C} \tag{2}$$

In Equation 2, it is assumed that the images contain only façades. Though there are a few exceptions, we still use the the pixel number of the image to approximate the total pixel number of the façade.

IV. Experiment

A. Dataset

We evaluate and compare these two approaches on three benchmark datasets, including **ECP** dataset [15], **CMP** dataset [16] and **Graz50** dataset [17]. As original images in each dataset have different shapes, we resize all images as well as masks into patches of 300×300. For each dataset, we follow the same design proposed in [18], i.e., data is randomly split into 80% for training and 20% for testing. The comparison experiments are carried out on each dataset.

Table I lists a few WWR-related statistics on these datasets. More specifically, **Image number** refers to the number of test images in each dataset, **Wall pixel number** refers to the pixel number of building walls, **GT window pixel number** refers to the total pixel number of all windows in the ground-truth data. It can be seen that the WWR predicted by the regression network is very close to the ground-truth value, proving the effectiveness of the regression approach.

The ground-truth data for the regression network is created from the pixelwise window masks provided in the datasets. Given a set of window masks **S**, the WWR of each individual image S_i is defined as follows:

$$WWR_i = \frac{area\_windows_i}{area\_image_i} \tag{3}$$

	CMP	ECP	Graz50
Image number	122	20	10
Wall pixel number	10980000	1800000	900000
GT window pixel number	1558242	259021	152909

TABLE I: WWR-related statistics on CMP dataset, ECP dataset and Graz50 dataset.

	RMSE			**MAE**		
	ECP	CMP	Graz50	ECP	CMP	Graz50
Seg [4]	0.032	0.048	0.089	0.016	0.030	0.052
Two-fold [14]	**0.027**	**0.042**	0.082	0.013	**0.026**	**0.047**
Reg	0.029	**0.042**	**0.080**	**0.012**	0.028	0.048

TABLE II: Comparison of RMSE value and MAE value on ECP, CMP and Graz50 datasets

where, WWR_i denotes the WWR of image S_i, $area\_window_i$ denotes the total pixel number of all windows in S_i, and $area\_image_i$ denotes the number of image pixels, namely the pixel number of wall, in image S_i.

B. Metrics

Most recent work about numerical regression like the person density estimation is based on the standard evaluation strategy [19]–[21]. Two most commonly used metrics are Mean Absolute Error (MAE) and Mean Square Error (RMSE), which are defined as follows:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(R_{I_i}^{pred} - R_{I_i}^{gt})^2} \tag{4}$$

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|R_{I_i}^{pred} - R_{I_i}^{gt}| \tag{5}$$

where N is the number of images, $R_{I_i}^{gt}$ is the ground-truth for the image number i and $R_{I_i}^{pred}$ denotes the predicted value. Roughly speaking, MAE determines the accuracy of the estimates, while RMSE indicates the robustness of the estimation.

C. Results

In order to validate the effectiveness of the proposed method, we compare the regression results with previous works of Touzani et al. [4] and Zhuo et al. [14]. For the sake of fair comparison, we use the UNet as the backbone for baseline and the proposed networks. We test the proposed approach and the baseline approach on three datasets and evaluate the results using RMSE and MAE as the metric. Table II lists the comparison of the RMSE value on ECP, CMP and Graz50 datasets, where the row **Seg** refers to the segmentation-based WWR calculation approach proposed in [4]; the row **Two-fold** refers to the two-fold window prediction approach proposed in [14], which combines a segmentation network and a post-processing vectorization network and achieves significantly higher accuracy than single segmentation network; the row **Reg** refers to the regression-based approach proposed in our paper. Table II lists the comparison of the three approaches on test datasets in terms of the RMSE and MAE value. From this table, it can be seen that accuracy of the regression-based method is higher than the segmentation-based method [4] on all datasets in terms of both the RMSE score and the MAE score. In addition, the accuracy in terms of RMSE of the regression-based method is no less than that of the two-fold method [14] on CMP dataset and Graz50 dataset, and the accuracy in terms of MAE of the regression-based method is higher than the two-fold method [14] on ECP dataset. Overall, the prediction accuracy of the regression-based method and the two-fold method [14] are balanced. As for the training time, the training of the two-fold approach [14] on CMP dataset takes c.a. 1.5 hours per epoch on 4 1080Ti GPUs, while the training of the segmentation-based method and the regression-based method both take around 1 hour. Besides, the regression network can be trained end-to-end while the two-fold method has to be implemented in two separate steps. Therefore the regression network is more easy to implement and efficient than the two-fold approach.

Figure 3 illustrates the WWR calculated from the two-fold method [14] and that predicted by the proposed regression-based method on three sample images. The numbers in blue show the WWR that is calculated by the two-fold approach proposed in [14], numbers in red shows the WWR that is predicted by the regression-based network, and numbers in green shows the ground-truth WWR. It can be seen that the accuracy of regression-based method is slightly lower, that is because the two-fold method has significantly higher segmentation accuracy due to its complicated network architecture.

V. Conclusion

Traditionally, WWR is estimated by calculating the number of pixels of windows and walls from the result of façade segmentation as a post-processing step. In this paper, we propose a novel end-to-end WWR prediction method using a regression-based deep neural network and demonstrate the feasibility of directly predicting the invisible WWR value from façade imagery. In order to validate the effectiveness of the proposed method, we compare the performance with the traditional segmentation-based method. In the case study applied to three public building façade datasets, the proposed method has achieved higher accuracy on all datasets compared to the segmentation-based method, which calculates the WWR value based on the semantic segmentation of the building façades. Moreover, the regression-based method has achieved balanced performance as the two-fold method proposed in [14]. However, the regression-based method can be trained end-to-end and is more efficient than the two-fold method. In addition, the regression-based method has the potential to improve when using a larger backbone (e.g. DeepLabV3 [22]). In has to be mentioned that the two-fold method relies greatly on the image quality and window types, as it requires accurate pixelwise annotations of windows and walls, and the images have to be rectified whereas the windows have to be

(a) Two-fold: 0.140 Reg: 0.139 GT: 0.172 (b) Two-fold: 0.179 Reg: 0.181 GT: 0.150 (c) Two-fold: 0.231 Reg: 0.229 GT: 0.200

Fig. 3: WWR of three sample images. The numbers in blue show the WWR that is calculated by the two-fold approach proposed in [14], numbers in red shows the WWR that is predicted by the proposed regression-based network, numbers in green shows the ground-truth WWR.

rectangular. By contrast, the regression-based method can be applied on any types of façade imagery and window types as long as corresponding training data is provided. Furthermore, the proposed regression model has also potential to predict other invisible building information, such as building age, storey number and storey height, if proper training data are provided. In the future, we plan to extend the applicability of the proposed method to unrectified images such as street view imagery and oblique aerial imagery in different global cities, and update the WWR prediction method accordingly. In addition, we aim to test correlations with other building attributes like building age, construction materials, etc.

REFERENCES

[1] A. Malhotra, J. Bischof, A. Nichersu, K.-H. Häfele, J. Exenberger, D. Sood, J. Allan, J. Frisch, C. van Treeck, J. O'Donnell *et al.*, "Information modelling for urban building energy simulation—a taxonomic review," *Building and Environment*, p. 108552, 2021.

[2] H. Liu, J. Zhang, J. Zhu, and S. C. Hoi, "Deepfacade: A deep learning approach to facade parsing," 2017.

[3] L. Troup, R. Phillips, M. J. Eckelman, and D. Fannon, "Effect of window-to-wall ratio on measured energy consumption in us office buildings," *Energy and Buildings*, vol. 203, p. 109434, 2019.

[4] S. Touzani, M. Wudunn, S. Fernandes, A. Zakhor, R. Najibi, and J. Granderson, "A machine learning approach to estimate windows-to-wall ratio using drone imagery," in *Remote Sensing Technologies and Applications in Urban Environments VI*, vol. 11864. SPIE, 2021, pp. 62–69.

[5] J. T. Szcześniak, Y. Q. Ang, S. Letellier-Duchesne, and C. F. Reinhart, "A method for using street view imagery to auto-extract window-to-wall ratios and its relevance for urban-level daylighting and energy simulations," *Building and Environment*, vol. 207, p. 108108, 2022.

[6] N. Tarkhan, S. Letellier-Duchesne, and C. Reinhart, "Capturing façade diversity in urban settings using an automated window to wall ratio extraction and detection workflow," in *2022 Annual Modeling and Simulation Conference (ANNSIM)*. IEEE, 2022, pp. 706–717.

[7] H. A. Amirkolaee and H. Arefi, "Height estimation from single aerial images using a deep convolutional encoder-decoder network," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 149, pp. 50–66, 2019.

[8] B. Ptak, D. Pieczyński, M. Piechocki, and M. Kraft, "On-board crowd counting and density estimation using low altitude unmanned aerial vehicles—looking beyond beating the benchmark," *Remote Sensing*, vol. 14, no. 10, p. 2288, 2022.

[9] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 801–818.

[10] M. Islam, V. Vibashan, V. Jose, N. Wijethilake, U. Utkarsh, and H. Ren, "Brain tumor segmentation and survival prediction using 3d attention unet," in *International MICCAI Brainlesion Workshop*. Springer, 2019, pp. 262–272.

[11] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[12] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[13] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2015. [Online]. Available: http://arxiv.org/abs/1412.6980

[14] X. Zhuo, J. Tian, and F. Fraundorfer, "Cross field-based segmentation and learning-based vectorization for rectangular windows," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2022.

[15] 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*. IEEE, 2010, pp. 3105–3112.

[16] R. Tyleček and R. Šára, "Spatial pattern templates for recognition of objects with regular structure," in *Proc. GCPR*, Saarbrucken, Germany, 2013.

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

[18] G. Zhang, Y. Pan, and L. Zhang, "Deep learning for detecting building façade elements from images considering prior knowledge," *Automation in Construction*, vol. 133, p. 104016, 2022.

[19] D. Helbing and A. Johansson, "Pedestrian, crowd, and evacuation dynamics," *arXiv preprint arXiv:1309.1609*, 2013.

[20] V. Lempitsky and A. Zisserman, "Learning to count objects in images," *Advances in neural information processing systems*, vol. 23, 2010.

[21] V. A. Sindagi and V. M. Patel, "Generating high-quality crowd density maps using contextual pyramid cnns," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 1861–1870.

[22] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," *arXiv preprint arXiv:1706.05587*, 2017.