Proposal

**Window-to-Wall Ratio Estimation using Oblique Aerial Imagery and 3D city models: A Comparison between Supervised Semantic Segmentation Models and Zero Shot Deep Learning Models**

Mohamad Albaaj

September

Primary Supervisor: Univ.-Prof. Dr.-Ing. Jörg Blankenbach

Secondary Supervisor: Dr.-Ing. Raimund Schwermann

Advisor: M.Sc. Fadi Moubayed

# Summary

Urbanization has accelerated in recent years, driving the need for better control and optimization of energy consumption. Urban Building Energy Modeling (UBEM) provides a method to analyze energy demand and consumption in rapidly urbanizing areas. However, performing large scale energy simulations for existing buildings is challenging due to the time and effort required to obtain the WWR through field measurements for buildings on a large scale. In all cases, Urban Building Energy Modeling (UBEM) requires multiple input parameters. One very important parameter is the WWR which plays a crucial role in building energy simulations. Furthermore, WWR significantly influences energy performance in terms of heating, cooling, and lighting loads. This thesis employs deep learning models to analyze building facade images. The aim is to extract valuable information in buildings facades in order to estimate the window to wall ratio (WWR). Using a dataset of oblique aerial images provided by the municipalities of Soest and Düsseldorf. The research experiments with models like DeepLabV3 and U-Net for semantic segmentation under the deep learning approach. Additionally, this thesis tests the performance of zero shot models such as Grounding DINO for object detection and SAM for segmentation in detecting façade elements that could be utilized to estimate the WWR. Additionally, this thesis compares the performance of conventional deep learning models with the zero shot models, highlighting the advantages and disadvantages of each. Furthermore, This thesis aims at providing window to wall ratio (WWR) as an input parameter for large scale energy simulations by automating the process of (WWR) estimation.

Contents

[Summary 4](#_Toc175781181)

[1 Introduction 7](#_Toc175781182)

[2 Research Background 8](#_Toc175781183)

[3 Thesis goal 10](#_Toc175781184)

[4 Preliminary Results 11](#_Toc175781185)

[4.1 Semantic Segmentation Approach 11](#_Toc175781186)

[5 Work plan 14](#_Toc175781187)

[Bibliography 15](#_Toc175781188)

[list of figures 16](#_Toc175781189)

# Introduction

The rapid expansion of urbanization is a major challenge of the current century. Man made environment is responsible for approximately 40% of the total co2 emissions at its highest point in 2019 and cities make up to 70% percent of the world gas emissions, including co2. Fortunately, the energy use intensity on the decline but at a very low rate, the efforts on optimizing and managing the energy usage are among the main contributing reasons on the reduction (cite, UBEM.io). Urban areas are expanding and the larger the expansion the higher the energy consumption, which require more energy management and optimization [1].

One of the famous method for optimizing the energy consumption within urban areas is the Urban Building Energy Modelling (UBEM). It effectively analyzing the energy demands and consumption resulting from rapid urbanization and city development [2]. It contains several tools to help cities use detailed data about building energy consumptions and environmental factors to make decisions about energy management and efficiency improvements at a city level. It has grown in recent years and is used for various purposes, including analyzining existing buildings, planning new buildings and upgrading some specific buildings (cite, UBEM.io).

When discussing UBEM, the relationship between energy consumption and building design must be taken into account, particularly the external walls, and more specifically, the windows, which play a crucial role in transferring and transmitting energy. The window to wall ratio (WWR) is a key parameter in energy studies of buildings. It affects heating, cooling, lighting, and overall energy efficiency. WWR is crucial for designing energy efficient buildings and reducing environmental impact [1], [2].

Building designs are often inaccessible, especially old drawings for old buildings. Therefore, performing energy simulations for existing buildings on a large scale, such as in cities, is challenging without accessible building designs. Facade measurements are time consuming and costly, particularly when building designs are unavailable. Analyzing building façade images could help address this issue [1]. Recently, various research approaches in segmenting facades have been increasing, producing acceptable results with the utilization of modern machine/deep learning models [3], [4]. This opens the door to leveraging these approaches and optimizing them.

Developing an automated solution for detecting windows and walls in large scale urban areas requires a diverse and extensive dataset of building facades. Such datasets should include images representing different types of buildings. Ensuring a variety of buildings in the dataset improves the automated solution's effectiveness [3]. The Objective of the above mentioned work is to develop a method for determining the window-to-wall ratio (WWR) for individual buildings and applying it in later stage as an input parameter in the energy simulation of buildings [4]. The dataset used in this thesis consists of rectified images obtained from oblique aerial images thanks to municipalities of Soest and Dusseldorf.

# Research Background

Images from which facade geometry can be extracted are generally divided into three categories: aerial images [5], street view images [6] and camera shots [4]. These image types are often used in datasets for facade analysis, with some datasets combining elements from multiple categories. CMP [7] , ECP [8], and Graz50 [9] for example are datasets commonly used for façade segmentation. These datasets contain images of various sizes that have been rectified and annotated. This makes these datasets suitable for multiple purposes such as façade segmentation as well as window to wall ratio calculation [2]. One approach to estimating the WWR is by applying a prediction model that extracts deep features from the façade dataset images using a regression based deep neural network. In this method, the WWR is computed by dividing the area of the windows by the total area of the image [2].

In addition to the three mentioned façade parsing datasets, two more eTRIMS [10] and ArtDeco [11] were used by [3] for extracting façade geometries. The eTRIMS images are not rectified, and most of the ArtDeco images contain significant vegetation. These datasets are generally designed to simplify façade parsing and segmentation tasks. [4] Exercised camera shots and available web images for façades, ensuring a diverse collection that included different building functions and various façade models, such as hotels, schools, dormitories, hospitals, residences, and shopping centers to diversify the dataset. Manual annotation and distortion correction (rectification) were performed during the segmentation process. Several deep learning models were tested, with the custom SOLOv2 algorithm yielding the best results. The goal was not only to estimate the window-to-wall ratio (WWR) but also to apply it to urban areas on a city scale.

Additionally, [5] used drones to collect images of blocks of buildings and utilized photogrammetry to extract 3D models from these images. Through capturing images with a high overlap percentage and using photogrammetry software, matching points were identified to form the 3D models of the buildings. The buildings in the models were isolated by zooming in on the targeted building and cropping out all surrounding objects, vegetation linked to the building was also cropped. Then texture mapping was extracted. Unlike methods which rely on rectified images, all façades were compiled into a single texture map for the entire envelope of each building. Manual labeling of the texture maps was performed, followed by the application of a combination of conditional General Adversarial Networks (cGANs) [12] and Segment Anything Model (SAM) [13] to detect windows in the façades for calculating the window-to-wall ratio (WWR). This approach eliminates the need to rectify each building's façades individually, thereby increasing data processing efficiency.

Street view images have proven to be a valuable data source for collecting and creating datasets. [6] proposes an approach where a large amount of data is collected, but only a portion is manually selected for use. The algorithm aims at estimating the Window-to-Wall Ratio (WWR) without relying on any machine learning methods. The main goal was to build an algorithm to filter street view images as an initial step, then rectify the buildings in the images to prepare them for the next step, which involved detecting the edges using an edge filter such as the Sobel filter. The final step was calculating the resultant mask drawn for the windows and the wall. The disadvantage of this method is that the windows must have hardened edges to be detected.

To diversify the dataset with images that include obstacles such as trees, [14] utilized a combination of publicly available façade datasets and street view images. This approach aims at enhancing the dataset's variety in resolution, quality, and the presence of obstacles, thereby improving the model's ability to generalize and learn from different types of images. The dataset combined elements from two distinct categories, resulting in images of various sizes, some of which were prelabeled while others required manual labeling. A rectifying algorithm was employed, involving steps such as taking the footprints of the buildings' polygons, identifying the corner points, calculating the camera angle based on the midpoint between these points and the camera location, and applying homography to rectify the extracted images into an orthogonal view. This processed dataset was then subjected to a CNN architecture for machine learning applications. After testing the method on around 864 images from the dataset, the successful detection rate reached up to 72% of the buildings with 10 % error rate, highest when windows are typology and punched, lowest when windows are glazed.

# Thesis goal

Bridging computer vision with practical energy studies through machine learning and focusing on the automatic estimation of the window-to-wall ratio (WWR) on a large scale. The WWR is a crucial parameter in energy studies. It significantly impacts building energy performance. Automating WWR estimation on a large scale will reduce both the time and cost associated with manual measurements [1]. This integration could enhance building energy performance on a large scale and results in better and more feasible energy modelling.

Traditional deep learning models can be employed for semantic segmentation. Experimenting with various architectures is key to achieving optimal performance and identifying the most suitable model for the task. An evaluation framework is crucial for assessing model performance. Using metrics such as accuracy, precision, and recall [4]. Models like DeepLabV3 and U-Net are promising in this context. Additionally, Expanding the dataset through the addition of other limited available datasets is likely to provide improvements in model performance.

In addition, testing with zero shot object detection model such as Grounding DINO [15] and zero shot segmentation model such as SAM [13], shows potential in this context. These models are valuable because they do not require any prior training on the specific objects in the dataset. These models represent a significant advancement in the field of object detection and semantic segmentation [13], [15]. Exploring whether these models can be effectively utilized for calculating the window-to-wall ratio (WWR). Comparing their performance with the performance of the supervised models.

# Preliminary Results

In this section, four different distinct approaches for detecting windows on façade will be presented, each leveraging different machine learning techniques. Two of the four approaches are supervised semantic segmentation and supervised object detection, both of which rely on labeled data to train the models. In contrast, the other two approaches, which are zero shot object detection and zero shot semantic segmenation, do not require labeled data. The two approaches, supervised semantic segmentation and the one shot semantic segmentation aiming to classify each pixel in an image, while the supervised object detection and the one shot object detection focus on identifying and localizing windows as bounding boxes. As shown in the figure (number) which illustrates the roadmap of the previous presented approaches. Results obtained of each method are the accuracy, precesion, recall, F1 score, IOU, TP, TN, FP, FN, (size/area to see the bounding box is it over or underestimates the window size compared to the segmented one, and at what percentage),

## Semantic Segmentation Approach

A sample of the dataset consisting of 49 images was labeled using the makesense.ai labeling tool [16]. The images were annotated into two classes: background and windows. Polygons were used to define the boundaries of the objects, ensuring compatibility with semantic segmentation methods. The annotations for each image were exported as a file containing points that define the polygon boundaries around each object class. These points are associated with their respective categories and the image ID. After reading the exported annotations file, each class was assigned a unique color from the grayscale range to create the masks. The masks were then exported in PNG format, which is ideal for preserving exact pixel values. A visual inspection was conducted to verify that the masks corresponded accurately to the expected annotations as shown in Figure 1.

A collage of different images of a building

Description automatically generated

**Figure 1**: images with mask

The dataset was Split into training, validation, and testing subsets with a distribution ratio of 0.7, 0.2, and 0.1 respectively. This ensures that each image is paired with its corresponding mask. The training images were resized to a uniform dimension to maintain consistency in the input data. For segmentation, a Fully Convolutional Network (FCN) based on the ResNet-50 architecture was selected, utilizing pretrained weights. The model was customized by adjusting Key training parameters, including setting the number of epochs to 50, choosing the Adam optimizer, and using Sigmoid as a loss function. The model’s performance was assessed on the testing subset through visual inspections of the predicted segmentations as shown in Figure 2. This evaluation was crucial for determining the model's accuracy and its ability on unseen data.

A screenshot of a computer generated image

Description automatically generated

**Figure 2**: image with mask and predictions

As shown in Figure 2, the results were inaccurate in the first and second images, but a bit improved in the third. One clear reason for this is the limited size of the dataset used for training. To improve the performance of the model, several techniques can be applied, with the most effective likely being to expand the dataset. As mentioned earlier, there are a few available datasets that can be incorporated to increase the dataset size and improve the model performance.

The following intended sections of the work will compare the performance of this model with zero shot semantic segmentation, as well as object detection with zero shot object detection. Each model's results will be discussed, and the optimal approach will be highlighted.

# Work plan

|  |  |
| --- | --- |
| 1-2 weeks | Collecting data, research papers about the topic |
| 3-4 weeks | implementation of the first semantic segmentation model fcn\_resnet50 |
| 5-6 weeks | Documenting the related information of the first implementation |
| 7-8 weeks | Implementing more suitable models to the case such as U\_net |
| 9-10 weeks | Improving the performance of the models until reaching satisfactory results |
| 11-12 weeks | Labeling the rest of the dataset for the semantic segmentation purpose |
| 13-14 weeks | Trying with zero shot semantic segmentation model such as SAM and zero shot object detection model such as Grounding DINO |
| 15-16 weeks | Implementing the object detection model such as R-CNN |
| 17-18 weeks | Improving the performance of the model until reaching satisfactory results |
| 19-20 weeks | Comparing the results of the semantic segmentation and the object detection to the ground truth |
| 21-22 weeks | Documenting the models and the used scripts |
| 23-24 weeks | Preparing proper visualization of the results |

# Bibliography

[1] X. Shi, C. Wang, M. Wang, and P. Tang, “An Innovative Approach to Determine Building Window- To-Wall Ratios for Urban Energy Simulation”.

[2] X. Zhuo, J. Tian, and K.-H. Häfele, “Direct Window-to-Wall Ratio Prediction Using Deep Learning Approaches,” in *2023 Joint Urban Remote Sensing Event (JURSE)*, Heraklion, Greece: IEEE, May 2023, pp. 1–4. doi: 10.1109/JURSE57346.2023.10144162.

[3] H. Liu, Y. Xu, J. Zhang, J. Zhu, Y. Li, and S. C. H. Hoi, “DeepFacade: A Deep Learning Approach to Facade Parsing With Symmetric Loss,” *IEEE Trans. Multimedia*, vol. 22, no. 12, pp. 3153–3165, Dec. 2020, doi: 10.1109/TMM.2020.2971431.

[4] Y. Lu, W. Wei, P. Li, T. Zhong, Y. Nong, and X. Shi, “A deep learning method for building façade parsing utilizing improved SOLOv2 instance segmentation,” *Energy and Buildings*, vol. 295, p. 113275, Sep. 2023, doi: 10.1016/j.enbuild.2023.113275.

[5] A. J. Su, K. C. Xu, A. Ren, T. Liu, and T. Dogan, “Facade Scanner: A scalable workflow for building geometry and window-to-wall ratio capture for urban building energy modeling.”.

[6] 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, Jan. 2022, doi: 10.1016/j.buildenv.2021.108108.

[7] R. Tyleček and R. Š. ara, “Spatial Pattern Templates for Recognition of Objects with Regular Structure,” in *Proc. GCPR*, Saarbrucken, Germany, 2013.

[8] 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*, San Francisco, CA, USA: IEEE, Jun. 2010.

[9] H. Riemenschneider *et al.*, “Irregular lattices for complex shape grammar facade parsing,” in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, Providence, RI: IEEE, Jun. 2012.

[10] F. Korč and W. F\" orstner, “eTRIMS Image Database for Interpreting Images of Man-Made Scenes,” TR-IGG-P-2009-01, Apr. 2009. [Online]. Available: http://www.ipb.uni-bonn.de/projects/etrims\_db/

[11] R. Gadde, R. Marlet, and N. Paragios, “Learning grammars for architecture-specific facade parsing,” *International Journal of Computer Vision*, vol. 117, no. 3, pp. 290–316, 2016.

[12] I. J. Goodfellow *et al.*, “Generative Adversarial Networks.” 2014. [Online]. Available: https://arxiv.org/abs/1406.2661

[13] A. Kirillov *et al.*, “Segment Anything.” 2023. [Online]. Available: https://arxiv.org/abs/2304.02643

[14] 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)*, San Diego, CA, USA: IEEE, Jul. 2022, pp. 706–717. doi: 10.23919/ANNSIM55834.2022.9859521.

[15] S. Liu *et al.*, “Grounding DINO: Marrying DINO with Grounded Pre-Training for Open-Set Object Detection.” 2024. [Online]. Available: https://arxiv.org/abs/2303.05499

[16] P. Skalski, “Make Sense.” 2019. [Online]. Available: https://github.com/SkalskiP/make-sense/

# list of figures

[**Figure 1**: image with mask 12](#_Toc175781270)

[**Figure 2**: image with mask and predictions 13](#_Toc175781271)