

## OccFaçade: enabling precise building façade parsing in large urban scenes with occlusion

Yongjun Zhang, Dongdong Yue, Xinyi Liu, Siyuan Zou, Weiwei Fan & Zihang Liu

**To cite this article:** Yongjun Zhang, Dongdong Yue, Xinyi Liu, Siyuan Zou, Weiwei Fan & Zihang Liu (2024) OccFaçade: enabling precise building façade parsing in large urban scenes with occlusion, *International Journal of Remote Sensing*, 45:18, 6651-6674, DOI: [10.1080/01431161.2024.2391589](https://doi.org/10.1080/01431161.2024.2391589)

**To link to this article:** <https://doi.org/10.1080/01431161.2024.2391589>



Published online: 01 Sep 2024.



Submit your article to this journal [↗](#)



View related articles [↗](#)



View Crossmark data [↗](#)



# OccFaçade: enabling precise building façade parsing in large urban scenes with occlusion

Yongjun Zhang, Dongdong Yue , Xinyi Liu , Siyuan Zou, Weiwei Fan and Zihang Liu

School of Remote Sensing and Information Engineering, Wuhan University, Wuhan, China

## ABSTRACT

Building façade parsing is to recognize the building façade image into different categories of individuals including walls, doors, windows, balconies, etc. However, obstructions such as trees present a significant challenge to conducting façade parsing. In this paper, we designed OccFaçade to achieve high-precision parsing of occluded building façades in large urban scenes. OccFaçade primarily incorporates two modules, Multi-layer Dilated Convolution Module (MD-Module) and Multi-scale Row-Column Convolution Module (MRC-Module), to capture repeated texture in local and row-column directions. This aims to leverage repetitive textures to address occlusion challenges in building façade parsing. Besides, we introduce our building façade dataset MeshFaçade from the Mesh data generated by drone imagery to study the occlusion problem of missing textures. The experimental results demonstrate that OccFaçade achieves state-of-the-art performance with mIOU of 85.01%, 84.09%, 72.95%, and 88.83% on the ENPC2014 dataset, ECP dataset, RueMonge2014 dataset, and our MeshFaçade dataset, respectively. The code and data are available at <https://github.com/yueyisui/OccFaçade>.

## ARTICLE HISTORY



Received 12 April 2024  
Accepted 3 August 2024

## KEYWORDS

Façade parsing; large urban scene; occlusion; repetitive texture; row-column feature

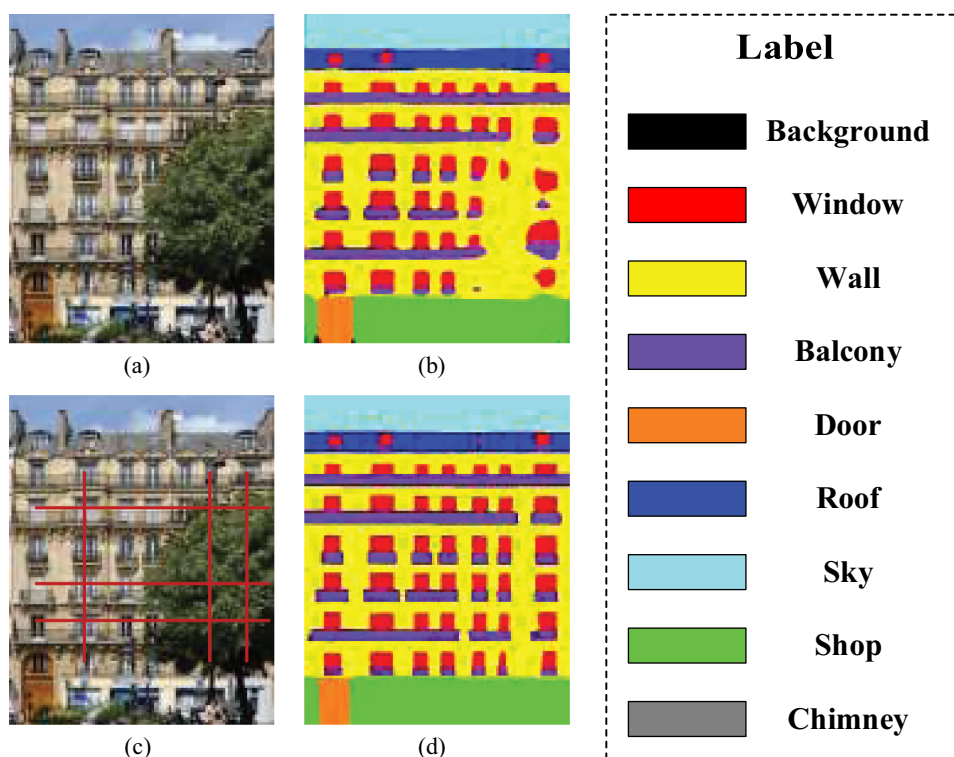
## 1. Introduction

Segmenting building façade elements, including doors, windows, and balconies, is important in parsing building façade images. This task is important for understanding urban environments, and its results are widely used in 3D building reconstruction (C. Li, Zhang, and Zhang 2016), autonomous driving (Geiger et al. 2013), urban planning (Gonzalez-Aguilera et al. 2013), and so on. The Level of Detail 3 (LoD3) model in CityGML (Gröger and Plümer 2012) particularly emphasizes the importance of detailed components such as windows (Hu et al. 2022). Therefore, it is crucial to adopt a high-precision building façade parsing method, as it contributes to superior building model reconstruction and a deeper understanding of urban semantic information. In essence, the meticulous segmentation of building façade elements is the cornerstone for advancing various applications that rely on a meticulous grasp of the urban landscape.

**CONTACT** Xinyi Liu  [liuxy0319@whu.edu.cn](mailto:liuxy0319@whu.edu.cn)  School of Remote Sensing and Information Engineering, Wuhan University, Luoyu Road, Hongshan District, Wuhan, Hubei, China

© 2024 Informa UK Limited, trading as Taylor & Francis Group

Although semantic segmentation techniques in computer vision can be directly applied to building façade parsing, there are still some challenges. The most difficult problem to solve is the occlusion problem caused by tall objects such as trees as shown in [Figure 1\(a\)](#). Early methods (Cohen et al. 2017; Cohen, Schwing, and Pollefeys 2014; Kozinski et al. 2015) discovered that building façades have repetitive artificial structural features, and they used rules and shape grammar to solve the occlusion problem in appearance parsing. However, they have limited accuracy and lack generalizability. In recent years, convolutional neural networks (Chen et al. 2018; Long, Shelhamer, and Darrell 2015; Ronneberger, Fischer, and Brox 2015; K. Sun et al. 2019) and Transformer-based deep learning methods (Carion et al. 2020; Xie et al. 2021) have mainly focused on improving the accuracy of semantic segmentation without considering the unique structural characteristics of building façades. Unlike other semantic segmentation tasks, building façade parsing requires obtaining the real building semantic information behind occluded areas (as shown in [Figure 1\(d\)](#)). This requirement makes it difficult for conventional semantic segmentation models to achieve pleasant results on building façade images. In addition, available datasets for studying building façades include ECP2011 (Teboul et al. 2011), CMP2013 (Tyleček and Šára 2013), ENPC2014 (Gadde, Marlet, and Paragios 2016), RueMonge2014 (Riemenschneider et al. 2014), etc. Most of these datasets consist of images that have been rectified and aligned to axes. This time-consuming and



**Figure 1.** Building façade parsing. (a) Building façade with occluded signal. (b) Result of building façade parsing by DeepFaçade. (c) Repeatability in row-and-column direction. (d) Result of building façade parsing by OccFaçade.

labour-intensive approach to image acquisition results in less data, with almost no large-scale urban building façade data. Therefore, before building façade parsing, we first make a large-scale urban building façade dataset MeshFaçade. The production process is simple, efficient, and fully automated. Initially, a large-scale building model reconstruction is achieved through an automated modelling process (X. Liu et al. 2023). Subsequently, the mesh textures are projected onto the corresponding building planes using parallel projection to obtain façade images.

To solve the interference caused by occlusion, we design a new network called OccFaçade, which uses encoding and decoding structures to capture local information and row and column direction information on the building façade to achieve high-precision building façade parsing. We divide the building feature extraction behind the occlusion signal into the local scope and the global scope. First, a Multi-layer Dilated Convolution Module (MD-Module) is introduced to extract the local features of the occlusion area. We then introduce a Multi-scale Row-Column Convolution Module (MRC-Module) to capture the repeated texture features in the row and column directions of the building façade (shown in Figure 1(c)) and integrate it into each sampling stage. Experimental results show that compared with the DeepFaçade (H. Liu et al. 2020) method, our OccFaçade not only effectively extracts the semantic information behind the occluded areas but also achieves better results in boundary detail segmentation, as shown in Figure 1(b) and (d). The contributions of this paper are summarized below:

- (1) We propose OccFaçade to address occlusion challenges in building façade parsing, enabling high-precision semantic segmentation of urban façades.
- (2) The MD-Module is designed to capture texture features similar to those near occluded areas to reduce the adverse impact of local occluded areas on segmentation performance.
- (3) The MRC-Module is designed to capture repeating texture features in both horizontal and vertical directions to solve the challenge of occlusion and refine the boundaries of building elements such as doors and windows.
- (4) We propose a building façade dataset named MeshFaçade, which is derived from Mesh data texture mapping to LoD building models for building façade related research.

The rest of this paper is organized as follows. Section 2 provides an overview of the related research work relevant to this study. In Section 3, OccFaçade, along with its corresponding MD-Module and MRC-Module, are comprehensively discussed. In Section 4, we introduce the classical dataset used for the experiments as well as the MeshFaçade dataset mentioned in this paper. Section 5 and Section 6 outline the experimental setup and present the comparative results, respectively. Finally, Section 7 concludes and summarizes the proposed approach.

## 2. Related works

### 2.1. Traditional building façade parsing methods

Since the shapes of buildings are human-designed and have strong regularity and repetitiveness, many studies have employed the shape grammar and structural



information of buildings to parse façades with occlusions. For instance, Kozinski et al. (2015) proposed a new shape prior form capable of segmenting both visible and occluded objects and restoring the occluded façade structure. Koutsourakis et al. (2009) proposed the use of primitive shapes and parameterized rules for single-view modelling, which allows for adjustment and adaptation to different architectural styles. Teboul et al. (2010) proposed an approach for building façade parsing that combines shape grammar, supervised classification, and random walks. They also introduced a reinforcement learning method for façade parsing, which resulted in improved computational speed (Hu et al. 2022). Gadde et al. (2018) proposed a façade parsing method that uses decision trees trained with automatic contextual features, leveraging the highly structured prior information of buildings. Furthermore, some methods directly incorporate the constraints of repetitive textures found in building façades. Wendel, Donoser, and Bischof (2010) proposed utilizing prior knowledge of repetitive regions in buildings to segment a single façade. For example, if windows in the façade image exhibit similarities, they described the window features and searched for all locations with similar features to segment the façade. Cohen et al. (2017) also utilized prior information on building symmetry and repetitiveness to address building occlusion issues.

Moreover, dividing the task of façade parsing into distinct steps is also a common approach. Cohen, Schwing, and Pollefeys (2014) proposed a sequential optimization technique that can correct the semantic shape of a building façade and optimize semantic categories. Ripperda and Brenner (2006) analysed building façade from a structural description perspective, constructing a parsing tree to derive their method based on a reversible jump Markov chain Monte Carlo process. Han and Zhu (2009) proposed a top-down and bottom-up algorithm, utilizing it to infer the parsing of the building façade. P. Zhao et al. (2010) proposed a method for parsing building façade images into building units. They first segment the environment into three objects: ground, building, and sky, and then separately partition the building façade as an independent exterior facet. Mathias, Martinović, and Gool (2016) used a three-layer system, including semantic segmentation, object parsing, and building parsing, which entails a step-by-step and coarse-to-fine segmentation process.

In conclusion, while these traditional approaches have contributed to building façade parsing, their precision is generally limited, and some methods rely heavily on manual intervention. Consequently, these traditional methods are less frequently considered in modern building façade parsing.

## ***2.2. Deep learning-based building façade parsing methods***

In recent years, deep learning technologies have found widespread application in the field of image processing, achieving state-of-the-art (SOTA) results in areas such as semantic segmentation (Chen et al. 2018; Long, Shelhamer, and Darrell 2015) and object detection (Girshick et al. 2014; Redmon et al. 2016). Semantic segmentation and target detection are commonly used deep learning methods for building façade parsing. Researchers have successfully employed deep convolutional networks with rich contextual information or transformer networks with comprehensive contextual understanding to segment building façades into distinct regions, facilitating the identification of individual elements.

Due to the inherent characteristics of building elements such as doors and windows, which usually appear as rectangular shapes, building façade parsing is usually performed through pixel-level semantic segmentation or area-level object detection. Methods in deep learning for object detection, such as Mask R-CNN (He et al. 2018), YOLO (Redmon et al. 2016), and DETR (Carion et al. 2020), can generate multiple bounding boxes on images. Each bounding box reflects the detected position and rectangular extent of the object. Assigning semantic attributes to the corresponding rectangular bounding boxes facilitates the parsing of building façades. For example, Nordmark and Ayenew (2021) employ Mask R-CNN (He et al. 2018) to detect windows, while Y. Sun et al. (2022) enhance detection accuracy by incorporating attention modules into Mask R-CNN. Although their method partially addresses the issue of small area occlusion, it still struggles to detect the complete outline of windows in scenes with significant occluded areas. In addition to bounding boxes, window detection can be achieved by determining the positions of keypoints. There are two approaches to determining keypoints: directly getting the four corners of the window to establish its scope (C. Li et al. 2020) or determining the centre and size of the window to get its scope (Tao, Zhang, and Chen 2022). However, methods for keypoint detection are not always stable, and errors in the localization and positional relationships of keypoints will occur in complex scenes.

Building façade parsing based on semantic segmentation achieves semantic information parsing at the pixel level. Conventional semantic segmentation models, such as FCN (Long, Shelhamer, and Darrell 2015), U-Net (Ronneberger, Fischer, and Brox 2015), Deeplab v3+ (Chen et al. 2018), HRNet (K. Sun et al. 2019), and Segformer (Xie et al. 2021), can be directly applied to building façade parsing. Among them, U-Net (Ronneberger, Fischer, and Brox 2015) is a promising option for building façade parsing, as it demonstrates superior generalization for small-scale datasets and can help reduce overfitting. Based on the achievements above, many excellent works have applied deep learning technology to the field of remote sensing. Examples include the Multistage Information Complementary Fusion Network (J. Wang et al. 2024), the Spatial-Logical Aggregation Network (SLA-NET) (M. Zhang et al. 2023), and the Graph-Feature-Enhanced Selective Assignment Network (W. Li and Tao 2022). However, these methods are not specifically tailored to datasets of building façades and often fail to optimize performance due to ignoring the inherent structural characteristics of building façades. ALKNet (Ma et al. 2021) is designed to represent the relationships between building elements in multi-scale feature maps and improve the accuracy of building façade parsing through the contextual information. Meanwhile, H. Liu et al. (2020) propose DeepFaçade, which uses the symmetry of building elements as a loss constraint in training FCN (Long, Shelhamer, and Darrell 2015) networks to improve the precision of symmetric elements such as doors and windows. Based on DeepFaçade, Zhang et al. (G. Zhang, Pan, and Zhang 2022) design a novel hierarchical deep-learning framework for building façade parsing by integrating PSPNet (H. Zhao et al. 2017), DANet (Fu et al. 2019), and DETR (Carion et al. 2020). Kong and Fan (2021), also integrating the advantages of semantic segmentation networks and object detection networks, designed a new pipeline for street-level building façade segmentation. RTFP (B. Wang et al. 2024) utilizes Vision Transformer (ViT) (Dosovitskiy et al. 2021) and line feature extraction to capture semantic information of building façades. However, their methods are difficult to effectively solve the problem of building information loss caused by tree occlusion or loss of texture itself.

Given the above challenges, existing methods make it difficult to solve the occlusion problem, especially when trees block buildings or textures are lost, and building information in the corresponding area cannot be directly obtained from the image. To take advantage of the inherent structural advantages of buildings and leverage the end-to-end capabilities of deep learning methods in semantic segmentation, OccFaçade is designed to learn both the repetitive texture information along the row and column directions of building façade images and the details of local texture information to obtain Building façade information behind the occluded area. OccFaçade focuses on the following key issues:

- Maximizing the structural advantages of buildings and the end-to-end capabilities of deep learning to achieve semantic segmentation of building façades.
- Addressing the issue of missing semantic information in building façades caused by obstructions like trees or texture loss.

### 3. Methodology

This section first introduces the network framework structure of OccFaçade. Then, it details the MD-Module and MRC-Module, which are adept at addressing occlusion issues. The MD-Module focuses on extracting similar texture information in local areas of the final feature map layer. Meanwhile, the MRC-Module captures feature information in row and column directions at different scales during each upsampling process using a rectangular convolution kernel. Finally, we introduce the loss function employed for training the network.

#### 3.1. Network overview

Figure 2 shows the overall framework of the OccFaçade designed in this paper, which adopts an encoder-decoder structure. The encoding process uses many downsampling operations, which is advantageous for the network to extract features from images at different scales. To improve the generalization ability and reduce the training time of the

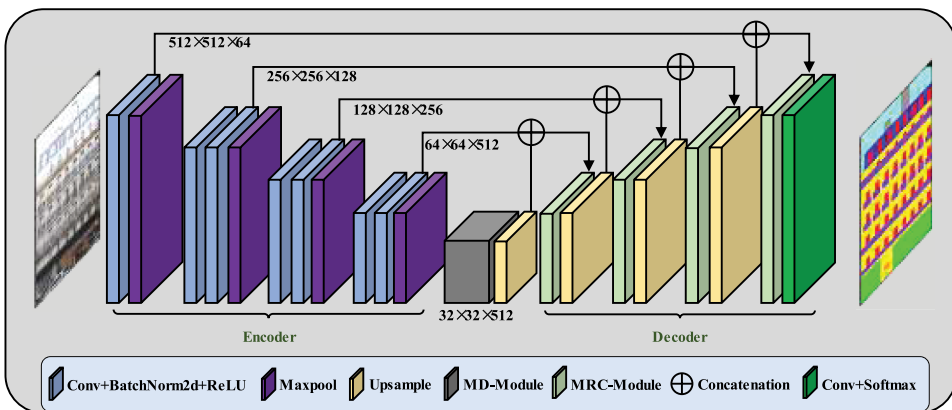


Figure 2. Overview of the proposed OccFaçade framework.

network, the encoding stage of OccFaçade uses the VGG16 (Simonyan and Zisserman 2015) network trained on the ImageNet (Deng et al. 2009) dataset as the backbone.

After cropping or resizing, the input image is adjusted to a fixed size of  $512 \times 512 \times 3$ . After the encoding process, a feature map of  $32 \times 32 \times 512$  is obtained, which is subsequently transmitted to the Multi-layer Dilated Convolution Module (MD-Module) for extracting similarity texture features within a constrained scope. The MD-Module produces a feature map of the same size ( $32 \times 32 \times 512$ ) as the input map. However, the occluded area is now enriched with information from the neighbouring non-occluded regions after the MD-Module.

During the decoding process of the framework, a sequence of upsampling operations is used. Each upsampling process involves inputs that include features of the corresponding size from the encoder and the output from the previous process, utilizing a skip connection. MD-Module can be understood as the process of obtaining local similar texture features within a small isotropic area, while MRC-Module is the process of obtaining local features with repetitive texture positions in both the row and column directions. As shown in Figure 1, windows and balconies of buildings have a high degree of repetitiveness, and adding this feature to the network can effectively solve the occlusion problems.

### 3.2. Multi-layer dilated convolution module (MD-Module)

To better capture the information around the occluded region, we design a multilayer dilated convolution module, the input and output of the MD-Module are feature maps with the size of  $32 \times 32 \times 512$ . As shown in the schematic diagram of the MD-Module structure in Figure 3, the MD-Module has 5 dilated convolution operations, and the convolution kernel is a square of size 3, but they have different dilation rates: 1, 2, 4, 8, 16. Larger convolution kernels can increase the receptive field (Peng et al. 2017), but directly increasing the kernel size will rapidly increase the number of parameters. Assuming a normal

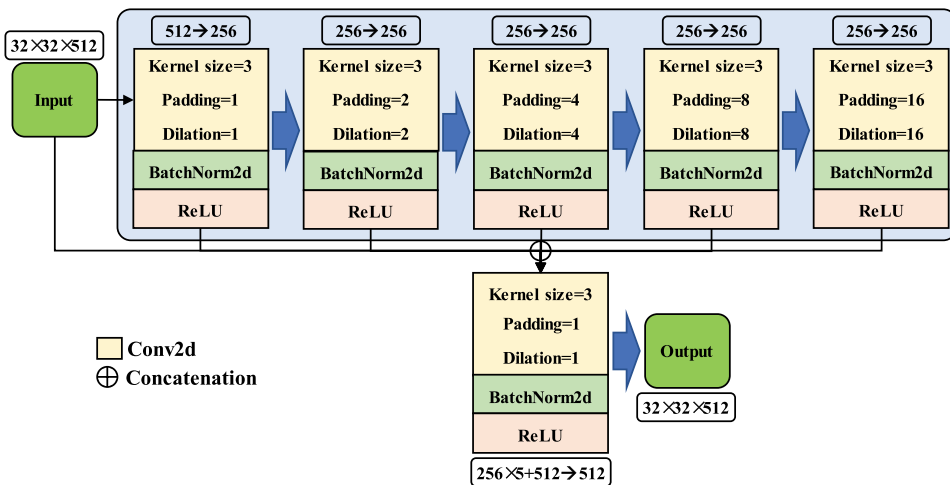
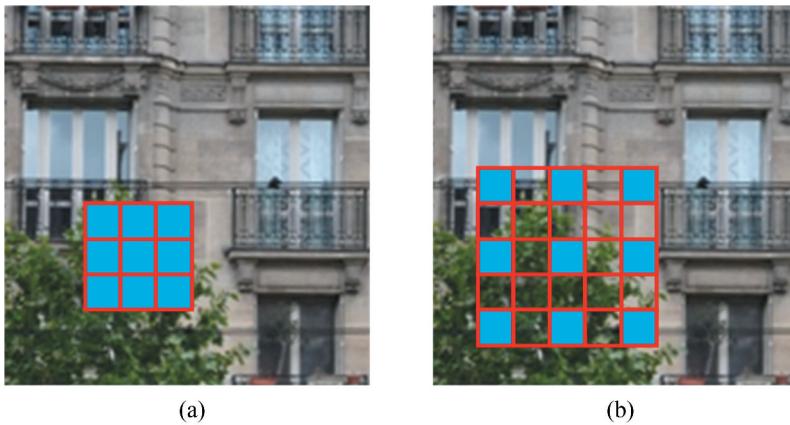


Figure 3. Schematic diagram of the multi-layer dilated convolution module (MD-Module).



**Figure 4.**  $3 \times 3$  convolution kernels with different dilation rates. Kernel (a) has 9 parameters, while kernel (b) has a dilation rate of 2 and 9 parameters. With the same number of parameters, kernel (b) has a larger receptive field.

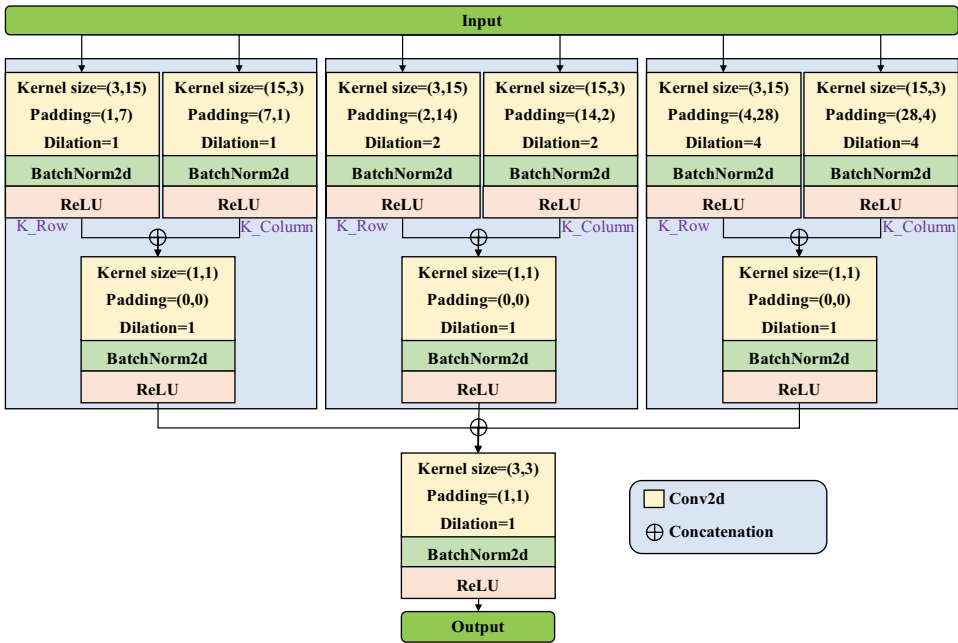
convolution kernel size is a square of  $x$ , an increase of 1 in kernel size corresponds to an increase of  $2x + 1$  in parameters.

Therefore, we chose to use dilated convolutions with different expansion rates instead of increasing the kernel size. As shown in Figure 4, using the same number of parameters can also expand the receptive field of the convolution. The input to each dilated convolution module is the output of the previous step. Finally, the input of the MD-Module is connected to the outputs of the five-hole convolution modules, and the resulting output is passed through a regular convolution layer to adjust its dimensionality before being sent to the upsampling process.

### 3.3. Multi-scale row-column convolution module (MRC-Module)

Unlike other semantic segmentation works, building facades have strong repetitive structures in both horizontal and vertical directions, such as windows, balconies, and other objects, as shown in Figure 1(c). Inspired by (Mei et al. 2020), we leverage this unique characteristic of building facade data to obtain similar features from occluded regions on the same row and column, to complement the information in those regions. Therefore, we have designed the MRC-Module. After each upsampling operation, a constant MRC-Module is applied to capture row and column orientation features at varying scales.

As shown in Figure 2, the output of the previous operation is concatenated with the corresponding-sized features from the encoder as the input to the MRC-Module. As illustrated in Figure 5, the MRC-Module consists of three distinct sub-modules directly connected to the input feature layer. After passing through these three parallel sub-modules, the features are concatenated together and outputted through a  $3 \times 3$  convolutional layer, as shown in Figure 5. Each sub-module has two convolutional kernels,  $K_{Row}$  and  $K_{Column}$ , with the same size but different orientations:  $(3, 15)$  for the row direction and  $(15, 3)$  for the column direction. These kernels are used to capture features



**Figure 5.** Schematic diagram of the multi-scale row-column convolution module.

in the row and column directions, respectively. The row and column information are obtained through  $3 \times 3$  convolutional layers. To capture features at different scales in the row and column directions, the dilation rates of  $K\_Row$  and  $K\_Column$  in the three sub-modules are set to 1, 2, and 4, respectively.

### 3.4. Loss function

Given that the building façade parsing belongs to the field of semantic segmentation in computer vision, we employ cross-entropy loss as the loss function. Specifically, the loss function is defined as follows:

$$L = \frac{1}{N} \sum_i^N (y_i \log \hat{y}_i + (1 - y_i) \log(1 - \hat{y}_i)) \quad (1)$$

where  $N$  is the number of all pixels,  $y_i$  is the true probability distribution value of the  $i$ th pixel, and  $\hat{y}_i$  is the probability division value of the  $i$ th pixel predicted by the network.

## 4. Experimental dataset

To demonstrate the performance of our MD-Module and MRC-Module in building façade parsing, we initially evaluate their effectiveness using three different types of publicly available datasets: (1) building façade datasets without occlusions: ECP dataset (Teboul et al. 2011); (2) building façade datasets with occlusions: ENPC2014 dataset (Gadde, Marlet, and Paragios 2016); (3) multi-view building façade datasets: RueMonge2014 dataset (Riemenschneider et al. 2014). Additionally, to facilitate the acquisition of building

façade images in large urban scenes, we create MeshFaçade dataset. This dataset is derived from the projection of a mesh model generated from drone imagery. Furthermore, due to Mesh quality issues, the loss of texture in this dataset also presents a form of occlusion challenge.

#### **4.1. ECP dataset**

The ECP dataset comprises façade images of the Haussmann-style buildings in Paris, collected by the École Centrale Paris. There are a total of 104 usable images, each with varying sizes, ranging from a maximum resolution of  $628 \times 554$  to a minimum resolution of  $186 \times 486$ . There are two versions of annotations available for the ECP dataset. The first version is annotated while ensuring the translational symmetry of elements such as doors, windows, and balconies, which results in a visually pleasing appearance. However, in this version, the balconies are connected to each other to maintain aesthetic consistency and regularity, which deviates from the actual appearance. The second version, annotated by Mathias, Martinović, and Gool (2016), aims to annotate the images based on the actual appearance and includes nine categories: background, window, wall, balcony, door, roof, sky, shop, and chimney. This version includes an additional category of chimneys compared to the first version. Given that semantic segmentation and object detection tasks require standardized datasets, this paper chooses to use the second version of the ECP dataset. Since semantic segmentation and object detection tasks require more standardized datasets, this paper opts to use the second version of the ECP dataset.

#### **4.2. ENPC2014 dataset**

The ENPC2014 dataset includes 79 rectified façade images captured from Parisian buildings with Art Nouveau style. This dataset comprises eight categories: background, window, wall, balcony, door, roof, sky, and shop. In contrast to the ECP dataset, the ENPC2014 dataset contains many images which are partially occluded by objects such as trees, streetlights, and power poles, which poses a significant challenge for façade parsing.

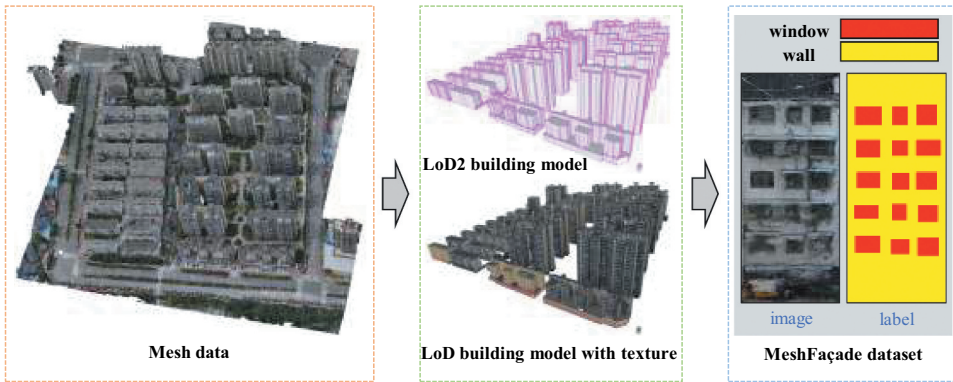
#### **4.3. RueMonge2014 dataset**

The RueMonge2014 dataset includes 428 images captured from a street in Paris, but only 219 of them are manually annotated with semantic information. The dataset consists of 8 categories, including background, window, wall, balcony, door, roof, sky, and shop. Since the RueMonge2014 dataset was captured along Rue Monge Street in Paris, it has the characteristic of having multiple viewpoints.

#### **4.4. MeshFaçade dataset**

The MeshFaçade dataset comprises 3120 images generated by projecting the mesh data of buildings. Among them, there are 306 images of roof façades and 2814 images of building façades. We annotate 184 building façade images to obtain semantic labels for training and predicting the network. The classification information of the MeshFaçade data set is relatively simple, with only two categories: wall





**Figure 6.** The process of MeshFaçade dataset production.

and window. Different from the acquisition methods of the above three data sets, the images generated by Mesh do not require additional correction, and it is easier to acquire large-scale data. To acquire the data for the entire city, corresponding aerial images are needed for photogrammetry to generate the Mesh model. Subsequently, an automated modelling algorithm (X. Liu et al. 2019, 2023) is employed to generate LoD2 models (Gröger and Plümer 2012). The mesh is finally projected onto the corresponding building planes to obtain images of building façades as shown in Figure 6. Simultaneously, the building façade images in MeshFaçade exhibit a certain lack of texture. The lack of texture can be seen as the effect of the building façade being occlusive, which poses a new challenge to the task of building façade parsing.

## 5. Experimental settings

### 5.1. Evaluation metrics

To quantitatively evaluate the proposed OccFaçade in this paper, we utilize Intersection Over Union (IOU) and accuracy for each class, mean Intersection Over Union (mIOU), class average accuracy (mPA), F1-score and total pixel accuracy (Acc) as evaluation metrics, and compare OccFaçade with existing state-of-the-art methods.

The IOU for each class is expressed using Formula (2):

$$IOU = \frac{TP}{TP+FP+FN} \quad (2)$$

where TP represents true positives, which are true samples predicted as true; FN represents false negatives, which are true samples predicted as false; FP represents false positives, which are false samples predicted as true; and TN represents true negatives, which are false samples predicted as false. The mIOU is the average of IOUs of all classes.

The pixel accuracy (PA) of each class is represented by Formula (3):

$$PA = \frac{TP}{TP+FN} \quad (3)$$

The average value of PA is mPA. The total pixel accuracy (Acc.) is defined as the ratio of correctly classified pixels to the total number of pixels in the image, and is calculated using Formula (4):

$$Acc. = \frac{TP+TN}{TP+TN+FP+FN} \quad (4)$$

The formula for calculating the F1 score is:

$$F1 = \frac{2TP}{2TP+FN+FP} \quad (5)$$

## 5.2. Experimental settings

The network architecture is implemented in PyTorch, and experiments are executed on an NVIDIA GeForce RTX 3090 GPU. To overcome the limited availability of building façade data, we adopt K-fold cross-validation, dividing the dataset into five parts for comprehensive experimentation. Our results are compared with existing state-of-the-art deep learning-based façade parsing methods. The optimization utilizes the Adam optimizer with an initial learning rate ( $lr$ ) of 0.0001 and a weight decay rate of 0.0001. The training spans 300 epochs, adjusting the learning rate by a constant factor  $wt$  at each epoch:

$$wt = 1000^{-\frac{1}{1000}} \quad (6)$$

To increase the randomness of the data, we apply data augmentation operations such as random scaling, cropping, flipping, etc. to the data.

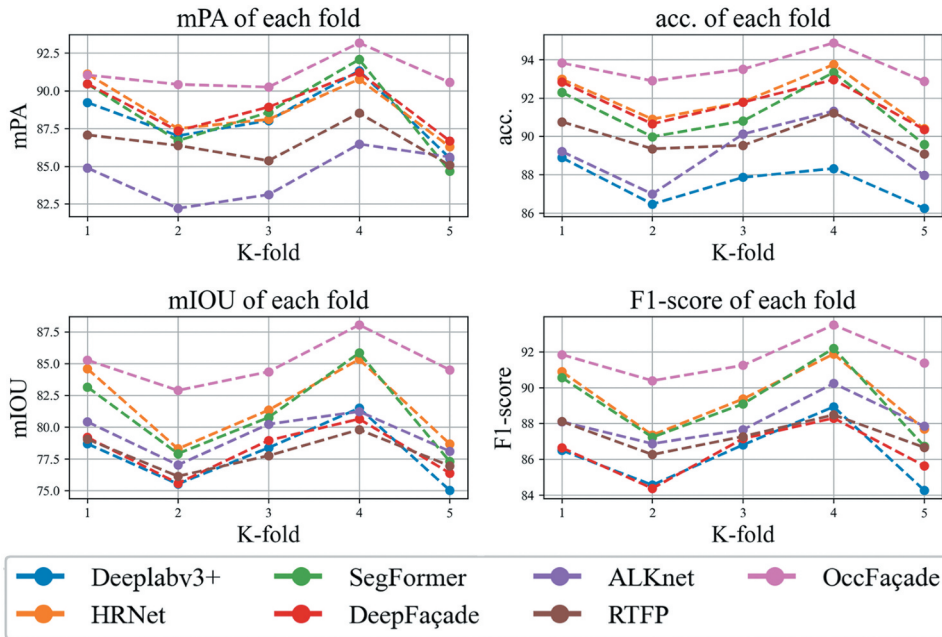
## 6. Experimental results

### 6.1. Comparison of state-of-the-art methods

We compare OccFaçade with several popular semantic segmentation models, including deeplabv3+ (Chen et al. 2018), HRNet (K. Sun et al. 2019), SegFormer (Xie et al. 2021), and three deep-learning models for façade parsing: namely DeepFaçade (H. Liu et al. 2020), ALKNet (Ma et al. 2021) and RTFP (B. Wang et al. 2024). For the ENPC2014 Dataset and ECP Dataset, we conducted five experiments using 5-fold cross-validation. Taking the ENPC2014 Dataset as an example, the overall accuracy comparison results of various deep learning methods and the proposed OccFaçade are shown in Figure 7. For ease of presentation, the experimental results for the ENPC2014 Dataset and ECP Dataset below are the averages of the 5-fold cross-validation experiments. Our experimental results show that the OccFaçade achieves state-of-the-art (SOTA) on occluded, non-occluded, multi-view data, and mesh texture.

#### 6.1.1. Results of comparison on ENPC2014 dataset

As shown in Tables 1 and 2, for the occluded ENPC2014 dataset, OccFaçade achieves almost the highest scores for both individual and overall evaluation metrics. Specifically, OccFaçade achieves 91.09% mPA, 93.60% Acc., and 85.01% mIOU. The most remarkable finding is that the mIOU of OccFaçade is at least 3% points higher than other methods. The visualization results are shown in Figure 8. We can see that OccFaçade can better predict the semantic information of the occluded region, thanks to the fact that the MD-



**Figure 7.** The mPA, acc., mIOU and F1-score of 5-fold cross-validation on the ENPC2014 dataset.

**Table 1.** The quantitative comparison results of PA for each category on the ENPC2014 dataset.

Method	Deeplabv3+	HRNet	SegFormer	DeepFaçade	ALKnet	RTFP	OccFaçade
Door	77.13±1.09	76.92±1.02	79.62±1.67	82.51±1.82	68.07±1.21	78.33±3.36	<b>82.90±1.58</b>
Shop	95.63±1.94	96.28±1.43	96.68±1.52	94.94±2.38	90.61±1.77	94.28±1.28	<b>97.12±0.48</b>
Balcony	83.93±2.11	86.81±2.98	84.23±2.78	84.15±2.78	74.18±1.85	79.45±2.82	<b>88.66±1.51</b>
Window	84.86±2.60	85.67±2.48	84.05±2.09	85.84±1.29	81.67±0.98	83.10±2.29	<b>88.36±0.74</b>
Wall	94.11±0.77	94.36±1.17	93.62±1.01	93.78±0.47	94.32±1.27	93.85±1.01	<b>95.66±0.72</b>
Sky	96.46±1.10	97.00±1.63	96.44±2.54	94.67±0.32	<b>97.78±1.42</b>	96.22±1.58	<b>97.59±1.63</b>
Roof	85.47±1.00	84.26±1.16	84.82±1.56	86.62±1.87	84.57±1.52	80.15±2.86	<b>87.37±1.20</b>

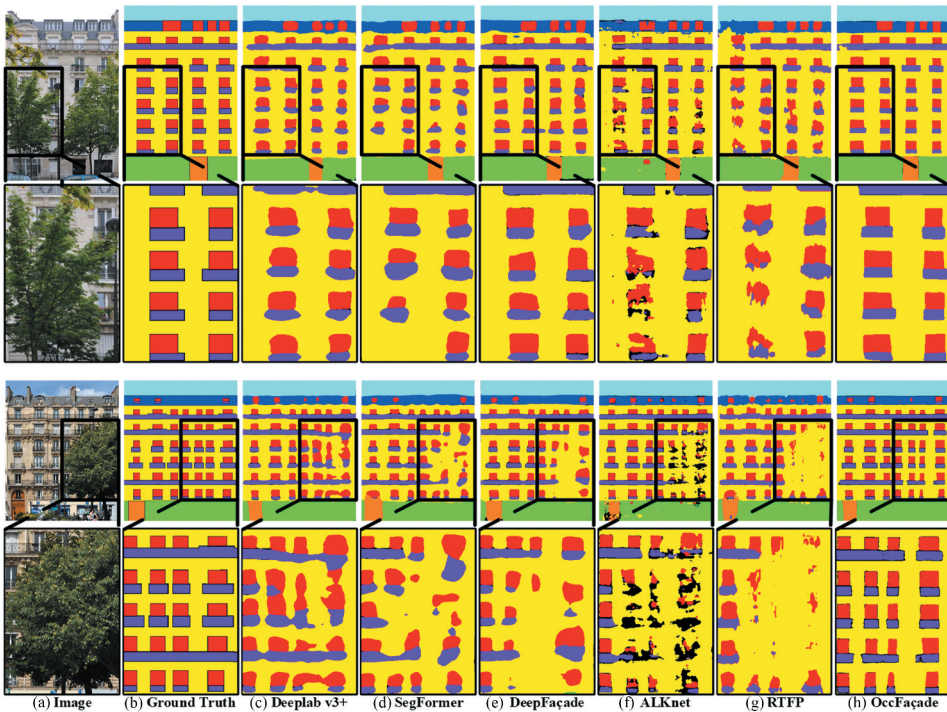
**Table 2.** The quantitative comparison results of overall metrics on the ENPC2014 dataset.

Method	Deeplabv3+	HRNet	SegFormer	DeepFaçade	ALKnet	RTFP	OccFaçade
mPA	88.23±2.20	88.76±2.10	88.50±1.94	88.93±1.95	84.46±1.76	86.48±1.25	<b>91.09±1.20</b>
Acc.	87.55±1.16	91.97±1.40	91.20±1.59	91.72±1.20	88.14±1.71	89.99±0.85	<b>93.6±0.82</b>
mIOU	77.81±2.63	81.64±1.26	80.98±1.59	78.14±2.10	79.39±1.76	77.99±1.27	<b>85.01±1.90</b>
F1	86.21±1.89	89.44±1.97	89.77±1.29	86.42±1.50	88.14±1.26	89.99±0.84	<b>91.68±1.03</b>

Module and the MRC-Module have respectively obtained local and row-column direction texture features. Furthermore, OccFaçade outperforms other methods in accurately resolving the boundaries of elements such as doors and windows and predicting their overall contours.

### 6.1.2. Results of comparison on ECP dataset

As shown in Tables 3 and 4, OccFaçade achieves the state-of-the-art (SOTA) on the ECP dataset even without occlusion. Specifically, OccFaçade achieves 91.07% mPA, 93.39% Acc., and 84.09% mIOU, which are the highest scores among all compared methods. The



**Figure 8.** Visualization comparison results of various state-of-the-art methods on ENPC2014 dataset.

visual results displayed in Figure 9 suggest that OccFaçade can neatly segment the edges of building elements, even on the ECP dataset without any occlusion issues, and it stands out as the most elegant among all the compared methods.

### 6.1.3. Results of comparison on RueMonge2014 dataset

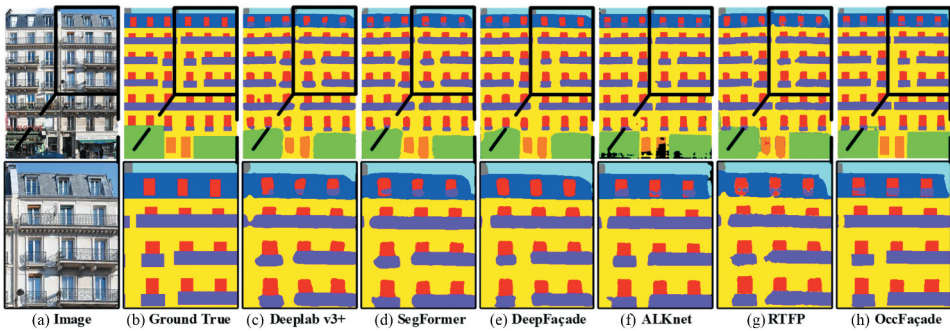
Besides the occluded and non-occluded datasets, we evaluate OccFaçade against state-of-the-art façade parsing approaches on a multi-view dataset. As shown in Tables 5 and 6,

**Table 3.** The quantitative comparison results of PA for each category on the ECP dataset.

Method	Deeplabv3+	HRNet	SegFormer	DeepFaçade	ALKnet	RTFP	OccFaçade
Window	85.04±2.22	86.52±1.69	86.12±1.62	<b>87.46±2.00</b>	81.81±2.71	84.01±1.40	86.83±0.82
Wall	94.32±0.30	94.69±0.85	94.35±0.43	94.72±0.28	<b>96.30±0.50</b>	94.81±0.23	94.89±0.36
Balcony	90.89±1.17	91.75±1.72	90.71±2.56	91.07±0.42	90.07±2.13	89.97±2.77	<b>92.11±2.52</b>
Door	79.66±1.94	82.50±1.63	80.17±1.38	82.10±1.51	71.26±1.73	80.42±1.97	<b>82.75±1.27</b>
Roof	90.85±1.89	91.19±1.80	90.13±1.59	<b>92.02±0.51</b>	90.28±1.31	86.74±1.91	90.99±1.57
Sky	93.19±1.11	94.90±1.99	95.60±0.85	95.62±0.37	92.75±2.22	95.14±1.23	<b>95.96±1.16</b>
Shop	95.45±1.11	95.22±1.88	<b>96.35±1.71</b>	95.26±1.63	85.53±1.86	93.37±1.74	96.25±1.58
Chimney	82.80±1.41	85.70±1.98	84.39±1.10	81.09±1.55	<b>94.76±3.20</b>	80.30±1.56	88.86±1.47

**Table 4.** The quantitative comparison results of overall metrics on the ECP dataset.

Method	Deeplabv3+	HRNet	SegFormer	DeepFaçade	ALKnet	RTFP	OccFaçade
mPA	89.03±1.16	90.31±1.10	89.73±0.74	89.92±0.67	87.85±1.02	88.09±1.48	<b>91.07±1.28</b>
Acc.	89.52±0.87	92.98±0.72	92.71±0.67	92.37±0.34	92.44±0.82	91.93±0.89	<b>93.39±0.94</b>
mIOU	79.92±1.12	82.68±1.60	82.28±0.71	80.09±1.27	81.84±1.35	79.50±1.60	<b>84.09±1.84</b>
F1	87.53±0.69	90.36±0.96	90.11±0.48	89.00±0.71	89.79±0.79	88.41±1.00	<b>91.22±1.10</b>



**Figure 9.** Visualization comparison results of various state-of-the-art methods on ECP dataset.

mPA, Acc., and mIOU of the proposed method are very similar to those of SegFormer, and almost achieve the highest values. Among them, our mIOU achieves the highest score of 72.95% among all methods. Due to the RueMonge2014 dataset being a multi-view data without image rectification, where elements such as doors and windows are not neatly arranged in the image, MD-Module and MRC-Module cannot achieve their best performance. Nevertheless, OccFaçade still produces pleasant parsing results, as demonstrated in Figure 10.

#### 6.1.4. Results of comparison on MeshFaçade dataset

As shown in Figure 11, due to the quality of Mesh data, there are some holes and missing problems in the building façade of MeshFaçade. These problems can be viewed as building occlusion problems, which can be effectively solved by our proposed OccFaçade. In terms of visual comparison, OccFaçade can still segment

**Table 5.** The quantitative comparison results of PA for each category on the RueMonge2014 dataset.

Method	Deeplabv3+	HRNet	SegFormer	DeepFaçade	ALKnet	RTFP	OccFaçade
Window	68.91±0.89	77.43±0.12	74.96±1.21	81.28±2.01	82.77±1.14	72.16±1.44	<b>89.91±0.93</b>
Wall	80.13±1.25	91.05±0.56	90.41±1.63	79.73±1.22	85.44±1.85	89.69±1.64	<b>92.88±1.08</b>
Balcony	77.01±1.58	<b>87.66±1.52</b>	87.33±0.85	75.43±1.37	75.33±1.63	79.32±1.09	83.55±1.41
Door	55.08±1.41	56.45±1.14	60.79±1.64	62.10±1.45	49.43±0.59	58.74±0.96	<b>62.21±1.33</b>
Roof	70.87±1.03	85.03±1.97	87.44±1.77	83.76±0.86	82.99±0.87	77.21±0.87	<b>87.89±0.56</b>
Sky	80.66±1.11	93.51±1.23	94.10±1.44	84.08±0.47	85.53±1.87	<b>95.77±1.96</b>	90.17±0.87
Shop	80.76±0.96	88.41±0.88	<b>92.09±0.55</b>	73.41±0.97	74.86±0.74	83.99±1.21	77.26±0.72

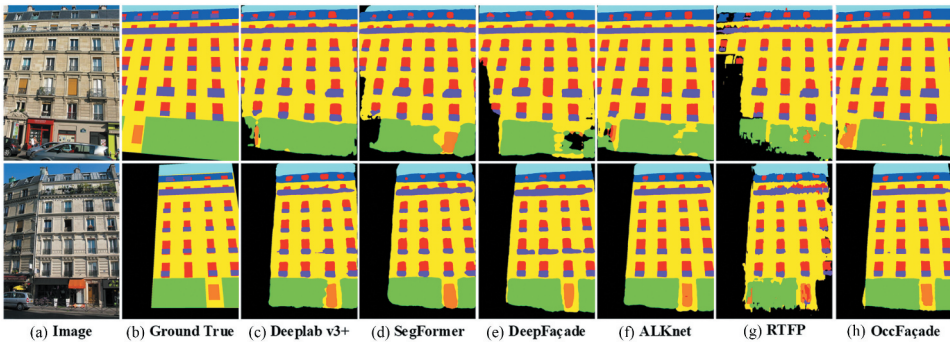
**Table 6.** The quantitative comparison results of overall metrics on the RueMonge2014 dataset.

Method	Deeplabv3+	HRNet	SegFormer	DeepFaçade	ALKnet	RTFP	OccFaçade
mPA	73.35±0.65	82.79±0.74	<b>83.87±0.77</b>	77.11±0.99	76.62±1.01	79.55±0.98	83.41±0.61
Acc.	81.30±0.74	87.84±0.54	<b>87.95±0.59</b>	79.53±0.93	82.38±0.54	84.73±1.27	87.64±0.71
mIOU	66.76±0.67	72.40±0.61	72.75±0.93	58.97±0.47	68.37±0.67	66.45±0.65	<b>72.95±0.69</b>
F1	70.23±0.69	83.33±0.72	83.61±0.74	70.03±0.61	80.62±0.82	79.01±0.81	<b>86.43±0.47</b>

**Table 7.** The quantitative comparison results of PA for each category on the MeshFaçade dataset.

Method	Deeplabv3+	SegFormer	DeepFaçade	ALKnet	RTFP	OccFaçade
Wall	91.85±0.32	<b>96.77±0.61</b>	92.31±0.74	94.63±0.14	94.46±0.62	96.41±0.44
Window	86.87±0.74	87.09±0.47	86.82±0.66	88.42±0.71	<b>91.74±0.41</b>	91.22±0.65

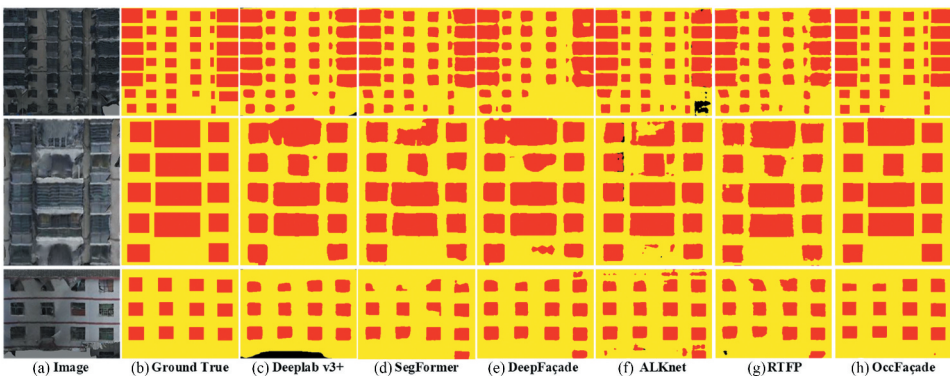




**Figure 10.** Visualization comparison results of various state-of-the-art deep learning methods on RueMonge2014 dataset.

complete and nearly rectangular windows with almost no noise. However, it is difficult for other methods to completely restore semantic information where the image is missing. Comparing the classification accuracy, OccFaçade achieves the highest scores of 93.82%, 94.36 and 88.83% on the three overall evaluation indicators of mPA, Acc., and mIOU respectively, as shown in Table 8. The accuracy for each category of walls and windows also achieves good results as shown in Table 7.

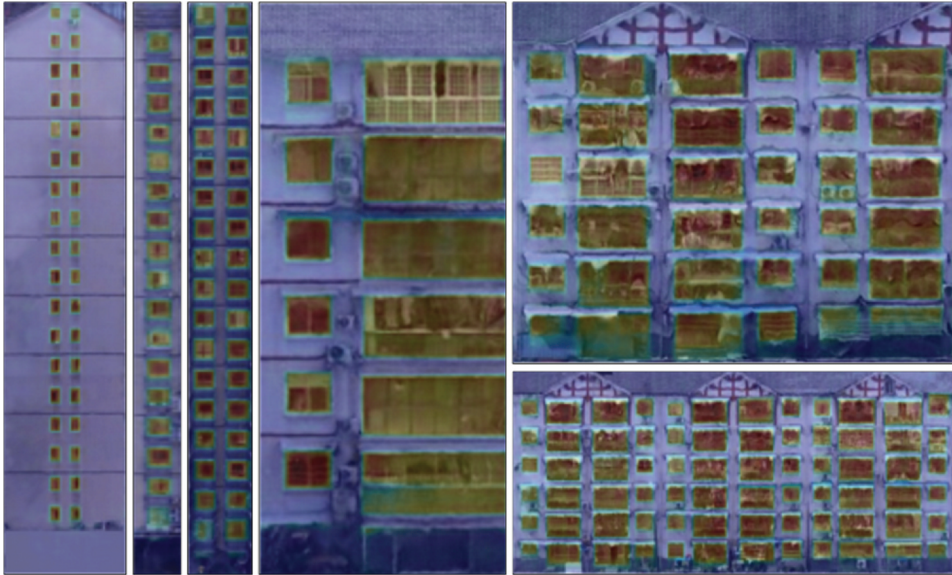
In Figure 12, we visualize the Gradient-weighted Class Activation Mapping (Grad CAM) (Selvaraju et al. 2017) on the MeshFaçade, illustrating the superior feature-extracting capabilities of OccFaçade. The left side of Figure 12 shows some higher-floor building façades. Due to resolution limitations, the windows appear as smaller targets on the façade image, but OccFaçade can capture the characteristics of every small target very well. The right side of Figure 12 shows some building façades that are not very high but have serious textures. OccFaçade can still extract relatively complete window features.



**Figure 11.** Visualization comparison results of various state-of-the-art deep learning methods on MeshFaçade dataset.

**Table 8.** The quantitative comparison results of overall metrics on the MeshFaçade dataset.

Method	Deeplabv3+	SegFormer	DeepFaçade	ALKnet	RTFP	OccFaçade
mPA	89.36±0.66	91.93±0.47	89.57±0.77	91.52±0.57	93.10±1.33	<b>93.82±0.66</b>
Acc.	92.16±0.14	92.16±1.08	91.34±1.28	92.17±1.49	93.41±0.57	<b>94.36±0.71</b>
mIOU	82.99±0.67	86.12±1.01	82.26±0.44	85.51±0.54	87.06±0.63	<b>88.83±0.62</b>
F1	89.25±0.59	92.51±0.63	89.12±0.61	92.17±0.81	93.06±0.95	<b>94.07±0.54</b>

**Figure 12.** Visualization of grad CAM on the MeshFaçade dataset.

## 6.2. Ablation study

To verify the effectiveness of our designed MD-Module and MRC-Module in resolving the occlusion, we conduct ablation experiments on the ENPC2014 dataset. Our baseline is the U-Net network with an encoder of Vgg trained in ImageNet. +MD-Module means that only the MD-Module is used based on the baseline. +MD+MRC means that the MD-Module is used after the encoder and only one MRC-Module is used in the decoder. +MD+MRCs (OccFaçade) means that the MD-Module is used after the encoder and the MRC-Module is used in each stage of upsampling of the decoder, and the network structure is shown in Figure 2.

As demonstrated in Tables 9 and 10, the addition of the MD-Module increases mPA by 1.65% points and mIOU by 2.98% points. After adding a single MRC-Module, the segmentation accuracy of each element of the building façade has been significantly improved. However, as can be seen from Figure 13, the boundary details of elements such as doors and windows are not very good. Each sample on the decoder After adding MRC-Module to the process, both segmentation accuracy and visualization effects have been further improved. Therefore, OccFaçade adopts the +MD+MRCs strategy to achieve an mPA of 93.60 and an mIOU of 85.01. In terms of visual effects, it can not only predict the building features behind the occlusion texture but also achieve segmentation with richer details. The increase in

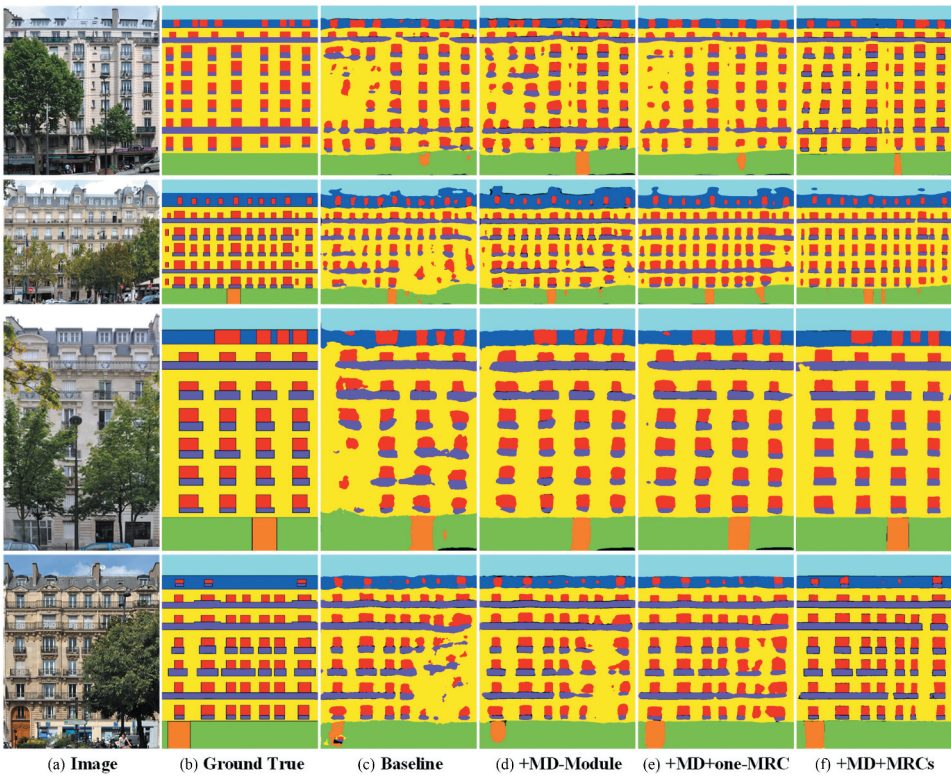


**Table 9.** mPA of OccFaçade ablation study on ENPC2014 dataset.

Model	Fold1	Fold2	Fold3	Fold4	Fold5	mean
Baseline	92.05	88.96	89.41	91.41	88.30	90.03
+MD-Module	91.91	91.29	91.58	92.99	90.61	91.68
+MD+one-MRC	92.48	90.85	92.14	92.87	91.01	91.87
<b>+MD+MRCs</b>	<b>93.83</b>	<b>92.90</b>	<b>93.51</b>	<b>94.88</b>	<b>92.87</b>	<b>93.60</b>

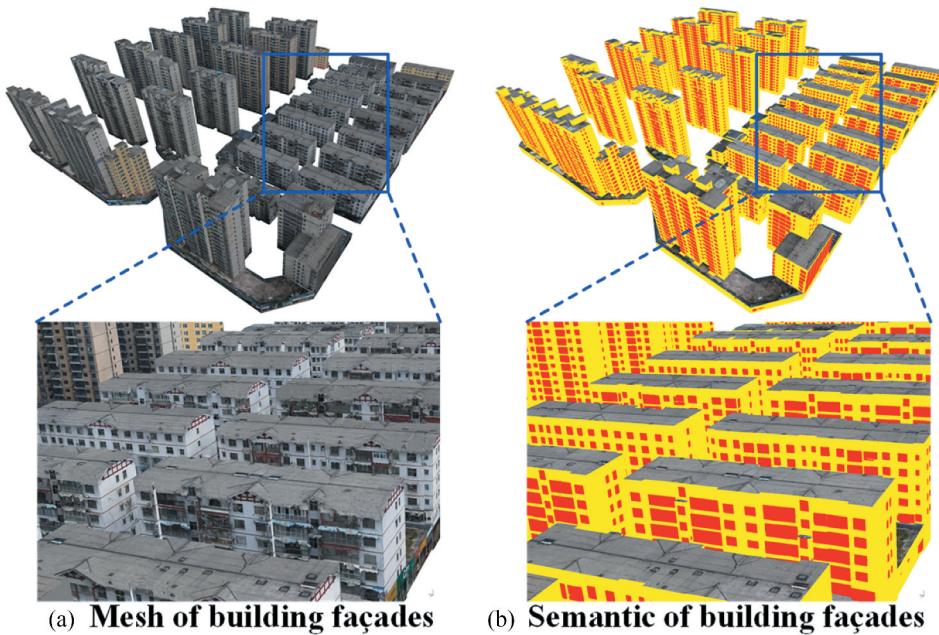
**Table 10.** mIOU of OccFaçade ablation study on ENPC2014 dataset.

Model	Fold1	Fold2	Fold3	Fold4	Fold5	mean
Baseline	81.85	75.46	77.65	80.40	75.36	78.14
+MD-Module	81.96	79.40	80.60	83.73	79.91	81.12
+MD+one-MRC	82.46	80.46	82.46	84.40	80.59	82.07
<b>+MD+MRCs</b>	<b>85.27</b>	<b>82.88</b>	<b>84.35</b>	<b>88.04</b>	<b>84.50</b>	<b>85.01</b>



**Figure 13.** Visualization results of ablation experiments on the ENPC2014 dataset.

metrics and enhancement of visual effects both testify to the effectiveness of our designed MD-Module and MRC-Module in addressing the issue of occlusion in building façade parsing.



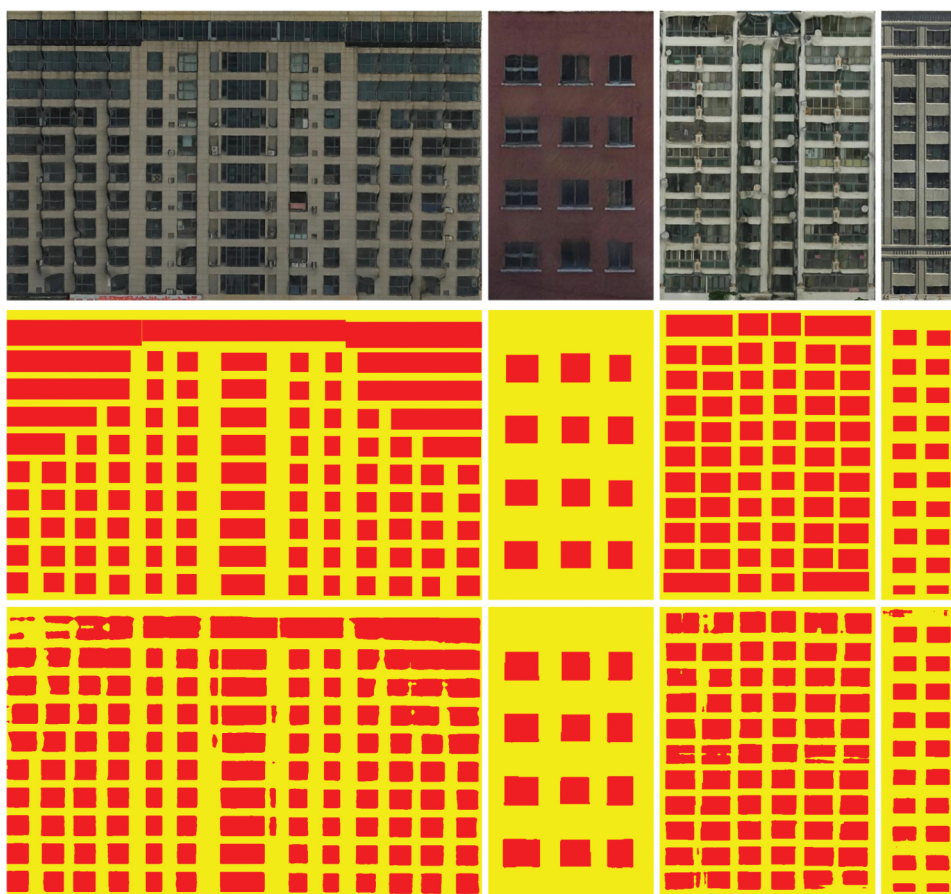
**Figure 14.** Semantic information parsing of building façades in larger urban scenes. (a) illustrates the projection of the mesh model of the building onto LoD2, while (b) illustrates the building façade parsing results obtained by OccFaçade.

### **6.3. Building façade parsing in large urban scenes**

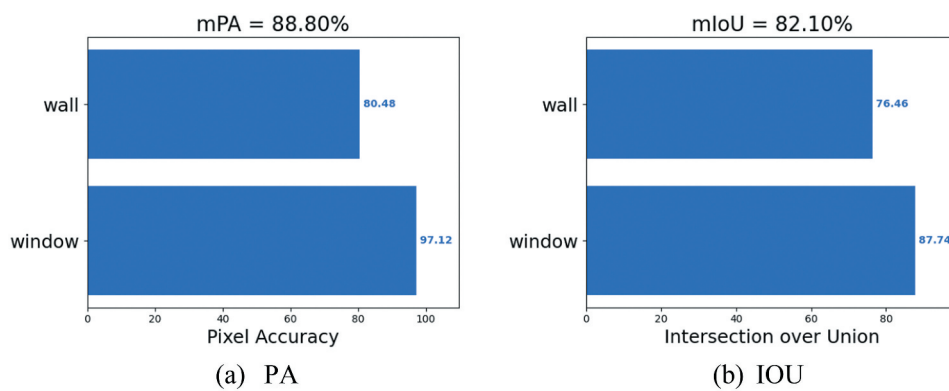
Through comparative experiments on existing public datasets and the MeshFaçade dataset, this paper determines that OccFaçade achieves state-of-the-art performance in building appearance parsing tasks. In addition, the introduced MeshFaçade dataset allows the extraction of building façades without a correction process, providing a strong data basis for parsing building façades in a wide range of urban scenes in a short time. As shown in Figure 14, by replacing the semantic information obtained by OccFaçade façade analysis with the building LoD2 model, the three-dimensional semantic information of the entire urban area can be obtained.

### **6.4. Evaluation of transferability of OccFaçade in MeshFaçade**

We use the OccFaçade trained on the MeshFaçade dataset to parse building façades in other cities to explore the transferability of OccFaçade in the MeshFaçade dataset. As shown in Figure 15, Even for building façades outside the MeshFaçade dataset, OccFaçade can achieve pleasant segmentation results. Compared with the MeshFaçade dataset, the mPA and mIOU of the transferability experiment only dropped by about 5 points as shown in Figure 16. The transferability experiment thoroughly demonstrates that our method and dataset exhibit strong adaptability in large-scale urban buildings. Furthermore, it proves our ability to achieve high-precision façade analysis even with limited annotated data.



**Figure 15.** Visualization results of transferability experiment on other urban mesh textures. The top row is the original building façade image, the middle row is the ground truth, and the bottom row is the predicted result.



**Figure 16.** Quantitative evaluation results of transferability experiment on other urban mesh textures.

**Table 11.** The time cost of each method.

Method	Deeplabv3+	HRNet	SegFormer	DeepFaçade	ALK	RTFP	OccFaçade
Time(s)	<b>0.0073</b>	0.0583	0.0125	0.0082	0.0848	0.5764	0.0080

### 6.5. Time cost

To better evaluate the performance of the OccFaçade algorithm, we measured the average time taken by OccFaçade and other methods to segment a single building façade image. The results are shown in Table 11. The shortest time cost is from the lightweight Deeplabv3+, at 0.0073 seconds. Our OccFaçade has a time cost of 0.0080 seconds, ranking second and demonstrating good performance.

## 7. Conclusion

To address the occlusion challenges in the parsing of large urban building façades, OccFaçade is introduced. This architecture consists two modules: MD-Module and MRC-Module, specifically designed to capture the local and row-column directional repeated texture features on building façades. Leveraging the inherent repetitive structural characteristics of buildings, OccFaçade excels in high-precision parsing of occluded regions. Through extensive comparative experiments on publicly available datasets, including ENPC2014, ECP, RueMonge2014, and the newly proposed MeshFaçade dataset, OccFaçade achieves state-of-the-art, and it can generate more regular and textured edges at the edges of building components such as doors and windows.

Simultaneously, a dataset named MeshFaçade is introduced, acquired by projecting mesh data onto building façades. Different from traditional image correction approaches, this dataset relies on automated modelling to quickly obtain building façade data in extensive urban scenes. It provides a novel and valuable resource for various applications, including building façade parsing.

## 8. Limitations

The MRC-Module of OccFaçade can capture features in the row and column directions of building façades, which is beneficial for extracting semantic information behind occluded areas. However, the strategy requires the façade images to be rectified to achieve horizontally and vertically aligned images to fully leverage the advantages of OccFaçade. The comparative results on the RueMonge2014 Dataset and the ENPC2014 Dataset have already demonstrated this.

## Acknowledgement

The work was supported by the National Natural Science Foundation of China (Grant No. 42201474, 42192581).

## Disclosure statement

No potential conflict of interest was reported by the author(s).

## ORCID

Dongdong Yue  <http://orcid.org/0009-0009-4477-1150>

Xinyi Liu  <http://orcid.org/0000-0001-5333-8054>

## References

- Carion, N., F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko. 2020. "End-To-End Object Detection with Transformers." *arXiv*. <http://arxiv.org/abs/2005.12872>.
- Chen, L., Z. Yu, G. Papandreou, F. Schroff, and H. Adam. 2018. "Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation." *Computer Vision – ECCV*: 833–851. [https://doi.org/10.1007/978-3-030-01234-2\\_49](https://doi.org/10.1007/978-3-030-01234-2_49).
- Cohen, A., M. R. Oswald, Y. Liu, and M. Pollefeys. 2017. "Symmetry-Aware Façade Parsing with Occlusions." *International Conference On 3D Vision (3DV)* 393–401. <https://doi.org/10.1109/3DV.2017.00052>.
- Cohen, A., A. G. Schwing, and M. Pollefeys. 2014. "Efficient Structured Parsing of Facades Using Dynamic Programming." *2014 IEEE Conference on Computer Vision and Pattern Recognition*, 3206–3213. <https://doi.org/10.1109/CVPR.2014.410>.
- Deng, J., W. Dong, R. Socher, L. J. Li, K. Li, and F. Li. 2009. "ImageNet: A Large-Scale Hierarchical Image Database." *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 248–255. <https://doi.org/10.1109/CVPR.2009.5206848>.
- Dosovitskiy, A., L. Beyer, A. Kolesnikov, D. Weissenborn, X. H. Zhai, T. Unterthiner, M. Dehghani, et al. 2021. "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale." *arXiv*. <http://arxiv.org/abs/2010.11929>.
- Fu, J., J. Liu, H. Tian, and Y. Li. 2019. "Dual Attention Network for Scene Segmentation." *IEEE Conference on Computer Vision and Pattern Recognition*, 3146–3154. <https://doi.org/10.1109/CVPR.2019.00326>.
- Gadde, R., V. Jampani, R. Marlet, and P. V. Gehler. 2018. "Efficient 2D and 3D Facade Segmentation Using Auto-Context." *IEEE Transactions on Pattern Analysis & Machine Intelligence* 40 (5): 1273–1280. <https://doi.org/10.1109/TPAMI.2017.2696526>.
- Gadde, R., R. Marlet, and N. Paragios. 2016. "Learning Grammars for Architecture-Specific Facade Parsing." *International Journal of Computer Vision* 117 (3): 290–316. <https://doi.org/10.1007/s11263-016-0887-4>.
- Geiger, A., P. Lenz, C. Stiller, and R. Urtasun. 2013. "Vision Meets Robotics: The KITTI Dataset." *The International Journal of Robotics Research* 32 (11): 1231–1237. <https://doi.org/10.1177/0278364913491297>.
- Girshick, R., J. Donahue, T. Darrell, and J. Malik. 2014. "Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation." *IEEE Conference on Computer Vision and Pattern Recognition*, 580–587. <https://doi.org/10.1109/CVPR.2014.81>.
- Gonzalez-Aguilera, D., E. Crespo-Matellan, D. Hernandez-Lopez, and P. Rodriguez-Gonzalvez. 2013. "Automated Urban Analysis Based on LiDAR-Derived Building Models." *IEEE Transactions on Geoscience & Remote Sensing* 51 (3): 1844–1851. <https://doi.org/10.1109/TGRS.2012.2205931>.
- Gröger, G., and L. Plümer. 2012. "CityGML – Interoperable Semantic 3D City Models." *Isprs Journal of Photogrammetry & Remote Sensing* 71 (July): 12–33. <https://doi.org/10.1016/j.isprsjprs.2012.04.004>.
- Han, F., and S. Zhu. 2009. "Bottom-Up/top-Down Image Parsing with Attribute Grammar." *IEEE Transactions on Pattern Analysis & Machine Intelligence* 31 (1): 59–73. <https://doi.org/10.1109/TPAMI.2008.65>.



- He, K., G. Gkioxari, P. Dollár, and R. Girshick. 2018. "Mask R-CNN." *arXiv*. <http://arxiv.org/abs/1703.06870>.
- Hu, H., B. Feng, B. Xu, Q. Zhu, X. Ge, and M. Chen. 2022. "Efficient Procedural Modelling of Building Façades Based on Windows from Sketches." *Photogrammetric Record* 37 (179): 333–353. <https://doi.org/10.1111/phor.12425>.
- Kong, G., and H. Fan. 2021. "Enhanced Facade Parsing for Street-Level Images Using Convolutional Neural Networks." *IEEE Transactions on Geoscience & Remote Sensing* 59 (12): 10519–10531. <https://doi.org/10.1109/TGRS.2020.3035878>.
- Koutsourakis, P., L. Simon, O. Teboul, G. Tziritas, and N. Paragios. 2009. "Single View Reconstruction Using Shape Grammars for Urban Environments." *IEEE 12th International Conference on Computer Vision*, 1795–1802. <https://doi.org/10.1109/ICCV.2009.5459400>.
- Kozinski, M., R. Gadde, S. Zagoruyko, G. Obozinski, and R. Marlet. 2015. "A MRF Shape Prior for Facade Parsing with Occlusions." *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2820–2828. <https://doi.org/10.1109/CVPR.2015.7298899>.
- Li, C., H. Zhang, J. Liu, Y. Zhang, S. Zou, and Y. Fang. 2020. "Window Detection in Facades Using Heatmap Fusion." *Journal of Computer Science and Technology* 35 (4): 900–912. <https://doi.org/10.1007/s11390-020-0253-4>.
- Li, C., Y. Zhang, and Z. Zhang. 2016. "Automatic Keyline Recognition and 3D Reconstruction for Quasi-Planar Façades in Close-Range Images." *Photogrammetric Record* 31 (153): 29–50. <https://doi.org/10.1111/phor.12141>.
- Li, W., and R. Tao. 2022. "Graph-Feature-Enhanced Selective Assignment Network for Hyperspectral and Multispectral Data Classification." *IEEE Transactions on Geoscience & Remote Sensing* 60:1–4. <https://doi.org/10.1109/TGRS.2022.3166252>.
- Liu, H., Y. Xu, J. Zhang, J. Zhu, Y. Li, and S. C. H. Hoi. 2020. "DeepFacade: A Deep Learning Approach to Facade Parsing with Symmetric Loss." *IEEE Transactions on Multimedia* 22 (12): 3153–3165. <https://doi.org/10.1109/TMM.2020.2971431>.
- Liu, X., Y. Zhang, X. Ling, Y. Wan, L. Liu, and Q. Li. 2019. "TopoLAP: Topology Recovery for Building Reconstruction by Deducing the Relationships Between Linear and Planar Primitives." *Remote Sensing* 11 (11): 1372. <https://doi.org/10.3390/rs11111372>.
- Liu, X., X. Zhu, Y. Zhang, S. Wang, and C. Jia. 2023. "Generation of Concise 3D Building Model from Dense Meshes by Extracting and Completing Planar Primitives." *Photogrammetric Record* 38 (181): 22–46. <https://doi.org/10.1111/phor.12438>.
- Long, J., E. Shelhamer, and T. Darrell. 2015. "Fully Convolutional Networks for Semantic Segmentation." *IEEE Conference on Computer Vision and Pattern Recognition*, 3431–3440. <https://doi.org/10.1109/CVPR.2015.7298965>.
- Ma, W., W. Ma, S. Xu, and H. Zha. 2021. "Pyramid ALKNet for Semantic Parsing of Building Facade Image." *IEEE Geoscience & Remote Sensing Letters* 18 (6): 1009–1013. <https://doi.org/10.1109/LGRS.2020.2993451>.
- Mathias, M., A. Martinović, and L. V. Gool. 2016. "ATLAS: A Three-Layered Approach to Facade Parsing." *International Journal of Computer Vision* 118 (1): 22–48. <https://doi.org/10.1007/s11263-015-0868-z>.
- Mei, H., X. Yang, Y. Wang, Y. Liu, S. He, Q. Zhang, X. Wei, and R. W. H. Lau. 2020. "Don't Hit Me! Glass Detection in Real-World Scenes." *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 3684–3693. <https://doi.org/10.1109/CVPR42600.2020.00374>.
- Nordmark, N., and M. Ayenew. 2021. "Window Detection in Facade Imagery: A Deep Learning Approach Using Mask R-CNN." *arXiv*. <http://arxiv.org/abs/2107.10006>.
- Peng, C., X. Zhang, G. Yu, G. Luo, and J. Sun. 2017. "Large Kernel Matters — Improve Semantic Segmentation by Global Convolutional Network." *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1743–1751. <https://doi.org/10.1109/CVPR.2017.189>.
- Redmon, J., S. Divvala, R. Girshick, and A. Farhadi. 2016. "You Only Look Once: Unified, Real-Time Object Detection." *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 779–788. <https://doi.org/10.1109/CVPR.2016.91>.

- Riemenschneider, H., A. Bódis-Szomorú, J. Weissenberg, and L. Van Gool. 2014. "Learning Where to Classify in Multi-View Semantic Segmentation." *Computer Vision – ECCV 2014* 516–532. [https://doi.org/10.1007/978-3-319-10602-1\\_34](https://doi.org/10.1007/978-3-319-10602-1_34).
- Ripperda, N., and C. Brenner. 2006. "Reconstruction of Façade Structures Using a Formal Grammar and RjMCMC." *Pattern Recognition* 4174:750–759. [https://doi.org/10.1007/11861898\\_75](https://doi.org/10.1007/11861898_75).
- Ronneberger, O., P. Fischer, and T. Brox. 2015. "U-Net: Convolutional Networks for Biomedical Image Segmentation." *arXiv*. <http://arxiv.org/abs/1505.04597>.
- Selvaraju, R. R., M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. 2017. "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization." *IEEE international conference on computer vision (ICCV)*, Venice, Italy, 618–626. <https://doi.org/10.1109/ICCV.2017.74>.
- Simonyan, K., and A. Zisserman. 2015. "Very Deep Convolutional Networks for Large-Scale Image Recognition." *arXiv*. <http://arxiv.org/abs/1409.1556>.
- Sun, K., B. Xiao, D. Liu, and J. Wang. 2019. "Deep High-Resolution Representation Learning for Human Pose Estimation." *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 5686–5696. <https://doi.org/10.1109/CVPR.2019.00584>.
- Sun, Y., S. Malihi, H. Li, and M. Maboudi. 2022. "DeepWindows: Windows Instance Segmentation Through an Improved Mask R-CNN Using Spatial Attention and Relation Modules." *ISPRS International Journal of Geo-Information* 11 (3): 162. <https://doi.org/10.3390/ijgi11030162>.
- Tao, Y., Y. Zhang, and X. Chen. 2022. "Element-Arrangement Context Network for Facade Parsing." *Journal of Computer Science and Technology* 37 (3): 652–665. <https://doi.org/10.1007/s11390-022-2189-3>.
- Teboul, O., I. Kokkinos, L. Simon, P. Koutsourakis, and N. Paragios. 2011. "Shape Grammar Parsing via Reinforcement Learning." *Cvpr 2011*: 2273–2280. <https://doi.org/10.1109/CVPR.2011.5995319>.
- Teboul, O., L. Simon, P. Koutsourakis, and N. Paragios. 2010. "Segmentation of Building Facades Using Procedural Shape Priors." *IEEE Computer Society Conference on Computer Vision and Pattern Recognition* 3105–3112. <https://doi.org/10.1109/CVPR.2010.5540068>.
- Tyleček, R., and R. Šára. 2013. "Spatial Pattern Templates for Recognition of Objects with Regular Structure." *Pattern Recognition* 8142:364–374. [https://doi.org/10.1007/978-3-642-40602-7\\_39](https://doi.org/10.1007/978-3-642-40602-7_39).
- Wang, B., J. Zhang, R. Zhang, Y. Li, L. Li, and Y. Nakashima. 2024. "Improving Facade Parsing with Vision Transformers and Line Integration." *Advanced Engineering Informatics* 60 (April): 102463. <https://doi.org/10.1016/j.aei.2024.102463>.
- Wang, J., M. Zhang, W. Li, and R. Tao. 2024. "A Multistage Information Complementary Fusion Network Based on Flexible-Mixup for HSI-X Image Classification." *IEEE Transactions on Neural Networks and Learning Systems* 1–13. <https://doi.org/10.1109/TNNLS.2023.3300903>.
- Wendel, A., M. Donoser, and H. Bischof. 2010. "Unsupervised Facade Segmentation Using Repetitive Patterns." *Pattern Recognition* 6376:51–60. [https://doi.org/10.1007/978-3-642-15986-2\\_6](https://doi.org/10.1007/978-3-642-15986-2_6).
- Xie, E., W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo. 2021. "SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers." In *Advances in Neural Information Processing Systems*, edited by M. Ranzato, A. Beygelzimer, Y. Dauphin, P. S. Liang, and J. Wortman Vaughan, 12077–12090. Vol. 34. Curran Associates, Inc. [https://proceedings.neurips.cc/paper\\_files/paper/2021/file/64f1f27bf1b4ec22924fd0acb550c235-Paper.pdf](https://proceedings.neurips.cc/paper_files/paper/2021/file/64f1f27bf1b4ec22924fd0acb550c235-Paper.pdf).
- Zhang, G., Y. Pan, and L. Zhang. 2022. "Deep Learning for Detecting Building Façade Elements from Images Considering Prior Knowledge." *Automation in Construction* 133 (January): 104016. <https://doi.org/10.1016/j.autcon.2021.104016>.
- Zhang, M., W. Li, X. Zhao, H. Liu, R. Tao, and Q. Du. 2023. "Morphological Transformation and Spatial-Logical Aggregation for Tree Species Classification Using Hyperspectral Imagery." *IEEE Transactions on Geoscience & Remote Sensing* 61:1–12. <https://doi.org/10.1109/TGRS.2022.3233847>.
- Zhao, H., J. Shi, X. Qi, X. Wang, and J. Jia. 2017. "Pyramid Scene Parsing Network." *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 6230–6239. <https://doi.org/10.1109/CVPR.2017.660>.
- Zhao, P., T. Fang, J. Xiao, H. Zhang, Q. Zhao, and L. Quan. 2010. "Rectilinear Parsing of Architecture in Urban Environment." *IEEE Computer Society Conference on Computer Vision and Pattern Recognition* 342–349. <https://doi.org/10.1109/CVPR.2010.5540192>.