International Journal of Geo-Information

MDPI

Article

DeepWindows: Windows Instance Segmentation through an Improved Mask R-CNN Using Spatial Attention and Relation Modules

Yanwei Sun [1], Shirin Malihi [2], Hao Li [1,*] and Mehdi Maboudi [3]

[1] School of Earth Sciences and Engineering, Hohai University, Nanjing 211100, China; sunyw0108@hhu.edu.cn
[2] Tandon School of Engineering, New York University, New York, NY 11201, USA; sh.malihi@kntu.ac.ir
[3] Institute of Geodesy and Photogrammetry, Technische Universitaet Braunschweig, 38106 Braunschweig, Germany; m.maboudi@tu-bs.de
* Correspondence: lihao@hhu.edu.cn

Abstract: Windows, as key components of building facades, have received increasing attention in facade parsing. Convolutional neural networks have shown promising results in window extraction. Most existing methods segment a facade into semantic categories and subsequently employ regularization based on the structure of manmade architectures. These methods merely concern the optimization of individual windows, without considering the spatial areas or relationships of windows. This paper presents a novel windows instance segmentation method based on Mask R-CNN architecture. The method features a spatial attention region proposal network and a relation module-enhanced head network. First, an attention module is introduced in the region proposal network to generate a spatial attention map, then the attention map is multiplied with the objectness scores of the classification branch. Second, for the head network, relation modules are added to model the spatial relationships between proposals. Appearance and geometric features are combined for instance recognition. Furthermore, we constructed a new window instance segmentation dataset with 1200 annotated images. With our dataset, the average precisions of our method on detection and segmentation increased from 53.1% and 53.7% to 56.4% and 56.7% compared with Mask R-CNN. A comparison with state-of-the-art methods also proves the predominance of our proposed method.

Keywords: windows; instance segmentation; spatial attention; relation module; Mask R-CNN

Citation: Sun, Y.; Malihi, S.; Li, H.; Maboudi, M. DeepWindows: Windows Instance Segmentation through an Improved Mask R-CNN Using Spatial Attention and Relation Modules. *ISPRS Int. J. Geo-Inf.* **2022**, *11*, 162. https://doi.org/10.3390/ijgi11030162

Academic Editor: Wolfgang Kainz

Received: 30 December 2021
Accepted: 20 February 2022
Published: 23 February 2022

1. Introduction

The three-dimensional (3-D) reconstruction of buildings has become an important research topic during the last 2 decades [1]. With a growing demand on the high Level-of-Detail (LoD) of building models [2], the detailed geometry of buildings and the semantics of their facade elements are both important. Windows are the most important elements of building facades. Window detection and segmentation have attracted a wide range of research interest in different applications, such as thermal inspections [3] and flood risk assessments [4]. In this paper, we address the research problem of windows instance segmentation from frontal facade images (see Figure 1). The accurate extraction of windows is challenging owing to the complexity of buildings in real scenes [5,6]. Specifically, the diversity of building styles usually results in a variety of window geometries. Facade decorations that look similar to windows may cause a false detection. Glass reflections and illumination changes also significantly impact the appearance of windows. In addition, the damaged furnishing materials of facades increase the diversity of facade textures and the difficulty of window recognition.

Images and point clouds are two widely used data types in window extraction. Point-cloud-based methods are normally based on the hypothesis that the most prominent features of facade components are planar [7]. This requirement is difficult to achieve for

some building styles. This paper focuses on image-based approaches. In past decades, hand-crafted based methods were dominant in facade semantic segmentation. Based on the repetitive and symmetric structures, grammar-based methods and pattern recognition algorithms are widely studied [8–15]. In recent years, however, deep learning methods have been introduced and applied in various application domains [16,17]. For image-processing applications, Convolutional Neural Networks (CNNs) show a powerful capability in image segmentation and object detection [18–21]. Unlike traditional methods, deep learning methods can deal with facades without strict structures. A number of CNN-based approaches have been proposed for facade segmentation [22–26] and window detection [27,28]. However, these methods only regard each window as an individual component. Although it is well believed that the modelling of spatial locations and relations will help object detection and segmentation, few researchers have applied this idea in window extraction.

Figure 1. Windows instance segmentation from facades.

In this paper, we propose a novel pipeline of instance segmentation for windows. Our method is based on Mask R-CNN [21], and is integrated with a spatial attention module and a relation module. The spatial attention and relation modules are first used in the application of windows instance segmentation. With these attention operations, our method can model the spatial relationships between windows. This is obviously helpful for the extraction of manmade structures. The contributions of this paper lie in three aspects:

1. We added a spatial attention module to the Region Proposal Network (RPN) and used channel-wise and spatial attention mechanisms to optimize the objectness scores of the RPN;
2. We embedded the relation modules into the head network of Mask R-CNN, and integrated appearance and geometric features for proposal recognition;
3. We standardized and concatenated different datasets and added some new images to create a new instance segmentation dataset for a window class with 1200 annotated images.

This paper is organized as follows: in the following section, some of the recent studies on window extraction, including traditional and CNN-based methods, are presented and some visual attention modules are introduced; Section 3 introduces our proposed method and the main innovations in detail; Section 4 describes the proposed window instance segmentation dataset and experiment results of the proposed method; in Section 5, we discuss our approach and the results that are obtained; and finally, some concluding remarks are presented in Section 6.

2. Related Work

Window extraction is one of the most important parts of facade parsing. This topic has been actively studied for several decades. Although some studies have employed laser scanning point clouds or photogrammetric point clouds for window extraction [29–33], this section presents a review of image-based approaches. We divide these methods into two categories: traditional and CNN-based methods.

Traditional methods usually rely on prior knowledge, such as the repetitive structures and symmetry of windows. Alegre et al. [8] constructed a Bayesian generative model from stochastic context-free grammars to encode knowledge regarding facades. This model takes a hierarchical structure into consideration, and uses Markov chain Monte Carlo sampling to approximate the posterior over partitions using an image. Müller et al. [9] combined the procedural modelling pipeline of shape grammars with image analysis to derive a meaningful hierarchical facade subdivision. Ali et al. [10] used the multiscale Haar wavelet representation to obtain facade tiles. These tiles are then fed into a cascaded decision tree classifier driven by Gentle Adaboost. Reznik and Mayer [11] used Implicit Shape Models [34] to detect and delineate windows. Then, combined with plane sweeping, the windows in rows or columns can be detected more precisely. Simon et al. [12] proposed a modular approach to build 3D modelling using procedural grammars. This approach is suitable for facades with many repetitions and regularities. A pixelwise random forest is used to find evidences when selecting grammar rules. Cohen et al. [13] applied dynamic programming to segment facade objects. The proposed method retrieves a parsing approach, which considers common architectural constraints and returns a certificate for global optimality. Jampani et al. [14] used auto-context features to connect a sequence of boosted decision trees. Structured prior information can be learnt using a stacked generalization. Their method is simple to implement and easy to extend. Mathias et al. [15] proposed a three-layered approach for facade parsing. These three layers represent different levels of abstraction in facade images: segments, objects, and architectural elements. The architectural rules of windows and doors are taken into consideration. As one limitation of traditional facade parsing methods, they assume that facade images have been orthorectified and cropped. They can therefore use much stronger architectural priors.

With the development of deep learning approaches, CNNs have achieved the state-of-the-art results in object detection and segmentation. CNN-based methods can learn image features from annotations. Many researchers have conducted some valuable studies on CNN-based facade segmentation. To the best of our knowledge, Schmitz and Mayer [22] are the first to apply deep learning on facade segmentation. They used AlexNet [35] as the backbone, and constructed an encoder–decoder-like structure. They trained the network using deformed patches of the images. However, they did not take advantage of the structure in facades. Fathalla and Vogiatzis [36] integrated appearance and layout cues in a single framework. They used a VGG-16 [37]-based Fully Convolutional Network (FCN) [18] to obtain coarse semantic segmentation results. The results are further improved through a probabilistic shape prior captured by trained Restricted Boltzmann Machines (RBMs). Femiani et al. [24] proposed three network architectures to achieve multilabel facade semantic segmentation. Each network is designed specially to solve a different type of problem. The first network, called MultiFacSegnet, aims to assign multiple labels to each pixel. The second network, which is called a Separable network, encourages the extraction of rectangular objects. In addition, a Compatibility network tries to eliminate errors by seeking segmentation across facade element types. Ma et al. [26] proposed a pyramid Atrous Large Kernel (ALK) Network (ALKNet) for the semantic segmentation of facade images. Their method can capture long-range dependencies among building elements by using ALK modules in multiscale feature maps. It makes full use of the regular structures of facades to aggregate useful nonlocal context information and is thus capable of dealing with challenging image regions caused by occlusions, ambiguities, and other factors.

The above methods still rely on semantic segmentation, without recognizing window instances. Liu et al. [23] proposed a DeepFacade network, which uses a symmetric regularizer for training a FCN. The authors used a clustering algorithm to divide the pixelwise segmentation results into individual windows. Moreover, they proposed a symmetric loss term to improve the results. Recently, the authors introduced a Region Proposal Network (RPN) into their symmetric loss term [25]. The distances between the clustered windows and the detected bounding boxes are treated as a loss metric. Li et al. [27] regard window detection as an issue of keypoint detection and grouping. Their method detects a window

as four keypoints, allowing it to deal with irregularly distributed windows and complex facades under diverse conditions. Ma et al. [28] designed an improved Faster R-CNN [20] architecture for window detection. The innovations include a window region proposal network, a Region of Interest (RoI) feature fusion, and a context-enhancement module. In addition, a postoptimization process is designed through the regular distribution of windows to refine the detection results obtained by the improved deep architecture.

The aforementioned methods only consider windows as individual objects, without integrating their spatial distribution and location relations into the end-to-end training process. Attention mechanisms have been proved effective in many visual computing tasks [38]. With attention modules, networks can capture long-range dependencies and model the global context information. Hu et al. [39] proposed a Squeeze-and-Excitation (SE) block to exploit the channel relationship of features. In addition to channel attention, Woo et al. [40] presented a Convolutional Block Attention Module (CBAM) that also considers spatial attention. Wang et al. [41] presented Non-Local (NL) operations for capturing long-range dependencies. Their non-local operation computes the response at a position as a weighted sum of the features at all positions. To overcome the heavy computation cost of non-local operations, Cao et al. [42] designed a Global Context (GC) block, which can obtain a better accuracy but with significantly fewer computations. Hu et al. [43] proposed an object relation module. The module can merge appearance and geometric features to model the relation of objects. Inspired by these attention modules, we propose a novel instance segmentation network that integrates spatial attention and relation modules into a Mask R-CNN.

3. Methodology

3.1. Network Architecture

Our improved Mask R-CNN is illustrated in Figure 2. It includes three parts: ResNet-50 and a Feature Pyramid Network (FPN) as the backbone; a Region Proposal Network (RPN) with spatial attention; and a head network with relation modules. First, as the original Mask R-CNN, ResNet-50 [44] and a FPN [45] are used as the backbone for the extraction of multiscale feature maps. Then, an RPN is utilized to predict the objectness scores and object bounds at each position. Meanwhile, an attention module is used to obtain a spatial attention map. The objectness scores and the spatial attention map are merged using elementwise multiplication. The proposals with higher scores are fed into the head network. In the head network, there exist two branches: a Fully Connected (FC) head for proposal recognition (classification and bounding box regression) and a mask head for segmentation using a small FCN. Relation modules are embedded after each fully connected layer of the FC head. The object location relations can be learned using this structure.

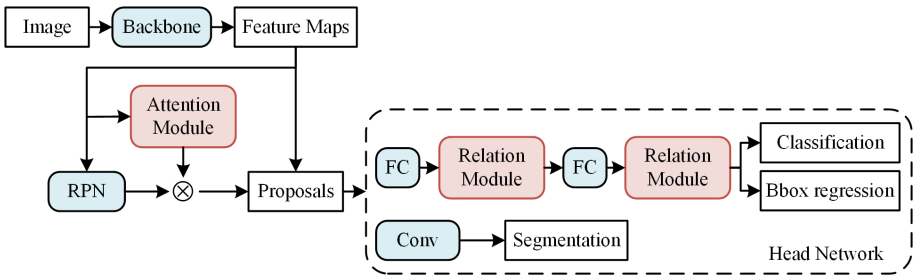

Figure 2. Pipeline of proposed method.

3.2. RPN with Spatial Attention

The RPN was first proposed by Ren et al. in the Faster R-CNN [20]. The RPN includes two branches: classification and bounding box regression. Because we use an FPN in the backbone, the RPN is applied on each level of the feature maps. At each position of a feature map, there exists three anchors of different shapes. The classification subnetwork

can predict an objectness score for each anchor. Thus, the output feature of the classification branch includes three channels. The bounding box regression subnetwork can predict the object bounds of the anchors. The number of output channels is 12, which corresponds to Δx, Δy, Δw, and Δh for each anchor, respectively.

Our spatial attention RPN is shown in Figure 3. An attention module is added as a new branch. Given an input feature map $\mathbf{F} \in \mathbb{R}^{C\times H\times W}$, our method sequentially generates a 1D channel attention map $\mathbf{M_C} \in \mathbb{R}^{C\times 1\times 1}$ and a 2D spatial attention map $\mathbf{M_S} \in \mathbb{R}^{1\times H\times W}$. Then, the 2D spatial attention map $\mathbf{M_S}$ and the objectness scores of the classification network are merged through an elementwise multiplication. The overall attention process can be summarized as follows:

$$\begin{aligned} \mathbf{F}' &= \mathbf{M_C}(\mathbf{F}) \otimes \mathbf{F}, \\ \mathbf{cls\_scores} &= \mathbf{M_S}(\mathbf{F}') \otimes \mathbf{cls\_scores}, \end{aligned} \tag{1}$$

where $\otimes$ denotes elementwise multiplication, $\mathbf{F}$ indicates the input feature map, $\mathbf{F}'$ indicates the feature map after being multiplied with channel attention, and **cls_scores** represents the objectness scores of the classification branch. During multiplication, the attention values are broadcasted (copied) accordingly: channel attention values are broadcasted along the spatial dimension; and spatial attention values are broadcasted along the channel dimension according to the outputs of the classification subnetwork.

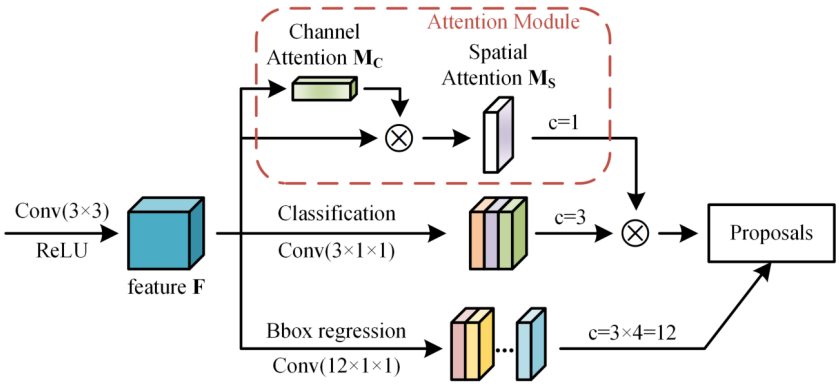

Figure 3. Diagram of the RPN with spatial attention.

The channel attention map $\mathbf{M_C} \in \mathbb{R}^{C\times 1\times 1}$ can express the interchannel relationship of the features. The spatial information of each feature map is aggregated by global average pooling and global max pooling operations, respectively, generating two different spatial context descriptors: $\mathbf{F^C_{avg}} \in \mathbb{R}^{C\times 1\times 1}$ and $\mathbf{F^C_{max}} \in \mathbb{R}^{C\times 1\times 1}$. Both descriptors are then forwarded to a shared network. The shared network is composed of Multilayer Perceptron (MLP) with two Fully Connected layers: FC_1 and FC_2. After the shared network is applied to each descriptor, the two output feature vectors are merged through an elementwise summation to produce our channel attention map $\mathbf{M_C} \in \mathbb{R}^{C\times 1\times 1}$. In short, the channel attention is computed as follows:

$$\begin{aligned} \mathbf{M_C}(\mathbf{F}) &= \sigma(\mathrm{MLP}(\mathrm{AvgPool}(\mathbf{F})) + \mathrm{MLP}(\mathrm{MaxPool}(\mathbf{F}))) \\ &= \sigma\left(\mathrm{FC}_2\left(\mathrm{ReLU}\left(\mathrm{FC}_1\left(\mathbf{F^C_{avg}}\right)\right)\right) + \mathrm{FC}_2\left(\mathrm{ReLU}\left(\mathrm{FC}_1\left(\mathbf{F^C_{max}}\right)\right)\right)\right), \end{aligned} \tag{2}$$

where σ denotes the sigmoid function. In addition, MLP indicates a Multilayer Perceptron, which includes two fully connected layers and a Rectified Linear Unit (ReLU) activation function. Here, FC_1 and FC_2 share the same weights for both inputs. AvgPool and MaxPool indicate the global average pooling and global max pooling, respectively.

The spatial attention map $\mathbf{M_S} \in \mathbb{R}^{1\times H\times W}$ indicates the interspatial relationship of the features. To compute the spatial attention, we first apply average-pooling and max-pooling operations along the channel axis and concatenate them to generate an efficient feature

descriptor $\left[\mathbf{F}^{\mathbf{S}}_{\mathbf{avg}}; \mathbf{F}^{\mathbf{S}}_{\mathbf{max}}\right] \in \mathbb{R}^{2\times H\times W}$. On the concatenated feature descriptor, we apply a convolution layer to generate a spatial attention map $\mathbf{M_S}(\mathbf{F}) \in \mathbb{R}^{1\times H\times W}$, which encodes where to emphasize or suppress. In short, the spatial attention is computed as

$$\begin{aligned}\mathbf{M_S}(\mathbf{F}) &= \sigma\left(\text{conv}^{7\times 7}([\text{AvgPool}(\mathbf{F}); \text{MaxPool}(\mathbf{F})])\right) \\ &= \sigma\left(\text{conv}^{7\times 7}\left(\left[\mathbf{F}^{\mathbf{S}}_{\mathbf{avg}}; \mathbf{F}^{\mathbf{S}}_{\mathbf{max}}\right]\right)\right),\end{aligned} \tag{3}$$

where σ denotes the sigmoid function. $\text{conv}^{7\times 7}$ represents a convolution operation with a filter size of 7×7. $[\cdot]$ indicates a concatenation of the feature maps. AvgPool and MaxPool are the average and max. pooling along the channel axis.

3.3. Head Network with Relation Modules

After applying the RPN, we can obtain some proposals that include the feature maps of the foreground objects. The feature map of each object is processed individually by the head network. The relations between these objects are not considered or learned by the network. However, there is no doubt that modelling relations among objects will improve the object detection and segmentation. Hence, after our spatial attention RPN, relation modules are embedded in the head network of the Mask R-CNN to learn the relations between window objects.

3.3.1. Relation Module

The object relation module was proposed by Hu et al. [43] in 2018. Their approach was inspired by a basic attention module, called Scaled Dot-Product Attention [46]. For one object, there exists an appearance feature $\mathbf{f}_A$ and a geometric feature $\mathbf{f}_G$. The appearance feature $\mathbf{f}_A$ indicates the clipped feature map in its bounding box. The geometric feature $\mathbf{f}_G$ indicates the four-dimensional object bounding box. Figure 4 shows the computation of the relation feature. For the nth object, its appearance feature $\mathbf{f}^n_A$ and the appearance features of other objects $\mathbf{f}^m_A$ are projected into subspaces through a dot product. An appearance weight, indicating their similarities, is then computed. The geometric features $\mathbf{f}^n_G$ and $\mathbf{f}^m_G$ are also embedded into a high-dimensional representation using sine and cosine functions of different wavelengths [46]. Finally, the appearance weight, geometry weight, and $\mathbf{f}^n_A$ are combined together to obtain a relation feature $\mathbf{f}^n_R$.

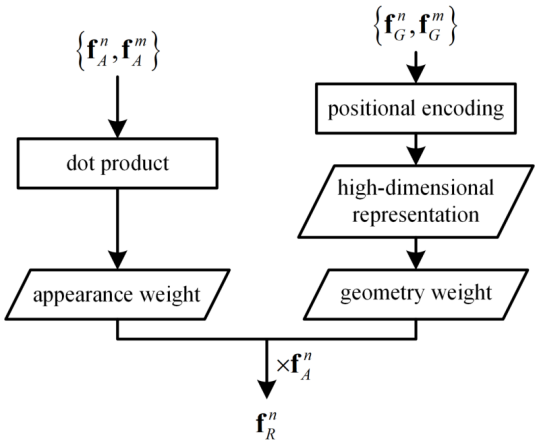

Figure 4. Relation feature computation.

After a total of N_r relation features are calculated, all relation features are concatenated together and augmented with the input appearance feature $\mathbf{f}^n_A$ through an addition, as shown in Equation (4).

$$\mathbf{f}^{n\,\prime}_A = \mathbf{f}^n_A + \left[\mathbf{f}^1_R(n), \cdots, \mathbf{f}^{N_r}_R(n)\right], \text{for all } n, \tag{4}$$

where $\mathbf{f}_A^n$ denotes the appearance feature of the nth object, $\mathbf{f}_R^{N_r}(n)$ indicates the N_rth relation feature of the nth object, $[\cdot]$ represents the concatenation of the feature maps, and $\mathbf{f}_A^{n\prime}$ indicates the new appearance feature after being augmented with relation modules.

3.3.2. Relation for Instance Segmentation

The relation module is lightweight and in-place. It does not require additional supervision and is easy to be embedded in existing networks. In this section, we embed relation modules into the head network of the Mask R-CNN. There include two branches in the head network. One branch uses two Fully Connected layers (2FC) to generate the final features for the proposal classification and bounding box regression. The other branch uses a list of convolutional layers for a binary segmentation of the objects.

Equation (5) shows the structure of the 2FC head. Given the RoI features for the nth proposal, two FC layers with 1024 dimensions are applied. Linear layers are then used for the instance classification $score_n$ and bounding box regression $bbox_n$.

$$\begin{aligned} RoI\_Feat_n &\xrightarrow{\text{FC}} 1024 \\ &\xrightarrow{\text{FC}} 1024 \\ &\xrightarrow{\text{LINEAR}} (score_n, bbox_n) \end{aligned} \tag{5}$$

Equation (6) shows the manner in which we embed the Relation Modules (RMs). Because relation modules can keep the dimensions of the input and output features, they can be used after either FC layer and repeated for an arbitrary number of times. Here, r_1 and r_2 indicate the repeated times of each relation module.

$$\begin{aligned} \{RoI\_Feat_n\}_{n=1}^{N} &\xrightarrow{\text{FC}} 1024 \cdot N \xrightarrow{\{\text{RM}\}^{r_1}} 1024 \cdot N \\ &\xrightarrow{\text{FC}} 1024 \cdot N \xrightarrow{\{\text{RM}\}^{r_2}} 1024 \cdot N \\ &\xrightarrow{\text{LINEAR}} \{(score_n, bbox_n)\}_{n=1}^{N} \end{aligned} \tag{6}$$

4. Experiments

In this section, we evaluate our approaches using window instance datasets. Our models were implemented using PyTorch and Detectron2 [47]. The codes will be publicly available (https://github.com/SunYW0108, accessed on 29 December 2021). For backbone networks, we used ResNet-50 with pretrained model parameters on ImageNet classification tasks [48]. The parameters of the first two stages were frozen, i.e., will not change during training.

We evaluated the experiment results through visualizations and numerical performance metrics, i.e., the mean Average Precision (mAP). In the Microsoft Common Objects in COntext (COCO) evaluation criteria [49], the AP is averaged over 10 Intersection over Union (IoU) values, which are 0.50–0.95 with a step size of 0.05. AP_{50} and AP_{75} represent the APs at IoUs of 0.50 and 0.75, respectively. AP_S, AP_M, and AP_L are the APs for small (area < 32^2), medium (32^2 < area < 96^2), and large (area > 96^2) objects, respectively. The mAP is averaged over all categories. In our approach, there is no distinction between AP and mAP because we focus on only one class, i.e., windows.

4.1. Our New Dataset

Currently, publicly available building facade datasets are mainly designed for image semantic segmentation tasks. To prepare them for windows instance segmentation, we extract window objects from the annotated images and encode the information of these windows in the COCO instance segmentation format. In this study, six facade datasets, CMP [50], eTRIMS [51], ECP [52], ICG Graz50 [53], RueMonge 2014 [54], and ParisArt-Deco [55], were selected. These facades are from different cities around the world and are of diverse architectural styles. The annotations of the ECP dataset are provided by Martinović

et al. [56], where their annotations better fit the actual ground truth based on the visual comparison. In addition, we manually annotated 82 images taken by ourselves.

All images and annotations are standardized and concatenated together to create a new instance segmentation dataset. The number of images in our concatenated dataset are shown in Table 1. The images and labels of each dataset are divided into five parts randomly, among which four are used for training and one is applied for testing. The total number of images in our dataset is shown in the last row of the table. The number of images for training is 959. The number of images for testing is 241.

Table 1. Number of images in our concatenated dataset.

Dataset	Number of Images	Training Set	Testing Set
CMP base	378	302	76
CMP extended	228	182	46
eTRIMS	60	48	12
ECP	104	83	21
ICG Graz50	50	40	10
RueMonge 2014	219	175	44
ParisArtDeco	79	63	16
TUBS	82	66	16
Merged dataset	1200	959	241

4.2. Three Variants of the RPN with Attention Modules

In this section, we compare the results of three combinations using attention modules and the RPN. The network architectures of different combinations are shown in Figure 5. In Figure 5, AM indicates the attention module, cls represents the classification subnetwork, reg indicates the bounding box regression sub-etwork, CA is the channel attention, SA is the spatial attention, and $\otimes$ indicates the elementwise multiplication. In the first variant (Figure 5a), the input feature is fed into three branches: an attention module, classification, and bounding box regression. In the subnetwork of the attention module, a channel attention map is first computed and merged with the input feature to obtain a new feature map. A spatial attention map is then generated for multiplication with the objectness scores of the classification branch. This architecture is labelled cls(AM)_reg. In the architecture shown in Figure 5b, labelled as cls(AM)_reg(AM), the spatial attention map is further merged with the output maps of the bounding box regression. In Figure 5c, the attention module is executed on the input feature. The output feature is then used for classification and bounding box regression. This architecture is labelled as AM_cls_reg.

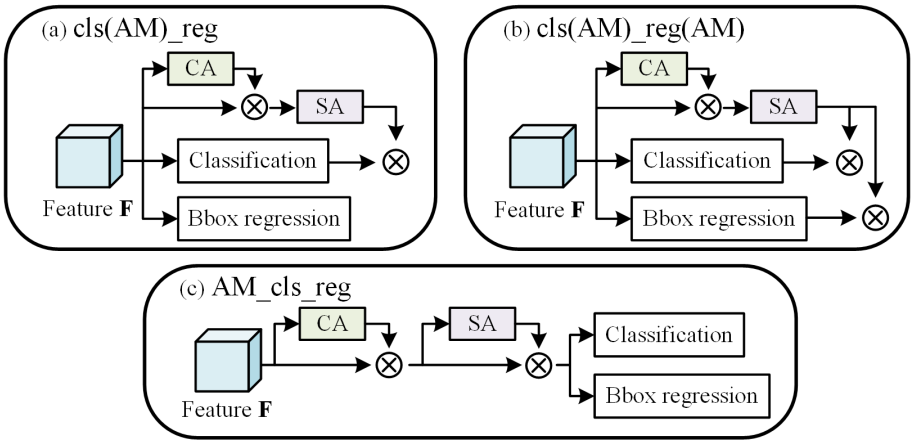

Figure 5. Diagram of three variants of the RPN with attention modules: (**a**) cls(AM)_reg; (**b**) cls(AM)_reg(AM); (**c**) AM_cls_reg.

The results of three different combinations and the original Mask R-CNN are shown in Table 2. The top scores are indicated in bold. Four rows in front of AP^{bb} denote the APs for object detection. Similarly, four rows in front of AP^{segm} represent the APs for object instance segmentation. The results of Mask R-CNN are used as the baseline. As can be seen in Table 2, our method cls(AM)_reg can achieve the best AP on both object detection and segmentation tasks, 0.7% and 0.7% higher than the original Mask R-CNN method, respectively. Excluding AP^{bb}_{75}, AP^{bb}_{L}, and AP^{segm}_{L}, our method has achieved the best results for other evaluation measurements. Although cls(AM)_reg(AM) can obtain better results than the original Mask R-CNN, its results are lower than our method. In addition, AM_cls_reg achieves a result worse than the original Mask R-CNN. The experiment results indicate that our spatial attention RPN, labelled cls(AM)_reg, is reasonable and effective. The method of the following section is implemented based on this architecture.

Table 2. Comparison of results using different combinations of attention modules with the RPN (unit: %).

	Network Architecture	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
AP^{bb}	Mask R-CNN	53.1	83.9	**61.2**	39.5	59.6	**61.2**
	cls(AM)_reg	**53.8**	**84.4**	61.0	**40.0**	**60.4**	59.4
	cls(AM)_reg(AM)	53.3	84.1	60.6	39.6	59.9	59.4
	AM_cls_reg	52.6	83.9	59.9	38.7	59.6	59.8
AP^{segm}	Mask R-CNN	53.7	83.0	62.0	40.6	60.2	**61.4**
	cls(AM)_reg	**54.4**	**83.6**	**62.6**	**40.9**	**60.8**	59.6
	cls(AM)_reg(AM)	53.9	83.2	62.4	40.7	60.6	60.0
	AM_cls_reg	53.2	83.2	60.6	39.8	60.2	59.1

4.3. Comparisons of Parameters for Relation Modules

In the head network with relation modules, there are two key parameters: the number of relations N_r and the number of modules $\{r_1, r_2\}$. Hu et al. [43] conducted some tests on these parameters. For the COCO detection datasets, their method achieves the highest AP when the number of relations N_r equals 16. For the number of modules $\{r_1, r_2\}$, they recommend $r_1 = 1$ and $r_2 = 1$ according to the tradeoff between the AP and computation complexity. Hence, in our experiments, $r_1 = 1$ and $r_2 = 1$ are applied, and the results of different relation number N_r are compared.

The experiment results using the spatial attention RPN and relation modules are shown in Table 3. We also show the results using relation modules without the spatial attention RPN, as listed in Table 4.

Table 3. Comparison of AP on different numbers of relations with the spatial attention RPN (unit: %).

Number of Relations		AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
1	bbox	55.0	85.7	62.8	41.4	61.0	61.3
	segm	55.3	84.6	63.5	41.7	61.3	61.8
2	bbox	55.2	84.9	63.3	40.5	61.7	63.8
	segm	55.6	84.0	64.8	40.6	62.1	62.7
4	bbox	55.3	84.8	63.5	41.9	61.4	61.4
	segm	55.7	83.9	64.9	42.4	61.8	61.3
8	bbox	55.7	85.7	63.7	42.2	61.6	62.4
	segm	56.3	84.7	65.4	43.1	62.1	62.1
16	bbox	55.5	84.8	64.1	42.2	61.5	62.3
	segm	55.9	84.6	64.7	42.3	61.8	62.0
32	bbox	56.4	87.0	64.7	42.4	62.4	62.1
	segm	56.7	86.1	65.5	42.8	63.1	62.6
64	bbox	56.2	84.7	65.6	41.2	62.6	63.5
	segm	56.2	84.6	65.9	41.5	62.6	62.3

Table 4. Comparison of AP on different numbers of relations with the original RPN (unit: %).

Number of Relations		AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
1	bbox	54.0	84.2	61.9	40.9	60.6	60.7
	segm	54.3	84.2	62.4	41.1	60.7	60.4
2	bbox	54.8	85.6	62.5	41.2	61.0	62.4
	segm	54.8	84.7	62.9	41.1	61.2	61.9
4	bbox	54.8	85.6	62.3	41.1	61.3	62.7
	segm	55.2	84.8	62.8	41.6	61.6	62.3
8	bbox	56.0	85.9	64.3	42.2	62.1	63.5
	segm	56.1	84.9	65.1	42.2	62.2	63.1
16	bbox	56.0	85.7	64.6	42.8	61.8	62.5
	segm	56.2	84.7	65.1	43.0	62.0	62.8
32	bbox	56.2	85.6	64.5	42.8	62.2	64.6
	segm	56.1	84.6	64.9	43.1	62.0	63.5
64	bbox	56.6	86.2	65.7	41.8	63.0	64.0
	segm	56.6	85.3	66.3	42.3	63.1	63.6

The AP values are depicted as a line chart in Figure 6. The horizontal axis represents the number of relations. The vertical axis represents the AP values of the different methods. The solid lines in red and blue represent the APs of our method for object detection and segmentation, respectively. The dash lines in red and blue indicate the AP^{bb} and AP^{segm} of the Mask R-CNN with relation modules, respectively. As shown in Figure 6, when the number of relations is small (less than 8), by increasing the number of relations, the AP^{bb} and AP^{segm} of these two algorithms increase, although the advantages of our method gradually decrease. When the number of relations equals 16 for our method and 32 for the Mask R-CNN with relation modules, the APs decreases slightly. When the number of relations equals 32, our method can achieve the highest AP results. As the number of relationships continues to increase, the AP values of our method begin to decrease. When the number of relations equals 64, the Mask R-CNN with relation modules obtains the highest AP results. Considering the tradeoff between the AP and computational complexity, $N_r = 32$ is used in our approach. The results of the experiments demonstrate the advantage of the spatial attention RPN. Thus, our method can achieve higher AP results with a smaller number of relations.

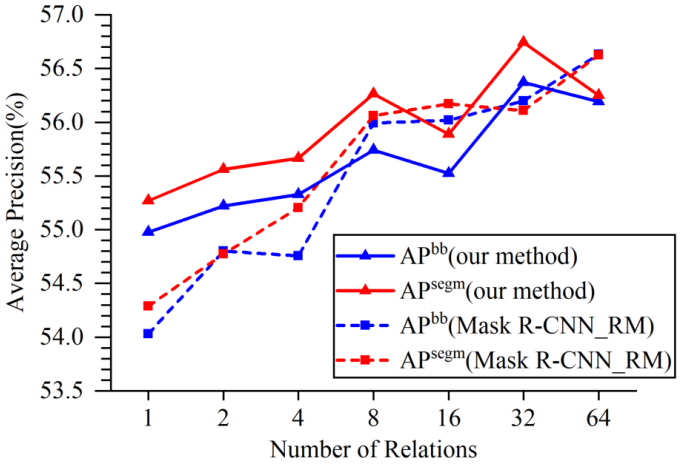

Figure 6. Comparison of AP on different numbers of relations (unit: %).

When the number of relations equals 32, AP^{bb} and AP^{segm} of our method reach 56.4% and 56.7%, respectively. Comparing these results with the results of the method without relation modules in Section 4.2 (cls(AM)_reg), AP^{bb} and AP^{segm} increase by 2.6% and 2.3%, respectively. This proves that relation modules have learnt information between objects. In

addition, the AP^{bb} and AP^{segm} of our method are higher than those of the Mask R-CNN by 3.3% and 3.0%, respectively.

4.4. Qualitative Results

This section shows a qualitative comparison of the proposed method and Mask R-CNN. We chose three facade images, as shown in Figures 7–9. Subgraph (a) shows the ground truth of window instances. Subgraph (b) shows the results of the Mask R-CNN. Subgraph (c) shows the results of the Mask R-CNN with a spatial attention RPN. Sub-graph (d) shows the results of the Mask R-CNN with a spatial attention RPN and relation modules. Different window instances are rendered in different colors. The segmentation errors are marked by the red rectangles.

Figure 7 shows the results of error detection by the Mask R-CNN. Compared with the ground truth, there exist some incorrectly detected windows by Mask R-CNN at the top of the facade. After adding a spatial attention mechanism for the RPN, the number of incorrectly detected windows diminishes to one. The improved method using relation modules results in no detection errors. As can be seen from Figure 7d, all detection errors are eliminated. The four windows in the lower part cannot be detected by any of the methods used in our experiments.

Figure 8 shows the windows undetected by the Mask R-CNN under different illumination conditions. There are some undetected windows and an incorrectly detected window at the top and ground floor of the facade using the Mask R-CNN. After applying a spatial attention mechanism for the RPN, some undetected windows can be correctly detected, but the incorrectly detected window still exists. Then, by adding relation modules to the head network, all undetected windows have been detected, and the error detection has been removed.

Figure 9 shows the results of different methods in the presence of large occlusions. As indicated in Figure 9a, the ground truth of this facade provides a hand-annotated ground truth for the labels behind the vegetation. In the red box of Figure 9b, only one window instance can be detected behind the vegetation by the Mask R-CNN. After adding a spatial attention and relation modules, the other two window instances are detected correctly. Comparing the windows in the middle of the red rectangle detected by the three methods, the size of the window detected by our method is more precise.

(a) (b) (c) (d)

Figure 7. Window segmentation results of facade 1: (**a**) ground truth of window instances; (**b**) result of Mask R-CNN; (**c**) result of Mask R-CNN with a spatial attention RPN; (**d**) result of Mask R-CNN with a spatial attention RPN and relation modules.

Figure 8. Window segmentation results of facade 2: (**a**) ground truth of window instances; (**b**) result of Mask R-CNN; (**c**) result of Mask R-CNN with a spatial attention RPN; (**d**) result of Mask R-CNN with a spatial attention RPN and relation modules.

Figure 9. Window segmentation results of facade 3: (**a**) ground truth of window instances; (**b**) result of Mask R-CNN; (**c**) result of Mask R-CNN with a spatial attention RPN; (**d**) result of Mask R-CNN with a spatial attention RPN and relation modules.

4.5. Comparisons with Other Attention-Based Methods

In this section, we make a quantitative comparison with four attention-based methods to verify the utility of our proposed approach. The attention modules we use for comparison include the Convolutional Block Attention Module (CBAM) [40], Non-Local (NL) module [41], Global Context (GC) module [42], and Relation Module (RM) [43]. The implementation details are the same as those in their original papers. For the network, called Mask R-CNN + CBAM, the CBAM is integrated with each residual block in ResNet [44]. For Mask R-CNN + NL, only one nonlocal module is added right before the last residual block of res_4 in ResNet. The architecture Mask R-CNN + GC denotes adding the GC module to all residual blocks of res_3, res_4, and res_5. In the Mask R-CNN + RM network, relation modules are added after both fully connected layers in the head network of the Mask R-CNN.

The average precisions of our method and the other attention-based methods are shown in Table 5. The top scores are indicated in bold. Compared with the baseline method (Mask R-CNN), all attention-based methods can obtain better results. Meanwhile, our method achieves a better performance than other attention-based methods except the detection and segmentation of small windows (AP_S^{bb} and AP_S^{segm}) and large windows (AP_L^{bb} and AP_L^{segm}). The results indicate that our method is more suitable for the instance segmentation of objects with medium and similar sizes. Windowlike facade elements are a perfect fit for this characteristic.

Table 5. Average precision of our method and other attention-based approaches (unit: %).

Network Architecture		AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Mask R-CNN	bbox	53.1	83.9	61.2	39.5	59.6	61.2
	segm	53.7	83.0	62.0	40.6	60.2	61.4
Mask R-CNN + CBAM	bbox	53.2	85.2	61.5	39.8	59.7	61.1
	segm	54.0	85.3	61.9	41.3	60.3	60.0
Mask R-CNN + NL	bbox	53.6	84.6	61.7	40.1	59.9	62.7
	segm	54.1	82.9	62.4	41.1	60.4	62.1
Mask R-CNN + GC	bbox	54.4	85.1	62.8	40.3	60.9	**63.5**
	segm	54.8	84.1	63.3	41.4	61.3	**63.2**
Mask R-CNN + RM	bbox	56.0	85.7	64.6	**42.8**	61.8	62.5
	segm	56.2	84.7	65.1	**43.0**	62.0	62.8
Our method	bbox	**56.4**	**87.0**	**64.7**	42.4	**62.4**	62.1
	segm	**56.7**	**86.1**	**65.5**	42.8	**63.1**	62.6

4.6. Comparisons with Other Window Extraction Methods

To compare our method with other window extraction approaches [22–25,27], we retrained and evaluated the proposed method on several datasets: eTRIMS, ECP, CMP, Graz50, and ParisArtDeco. The pixel accuracy is used as a metric in these previous studies, which can be calculated through Equation (7). True Positive (TP) means the pixels are correctly recognized as windows. True Negative (TN) means the pixels are correctly recognized as facades. False Positive (FP) means the pixels belonging to facades are incorrectly recognized as windows. False Negative (FN) means the pixels belonging to windows are incorrectly recognized as facades. The sum of TP and TN divided by the number of all pixels represents the pixel accuracy. The pixel accuracy is expressed as a percentage. Table 6 shows the pixel accuracy of the different methods. The top scores are indicated in bold. Here, "-" indicates that the authors did not conduct experiments on the corresponding dataset. The pixel accuracy of our method was evaluated using the window instances with confidence thresholds > 0.5.

$$\text{Pixel Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{7}$$

Table 6. Comparison of results of our method and other window extraction approaches on different datasets based on the pixel accuracy (unit: %).

	eTRIMS	ECP	CMP	Graz50	ParisArtDeco
Schmitz et al. [22]	86.0	-	-	-	-
Liu et al. [23]	90.9	93.0	89.0	87.7	94.2
Femiani et al. [24]	**97.1**	95.6	-	-	-
Liu et al. [25]	92.4	**97.6**	95.0	88.8	95.4
Li et al. [27]	84.0	95.0	-	90.0	-
Our method	96.5	97.2	**96.5**	**95.2**	**96.1**

The results in Table 6 indicate that the proposed method outperforms most of the other approaches. Although our method does not achieve the highest accuracy on the eTRIMS and ECP datasets, the values of the pixel accuracy only decrease by 0.6% and 0.4% compared with the best results.

5. Discussion

Most of the current CNN-based methods only concern the optimization of individual windows and ignore the spatial areas or relationships of windows. In this study, we improve the Mask R-CNN architecture by integrating a spatial attention module and a relation module, and present a novel pipeline of instance segmentation for windows. With the help of the spatial attention, the improved RPN gains the capability of generating proposals that cover window objects. The elimination of redundant background proposals

will contribute to further training task in the head network. On the other hand, after integrating the relation module into the head network, the head architecture can process a set of window objects simultaneously through interaction between their appearance feature and geometry. In this way, the relations of windows can be modelled during learning. With these attention operations, our method can model the spatial relationships between windows and achieves higher average precisions for object detection and segmentation than those of the original Mask R-CNN [21] by 3.3% and 3.0%.

In order to evaluate our results, we made comparisons from two different perspectives. First, as a network integrated with attention modules, our method was compared with several attention-based methods, as shown in Section 4.5. In the original papers of these attention modules, the authors made comparisons of their methods with Mask R-CNN [21] on the Microsoft COCO dataset. The experiment results on our window instance segmentation dataset, given in Table 5, show a similar tendency of AP as in these original papers. This also proves the advantage of our network architecture. Second, in the field of window extraction, most methods use pixel accuracy as a metric to evaluate their results. As shown in Section 4.6, the researchers trained and validated their networks on different datasets. Then they reported the pixel accuracies for different datasets. To compare with other window extraction methods, we evaluated our network on five datasets: eTRIMS, ECP, CMP, Graz50, and ParisArtDeco. The pixel accuracies of our method and other methods are shown in Table 6. Except eTRIMS and ECP, our method achieves the best results on the other datasets. On the eTRIMS and ECP datasets, the results of our method are still competitive. Compared with other window extraction methods that only optimize the shape of each individual window, our method first takes spatial locations and relations of windows into consideration. This comparison proves the effectiveness of the spatial and relation modules. We also standardized and concatenated together six publicly available datasets and added 82 new images to create a new publicly available standard windows instance segmentation dataset.

Because our method focused on only one class object, one limitation is that it is difficult to distinguish windows from balconies. We believe that the instance segmentation of multiclass objects will improve the precision of window extraction. Further trials with other attention modules, such as Efficient Channel Attention (ECA) module [57] and coordinate attention [58], will also be investigated. With these new attention modules, the improved RPN can further improve the precision of window extraction.

6. Conclusions

We proposed an end-to-end deep learning network for windows instance segmentation using facade images. In particular, our proposed network is defined by adding spatial attention mechanism and relation modules to the Mask R-CNN deep learning network. First, a 2D spatial attention map is multiplied with the objectness scores of the RPN. This operation is beneficial for the generation of proposals that are more likely to cover the window objects. Second, relation modules are embedded after each fully connected layer in the head network. The relation modules enhance the representation power of the geometry relationship between window instances.

The performance of the proposed method is tested on our window instance segmentation dataset. The new dataset is created by combining six publicly available datasets and 82 new images annotated by our team. The average precisions of our method for object detection and segmentation are 56.4% and 56.7%, which are higher than those of Mask R-CNN by 3.3% and 3.0%. The qualitative comparison shows that, benefiting from the representation power on spatial relationships, our method is robust to changes in texture. We also made a quantitative comparison with other attention-based methods. The results show that our method is more suited to extracting objects with medium and similar sizes, such as windows. Furthermore, to compare our method with other window extraction methods, we retrained our network on five public datasets, and used the pixel accuracy as

the metric. The comparison results show our method achieved the best performance using three datasets, and placed second using the other two datasets.

In future works, we intend to modify the network and expand the dataset to adapt to the instance segmentation of multiclass objects. Facade elements, such as windows, doors, shops, and balconies, are obviously interrelated and interacted with each other. Besides that, more attention modules will be studied and integrated into our network to obtain better results. In addition, the existing datasets contain only one image for each facade. The use of multiview images and 3D features of facades will be one possible future research effort in furthering the development of the instance segmentation approach.

Author Contributions: Conceptualization, Yanwei Sun and Mehdi Maboudi; methodology, Yanwei Sun; software, Yanwei Sun; resources, Mehdi Maboudi; writing—original draft preparation, Yanwei Sun; writing—review and editing, Shirin Malihi, Hao Li and Mehdi Maboudi; supervision, Hao Li and Mehdi Maboudi; funding acquisition, Hao Li. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China grant number 41471276.

Institutional Review Board Statement: Not applicable

Informed Consent Statement: Not applicable

Data Availability Statement: Publicly available datasets were analyzed in this study. This data can be found here: https://github.com/SunYW0108 (accessed on 29 December 2021).

Acknowledgments: The authors would like to thank Markus Gerke from the Technical University of Braunschweig, Germany for supporting this research.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Neuhausen, M.; Koch, C.; König, M. Image-based window detection: An overview. In Proceedings of the 23rd International Workshop of the European Group for Intelligent Computing in Engineering, Krakow, Poland, 29 June–1 July 2016.
2. Gröger, G.; Plümer, L. CityGML–Interoperable semantic 3D city models. *ISPRS J. Photogramm. Remote. Sens.* **2012**, *71*, 12–33. [CrossRef]
3. Kim, S.; Zadeh, P.A.; Staub-French, S.; Froese, T.; Cavka, B.T. Assessment of the impact of window size, position and orientation on building energy load using BIM. *Procedia Eng.* **2016**, *145*, 1424–1431. [CrossRef]
4. Amirebrahimi, S.; Rajabifard, A.; Mendis, P.; Ngo, T. A framework for a microscale flood damage assessment and visualization for a building using BIM–GIS integration. *Int. J. Digit. Earth* **2016**, *9*, 363–386. [CrossRef]
5. Perez, H.; Tah, J.H.M.; Mosavi, A. Deep Learning for Detecting Building Defects Using Convolutional Neural Networks. *Sensors* **2019**, *19*, 3556. [CrossRef]
6. Taoufiq, S.; Nagy, B.; Benedek, C. HierarchyNet: Hierarchical CNN-Based Urban Building Classification. *Remote Sens.* **2020**, *12*, 3794. [CrossRef]
7. Alshawa, M.; Boulaassal, H.; Landes, T.; Grussenmeyer, P. Acquisition and Automatic Extraction of Facade Elements on Large Sites from a Low Cost Laser Mobile Mapping System. In Proceedings of the ISPRS Workshop 3D Virtual Reconstruction and Visualization of Complex Architectures, Trento, Italy, 25–28 February 2009.
8. Alegre, F.; Dellaert, F. A Probabilistic Approach to the Semantic Interpretation of Building Facades. In Proceedings of the International Workshop on Vision Techniques Applied to the Rehabilitation of City Centres, Lisbonne, Portugal, 25–27 October 2004.
9. Müller, P.; Zeng, G.; Wonka, P.; Van Gool, L. Image-based procedural modeling of facades. *ACM Trans. Graph. (TOG)* **2007**, *26*, 85. [CrossRef]
10. Ali, H.; Seifert, C.; Jindal, N.; Paletta, L.; Paar, G. Window detection in facades. In Proceedings of the 14th International conference on Image Analysis and Processing, ICIAP 2007, Modena, Italy, 10–14 September 2007; pp. 837–842. [CrossRef]
11. Reznik, S.; Mayer, H. Implicit shape models, self-diagnosis, and model selection for 3D façade interpretation. *Photogramm. Fernerkund. Geoinf.* **2008**, *3*, 187–196.
12. Simon, L.; Teboul, O.; Koutsourakis, P.; Paragios, N. Random exploration of the procedural space for single-view 3D modeling of buildings. *Int. J. Comput. Vis.* **2011**, *93*, 253–271. [CrossRef]
13. Cohen, A.; Schwing, A.G.; Pollefeys, M. Efficient structured parsing of facades using dynamic programming. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Columbus, OH, USA, 23–28 June 2014. [CrossRef]
14. Jampani, V.; Gadde, R.; Gehler, P.V. Efficient facade segmentation using auto-context. In Proceedings of the 2015 IEEE Winter Conference on Applications of Computer Vision, WACV 2015, Waikoloa, HI, USA, 5–9 January 2015; pp. 1038–1045. [CrossRef]

15. Mathias, M.; Martinović, A.; Van Gool, L. ATLAS: A Three-Layered Approach to Facade Parsing. *Int. J. Comput. Vis.* **2016**, *118*, 22–48. [CrossRef]
16. Liu, W.; Wang, Z.; Liu, X.; Zeng, N.; Liu, Y.; Alsaadi, F.E. A survey of deep neural network architectures and their applications. *Neurocomputing* **2017**, *234*, 11–26. [CrossRef]
17. Mosavi, A.; Ardabili, S.; Varkonyi-Koczy, A.R. List of deep learning models. In *Engineering for Sustainable Future*; Springer: Cham, Switzerland, 2019; pp. 202–214. [CrossRef]
18. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3431–3440. [CrossRef]
19. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015*; Springer: Cham, Switzerland, 2015; pp. 234–241. [CrossRef]
20. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards real-time object detection with region proposal networks. *Adv. Neural Inf. Process. Syst.* **2015**, *2015*, 91–99. [CrossRef] [PubMed]
21. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask R-CNN. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2961–2969.
22. Schmitz, M.; Mayer, H. A convolutional network for semantic facade segmentation and interpretation. *Int. Arch. Photogramm. Remote. Sens. Spat. Inf. Sci. -ISPRS Arch.* **2016**, *41*, 709–715. [CrossRef]
23. Liu, H.; Zhang, J.; Zhu, J.; Hoi, S.C. Deepfacade: A deep learning approach to facade parsing. In Proceedings of the IJCAI International Joint Conference on Artificial Intelligence, Melbourne, Australia, 19–25 August 2017; pp. 2301–2307. [CrossRef]
24. Femiani, J.; Para, W.R.; Mitra, N.; Wonka, P. Facade Segmentation in the Wild. *arXiv* **2018**, arXiv:1805.08634.
25. Liu, H.; Xu, Y.; Zhang, J.; Zhu, J.; Li, Y.; Hoi, C.S. DeepFacade: A Deep Learning Approach to Facade Parsing with Symmetric Loss. *IEEE Trans. Multimed.* **2020**, *22*, 3153–3165. [CrossRef]
26. Ma, W.; Ma, W.; Xu, S.; Zha, H. Pyramid ALKNet for Semantic Parsing of Building Facade Image. *IEEE Geosci. Remote. Sens. Lett.* **2020**, *18*, 1009–1013. [CrossRef]
27. Li, C.K.; Zhang, H.X.; Liu, J.X.; Zhang, Y.Q.; Zou, S.C.; Fang, Y.T. Window Detection in Facades Using Heatmap Fusion. *J. Comput. Sci. Technol.* **2020**, *35*, 900–912. [CrossRef]
28. Ma, W.; Ma, W. Deep window detection in street scenes. *KSII Trans. Internet Inf. Syst. (TIIS)* **2020**, *14*, 855–870.
29. Wang, R.; Ferrie, F.P.; Macfarlane, J. A method for detecting windows from mobile lidar data. *Photogramm. Eng. Remote. Sens.* **2012**, *78*, 1129–1140. [CrossRef]
30. Zolanvari, S.I.; Laefer, D.F. Slicing Method for curved façade and window extraction from point clouds. *ISPRS J. Photogramm. Remote. Sens.* **2016**, *119*, 334–346. [CrossRef]
31. Malihi, S.; Valadan Zoej, M.J.; Hahn, M.; Mokhtarzade, M. Window Detection from UAS-Derived Photogrammetric Point Cloud Employing Density-Based Filtering and Perceptual Organization. *Remote Sens.* **2018**, *10*, 1320. [CrossRef]
32. Xia, S.B.; Wang, R.S. Facade Separation in Ground-Based LiDAR Point Clouds Based on Edges and Windows. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2019**, *12*, 1041–1052. [CrossRef]
33. Sun, Y.; Li, H.; Sun, L. Window detection employing a global regularity level set from oblique unmanned aerial vehicle images and point clouds. *J. Appl. Remote Sens.* **2020**, *14*, 024513. [CrossRef]
34. Leibe, B.; Leonardis, A.; Schiele, B. Combined object categorization and segmentation with an implicit shape model. In Proceedings of the Workshop on Statistical Learning in Computer Vision, ECCV 2004, Prague, Czech Republic, 11–14 May 2004 Volume 2, p. 7.
35. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: New York, NY, USA, 2012; Volume 2, pp. 1097–1105.
36. Fathalla, R.; Vogiatzis, G. A deep learning pipeline for semantic facade segmentation. In Proceedings of the British Machine Vision Conference 2017, BMVC 2017, London, UK, 4–7 September 2017; pp. 1–13. [CrossRef]
37. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. In Proceedings of the 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, 7–9 May 2015 .
38. Guo, M.H.; Xu, T.X.; Liu, J.J.; Liu, Z.N.; Jiang, P.T.; Mu, T.J.; Zhang, S.H.; Martin, R.R.; Cheng, M.M.; Hu, S.M. Attention Mechanisms in Computer Vision: A Survey. *arXiv* **2021**, arXiv:2111.07624.
39. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 7132–7141.
40. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. CBAM: Convolutional Block Attention Module. In *Computer Vision—ECCV 2018*; Springer: Cham, Switzerland, 2018; pp. 3–19. [CrossRef]
41. Wang, X.; Girshick, R.; Gupta, A.; He, K. Non-local neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 7794–7803.
42. Cao, Y.; Xu, J.; Lin, S.; Wei, F.; Hu, H. GCNet: Non-Local Networks Meet Squeeze-Excitation Networks and Beyond. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops, Seoul, Korea, 27 October–2 November 2019.
43. Hu, H.; Gu, J.; Zhang, Z.; Dai, J.; Wei, Y. Relation Networks for Object Detection. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 3588–3597. [CrossRef]

44. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; Volume 2016-December, pp. 770–778. [CrossRef]
45. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, 21–26 July 2016; Volume 2017, pp. 936–944. [CrossRef]
46. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: New York, NY, USA, 2017; pp. 5999–6009.
47. Wu, Y.; Kirillov, A.; Massa, F.; Lo, W.Y.; Girshick, R. Detectron2. 2019. Available online: https://github.com/facebookresearch/detectron2 (accessed on 29 December 2021).
48. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 248–255.
49. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*; Springer: Cham, Switzerland, 2014; pp. 740–755.
50. Tyleček, R.; Šára, R. Spatial pattern templates for recognition of objects with regular structure. In *German Conference on Pattern Recognition*; Springer: Berlin/Heidelberg, Germany, 2013; Volume 8142 LNCS, pp. 364–374. [CrossRef]
51. Korč, F.; Förstner, W. *eTRIMS Image Database for Interpreting Images of Man-Made Scenes*; Technical Report; 2009. Available online: http://www.ipb.uni-bonn.de/projects/etrims_db/ (accessed on 29 December 2021).
52. Teboul, O. Ecole Centrale Paris Facades Database. Available online: http://vision.mas.ecp.fr/Personnel/teboul/data.php (accessed on 29 December 2021).
53. Riemenschneider, H.; Krispel, U.; Thaller, W.; Donoser, M.; Havemann, S.; Fellner, D.; Bischof, H. Irregular lattices for complex shape grammar facade parsing. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012 ; pp. 1640–1647. [CrossRef]
54. Riemenschneider, H.; Bodis-Szomoru, A.; Weissenberg, J.; Van Gool, L. Learning Where to Classify in Multi-view Semantic Segmentation. In *Computer Vision—Eccv 2014, Pt V*; Springer: Cham, Switzerland, 2014; Volume 8693, pp. 516–532.
55. Gadde, R.; Marlet, R.; Paragios, N.; Marlet, R. Learning Grammars for Architecture-Specific Facade Parsing. *Int. J. Comput. Vis.* **2016**, *117*, 290–316. [CrossRef]
56. Martinović, A.; Mathias, M.; Weissenberg, J.; Van Gool, L. A three-layered approach to facade parsing. In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2012. [CrossRef]
57. Wang, Q.; Wu, B.; Zhu, P.; Li, P.; Zuo, W.; Hu, Q. ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks. *arXiv* **2020**, arXiv:1910.03151.
58. Hou, Q.; Zhou, D.; Feng, J. Coordinate attention for efficient mobile network design. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 19–25 June 2021; pp. 13713–13722.