remote sensing

MDPI

Article

Sequentially Delineation of Rooftops with Holes from VHR Aerial Images Using a Convolutional Recurrent Neural Network

Wei Huang [1,2,†], Zeping Liu [1,2,†], Hong Tang [1,2,*] and Jiayi Ge [1,2]

[1] State Key Laboratory of Remote Sensing Science, Faculty of Geographical Science, Beijing Normal University, Beijing 100875, China; huangwei@mail.bnu.edu.cn (W.H.); 202121051202@mail.bnu.edu.cn (Z.L.); 202021051203@mail.bnu.edu.cn (J.G.)

[2] Beijing Key Laboratory of Environmental Remote Sensing and Digital Cities, Faculty of Geographical Science, Beijing Normal University, Beijing 100875, China

* Correspondence: hongtang@bnu.edu.cn; Tel.: +86-10-5880-6401

† These authors contributed equally to this work.

Abstract: Semantic and instance segmentation methods are commonly used to build extraction from high-resolution images. The semantic segmentation method involves assigning a class label to each pixel in the image, thus ignoring the geometry of the building rooftop, which results in irregular shapes of the rooftop edges. As for instance segmentation, there is a strong assumption within this method that there exists only one outline polygon along the rooftop boundary. In this paper, we present a novel method to sequentially delineate exterior and interior contours of rooftops with holes from VHR aerial images, where most of the buildings have holes, by integrating semantic segmentation and polygon delineation. Specifically, semantic segmentation from the Mask R-CNN is used as a prior for hole detection. Then, the holes are used as objects for generating the internal contours of the rooftop. The external and internal contours of the rooftop are inferred separately using a convolutional recurrent neural network. Experimental results showed that the proposed method can effectively delineate the rooftops with both one and multiple polygons and outperform state-of-the-art methods in terms of the visual results and six statistical indicators, including IoU, OA, F1, BoundF, RE and Hd.

Keywords: building extraction; rooftop delineation with holes; convolutional neural network (CNN); recurrent neural network (RNN)

check for updates

Citation: Huang, W.; Liu, Z.; Tang, H.; Ge, J. Sequentially Delineation of Rooftops with Holes from VHR Aerial Images Using a Convolutional Recurrent Neural Network. *Remote Sens.* **2021**, *13*, 4271. https://doi.org/10.3390/rs13214271

Academic Editors: Pedro Melo-Pinto, Taïs Grippa, Claudio Persello and Arnaud Le Bris

Received: 31 August 2021
Accepted: 21 October 2021
Published: 24 October 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Mapping the distribution of buildings using remote sensing imagery is of great significance in land use management, urban planning, disaster emergency response, and more. Since convolutional neural networks (CNN) have achieved excellent performance in the field of computer vision [1], CNN-based deep learning methods have been widely applied to building extraction tasks from Very High Resolution (VHR) images [2–8] for: locating buildings using object detection [9]; the pixel-level classification of buildings using semantic segmentation [10]; and the delineation of rooftop outlines using instance segmentation [11].

Object detection methods are often applied in building localization tasks. In rural areas where buildings are sparse, a two-stage CNN [12] structure can also be used to improve the accuracy of building localization by first performing village detection and then building detection. However, this method only outputs the coordinate position of the external rectangle where the building is and cannot reflect the geometry of the building rooftops. In contrast, semantic segmentation can be used to assign a semantic label (e.g., building or not) to each pixel [13]. Consequently, each building can be localized in a more precise way. However, the buildings from semantic segmentation appear as

irregular blobs since the semantic label of each pixel is independently labeled based on learned features from an image [14]. To make a building appear as a regular polygonal geometry (i.e., high accuracy of building edge extraction), a standard set of pre-processing or post-processing, such as edge enhancing [15], multi-source data augmentation [15,16] and boundary regularization [17,18], needs to be developed. Xu et al. [15] utilize NDVI, NDSM and DEM as additional network input and enhance the edge information of original aerial images to improve the accuracy of building extraction. Huang et al. [16] fuse the RGB and NRG compositions to accurately locate buildings. Recently, the Frame Field Learning method [19] followed the multi-task learning network and used U-net [20] to output both segmentation maps and frame fields. The contours were then optimized using the Active Skeleton Model (ASM) to align with the frame field, and the corners were detected using the frame field to simplify non-corner vertices. The approach is a typical method of semantic segmentation followed by a series of post-processing to make the building edge contours more regular.

Although semantic segmentation methods achieve excellent accuracy on building extraction with the benefit of pre-processing or post-processing [4], the parameters of these procedures rely heavily on domain knowledge, which brings uncertainty to the output of models.

Distinct from semantic segmentation methods, instance segmentation is more in line with the cognition of buildings as independent individuals. Typical instance segmentation methods such as Mask Region-CNN [21] (Mask R-CNN) add a semantic segmentation branch to the building detection task to obtain more accurate building extraction results. The results extracted from the VHR images using Mask R-CNN are still irregular edges of the building rooftop, while the building distribution mapping is concerned with whether the contours of the rooftop have regular geometry. This makes it necessary for the building extraction method to reflect the geometric elements of the rooftop contours, such as line segments and corner points. Such situations can also be found in widely used contour-based building extraction methods, such as ACDRNet. ACDRNet [22] is an active contouring method based on neural network inference. The network uses an encoder-decoder to infer the shift map from the image and create polygons based on the vertices, which are shifted according to the 2D shift map in each iteration. It requires initialization of the building polygon outline, which is generally a circle with a fixed threshold set as the radius based on the size of the building object. As such, initialization makes the methods unable to handle the presence of the holes. Zhao et al. [23] use recurrent neural networks (RNN) guided by edge information to predict the coordinates of corners and connect these predicted corners to form building rooftops. Similar to [23], PolyMapper [24] aims to extract both road nets and building outlines, utilizing two parallel recurrent neural networks (RNN) to extract the vertex sequences of road nets and building outlines, finally getting the topology maps of aerial images.

OEC-RNN (i.e., object, edge and corner RNN) [25] is a recently proposed method that adds convolutional recurrent neural network structures to infer regular building rooftops based on instance segmentation (i.e., building object), specifically focusing on the edge and corner of a rooftop and delineating the polygon according to the sequence of corner points. In this method, the image features are encoded by CNN and the properties of rooftop geometry are captured by RNN [26,27]. Thus, the polygonal building rooftops are obtained directly without the post-procedure of boundary regularization. The OEC-RNN can delineate polygons accurately by predicting the position of corner points in sequence and achieving state-of-the-art performance in several aerial images. However, OEC-RNN and other state-of-the-art methods, including ACDRNet, Zhao et al. and PolyMapper, which represent instances of buildings as polygons, all have a crucial restriction to model the building with holes due to the very strong assumption that there exists only one exterior polygon along the boundary of a building. Hence, none of these models can handle buildings of complex shapes, such as quadrangles and the rooftops of buildings with holes, which are widely distributed in aerial image datasets. Figure 1 shows two

aerial image patches of complex buildings in the form of quadrangles in Austin and Vienna, respectively. It is evident from Figure 2 that the issue for building rooftop delineation in OEC-RNN is mainly caused by two aspects:

1. Edge confusing: Given an image of a building with holes as shown in Figure 2a, the external boundary of the rooftops tends to be well delineated, while the geometric features of the inner outline may be ignored. As shown in Figure 2c, the sequence delineation process of OEC-RNN stops with the rooftop boundary is finished;
2. Corner confusing: Internal corners and external corners might be clearly distinguished, leading to confusion in depiction. As shown in Figure 2d, the corner ambiguity leads to a failure in prediction.



Figure 1. Buildings in the form of quadrangles in aerial images: (**a**) Austin; (**b**) Vienna.

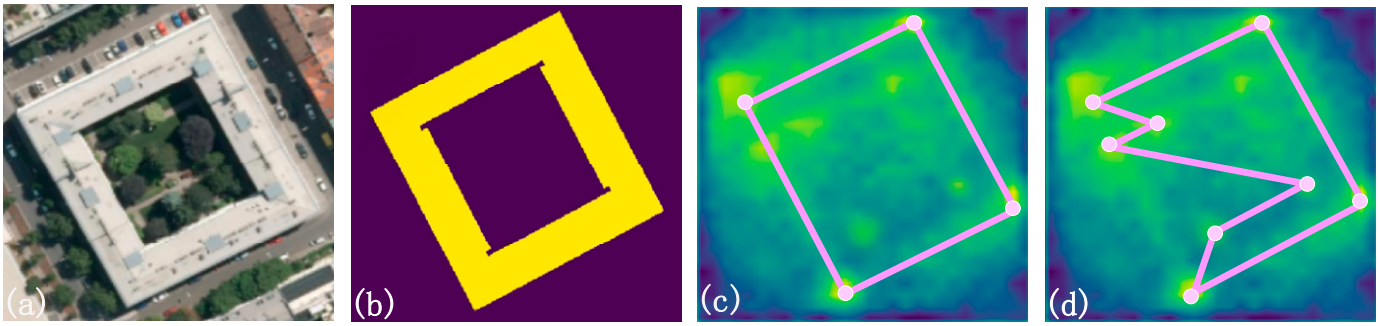


Figure 2. The failed case of Vienna in INRIA datasets: (**a**)aerial image; (**b**)ground-truth; (**c**) probability map for corner; and (**d**) results of the OEC-RNN.

To eliminate the limitation mentioned above, the OEC-RNN is extended to delineate both external and internal boundaries of building rooftops, referred to as buildings with holes in this paper. Specifically, we add a semantic segmentation branch to detect holes, and individually delineate the hole's boundary and the rooftop's exterior boundary. The polygons of the hole boundary can be more than one, defined as the internal contours of the building rooftop, while the rooftop boundary is defined as the external contours. The internal and external contours of the building rooftop are then delineated respectively by decoding the structure in OEC-RNN and finally combined into a polygon with holes.

In summary, we make the following contributions:

1. We propose a novel instance segmentation approach to delineate the boundary of rooftops with or without holes;

2. Both internal and external boundaries are learned in a unified framework and delineated by a same convolutional recurrent neural network.

The rest of this paper is organized as follows: the proposed method is described in Section 2; experimental results and performance evaluation are given in Section 3; and some conclusion is drawn in Section 4.

2. Methods

The proposed method is an extension of the OEC-RNN, an instance segmentation method that combines object detection, edge recognition and corner point sequence prediction for delineating building rooftop polygons. In [25], the author discusses some limitations of the method, especially in the case of polygons with holes in a building rooftop, where OEC-RNN may have more serious problems. In this paper, the mask branch of the object detection module in OEC-RNN is added to the overall network structure and the semantic segmentation results are used to detect holes and locate edge pixels of the internal and external contours of the box, and finally delineate the regular polygons by RNN.

In the following subsections, the three main steps will be described in detail, i.e., the preliminary work OEC-RNN, network structure of the proposed method in this paper, and delineation of rooftops by combining mask branch with the RNN branch.

2.1. Preliminary: OEC-RNN

OEC-RNN is a method to unify object detection, edge recognition and corner point sequence prediction in a single framework. The method first determines the location of the image where the building is located by object detection, and then perceives the edges of the rooftop by extracting local image features through CNN. The final polygon without holes is formed by sequentially predicting the location of the corner points and connecting them under the guidance of the rooftop edge.

As shown in Figure 3a, given a VHR image, the OEC-RNN firstly utilizes an object detection module (i.e., Mask R-CNN) to provide the bounding box (i.e., the coordinates of the upper left and lower right in the image coordinate system) of each building in the input image. Then, the OEC-RNN applies the predicted bounding box to extract every image patch which contains only one building. The second part of OEC-RNN adopts the encoder-decoder structure. Given the image patch, the network with the encoding structure uses a residual network to obtain the high-dimensional skip features of the building in the image crop (i.e., feature maps with 128 channels) and richer convolutional features (RCF) (i.e., feature maps with 5 channels) [28]. The edge possibility map, which works as edge attention of RNN, is obtained from RCF. Finally, the skip features with information of edges and corners are then passed into a two-layer convolutional long short-term memory (Conv-LSTM) [29], which outputs the coordinates of each corner point in order to form a polygon.

2.2. Network Structure of the Method

Following OEC-RNN, the workflow of our proposed method also consists of two connected modules—object detection and an encoder-decoder architecture. Differently from OEC-RNN, we utilize a mask branch from the object detection module, i.e., the red dashed box in Figure 3b. Before the encoder-decoder architecture, the structure of network is a complete Mask R-CNN. As a multi-task learning framework, it is evolved from the well-performing method Faster-RCNN. The intermediate convolutional layer of the CNN backbone is used as input of the region proposal network (RPN), and then the image feature maps of region proposals through the RoIAlign layer are used to generate a fixed-size feature map to execute box regression, classification and mask prediction.

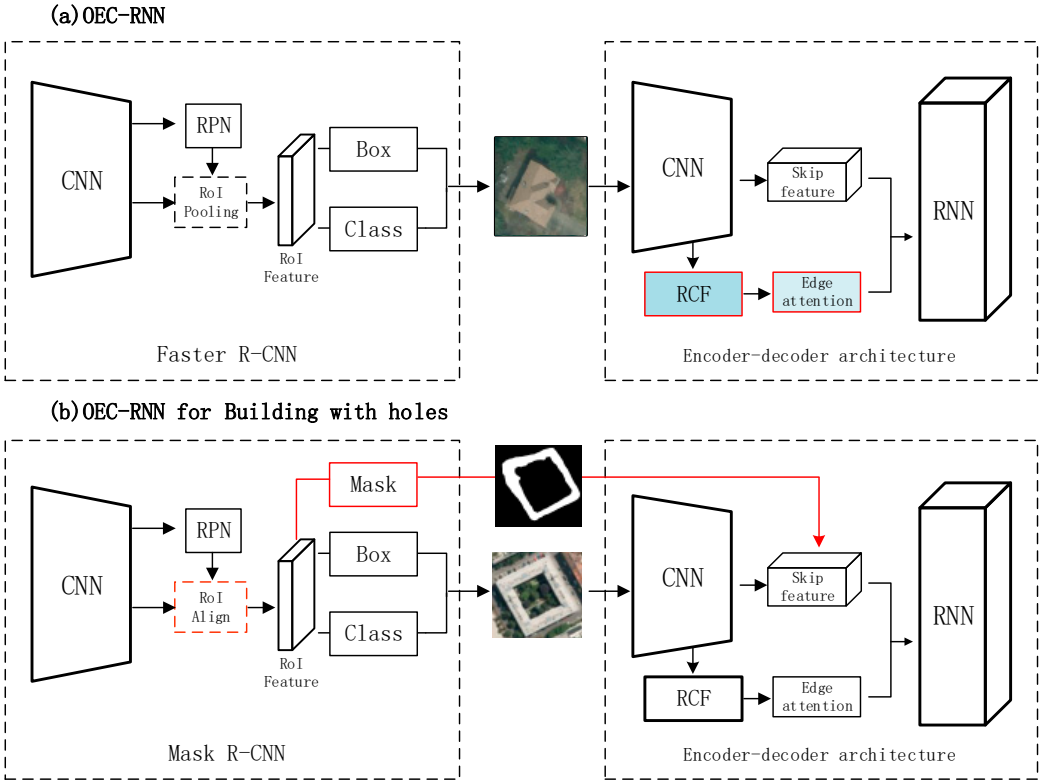


Figure 3. The network of OEC-RNN and OEC-RNN for buildings with holes. (**a**) Network structure of OEC-RNN. (**b**) Network structure of OEC-RNN for buildings with holes.

In this paper, the Backbone in Mask R-CNN is ResNet-101-FPN. Residual network (ResNet) [30] is a kind of convolutional neural network with superior performance, which solves the performance degradation of the deep convolutional neural network under depth condition and deepening the number of network layers helps to improve the correct rate. Feature pyramid networks (FPN) [31] propose a method to efficiently generate multi-dimensional feature representations for a single image view by exploiting the feature representation structure of different dimensions of the same scale image from the bottom to the top of each layer inside the conventional CNN model. This can effectively empower the regular CNN model to generate more expressive feature maps for the next stage of computer vision tasks, such as object detection and semantic segmentation.

The region of interest (RoI) feature in Figure 3 is obtained by region proposals and the features are converted by bilinear interpolation in the RoIAlign layer before being passed into the box head to output a multidimensional feature of size 7×7 with 256 channels. Box head and Class head return the exact range of coordinates where each building is located, while the Roi feature is fixed with a size of 14×14 before the mask head. The results of the mask brunch could be used for qualitative comparisons in Section 3.3. and provide a priori knowledge of the inner and outer contours for multi-polygon delineation.

The Encoder and RNN parts are the same as OEC-RNN, using ResNet to acquire the skip features of the image crop and acquire the rooftop edge attention. The building outline is delineated by a double layer convolutional RNN.

2.3. Delineation of Rooftops with Holes

Figure 4 shows the workflow of rooftop delineation in the OEC-RNN for buildings with holes, where the mask results are used to focus on the spatial position of the feature map where the inner and outer contour is located. The same RNN structure is used to delineate the contours of both internal and external boundaries of the building, respectively.

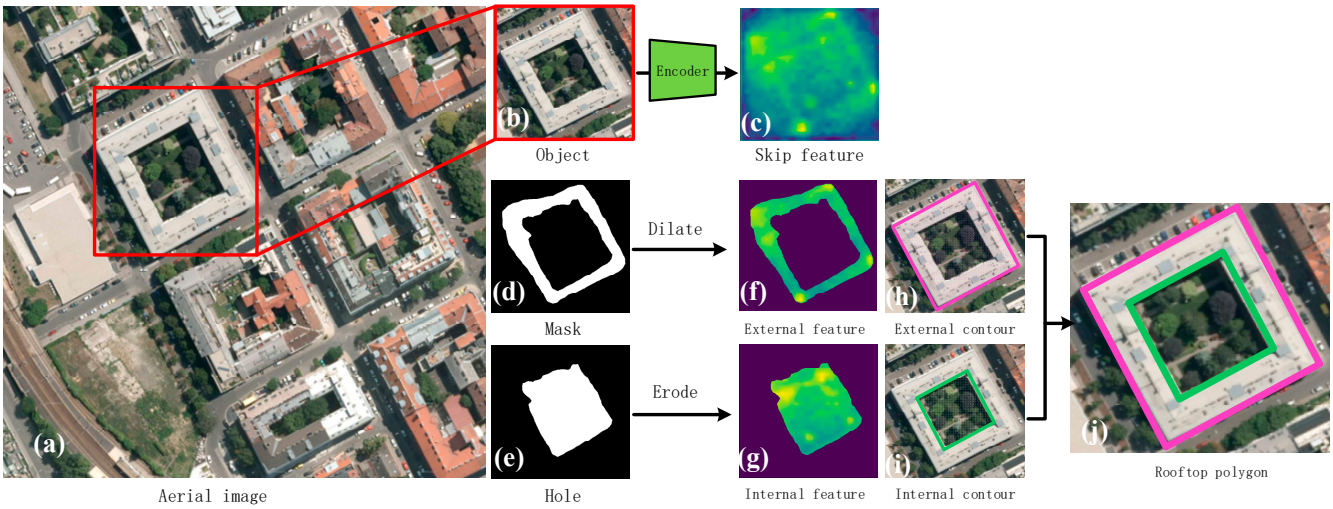


Figure 4. The framework of OEC-RNN for buildings with holes used in this paper: (**a**) original aerial image; (**b**) croped image; (**c**) feature map from the encoder; (**d**) mask result of the rooftop; (**e**) mask result of the hole; (**f**) feature map of the rooftop; (**g**) feature map of the hole; (**h**) polygon of outer contour; (**i**) polygon of hole; and (**j**) polygon of the rooftop with a hole.

The method incorporates two components. Firstly, it applies Mask R-CNN for object detection by providing a bounding box per building and object segmentation by predicting a pixel-wise probability map of both inner and outer contours. The bounding box helps crop aerial images into image patches that contain only one building. The pixel-wise probability map helps the model distinguish the location of both inner and outer contours through morphological processing to ensure that they do not interfere with each other. Secondly, based on the attention (i.e., the additional weight) given from segmentation results of inner and outer contour, the model utilizes CNN-RNN architecture to constitute the rooftop polygon with internal independent holes.

Based on the aerial images shown in Figure 4a, the bounding box (e.g., Figure 4b) of each building and the rooftop mask (Figure 4d) are obtained by instance segmentation. The bounding box crops aerial images into image patches that contain only one building. Then, the cropped image is fed into an encoder to generate a pixel-wise feature map (Figure 4c). In order to divide the external contour and the internal contour, morphological expansion and erosion with kernel size 3 are performed for each of the two masks, as shown in Figure 4d,e. The morphologically processed masks are used to intercept the feature maps of the external contour and the internal contour, as shown in Figure 4f,g, respectively. There is a difference between the proposed method and OEC-RNN—the latter is the way the features are intercepted. Mask branch performs to output the external and internal mask (obtained by the Hadamard product of the contour mask and the skip feature) of the building with the hole, as shown in Figure 2. The skip feature in the encoder-decoder architecture is the same as in the "Preliminary: OEC-RNN". Since the encoder uses a CNN with skip connections, the features extracted by the encoder are called "Skip features". In this paper, we use masks not only to distinguish the internal and external contours, but also to remove the influence of the corners of other buildings from the main object. Based on the feature maps, the decoder delineates the external contour and the internal contour of the building rooftop, shown in Figure 4h,i, respectively. Finally, all polygons are stacked to obtain a complete rooftop result as shown in Figure 4j.

3. Experimental Results

Firstly, this paper introduces the experimental data in order to solve the problem of delineating the rooftops of buildings with special morphology. The data needs to be screened in a targeted way. Then, some details and parameters of the model training are disclosed. Experimental results under two kinds of scenarios are given in the last

two subsections, respectively. The first scenario consists of qualitative comparison and quantitative evaluation between the proposed method with state-of-the-art methods, i.e., Mask R-CNN, OEC-RNN, ACDRNet and Frame Field Learning, under the assumption that the image of individual buildings has been correctly detected and cropped from the original aerial images using the given expanded ground-truth bounding box. The second scenario involves making an ablation experiment in order to analyze the impact of different components in the proposed method on the delineation of rooftops. Six metrics, including intersection over union (IoU), overall accuracy (OA), F1-score (F1), boundary F-score (BoundF [32]), rotation error (RE [25]) and Hausdorff distance (Hd [25]) are used for quantitative evaluation.

3.1. Datasets

INRIA-Austin and Vienna: The Inria Aerial Image Labeling Dataset (INRIA) [33] is mainly used for urban building mapping in aerial images for semantic segmentation tasks. We chose a subset of INRIA that contained aerial images of two cities, Austin and Vienna, i.e., INRIA-Austin and INRIA-Vienna, for training, validation and testing. The shapes of the buildings in INRIA-Austin were relatively simple, while those in INRIA-Vienna were more complex. The experimental dataset we used was composed by INRIA-Austin and INRIA-Vienna and contained 72 ortho-rectified color images of size 5000 $\times$ 5000 with a spatial resolution of 0.3 m in all. For the model to be trained effectively, we cropped 72 images of 5000 $\times$ 5000 into 3528 image patches of size 800 $\times$ 800, which contained a total of 88,926 buildings.

3.2. Model Training

We selected 80% of the image patches as training and validation samples, and the remaining 20% as test samples. The binary raster label was vectorized into the standard COCO dataset [11] format. For training, the method adopted a phased training strategy. First, Mask R-CNN was trained for instance segmentation, and then the encoder and RNN network in Figure 3 were trained for capturing features and polygon delineating. The annotations of both the outline of the hole and the rooftop were used as the ground truth to train the RNN network as the decoder. We used the Adam optimizer with a batch size of 8 and an initial learning rate of 1×10^{-4}. We then decayed the learning rate by a factor of 0.1 every 10 epochs. Taking INRIA-Vienna as an example, the training phase of the model spent about 16 h on Nvidia Tesla P40 GPU.

3.3. Comparison with State-of-the-Art Methods

Four state-of-the-art methods were selected for comparison with the proposed method in terms of both quantitative and qualitative performance. We chose these four methods as they represented four different typical methods for extracting buildings. Since the results were presented in different forms (either with the pixel classification labels of the image as the raster data of the results, or by returning the coordinates of the vertices on the contours in a certain order), we visualized all the results with a uniform post-processing (e.g., the Douglas-Peucker algorithm [34]) with a threshold value of 5 on the original aerial image and on a pure-color background.

3.3.1. Qualitative Comparison

An aerial image of a typical quadrangle building distribution in Vienna was selected for qualitative comparison in terms of delineation of the rooftop boundary. To facilitate visual comparison, the outlines of the different buildings were drawn in different colors and superimposed on the original aerial image and the solid-colored background, respectively. Figure 5 shows the performance of different methods for delineating the rooftops of buildings.

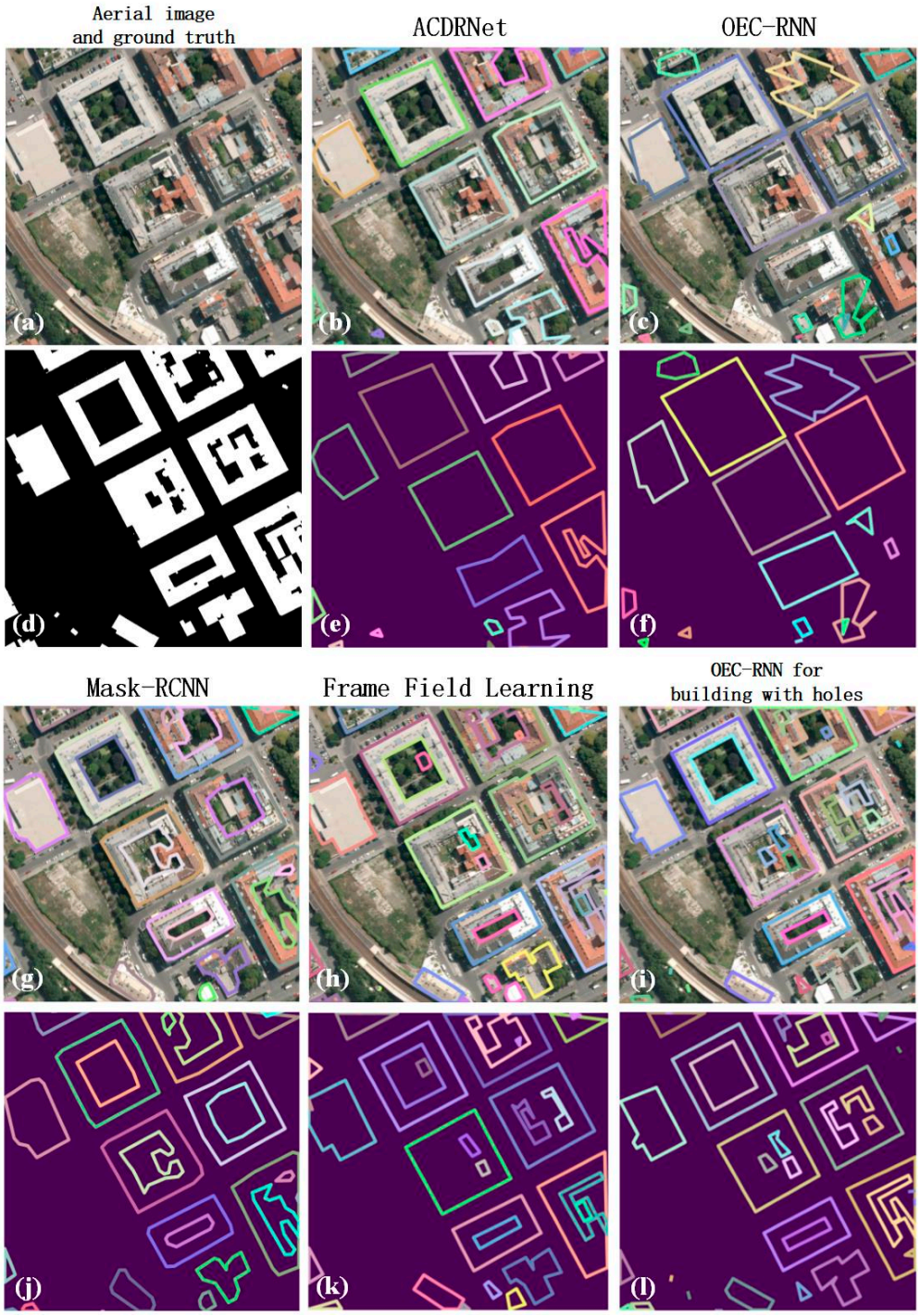


Figure 5. Qualitative comparison. These building instances are selected from Inria-vienna datasets: (**a**) original aerial image; (**d**) ground-truth. (**b**,**c**,**g**,**h**,**i**) are results with aerial images of ACDRNet, OEC-RNN, Mask R-CNN, frame field learning and OEC-RNN for buildings with holes, respectively. (**e**,**f**,**j**,**k**,**l**) are the results with a pure-color background.

Figure 5e,f are the results of the ACDRNet and OEC-RNN. Due to the inability to take into account the holes contained in the polygons, most of the buildings have only the outer contour of a rooftop, and part of the building rooftops have errors at the corner points. The results obtained by this type of method do not truly reflect the shape of building rooftops on urban ground. It could thus be elucidated that contour-based methods are unable to delineate polygons with holes.

Since the outputs of Mask R-CNN are semantic segmentation maps, they overcome the issue of the detection of holes, but the rooftop outlines they obtain are irregular boundaries. Although we used some regularization post-processing, we still had difficulty conforming the folded line rooftop contour of the buildings and the geometric properties of the building rooftop contours to human perception. In addition, the post-processing of edge regularization could lead to worse results due to the setting of algorithm thresholds. Furthermore, another general problem with semantic segmentation is that the aligned labels of image pixels easily adhere together. As shown in Figure 5j, a polygon may contain multiple holes that tend to fuse into one large hole when the holes are small and close together.

Figure 5h,k present the results of frame field learning. Visually, the method can be described as a satisfactory result: the rooftop edges are regular and accurately reflect the real condition of the building. The outline of the holes inside the polygon is clear and less adherent. Frame field learning also outputs a semantic segmentation map and then implements elaborate post-processing to align the edges of the semantic segmentation with the frame field. The method has been able to obtain satisfactory results on regularized edges, but possible errors in semantic segmentation cannot be avoided, such as the omission of holes and the misclassification of trees as buildings.

In contrast, the method we proposed in this paper is able to delineate polygons with one or more holes in the image, obtaining more regular geometric edges. The method also detects the presence of holes based on the results of semantic segmentation from the mask head, but when actually drawing the holes, it needs to detect whether there are corner points of the inner contour within the specific region. This means that even if there are objects misclassified as holes or buildings, they will not be drawn without obvious building corner points, which further reduces the error rate of the method. Compared to the true value, the results shown in Figure 5i,l still obtain regular and accurate edges and well-distributed holes.

3.3.2. Quantitative Evaluation

Table 1 lists the quantitative evaluation of the datasets using the four metrics. The intersection over union (i.e., IoU) is the common metric used to measure the semantic segmentation results, which is the ratio of intersection and union between the predicted results and the ground truth. F1 and OA are used to evaluate the accuracy of the pixel-wise classification. The Boundary F-score (i.e., BoundF [32]) is used to measure the boundary consistency between the polygon delineated by the model and the ground truth, where the boundaries are the boundary of rooftops and the boundary formed by holes. The Rotation Error (i.e., RE [25]) is the deviation angle between the orientations of the predicted polygon and the ground truth. The smaller the measurement value, the higher the accuracy of the predicted polygon. The Hausdorff distance (i.e., Hd [25]) is the maximum distance from the deviation point. When the value is lower, it indicates that the position of each corner of the polygon is more accurate.

As shown in Table 1, the OEC-RNN for buildings with holes obtained higher precision than other methods in terms of both IoU and BoundF. Fewer errors are committed than with other methods in terms of RE and Hd. For the Austin dataset, the buildings are evenly distributed and mostly rural detached residential dwellings. They have regular roof shapes and rarely contain leaky structures. The performance of the different methods in this dataset varies little. In the Vienna dataset, many of the buildings share walls and form a building structure with an open courtyard, which appears as a polygonal rooftop with holes in it. It can be summarized from Table 1 that the contour-based methods (e.g., ACDRNet and OEC-RNN) do not detect the presence of holes, which makes the performance of these methods on this kind of building worse compared to semantic segmentation methods (e.g., Mask R-CNN and Frame Field Learning). On average, semantic segmentation methods outperform contour methods by 15% on the IoU metric. Since Mask R-CNN does not have a post-processing regularization scheme specifically designed for it, the metric BoundF

is significantly low. The performance of Frame Field Learning is satisfactory in both qualitative comparison and quantitative evaluation, but the post-processing process is rather complex.

Table 1. Quantitative evaluation results.

Datasets	Methods	IoU	F1	OA	BoundF	RE	Hd
Austin	ACDRNet	0.7881	0.7144	0.9245	0.7163	0.6301	3.26
	OEC-RNN	0.8144	0.7521	0.9366	0.7293	**0.5653**	2.72
	Mask R-CNN	0.7812	0.819	0.9298	0.7067	0.7397	4.37
	Frame Field Learning	**0.8205**	**0.8321**	**0.9542**	0.7541	0.6126	**2.69**
	mOEC-RNN	0.8125	0.8297	0.9401	**0.7625**	0.5823	2.88
Vienna	ACDRNet	0.3717	0.677	0.8747	0.6471	1.0569	6.82
	OEC-RNN	0.3558	0.7071	0.8656	0.6595	1.1735	7.62
	Mask R-CNN	0.6238	0.7724	0.9078	0.6941	0.7507	6.75
	Frame Field Learning	0.6403	**0.8073**	**0.9282**	0.7197	0.8724	5.85
	mOEC-RNN	**0.6527**	0.7997	0.9247	**0.7199**	**0.8695**	**5.34**
Austin & Vienna	ACDRNet	0.5799	0.6957	0.8996	0.6817	0.8435	5.04
	OEC-RNN	0.5851	0.7296	0.9011	0.6944	0.8694	5.17
	Mask R-CNN	0.7225	0.7957	0.9188	0.7004	0.7452	4.56
	Frame Field Learning	0.7304	**0.8197**	**0.9412**	0.7369	0.7425	4.27
	mOEC-RNN	**0.7326**	0.8147	0.9324	**0.7412**	**0.7259**	**4.11**

3.4. Performance of Hole Detection

Hole detection relies on the natural formation of closed graphs by a pixel-by-pixel classification, which means that the existence and location of holes depends entirely on the performance of semantic segmentation. We measure hole detection performance in terms of number error rates and localization performance in terms of IoU. To define the holes, we set the set of pixels that are classified as background pixels with a number greater than 25 and are presented as closed shapes.

Table 2 lists the quantitative evaluation results of holes, since ACDRNet and OEC-RNN cannot obtain the information of holes. Holes-IoU is equal to 0 and the number errors rate of the holes in the test set is 1. Mask R-CNN can be used as a baseline, which has both errors that may be introduced by object detection and errors caused by semantic segmentation tasks. In the task of building instance segmentation, it is inevitable that a small part of the building will be lost in the object detection, so that mask branches will no longer segment the holes that may exist in the building.

Table 2. Quantitative evaluation results of holes.

Datasets	Methods	Holes-IoU	Number Errors Rate
Vienna	ACDR-Net	0	1
	OEC-RNN	0	1
	Mask R-CNN	0.7548	0.2461
	Frame Field Learning	0.8072	0.1895
	Our	**0.8624**	**0.1477**

The metrics of Frame Field Learning are slightly higher than Mask R-CNN in Table 2, because its output is a segmentation map based on the whole image, thus reducing the cases of missing buildings due to object detection. Our method is significantly better than Frame Field Learning in performance testing for holes, partly because the holes are relatively small compared to the area of the building rooftop and so are not prominent enough in the overall performance metric. In addition, the method proposed in this paper is more error-tolerant and the result of semantic segmentation only partially proves whether holes

exist and the approximate location of the existence; however, the real drawing of holes still depends on the detection of the internal contour corner points of the rooftop.

3.5. Ablation Experiments

In this section, another four experiments are added to illustrate the impact of both object detection and the use of masks on building rooftop extractions in OEC-RNN for buildings with holes. The proposed method in this paper is still a two-stage approach, and the performance of object detection is critical for the final building extraction results. We use the OEC-RNN as the baseline and replace the object detection step with an object box of ground truth. In addition, the mask from the semantic segmentation branch provides information about the possible holes in rooftops. Therefore, we continue to use the ground truth pixel annotation instead of the mask to explore the upper-performance limit of the method.

In Table 3, "BOX-RNN" means a variant of the OEC-RNN, where the ground truth box is used instead of the detected location of the building. This method is designed to reveal the impact of object detection on the delineation of rooftops. The "label" represents the pixel-wise label that uses the ground truth to distinguish where the interior and exterior edges of the building rooftops are located, while the "mask" represents the result obtained using the mask branch of the Mask R-CNN. In the second row of Table 3, the "OEC-RNN + mask" is the method proposed in this paper.

Table 3. Ablation experiment evaluation results.

Methods	IoU	F1	OA	BoundF	RE	Hd
OEC-RNN	0.5851	0.7425	0.9047	0.6944	0.8694	5.17
OEC-RNN + mask	**0.7326**	0.8147	0.9324	**0.7412**	**0.7259**	**4.11**
BOX-RNN + mask	0.7468	0.8179	0.9452	0.7724	0.6985	3.98
OEC-RNN + label	0.7996	0.8510	0.9587	0.7825	0.6687	3.56
BOX-RNN + label	0.8131	0.8818	0.9612	0.8189	0.6324	2.96

It is evident from Table 3 that the performance of these methods increases one by one from the OEC-RNN to the "Box-RNN + label" in terms of the four indexes. The proposed method achieves the largest margin over the OEC-RNN, e.g., more than 14% in terms of IoU. In addition, the "OEC-RNN + label" also achieves a rather large margin over the "OEC-RNN + mask", e.g., more than 5% in terms of IoU. They only differ in the pixel-wise label of buildings. These results show that the quality of pixel-wise labels is very important for the accurate delineation of buildings. This also validates the necessity to combine both the mask branch and the RNN to achieve a better result.

In contrast, the margin between "OEC-RNN + mask" and "BOX-RNN + mask" is less than 2% in the IoU, which is the same as that between "OEC-RNN + label" and "BOX-RNN + label". These results show that better object detection performs well in improving the delineation of buildings. However, its impact is not so significant as that of semantic segmentation results. As for the reason why a mask is more important than a box, it might be mainly due to the characteristics of the building itself in the dataset. In Vienna datasets, the buildings, whether single buildings or collections of multiple buildings, are larger objects, and there are not too many missed or false detections in object detection. Moreover, the area of holes in rooftop polygons is larger, and the error of mask branch prediction is magnified; thus, it is especially important to have a more accurate pixel-level annotation.

3.6. Computational Cost

In this section, we evaluate the computational cost of our proposed model and other state-of-the-art methods. All models are tested at batch sizes of eight by NVIDIA V100 GPU on our test set. The model sizes and inference times are listed in Table 4.

Table 4. The comparison of computational cost.

Methods	Model Size	Times
ACDRNet	0.31 GB	15 s
OEC-RNN	2.24 GB	84 s
Mask R-CNN	1.36 GB	50 s
Frame Field	**0.45 GB**	**17 s**
Our	2.52 GB	91 s

The model size means a memory usage of GPU at a batch size of eight on inference procedures. The time is the average inference time of models in the test dataset. As shown in Table 4, although the computational burden of ACDRNet, OEC-RNN and Mask R-CNN is lower than our method, they cannot delineate the rooftop with holes, which means their topology of output is limited. The frame field model shows the least time-consuming ability, while our model owns the highest computational cost. However, the Frame Field is a semantic segmentation method with a relatively complex post-processing procedure. It relies heavily on domain knowledge to get the fine hyperparameters, while our model is trained end-to-end without complex parameter settings.

4. Conclusions

In this paper, we proposed a novel method to delineate the inner and outer contours of building rooftops with holes. The workflow of our proposed method followed the objection detection module and encoder-decoder architecture. Our method utilized Mask R-CNN to detect the bounding boxes of buildings in the aerial image and crop these buildings into image patches. The encoder-decoder architecture was applied to extract high dimensional features of buildings and predict the coordinates of corners on the outer and inner outlines of building rooftops.

Different to the work of OEC-RNN [25], we utilized the addition of the semantic segmentation task of the objection detection module (i.e., Mask R-CNN) to provide information on the presence or absence of holes and their locations, allowing the RNN to focus more accurately on the spatial location of the features when delineating contours and thus improving the accuracy of rooftops.

Apart from the commonly used metrics IoU, OA and F1, we also applied RE, BoundF and Hd to evaluate the results by considering the point accuracy and geometric deviation. Our experimental results on INRIA-Austin and Vienna showed that the OEC-RNN for buildings with holes achieved the highest accuracy in inner and outer contour delineation, considering both quantitative and qualitative criteria. Our methods showed the best performance in areas where the buildings had holes inside compared with other models that could detect holes.

The ablation experiments showed that the accuracy of object detection (i.e., the precision of the location of the building bounding box) and the mask branch (i.e., the mask map of the external and internal contour of the building rooftop) could severely influence the performance of rooftop delineation. Improving the accuracy of these two branches may benefit the performance of the whole workflow.

Although the performance of our proposed method was superior to the others, the computation cost was higher, which means it required more time for tuning the parameters. For future work, there are several ways to continue, such as reducing the computation cost, exploring more effective methods to detect building objects and generating more accurate mask predictions of the external and inner outlines of building rooftops.

Author Contributions: W.H. and Z.L. made equal contributions to this work. conceptualization, H.T.; funding acquisition, H.T.; investigation, H.T., W.H., Z.L. and J.G.; methodology, W.H.; software, W.H. and Z.L.; and supervision, H.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research has been supported by the National Key R&D Program of China (grant no. 2017YFB0504102) and National Natural Science Foundation of China (grant no. 41971280).

Data Availability Statement: Publicly available datasets were analyzed in this study. This data (INRIA Dataset) can be found here: https://project.inria.fr/aerialimagelabeling/ (accessed on 20 October 2021). All of the source code for this study is available from the corresponding author upon request.

Acknowledgments: This work was supported by the National Natural Science Foundation of China (No. 41971280).

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

Abbreviations

CNN	convolutional neural networks
VHR	very high resolution
ASM	active skeleton model
Mask R-CNN	mask region-cnn
RNN	recurrent neural network
OEC-RNN	object, edge and corner recurrent neural network
RCF	richer convolutional features
Conv-LSTM	convolutional long-short term memory
RPN	region proposal network
ResNet	residual network
FPN	feature pyramid networks
RoI	region of interest
IoU	intersect over union
BoundF	boundary f-score
RE	rotation error
HD	hausdorff distance

References

1. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Syst.* **2012**, *25*, 1097–1105. [CrossRef]
2. Yuan, J. Learning building extraction in aerial scenes with convolutional networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *40*, 2793–2798. [CrossRef] [PubMed]
3. Li, Q.; Shi, Y.; Huang, X.; Zhu, X.X. Building footprint generation by integrating convolution neural network with feature pairwise conditional random field (FPCRF). *IEEE Trans. Geosci. Remote Sens.* **2020**, *58*, 7502–7519. [CrossRef]
4. Yang, N.; Tang, H. Semantic Segmentation of Satellite Images: A Deep Learning Approach Integrated with Geospatial Hash Codes. *Remote Sens.* **2021**, *13*, 2723. [CrossRef]
5. Yang, H.L.; Yuan, J.; Lunga, D.; Laverdiere, M.; Rose, A.; Bhaduri, B. Building extraction at scale using convolutional neural network: Mapping of the united states. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2018**, *11*, 2600–2614. [CrossRef]
6. Ji, S.; Wei, S.; Lu, M. Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set. *IEEE Trans. Geosci. Remote Sens.* **2018**, *57*, 574–586. [CrossRef]
7. Majd, R.D.; Momeni, M.; Moallem, P. Transferable object-based framework based on deep convolutional neural networks for building extraction. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2019**, *12*, 2627–2635. [CrossRef]
8. Xie, Y.; Zhu, J.; Cao, Y.; Feng, D.; Hu, M.; Li, W.; Zhang, Y.; Fu, L. Refined extraction of building outlines from high-resolution remote sensing imagery based on a multifeature convolutional neural network and morphological filtering. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2020**, *13*, 1842–1855. [CrossRef]
9. 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**, *28*, 91–99. [CrossRef] [PubMed]
10. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 8–10 June 2015; pp. 3431–3440.
11. 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 Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014; pp. 740–755.
12. Sun, L.; Tang, Y.; Zhang, L. Rural building detection in high-resolution imagery based on a two-stage CNN model. *IEEE Geosci. Remote Sens. Lett.* **2017**, *14*, 1998–2002. [CrossRef]
13. Yang, N.; Tang, H. GeoBoost: An incremental deep learning approach toward global mapping of buildings from VHR remote sensing images. *Remote Sens.* **2020**, *12*, 1794. [CrossRef]

14. Cheng, D.; Liao, R.; Fidler, S.; Urtasun, R. Darnet: Deep active ray network for building segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 7431–7439.
15. Xu, Y.; Wu, L.; Zhong, X.; Chen, Z. Building Extraction in Very High Resolution Remote Sensing Imagery Using Deep Learning and Guided Filters. *Remote Sens.* **2018**, *10*, 144. [CrossRef]
16. Huang, Z.; Cheng, G.; Wang, H.; Li, H.; Shi, L.; Pan, C. Building extraction from multi-source remote sensing images via deep deconvolution neural networks. In Proceedings of the IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Beijing, China, 10–15 July 2016; pp. 1835–1838.
17. Zhao, K.; Kang, J.; Jung, J.; Sohn, G. Building extraction from satellite images using mask R-CNN with building boundary regularization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Salt Lake City, UT, USA, 18–22 June 2018; pp. 247–251.
18. Wei, S.; Ji, S.; Lu, M. Toward automatic building footprint delineation from aerial images using cnn and regularization. *IEEE Trans. Geosci. Remote Sens.* **2019**, *58*, 2178–2189. [CrossRef]
19. Girard, N.; Smirnov, D.; Solomon, J.; Tarabalka, Y. Regularized Building Segmentation by Frame Field Learning. In Proceedings of the IGARSS 2020–2020 IEEE International Geoscience and Remote Sensing Symposium, Brussels, Belgium, 26 September 2020; pp. 1805–1808.
20. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, 5–9 October 2015; pp. 234–241.
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. Gur, S.; Shaharabany, T.; Wolf, L. End to end trainable active contours via differentiable rendering. *arXiv* **2019**, arXiv:1912.00367.
23. Zhao, W.; Persello, C.; Stein, A. Building outline delineation: From aerial images to polygons with an improved end-to-end learning framework. *ISPRS J. Photogramm.* **2021**, *175*, 119–131. [CrossRef]
24. Li, Z.; Wegner, J.; Lucchi, A. Topological map extraction from overhead images. In Proceedings of the IEEE International Conference on Computer Vision, Seoul, Korea, 27–31 October 2019; pp. 1715–1724.
25. Huang, W.; Tang, H.; Xu, P. OEC-RNN: Object-oriented delineation of rooftops with edges and corners using the recurrent neural network from the aerial images. *IEEE Trans. Geosci. Remote Sens.* **2021**. Online Early Access. [CrossRef]
26. Castrejon, L.; Kundu, K.; Urtasun, R.; Fidler, S. Annotating object instances with a polygon-RNN. *arXiv* **2017**, arXiv:1704.05548.
27. Acuna, D.; Ling, H.; Kar, A.; Fidler, S. Efficient interactive annotation of segmentation datasets with polygon-rnn++. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 859–868.
28. Liu, Y.; Cheng, M.-M.; Hu, X.; Wang, K.; Bai, X. Richer convolutional features for edge detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 3000–3009.
29. Xingjian, S.H.I.; Chen, Z.; Wang, H.; Yeung, D.-Y.; Wong, W.-K.; Woo, W. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 7–12 December 2015; pp. 802–810.
30. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
31. Lin, T.-Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.
32. Perazzi, F.; Point-Tuset, J.; McWilliams, B.; Van Gool, L.; Gross, M.; Sorkine-Hornung, A. A benchmark dataset and evaluation methodology for video object segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 724–732.
33. Maggiori, E.; Tarabalka, Y.; Charpiat, G.; Alliez, P. Can semantic labeling methods generalize to any city? The inria aerial image labeling benchmark. In Proceedings of the 2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Fort Worth, TX, USA, 23–28 July 2017; pp. 3226–3229.
34. Wu, S.-T.; da Silva, A.C.G.; Márquez, M.R.G. The Douglas-peucker algorithm: Sufficiency conditions for non-self-intersections. *J. Braz. Comput. Soc.* **2004**, *9*, 67–84. [CrossRef]