



A Novel Framework Based on Mask R-CNN and Histogram Thresholding for Scalable Segmentation of New and Old Rural Buildings

Ying Li ¹, Weipan Xu ¹, Haohui Chen ² , Junhao Jiang ¹ and Xun Li ^{1,*} 

¹ Department of Urban and Regional Planning, School of Geography and Planning, China Regional Coordinated Development and Rural Construction Institute, Urbanization Institute, Sun Yat-sen University, Guangzhou 510275, China; liying268@mail2.sysu.edu.cn (Y.L.); xuweipan@mail2.sysu.edu.cn (W.X.); jiangjh26@mail2.sysu.edu.cn (J.J.)

² Data61, Commonwealth Scientific and Industrial Research Organisation (CSIRO), Canberra 2601, Australia; caronhaohui.chen@data61.csiro.au

* Correspondence: lixun@mail.sysu.edu.cn

Abstract: Mapping new and old buildings are of great significance for understanding socio-economic development in rural areas. In recent years, deep neural networks have achieved remarkable building segmentation results in high-resolution remote sensing images. However, the scarce training data and the varying geographical environments have posed challenges for scalable building segmentation. This study proposes a novel framework based on Mask R-CNN, named Histogram Thresholding Mask Region-Based Convolutional Neural Network (HTMask R-CNN), to extract new and old rural buildings even when the label is scarce. The framework adopts the result of single-object instance segmentation from the orthodox Mask R-CNN. Further, it classifies the rural buildings into new and old ones based on a dynamic grayscale threshold inferred from the result of a two-object instance segmentation task where training data is scarce. We found that the framework can extract more buildings and achieve a much higher mean Average Precision (mAP) than the orthodox Mask R-CNN model. We tested the novel framework's performance with increasing training data and found that it converged even when the training samples were limited. This framework's main contribution is to allow scalable segmentation by using significantly fewer training samples than traditional machine learning practices. That makes mapping China's new and old rural buildings viable.

Keywords: deep learning; rural buildings; instance segmentation; Mask R-CNN; histogram thresholding



Citation: Li, Y.; Xu, W.; Chen, H.; Jiang, J.; Li, X. A Novel Framework Based on Mask R-CNN and Histogram Thresholding for Scalable Segmentation of New and Old Rural Buildings. *Remote Sens.* **2021**, *13*, 1070. <https://doi.org/10.3390/rs13061070>

Academic Editors: Sawaid Abbas, Janet E. Nichol, Faisal M. Qamer and Jianchu Xu

Received: 8 February 2021

Accepted: 8 March 2021

Published: 11 March 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

Monitoring the composition of new and old buildings in rural area is of great significance to rural development [1]. In particular, China's recent rapid urbanization has tremendously transformed its rural settlements over the last decades [2]. However, unplanned and poorly-documented dwellings have posed significant challenges for understanding rural settlements [3,4]. Traditionally, field surveys had been the major solutions, but they require intensive labour inputs and could be time-consuming, especially in remote areas. The recent breakthroughs of remote sensing technologies provide the growing availability of high-resolution remote sensing images such as low-altitude aerial photos and Unmanned Aerial Vehicle (UAV) images. That allows manual mappings of the rural settlements at a lower cost and with broader coverage, but they are still time-consuming. Therefore, to map the settlements for nearly 564 million rural population in China [5], a scalable, intelligent and image-based solution is urgently needed.

Remote sensing-based mapping of buildings has been a popular research topic for decades [6–10]. Since the launch of IKONO, QuickBird, WorldView and most recently

UAVs, the remote sensing images' spatial resolution grows significantly. That encourages applications of machine learning (ML) technologies in computer vision to push image segmentation forefronts. ML-based methods, such as Markov Random Fields [11], Bayesian Network [12], Neural Network [13], SVM [14], and Deep Convolution neural network (DCNN) [15] achieved impressive results. These supervised methods learn the spatial and hierarchical relationships between pixels and objects, utilizing remote sensing technologies' resolution gains in recent years. There are two kinds of image segmentation. The semantics segmentation, such as U-Net [16], SegNet [17], and DeepLab [18], treats multiple objects of the same class as a single entity. In contrast, instance segmentation treats multiple objects of the same class as distinct individual entities (or instances) [19,20]. The understanding of rural settlements would involve instance segmentation instead of semantic segmentation, as settlement's numbers and areas are needed. Amongst the instance segmentation methods, Mask Region-Based Convolutional Neural Network (R-CNN) has been unprecedentedly popular. It predicts bounding boxes for the target objects and then segments the objects inside the predicted boxes [21]. Scholars have used Mask R-CNN to carry out building extraction work and regularize the extraction results. Li et al. [22] improved Mask R-CNN by detecting key points to segment individual building and preserve geometric details. Zhao et al. introduced building boundary regularization to the orthodox Mask R-CNN, which benefited many cartographic and engineering applications [23]. In addition, some scholars have used Mask R-CNN for object detection [24] and mapping other land use objects, such as ice-wedge [25,26]. The empirical studies mostly focus on the optimization of object extraction but pay little attention to the challenges where training samples are scarce.

Compared to the urban buildings, rural ones attracted much less attention from the remote sensing community [27,28]. Only a few rural datasets were open to the public. The Wuhan dataset (WHU) [29] extracts 220,000 independent buildings from high-resolution remote sensing images covering 450 km² in Christchurch, New Zealand. The Massachusetts dataset consists of 151 aerial images of the Boston area, covering roughly 340 km² [30]. Training on these datasets, the ML-based models achieved impressive segmentation results [31–33]. However, these data only cover a relatively small area. As a result, the models might not generalize well in other regions where the geographical environments differ significantly. Therefore, the lack of training data specifically for rural environments and the varying geographical environments across regions have posed challenges for building a robust and generalized algorithm. To achieve human-like segmentation for China's vast and varying geography, we might need to build models dedicated to different regions. In this regard, the bottleneck is the manual effort for annotating a large amount of training data. In this study, we proposed a novel framework that could significantly reduce data annotation efforts while retaining the classification capability.

Humans annotate the new and old buildings in high-resolution remote sensing images by the difference of pixel color, because the histogram of grayscales would vary significantly across new and old buildings. When new and old buildings' grayscale histogram exhibits bimodal distributions, the valley point can be used as the threshold for discriminating them. This methodology is called histogram thresholding, which is widely used for image object extraction, such as the extraction of surface water area [34], built-up areas [35], and brain image fusion [36]. If all building footprints are given, a few predicted labels of new and old buildings in the same remote sensing image could validate if such a bimodal distribution exists and consequently find the valley point. In this regard, we can reduce the number of training samples while retaining the algorithm's capability. It should be noticed that the segmentation of buildings is more manageable than the segmentation of new and old ones in machine learning practices. That is partially due to the reduced efforts for labelling one class instead of two. Another reason is that binary classifiers are more accurate than multi-class classifiers in general. For example, classifying dogs is easier than classifying dog breeds. Therefore, the proposed framework uses histogram thresholding as an add-on to the state-of-the-art deep learning algorithm to achieve impressive segmentation results. In the methods section, we will address the proposed framework in detail. The proposed

framework's contribution to the building extraction research area is to achieve a promising classification capability while annotation efforts can be significantly reduced. This study uses rural areas in Xinxing County, Guangdong Province, as the case study to test the proposed framework's performance.

2. Study Area and Data

2.1. Study Area

To test the proposed framework's performance, we collected data samples from high-resolution satellite images covering rural Xinxing County, Guangdong Province, China (see Figure 1). Xinxing is a traditional mountainous agricultural county, with a large agricultural population and a relatively complete landscape, forest, and land city. Moreover, Xinxing is a rural revitalization pilot area and has made much rural development and governance achievements [37]. The extraction of new and old buildings is of great significance for understanding rural development in Xinxing.

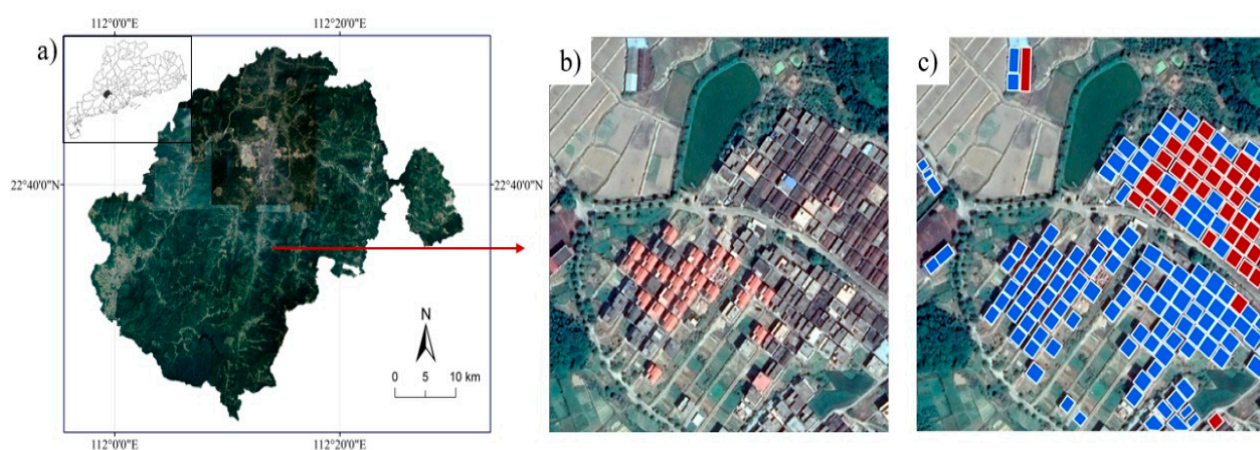






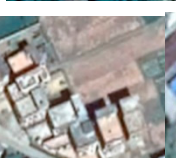
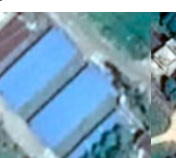




Figure 1. The study area. (a) The location of Xinxing in Guangdong Province, China, (b) a village of Xinxing (c) building labels of the same region in panel c (new and old buildings are masked in blue and red respectively).

2.2. Data Collection and Annotation

Table 1 shows the new and old buildings in high-resolution satellite images. Most of the new buildings are brick-concrete structures, with roofs made of cement or colored tiles. Old houses are mainly Cantonese-style courtyards in Xinxing, where the roof materials are dark tiles. Moreover, the outline of their footprints is less clear than new houses. New buildings are mostly distributed along the streets, while the old buildings still retain a compact comb pattern.

Table 1. The image characteristics of new and old buildings in rural areas.

Type	Image Characteristics				
old buildings					
new buildings					

For model training purposes, we collected 68 images with a resolution of 0.26 m. Each image has a size ranging from 900×900 to 1024×1024 pixels in the RGB color space. We use the open-source image annotation tool VIA [38] to delineate the building footprints (see Figure 2). We edited and checked all the building samples of the original vector file using the ArcGISTM software to produce a high-quality dataset. All building samples from those 68 images were compiled as dataset called one-class samples. We annotated only 34 out of 60 images with new and old labels (called two-class samples hereafter). Finally, the annotated imagerys were randomly divided into a training, validation and test set (see Table 2).

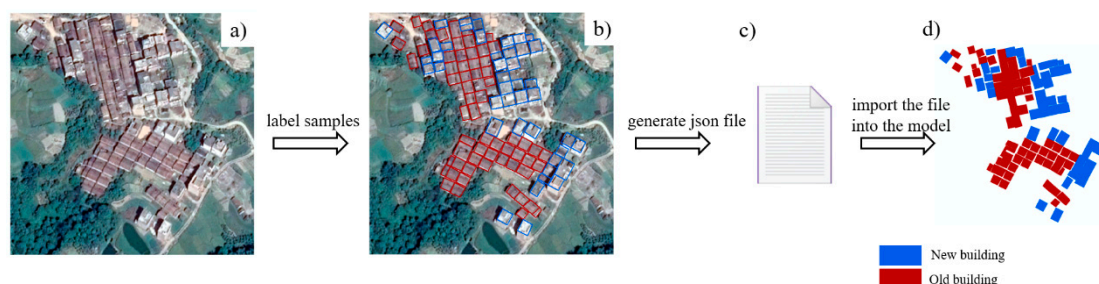


Figure 2. The construction of image dataset. (a) a village image; (b) the building labels; (c) the label json file; (d) the mask with building categories.

Table 2. The training, validation, and test dataset.

	One-Class Samples		Two-Class Samples	
	Buildings	New Buildings	Old Buildings	Total
training	54 pic/5359 poly	1340	1081	20 pic/2421 poly
validation	8 pic/817 poly	439	378	8 pic/817 poly
test	6 pic/892 poly	462	454	6 pic/906 poly
total	68 pic/7068 poly	2241	1913	34 pic/4154 poly

3. Methods

3.1. HTMask R-CNN

Mask R-CNN [21] has been proven to be a powerful and adaptable model in many different domains [23,39]. It operates in two phases, generation of region proposals and classification of each generated proposal. In this study, we use Mask R-CNN as our baseline model for benchmarking. As discussed before, we propose a novel segmentation framework that can utilize the histogram thresholding and deep learning's image segmentation capability to extract the new and old rural buildings. We call the proposed framework HTMask R-CNN, abbreviating Histogram Thresholding Mask R-CNN. The workflow of the framework is addressed as follows (see Figure 3 for illustration):

- We built two segmentation models (one-class model and two-class model) based on the one-class and two-class samples' training sets (Figure 3a). The one-class model can extract rural buildings, while the two-class model can classify new and old rural buildings. All models used the pre-trained weights trained on the COCO dataset [40] as the base weights.
- An satellite image (Figure 3b) is classified by the one-class and two-class model separately, leading to a map of building footprints (R1 in Figure 3c), and a map of new and old buildings (R2 in Figure 3d).
- Grayscale histograms were built using the pixels from the new and old building footprints (R2). The average grayscale levels for new and old buildings were computed as N and O, respectively (Figure 3e). A valley point is determined by $\theta = (N + O) / 2$.
- The valley point θ is used as the threshold to determine the type of building in R1. Finally, we get a map of the old and new buildings in R3 (Figure 3f).

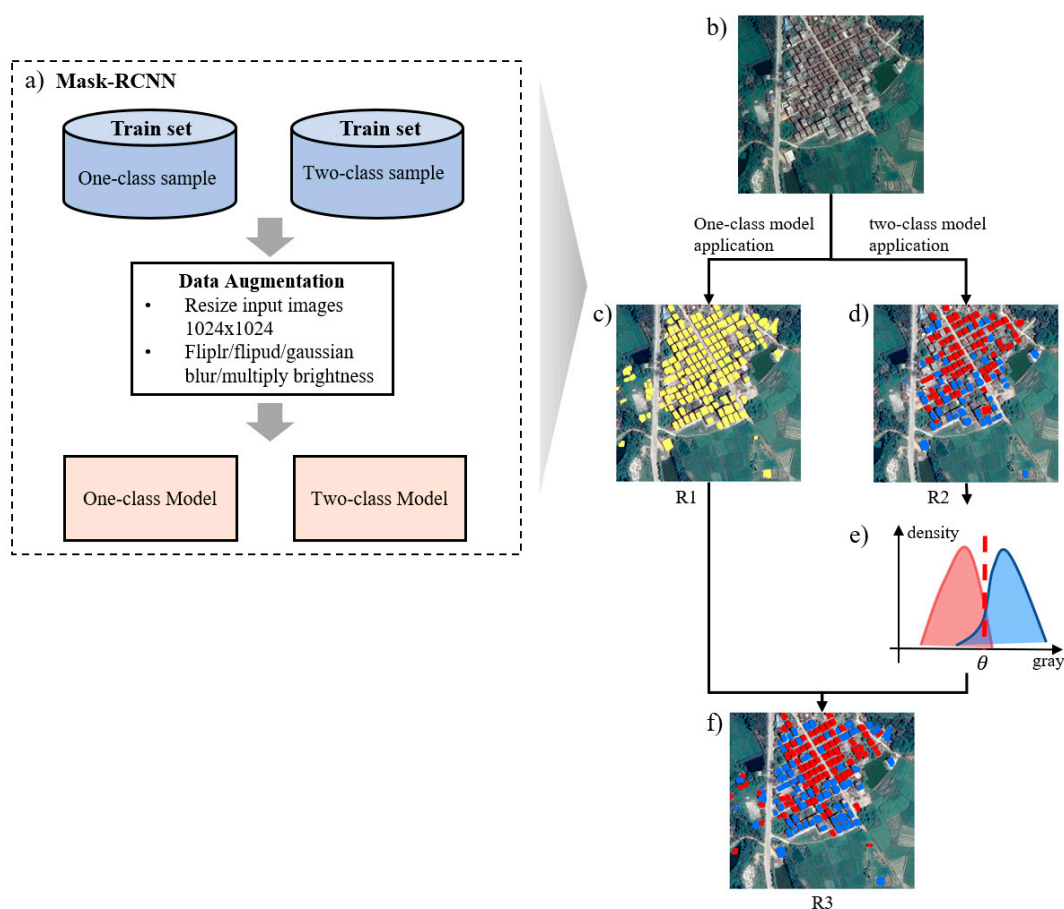


Figure 3. The workflow of HTMask R-CNN (Histogram Thresholding Mask R-CNN). (a) the training process of two segmentation models based on the one-class and two-class samples' training sets; (b) a sample village image in the validation and test set; (c) R1: the prediction result of the one-class model; (d) R2: the prediction result of the two-class model; (e) the calculation of the threshold from the grayscale histogram; (f) R3: the prediction result of HTMask R-CNN.

The hypothesis is that R3 performs better than R2. Specifically, R3 can take advantage of the capability of R1, while utilizing the grayscale difference of the new and old buildings in R2. The two-class model's performance depends on the numbers of training samples. Assumably, the more the training data are added, the more robust the network training, the better the segmentation results. This study also tests how the numbers of training samples could affect R2 and R3's performance to evaluate how HTMask R-CNN can save the annotation efforts while retaining the segmentation capability.

3.2. Experiment

We used R2, the prediction results of the two-class model, as the benchmarking. R3 is the result of the proposed framework. We compared R2 and R3 to test how much accuracy improvements in the proposed framework.

We randomly selected 50% of the images from the one-class and two-class training set for data augmentations, resulting in 1.5 times of the original training size. The augmentations included rotating, mirroring, brightness enhancement, and adding noise points to the images. In the training stage for the one-class and two-class models, 50 epochs with two batches per epoch were applied, and the learning rate was set at 0.0001. The Stochastic Gradient Descent (SGD) optimization algorithm was adopted as the optimizer [41]. We set the weight decay to 0.000. The loss function is shown in Equation (S1). The learning momentum was set at 0.9, which was used to control to what extent the model remains in the original updating direction. We used cross-entropy as a loss function to evaluate the train-

ing performance. We performed hyperparameters tuning and the settings addressed above achieved the best performance (refer to Table S2 for details of hyperparameters tuning).

To test how HTMask R-CNN can achieve a converged performance with a limited amount of training data, the training process has involved an incremental number of samples (from 5 to 20 satellite images). Afterward, we compared the baseline Mask R-CNN and the HTMask R-CNN by comparing R2 and R3.

3.3. Accuracy Assessment

We use the average precision (AP) to quantitatively evaluate our framework on the validation dataset. The AP is equal to taking the area under the precision-recall (PR) curve, Equation (1). The mAP_{50} represents the AP value when the threshold of intersection over union (IoU) is 0.5.

$$\begin{aligned} \text{Precision} &= TP / (TP + FP) \\ \text{Recall} &= TP / (TP + FN) \\ AP &= \int_0^1 P(R) dR \end{aligned} \quad (1)$$

IoU means the ratio of intersection and union of the prediction and the reference. When a segmentation image is obtained, the value of IoU is calculated according to Equation (2).

$$IoU = TP / (TP + FP + FN) \quad (2)$$

4. Results

Figure 4 shows the result of an example image Site1 (Figures S1 and S2 presents additional examples for other sites with gradually increasing building density). In terms of the building footprints mapping, the one-class model has identified most of the buildings. More importantly, it can accurately outline individual buildings and the boundaries of adjacent buildings being correctly separated, which allows the texture of the building to be captured.

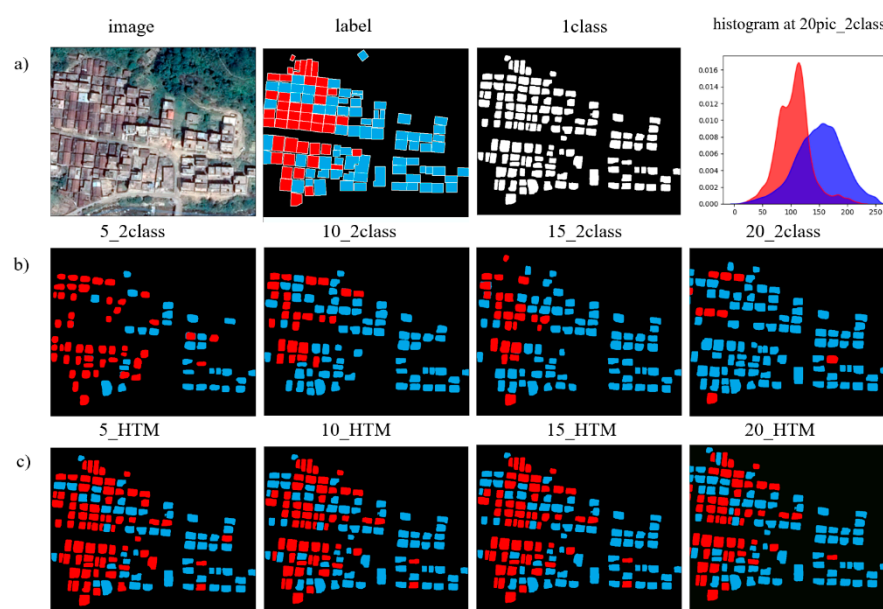


Figure 4. The result comparison of Site 1 between the baseline Mask R-CNN model and the HTMask R-CNN framework. (a) the satellite image from the test set, the second column shows the annotations, the third column shows the result of one-class model (R1), and the fourth column shows the grayscale histograms of the new and old building footprints from R2 when training images = 20. (b) the result of the two-class model (R2), with incremental training samples from 5 to 20. (c) the result of the HTMask R-CNN framework (R3) with incremental training samples.

The baseline model (two-class model) performed better and better with the growing numbers of training samples in building extraction (Figure 4b). However, the numbers of buildings in R2 are still significantly fewer than R1, especially if the buildings are very dense (see Site3), which aligns with our assumption. In R3, the proposed framework uses R1 as the base map, so the numbers of buildings are equal between R1 and R3. That means the proposed framework outperforms the baseline model. In terms of the new and old building segmentation, R3 is significantly better than R2 at all levels of training samples. When the number of training samples is very limited, e.g., five, the baseline model nearly misidentified most of the new and old buildings, while the proposed framework still produced a reasonable result.

Figure 5 shows the performance of the one-class model. We noticed that the performance of the one-class model converges at the 25th epoch, where it identified most of the buildings. The mAP_{50} could reach 0.70.

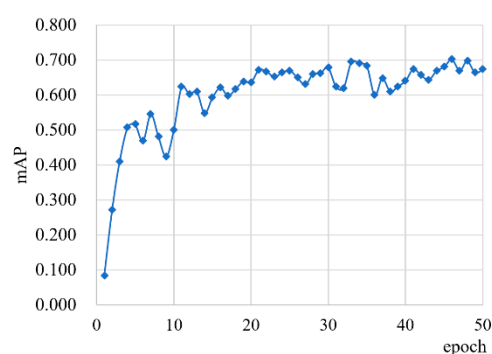


Figure 5. The mAP_{50} of one-class model on the test dataset at different training iterations.

The baseline two-class model and the HTMask R-CNN also converge at the 25th epoch (Table 3; Figure 6). When the training size is small (image_num = 5), the mAP_{50} of the baseline two-class model is very low (0.24), while the HTMask R-CNN can significantly improve the recognition (0.46). With the increasing training size, the baseline two-class model and HTMask R-CNN's performance gap become narrower (see Figure 6d). Finally, the mAP_{50} of the baseline two-class reached 0.51 when the training size is 20. More importantly, HTMask R-CNN performs consistently ($mAP_{50} \approx 0.48$), no matter which levels of training size.

Table 3. The mAP_{50} between R2 and R3 at all levels of training.

Models\ Epoch	1	5	10	15	20	25	30	35	40	45	50
5_2 class	0.00	0.01	0.09	0.16	0.17	0.14	0.16	0.22	0.22	0.24	0.24
5_HTM	0.04	0.31	0.33	0.41	0.45	0.45	0.48	0.47	0.45	0.48	0.46
10_2 class	0.00	0.06	0.18	0.29	0.33	0.35	0.37	0.37	0.39	0.41	0.41
10_pic_HTM	0.05	0.37	0.32	0.43	0.45	0.46	0.48	0.49	0.46	0.48	0.47
15_2 class	0.00	0.11	0.28	0.29	0.30	0.35	0.39	0.39	0.44	0.42	0.44
15_pic_HTM	0.05	0.35	0.34	0.41	0.46	0.47	0.48	0.49	0.46	0.49	0.48
20_2 class	0.02	0.16	0.27	0.40	0.43	0.46	0.42	0.49	0.45	0.50	0.51
20_HTM	0.05	0.36	0.34	0.42	0.45	0.47	0.48	0.50	0.46	0.49	0.48

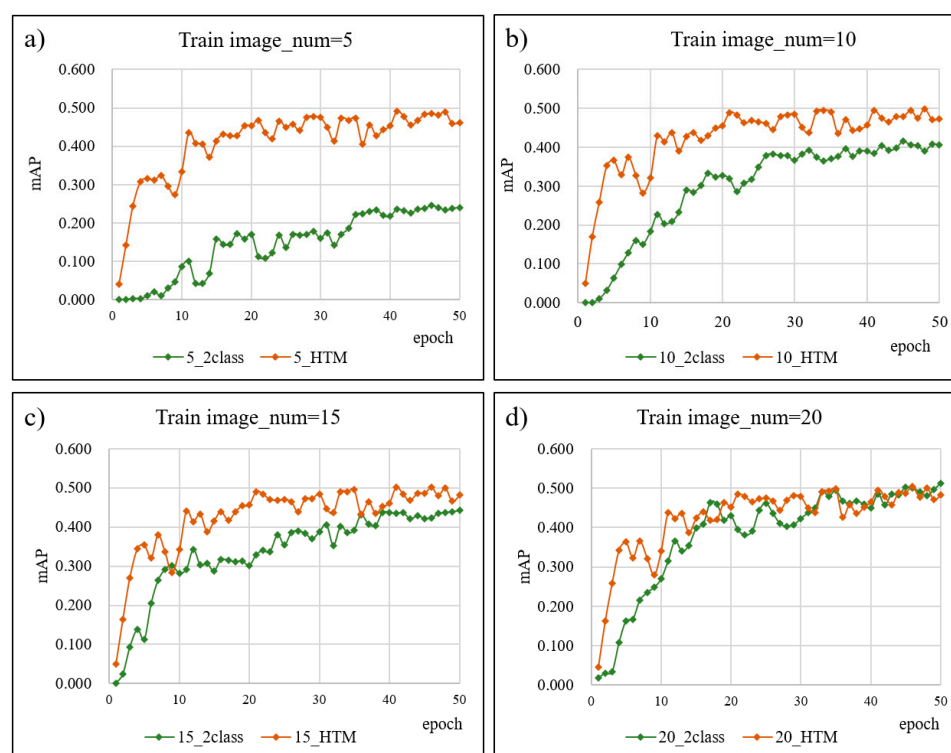


Figure 6. The mAP_{50} for each model. (a). sample size = 5; (b) sample size = 10; (c) sample size = 15; (d) sample size = 20.

It confirms our assumption that HTMask R-CNN can perform well in a small number of samples, which means that it can significantly reduce annotation efforts while retaining the segmentation capability. In contrast, the baseline two-class model performed poorly in extracting old-new two-category buildings.

5. Discussions

With the advance of deep learning, the extraction of building footprint from satellite imagery has made notable progress, contributing significantly to settlements' digital records. However, the scarcity of training data has always been the main challenge for scaling building segmentation. Therefore, this study proposes a novel framework based on the Mask R-CNN model and histogram thresholding to extract old and new rural buildings even when the label is scarce. We tested the framework in Xinxing County, Guangdong Province, and achieved promising results. This framework provides a viable solution for mapping China's rural buildings at a significantly reduced cost.

Mask R-CNN models have been proven useful in many applications. However, this study found that the orthodox Mask R-CNN model performed poorly in extracting old-new two-category buildings. When the training samples are limited, the mAP_{50} is only 0.24, respectively. We believe the varying geographical environments lead to the poor generalization of the segmentation model when the training samples cannot cover the most distinctive spatial and spectrum features. For instance, the model might not classify a building with an open patio as either a new or old building if none of the training samples contains this unique shape. Meanwhile, the single-category classification task using Mask R-CNN could reach mAP_{50} at 0.70, respectively. That means utilizing one-class model's capability in mapping building footprints could improve the recall rate for the old-new two-category classification task, especially in high-density areas. Hence, we propose such a novel framework.

While tested the framework with increasing training samples, we found that it converges at a very early stage when the numbers of training images are only five. That means the framework can be applied on a large scale, to map all rural buildings in China.

Before then, more careful studies should be undertaken to understand the limitations of the framework. We have applied the framework in Fuliang County, Jiangxi Province, China, and found that the performance of the framework is worse than the benchmarking two-class model R2 (see Figure S4). When the histogram of the R2 result does not exhibit a clear valley, the pixel grayscale of the new and old buildings is similar. In this regard, the thresholding method loses its advantage against the orthodox model. In this case, the prediction should come from the output of R2.

Moreover, polygons produced from the proposed framework have irregular shapes, slightly different from the building footprint boundaries. Therefore, downstream regularization is needed in future studies. The recent advances of multi-angle imaging technologies and vision technologies integrated with deep learning that emerged in civil engineering provide new opportunities in the 3D reconstruction of rural building models [42,43].

6. Conclusions

Nearly half of the Chinese population live in the rural areas of China. The lack of a digital record of the new and old buildings has posed challenges for the governments to realize the socio-economic state. Under China's central government's current rural revitalization policy, many migrant workers will return to the villages. Therefore, a scalable, intelligent, and accurate building mapping solution is urgently needed. The proposed framework in this study achieved a promising result even when the training samples are scarce. As a result, we can scale the mapping process at a significantly reduced cost. Therefore, we believe this framework could map every settlement in the rural areas, help policymakers establish a longitudinal digital building record, and monitor socio-economics across all rural regions.

Supplementary Materials: The following are available online at <https://www.mdpi.com/2072-4292/13/6/1070/s1>, Figure S1: The result comparison of Site 2 between the baseline Mask R-CNN model and the HTMask R-CNN framework, Figure S2: The result comparison of Site 3 between the baseline Mask R-CNN model and the HTMask R-CNN framework, Figure S3: The feature maps of the three sites in the two-class model at the 50th epoch, Figure S4: The baseline Mask R-CNN model and the HTMask R-CNN framework were tested in Fuliang County, Jiangxi Province, China, Figure S5: The loss function per training iteration, Table S1: The mAP₇₅ between R2 and R3 at all levels of training, Table S2: Tuning hyperparameters, Equation S1: Loss function of the two-class model R2.

Author Contributions: Conceptualization, X.L.; Data curation, Y.L. and W.X.; Formal analysis, J.J.; Methodology, Y.L. and H.C.; Project administration, W.X. and X.L.; Resources, X.L.; Validation, H.C.; Visualization, J.J.; Writing – original draft, Y.L. and H.C. All authors reviewed and edited the draft, approved the submitted manuscript, and agreed to be listed and accepted the version for publication. 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 (41971157), the Key R&D Programs in Guangdong Province aligned with Major National Science and Technology Projects of China (2020B0202010002)

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study. Written informed consent has been obtained from the patient(s) to publish this paper.

Data Availability Statement: The data presented in this study are available at <https://github.com/liying268-sysu/HTM-R-CNN>, accessed on 8 February 2021.

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

References

1. Zhao, X.; Sun, H.; Chen, B.; Xia, X.; Li, P. China's rural human settlements: Qualitative evaluation, quantitative analysis and policy implications. *Ecol. Indic.* **2018**, *105*, 398–405. [CrossRef]
2. Yang, R.; Xu, Q.; Long, H. Spatial distribution characteristics and optimized reconstruction analysis of China's rural settlements during the process of rapid urbanization. *Rural Stud.* **2016**, *47*, 413–424. [CrossRef]
3. Kuffer, M.; Pfeffer, K.; Sliuzas, R. Slums from space—15 years of slum mapping using remote sensing. *Remote Sens.* **2016**, *8*, 455. [CrossRef]
4. Kuffer, M.; Persello, C.; Pfeffer, K.; Sliuzas, R.; Rao, V. Do we underestimate the global slum population? Joint Urban Remote Sensing Event (JURSE). *IEEE* **2019**, 2019, 1–4.
5. National Bureau of Statistics of China. *China Statistical Yearbook 2018*; China Statistics Press: Beijing, China, 2019.
6. Patino, J.E.; Duque, J.C. A review of regional science applications of satellite remote sensing in urban settings. *Comput. Environ. Urban Syst.* **2013**, *37*, 1–17. [CrossRef]
7. Jin, X.; Davis, C.H. Automated building extraction from high-resolution satellite imagery in urban areas using structural, contextual, and spectral information. *EURASIP J. Adv. Signal Process.* **2005**, 2005, 745309. [CrossRef]
8. Huang, X.; Zhang, L. Morphological building/shadow index for building extraction from high-resolution imagery over urban areas. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2011**, *5*, 161–172. [CrossRef]
9. Ghanea, M.; Moallem, P.; Momeni, M. Building extraction from high-resolution satellite images in urban areas: Recent methods and strategies against significant challenges. *Int. J. Remote Sens.* **2016**, *37*, 5234–5248. [CrossRef]
10. Bachofer, F.; Braun, A.; Adamietz, F.; Murray, S.; D'Angelo, P.; Kyazze, E.; Mumuhire, A.P.; Bower, J. Building stock and building typology of kigali, rwanda. *Data* **2019**, *4*, 105. [CrossRef]
11. Tupin, F.; Roux, M. Markov random field on region adjacency graph for the fusion of SAR and optical data in radar grammetric applications. *IEEE Trans. Geosci. Remote Sens.* **2005**, *43*, 1920–1928. [CrossRef]
12. Zhang, L.; Ji, Q. Image segmentation with a unified graphical model. *IEEE Trans. Pattern Anal. Mach. Intell.* **2010**, *32*, 1406–1425. [CrossRef] [PubMed]
13. Kurnaz, M.N.; Dokur, Z.; Ölmez, T. Segmentation of remote-sensing images by incremental neural network. *Pattern Recognit. Lett.* **2005**, *26*, 1096–1104. [CrossRef]
14. Mitra, P.; Uma Shankar, B.; Pal, S.K. Segmentation of multispectral remote sensing images using active support vector machines. *Pattern Recognit. Lett.* **2004**, *25*, 1067–1074. [CrossRef]
15. Audebert, N.; Le Saux, B.; Lefèvre, S. *Semantic Segmentation of Earth Observation Data Using Multimodal and Multi-Scale Deep Networks*; Lai, S., Lepetit, V., Nishino, K., Sato, Y., Eds.; Asian Conference on Computer Vision; Springer: Cham, Germany, 2016; pp. 180–196. [CrossRef]
16. Guo, M.; Liu, H.; Xu, Y.; Huang, Y. Building extraction based on U-Net with an attention block and multiple losses. *Remote Sens.* **2020**, *12*, 1400. [CrossRef]
17. Chen, H.; Lu, S. Building Extraction from Remote Sensing Images Using SegNet. In Proceedings of the 2019 IEEE 4th International Conference on Image, Vision and Computing (ICIVC), Xiamen, China, 5–7 July 2019; pp. 227–230.
18. Murphy, K.; Yuille, A.L. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *40*, 834–848.
19. Arnab, A.; Zheng, S.; Jayasumana, S.; Romera-Paredes, B.; Larsson, M.; Kirillov, A.; Savchynskyy, B.; Rother, C.; Kahl, F.; Torr, P.H. Conditional random fields meet deep neural networks for semantic segmentation: Combining probabilistic graphical models with deep learning for structured prediction. *IEEE Signal Process. Mag.* **2018**, *35*, 37–52. [CrossRef]
20. Pan, Z.; Xu, J.; Guo, Y.; Hu, Y.; Wang, G. Deep Learning Segmentation and Classification for Urban Village Using a Worldview Satellite Image Based on U-Net. *Remote Sens.* **2020**, *12*, 1574. [CrossRef]
21. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask r-cnn. In Proceedings of the IEEE International Conference on Computer Vision; 2017; pp. 2961–2969. Available online: https://openaccess.thecvf.com/content_iccv_2017/html/He_Mask_R-CNN_ICCV_2017_paper.html (accessed on 8 February 2021).
22. Li, Q.; Mou, L.; Hua, Y.; Sun, Y.; Jin, P.; Shi, Y.; Zhu, X.X. Instance segmentation of buildings using keypoints. *arXiv* **2020**, arXiv:2006.03858.
23. Zhao, K.; Kang, J.; Jung, J.; Sohn, G. Building Extraction from Satellite Images Using Mask R-CNN With Building Boundary Regularization. In *CVPR Workshops*; IEEE: New York, NY, USA, 2018; pp. 247–251. [CrossRef]
24. Mahmoud, A.; Mohamed, S.; El-Khoribi, R.; Abdelsalam, H. Object Detection Using Adaptive Mask RCNN in Optical Remote Sensing Images. *Int. Intell. Eng. Syst.* **2020**, *13*, 65–76.
25. Zhang, W.; Liljedahl, A.K.; Kanevskiy, M.; Epstein, H.E.; Jones, B.M.; Jorgenson, M.T.; Kent, K. Transferability of the deep learning mask R-CNN model for automated mapping of ice-wedge polygons in high-resolution satellite and UAV images. *Remote Sens.* **2020**, *12*, 1085.
26. Bhuiyan, M.A.E.; Witharana, C.; Liljedahl, A.K. Use of Very High Spatial Resolution Commercial Satellite Imagery and Deep Learning to Automatically Map Ice-Wedge Polygons across Tundra Vegetation Types. *J. Imaging* **2020**, *6*, 137. [CrossRef]
27. Kaiser, P.; Wegner, D.; Lucchi, A.; Jaggi, M.; Hofmann, T.; Schindler, K. Learning Aerial Image Segmentation from Online Maps. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 6054–6068. [CrossRef]

-
28. 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); 2017; pp. 3226–3229. [\[CrossRef\]](#)
 29. 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. [\[CrossRef\]](#)
 30. Mnih, V. *Machine Learning for Aerial Image Labeling*; University of Toronto (Canada): Toronto, ON, Canada, 2013.
 31. Wang, S.; Hou, X.; Zhao, X. Automatic building extraction from high-resolution aerial imagery via fully convolutional encoder-decoder network with non-local block. *IEEE Access* **2020**, *8*, 7313–7322. [\[CrossRef\]](#)
 32. Kang, W.; Xiang, Y.; Wang, F.; You, H. EU-net: An efficient fully convolutional network for building extraction from optical remote sensing images. *Remote Sens.* **2019**, *11*, 2813. [\[CrossRef\]](#)
 33. Chen, M.; Wu, J.; Liu, L.; Zhao, W.; Tian, F.; Shen, Q.; Zhao, B.; Du, R. DR-Net: An Improved Network for Building Extraction from High Resolution Remote Sensing Image. *Remote Sens.* **2021**, *13*, 294. [\[CrossRef\]](#)
 34. Sekertekin, A. A survey on global thresholding methods for mapping open water body using Sentinel-2 satellite imagery and normalized difference water index. *Arch. Comput. Methods Eng.* **2020**, 1–13. [\[CrossRef\]](#)
 35. Li, C.; Duan, P.; Wang, M.; Li, J.; Zhang, B. The Extraction of Built-up Areas in Chinese Mainland Cities Based on the Local Optimal Threshold Method Using NPP-VIIRS Images. *J. Indian Soc. Remote Sens.* **2020**, 1–16. [\[CrossRef\]](#)
 36. Srikanth, M.V.; Prasad, V.; Prasad, K.S. An improved firefly algorithm-based 2-d image thresholding for brain image fusion. *Int. J. Cogn. Inform. Nat. Intell. (IJCINI)* **2020**, *14*, 60–96. [\[CrossRef\]](#)
 37. Qi, Z. Rural revitalization in Xinxing County. *China Econ. Wkly.* **2018**, 78–79. (In Chinese)
 38. Dutta, A.; Zisserman, A. The VIA annotation software for images, audio and video. In Proceedings of the 27th ACM International Conference on Multimedia; Association for Computing Machinery: New York, NY, USA, 2019; pp. 2276–2279.
 39. Wu, T.; Hu, Y.; Peng, L.; Chen, R. Improved Anchor-Free Instance Segmentation for Building Extraction from High-Resolution Remote Sensing Images. *Remote Sens.* **2020**, *12*, 2910. [\[CrossRef\]](#)
 40. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Zitnick, C.L.; Dollár, P. *Microsoft Coco: Common Objects in Context*; European Conference on Computer Vision; Springer: Cham, Germany, 2014; pp. 740–755.
 41. Bottou, L. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010*; Physica-Verlag HD: Heidelberg, Germany, 2010; pp. 177–186.
 42. Luo, L.; Tang, Y.; Zou, X.; Ye, M.; Feng, W.; Li, G. Vision-based extraction of spatial information in grape clusters for harvesting robots. *Biosyst. Eng.* **2016**, *151*, 90–104. [\[CrossRef\]](#)
 43. Tang, Y.; Li, L.; Wang, C.; Chen, M.; Feng, W.; Zou, X.; Huang, K. Real-time detection of surface deformation and strain in recycled aggregate concrete-filled steel tubular columns via four-ocular vision. *Robot. Comput. Integr. Manuf.* **2019**, *59*, 36–46. [\[CrossRef\]](#)