A CNN-BASED CHANGE DETECTION METHOD FOR SQUATTER STRUCTURE RECOGNITION FROM AERIAL IMAGES AND DSM

Min Zhang [1, 2], Wenhao Li [1], Wenzhong Shi [1, 2, *]

[1] The Department of Land Surveying and Geo-Informatics, The Hong Kong Polytechnic University, Hong Kong, 999077, lsgi-min.zhang@polyu.edu.hk, wen-hao.li@connect.polyu.hk, lswzshi@polyu.edu.hk
[2] Otto Poon Charitable Foundation Smart Cities Research Institute, The Hong Kong Polytechnic University, Hong Kong, 999077

Commission III, WG III/7

KEY WORDS: Squatter Structure, Deep Learning, Digital Surface Model, Change Detection, Attention Mechanism

ABSTRACT:

Squatter structures have been a serious threat to human safety and health for a long time. And monitoring their changes is important to facilitate government management of squatters. However, existing methods are still not automatic, accurate and fast enough to meet the actual needs of practical applications. In this paper, we propose a novel deep learning-based method to detect squatter structure changes from bi-temporal remotely sensed (RS) images and digital surface models (DSMs). The proposed convolutional neural network (CNN) takes the advantages of the spectral information from high resolution image and the height information from the DSM, so as to detect changes more accurately in type and height of squatter structures. Moreover, we create a data set for deep learning model training, covering a variety of squatter structures in Hong Kong. Compared with three existing representative methods, Our model performs the best, with Kappa of 0.6786 and 0.6458 in the detection results of the two test regions, respectively, which indicates that it has application potential.

1. INTRODUCTION

The squatter problem is a socialization problem caused by the inability of cities to meet the needs of rapid population growth. It usually occurs on the outskirts of fast-developing cities. In Hong Kong, squatter structures generally refer to illegally occupied structures or temporary residences on government or private land. Since 1940, a large number of immigrants from the mainland have poured into Hong Kong, and the population living in squatter areas once reached a quarter of the total (Wong, 1978). The ensuing public health and public safety issues have seriously hindered the development of Hong Kong. Most memorably, the Shek Kip Mei fire in 1953 caused the displacement of more than 58,000 people. After the fire, the Hong Kong government fundamentally changed its housing policy and began to build public housing to provide housing benefits for the lower class, aiming to reduce squatter settlements.

In recent years, the government has strengthened squatter control measures, but the survey of squatter structures has always been a difficult problem to solve in the management of squatters. There are two commonly used methods, i.e., field survey and visual interpretation based on remotely sensed (RS) images (Smart, 2001). The former requires staff to go to the squatter area and take photos for evidence, while the latter requires staff to identify new squatter structures and monitor the ones that need to be demolished from multi-period RS images. These methods usually achieve good accuracy but require a great deal of times, manpower, and material resources, and it is difficult to monitor squatter structures in a large area, especially in a timely manner. Therefore, developing automatic methods to identify squatter structures from RS images can help improve efficiency, but unfortunately, there is currently a lack of research on this issue.

The squatter structure can be regarded as a kind of building, and to some extent, existing building extraction method have the potential to identify it. But unlike ordinary buildings, squatter structure is simple in components and small in size, mostly made of iron sheets, wooden boards, and containers. Moreover, its various styles and textures make it difficult to distinguish it from ordinary buildings. To address this problem, a deep learning-based method is proposed in this paper to detect squatter structure changes from bi-temporal RS images and digital surface models (DSMs). And our experiments shows that the proposed method achieves better performance than several representative deep learning-based change detection methods.

Generally, the changes of buildings can be divided into three categories, i.e., changes in building height, changes in building into other land cover, and vice versa, and the same is true for the changes of squatter structure. Existing building extraction and change detection methods mainly focus on ordinary buildings. To detect changes in building height, the DSM is the key component, which can usually be generated by satellite stereo images, oblique aerial photographs, and airborne lidar point cloud data. For example, Tian et al. (Tian et al., 2013) proposed a building change detection method based on satellite stereo images and DSMs, which improved the accuracy by removing non-building areas such as vegetation and shadows. Experiments show that the method has good performance but is limited by the quality of DSM. Huang et al. (Huang et al., 2020) developed an automated method to monitoring newly constructed building areas from multi-view Ziyuan-3 satellite images. The method extracted various features of the building, including planar features, i.e., MBI (Huang and Zhang, 2011b), HARRIS (Harris and Stephens, 1988), and PENTAX (Pesaresi et al., 2008), and vertical features, i.e., MSI (Huang and Zhang, 2011a), nDSM (Qin and Fang, 2014), and MABI (Liu et al., 2019). It achieved good results in Shanghai and Beijing test areas, but not good for reconstructed buildings, mainly because some complex changes, such as

* Corresponding author

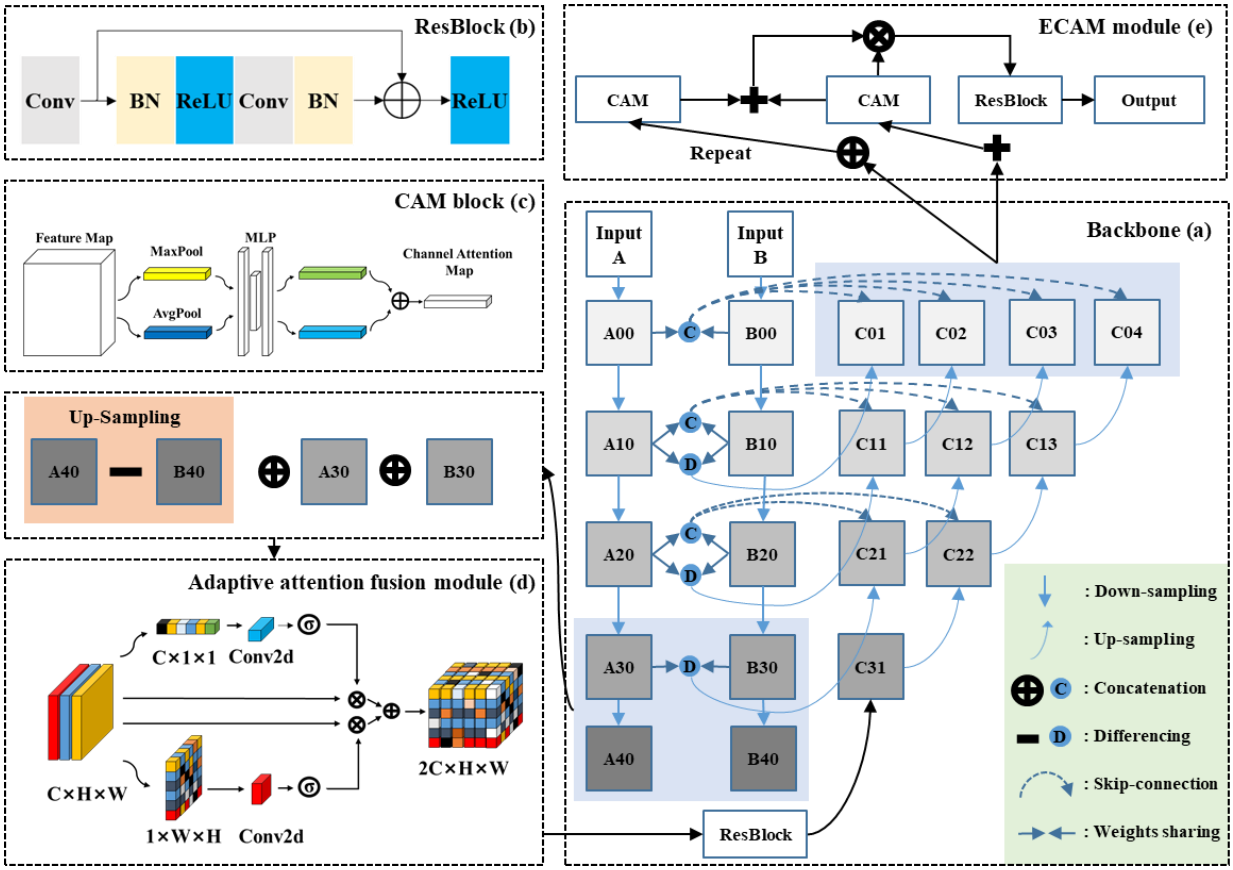

Figure 2. The proposed network architecture. Part (a) is the backbone of the model; Part (b) is the ResBlock; Part (c) is the CAM block used in Part (e), the ECAM; Part (d) illustrates the adaptive attention fusion module.

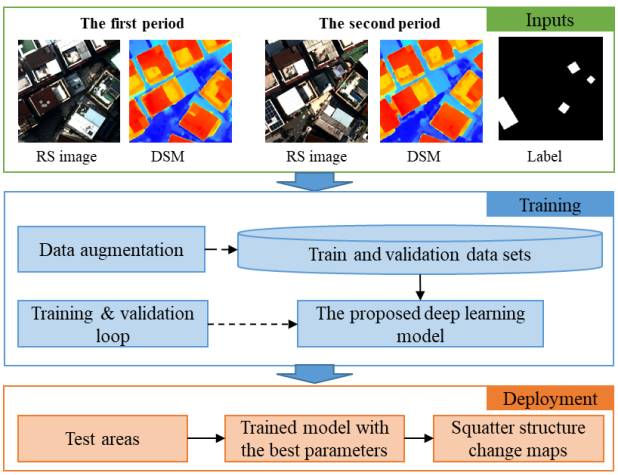

Figure 1. The schematic workflow of proposed method for squatter structure change detection.

changes caused by vegetation around buildings and on top of buildings, are not well considered.

Compared with traditional change detection methods mentioned above, deep learning-based methods can effectively eliminate the dependence of detection results on the difference map. Advantageously, it can process RS data obtained by multiple sensors with strong robustness and applicability. Zhu et al. (Zhu et al., 2018) innovatively introduced the SegNet into building change detection and Liu et al. (Liu et al., 2020) employed a dual attention module (DAM) to improve feature recognition capabilities and implemented a change detection loss to solve the problem of sample imbalance. It has good detection accuracy on the WHU building change detection data set. To achieve fine-grained building change detection, Sun et al. (Sun et al., 2020) proposed a multi-task learning framework (MTL-CD) from high-resolution RS image captured by unmanned aerial vehicle (UAV). And it benefits two semantic segmentation tasks in the framework, the model has good detection results on the Guangzhou dataset their proposed. Besides, to update the Serbian cadastral information system and detect illegal buildings, an object-based and pixel-based change detection method was proposed to detect newly built, modified or demolished buildings (Jovanović et al., 2021).

These deep learning-based methods can achieve good performance in the corresponding building change detection tasks, but their use in squatter structure change detection is mainly limited by the low performance of small object recognition and low learning ability of complex features of squatter structures. In other words, none of these methods are optimized for squatter structure detection. To detect squatter structure changes from RS images more accurately and automatically, a new deep learning-based method is proposed in this paper. The main contributions are as follows:

(i) We employ deep learning techniques to address this issue for the first time, and subsequently, produce a data set with diverse changes of squatter structures.

(ii) A new CNN-based model is proposed, which uses an adaptive attention fusion module (AAFM) and an ensemble channel attention module (ECAM) to enhance the extraction of

spatial and channel information, thereby suppressing useless information and speeding up the convergence of the model.

Finally, the experiments were conducted in a typical squatter area in Yuen Long, Hong Kong, and the results show that the proposed method achieves good performance.

2. METHOD

The schematic workflow of proposed method for squatter structure change detection is shown in Figure 1. Two period RS images and DSMs are used to generate a data set for all deep learning model training. With data augmentation, the model is trained and verified in a loop to find the best model parameters. Finally, the trained model will be deployed for the practical application. In general, well-designed network architecture is the key to good performance of the proposed method.

2.1 Network architecture

The schematic diagram of the proposed network architecture is presented in Figure 2, where part (a) shows the backbone of the model. It consists of two branches with a shared weight, which is made up of residual blocks (ResBlocks), as shown in part (b). Between each ResBlock of the feature encoder A (i.e., A00-A40) and B (i.e., B00-B40), a max pooling layer is used for down-sampling. After each down-sampling, the number of input channels will be doubled, and the image size will be half of the previous ResBlock. The overall structure is inspired by UNet++ (Zhou et al., 2018), which uses a lot of dense skip connections. The backbone of our model as a feature extractor needs to learn multiscale and different semantic information. In addition, many differencing operations are used in up-sampling stage to detect changes, so that the model can learn the differential information of height and spectral information easily and directly.

The spectral and texture information of the squatter structures are diverse, and the background environment are very complex, especially in high-resolution aerial images, with many noises caused by shadows, occlusion, or illumination differences. To identify the key features and suppress unimportant features from the large amount of extracted features, an attention mechanism is adopted. Specifically, two types of attention modules are embedded into the proposed model: i) adaptive attention fusion module (AAFM), used in the up-sampling process, and ii) ensemble channel attention module (ECAM), used in feature fusion. They are explained separately in the following sections.

2.2 Adaptive attention fusion module

The diagram of the AAFM is shown in Figure 2 (d), which is proposed by (Wang et al., 2021). Since the input of the model is the concatenation of RS image and DSM, the extraction of information between channels is particularly important. On the other hand, many squatter structures are small in size, it is very crucial to distinguish their features from those of the background environment, which means that richer information needs to be extracted both spectrally and spatially. Thus, AAFM is designed to solve this issue, which allows the model to pay more attention to the important features and suppress redundant information.

In our implementation, an AAFM consists of two sub-modules, i.e., channel attention module and spatial attention module. The kernel size of the convolution layer in the two attention modules is changed based on the number of channels to achieve multi-scale and multilevel information extraction. The channel attention module assigns different weights to each channel by training. Its output is the channel attention feature map, and can be calculated by:

$$y_1 = \sigma\left(Conv1\left(Avgpool(x)\right)\right) \otimes x \tag{1}$$

where x is the input feature map, σ is the sigmoid activation function, $\otimes$ means element-wise multiplication. The kernel size of the convolution *Conv1* is calculated by the number of input channels. The formula can be written as:

$$kernel_1 = \left|\frac{log_2(C) + b}{a}\right|_{odd} \tag{2}$$

where C is the number of channels of the input feature map. Here a and b are hyper parameters and set to 2 and 1 respectively in our experiments. Unlike the channel attention module, spatial attention module is used to learn the importance of each pixel. The output spatial attention feature map can be calculated by:

$$y_2 = \sigma\left(Conv2\left(Maxpool(x)\right)\right) \otimes x \tag{3}$$

Similarly, the kernel size of the convolution of *Conv2* is calculated by the number of input channels, that is:

$$kernel_2 = \left|\frac{log_2(C) + b}{a}\right|_{odd} \tag{4}$$

here, a and b are set to 2 and 3 respectively. Finally, the generated channel feature maps and spatial feature maps are concatenated by channels and then output as the attention fusion maps.

2.3 Ensemble channel attention module

In the last part of the model, four sets of output feature maps (i.e., *C01-C04*), which have the same size, but different semantic and spatial representations, are input to the ECAM module, as it is shown in Figure 2 (e). This is inspired by (Fang et al., 2021). Features derived from shallow layers contain fine-grained features and more accurate position information, while features obtained from deep layers have coarse-grained features and more semantic information. Therefore, the fusion of multi-feature maps requires a module to identify useful information from such feature maps, and ECAM can do this well.

As shown in Figure 2 (c), ECAM is a fusion module based on the traditional CAM (Woo et al., 2018), and its basic idea is deep supervision and ensemble learning. It can be formulated as:

$$X = CAM(C01 + C02 + C03 + C04) \tag{5}$$

$$Y = CAM(C01 \oplus C02 \oplus C03 \oplus C04) \tag{6}$$

$$out = Conv2d\left(X \otimes \left(Y + repeat_{(4)}(Y)\right)\right) \tag{7}$$

where *C01*, *C02*, *C03*, *C04* are the output of feature maps, $repeat_{(4)}(Y)$ represents the operation of repeating the attention map Y 4 times and then concatenating them in channel.

2.4 Loss function

Generally, there is a low probability of squatter structure changes, that is, the unchanged pixels are far more than changed. Thus, a hybrid loss function combining dice loss and BCE loss is used in this task. Dice loss performs very well in scenes with serious

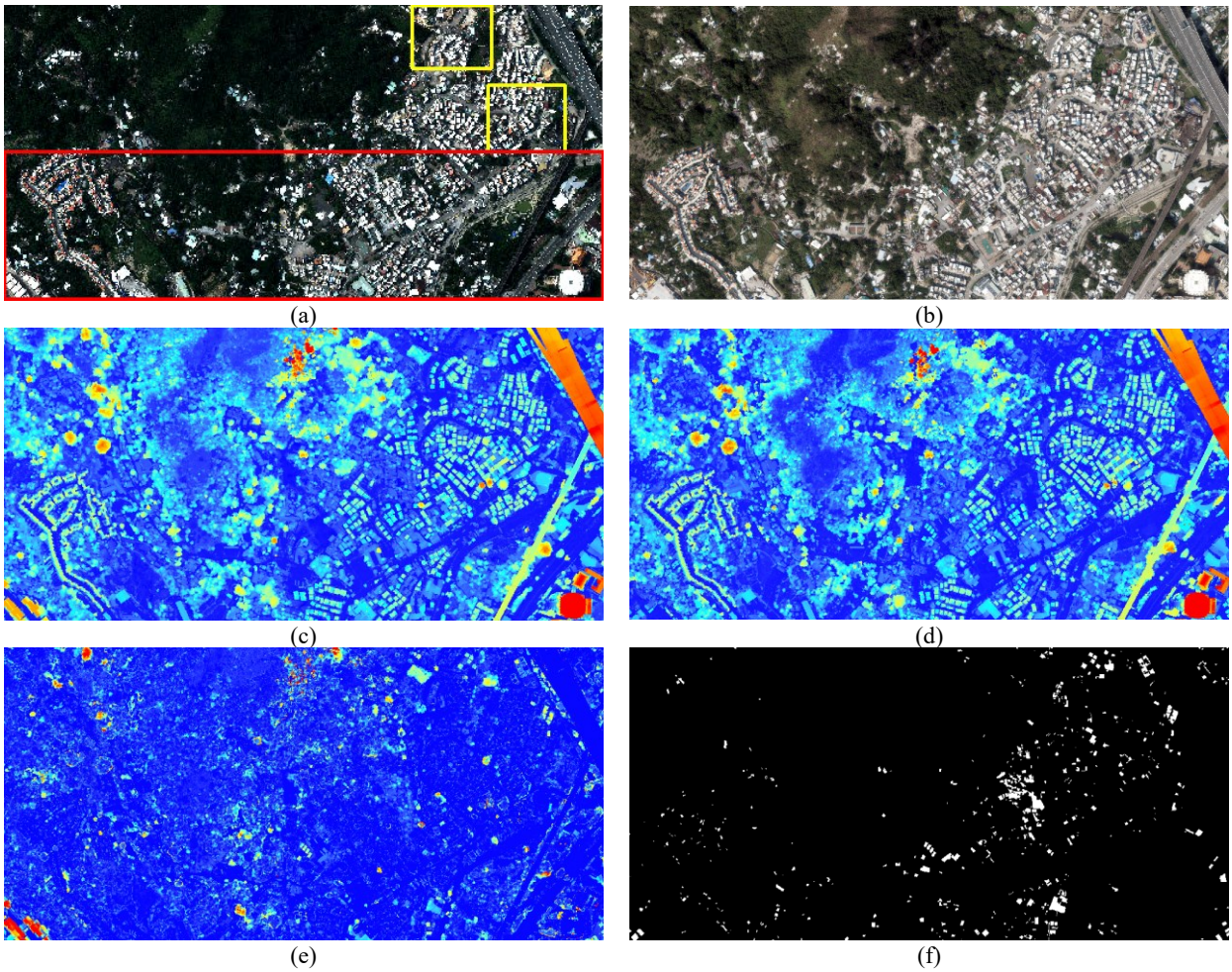

Figure 3. The experimental area. (a) The first period RS image. (b) The second period RS image. (c) The first period DSM. (d) The second period DSM. (e) The difference DSM. (f) Ground truth.

imbalance of positive and negative samples, while BCE loss helps the training to be more stable. The hybrid loss is defined as:

$$Loss = Loss_{Dice} + Loss_{BCE} \tag{8}$$

where dice loss can be written as:

$$L_{Dice} = 1 - \frac{2|X \cap Y|}{|X| + |Y|} \tag{9}$$

where $|X|$ and $|Y|$ are the detected change result and ground truth respectively, $|X \cap Y|$ denotes the intersection of $|X|$ and $|Y|$, representing the area located in both X and Y. And BCE loss is written as:

$$Loss_{BCE} = -\frac{1}{N}\sum_{i=1}^{N}[y_i\, log(p_i) + (1 - y_i)\, log(1 - p_i)] \tag{10}$$

where N is the total number of samples, y_i is the ground truth value of the *i-th* sample, p_i is the predicted value of the *i-th* sample.

3. EXPERIMENTS

3.1 Experimental area

The experiment area is in Yuen Long, Hong Kong, with an area of 874,398 m^2. It can be seen from Figure 3, there are many squatters in this area, which has attracted the attention of the government for a long time. Two period DSMs and aerial photos with 4 bands (i.e., red, green, blue, and infer red) are used in our study, collected in 2016 and 2020 respectively. The ground truth map is manually labeled. As shown in Figure 3 (a), the region in red rectangle is selected as the training set, and the rest is used to generate the validation set. For better performance analysis, two test sites with typical squatter structure changes are selected, see the region in the yellow rectangle in Figure 3 (a).

3.2 Data set generation

In our experiment, the input RS image and DSM size of the model are both 256×256 pixels. Thus, the selected training area are cropped to this size with a 50% horizontal and vertical overlap. Consider the varying sizes of squatter structures, we down-sample the original image by two scales, and then crop the same, the spatial resolution is 2 times and 4 times that of the original image. For the validation area, there is no overlap used when cropping. Finally, 9565 samples for training and 1536 for validation are generated. For each training sample, it consists of two RS images, two DSMs and a binary map reflecting the squatter structures changes.

3.3 Implementation detail

The proposed method is implemented by Pytorch with python 3.6 and trained by a workstation with two NVIDIA Quadro GV100 GPUs. AdamW optimizer with an initial learning rate of 0.001 is

Method	Test area 1				Test area 2			
	Precision	Recall	F1-Score	Kappa	Precision	Recall	F1-Score	Kappa
FC-Siam-diff	78.4565	33.0358	0.4624	0.4399	71.1353	29.715	0.4173	0.3951
CDNet	77.2666	62.1839	0.6885	0.6684	73.2972	58.8555	0.6524	0.6321
MFPNet	80.831	58.9604	0.6813	0.6619	75.559	55.4034	0.6389	0.6189
Ours	77.6685	**63.7139**	**0.6982**	**0.6786**	75.3145	**59.8831**	**0.6652**	**0.6458**
w/o AAFM	82.5631	43.5269	0.5691	0.5479	82.1488	44.1121	0.5732	0.5521
w/o ECAM	**82.9638**	53.4904	0.6502	0.6304	**82.8532**	53.4223	0.6494	0.6299

Table 1. Quantitative results of experimental comparison methods in the two test areas.

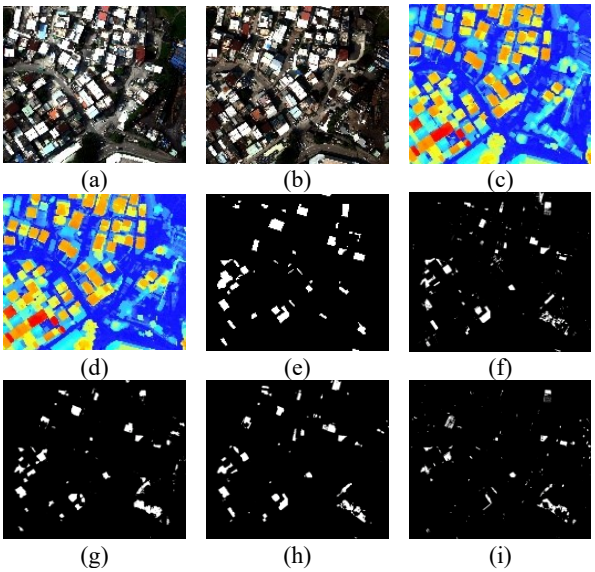

Figure 4. Squatter structures change detection results in test area 2. (a) The first period RS image. (b) The second period RS image. (c) The first period DSM. (d) The second period DSM. (e) Ground truth. The change maps of (f) ours, (g) CDNet, (h) MFPNet, and (i) FC-Siam-diff.

adopted in the training process and the learning rate decays to half of the previous one after every 20 iterations. The training process is terminated after 150 epochs. The batch size is set to 32. Online data augmentation is also applied during training, including random rotation, random cropping, random scaling, and flipping horizontally and vertically.

We selected three representative change detection models for comparative experiments. They are i) CDnet (Wang et al., 2014), a classical model for change detection, which is a combination of FCN and Siamese network; ii) FC-Siam-diff (Daudt et al., 2018), a combination of UNet and Siamese network; and iii) MFPNet (Lei et al., 2021), the latest state-of-the-art (SOTA) model with various novel modules. In the experiments, the training data and training strategies of all methods are the same for fairness.

To quantitatively compare the performance of change maps generated by the different models, four accuracy evaluation metrics are used, i.e., precision, recall, F1-score, and Kappa. F1-score is the harmonic mean of precision and recall. It considers both precision and recall. Kappa coefficient can detect whether the model prediction results are consistent with the true value. Thus, the overall performance of the model can be reflected by the F1-Score and Kappa coefficients.

3.4 Result

As we can see from the Table 1, the proposed model achieves optimal values in the recall, F1-score and kappa coefficient.

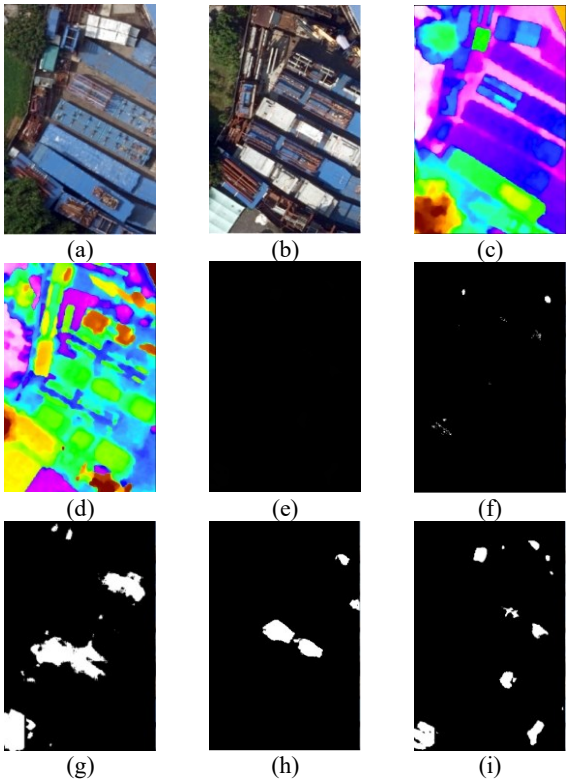

Figure 5. The results in a parking lot. (a)-(d) The input data. (e) Ground truth. The change maps of (f) ours, (g) CDNet, (h) MFPNet, and (i) FC-Siam-diff.

Compared with the result of CDNet and MFPNet, the Kappa reaches 67.82% in test area 1. And in test area 2, our method shows similar advantages. Both Kappa and F1-score are higher than the remaining three methods. FC-Siam-diff gets the best precision in test area 1, but the recall rate is the lowest. It can be seen clearly from the Figure 4 that the change map of the proposed model has the best performance. According to visual interpretation, there are four missing objects in the result of our model in terms of object-level accuracy. However, the CDNet, which has the closest recall rate to the proposed model, also has 9 changed objects undetected.

Since the geometric and spectral characteristics of containers, the fronts of large trucks, and large vans are very similar to those of squatters in RS images, the models are prone to misdetecting these changes as squatter changes. As shown in Figure 5, there is an area where quite a few super-long trucks are parked. Unlike the other three models that have a large number of false positives, the proposed model has almost none. In addition, the attention module enables the proposed model to detect edges better than the other three models. Thanks to better use of differential information, the proposed model can detect many changed areas that CDNet and MFPNet fail to detect. Moreover, compared with

FC-Siam-diff, our results are better due to the use of AAFM.

Ablation studies were also performed on the ECAM and AAFM modules for the proposed method. The results show that the ECAM module improves the Kappa value of the method on the test area A by 0.04, while the AAFM module can improve it by 0.13. Furthermore, with the help of ECAM, the proposed method performs better on the detection of small changes and edges of the changed area.

Overall, the proposed model can detect most of the changes of squatter structures and achieve good edge extraction for independent ones. Due to the variety of squatter structure changes, it is difficult for training sets to cover all types of changes. Changes that occur in the test areas may not be similar to any changes in the training set. This is one of the reasons why the recall is not very high. Although the proposed model performs better than others in our experiment, there is still room for improvement in the detection of some small changes.

4. CONCLUSION

Changes in squatter structures generally involve legal issues, such as unauthorized construction on government or private land. And these unpredictable changes pose problems for the government's management. Therefore, we propose an efficient and time-saving method to solve this problem. This method is based on the latest deep learning techniques and able to detect squatter structure changes from RS images and DSM automatically. Compared with other deep learning models, our well-designed model achieves the highest F1-score and kappa coefficients in both test regions, the F1-score is at least 1.5% higher than the second best model. That means it has good practical value. In our future work, considering the incompleteness of extracting changed objects, we will improve the recall rate of the method by using morphological operations and random field models, and design an new loss function for edge detection to improve the boundary accuracy.

ACKNOWLEDGEMENTS

We would like to thank the Survey and Mapping Office of the Lands Department in Hong Kong for providing aerial photos described in this paper.

REFERENCES

Daudt, R.C., Le Saux, B., Boulch, A., 2018. Fully convolutional siamese networks for change detection, 2018 25th IEEE International Conference on Image Processing (ICIP). IEEE, pp. 4063-4067.

Fang, S., Li, K., Shao, J., Li, Z., 2021. SNUNet-CD: A Densely Connected Siamese Network for Change Detection of VHR Images. IEEE Geoscience and Remote Sensing Letters. 19, 1-5.

Harris, C., Stephens, M., 1988. A combined corner and edge detector in Alvey vision conference. 1988. Manchester, UK.

Huang, X., Cao, Y., Li, J., 2020. An automatic change detection method for monitoring newly constructed building areas using time-series multi-view high-resolution optical satellite images. Remote Sens. Environ. 244, 111802.

Huang, X., Zhang, L., 2011a. Morphological building/shadow index for building extraction from high-resolution imagery over urban areas. IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens. 5, 161-172.

Huang, X., Zhang, L., 2011b. A multidirectional and multiscale morphological index for automatic building extraction from multispectral GeoEye-1 imagery. Photogrammetric Engineering & Remote Sensing 77, 721-732.

Jovanović, D., Gavrilović, M., Sladić, D., Radulović, A., Govedarica, M., 2021. Building Change Detection Method to Support Register of Identified Changes on Buildings. Remote Sens. 13, 3150.

Lei, T., Zhang, D., Wang, R., Li, S., Zhang, W., Nandi, A.K., 2021. MFP‐Net: Multi‐scale feature pyramid network for crowd counting. IET Image Processing. 15(14), 3522-3533

Liu, C., Huang, X., Zhu, Z., Chen, H., Tang, X., Gong, J., 2019. Automatic extraction of built-up area from ZY3 multi-view satellite imagery: Analysis of 45 global cities. Remote Sens. Environ. 226, 51-73.

Liu, Y., Pang, C., Zhan, Z., Zhang, X., Yang, X., 2020. Building Change Detection for Remote Sensing Images Using a Dual-Task Constrained Deep Siamese Convolutional Network Model. IEEE Geoscience and Remote Sensing Letters 18, 811-815.

Pesaresi, M., Gerhardinger, A., Kayitakire, F., 2008. A robust built-up area presence index by anisotropic rotation-invariant textural measure. IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens. 1, 180-192.

Qin, R., Fang, W., 2014. A hierarchical building detection method for very high resolution remotely sensed images combined with DSM using graph cut optimization. Photogrammetric Engineering & Remote Sensing 80, 873-883.

Smart, A., 2001. Unruly places: Urban governance and the persistence of illegality in Hong Kong's urban squatter areas. American Anthropologist 103, 30-44.

Sun, Y., Zhang, X., Huang, J., Wang, H., Xin, Q., 2020. Fine-Grained Building Change Detection From Very High-Spatial-Resolution Remote Sensing Images Based on Deep Multitask Learning. IEEE Geoscience and Remote Sensing Letters. 19, 1-5.

Tian, J., Cui, S., Reinartz, P., 2013. Building change detection based on satellite stereo imagery and digital surface models. IEEE Trans. Geosci. Remote Sensing 52, 406-417.

Wang, D., Chen, X., Jiang, M., Du, S., Xu, B., Wang, J., 2021. ADS-Net: An Attention-Based deeply supervised network for remote sensing image change detection. International Journal of Applied Earth Observation and Geoinformation 101, 102348.

Wang, Y., Jodoin, P.-M., Porikli, F., Konrad, J., Benezeth, Y., Ishwar, P., 2014. CDnet 2014: An expanded change detection benchmark dataset, Proceedings of the IEEE conference on computer vision and pattern recognition workshops, pp. 387-394.

Wong, L.S.K., 1978. Housing in Hong Kong: A Multi-disciplinary Study. Heinemann Educational Books (Asia).

Woo, S., Park, J., Lee, J.-Y., Kweon, I.S., 2018. Cbam: Convolutional block attention module, Proceedings of the European conference on computer vision (ECCV), pp. 3-19.

Zhou, Z., Siddiquee, M.M.R., Tajbakhsh, N., Liang, J., 2018. Unet++: A nested u-net architecture for medical image segmentation, Deep learning in medical image analysis and multimodal learning for clinical decision support. Springer, pp. 3-11.

Zhu, B., Gao, H., Wang, X., Xu, M., Zhu, X., 2018. Change detection based on the combination of improved SegNet neural network and morphology, 2018 IEEE 3rd International Conference on Image, Vision and Computing (ICIVC). IEEE, pp. 55-59.