Comparison Between UNet, Modified UNet and Dense-Attention Network (DAN) for Building Extraction from TripleSat Imagery

Xingjue Wang (1), Xiaoman Huang (1), Chong Chen (1), Bin Zhou (1), Jianjun He (2), Ting Chen (2)

[1] Twenty First Century Aerospace Technology (Asia) Pte. Ltd., 61 Science Park Road, #05-17 The Galen, Singapore Science Park II, Singapore 117525
[2] Twenty First Century Aerospace Technology Co., Ltd., No. 26 Jiancaicheng East Road, Haidian District, Beijing 100096, China
Email: wangxj@21at.sg; huangxm@21at.sg; chenchong@21at.sg; zhoubin@21at.sg; hejj@21at.com.cn; chenting@21at.com.cn

KEY WORDS: Building extraction; TripleSat; deep learning; UNet; dense-attention network (DAN)

ABSTRACT: Building extraction from high resolution remote sensing imagery is of great importance for land use analysis, urban planning, and many other applications. Notably convolutional neural networks (CNN) have shown significant advantage over traditional methods for this task. Among various CNN models, UNet has gained popularity due to its simplicity, efficiency and robustness while many modified versions have been proposed. More recently, a model called the dense attention network (DAN) based on DenseNets and attention mechanism was proposed. This model achieved good performance in building extraction from very high resolution imagery.
Based on these developments, in this paper, we compared three architectures (UNet, modified UNet (with residual blocks and recurrent feature), and DAN) for building extraction in Kuala Lumpur, Malaysia using 0.8m TripleSat imagery. For the modified UNet, skip connections were implemented in each encoder blocks to mix features of different levels. Output was multiplied to input and feed to the same UNet again. The comparison results showed that the modified UNet achieved the highest F1-score, while the DAN achieved average higher F1-score than the UNet. But DAN had the highest accuracy for validation patches with large buildings.

1. INTRODUCTION

Building extraction is one of the most important tasks in remote sensing. The automation of building extraction based on high spatial resolution remote sensing images can save significant labor cost in many areas such as urban planning, change detection, disaster management, land cover mapping and population estimation. However, there are currently many challenges due to variation of building shape and size, variation of color and texture of rooftop, and similar band values to bare soil. Researchers have tried to tackle these issues with various methods.

Traditional building extraction methods are usually unsupervised. The common workflow is image segmentation followed by post processing. The building pixel segmentation is the core part which is achieved by using some of the common building characteristics such as uniform color, high contrast to surrounding, morphological characteristics, brightness, and shadow. Certain index such as morphological building index, morphological shadow index and texture-derived built-up presence index (PanTex) were invented in order to quantitatively represent typical characteristic of building shape and shadow (Huang and Zhang, 2011; Pesaresi et al., 2008). Level set, mean shift, watershed, k-means or active contour based image segmentation were combined in various ways to distinguish building pixels from backgrounds (Jiang et al., 2008; Song and Shan, 2008; Yang and Wang, 2012). However, the traditional methods are usually limited by certain criteria and specific types of buildings and hard to be generalized.

Deep learning methods gain great popularity recently in many computer vision tasks such as image classification, object detection and semantic segmentation. By stacking convolutional layers and max-pooling layers, the deep learning model is able to collect and process features of different levels hierarchically and for semantic segmentation of buildings, with proper decoders, it is able to recover building contour even in complex shape.

UNet, proposed in Ronneberger et al. (2015), is one of the most important deep learning models for semantic segmentation. It mainly consists of a sequence of encoder blocks and decoder blocks with skip connection similar to fully convolutional networks. However, for skip connections, it applies concatenation instead of addition making it more efficient in processing features. The architecture is simple yet effective in segmentation problems. Its simplicity and robustness soon give it great popularity in segmentation applications in remote sensing. Based on UNet, people developed various modified versions for building extraction and achieved good results. For example, Iglovikov et al. (2017) won the third place in Defence Science and Technology Laboratory (DSTL) Satellite Imagery Feature Detection challenge run by Kaggle with standard U-Net architecture but modified activation function and

loss function. The accuracy in term of IoU for the building was 0.7453 for public evaluation set and 0.6290 for private evaluation set. In 2017, Chhor et al. applied UNet mainly on task of building extraction and achieved F1-score around 0.75. In 2018, Hamaguchi and Hikosaka proposed a multi-task UNet model on building extraction based on UNet. Specifically, the multi-task model consisted of a shared feature extractor and successive multitask branches for specific building sizes in order to solve building size variation problem. They also included a branch to detect road. The average F1-score reached was 73.91%.

Based on UNet, researches tried to come up with more efficient models. There are many modifications of UNet which are proven to be useful. Residual connection is one of the most important and efficient modifications (He et al., 2016). By introducing residual connections in each encoder decoder blocks, the UNet is able to stack more layers without the vanishing gradient problem. The residual connection also enables mixing of low level and high level features and improve generalization significantly. In 2018, Hamaguchi and Hikosaka compared the VGG-U-Net and Res-U-Net and concluded that there was a significant accuracy boost by applying the residual connection. Aside from the residual connection, at the end of 2018, another interesting model was proposed which is called the stack UNet (Sevastopolsky et al., 2018). It applies two UNet connected in sequence. The output of the first UNet is concatenated with input image for the second UNet. The benefit is that the first UNet can provide additional information for the second UNet and the second UNet can refine the segmentation result. The model was used in optic disc cup segmentation and it was reported to outperform UNet and other state-of-the-art methods.

Taking the idea of ResNet one step further, DenseNet was proposed by replacing identity mapping to concatenation for shortcut connections (Huang et al., 2016). Within each block in DenseNet, each layer directly connects to all subsequent layers in the way of concatenation. This setup can process features from different levels more efficiently, which enables the network to achieve the same or better accuracy with much fewer parameters. Later, DenseNet was extended to Fully Convolutional DenseNets for semantic segmentation problems and achieved state of the art accuracy in CamVid and Gatech dataset (Jégou et al., 2016). Several work has been done to explore the usage of DenseNet in satellite image segmentation problems (Li et al., 2018). In the end of 2018, Dense-Attention Network (DAN) was proposed for the task of building extraction (Yang et al., 2018). In addition to the similar but simplified structure of fully convolutional DenseNet structure, it also introduced an attention mechanism which used the higher-level semantic information to re-weight the low-level information in skip connections. The proposed model achieved average F1-score around 92.5% on the ISPRS 2D semantic labeling contest (Potsdam) with 5cm spatial resolution. The accuracy of DAN was reported to outperform Deeplab-V3 which had average F1-score of 83.36%.

In this paper, we implemented some proven modifications to UNet to improve the model, and compared the UNet, the modified UNet and the DAN in 0.8m resolution TripleSat images. The objective is to give us more insights into model selection for building extraction task using 0.8m resolution satellite imagery.

2. DATASET AND METHODS

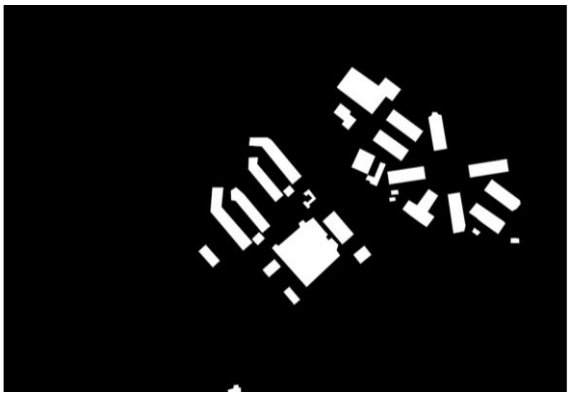
2.1 Dataset

We used 4 TripleSat scenes acquired in June 2016 covering the great Kuala Lumpur area with spatial resolution of 0.8m in this study. The image data has four bands: red, green, blue and near-infrared (NIR). The TripleSat constellation was launched on July 10, 2015 and operated by the Twenty First Century Aerospace Technology Co. Ltd. (21AT), a commercial Earth observation satellite operator based in Beijing, China. The constellation's three satellites are spaced 120 degrees apart, orbiting the Earth in a sun-synchronous orbit at 651km. They deliver 3.2m multi-spectral and 0.8m panchromatic imagery. The datasets used in this paper are 0.8m multi-spectral image product that has been pansharpened.

64 training patches were selected to contain various types of buildings such as big factory, row houses, villas, high rise buildings in urban area, small and medium sized single houses. The size of buildings ranges from $70m^2$ to $60000\,m^2$. The distance between nearby buildings can be as small as 3 pixels which caused difficulty in distinguishing nearby buildings. For each patch, all building polygons were manually delineated in QGIS software and rasterized to produce the ground truth masks. The training patches covered total area of around $63km^2$ and were cut into 167 768x768 smaller patches for model input.

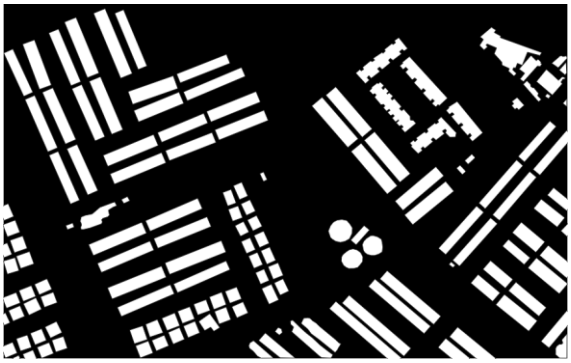
For validation dataset, 8 patches and corresponding ground truth masks were created with total area of $7.2km^2$. Among the 8 validation patches, patch 5 and 6 contain mainly big factories while the rest of the patches mainly contain small and medium sized single houses and row houses. Figure 1 shows some examples of training patches and their ground truth masks.

(a) Original (villas)

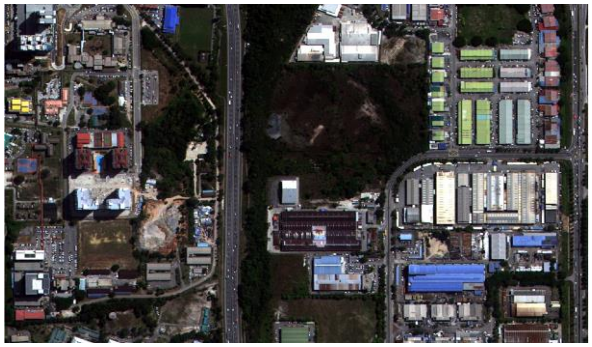


(b) Ground Truth

(c) Original(row houses)



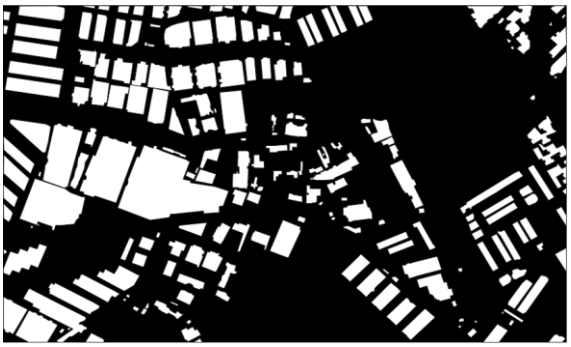
(d) Ground Truth



(e) Original (big factories)

(f) Ground Truth

(g) Original (big factories)



(h) Ground Truth

Figure 1. Examples of training patches and ground truth masks.

2.2 Methods

We compared the performance of the UNet, the modified UNet, and the DAN in building extraction in the KL area. The experiment was conducted so that the three models were compared in a relatively fair way.

UNet

For the UNet, in order to match the building size, the UNet model was designed to have 5 encoders block and 5 decoder blocks. Each encoder block consisted of 2 convolutional layers and each convolutional layer was followed by batch norm layer and dropout layer. The number of filters for each encoder blocks were 64, 128, 256 ,512, 1024. The activation function of each convolutional layers were chosen to be leaky ReLU instead of ReLU in order to overcome the dead ReLU problem.

Modified UNet

For the modified UNet, some of the proven modifications were implemented. First of all, residual connection was used in each encoder blocks. For each encoder block, there were 3 convolutional layers. The output of the previous encoder block was added to the input of the last convolutional layer. For the last convolutional layer, it had stride of 2 which was used to replace max pooling layer. Besides, the idea of stack UNet model was borrowed for the modification. However, instead of applying two UNet, only one UNet was implemented and the output of the UNet was redirected back to its input. By using recurrent design, it can use less parameters and save GPU memory significantly. The cost function was a weighted sum of the cost from the UNet in the first round and the UNet in the second round. By involving the intermediate result in the estimation of cost, the training can be effectively speeded up.

DAN

For the DAN, the original proposed model with some minor revision was implemented for comparison. Since the original model was used to extract building in the very high resolution imagery at 5cm spatial resolution, for 0.8m resolution TripleSat imagery, the kernel size of first convolutional layer was changed from 7x7 to 3x3. Each ReLU activation function was revised to leaky ReLU.

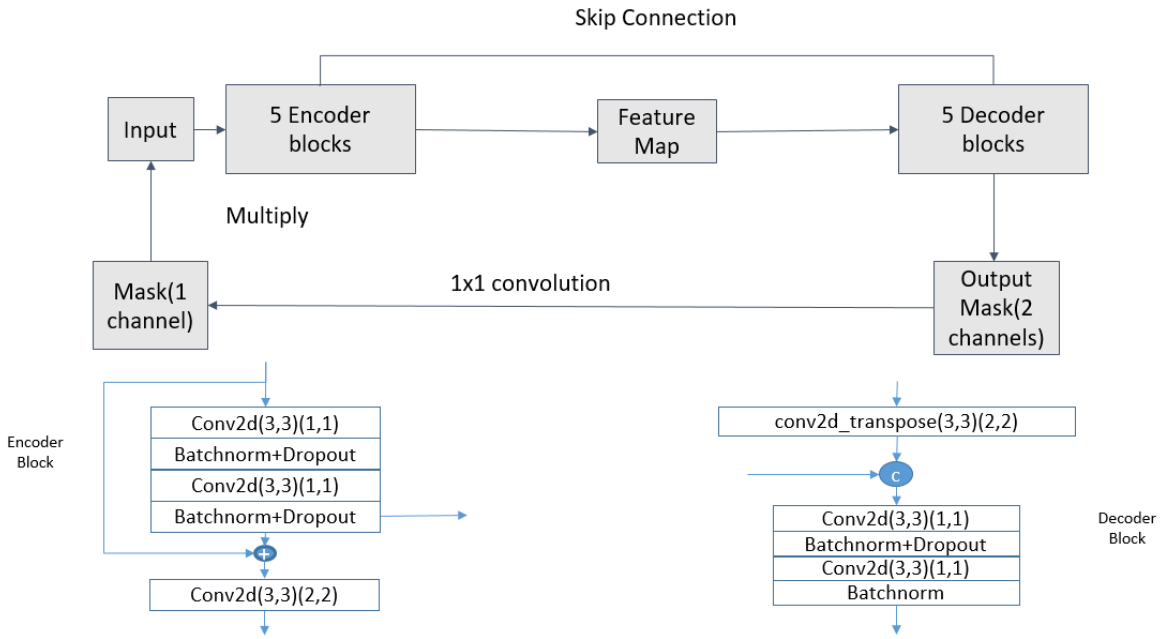


Figure 2. Architecture of the modified UNet.

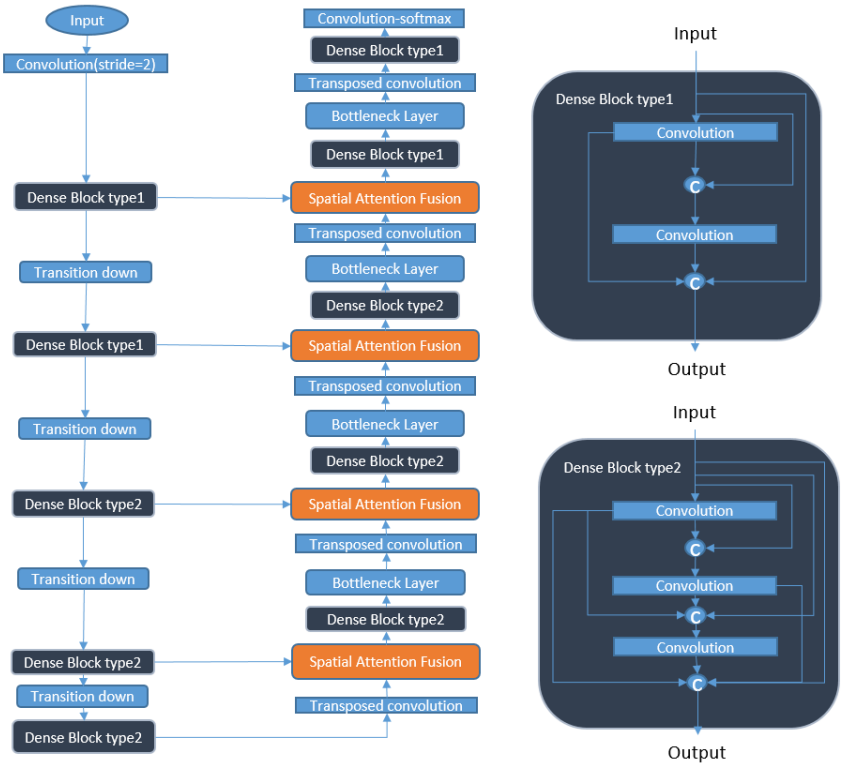


Figure 3. Proposed architecture of the DAN.

Model Comparison

In order to have a relatively fair comparison, the three models were compared without data preprocessing, data augmentation, and L2 regularization, and the model outputs were not post-processed. Same drop-out rate was applied to the three models. Each model was trained for 3000 epochs and the validation was conducted every 50 epochs. The best F1-score achieved by each model from validation dataset was used for comparison.

The accuracy was evaluated pixel-wise by using the F1-score and Intersection over union (IoU) to assess the performance of models quantitatively. The F1-score follow the below equation

$$F1 - score = \frac{2 \times precision \times recall}{precision + recall} \tag{1}$$

where precision and recall were calculated as

$$precision = \frac{true\ positive}{true\ positive + false\ positive} \tag{2}$$

$$recall = \frac{true\ positive}{true\ positive + false\ negative} \tag{3}$$

IoU was calculated as

$$IoU = \frac{true\ positive}{true\ positive + false\ positive + false\ negative} \tag{4}$$

3. RESULTS AND DISCUSSION

To compare the models, validation was conducted every 50 epochs and the average F1-score of each model along the training process is shown in Figure 4. It shows that the modified UNet has the fastest accuracy increase with the F1-score reaching around 0.8 within first 500 epochs, while the other two models took more than 2500 epochs to reach the same level.

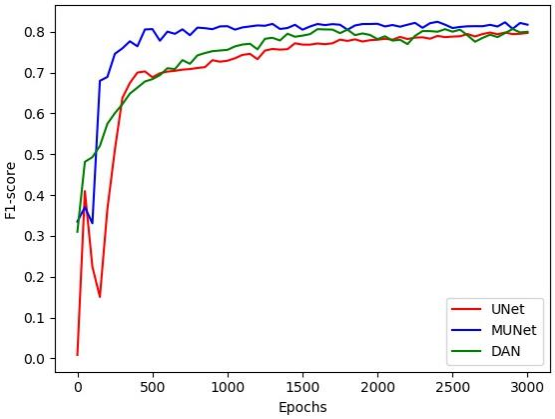
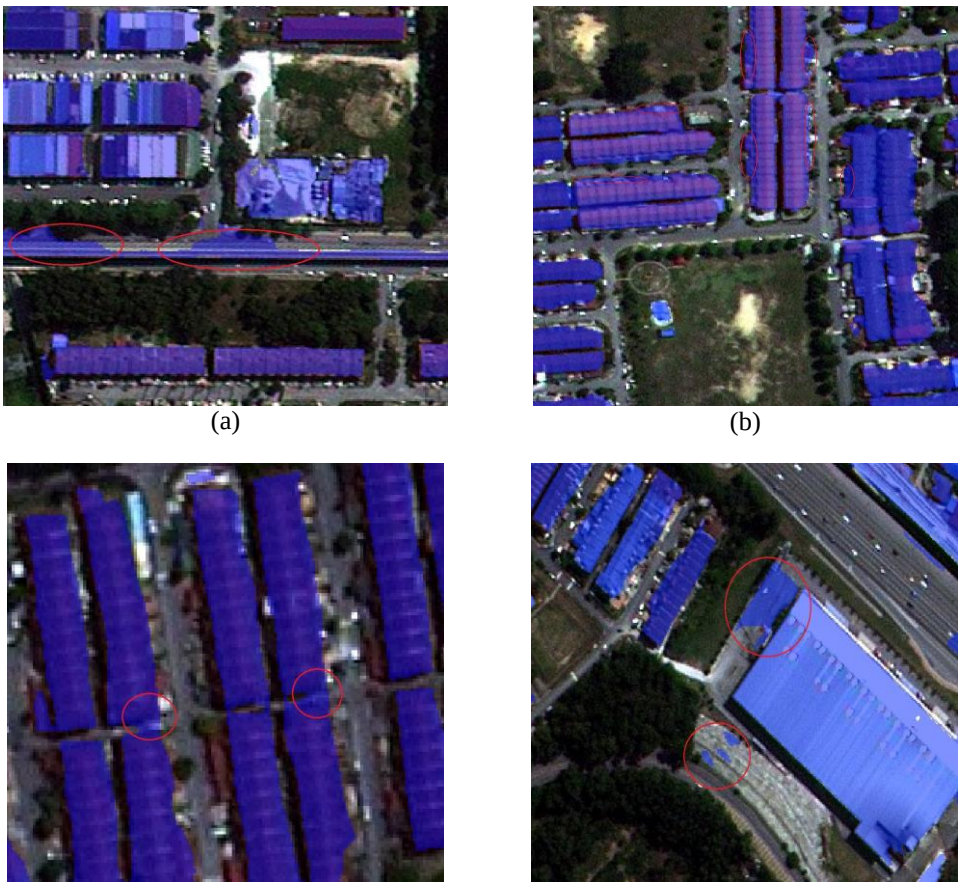


Figure 4. Evolution of F1-scores of validation dataset for each model

For all three models, recall tends to be higher than precision, which means, the models tend to predict more false positive. One of the reason is that the ground truth masks follow the real building footprint which has 90-degree angle in corners. However, limited by resolution, the buildings represented on the satellite image has blurred edges and the predicted buildings are usually larger than the ground truth. Nearby buildings are another source of false positive that when the building distance is less than 6 pixels, the deep learning models tend to predict them as one building. Besides, some roads and bare soil areas are confused as buildings.



(a) (b)

(c) (d)

Figure 5. Some examples of false positives caused by (a) misclassification of road, (b) edges of building, (c) nearby buildings and (d) misclassification of bare soil areas. The blue color represents the predicted results.

The average F1-score and IoU of the models are compared in Table 1. The results show that the accuracy of the three models are very close and the DAN has relatively higher accuracy than UNet in 0.8m resolution TripleSat image, while the modified UNet shows the highest average accuracy. The detailed comparison for each validation dataset is shown in Table 2.

Table 1. Comparison of average F1-score and IoU

	F1-SCORE(%)	IoU(%)
UNet	79.82	66.60
Modified UNet	82.45	70.35
DAN	80.68	67.94

Table 2. Accuracy comparison for each validation dataset.

Index		Precision(%)	Recall(%)	F1-Score(%)
1	UNet	64.68	89.29	75.02
	Modified UNet	76.62	78.06	**77.33**
	DAN	65.12	89.47	75.38
2	UNet	77.22	84.39	**80.64**
	Modified UNet	84.80	73.62	78.82
	DAN	72.94	87.16	79.42
3	UNet	76.78	89.77	82.77
	Modified UNet	83.15	85.55	**84.33**
	DAN	75.42	92.10	82.93
4	UNet	59.63	92.58	72.54
	Modified UNet	72.37	85.43	**78.36**
	DAN	61.49	89.17	72.79
5	UNet	78.97	84.53	81.65
	Modified UNet	86.93	84.32	85.61
	DAN	80.59	92.19	**86.00**
6	UNet	83.75	83.20	83.47
	Modified UNet	92.62	82.29	87.15
	DAN	87.68	90.34	**88.99**
7	UNet	77.95	91.51	84.19
	Modified UNet	87.50	88.55	**88.02**
	DAN	74.11	94.47	83.06
8	UNet	69.55	89.55	78.29
	Modified UNet	76.88	83.40	**80.01**
	DAN	66.57	91.03	76.90

From the detail comparison in Table 2, it can be found that the DAN has very high accuracy for patch 5 and 6 where there are mainly big factories. The average F1-score of the DAN for these two validation patches is 87.50%, while for the UNet and the modified UNet, it is 82.56% and 86.38% respectively. The modified UNet is close to the DAN but still cannot match the DAN. However, for validation patches where there are mainly small and medium sized buildings, DAN has relatively lower accuracy. For patch 1, 2, 3, 4, 7 and 8, the average F1-score of the DAN is 78.41% which is a little bit lower than the UNet with average F1-score of 78.91%. In the original paper of DAN, it was reported that the DAN might miss small buildings. It seems the DAN is good at predicting large buildings on 0.8m resolution satellite image. The modified UNet achieved the highest average F1-score of 81.15% for patches containing mainly small and medium buildings.

Visual comparison of the predicted results is shown in Figure 6. For large buildings, the UNet has cavity issues, while the predicted result of DAN in this case is more intact. However, the DAN has more false positives for small and medium buildings. The modified UNet has the best performance for edges of small and medium buildings.

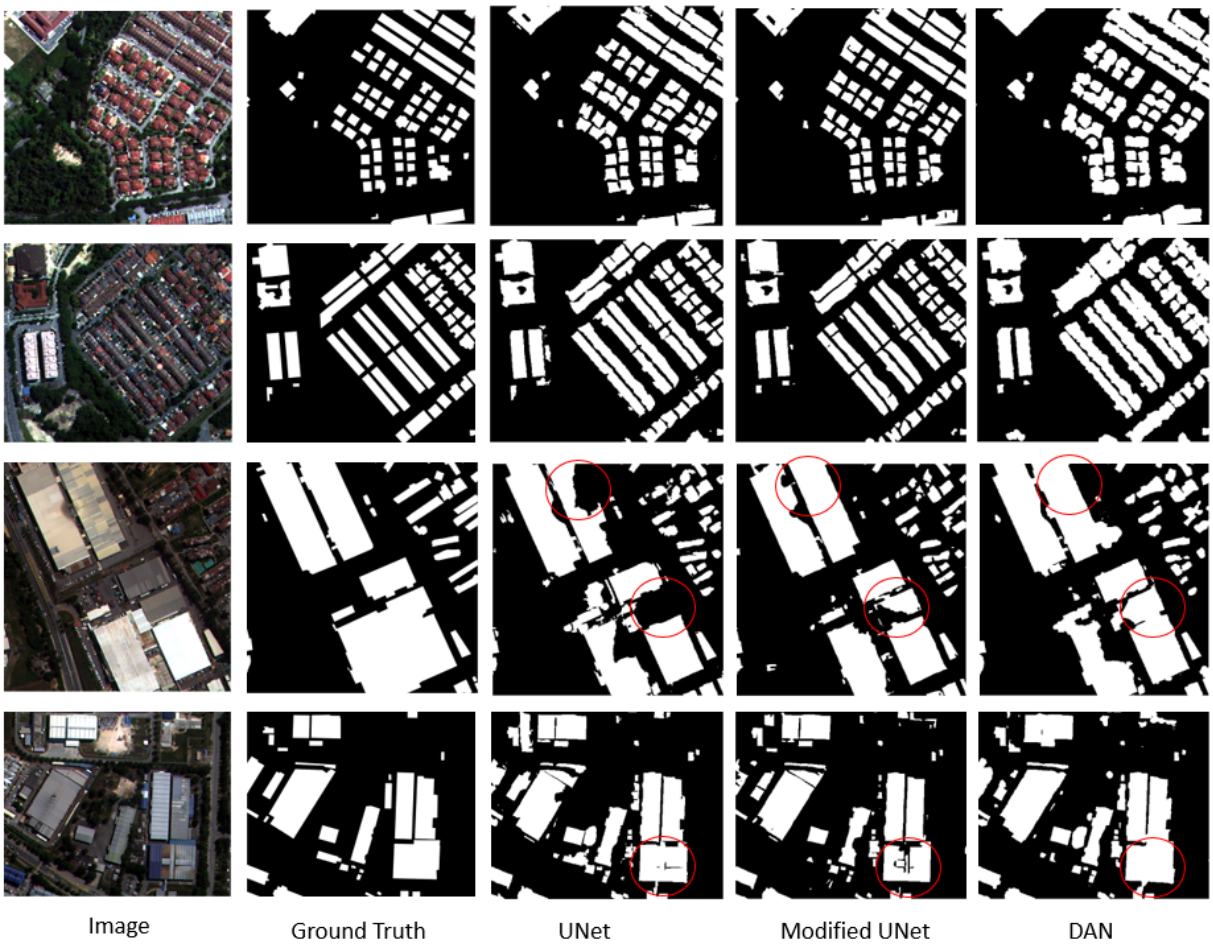


Figure 6. Visual comparison of segmentation results from the three models. The modified UNet is the most precise for small and medium buildings. But for large buildings, the UNet and the modified UNet have cavity issues and the DAN works relatively better.

The results suggest that different models may tend to be good at prediction of buildings of specific size. Thus combination of models for buildings of different sizes can help to further boost building extraction accuracy.

4. CONCLUSION

In this paper, we compared the performance of the UNet, the modified UNet and the DAN for building extraction with 0.8m TripleSat imagery. The modified UNet achieved the highest accuracy with average F1-score of around 82.45%, while the DAN achieved F1-score of 80.68% which surpassed the UNet with F1-score of 79.82%. However, the DAN has the highest accuracy for larger buildings. It suggests that DAN is very good at predicting large buildings for 0.8m resolution TripleSat imagery, while the modified UNet is the most precise for small and medium buildings

and a combination of models may improve the accuracy of segmentation results significantly.

5. REFERENCES

Chhor, G., Aramburu, C. B., Lambert, I. B., Satellite Image Segmentation for Building Detection using U-net, from http://cs229.stanford.edu/proj2017/final-reports/5243715.pdf.

Hamaguchi, R., Hikosaka, S., 2018. Building detection from satellite imagery using ensemble of size-specific detectors, Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW), pp. 223-227.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Identity mappings in deep residual networks, from https://arxiv.org/pdf/1603.05027.pdf.

Huang, G., Liu, Z., Weinberger, K. Q., 2016. Densely connected convolutional networks, from https://arxiv.org/pdf/1608.06993.pdf.

Huang, X., Zhang, L., 2011. A Multidirectional and Multiscale Morphological Index for Automatic Building Extraction from Multispectral GeoEye-1 Imagery[J]. Photogrammetric Engineering & Remote Sensing, 77(7), pp. 721-732.

Iglovikov, V., Mushinskiy, S., Osin, V., 2017. Satellite imagery feature detection using deep convolutional neural network: A kaggle competition, from https://arxiv.org/pdf/1706.06169.pdf.

Jégou, S., Drozdzal, M., Vazquez, D., Romero, A., Bengio, Y., 2016. The one hundred layers tiramisu: Fully convolutional DenseNets for semantic segmentation, from https://arxiv.org/abs/1611.09326.

Jiang, N., Zhang, J. X., Li, H. T., Lin, X. G., 2008. Semi-automatic building extraction from high resolution imagery based on segmentation, Proc. Int. Workshop EORSA, pp. 1-5.

Li, L., Liang, J., Weng, M., Zhu, H., 2018 A Multiple-Feature Reuse Network to Extract Buildings from Remote Sensing Imagery. Remote Sens., 10, 1350.

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 Obs. Remote Sens. 2008, 1, pp. 180–192.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation, Proc. Int. Conf. Medical Image Comput. Comput.-Assisted Intervention, pp. 234-241.

Sevastopolsky, A., Stepan, D., Konstantin, K., Snyder, B. M., Anastasia, G., 2018. Stack-u-net: Refinement network for image segmentation on the example of optic disc and cup, from https://arxiv.org/pdf/1804.11294.pdf.

Song, Y., Shan, J., 2008. Building extraction from high resolution color imagery based on edge flow driven active contour and JSEG. Proceedings of the XXI International Congress of ISPRS. Beijing, China. pp.185–190.

Yang, H., Wu, P., Yao, X., Wu, Y., Wang, B., Xu, Y., 2018. Building extraction in very high resolution imagery by dense-attention networks. Remote Sens. 10, 1768.

Yang, J., Wang, Y. H., 2012. Towards automatic building extraction: Variational level set model using prior shape knowledge, Proc. Int. Conf. Image Anal. Signal Process., pp. 1-6.