Automated Building Extraction from Aerial Images with An Improved End-To-End Deep-Learning-Based Approach

Hailun YAN and Ruisheng WANG, Canada

Key words: Building segmentation; Boundary regularization; Deep learning; End-to-end

SUMMARY

Automatically extracting high-quality building polygons from Aerial images is crucial for supporting various land use and land cover mapping applications as the conventional object extraction process requires hand-crafted feature and human interventions which often has limited generalization capability and is time consuming. In this paper, we introduce an improved end-to-end deep-learning-based building extraction method based on PolygonCNN. First, after comparing the building segmentation performances of several famous segmentation networks, we replace the PSPNet of the original PolygonCNN with the Swin Transformer-based Mask R-CNN which has shown a significantly improved building segmentation capability. Next, we integrate the Mask-R-CNN-based PolygonCNN with the Feature Pyramid Network (FPN) which exploits the multi-scale, pyramidal hierarchy feature maps of Swin Transformer. The Integration of FPN has shown to significantly improve both the segmentation performance of Mask-RCNN and the ragularization ability of the modified PointNet. Lastly, we further modify the original modified PointNet (BregNet) into a wider and deeper version to utilize the multi-scale feature maps of the FPN, thus to achieve better regularization effects. Our modified PolygonCNN has achieved state-of-the-art results when comapred with other end-to-end deep-learning-based building extraction methods.

Automated Building Extraction from Aerial Images with An Improved End-To-End Deep-Learning-Based Approach

Hailun YAN and Ruisheng WANG, Canada

1. INTRODUCTION

The aerial photography technologies in the recent decade have enabled massive amount of data to be acquired every day and opened possibilities of many new land use and land cover (LULC) mapping applications. These applications often require digitization and interpretation of objects within the raw image data so that analysis can be further done with geographic information systems.
Building, as one of the most frequently appearing object in urban scenes, its extraction from aerial images has played an important role in supporting LULC applications such as urban planning, change detection, and disaster management. Traditionally, this building extraction process requires human image interpreters which is extremely labour intensive and time consuming. Thus, automated extraction methods have been developed. In the early years, these methods often consist of manually extracting features (e.g., spectral, spatial, textural) followed by traditional machine learning classification methods. However, manual feature extraction usually requires experienced experts which may not always be feasible; moreover, since the extracted features are often designed to model spefic building types, the generalization capabilities of these methods are highly limited. As a result, many valuable information contained in this enormous amount of new data are not available.

In the recent years, due to the fast-paced development of computation capability and the availability of vast training data, deep learning techniques have shed light on this problem. Techniques such as convolutional neural networks (CNN) and fully convolutional networks (FCN) have shown dominancy over conventional methods in terms of the level of automation, the building segmentation accuracy, and the generalization capabilities. However, the process of converting the predicted building segmentations, which often have irregular shapes that differ significantly from real-world building footprints, into regularized (i.e. straight edges and right-angled corners) building polygons continued to rely on handcrafted features and high human interventions. Despite a few works have been done on developing end-to-end deep-learning-based methods to extract regularized building boundaries, these methods often have significantly lower accuracy compared to the semi-automated methods.

Thus, to facilitate the development of the fully automated methods, this paper provides an improved end-to-end deep-learning-based method to extract building footprint polygons from aerial images. Based on the work of PolygonCNN (Chen et al., 2020), we introduce several improvements. First, we replace the PSPNet in the original PolygonCNN with the Swin Transformer-based Mask R-CNN which has shown to have significantly improved building segmentation capability. Next, we integrate PolygonCNN with the Feature Pyramid Network (FPN) which exploits the multi-scale, pyramidal hierarchy feature maps of Swin Transformer,

and significantly improves both the segmentation performance of Mask-RCNN and the ragularization ability of the modified PointNet. Lastly, we further modify the original modified PointNet into a wider and deeper version to utilize the multi-scale feature maps of the FPN, thus to achieve better regularization effects.

2. RELATED WORKS

In general, the existing methods of building extraction from aerial images can be categorized into two-stage methods and end-to-end methods. The two-stage methods often consist of a building segmentation step and a boundary regularization step. The end-to-end methods aim to take aerial images as input and directly output the building vectors. Therefore, this section will review the related works in the following three sub-sections: building segmentation, building boundary regularization, and end-to-end building polygon prediction.

2.1 Building segmentation

The studies on building segmentation have transitioned significantly from the early days' conventional methods into the nowadays' popular deep learning-based methods. The conventional methods usually involve manually extracting features based on the spatial (i.e., key points, corner points, edges), textual, and/or spectral characteristics of the image, then segmenting the building footprints by applying methods such as template matching (Sirmacek and Unsalan, 2009), graph cut (Manno-Kovacs and Ok, 2015), random forest classifier (Pelizari et al., 2018) or support vector machine classifier (Turker and Koc-San, 2015) to the extracted features. Despite that many significant achievements have been made, the feature extraction step of these methods is often designed for specific building types which has highly limited generalization capability. In addition, the hand-crafted features rely heavily on human intervention which is time consuming and may not always be practical.

In the recent years, deep-learning-based building segmentation methods have become widely popular mainly due to the breakthrough of convolutional neural networks (CNN) on the ImageNet classification contest in 2012 (Krizhevsky et al., 2012). Deep-learning-based methods overcome the explicit feature design problem of the conventional methods by allowing adaptive feature learning from labeled training data. The early building segmentation CNNs achieve pixel-level segmentations by splitting a high resolution image into small patches (Alshehhi et al., 2017; Guo et al., 2017). Despite of the promising results, these methods are limited to overly redundant computations due to the heavy overlaps between patches. Thus, to over come the problem, fully convolutional networks (FCN) have been proposed to achieve pixels-to-pixels predictions (Long et al., 2015; Maggiori et al., 2017). In addition, He et al. (2017) have proposed Mask R-CNN for instance segmentation which combines semantic segmentation with object detection to predict the pixel masks for each individual building instance.

2.2 Building boundary regularization

Despite of the great performances of FCN networks, there often tends to exist slightly mis-predicted pixels along the predicted building boundaries, resulting in irregular shapes of the directly traced building polygons. Therefore, many studies have focused on regularizing the building segmentations into accurate, simple and regular polygons. Typically, building regularization require additional data sources such as airborne lidar scanning or public GIS data as assistances for precised regularizations (Boehm, 2019; Li et al., 2019). When only image data is available, usually hand-crafted features or pre-specified constraints such as 90-degree corners and principle orientations are applied to regularize and simplify the building boundaries (Ling et al., 2012; Zhang et al., 2018). However, these low-level features has highly limited generalization capability on diversified building shapes. In addition, although they produce well-regularized building polygons, the regularized polygons often have significantly decreased accuracies.

Due to the significant drawbacks of the conventional methods, seeking deep-learning-based building regularization methods has become a new trend. Girard and Tarabalka (2018) developed CNN-based methods to produce vectorial boundary labels of an image directly; Marcos et al. (2018) developed a CNN to predict building polygons close to the ground truths by learning parameters of an active countour model. The work was further improved by deep active ray network (DARNet) (Cheng et al., 2019). Although these methods have susscessfully improved the generalization capability and level of automation compared to the conventional building regularization methods, the regularized building vectors often have lacked simplicity and regularity.

2.3 End-to-end building polygon prediction

The end-to-end building extraction methods aim to take the aerial images as input, and directly output the regularized building polygons. Cheng et al. (2019) proposed an end-to-end deep neural network named deep active ray network (DARNet) for building polygon extraction. They use a backbone CNN to predict energy maps which are utilized to construct an energy function, and the building polygons are derived by minimizing the energy function. Despite the end-to-end structure, the network fails to take into consideration the simplicity and regularity of building polygons. Castrejon et al. (2017) developed an deep neural network named PolygonRNN to sequentially predict the object contour points. The network is consisted of a CNN feature extractor, and a recurrent neural network (RNN) which decodes one polygon vertex at a time. Despite the end-to-end structure, the network is limited by its high memory requirement. Motivated by PolygonRNN, Li et al. (2019) developed PolyMapper which focuses on delineating the road and building vector boundaries from a given image. Manual bounding box labels are no longer required due to the integration of the Feature Pyramid Network (FPN) detection module on top of PolygonRNN. However, PolyMapper lacks ability in predicting objects with complexed shapes; moreover, its convolutional Long Short-Term Memory module is computationally expensive. To overcome the expensive memory issue of these RNN-based networks, Chen et al. (2020) proposed an end-to-end network based on a

segmentation CNN and a boundary regularization CNN. The sementic segmentation network PSPNet (Zhao et al., 2017) is used to generate the initial building contour, while a modified PointNet predicts the coordinate offsets of the polygon vertices to generate the regularized buildings.

3. METHODS

Following the work of PolygonCNN (Chen et al., 2020), we introduced several enhancements. First, after evaluating the performances of several famous segmentation networks on the SpaceNet2 Building Detection Dataset, we replaced the PSPNet (Zhao et al., 2017) of PolygonCNN with the significantly improved Mask R-CNN (He et al., 2017) which was mounted with the state-of-the-art Swin Transformer (Liu et al., 2021) backbone. Next, we integrated PolygonCNN with the Feature Pyramid Network (FPN) which exploits the multi-scale, pyramidal hierarchy feature maps of Swin Transformer, and significantly improves the segmentation performance of Mask-RCNN and the regularization ability of the modified PointNet. Lastly, we further modified the original modified PointNet into a wider and deeper version to utilize the multi-scale feature maps of the FPN to achieve better regularization effects.

3.1 Building segmentation network

To determine the segmentation network that can most precisely segment building footprints from aerial images, we have selected several famous segmentation networks from different application domains, and trained and compared their performances on the SpaceNet2 Building Detection Dataset. These selected networks are: U-Net (Ronneberger et al., 2015), PSPNet (Zhao et al., 2017), DeepLabV3+ (Chen et al., 2018), UPerNet (Xiao et al., 2018), HRNet (Wang et al., 2020), and Mask R-CNN (He et al., 2017). Since PSPNet, DeepLabV3+, UperNet, and Mask R-CNN take custom backbones, ResNet50 (He et al., 2016) is used for fare comparison. In addition, due to the impressive performance of Mask R-CNN, the Mask R-CNN models with different backbone networks were also compared. The selected backbone networks were: ResNet50 and ResNet101 (He et al., 2016) , DenseNet121 and DesNet161 (Huang et al., 2017), HRNet (Wang et al., 2020), and Swin Transformer (Liu et al., 2021). The best performing model was used to replace the PSPNet in the original PolygonCNN.

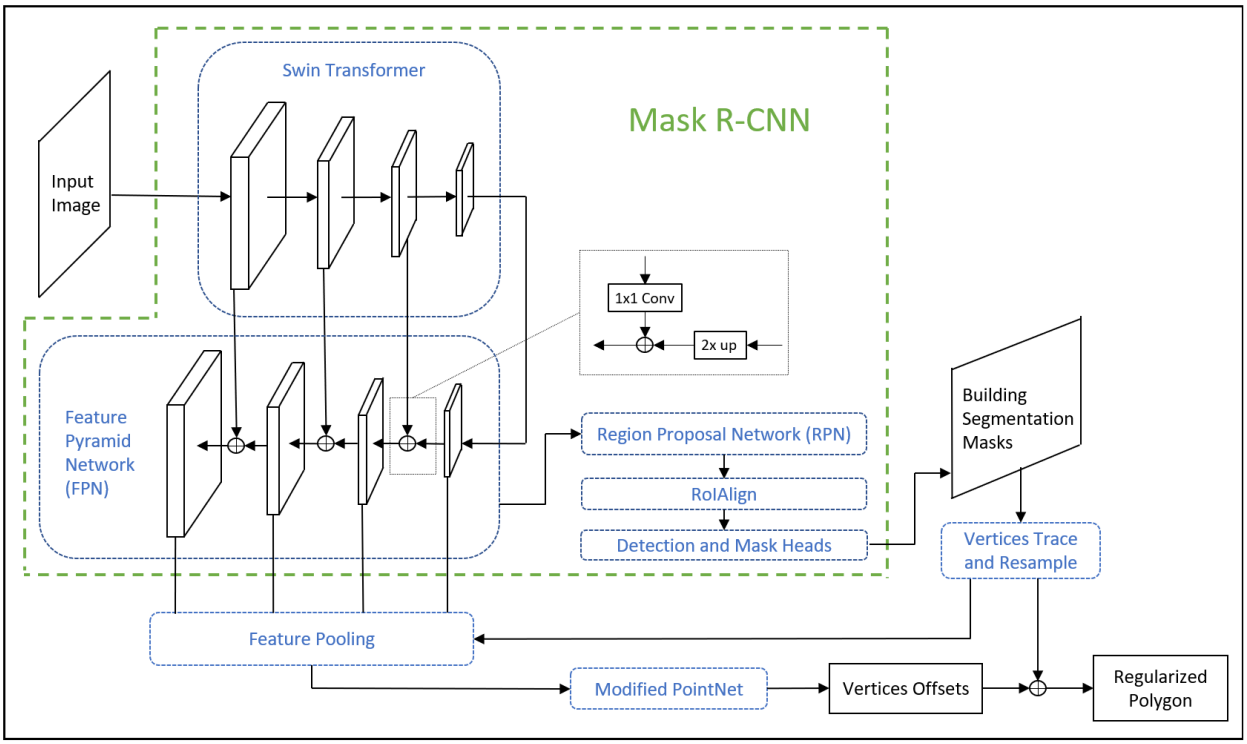

Figure 1. Integration of the Feature Pyramid Network with PolygonCNN.

3.2 Integrating FPN into PolygonCNN

As shown in Figure 1, based on the results from 3.1, we replaced the PSPNet in the original PolygonCNN with the Swin Transformer-based Mask R-CNN, and integrated the new PolygonCNN with the Feature Pyramid Network (FPN) (Lin et al., 2017). The FPN enhances the performance of PolygonCNN in the following two ways.

First, the FPN greatly enhances the region proposal network (RPN) used by Mask R-CNN by exploiting the multi-scale, pyramidal hierarchy feature maps of Swin Transformer. The RPN in the original Mask R-CNN uses a single-scale feature map to create Region of Interests (RoIs); with the integration of FPN, the RPN is applied to feature maps at mutiple levels to generate multi-scale RoIs. Based on the size of the RoI, the feature map in the most proper scale is used to extract the feature patches. This procedure greatly enhances the segmentation performance of Mask R-CNN.

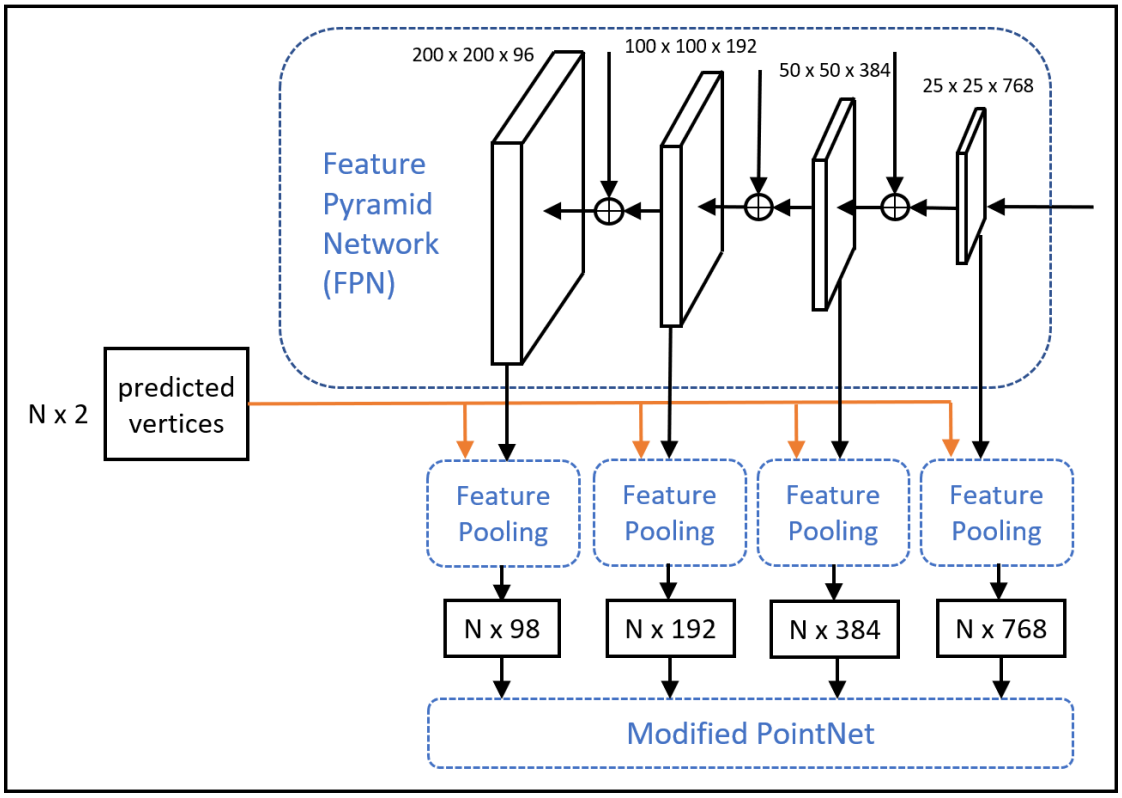

Figure 2. Detailed structure of the improved feature pooling process.

Furthermore, the original PolygonCNN uses a single-scale feature map for the feature pooling process, while with the integration of FPN, we utilize all four levels of feature maps in the FPN for the feature pooling. For example, a detailed feature pooling process is illustrated in Figure 2. Assuming there are N predicted vertices (a N x 2 tensor), and four levels of feature maps in the FPN (with dimensions of 768, 384, 192, and 96). The new feature pooling process extracts every predicted polygon vertex's corresponding feature vector in each level of the feature maps. Then, the predicted vertex vectors (N x 2) are concatenated with their corresponding feature vectors at each level (with sizes of N x 768, N x 384, N x 192, and N x 96), resulting in four tensors of sizes N x 770, N x 386, N x 194, and N x 98. The four tensors are fed concurrently into our improved PointNet (see section 3.3) for the improved buiding polygon optimization.

3.3 Modified PointNet

We improved the original modified PointNet of PolygonCNN to enable the utilization of multi-scale feature maps of FPN. For naming convenience, we named the original modified PointNet of PolygonCNN as BRegNet which stands for Building Regularization Network. The BregNet, as described in Chen et al., 2020) takes a single input which is a concatenation of the predicted polygon vertices with their corresponding feature vectors extracted from a single-scaled feature

map. However, our improved BRegNet takes four inputs consisting of concatenations of predicted polygon vertices with their corresponding feature vectors extracted from the four levels of the FPN feature maps.

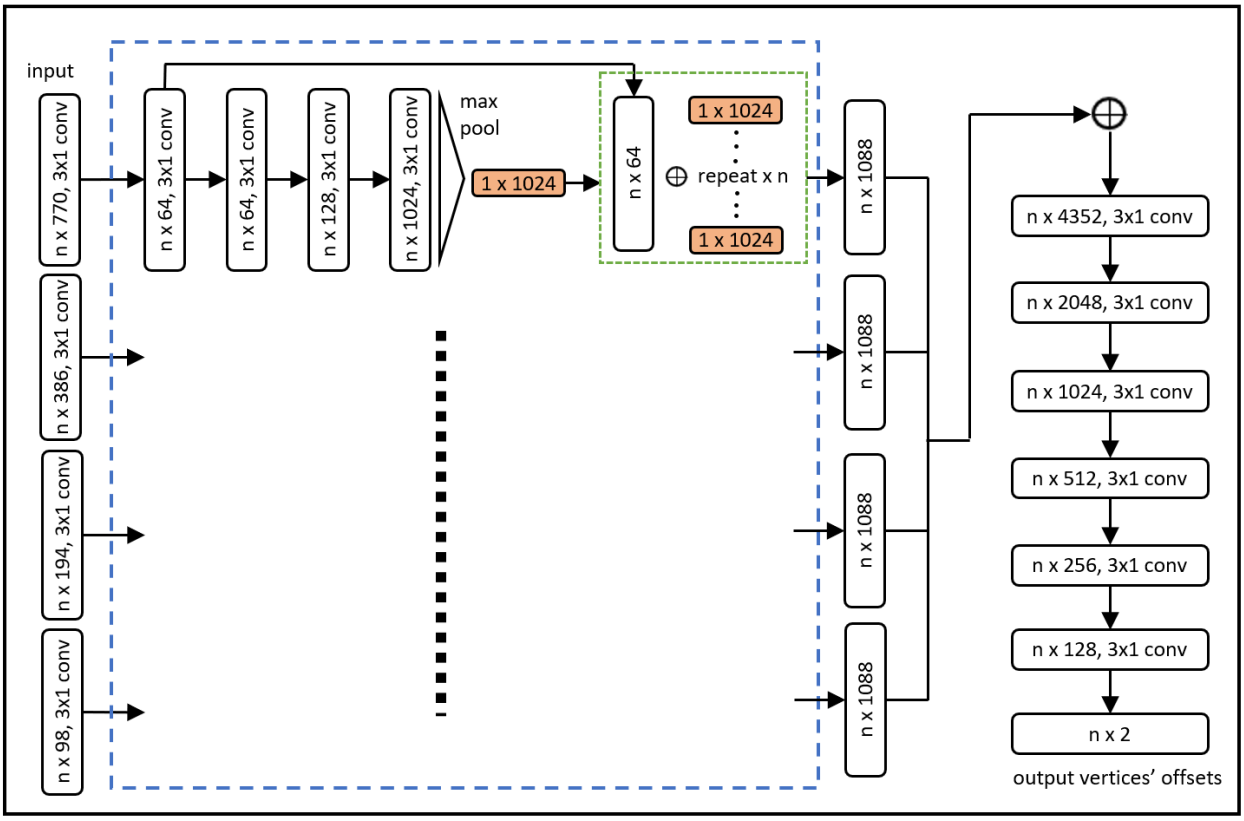

Figure 3. The improved BRegNet which enables multi-scale feature vector inputs

As shown in Figure 3, to enable multi-scale inputs, we first widened the original network by adding three more duplicated paths of feature transform and global pooling layers (as shown in the blue-dotted box). The outputs (n x 1088) of the four duplicated paths consist of local and global information of the multi-scale inputs. These ouputs are concatenated into one tensor of size n x 4352. Next, to reduce the large dimension size of the feature aggregation, we deepened the network by adding two more feature transform layers with output sizes of n x 2048 and n x 1024. This way, the dimension of the feature aggregation is reduced gradually with less information loss.

4. Experimental settings

4.1 Dataset

We evaluate both the segmentation networks and the building extraction networks on the Round 2 dataset of the SpaceNet Building Detection Challenge. The dataset provides satellite images of four urban cities including Las Vegas, Paris, Shanghai, and Khartoum, contains over

302,701 building footprints. All images were pan-sharpened and have uniformed properties: the spatial resolution is 0.3m by 0.3m per pixel, channel is RGB, and size is 650×650. To train the networks with a reasonable sized dataset, we randomly sampled 4000 images among the whole dataset. The sampled images was split into 3500 training images and 500 validation images.

4.2 Network settings

The eleven segmentation networks (i.e. U-Net, PSPNet, DeepLabV3+, UperNet, HRNet, and six Mask R-CNN models with different backbone networks) as well as the five building extraction networks (i.e. Polygon-RNN, DARNet, PolyMapper, PolygonCNN, proposed) share the following settings

- The learning rate scheduler (ReduceLROnplateau) was set with a starting learning rate of 0.1, reduce factor of 0.2, and patience of 10.
- the SGD optimization strategy was used with a weight decay of 0.05 and momentum of 0.9.
- The batch size for the segmentation networks were set to 4, and that for the extraction networks were set to 1.
- All the generated building polygons are processed by the Doglus-Peuker (DP) algorithm with ε of 1 pixel for fair comparisons of the building extraction networks.

In addition, the anchor box size of Mask R-CNN was reduced to{8, 16, 32, 64, 128} instead of the default values considering the general sizes of the building footprints. Other non-specified parameters of all models were left with the default values as that were discribed in the papers. Furthermore, all models were implemented and tested in Pytorch (Steiner et al., 2019) on a 64-bit Ubuntu system equipped with an NVIDIA TITAN X GPU.

4.3 Evaluation metrics

To evaluate the performances of the building segmentation networks, the mean Average Precision (AP) and mean Average Recall (AR) were calculated based on the intersection over union (IoU) (Jaccard, 1912) metric. Specifically, AP and AR were averaged over ten IoU values with thresholds from .50 to 0.95 with steps of 0.05. In addtition, the performances of the building extraction networks were evaluated using the metrics of F1-score which is a harmonic average of the polygon-based precision and recall, and was provided by the SpaceNet Building Detection Challenge.

5. Results and Discussion

5.1 Building segmentation results and comparison

Model	Backbone	AP	AR
U-Net	-	0.391	0.404
PSPNet	ResNet50	0.422	0.439
DeepLabV3+	ResNet50	0.418	0.451
UPerNet	ResNet50	0.409	0.431
HRNet	-	0.427	0.455
Mask R-CNN	ResNet50	**0.473**	**0.502**

Table 1. Comparison of the evaluation results of the selected segmentation networks.

To determine the segmentation network that can most precisely segment building footprints from aerial images, we first trained and compared the six selected segmentation networks on the SpaceNet2 Building Detection Dataset.As shown in Table 1, Mask R-CNN significantly outperformed all other segmentation networks by a large margin with an AP of 0.473 and AR of 0.502; HRNet, PSPNet and DeepLabV3 had shown similar performances with APs ranged between 0.418 to 0.427 and ARs ranged between 0.439 to 0.455; UPerNet had a slightly worse performance with an AP of 0.409 and AR of 0.431; U-Net had the lowest 0.391 AP and 0.404 AR.

Due to the outstanding performance of Mask R-CNN, we futher evaluated several popular backbone networks on Mask R-CNN. Their evaluation results are shown in Table 2. Compared to ResNet50, the ResNet101 model had shown significantly improved AP and AR of 0.506 and

Model	Backbone	AP	AR
Mask R-CNN	ResNet50	0.473	0.502
Mask R-CNN	ResNet101	0.506	0.555
Mask R-CNN	DenseNet121	0.456	0.485
Mask R-CNN	DenseNet161	0.512	0.562
Mask R-CNN	HRNet	0.556	0.601
Mask R-CNN	Swin Transformer	**0.564**	**0.614**

Table 2. Evaluation results of Mask R-CNN with different backbones.

0.555, respectively, due to its larger depth and a sacrifice to the more expensive computations. In addition, the DenseNet161 had similar results compared to ResNet101. Surprisingly, the HRNet model had outperformed the ResNet and DenseNet models by a large margin, indicating that the rich high resolution features and multi-scale fusion of HRNet is effective in extracting features of building objects from high resolution aerial images. Furthermore, the Swin Transformer model reached the highest AP and AR of 0.564 and 0.614 with its novel non-overlapped window-based self-attention and shifted-window operation. Based on the comparison results, the Swin Transformer-based Mask R-CNN was used as the segmentation network in the PolygonCNN model.

5.2 Building extraction results and comparison

Method	Segmentation Network		FPN	Enhanced feature pooling	Improved BRegNet	F1	AP	AR
	PSPNet	Mask R-CNN						
PolygonCNN (Chen et al., 2020)	×					0.421	0.415	0.428
		×				0.474	0.463	0.486
		×	×			0.499	0.488	0.510
		×	×	×		0.514	0.501	0.527
PolygonCNN (ours)		×	×	×	×	**0.530**	**0.519**	**0.541**

Table 3. Comparison of the polygon-based F1-score, AP, and AR between the original PolygonCNN, the PolygonCNNs with gradually added modules, and the final upgraded PolygonCNN.

In Table 3, we compare the polygon-based F1-score (calculated using on the polygon-based AP and AR as shown in Table 3) between the original PolygonCNN, the PolygonCNNs with gradually added modules, and our final upgraded PolygonCNN. With the proposed modifications made on PolygonCNN, we achieved absolute improvements of 0.109 F1-score compared to the original PolygonCNN. By gradually adding on modules to the original PolygonCNN, the changes of the F1-score are as the following.

- Replacing the PSPNet in the original PolygonCNN with the Swin Transformer-based Mask R-CNN, the F1-score was improved significantly from 0.421 to 0.474.
- Incorporating the FPN module into the Mask R-CNN network further boosts the F1-score to 0.499. Note that the original feature pooling process was applied to the last layer of feature map in the top-down pathway of the FPN.
- Applying the improved feature pooling process on all four levels of feature maps in the top-down pathway of the FPN (as shown in Figure 1) further increases the F1-score to 0.514. Note that the pooled feature representations of the four scales were concatenated along with the predicted polygon vertices (e.g., in Figure 2, n feature vectors are extracted from the four feature maps resulting in four feature representations of sizes n x 768, n x 384, n x 192, n x 96. They are concatenated along with the predicted n x 2 polygon vertices, resulting in a n x 1442 tensor). The resulted tensor was sent into the original BRegNet which takes a single input.
- With the modifications made to the BRegNet as described in section 3.3, the F1-score was increased to 0.530.

In addition, we evaluated other state-of-the-art end-to-end deep-learning-based methods such as the Polygon-RNN, DARNet, and PolyMapper on the SpaceNet 2 Building Detection

Dataset, and compared their results with our improved PolygonCNN as shown in Table 4. The results show that our improved PolygonCNN significantly outperform other state-of-the-art methods with a F1-score of 0.530. Polygon-RNN has the worst performance in terms of building extraction with a F1-score of 0.375; PolyMapper has the second lowest F1-score of 0.402; PolygonCNN is placed at third with a F1-score of 0.421; DARNet has the second highest F1-score of 0.432.

Method	F1	AP	AR
Polygon-RNN	0.375	0.473	0.310
DARNet	0.432	0.422	0.443
PolyMapper	0.402	0.408	0.396
PolygonCNN	0.421	0.415	0.428
Improved PolygonCNN (Ours)	**0.530**	**0.519**	**0.541**

Table 4. Comparison of the evaluation results between Polygon-RNN, DARNet, PolyMapper, PolygonCNN, and our improved PolygonCNN on the SpaceNet2 Building Detection Dataset.

To better understand the causes of this ranking, a comparison of the building polygons extracted by these networks is shown in Figure 4.

- Both Polygon-RNN and PolyMapper tend to make accurate prediction on the building parts with simple structures. However, they fail to predict buildings with more complexed shapes. This may be caused by the insufficient capability of their segmentation modules.
- DARNet is capable of capturing the main structure of various types of buildings; however, the predicted polygons show lacks of regularization and simplification (i.e. fails to capture sharp corners and always contain redundant vertices).
- The extracted building polygons of the original PolygonCNN are generally acceptable despite that some complexed building parts are mis-predicted. This is mainly caused by the lack of segmentation capability of the PSPNet considering the large performance difference between PSPNet and Mask R-CNN (as shown in Table 1) in segmenting building footprints.
- Our improved PolygonCNN, benefited from its Swin Transformer-based Mask R-CNN, the multiscale feature maps of the FPN, and the enhanced BRegNet, is capable of generating well regularized and simplified polygons for various types of buildings.

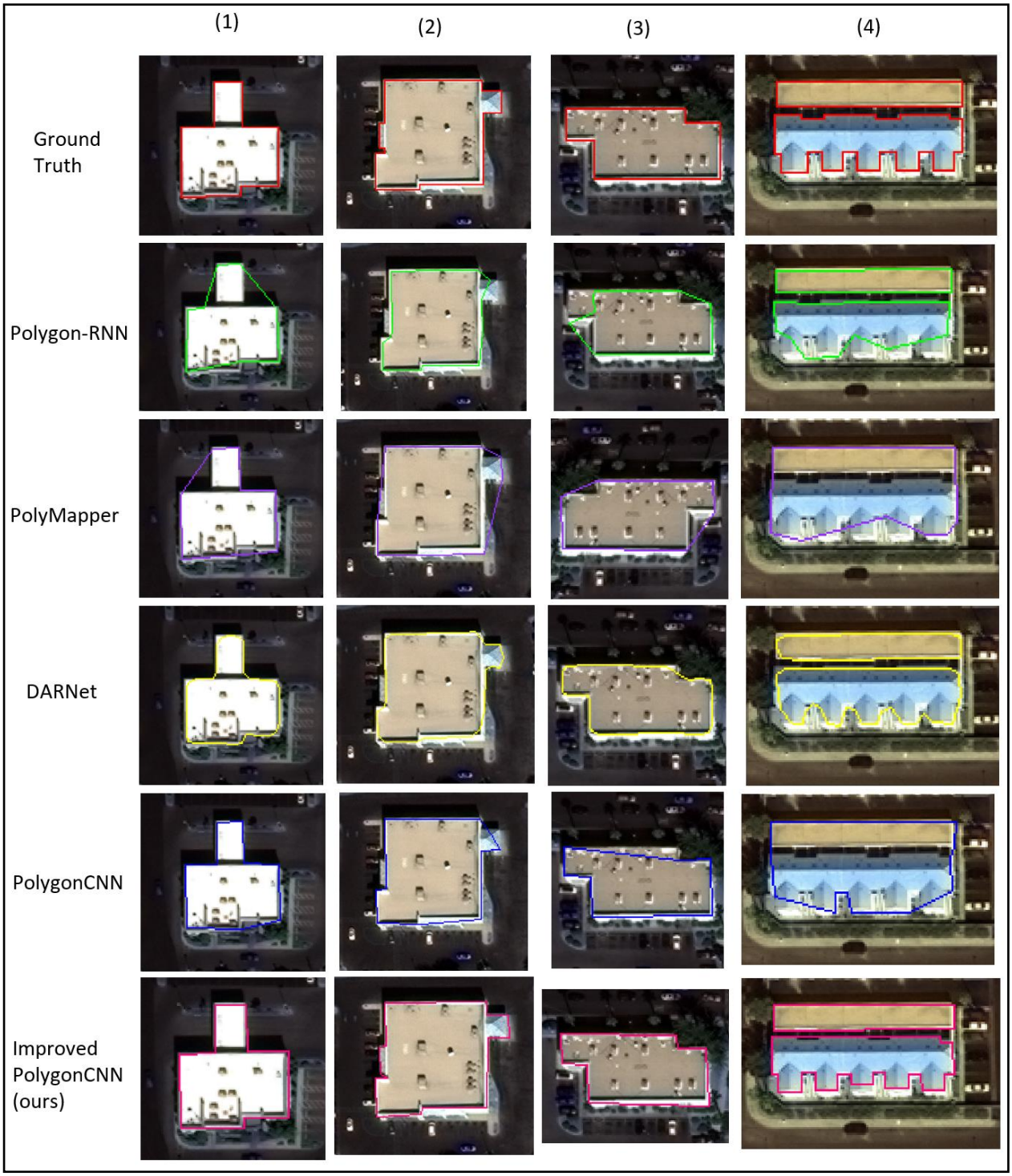

Figure 4. Examples of building vectors generated by Polygon-RNN, PolyMapper, DARNet, PolygonCNN, and our improved PolygonCNN.

6. Conclusion

In this work, we have provided an improved end-to-end deep-learning-based method to extract building footprint polygons from aerial images. Following the work of PolygonCNN (Chen et al., 2020), we introduced several improvements. First, by evaluating the performances of several famous segmentation networks on the SpaceNet2 Building Detection Dataset, we replaced the PSPNet in the original PolygonCNN with the Swin Transformer-based Mask R-CNN which had shown to have significantly improved building segmentation capability. Next, we integrated PolygonCNN with the Feature Pyramid Network (FPN) which exploits the multi-scale, pyramidal hierarchy feature maps of Swin Transformer, and significantly improves both the segmentation performance of Mask-RCNN and the ragularization ability of the modified PointNet. Lastly, we further modified the original modified PointNet (BregNet) into a wider and deeper version to utilize the multi-scale feature maps of the FPN, thus to achieve better regularization effects. Our modified PolygonCNN had achieved state-of-the-art results when comapred with other end-to-end deep-learning-based building extraction methods.

Since our modifications to PolyCNN mainly focus on improving the extraction accuray instead of the extraction/training speed, the improved BRegNet had shown significantly higher requirements in the computation power. Moreover, despite the great extraction capability of the current model on most commonly-seen building types, significant mis-predictions can still be found on certain biuldings with curved edges and very complicated shapes. Thus, our future work will focus on: (1) developing lighter weighted regularization networks; (2) developing methods that can precisely extract building polygons with curved edges and very complexed shapes.

REFERENCES

Alshehhi, R., Marpu, P.R., Woon, W.L. and Dalla Mura, M., 2017. Simultaneous extraction of roads and buildings in remote sensing imagery with convolutional neural networks. ISPRS Journal of Photogrammetry and Remote Sensing, 130, pp.139-149.

Bittner, K., Adam, F., Cui, S., Körner, M. and Reinartz, P., 2018. Building footprint extraction from VHR remote sensing images combined with normalized DSMs using fused fully convolutional networks. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 11(8), pp.2615-2629.

Castrejon, L., Kundu, K., Urtasun, R. and Fidler, S., 2017. Annotating object instances with a polygon-rnn. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 5230-5238).

Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F. and Adam, H., 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation. In Proceedings of the European conference on computer vision (ECCV) (pp. 801-818).

Chen, Q., Wang, L., Waslander, S.L. and Liu, X., 2020. An end-to-end shape modeling framework for vectorized building outline generation from aerial images. ISPRS Journal of Photogrammetry and Remote Sensing, 170, pp.114-126.

Cheng, D., Liao, R., Fidler, S. and Urtasun, R., 2019. Darnet: Deep active ray network for building segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 7431-7439).

Girard, N. and Tarabalka, Y., 2018, July. End-to-end learning of polygons for remote sensing image classification. In IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium (pp. 2083-2086). IEEE.

Griffiths, D. and Boehm, J., 2019. Improving public data for building segmentation from Convolutional Neural Networks (CNNs) for fused airborne lidar and image data using active contours. ISPRS journal of photogrammetry and remote sensing, 154, pp.70-83.

Guo, Z., Chen, Q., Wu, G., Xu, Y., Shibasaki, R. and Shao, X., 2017. Village building identification based on ensemble convolutional neural networks. Sensors, 17(11), p.2487.

He, K., Zhang, X., Ren, S. and Sun, J., 2016. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).

He, K., Gkioxari, G., Dollár, P. and Girshick, R., 2017. Mask r-cnn. In Proceedings of the IEEE international conference on computer vision (pp. 2961-2969).

Huang, G., Liu, Z., Van Der Maaten, L. and Weinberger, K.Q., 2017. Densely connected convolutional networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4700-4708).

Jaccard, P., 1912. The distribution of the flora in the alpine zone. 1. New phytologist, 11(2), pp.37-50.

Krizhevsky, A., Sutskever, I. and Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems, 25.

Li, W., He, C., Fang, J., Zheng, J., Fu, H. and Yu, L., 2019. Semantic segmentation-based building footprint extraction using very high-resolution satellite images and multi-source GIS data. Remote Sensing, 11(4), p.403.

Li, Z., Wegner, J.D. and Lucchi, A., 2019. Topological map extraction from overhead images. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 1715-1724).

Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B. and Belongie, S., 2017. Feature pyramid networks for object detection. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2117-2125).

Ling, F., Li, X., Xiao, F., Fang, S. and Du, Y., 2012. Object-based sub-pixel mapping of buildings incorporating the prior shape information from remotely sensed imagery. International Journal of Applied Earth Observation and Geoinformation, 18, pp.283-292.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S. and Guo, B., 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 10012-10022).

Long, J., Shelhamer, E. and Darrell, T., 2015. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 3431-3440).

Marcos, D., Tuia, D., Kellenberger, B., Zhang, L., Bai, M., Liao, R. and Urtasun, R., 2018. Learning deep structured active contours end-to-end. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 8877-8885).

Manno-Kovács, A. and Ok, A.O., 2015. Building detection from monocular VHR images by integrated urban area knowledge. IEEE Geoscience and Remote Sensing Letters, 12(10), pp.2140-2144..

Pelizari, P.A., Spröhnle, K., Geiß, C., Schoepfer, E., Plank, S. and Taubenböck, H., 2018. Multi-sensor feature fusion for very high spatial resolution built-up area extraction in temporary settlements. Remote Sensing of Environment, 209, pp.793-807.

Ronneberger, O., Fischer, P. and Brox, T., 2015, October. U-net: Convolutional networks for biomedical image segmentation. In International Conference on Medical image computing and computer-assisted intervention (pp. 234-241). Springer, Cham.

Sirmacek, B. and Unsalan, C., 2009. Urban-area and building detection using SIFT keypoints and graph theory. IEEE transactions on geoscience and remote sensing, 47(4), pp.1156-1167.

Turker, M. and Koc-San, D., 2015. Building extraction from high-resolution optical spaceborne images using the integration of support vector machine (SVM) classification, Hough transformation and perceptual grouping. International Journal of Applied Earth Observation and Geoinformation, 34, pp.58-69.

Wang, J., Sun, K., Cheng, T., Jiang, B., Deng, C., Zhao, Y., Liu, D., Mu, Y., Tan, M., Wang, X. and Liu, W., 2020. Deep high-resolution representation learning for visual recognition. IEEE transactions on pattern analysis and machine intelligence, 43(10), pp.3349-3364.

Xiao, T., Liu, Y., Zhou, B., Jiang, Y. and Sun, J., 2018. Unified perceptual parsing for scene understanding. In Proceedings of the European Conference on Computer Vision (ECCV) (pp. 418-434).

Zhang, C., Hu, Y. and Cui, W., 2018. Semiautomatic right-angle building extraction from very high-resolution aerial images using graph cuts with star shape constraint and regularization. Journal of Applied Remote Sensing, 12(2), p.026005.

Zhao, H., Shi, J., Qi, X., Wang, X. and Jia, J., 2017. Pyramid scene parsing network. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2881-2890).

BIOGRAPHICAL NOTES

Hailun Yan is a graduate student in Geomatics Engineering of the University of Calgary, Canada. He holds a Bachelor's degree in geographic information science from the University of Toronto at Mississauga, Canada.

Ruisheng Wang is a professor in the department of geomatics engineering at the University of Calgary. Dr. Wang holds a Ph.D. in Electrical and Computer Engineering from McGill University, an M.Sc.E.in Geomatics Engineering from University of New Brunswick, and a B.Eng. in Photogrammetry and Remote Sensing from Wuhan University, respectively. His research interests are photogrammetry, remote sensing and computer vision.

CONTACTS

Mr. Hailun Yan
University of Calgary
2416 16 Ave NW
Calgary
CANADA
Tel. +1 (226)978-3703
Email: hailun.yan@ucalgary.ca