MAPPING INACCESSIBLE AREAS USING DEEP LEARNING BASED SEMANTIC SEGMENTATION OF VHR SATELLITE IMAGES WITH OPENSTREETMAP DATA

Minho Kim, Taehong Kwak, Jaeung Jung, Yongil Kim*
Seoul National University, mhk93@snu.ac.kr, mwsa302@snu.ac.kr, roosa0210@snu.ac.kr, yik@snu.ac.kr

ABSTRACT Remote sensing is crucial for mapping and developing geospatial information of inaccessible areas. In particular, supervised classification or semantic segmentation of very high resolution (VHR) satellite images are used to extract key features such as buildings, roads, vegetation, and water bodies, but these methods are limited by the need for ground truth data, which is physically unobtainable for remotely located areas. To address this limitation, OpenStreetMap (OSM) data can provide ground truth labels that can be modified for use in VHR satellite images. In this study, Geoeye-1 VHR satellite imagery and refined OSM labels were acquired in urban regions situated in Pyeongyang, North Korea and are integrated into a feature pyramid network-based segmentation model with a pre-trained EfficientNet-B1 backbone. Building and road extraction yielded an F1-score of 0.8806 and 0.9580, respectively. Building and road segmentation results are combined with vegetation and waterbody features from spectral index thresholding to map four fundamental spatial data that are crucial for the development and updating of geospatial information in inaccessible urban areas.

KEY WORDS: Very High Resolution Satellite Images, OpenStreetMap, inaccessible areas, semantic segmentation

1. INTRODUCTION

According to the Fundamental Act on National Spatial Data Infrastructure in South Korea, fundamental spatial data can be divided into 12 large categories based on the requirements for the creation and integration of geospatial information such as thematic maps and GIS products. For inaccessible areas, establishing and continuously updating fundamental geospatial information and maps are crucial for future development and decision-making. Airborne surveys and unmanned aerial vehicles have been widely used to acquire earth observations with finer detail. When mapping inaccessible and remotely situated areas, however, airborne sensors are limited by flight restrictions, flight time capacity, and physical distance. Alternatively, mid-resolution satellite images are used for land cover mapping and are effective for nationwide coverage, but lack the fine-grained spatial detail required for urban areas. To address these limitations, very high resolution (VHR) satellite images contain detailed, key information on the earth's surface such as built area, vegetation, impervious surface, water bodies. Furthermore, information extracted from VHR satellite images can be used to update outdated maps and geospatial data in inaccessible areas.

Semantic segmentation is an effective method to extract features from remote sensing images by allocating a semantic label to each coherent region of an image (Neupane et al., 2021). In recent studies, convolutional neural network (CNN)-based semantic segmentation models have been used to accurately classify urban objects, such as buildings and roads, which typically require a higher spatial resolution. Fully convolutional networks (FCN) use an encoder-decoder setup to extract meaningful feature maps and convert them into dense label maps (Neupane et al., 2021). Sun and Wang (2018) used FCN for semantic segmentation of VHR remotely-sensed images and DSM height information from the ISPRS Vaihingen dataset. Li et al. (2019) used a U-Net model for semantic segmentation-based building footprint extraction from Worldview-3 satellite images and multi-source GIS data. Wu et al. (2019) adapted a LinkNet architecture based on feature-forwarding from the encoder to the decoder. Lastly, Gao et al. (2018) used a multiple feature pyramid network (FPN) to exploit multi-level semantic features inherent in VHR satellite images for road extraction.

The aforementioned studies are based on supervised learning for semantic segmentation, but acquiring ground truth semantic labels in inaccessible regions for supervised learning is limited. The OpenStreetMap (OSM) platform provides labeled data on fundamental spatial data such as buildings, roads, water bodies, and vegetation at a global scale – even in secluded areas such as North Korea. However, OSM labels in inaccessible areas suffer from low to no updates and sparse annotation. In addition, OSM labels may not match with VHR satellite images due to the geometric location and sensor specifications.

In response to the need for accurate, high-resolution mapping of inaccessible areas, this study investigates deep learning-based semantic segmentation of fundamental spatial data in inaccessible urban areas in Pyeongyang, North Korea using VHR satellite imagery with refined OSM labels. Among the fundamental 12 spatial data categories, four classes (buildings, roads, vegetation, water) were selected based on the high number of related studies in the literature (Neupane et al., 2021). For this study, semantic segmentation is performed on the former two classes (buildings and roads) using UNet, LinkNet, and FPN, given the model's usage in related building and road extraction studies. The remaining classes (vegetation and water) were obtained using vegetation index thresholding. Each class was extracted individually and combined together in the final map.

2. DATA AND METHODS

2.1 Data

Geoeye-1 consists of five spectral bands and has a spatial resolution of 0.46 m (panchromatic) and 1.84 m (multispectral). The Geoeye-1 image used in this study was acquired on April 27th, 2017 in Pyeongyang, North Korea covering an area of approximately 7 km x 6.2 km. Only the multispectral bands were used in the experiments.

Ground truth labels were initially obtained from OSM. In more detail, building, road, vegetation, and water OSM layers were acquired. The OSM labels were refined via visual inspection using the Geoye-1 image and comparison with VHR satellite images provided by Google Satellite and Earth imagery. Refined labels considered the different acquisition conditions of the VHR image. Since OSM road layers are distributed as line vectors, primary and secondary road layers, typically consisting of highways and large arterial roads, were given a buffer of 6 pixels, while tertiary road layers such as collector and local roads were assigned a buffer of 2.5 pixels.

2.2 Model

For this study, three models were selected based on their widespread usage in building and road extraction via supervised semantic segmentation. The convolutional layers and kernels in these models use multi-scale and multi-level information from the image, which is important to distinguish various sizes of urban features. The model input layer receives a 256 x 256 image for three spectral bands and incorporates the recently proposed EfficientNet-B1 trained on ImageNet as a backbone model (Tan and Le, 2019).

2.2.1 U-Net

U-Net was initially designed for biomedical imaging, but has also been adapted for satellite image segmentation (Li et al., 2019). The U-Net architecture consists of convolutional layers with multiple 3x3 kernels, batch normalization and non-linear rectified linear unit activation, max-pooling layers for downsampling, upsampling layers for recovering feature map sizes, and a concatenation layer for combining multilevel features.

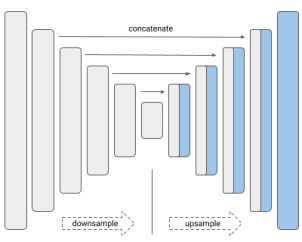

Figure 2. U-Net architecture (Yakubovskiy, 2019).

2.2.2 LinkNet

LinkNet is adapted from U-Net and is composed of ResBlocks from residual modules (He et al., 2016) and integrates shallow and deep feature maps using concatenation in contrast to simple addition.

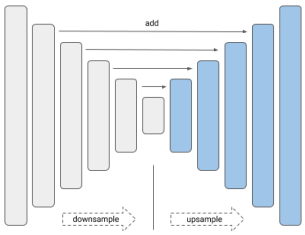

Figure 3. LinkNet architecture (Yakubovskiy, 2019).

2.2.3 FPN

FPN combines multi-level feature maps using bottom-up and top-down paths linked by lateral connections which merge feature maps of the same spatial size. These combined feature maps are concatenated and used to generate the final segmentation result.

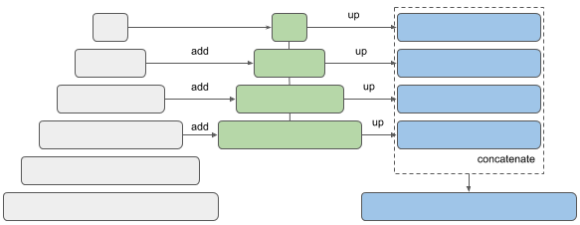

Figure 4. FPN architecture (Yakubovskiy, 2019).

2.3 Implementation Details

The training and test datasets are split into a 70% to 30% ratio and are composed of multispectral Geoeye-1 and refined OSM labels. The datasets are geographically split to avoid data redundancy. Data augmentation is used to increase the dataset size by using random cropping of 256 x 256 pixels (from 512 x 512 patches), random scaling, flipping, as well as brightness and contrast changes. The model was implemented in Tensorflow 2.0 and trained with an RTX 2070 Super GPU for 500 epochs using the Adam optimizer with a batch size of 4 and a learning rate of 0.0002 with an exponential decay at 0.96. Cross entropy (CE) was used as the loss function as shown in Eq. 1.

$$\mathrm{CE}(i,j) = -\frac{1}{N}\sum_{n}^{N}[y_{ij}\log(x_{ij}) + (1-y_{ij})\log(1-x_{ij})] \quad (1)$$

where N is the number of classes, x_i is the segmentation result, and y_i is the ground truth label.

Two experiments were conducted to compare model performance and input band combinations. In the first experiment, three segmentation models (U-Net, LinkNet, FPN) with an EfficientNet-B1 backbone were used. In the second experiment, RGB and NIR-R-G band combinations were inputted in the best-performing segmentation model.

2.4 Evaluation Metrics

F1-score and intersection of union (IoU) were used to quantitatively evaluate the segmentation models based on their widespread usage in related studies (Neupane et al., 2021). F1-score is the harmonic combination of precision and recall scores, while IoU is the intersection area of the segmentation result (denoted *A*) and the ground truth label (denoted *B*) divided by the union of the two.

$$F1 = \frac{2(Precision \cdot Recall)}{Precision + Recall} \quad (2)$$

$$IoU = \frac{Area(A \cap B)}{Area(A \cup B)} \quad (3)$$

3. RESULTS AND DISCUSSION

3.1 Experiment 1: Buildings

Building segmentation results are displayed in Table 1. FPN returned the highest IoU and F1 scores of 0.8017 and 0.8806, respectively. The effectiveness of FPN can be attributed to the integration of rich semantic information from multi-level layers to better distinguish multi-scale building shapes. Despite the small number of training samples, the high F1-score indicates the importance of the pre-trained EfficientNet-B1 backbone to extract deep feature information for the segmentation models. The CE loss for all models shows a continuously decreasing trend over time, suggesting a potential margin for improvement with hyperparameter tuning.

Table 1. Segmentation results for building features.

Model	IoU	F1
LinkNet	0.7794	0.8641
U-Net	0.7985	0.8778
FPN	**0.8017**	**0.8806**

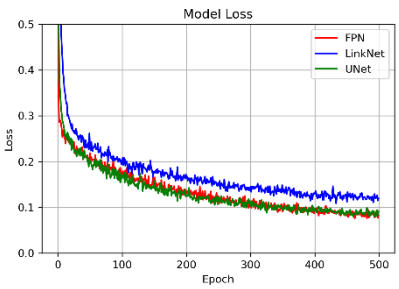

Figure 5. CE Loss graph for building segmentation.

3.2 Experiment 1: Roads

Due to the combination of buffered large and narrow roads, this experiment can be considered a multi-scale segmentation problem. However, there was not much variability in the results from the segmentation models, as shown in Table 2. LinkNet and U-Net recorded higher F1-score and IoU values, respectively, but only to a slim extent. Model loss continues to decrease, albeit at small increments per epoch, as shown in Figure 6.

Table 2. Segmentation results for road features.

Model	IoU	F1
LinkNet	0.9353	**0.9583**
U-Net	**0.9359**	0.9580
FPN	0.9356	0.9580

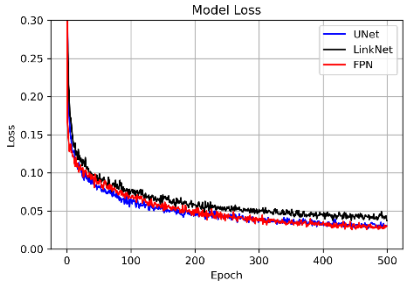

Figure 6. CE Loss graph for road segmentation.

3.3 Experiment 1: Visual Inspection

Qualitative assessment of the segmentation results is also important to evaluate model performance. Visual inspection is conducted using results from FPN with the pre-trained EfficientNet-B1 backbone since this model yielded the best results for building extraction, while results were variable for road segmentation. Figures 7 and 8 show the results from building and road segmentation, respectively. The left image overlays the refined ground truth labels and segmentation results over the Geoeye-1 satellite image. The right image compares OSM, refined ground truth labels, and segmentation results.

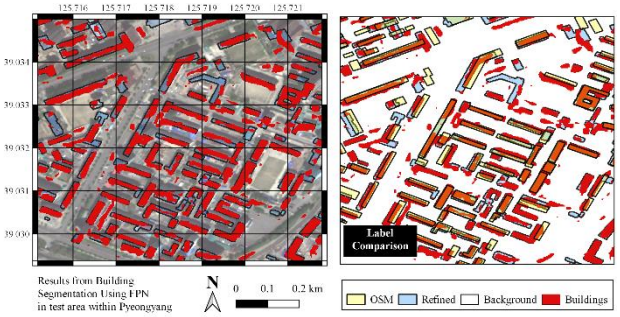

Figure 7. Building segmentation results using FPN.

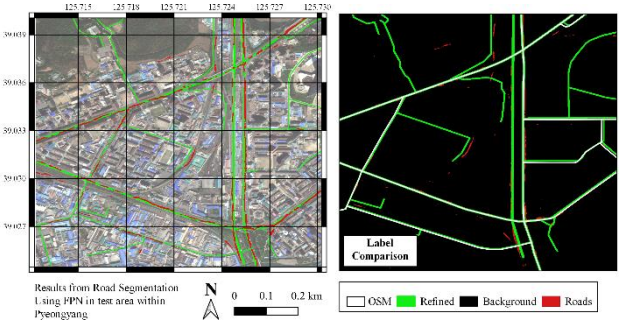

Figure 8. Road segmentation results using FPN.

Building segmentation results were found to be well-classified within or near the ground truth labels. FPN was robust to the distribution of differently sized and shaped buildings in the VHR image. In addition, FPN was able to avoid side-views and shadows caused by high-rise buildings. In contrast, while the road segmentation results were generally on the same line as the ground truth labels, the result contained many discontinuities and noise.

Label refinement and FPN enabled the mapping of many additional buildings that were originally omitted in the OSM dataset. However, many of the building segmentation results lost their rigid shape, which is common with convolution-based models. Post-processing using conditional random fields and graph convolutional networks can help improve visual quality. Moreover, the majority of the road segmentation results appear sparse or detached. This result can be explained by the relatively small input image size and the lack of regional and contextual information used in the model. Dilated convolutions and context aggregation methods can be explored to improve the linkage of segmentation results.

Building and road segmentation results are combined with vegetation and water to create a multi-class map of Pyeongyang consisting of four of the fundamental spatial data categories, as shown in Figure 9. Vegetation areas were thresholded at NDVI values greater than 0.4 and post-processed using a majority filter. Water bodies were thresholded using an NDWI value greater than 0.25.

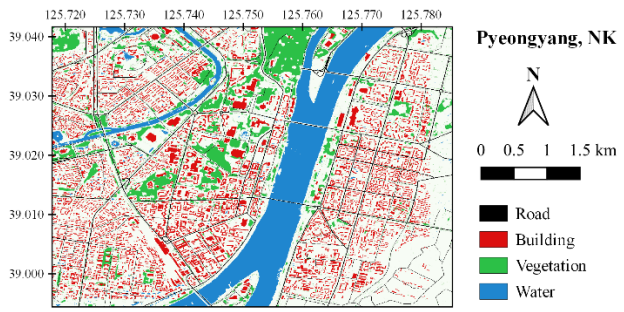

Figure 9. 4-class map using FPN and index thresholding.

3.4 Experiment 2: Band Comparison

This experiment also compared the effect of RGB and false composite (NIR-R-G) combinations. FPN with a pre-trained EfficientNet-B1 backbone was used as the main comparison model. The results organized in Table 3 demonstrate that the NIR band does not necessarily help improve segmentation performance.

Table 3. Comparison of band combinations.

Feature	Bands	IoU	F1
Building	RGB	**0.8017**	**0.8806**
	NIR-R-G	0.7488	0.8374
Road	RGB	**0.9356**	**0.9580**
	NIR-R-G	0.9308	0.9525

Future work should include additional VHR satellite images from different sites and acquisition conditions to assess model robustness and expand the training dataset. While this study used refined OSM labels, current OSM labels can be maintained if high-resolution orthoimages or orthorectified VHR satellite images can be used.

4. CONCLUSION

Acquiring fundamental spatial data is crucial to develop and update geospatial information in inaccessible areas. Geoeye-1 satellite imagery and refined OSM labels were processed using a pre-trained EfficientNet-B1 backbone with an FPN segmentation model to extract building and road features to a high level of accuracy. These urban features were combined with vegetation and water features from spectral index thresholding to generate multi-class maps that can be used as a basis for future monitoring and map updating applications.

References

Gao, X., Sun, X., Zhang, Y., Yan, M., Xu, G., Sun, H., ... & Fu, K. (2018). An end-to-end neural network for road extraction from remote sensing imagery by multiple feature pyramid network. IEEE Access, 6, 39401-39414.

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

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

Neupane, B., Horanont, T., & Aryal, J. (2021). Deep Learning-Based Semantic Segmentation of Urban Features in Satellite Images: A Review and Meta-Analysis. Remote Sensing, 13(4), 808.

Sun, W., & Wang, R. (2018). Fully convolutional networks for semantic segmentation of very high resolution remotely sensed images combined with DSM. IEEE Geoscience and Remote Sensing Letters, 15(3), 474-478.

Tan, M., & Le, Q. (2019). Efficientnet: Rethinking model scaling for convolutional neural networks. In International Conference on Machine Learning (pp. 6105-6114). PMLR.

Wu, M., Zhang, C., Liu, J., Zhou, L., & Li, X. (2019). Towards accurate high resolution satellite image semantic segmentation. IEEE Access, 7, 55609-55619.

Yakubovskiy, P. (2019) Segmentation Models. Github repository.

Acknowledgements

This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(NRF-2019R1I1A2A01058144). This research was supported by the project “Laying the Groundwork for Peace and Unification” funded by the Institute for Peace and Unification Studies (IPUS) at Seoul National University.