Accredited Ranking SINTA 2
Decree of the Director General of Higher Education, Research, and Technology, No. 158/E/KPT/2021
Validity period from Volume 5 Number 2 of 2021 to Volume 10 Number 1 of 2026

Published online on: **http://jurnal.iaii.or.id**

JURNAL RESTI
(Rekayasa Sistem dan Teknologi Informasi)
Vol. 6 No. 3 (2022) 376 - 379 ISSN Media Electronic: 2580-0760

Segmentation of Small Objects in Satellite Imagery Using Dense U-Net in Massachusetts Buildings Dataset

Muhammad Iqbal Izzul Haq[1], Aniati Murni Arymurthy[2], Irham Muhammad Fadhil[3]
[1,2,3]Ilmu komputer, Fakultas Ilmu Komputer, Universitas Indonesia
[1]iqbal08@ui.ac.id, [2]aniati@cs.ui.ac.id, [3]irham.muhammad@ui.ac.id

Abstract

Class imbalance is a serious problem that disrupts the process of semantic segmentation of satellite imagery in urban areas in Earth remote sensing. Due to the large objects dominating the segmentation process, small object are consequently limited, so solutions based on optimizing overall accuracy are often unsatisfactory. Due to the class imbalance of semantic segmentation in Earth remote sensing images in urban areas, we developed the concept of Down-Sampling Block (DownBlock) to obtain contextual information and Up-Sampling Block (UpBlock) to restore the original resolution. We proposed an end-to-end deep convolutional neural network (DenseU-Net) architecture for pixel-wise urban remote sensing image segmentation. this method to segmentation the small object in satellite imagery.The accuracy of the small object class in this study was further improved using our proposed method. This study used data from the Massachusetts Buildings dataset using Dense U-Net method and obtained an overall accuracy of 84.34%.

Keywords: class imbalance, dataset, end-to-end, convolution, denseU-net

1. Introduction

The semantic segmentation task of Earth remote sensing images classifies each pixel of the Earth remote sensing image. In remote sensing of a very high spatial resolution (VHR) image, spectral resolution and spatial resolution may be limited [1], so the sensor sells spectral resolution to obtain spatial resolution. Therefore, when analyzing pixel spectral information, the spatial context must take into spatial functions such as texture [2] or morphology[3].

The above spatial features require manual extraction, and the purpose of the in deep learning is to train a parametric function extraction system with end-to-end classification to avoid manual extraction of spatial functions. With the development of deep learning, the convolutional neural network (CNN)[4] Researchers have also made a number of advances in computer vision, such as image classification, object analysis, semantic segmentation, and other tasks. CNN huge success is largely due to its excellent visual data feature. Deep Networks extract has a better function than artificial feature engineering[5].

In 2017, Iglovikov [6] implemented the idea of bypassing U-Net connections and combining shallow and deep layer components to improve the use of component information. Liu and others [7]designed Hourglass-Shape Networks (HSN) and introduced the first module to provide networks with multi-level receiving regions with different contexts. In the same year,Volpi & Tuia [8] proposed a CNN-FPL-based U-Net model for mapping with the benefits of a contracting path, followed by a multilayer deconvolution layer to return functions to resolution original photo. This technique helps keep details out of the drawing.

In 2018, Gao et al.[9] proposed a Multiple Feature Pyramid Network (MFPN), which is used to record video paths for remote recognition and provides a weighting function to solve the class imbalance problem caused by fewer flows.

In 2020, the CNN AlexNet [10] has been porposed . This classification is performed by Lung CT image recognition. In addition, ZFNet[11], VGGNet[12] and GoogleNet [13] are proposed. Deep Learning has been applied to remote semantic image segmentation. Daniel Gritzner,et al [14] used deep neural networks for semantic segmentation of aerial images, Machine Learning for Aerial Image performed an experiment on a data file published by the state of Massachusetts [15]and a dataset published by the state of New York.

Accepted: 02-04-2022 | Received in revised: 13-06-2022 | Published: 30-06-2022

We propose end-to-end (DenseU-Net) based on U-Net [16]. DenseU-Net interconnects CNN functions through a cascade and through a symmetric structure composed of continuous downBlocks and UpBlocks. Shallow layer properties, such as color and texture, combine with the abstract semantic features of the deep layer by removing connections.

2. Research Methods

2.1 System Architecture

In this paper, we use Dense U-Net architecture based on convolutional neural network (CNN). The main idea of the DenseU-Net is to connect convolutional neural network features through cascade operations and use its symmetrical structure to fuse the detail features in shallow layers and the abstract semantic features in deep layers. The dense U-Net layer is divided into three parts, namely the downblock, bridging, and upblock sections. Each layer contains convolution, max pooling, and dropout. In the downblock, there are 4 layers, the first layer with the number of filters 1 and a resolution of 1024x1024, followed by the second layer with the number of filters 2 and a resolution of 512x512, followed by the third layer with the number of filters 4 and a resolution of 256x256, and ending with the fourth layer with the number of filters 8 and 128x128 resolution. After the downblock continued with the bridging which functions as a bridge between the downblock and upblock with a filter number of 16 and a resolution of 64x64.

Next there is the upblock which consists of 4 layers, the first layer with 8 filters and 128x128 resolution, followed by the second layer with 4 filters and 256x256 resolution, followed by the third layer with 2 filters and 512x512 resolution, and ends with the fourth layer with 1 filters and a resolution of 1024x1024.

In each convolution, both in the downblock and upblock sections there are three layers, each of which contains a Conv2D layer, batch normalization, and a ReLU activation layer. Between the convolution second convolution third dense layer. Figure 1 describes the proposed dense U-net system architecture.

2.2 Methodology

The first step is toload a dataset containing satellite images and ground truth which are the basis for segmentation. A total of 151 data are then divided into training data, validation, and test data with a proportion of 70% training data, 21% validation data, and 9% test data and converting the image to grayscale.

The second step is to resize the image to a size of 1024x1024 because it pays attention to the limitations of the Google Colab GPU. The third step is to create a border around the marker indicated by the ground truth.

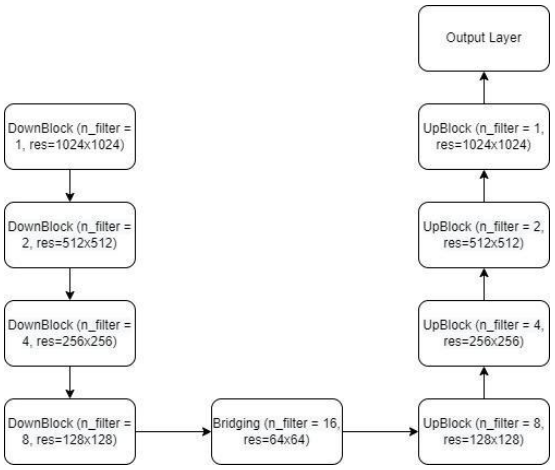

Figure 1. Dense U-Net architecture

The next step is to define the Dense U-Net architecture that has been described previously and then conduct training using training and validation data of 50 epochs. Early stopping is used to prevent overfitting (a huge difference between accuracy during training and testing).which if the learning rate is less than the specified then the training process will stop.

After the training was completed, the model was tested using the test data previously described and then the performance of the model that had been created was measured based on the parameters of accuracy, precision, recall, and F1-score.

2.3. Dataset

Dataset used is the Massachusetts Building Dataset. Dataset sourced from Kaggle. The dataset contains 151 satellite images of the urban area of Boston, United States. Each image has a resolution of 1500x1500 with an area of 2.25 km2. Small objects that are classified in this study are houses objects found in satellite images.

The following is an example of an image contained in the dataset and ground truth, Figure 2.

Figure 2 Example of a satellite image depicting a house

The dataset covers mostly urban and suburban areas and buildings of all sizes, including individual houses and garages, are included in the labels. The datasets make

DOI: https://doi.org/10.29207/resti.v6i3.3993

use of imagery released by the state of Massachusetts. All imagery is rescaled to a resolution of 1 pixel per square meter. The target maps for the dataset were generated using data from the OpenStreetMap project. Target maps for the test and validation portions of the dataset were hand-corrected to make the evaluations more accurate. The following are the example of a ground truth in dataset, Figure 3.

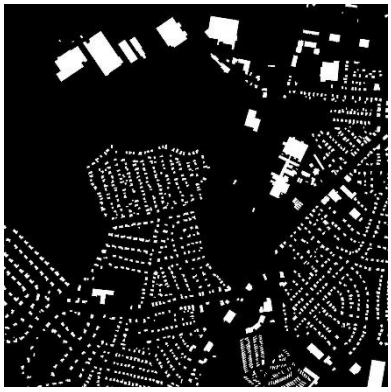

Figure 3 Example of a Ground Truth that Describes The Part to be Segmented

3. Results and Discussions

3.1 Experimental Setup

In this study we use several parameters in its implementation . We describe these parameters below, Table 1.

Table 1. Parameter Implementation

Parameter	value
Dataset	Massachusetts Building Dataset
Resolution	1024 x 1024
Number of filters	16
Epoch	50
Kernel size	3
Learning rate	1e-3(1x10-3)
Optimizer	Adam
Dropout	0.05
Loss Function	Binary Crossentrophy

This study uses early stopping which process stops the training when the learning rate is less than 1x10^-5 to prevent overfitting. While the following table describes the implementation environment.

Table 2. Implementation Environtment

Parameter	value
Text Editor	Google Collaboratory
Runtime	GPU
Tensorflow Version	2.7.0

Segmentation of small objects in the form of houses is carried out using the Dense U-Net architecture. Training accuracy, test accuracy, precision, recall, F1-score, and loss are used in evaluating the model created. The following table shows the values of accuracy, precision, recall, F1-score, and loss in the training, validation, and testing phases.

Table 3. Results of accuracy, precision, recall, F1-score and loss on training data, validation, and test data

Parameter	Training	Validation	Testing
Accuracy	0.8346	0.8375	0.8434
Precision	0.6966	0.7433	0.7483
Recall	0.5069	0.4434	0.4054
F1-Score	0.5831	0.5503	0.5259
Loss	0.2413	0.2380	0.2287

The result from the experiment show that the accuracy of testing data 0.8434 and the loss of testing data 0.2287. The following figure shows the relationship between training loss and validation loss in each epoch.

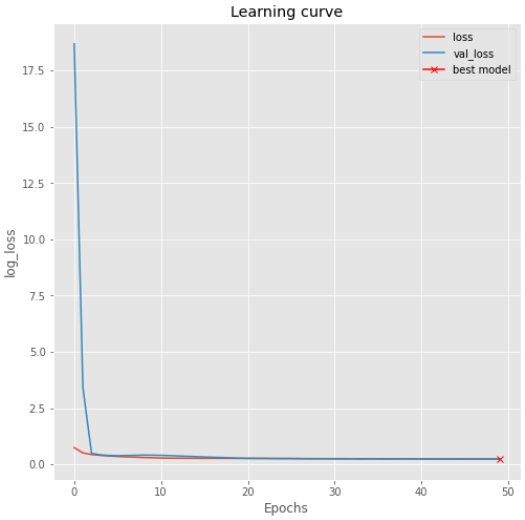

Figure 4 Comparison of *loss* and *validation loss* against epochs

the *loss* and *validation loss* decreases with each *epoch* and the best results are found in the 50th epoch. This result shows that there is no overfitting on *loss* and validation *loss*.The following are the results of segmentation using the proposed Dense U-Net model.

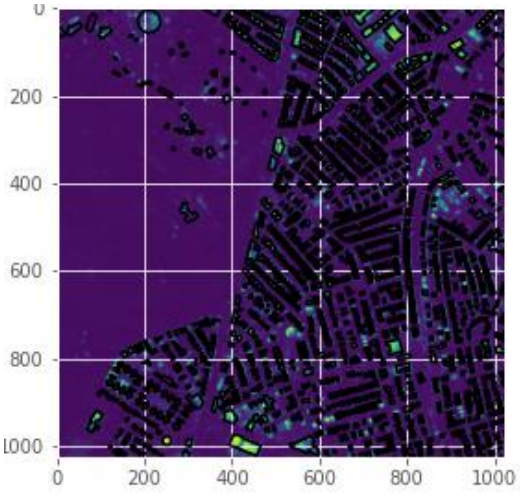

Figure 5 The results of segmentation of small objects on the test data

DOI: https://doi.org/10.29207/resti.v6i3.3993

The Dense U-Net model are proven to be able to segment small objects in the form of houses in massachusetts buildings with 50 epoch , 16 filters and optimizer adam produce an overall accuracy of 84.34%. This result better than previous method because our method use cascade operations and its symmetrical structure to fuse the detail features in shallow layers and the abstract semantic features in deep layers.

4. Conclusion

To solve the problem of segmentation class imbalance in urban remote sensing imagery, Dense U-Net architecture convolutional neural network for pixel based urban remote sensing image segmentation is proposed in this paper. The DenseU-Net applies DownBlocks inthe contracting path to extract CNN features, and applies UpBlocks to restore image resolution. This method preserves details such as the color and texture of the image. Combined with the abstract semantic features of the deep layer by removing connections. We implement using google collaboratory and GPU runtime. Experiments on the Massachusetts Building Dataset show that DenseU-Net can identify small objects well for the "Building" class with an overall accuracy 84,34% with 50 epoch and loss in testing data 0,2287 with no overfitting. Decreasing in loss and validation loss in each epoch , for further experiments try other methods such as VGG U-Net to improve the accuracy.

Reference

[1] S. Liu, Y. Li, and X. Tong, "Superpixel-based multiple change detection in very-high-resolution remote sensing images," *RSIP 2017 - Int. Work. Remote Sens. with Intell. Process. Proc.*, no. 3, pp. 25–27, 2017, doi: 10.1109/RSIP.2017.7958817.

[2] O. Regniers, L. Bombrun, V. Lafon, and C. Germain, "Supervised Classification of Very High Resolution Optical Images Using Wavelet-Based Textural Features," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 6, pp. 3722–3735, 2016, doi: 10.1109/TGRS.2016.2526078.

[3] X. Cao, R. Li, L. Wen, J. Feng, and L. Jiao, "Deep Multiple Feature Fusion for Hyperspectral Image Classification," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 11, no. 10, pp. 3880–3891, 2018, doi: 10.1109/JSTARS.2018.2866595.

[4] Y. Seo and K. S. Shin, "Image classification of fine-grained fashion image based on style using pre-trained convolutional neural network," *2018 IEEE 3rd Int. Conf. Big Data Anal. ICBDA 2018*, pp. 387–390, 2018, doi: 10.1109/ICBDA.2018.8367713.

[5] Z. Wang, X. Xiang, Z. Zhao, and F. Su, "Deep Image Retrieval: Indicator and Gram Matrix Weighting for Aggregated Convolutional Features," *Proc. - IEEE Int. Conf. Multimed. Expo*, vol. 2018-July, pp. 1–6, 2018, doi: 10.1109/ICME.2018.8486547.

[6] V. Iglovikov, S. Mushinskiy, and V. Osin, "Satellite Imagery Feature Detection using Deep Convolutional Neural Network: A Kaggle Competition," 2017, [Online]. Available: http://arxiv.org/abs/1706.06169

[7] Y. Liu, D. M. Nguyen, N. Deligiannis, W. Ding, and A. Munteanu, "Hourglass-shape network based semantic segmentation for high resolution aerial imagery," *Remote Sens.*, vol. 9, no. 6, pp. 1–24, 2017, doi: 10.3390/rs9060522.

[8] M. Volpi and D. Tuia, "Dense semantic labeling of subdecimeter resolution images with convolutional neural networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 2, pp. 881–893, 2017, doi: 10.1109/TGRS.2016.2616585.

[9] X. Gao *et al.*, "An End-to-End Neural Network for Road Extraction from Remote Sensing Imagery by Multiple Feature Pyramid Network," *IEEE Access*, vol. 6, pp. 39401–39414, 2018, doi: 10.1109/ACCESS.2018.2856088.

[10] T. Wang, Y. Zhao, L. Zhu, G. Liu, Z. Ma, and J. Zheng, "Lung CT image aided detection COVID-19 based on Alexnet network," *Proc. - 2020 5th Int. Conf. Commun. Image Signal Process. CCISP 2020*, pp. 199–203, 2020, doi: 10.1109/CCISP51026.2020.9273512.

[11] S. S. Kaddoun, Y. Aberni, L. Boubchir, M. Raddadi, and B. Daachi, "Convolutional Neural Algorithm for Palm Vein Recognition using ZFNet Architecture," *BioSMART 2021 - Proc. 4th Int. Conf. Bio-Engineering Smart Technol.*, no. iv, 2021, doi: 10.1109/BioSMART54244.2021.9677799.

[12] X. Liu, M. Chi, Y. Zhang, and Y. Qin, "Classifying High Resolution Remote Sensing Images by Fine-Tuned VGG Deep Networks," in *IGARSS 2018 - 2018 IEEE International Geoscience and Remote Sensing Symposium*, Jul. 2018, pp. 7137–7140. doi: 10.1109/IGARSS.2018.8518078.

[13] Y. Yang, P. Bi, and Y. Liu, "License Plate Image Super-Resolution Based on Convolutional Neural Network," *2018 3rd IEEE Int. Conf. Image, Vis. Comput. ICIVC 2018*, pp. 723–727, 2018, doi: 10.1109/ICIVC.2018.8492768.

[14] D. Gritzner and J. Ostermann, "Minimizing Manual Labeling Effort for The Semantic Segmentation of Aerial Images," in *2021 IEEE Statistical Signal Processing Workshop (SSP)*, Jul. 2021, pp. 81–85. doi: 10.1109/SSP49050.2021.9513774.

[15] Z. Cheng and D. Fu, "Remote Sensing Image Segmentation Method based on HRNET," in *IGARSS 2020 - 2020 IEEE International Geoscience and Remote Sensing Symposium*, Sep. 2020, vol. 2507, no. February, pp. 6750–6753. doi: 10.1109/IGARSS39084.2020.9324289.

[16] R. Dong, X. Pan, and F. Li, "DenseU-Net-Based Semantic Segmentation of Small Objects in Urban Remote Sensing Images," *IEEE Access*, vol. 7, pp. 65347–65356, 2019, doi: 10.1109/ACCESS.2019.2917952.

DOI: https://doi.org/10.29207/resti.v6i3.3993