Cau Giay: A Dataset for Very Dense Building Extraction from Google Earth Imagery

Anh Nguyen[1], Hung Luu[1,2], Anh Phan[1], Hung Bui[1], and Thanh Nguyen[1]
[1]Vietnam National University of Engineering and Technology
Hanoi, Vietnam
[2] School of Electrical and Data Engineering, University of Technology Sydney
New South Wale, Australia
*Correspondence author: hunglv@fimo.edu.vn

***Abstract*—One of the major topics in photogrammetry is the automated extraction of building from data acquired by airborne sensors. What makes this task challenging is the very heterogeneous appearance and dense distribution of buildings in urban areas. While many dataset have been established, none of them pay attention to developing cities where buildings are not well planned. To complement the development of building extraction algorithms, a dataset of high resolution satellite image is constructed in this paper covering Cau Giay district, Hanoi, Vietnam. The dataset consists of 2100 images of size 1024×1024 pixels extracted from Google Earth. Shape, size, and construction material differ greatly from building to building, thus make it challenging for state-of-the-art algorithm to accurately extract building location. Some baselines are provided using Convolutional Neural Networks (CNNs). Experimental results show that U-Net model trained with Mean Square Error loss is able to achieve comparable results ($OA = 92.04$).**

***Index Terms*—building extraction, semantic segmentation, open source**

I. Introduction

Recently, with the advantages of large scale monitoring and fast-updated, high resolution satellite image has been widely used for building extraction. The established building maps has many applications in infrastructure monitoring and management, urban planing, as well as city understanding. Since high resolution satellite image has become more accessible and affordable [1], many dataset for building extraction have been established, providing high quality images with high spatial resolution of less than 1 meter and rich spectral information. However, there remains limitation in establishing a more diversity dataset for building extraction. Most of available dataset such as ISPRS Vaihingen [2], ISPRS Postdam [3], SpaceNet [4], and Microsoft US Building Footprint [5] pay their interest in developed cities where buildings are well planned. Meanwhile, cities in developing countries where rapid urbanization are happening without restricted planning receive less focus. A dataset of highly dense and complex structure of buildings in these areas may benefit state-of-the-art algorithms for better generalization.

One of the main problem for constructing dataset in developed cities is that they can not afford the price for high resolution satellite image at scale. Thus, obtaining these data from free and open source might be considered. Recently, satellite image extracted from Google Earth received a lot of attention for various applications (e.g. scattered shrub detection [6]; ship detection [7]) including rooftop and road extraction [8]. While these images are freely available for research purpose [9], the image quality are nowhere comparable to established dataset. Thus, it requires further analysis and investigation to develop more sophisticated model for building extraction.

Recent developments in deep convolutional neural networks (CNNs) provide an unique opportunity to achieve remarkable building extraction performance in the remote sensing society [1]. Building extraction can be formulated as semantic segmentation task where there are only two label building and non-building. Since then, many works have been proposed based on the architecture of well-known semantic segmentation networks such as U-Net [14], FCN [12], Convolutional and Deconvolutional Networks [13].

Based on discussions above, a dataset for very dense building rooftop extraction is constructed with image from Google Earth. Specifically, it contains 2100 images of size 1024×1024 pixels cover Cau Giay district, Hanoi, Vietnam. Our contributions are as follows:

- A dataset for very dense building rooftop extraction is constructed. Unlike other dataset which focus on developed cities with sparse and well planned buildings, our dataset covers very dense building area with high variation in term of building rooftop shape and size. The detailed data information will be presented in Section II.
- Second, some results based on U-Net, a widely used CNN architecture for semantic segmentation, are provided as baselines.

This paper is organized as follows. Section II presents the details of the dataset. Section III contains the brief descriptions of baseline methods. Finally, section IV and Section V present the experimental results and conclusions, respectively.

II. Google Earth Dataset

A. Study Area

The dataset covers the administrative boundaries of Cau Giay district, Hanoi, Vietnam (see Fig. 1) with the area of $12.03km^2$ and the population density of $20,931$ people per square kilometer as of 2017 [10]. It's ten times higher than average population density of Hanoi ($2,239$ people per

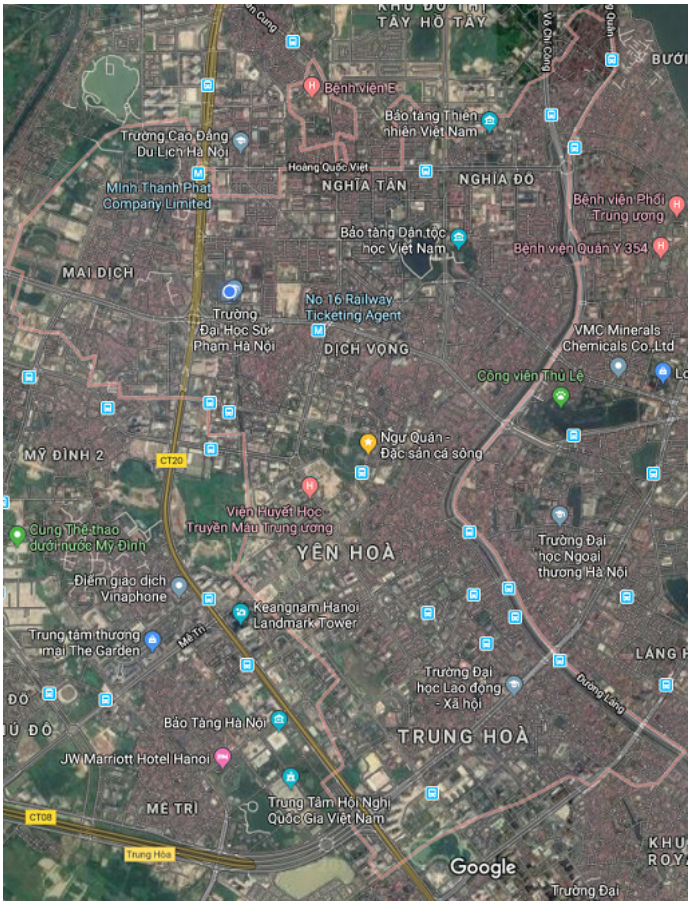
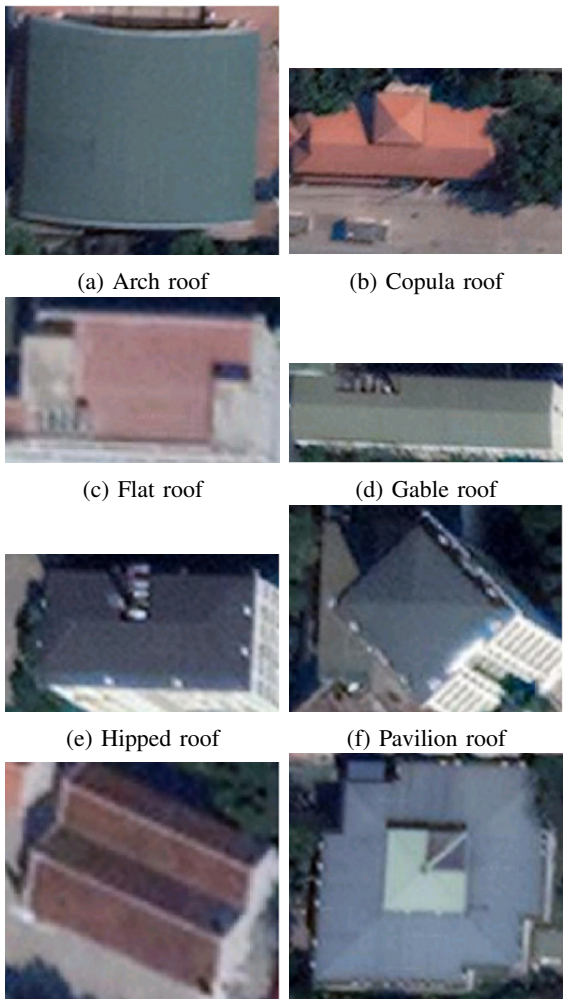


Fig. 1: The administrative boundaries of Cau Giay district, Hanoi, Vietnam.



(a) Arch roof (b) Copula roof

(c) Flat roof (d) Gable roof

(e) Hipped roof (f) Pavilion roof

(g) Saw-tooth roof (h) Combination

Fig. 2: Nine different roof types in Cau Giay area.

square kilometer), and 73 times higher than average population density of Vietnam (286 people per square kilometer) [11]. As such, this area is one of the densest urban area in Vietnam.

Due to high population density, tube-house is the most common architecture in this area with the narrow-shaped facade and great length. Meanwhile, roof shapes and roof materials differ greatly from building to building. In total, nine roof types have been observed (see Fig. 2).

B. Dataset Description

The images are extracted from Google Earth at zoom level of 22, and come as 24-bit files in Red-Green-Blue (RGB) format. Since Google Earth imagery are mosaic-ed from various sources, we can not guarantee as much in terms of quality or appearance. Many images are affected by a variety of artifacts such as cloud shadow, blurring effect, or non-ortho view (see Fig. 3).

Buildings rooftop in each image have been manually annotated and the ground truth data (label images) are provided together with Google Earth image (see Fig. 4). Occasionally, parts of some buildings are highly ambiguous (be covered by shadow or may be distorted in the original image). They are included as long as the annotator is reasonably sure the pixels belong to the buildings. Besides, the side-wall of buildings may appear in the image since many of them have non-ortho view. In this dataset, only building rooftop is considered, while the side-wall is ignored.

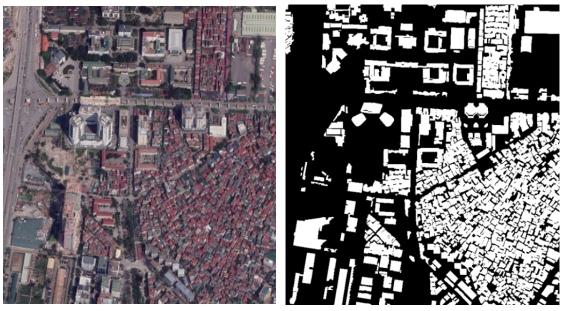
The area is manually divided into training, validation, and testing regions. The Google Earth image were subdivided into patches of size 1024×1024 pixels and were automatically assigned as training, validation, and testing set according to its corresponding region. The patches in training set cannot overlap with other patches in validation and test set, and vice versa. However, two patches in the same set can be overlapped. This helps increase the volume of dataset which is pre-requisite for deep learning model to learn. In total, the data set contains 2100 patches of size 1024×1024 pixels in which 1260 patches are used for training, 140 patches are used for validation, and 700 patches are used for testing

To this end, some properties of our dataset that make it challenging for building extraction algorithms are that:

- The diversity in shape, size and construction material of

(a) (b)

Fig. 3: Visualization quality of extracted images. (a) Good quality image with near-ortho view and high resolution (b) Bad quality image with non-ortho view and is affected by cloud shadow.



(a) (b)

Fig. 4: Example patch of Cau Giay dataset (a) Google Earth image (b) Ground truth.

roof top.

- The variation in resolution, incident angle, and quality of the Google Earth image.
- The high density of buildings.

III. Baseline Methods

Currently, there are many semantic segmentation methods in deep learning for building footprints extraction such as Fully Convolutional Network (FCN) [12], Convolutional and Deconvolutional Networks [13], U-Net [14]. These models often composed of two linked parts. The first part is a encoder network which computes feature maps at different depth layers. The second part is a decoder network which up-sampling the feature maps and then generating a map of pixel-wise probabilities at original resolution. In this paper, U-Net with ResNet backbone was used as our baselines.

A. U-Net with ResNet backbone

1) ResNet: ResNet is a Convolutional Neural Network (CNN) architecture, made up of series of residual blocks (ResBlocks) with skip connections [15]. Fig. 5 represents the architecture of a ResBlock. Let H_{i-1} denotes the output of $i-1^{th}$ block, $f_i(.)$ represents a series of convolutions, batch normalisation and linear functions in i^{th} block, we obtain:

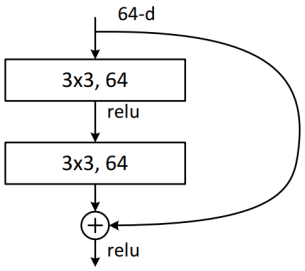


Fig. 5: The architecture of ResBlock (image from [15]).

$$H_i = ReLU(f_i(H_{i-1}) + id(H_{i-1})) \tag{1}$$

where $id(.)$ is identity transformation, and we assume a ReLU [16] activation function.

2) U-Net: U-Net was first developed for medial image segmentation [14]. It consists of an encoder part and a decoder part. The encoder part follows the typical architecture of a convolutional network (ResNet-50 in this case) which is used to learn the image features. The decoder part uses transposed convolutions to up-sampling the learned features map to original resolution. At the final layer a 1x1 convolution is used to map each feature vector to the desired number of classes (building or non-building).

B. Loss Functions

Mean Squared Error Loss (MSE) and Cross Entropy Loss (CE) are widely used for training semantic segmentation model. In this work, we trained two identical U-Net models with MSE and CE loss as baselines.

1) Cross Entropy Loss: Let $P(Y=0) = p$ and $P(Y=1) = 1p$. The predictions are given by the logistic/sigmoid function $P(\hat{Y}=0) = 1 - \frac{1}{1+e^{-x}} = \hat{p}$ and $P(\hat{Y}=1) = 1 - \frac{1}{1+e^{-x}} = 1 - \hat{p}$. Then cross entropy (CE) can be defined as follows:

$$CE(p, \hat{p}) = -(p \log \hat{p} + (1-p) \log 1 - \hat{p}) \tag{2}$$

2) Mean Squared Error Loss: Let N is the number of pixels, y_i is the ground truth (0 or 1), and $\hat{y}_i$ is predicted probability. MSE loss is defined as:

$$MSE = \frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 \tag{3}$$

IV. Results for baselines

A. Training Details

Both U-Net models with CE and MSE loss are trained using stochastic gradient descent (SGD) optimizer. Weights are randomly initialized and updated with the learning rate set by 0.05, momentum parameter set by 0.9, and weight decay set by 0.001. Learning rate is reduced by a factor of 0.05 every ten epochs. During training, image patches are augmented using randomly flip horizontal and flip vertical.

TABLE I: Results comparison.

Method	Precision	Recall	F1 score	OA
U-Net + CE Loss	82.97	85.67	84.30	91.48
U-Net + MSE Loss	83.39	87.67	85.48	92.04

B. Evaluation Metrics

F1-score and Overal Accuracy (OA) are used as evaluation metric, and is defined as follows:

$$precision = \frac{tp}{tp + fp} \tag{4}$$

$$recall = \frac{tp}{tp + fn} \tag{5}$$

$$F1 = 2 \times \frac{precision \times recall}{precision + recall} \tag{6}$$

$$OA = \frac{tp + tn}{tp + fp + tn + fn} \tag{7}$$

where tp is the number of true positives, tn is the number of true negatives, fp is the number of false positives, and fn the number of false negatives.

C. Experimental Results

We compare U-Net models with CE and MSE loss. Quantitative comparisons are summarized in Table I. Both CNN models achieved comparative results. Model trained with MSE loss is slightly better than CE loss with F1 score of 85.48 and OA score of 92.04.

We give in Fig. 6 the final building extraction results for all models in some test images. Most of building rooftops can be mapped by both models trained with CE and MSE loss. Although the difference in mapping accuracy is insignificance, the model trained with MSE loss is much better than CE loss in term of detection rate. Besides, it's interesting to see that, both models are able to distinguish between building rooftop and side-wall and are able to work with degraded quality image (see the first and third row of Fig. 6).

V. Conclusions

In this study, we introduce a new dataset dedicated to building rooftop extraction from open-source Google Earth imagery. The buildings in this dataset have numerous types of rooftop with various shape and size. Besides, it's the first dataset to tackle the rooftop extraction within very dense building area. Besides, we provide some baselines using U-Net model in which different loss functions were evaluated. The experiment results showed that the models trained on these data are able to detect building rooftops with comparable accuracy and recall rate regardless of the image quality. We believe this dataset will contribute to the diversity of aerial dataset for building rooftop and building footprint extraction. Our future work would focus on the extraction of individual buildings from image.

Acknowledgment

This work has been supported by Vietnam National University Hanoi (VNU), under Project No. QG.18.36.

References

[1] Yang, H. L., Yuan, J., Lunga, D., Laverdiere, M., Rose, A., & Bhaduri, B. (2018). Building Extraction at Scale using Convolutional Neural Network: Mapping of the United States. Retrieved from http://arxiv.org/abs/1805.08946

[2] International Society for Photogrammetry and Remote Sensing. (n.d.). 2D Semantic Labeling - Vaihingen data. Retrieved from http://www2.isprs.org/commissions/comm3/wg4/2d-sem-label-vaihingen.html

[3] International Society for Photogrammetry and Remote Sensing. (n.d.). 2D Semantic Labeling Contest - Potsdam. Retrieved from http://www2.isprs.org/commissions/comm3/wg4/2d-sem-label-potsdam.html

[4] SpaceNet. (n.d.). SpaceNet Challenge. Retrieved from https://spacenetchallenge.github.io/datasets/datasetHomePage.html

[5] Microsoft. (n.d.). US Building Footprints. Retrieved from https://github.com/microsoft/USBuildingFootprints

[6] Guirado, E., Tabik, S., Alcaraz-Segura, D., Cabello, J., & Herrera, F. (2017). Deep-Learning Convolutional Neural Networks for scattered shrub detection with Google Earth Imagery, (November). doi:10.3390/rs9121220.

[7] Luu, V. H., Dinh, V. K., Luong, N. H. H., Bui, Q. H., & Nguyen, T. N. T. (2019). Improving the Bag-of-Words model with Spatial Pyramid matching using data augmentation for fine-grained arbitrary-oriented ship classification. Remote Sensing Letters, 10(9), 826834. doi:10.1080/2150704X.2019.1616123

[8] Guirado, E., Tabik, S., Alcaraz-Segura, D., Cabello, J., & Herrera, F. (2017). Deep-Learning Convolutional Neural Networks for scattered shrub detection with Google Earth Imagery, (November). doi:10.3390/rs9121220.

[9] Google. (n.d.). Google Maps & Google Earth GeoGuidelines. Retrieved from https://www.google.com/permissions/geoguidelines/.

[10] Hanoi Promotion Agency (2017). Retrieved from: http://www.hpa.hanoi.gov.vn/dau-tu/thong-tin-dau-tu/ha-noi-va-nhung-con-so/quy-mo-dan-so-va-dien-tich-30-quan-huyen-cua-ha-noi-a2144. (In Vietnamese)

[11] GENERAL STATISTICS OFFICE of VIET NAM (2018). Population and Employment.

[12] Long, J., Shelhamer, E., & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 34313440). IEEE. doi:10.1109/CVPR.2015.7298965.

[13] Noh, H., Hong, S., & Han, B. (2015). Learning Deconvolution Network for Semantic Segmentation. Retrieved from http://arxiv.org/abs/1505.04366.

[14] Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 9351, 234241. doi:10.1007/978-3-319-24574-4_28.

[15] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. In 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 770778). IEEE. doi:10.1109/CVPR.2016.90

[16] Nair, V., & Hinton, G. E. (2010). Rectified Linear Units Improve Restricted Boltzmann Machines. In Proceedings of the 27th International Conference on International Conference on Machine Learning (pp. 807814). USA: Omnipress.

(a) Google Earth Image (b) Ground truth (c) U-Net + CE loss (d) U-Net + MSE loss

Fig. 6: Result visualization