FPT UNIVERSITY

Building Detection on Satellite Images

Dam Hong Phuc, Nguyen Hai Dang, Nguyen Huu Thai Minh, Phan Duy Hung

FPT University, Hanoi, Vietnam

phucdhhe181106@fpt.edu.vn; dangnhhe180564@fpt.edu.vn;
minhnhthe182044@fpt.edu.vn; hungpd2@fe.edu.vn

Abstract. This paper presents a comprehensive study on the automatic detection of buildings in satellite images using deep learning techniques. The focus is on developing and evaluating Convolutional Neural Network (CNN) architectures, specifically U-Net and LinkNet, for the task of building detection and segmentation. The study leverages high-resolution satellite imagery from Hanoi, Vietnam, and employs data augmentation techniques to enhance the training dataset. Experimental results demonstrate the effectiveness of the proposed models, achieving high accuracy and robustness in building detection. This research highlights the potential of deep learning models in urban planning, state cadastral inspections, and disaster management in Vietnam.

Keywords. Building Detection,, Satellite images, U-net model, Linknet model.

1. Introduction

In recent years, the task of automatically identifying buildings from satellite imagery has become a significant and complex challenge due to the rapid expansion of urban areas. Extracting building information from these images is computationally intensive but crucial for various applications, including city planning, infrastructure development, urban mapping management, marketing, and population estimation. Moreover, it is invaluable during emergency response efforts, such as natural disaster relief, and in military operations. [1] The success of these applications heavily depends on the reliable and consistent extraction of information from satellite images. However, relying solely on manual efforts for building detection in satellite images is both time-consuming and inefficient. As a result, there is a growing need to develop automated building detection techniques to address these challenges.

On the other hand, when extracting urban information features from aerial photos, spatial resolution is thought to be more essential than spectral resolution. This is why images captured by commercial satellites and unmanned aerial vehicles (UAVs), which are also known for remote sensing technologies, are so popular nowadays [2]. It delivers high-resolution, accurate satellite images of huge regions on a regular basis and at different times, helping to enhance environmental research and studies, sectors, natural resource exploration, city planning, agricultural crop management, and other key fields [3]. As a result, urban objects, especially buildings, can be clearly seen in satellite images, which facilitates the construction of advanced machine learning models.

In the field of machine learning, the problem of image segmentation is usually reformulated as a classification on a pixel-wise level. However, it is a time-consuming process, which is subjected to human errors due to monotonous manual work. Therefore, the great interest aims at automatic image segmentation. This article presents developed convolutional neural networks (CNNs). The structure of these models is parallel and fits the architecture of graphics processing units (GPUs)

which consists of thousands of cores to perform several tasks simultaneously. Although CNNs have been known for decades, only recent achievements in the development of high performance computers with GPUs have allowed researchers to launch CNNs, which have millions of parameters. Currently, in tasks of computer vision CNNs excel traditional machine learning algorithms and even some experts in speed and quality [4].

This article is structured into six distinct sections. Initially, it addresses the challenge of segmenting satellite images and elucidates the benefits of employing Convolutional Neural Networks (CNNs) in comparison to conventional machine learning techniques. The second section reviews related literature on image segmentation. The third section provides insights and methodologies tailored to the characteristics of Vietnam facing challenges in urban planning that differ from developed nations. The fourth section discusses the dataset of satellite images. The fifth section details the developed CNN architectures for building detection in aerial imagery and highlights specific aspects of model training. The sixth section presents the results of numerical experiments conducted with the developed models. Finally, the conclusion summarizes the research findings.

2. Related Work

A literature review is undertaken to provide comprehensive details on various studies and their contributions concerning the existing methods employed for the building detection in satellite images. This section also examines and determines the most appropriate deep learning-based approach for building detection and segmentation in satellite imagery.

Artificial intelligence methodologies for pattern recognition and computer vision, including K-means, neural networks (NN), Support Vector Machines (SVM), and Random Forests (RF), have demonstrated their efficacy in classifying remote sensing images. In 2006, Hinton et al. [5] introduced the concept of deep learning. Unlike traditional machine learning techniques such as NN, SVM, and RF, deep learning models focus on automatic feature extraction from large datasets. These models learn to identify and extract distinctive features during the training process, eliminating the need for manual feature design.

In the field of computer vision, deep learning, particularly Convolutional Neural Networks (CNNs), has been successfully employed for tasks such as image classification, object detection, and scene interpretation. [6] One of the most effective algorithms for image segmentation is based on fully convolutional networks (FCNs). The core concept of FCNs involves using a fully connected layer combined with a convolution layer at the end, while other layers focus on extracting essential features from the input data. This design enables the application of FCNs to image segmentation tasks . FCNs have since evolved into what are now known as Feature Pyramid Networks (FPNs). FPNs utilize a pyramid architecture within CNNs to construct complex features at various scales through a bottom-up pathway, subsequently extracting the necessary features via a top-down process. This architecture has demonstrated significant improvements in numerous applications, particularly in object detection within satellite images.

The FCN method was further extended to the U-Net architecture, which was initially introduced for biomedical image segmentation in [7] and was later adapted for pixel-wise classification of satellite images [8]. U-Net is a specialized type of FCN that integrates low-level and high-level feature maps to enhance object localization.

In paper [9], the authors introduce LinkNet, a specialized CNN architecture featuring an encoder and a decoder, similar to U-Net. LinkNet efficiently shares information learned by the encoder with the decoder after each downsampling block. In certain scenarios, this method surpasses Fully Convolutional Networks (FCNs) in the decoder phase.

3. Contribution

While research on Land Use and Land Cover (LULC) exists for various developed nations across different continents (Europe [10], Africa [11], Oceania [12],...), there remains a significant gap in

studies focusing on developing countries like Vietnam. This paper aims to bridge that gap by providing valuable insights and methodologies tailored to the unique characteristics found in Vietnam.

Urban planning and housing development significantly differ between developed countries such as the United Kingdom, the United States, and Australia, and developing countries like Vietnam. In developed nations, urban planning is often characterized by comprehensive and long-term strategies, which emphasize sustainability, infrastructure efficiency, and quality of life. These countries typically have stringent zoning laws, advanced public transportation systems, and well-maintained public spaces. In contrast, urban planning in developing countries like Vietnam faces numerous challenges that stem from rapid urbanization, population growth, and limited resources. In Vietnam, urban development is often reactive rather than proactive, leading to issues such as overcrowding, inadequate infrastructure, and informal settlements. Consequently, the disparity in urban planning approaches reflects broader socio-economic differences, where developed countries benefit from greater financial resources, technological advancements, and institutional frameworks that support sustainable urban growth, whereas developing countries must navigate the complexities of rapid development with more constrained capacities.

The findings and techniques presented here can serve as a foundational resource for Vietnamese government and enterprises, facilitating more informed decision-making in urban planning, infrastructure development, disaster management and detecting violated cadastral regulations. By adapting sophisticated semantic segmentation models such as U-Net and LinkNet, this research underscores the potential for high-accuracy building detection even in areas with less technological resources.

4. Data Preparation

4.1. Data description

The dataset utilized for this study encompasses an area of 10,720 hectares across four districts in Hanoi: Ha Dong, Cau Giay, Hoang Mai and Hoan Kiem [Figure 1]. These districts have been designated for future land use adjustments and supplementary planning by the People's Committee of Hanoi, which includes major projects such as land clearance, ecological zones construction. The dataset consists of 77 images with the size of 1096x768 pixels, at a viewing height of 2000 feet, captured from Google Earth Pro (a computer program that renders a 3D representation of Earth based primarily on satellite imagery) [Figure 2]. The dataset was randomly divided into two parts: 80% for the training set and 20% for the validation set, facilitating effective model training and performance evaluation. In particular, the images were acquired between October 7, 2023, and December 23, 2023, providing up-to-date and relevant data for the study.

Areas	**# train**	**# test**	**Total**
Ha Dong	152	38	190
Cau Giay	128	32	160
Hoan Kiem	102	26	128
Hoang Mai	234	58	292
Total	616	154	770

Table 1. Review satellite images datasets

Using huge-dimension images to train the model will cause many obstacles such as requiring more computational resources, leading to the need for more GPU memory and longer training time. Because of limited resources, we used two methods to overcome the problem. On the one hand, we cut the image into 4 smaller parts of size 548*384, which not only reduces the memory and computational resource requirements but also helps increase the

training set size (each cut part of the image can be treated as a separate training sample) and does not lose detailed information in each small cut. On the other hand, resizing the image to a size similar to the previous one, which is easy and quick to do during data preprocessing. Furthermore, to enhance the sample size, each image was augmented by a 180-degree rotation, resulting in a total of 770 images. [Table 1]. This augmentation ensures a more robust dataset for training and validation purposes.

Figure 1. Study Area in 4 Districts of Hanoi, respectively Ha Dong, Cau Giay, Hoang Mai, Hoan Kiem

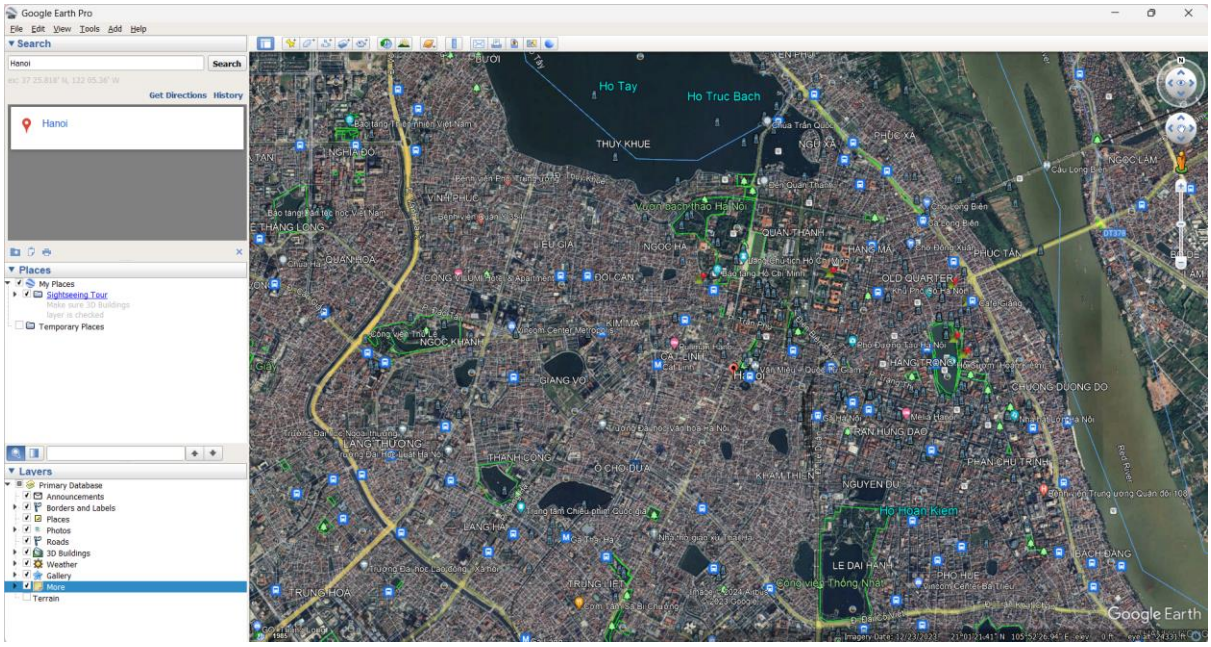

Figure 2. Google Earth Pro user interface (free version)

4.2. Data labeling

The next crucial step in our data preparation pipeline involved labeling the collected images. We utilized the cvat.ai [Figure 3] platform for this task, which facilitated the annotation process. However, labeling proved to be labor-intensive, requiring manual intervention to accurately delineate boundaries around classes within the images. In order to train and validate the models, all the image patches from Google Earth Pro are cropped and resized respectively. They are fully annotated as ground truth with only one class named "building" which is identified for each pixel of the image patch. To ensure the robustness of the dataset, extensive quality checks were performed to verify the accuracy of the annotations. Examples of images from the dataset and generated masks are shown in Figure 4.

Figure 3. Labeling building class with cvat.ai

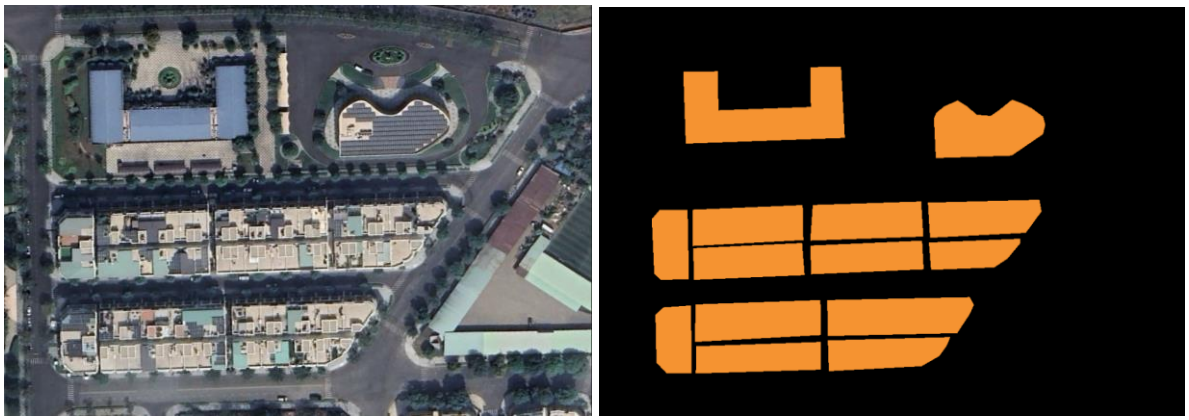

Figure 4. Image of part of Ha Dong and its generated mask

5. Methodology

5.1. Deep Learning Model: U-Net and Linknet

In the new millennium, deep neural networks (NNs) have finally garnered widespread attention, primarily due to their superior performance in numerous critical applications when compared to alternative machine learning methods such as kernel machines [13][14]. Among these successful neural networks, the UNet architecture stands out in the realm of image segmentation. The architectural innovations of UNet are twofold: it integrates upsampling and downsampling layers in equal measure, and employs skip connections between corresponding convolutional and deconvolutional layers. This approach results in the concatenation of features from both the contracting and expanding paths. [7]

By incorporating these features, UNet processes the entire image in a single forward pass, yielding a segmentation map of the input image. The architecture's design resembles a 'U' shape, as depicted in Figure 5, and consists of three primary sections: the contraction section, the bottleneck, and the expansion section. The contraction section comprises several contraction blocks, each of which applies two 3x3 convolutional layers followed by a 2x2 max pooling operation. The number of kernels or feature maps doubles after each block, enabling the architecture to effectively learn complex structures. The bottleneck layer serves as an intermediary between the contraction and expansion layers, utilizing two 3x3 convolutional layers followed by a 2x2 up-convolution layer.

Similar to the U-Net architecture, LinkNet comprises both an encoder and a decoder. As illustrated in Figure 6, each of these subnets is composed of four blocks. Each encoder block includes four convolutional layers, two merging operations, and a max-pooling operation with a 2×2 filter and a stride of 2. The corresponding decoder blocks mirror this structure but replace the merging and max-pooling operations with upsampling operations using a 2×2 filter. Before the feature map from the input data is fed into the first encoder block, it undergoes a series of operations: two batch normalization steps, a ReLU activation function, a convolution with a 2×2 filter, and a max-pooling operation with a 2×2 filter. After completing the final decoder block, the network performs an upsampling operation with a 2×2 filter, followed by two batch normalization steps, a ReLU activation function, and a convolutional layer with a 2×2 filter. The LinkNet architecture and the encoder scheme are depicted in Figure 6.

The U-Net and LinkNet models represent two distinct influential architectures in the realm of convolutional neural networks (CNNs) for image segmentation tasks. U-Net, employs a symmetric encoder-decoder structure that facilitates precise localization by combining low-level and high-level features through skip connections. This architecture allows the U-Net to effectively capture fine-grained details and structural information, making it particularly adept at handling complex segmentation challenges. In contrast, LinkNet, while also utilizing an encoder-decoder framework, introduces a more streamlined approach by linking encoder and decoder layers directly, thereby reducing the computational complexity and memory footprint. This design choice not only accelerates the training process but also retains critical spatial information, albeit with potentially less granularity compared to U-Net. Furthermore, LinkNet's efficiency in sharing learned features across layers enhances its performance in scenarios where computational resources are limited. Consequently, while U-Net excels in applications requiring high precision and detailed segmentation, LinkNet offers a balanced trade-off between accuracy and computational efficiency, making it suitable for real-time applications or environments with constrained resources.

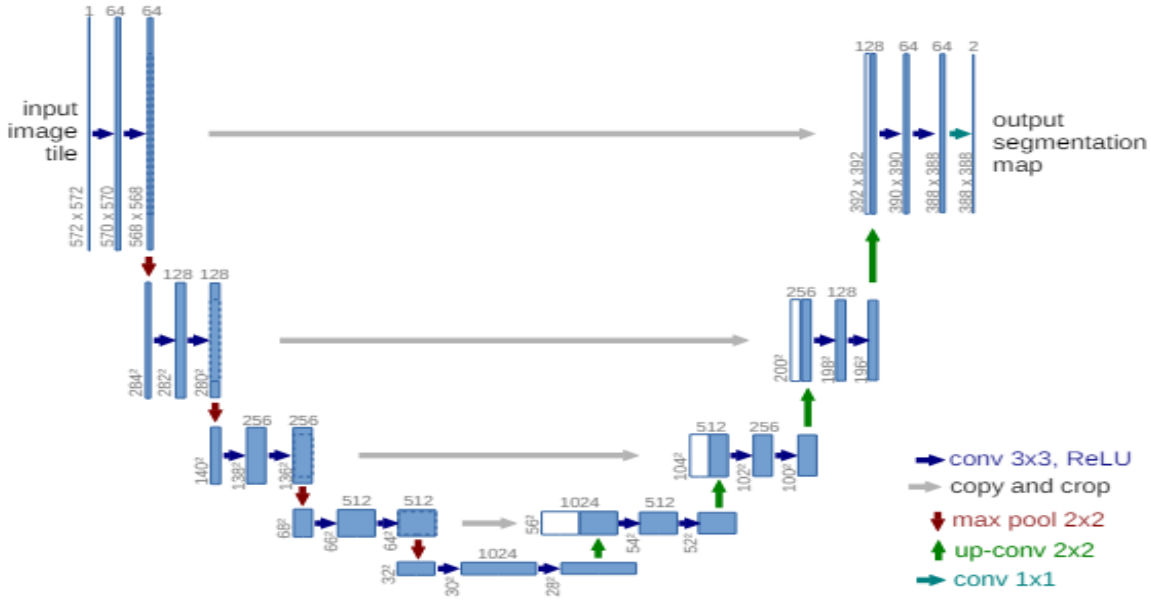

Figure 5. U-net architecture

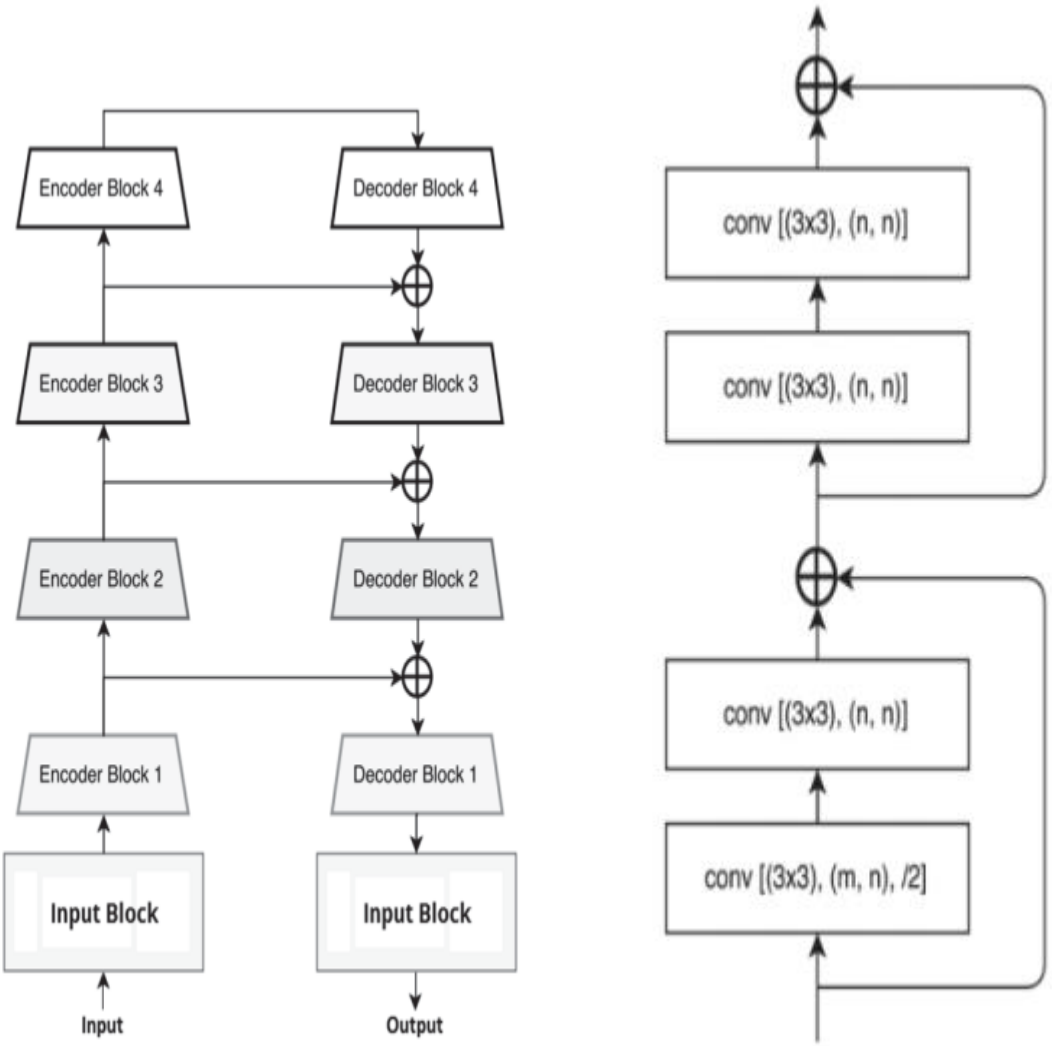

Figure 6. Linknet Encoder and Decoder architectures

5.2. Model Training and Validation

The U-Net model will be trained on the labeled ground truth data. The training process involves optimizing the model parameters to minimize the difference between predicted and actual land cover classes. The validation set will be used to monitor model performance during training and prevent overfitting.

All developed networks were created using Keras library with Tensorflow framework as a backend. Keras is an open source library written in Python. This library contains many implementations of structural blocks of neural networks such as layers, activation functions, optimizers, and ready tools for preprocessing of images and text data. Furthermore, this library allows us to train and test networks on GPU. On the other hand, the resources to train and test the model will be provided by Google Colab. Colab supplies free access to their in-house TPU v2, each instance of free runtime may last up to 3 hours and 20 minutes.

A rigorous accuracy and loss assessment will be conducted to evaluate the performance of the semantic segmentation models. This will involve comparing the predicted land cover map with the independent testing set of ground truth data

6. Result and discussion

In this project, we mainly used two types of metrics: "Accuracy" and "Binary Cross-Entropy Loss" to evaluate the results of models. Firstly, Accuracy (A) was calculated with the following formula:

$$A = \frac{P}{N'}$$

where P is a quantity of right classified objects and ŏN' is the count of objects for classification [15]. The results of numerical experiments on the test set are cited in Table 2.

Model	Accuracy (A)
U-Net	92.25%
LinkNet	91.99%

Table 2. TESTING RESULTS OF U-NET AND LINKNET IN TERM OF ACCURACY

Secondly, Binary Cross-Entropy Loss is a popular choice for image segmentation problems, where each pixel is classified into one of two classes: building or non-building. The formula for Binary Cross-Entropy Loss as follows:

$$\text{Loss} = -\frac{1}{N}\sum_{i=1}^{N} [y_i log(p_i) + (1 - y_i) log(1 - p_i)]$$

In detail, N is the number of pixels, y_i is the actual label of the i^{th} pixel, p_i is the model's prediction probability for the th pixel. Dependencies of Binary Cross-Entropy Loss from 100 epochs for each developed algorithm on both training and validation sets are shown in Figure 7. Both U-Net and LinkNet architectures exhibit a common trend where the training loss decreases rapidly and stabilizes at a low value. However, the validation loss for both models does not decrease to the same extent and shows fluctuations, indicating potential overfitting issues. The initial rapid decline in validation loss followed by fluctuations suggests that both models learn effectively during the early epochs but may struggle with generalization to unseen data as training progresses.

For the visual inspection, fourth image subsets from different areas of Hanoi are chosen. The images present a series of comparative results between original satellite images, ground truth masks, and

predicted segmentation masks generated by the U-Net and Linknet model. Overall, the model effectively captures the general shape and location of buildings, though minor discrepancies in boundary precision and smaller structure identification suggest areas for potential improvement. Even if the buildings are very small, the method performs very well in segmenting the buildings as shown in the last subset of images in Figure 8.

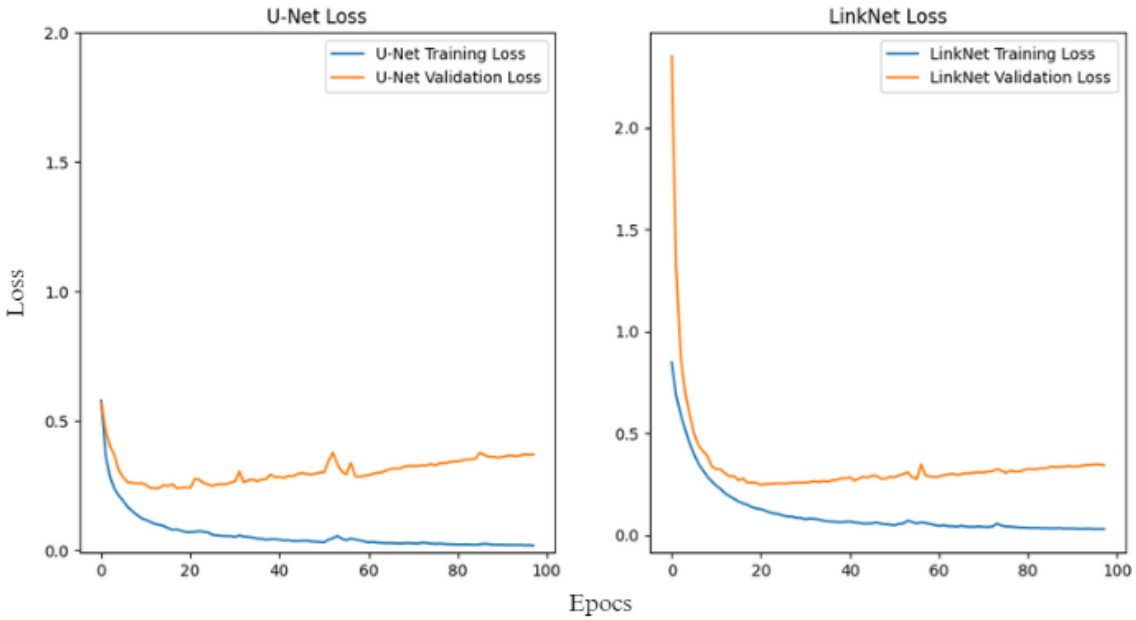

Figure 7. U-Net Loss and LinkNet Loss across epochs

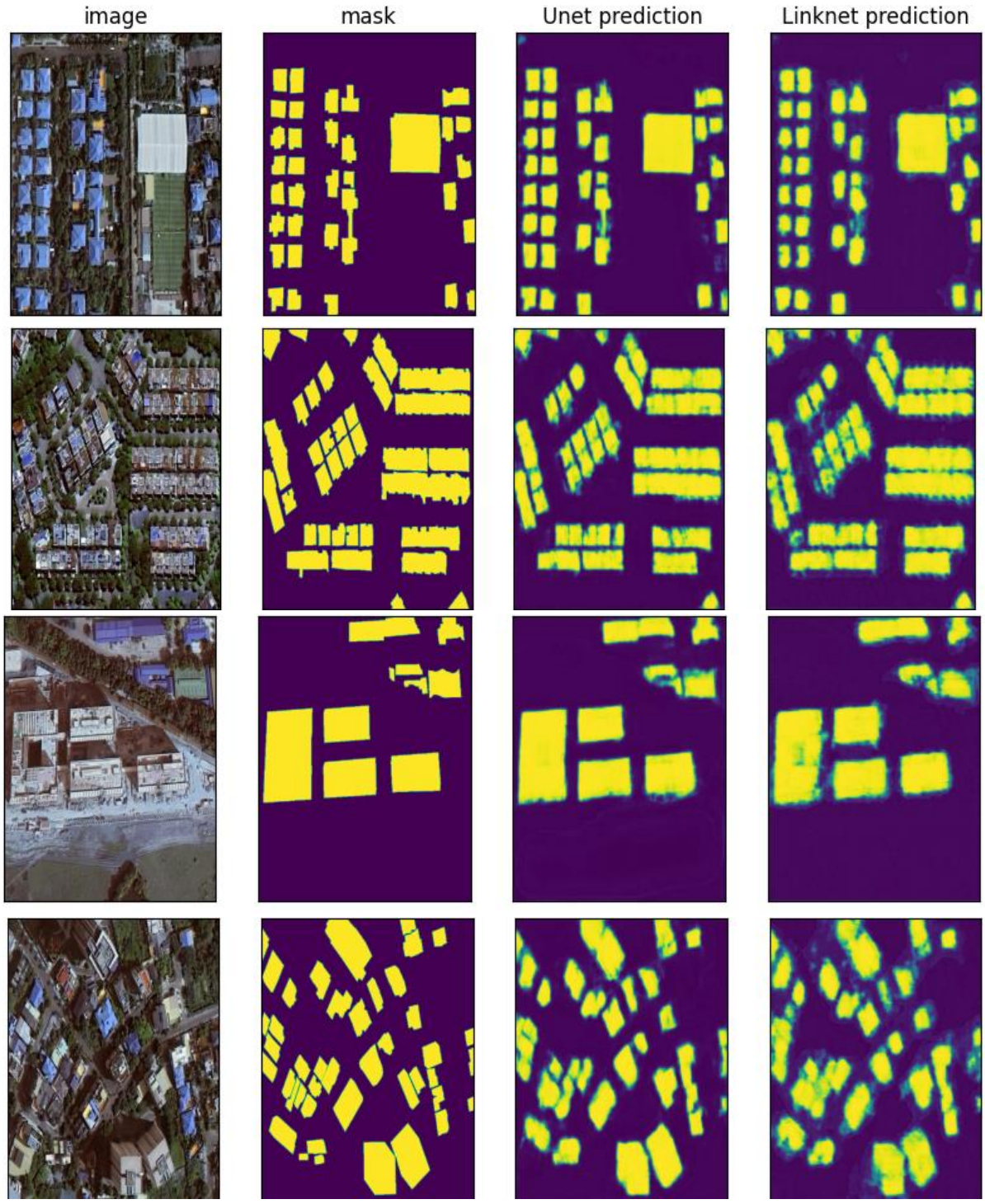

Figure 8. Comparison between ground truth, U-Net and Linknet prediction

During the implementation of our project, we encountered several challenges. One of the primary issues was the lack of diverse specialized satellite data. The available dataset was limited in terms of building types, shooting angles, and lighting conditions, which significantly impacted the model's ability to generalize well on new data. The dataset's relatively small size further exacerbated this problem, as it limited the model's exposure to varied scenarios, thereby reducing its robustness and accuracy in real-world applications. Another critical challenge was the limitation in computational resources. While Google Colab provides a valuable platform for model training, the free tier's runtime and RAM capacity constraints posed significant limitations. Despite these challenges, the project

demonstrates substantial potential for further development.

The potential applications of this project are vast and diverse, spanning multiple fields and providing substantial benefits. It proves significant for city planning, state cadastral inspections, and the provision of municipal services. Moreover, the project plays a crucial role in monitoring the rate of urbanization, offering insights into the dynamic changes occurring in urban landscapes. These insights are essential for managing urban growth and ensuring sustainable development. Telecommunication companies also stand to benefit significantly from this project. By providing precise data on building locations and other geographical features, the project aids in making informed decisions regarding the placement of transmitter stations.. The project's outputs can thus contribute to enhancing communication networks, leading to better service quality and coverage.

7. Conclusion

In conclusion, this study successfully demonstrates the application of deep learning models, specifically U-Net and LinkNet, for the automatic detection and segmentation of buildings in satellite images. The research addresses a significant gap in the existing literature by focusing on the unique characteristics of a developing country like Vietnam. Using the metrics of similarity between expert markup and predicted masks, it was shown that U-Net got better results compared with LinkNet. In detail, for U-Net and Linknet the accuracy is equal to 92,25% and 91,99% respectively. Moreover, the study sets the stage for future advancements by suggesting the integration of multi-spectral data and more complex architectures to enhance model robustness and versatility. The deployment of advanced deep learning models are proven to efficiently yield substantial improvements in satellite image analysis, contributing to more informed decision-making and efficient resource management in Vietnam.

References

[1] Parvaneh Saeedi and Harold Zwick. "*Automatic building detection in aerial and satellite images*". In 2008 10th International Conference on Control, Automation, Robotics and Vision, pages 623–629, Hanoi, Vietnam, December **2008**.

[2] Neupane, B.; Horanont, T.; Aryal, J. "*Deep Learning-Based Semantic Segmentation of Urban Features in Satellite Images: A Review and Meta-Analysis. Remote Sens.*" **2021**, 1, 808.

[3] Jamil, Abdlhamed & Al-Sharif, Abdulmohsen & Al-Thubaiti, Amer. **(2020)**. "*Classifications of Satellite Imagery for Identifying Urban Area Structures. Advances in Remote Sensing.*" 09. 12-32.

[4] Y. Goodfellow, Y. Bengio, A. Courville, "*Deep Learning*", The MIT Press, **2016**, 80.

[5] Geoffrey Hinton, Simon Osindero, Max Welling, Yee-Whye Teh, *Unsupervised Discovery of Nonlinear Structure Using Contrastive Backpropagation*, **2010.**

[6] Evan Shelhamer , Jonathan Long , and Trevor Darrell, Member, IEEE, *Fully Convolutional Networks for Semantic Segmentation*, **2016**.

[7] O. Ronneberger, P. Fischer, T. Brox, "*U-Net: Convolutional Networks for Biomedical Image Segmentation*", Medical Image Computing and Computer-Assisted Intervention (MICCAI), Springer, LNCS, vol. 9351, **2015**, pp. 234–341.

[8] G. Chhor, C. Bartolome Aramburu, I. Bougdal-Lambert, "*Satellite Image Segmentation for Building Detection using U-net*".

[9] A. Chaurasia, E. Culurciello, "*LinkNet: Exploiting Encoder Representations for Efficient Semantic Segmentation*" .

[10] Witjes, Martijn, Leandro Leal Parente, Chris J. van Diemen, Tomislav Hengl, Martin Landa, Lukáš Brodský, Lena Halounová, J. Krizan, Luka Antonić, Codrina Maria Ilie, Vasile Craciunescu, Milorad Kilibarda, Ognjen Antonijević and Luka Glušica. "*A spatiotemporal ensemble machine learning*

framework for generating land use/land cover time-series maps for Europe (2000–2019) based on LUCAS, CORINE and GLAD Landsat." PeerJ 10 (**2021**).

[11] Megahed, Y.; Cabral, P.; Silva, J.; Caetano, M. *Land Cover Mapping Analysis and Urban Growth Modelling Using Remote Sensing Techniques in Greater Cairo Region—Egypt*. ISPRS Int. J. Geo-Inf. **2015**, 4, 1750-1769.

[12] Aryal, J., Sitaula, C. & Frery, A.C. *Land use and land cover (LULC) performance modeling using machine learning algorithms: a case study of the city of Melbourne*, Australia. Sci Rep 13, 13510 (**2023**).

[13] Vapnik V, Guyon I, Hastie T (**1995**) *Support Vector Machines*. Mach Learn 20(3):273–297.

[14] Schölkopf B, Smola A, Müller KR (**1998**) *Nonlinear component analysis as a kernel eigenvalue problem. Neural Compute* 10(5):1299–1319.

[15] J. Sanders, E. Kandrot, CUDA by Example: *An Introduction to General-Purpose GPU Programming, Addison-Wesley Professional*, **2010**, 320 p

Task Name	Priority	Owner	Start date	End date	Status	Issues
Find documents	High	Phuc, Minh, Dang	03/05	12/05	Finished	
Review related papers	Medium	Phuc, Minh, Dang	12/05	18/05	Finished	
Data Collecting	High	Phuc	18/05	25/05	Finished	Could not find data from specialized satellites, which are unavailable and expensive. However, Google Earth Pro supplies high-quality images.
Data Labeling	High	Phuc, Minh	29/05	22/06	Finished	Cvat.ai is time consuming because the server isn't capable of loading a large amount of data.
Evaluate potential method	Medium	Phuc, Dang, Minh	16/06	22/06	Finished	
Coding	High	Dang	22/06	13/07	Finished	
Compare results	Medium	Phuc, Dang	13/07	14/07	Finished	

Prepare slides and scripts for presentation	Medium	Minh	12/07	14/07	Finished	
Writing paper.	High	Phuc	10/05	14/07	Finished	
Writing appendix	Low	Phuc	14/07	14/07	Finished	