Current **A**pplied **S**cience and **T**echnology e-ISSN: 2586-9396

Curr. Appl. Sci. Technol. 2025, Vol. 25 (No. 1), e0260846

Research article

Comparative Analysis of Deep Learning Models for Building Extraction from High-resolution Satellite Imagery

Tachasit Chueprasert*, Akadej Udomchaiporn and Sarun Intagosum

Data Science and Computational Intelligence Laboratory (DSCI), Department of Computer Science, School of Science, King Mongkut's Institute of Technology Ladkrabang, Bangkok, Thailand

Received: 9 October 2023, Revised: 2 March 2024, Accepted: 27 June 2024, Published: 16 October 2024

Abstract

In this research, an approach to extract buildings from Google's satellite imagery was proposed. The performances of various deep learning models (U-Net, RIU-Net, U-Net++, Res-U-Net, and DeepLabV3+) on pre-processed datasets were compared. The models were trained using the similarity metrics of Intersection over Union (IoU) and Dice Similarity Coefficient (DSC). The best-performing models among the segmentation techniques were Res-U-Net and DeepLabV3+. Res-U-Net, an enhanced version of the traditional U-Net model that incorporates residual connections for improved feature propagation, achieved an F1 score of 85.43% when using the RGB dataset. Similarly, DeepLabV3+ also achieved high performance on the Enhanced RGB dataset, obtaining an F1 score of 85.18% after applying pre-processing techniques. This research highlights the significance of color as a dominant feature for accurate building extraction from satellite images. The findings contribute to improved methodologies for building identification, benefiting urban planning, and disaster management applications.

Keywords: building extraction; deep learning; image processing; satellite imagery; semantic segmentation

1. Introduction

The extraction of building areas from satellite imagery has wide-ranging applications in both public and private sectors. It serves diverse purposes, including monitoring residential expansion, estimating population based on growth rates and housing sizes. In some countries, population censuses are conducted every decade; however, these surveys had to be canceled or postponed due to the inability to carry out fieldwork caused by the COVID-19 pandemic. This unprecedented situation has posed challenges in gathering accurate population data. Recently, satellite imagery technology has undergone significant advancements (McCarthy & Halls, 2014), resulting in the improved capture, recording, and storage of high-quality images of Earth's surface. These images span various time periods

*Corresponding author: E-mail: 61605035@kmitl.ac.th
https://doi.org/10.55003/cast.2024.260846

and cover extensive global regions, regarding permissible boundaries for image acquisition. Access to such high-resolution (Boyle et al., 2014), satellite imagery is facilitated by multiple service providers. For this study, Google's imagery, for which permission has been granted for academic and research purposes, was used as the dataset.

The elements captured in satellite imagery vary based on the location and geography of the area. These elements include buildings, forests, barren lands, agricultural areas, roads, and others. There are certain challenges associated with satellite imagery such as differences in image quality that present due to varying time intervals among satellite orbits, which result in inconsistent color representation. Additionally, the availability of data has led to variations in satellite imagery data recorded over different years for the same area. Furthermore, the presence of cloud cover can obscure the clarity of the surface. Consequently, manually extracting specific locations and structures of buildings is time-consuming, costly, and prone to human error. Therefore, this research presents an automatic method for extracting building areas from satellite imagery.

This research study utilized satellite imagery from various areas in Thailand that faced the COVID-19 pandemic. The pandemic resulted in the postponement of the 2020 census. The study specifically focused on regions distant from the capital city that included a lot of rural landscapes. During the examination of multiple areas, it was observed that buildings exhibited variations in architectural design, roof color, size, and materials. Additionally, the presence of tall trees close to the buildings often obscured roof areas and resulted in indistinct shapes. However, the factors of color and light intensity generally facilitated clear differentiation between the buildings and the surrounding natural environment.

The primary objective of this research was to leverage satellite imagery when constructing a comprehensive dataset that facilitated the extraction of building boundaries. The methodology included a strong emphasis on a comparative analytical approach, beginning with the division of the image dataset into two categories: the original image dataset and the pre-processed image dataset. Subsequently, the data were fed into various deep learning models encompassing the prominent architectures: U-Net, RIU-Net, U-Net++, Res-U-Net, and DeepLabV3Plus. In addition, backbone experiments were conducted using ResNet-50, ResNet-101, and ResNet-152, which further enhanced the analysis and performance of the models.

With a focus on methodologies for building identification in satellite imagery, preprocessing techniques to enhance segmentation accuracy were developed and refined. A significant aspect of this work was the application of these methodologies to a unique dataset analyzed alongside four pre-processed variations, to discern the most effective configurations for precise building detection. The dataset, consisting of high-resolution satellite imagery from a specific region, was pivotal for evaluating the impact of various preprocessing strategies on segmentation performance. While new models were not introduced, the emphasis on methodological enhancement through strategic preprocessing represented a substantial step towards more reliable and effective segmentation outcomes. The creation and use of this meticulously labeled dataset highlighted the study's contribution to advancing image segmentation techniques.

The application of various techniques for extracting building areas from satellite imagery has been extensively studied and explored. These techniques encompass a range of approaches, including image processing methods, the utilization of deep learning algorithms, and the integration of image processing and deep learning frameworks. Therefore, the following section presents a thorough review of relevant literature concerned with these processes.

Process and techniques for extracting buildings from satellite imagery using mathematical morphology were presented by Gavankar and Ghosh (2018). Although the entire process did not involve the use of deep learning, the researchers highlighted the interesting aspects of the pre-processing and post-processing steps and the application of morphological operations to modify the basic characteristics of building structures. Specifically, the proposed technique incorporated a morphological top-hat filter and the K-means algorithm to extract buildings with bright and dark rooftops. By separately extracting segments of buildings with different rooftop characteristics and subsequently merging them, the final output consisted of accurately extracted building segments. In addition, post-processing steps to mitigate false detections were proposed to set up an appropriate threshold range for building area based on the characteristics of the study area. However, it was noted that the specific threshold range might have varied depending on the resolution of the satellite imagery and the urban scenario under investigation.

Previous studies have identified limitations in various areas, particularly when dealing with satellite images that contain dense natural components. These challenges hinder the efficiency of image processing methods for building extraction. An example of research emphasizing these findings is the study by Daranagama & Witayangkurn (2021). The study presented a methodology for extracting building footprints from high-resolution aerial and UAV imagery. The approach incorporated data pre-processing techniques and dataset merging to enhance accuracy and usability. Modified U-Net architecture and specific pre-processing techniques were applied to their dataset. A logarithmic correction image enhancing algorithm, which was applied to pre-processing steps, significantly improved building detection accuracy for aerial images, while the Contrast Limited Adaptive Histogram Equalization (CLAHE) algorithm proved effective for enhancing UAV images.

Recent research studies have explored the application of U-Net for building extraction from aerial satellite images. Ivanovsky et al. (2019) adopted the novel approach of categorizing the dataset into “Buildings” and “Not Buildings” to train the model. Results were compared with LinkNet, revealing U-Net's superior performance in terms of output quality. The Sorensen-Dice Coefficient was used to evaluate the effectiveness of the experiment, and the results indicated a coefficient value of 77%. Additionally, U-Net was widely applied in the research for building extraction from satellite imagery, with notable success compared to LinkNet.

Chhor et al. (2017) proposed the application of a Convolutional Neural Network (CNN) based on U-Net architecture, originally designed for biomedical image segmentation, to extract buildings from satellite imagery. The focal aspect was the implementation of data augmentation and the application of evaluation metrics, specifically the Jaccard Index and Dice Coefficient, which yielded scores of 59% and 74%, respectively. However, the study did not employ any post-processing techniques.

Apart from the recent applications of U-Net for automatic building extraction, new concepts regarding the development of the model's architecture to enhance feature extraction efficiency in the encoder part of the original model were explored. Sariturk & Seker (2022) presented the utilization of Residual-Inception. The research findings indicated that RIU-Net achieved the best results with the Inria dataset compared to several other models such as U-Net, Residual U-Net, and Trans U-Net. Therefore, how the RIU-Net concept performed with the provided data and methodology became worthy of study.

The concept of improving the encoder in building extraction from satellite images has been explored in various research studies. One engaging approach is the integration of Residual Network (ResNet) architectures into the U-Net framework, named Res-U-Net. This modification incorporates residual connections to enhance information flow and gradient propagation during training. In the research study proposed by Xu et al. (2018),

Res-U-Net incorporating ResNet and guided filters in post-processing achieved high accuracy with overall accuracy (OA) values of 97.71% and 96.91% on the Vaihingen and Potsdam datasets, respectively. Similarly, in the study proposed by Alsabhan et al. (2022), U-Net with ResNet as the backbone architectures were combined to help extract buildings and the results indicated improved performance compared to the original U-Net, as evaluated by Intersection over Union (IoU) and Dice Coefficients. Next, a research study proposed by Alsabhan & Alotaiby (2022) demonstrated that using ResNet50 and ResNet152 as backbone architectures led to better results than using the original CNN.

The application of CNN has also been discovered in another interesting model, namely DeepLabv3+. Several research studies utilized DeepLabv3+ for building extraction processes, using both aerial photography and satellite imagery. These studies achieved results that were comparable to U-Net with encoder modifications. For instance, Aslantaş et al. (2021) presented a technique for building extraction using DeepLabv3+, which employed a similar architecture to the original model. The experiments focused on tuning hyperparameters to achieve the best results for the Wuhan University (WHU) aerial building dataset. The results of this research indicated a high accuracy in prediction, with an IoU of 98.23%. Additionally, the study proposed by Han et al. (2022) aimed to enhance the performance of building extraction by addressing issues such as slow extraction speed and incomplete edge segmentation. This was achieved through modifications to the backbone and the space pyramid pooling module. In the experiments, the proposed method was compared with other prediction models, including U-Net, SegNet, PSPNet, and DeepLabV3+, using two datasets: WHU and Massachusetts. The results revealed that the proposed method outperformed the others.

It is worth noting that U-Net and DeepLabV3+ have been often mentioned in research on deep learning models for semantic segmentation in building extraction. For example, in the research proposed by Bakirman et al. (2022), an improved U-Net++ architecture utilizing an SE-ResNeXt101 encoder pre-trained with ImageNet was invented. Their extensive experiments were conducted to compare various encoders and optimizers and the results revealed IoU accuracies of 75.39% and 92.53% on the Inria and Massachusetts datasets, respectively.

Considering the limitations of traditional image processing in accurately segmenting buildings from satellite images, particularly in varied rural and suburban landscapes, this study pivoted to deep learning models known for their segmentation capabilities. The choice of U-Net, RIU-Net, U-Net++, and Res-U-Net was informed by their success in detailed segmentation tasks, where precision was paramount. Each of these U-Net variants introduces enhancements to the original architecture, RIU-Net for rotational invariance, U-Net++ for nested, dense skip pathways, and Res-U-Net for incorporating residual learning to better capture the complex spatial relationships in satellite imagery. DeepLabV3+, distinct from the U-Net family, was included for its effectiveness in utilizing pre-trained models such as ResNet, which are pre-trained on ImageNet, offering a robust framework for multi-scale feature extraction. The adoption of models capable of integrating pre-trained weights was aimed at leveraging vast amounts of learned features from diverse visual domains, enhancing model performance even before fine-tuning on our specific dataset. This strategic selection, discussed later in this paper, underscored the foundation of our methodology, with its focus on addressing the nuanced challenges of satellite-based building segmentation.

2. Materials and Methods

A comprehensive methodology is presented, starting with data acquisition and labeling, data preparation techniques such as data augmentation and preprocessing. This is followed by deep learning models and their implementation methods. The evaluation process is also reported in this section. Process overview diagram is shown in Figure 1.

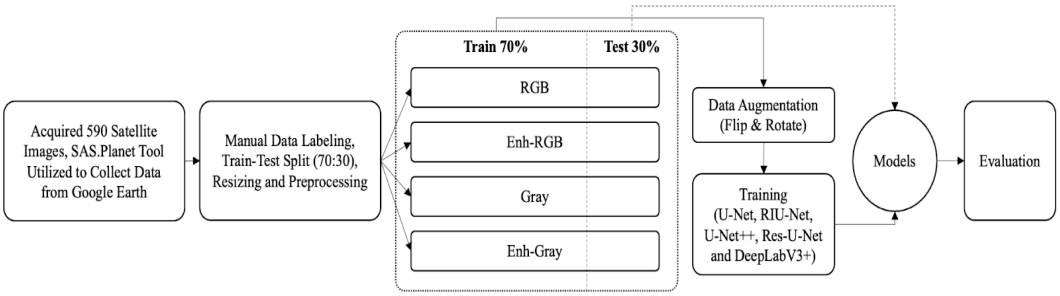

Figure 1. Process overview diagram

2.1 Study area and dataset

This research was focused on the collection of satellite imagery data in remote and geographically challenging mountainous areas that were primarily characterized by dense forested terrain. This led to challenges and issues encountered in planning, management, and development for both government and business sectors. For instance, remote areas often present difficulties in conducting population censuses due to limited accessibility and the high costs associated with transportation infrastructure challenges. Consequently, the predominant landscape of these areas typically comprises a juxtaposition of forested regions and human-made structures.

In this study, Loei Province in Thailand was selected as the focused area. The area represented a geographically and contextually relevant experimental site that adhered to geographical principles and was an appropriate one for addressing the challenges. The province holds the distinction of being a border region adjacent to Laos, further contributing to its unique characteristics. The selected experimental area is situated within a specific subdistrict, which exhibits a relatively higher population density compared to the urban core area. This geographical region is approximately bounded by latitude 17.59° to 17.69° N and longitude 101.67° to 101.81° E, covering an area of approximately 128.50 square kilometers.

The dataset utilized in this research comprised high-resolution satellite images obtained from the Google Earth (Google, 2022). It is acknowledged that previous studies also employed satellite imagery for similar purposes (Wen et al., 2019; Zhang et al., 2021; Chen et al., 2023). To collect the data, the SAS.Planet tool (GIS English, 2023), which facilitates the extraction of imagery from specified areas of interest, was used. The dataset consists of 590 color images, each with dimensions of 517 x 517 pixels and a spatial resolution of 0.6 m, equivalent to a 19x magnification of the original data provided by the data provider.

To facilitate subsequent analysis, the labeling process, specifically focusing on identifying building rooftops within the images, was manually conducted. The resulting

labels represented binary masks or ground-truth data, preserving the same dimensions as the original images (517x517x1). In these labels, areas corresponding to buildings were represented by white color, while non-building areas were represented by black color as shown in Figure 2. After a thorough examination and label creation for each image, the dataset was divided into two distinct groups. The first group consisted of images containing at least one building rooftop, while the second group comprised images without any building rooftops. The total number of images in these two groups were 294 and 296, respectively. Focusing on this area, we sought to contribute to the scientific understanding of remote and challenging environments. The selected area served as a representative case study that aligned with geographical principles and addressed pertinent issues in a comprehensive manner.

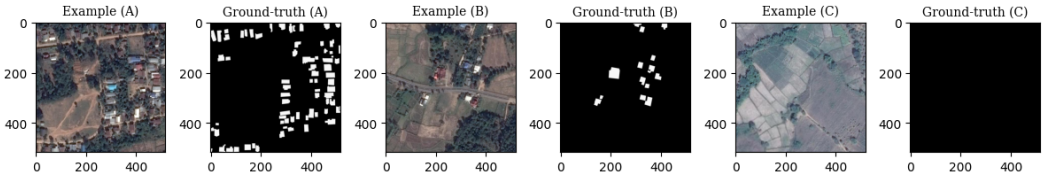

Figure 2. Example of satellite imaged for Area A (community), Area B (suburban), and Area C (agricultural), each with corresponding ground truth labels

The data collection approach in this study was different to approaches used in previous research as we personally gathered both the data and mask labels. However, this introduced a limitation for the learning process of the models as the data set was relatively small. Thus, a direct and precise comparison of the data with evaluation outcomes from other studies was not feasible. Nonetheless, experimental findings revealed that even with self-collected images, the deep learning models proved effective in tackling the task of building extraction from satellite imagery.

2.2 Data preparation

The dataset was divided into two groups: images containing buildings and images without buildings. To mitigate potential issues of biased training or overfitting towards a specific class, a random Train-Test Split technique was employed, allocating the data in a 70:30 ratio. Specifically, 70% of the dataset, comprising 413 images from both groups, was assigned for training purpose, while the remaining 30%, consisting of 177 images from both groups, was set aside for testing. This rigorous partitioning strategy was aimed at ensuring a balanced distribution of samples and facilitation of reliable evaluation of the model's performance.

2.2.1 Data preprocessing

The initial pre-processing technique applied in this study was image resizing, specifically the reduction of image dimensions. The purpose was to decrease the computational requirements and training time. Both the train set and test set underwent this image resizing process, which transformed the images from their original size of 517x517 pixels to a reduced size of 256x256 pixels. This technique was implemented to optimize the training process while preserving the key features and relevant information within the images. In

the subsequent stage, our objective was to explore the influence of pre-processing image quality prior to its utilization in the learning phase and its impact on the learning process and prediction outcomes. In a previous research study, Lin et al. (2017) showed the effectiveness of pre-processing image quality enhancement techniques, which demonstrated superior performance compared to the cases without enhancements. Therefore, a series of experiments was conducted in this research wherein different pre-processing steps were applied to datasets to create distinct dataset variations that were characterized by the following specifications:

a) RGB Color Image Dataset: This dataset comprised the original images resized to the dimensions of 256x256x3.

b) Enhanced RGB Color Image Dataset: In this dataset, RGB color images were further enhanced by applying the gamma correction technique. A fixed gamma value of 0.75 was experimentally chosen, resulting in darker image shades. This adjustment was aimed at addressing the fact that natural outdoor environments generally exhibit less light reflection compared to the rooftops of buildings. Subsequently, the images underwent a color balance process, specifically adjusting the CIELAB color channel values (International Commission on Illumination, 2012). This modification was motivated by the human perceptibility of CIELAB color values (Rosentrater & Evers, 2018), which are closely related to the digital RGB color values. Hence, an additional hypothesis was stated that if humans were able to identify buildings more easily in the images, the learning model might also exhibit improved predictive performance. Finally, the images were subjected to an unsharp masking process, combining Gaussian blur (Gedraite & Hadad, 2011) and weighted image adjustment techniques (Li et al., 2014). Consequently, the images underwent contrast stretching to improve the overall contrast and dynamic range.

c) Grayscale Image Dataset: This dataset involved converting the RGB color images to grayscale in the dimensions of 256x256x1.

d) Enhanced Grayscale Image Dataset: In this dataset, the grayscale images were enhanced using the unsharp masking technique to increase image sharpness. Additionally, Contrast Limited Adaptive Histogram Equalization (CLAHE) (Vidhya & Ramesh, 2017) was applied to improve the distribution of contrast differences.

By following this approach, the train and test sets were utilized to create the four datasets as shown in Figure 3.

Figure 3. Comparison of different datasets for building extraction

2.2.2 Data augmentation

Due to the constrained size of the learning dataset, the researchers aimed to overcome this constraint by utilizing data augmentation techniques. A previous study by Shorten &

Khoshgoftaar (2019) explored data augmentation using various color manipulation methods and other methods. However, this specific investigation focused on pre-processing techniques that entailed segmenting the dataset based on color values.

The augmentation strategies selected for implementation included two principal transformations: rotation and flipping. Rotation was applied to each image in the dataset in two distinct orientations; this involved a 90-degree rotation both clockwise and counterclockwise. This rotational augmentation was designed to ensure the model's invariance to the orientation of objects within the satellite imagery, thereby increasing its adaptability to diverse spatial configurations. Concurrently, flipping augmentation was executed along both the vertical and horizontal axes. Vertical flipping inverts the image top to bottom, while horizontal flipping mirrors the image along its vertical midline. These flipping operations further contributed to the model's resilience against variations in perspective and alignment, simulating a wider range of viewing angles and spatial arrangements.

Consequently, these strategies markedly improved the model's capacity for generalization across unfamiliar datasets. Such an approach guarantees that, despite alterations in the physical appearance of images through rotations and flips, the semantic content and contextual integrity are maintained. This methodology permitted the model to derive insights from an augmented, yet consistently meaningful dataset. The outcome of these transformations yielded a dataset of 2,065 images for the learning process. Figure 4 showcases illustrative examples of these transformations.

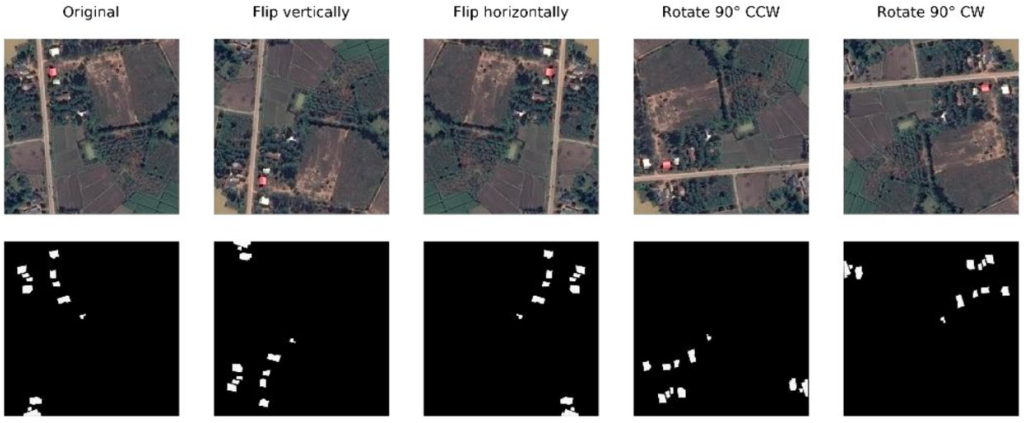

Figure 4. Comparison of data augmentation techniques: flip and rotate

2.3 Methods

The aim of this research was to compare and evaluate the prediction performance of deep learning models, namely U-Net, RIU-Net, U-Net++, Res-U-Net, and DeepLabV3+. These models have been widely studied and applied in various fields. Our focus was on extracting buildings from specific areas of interest. With respect to Res-U-Net, it stood out with its incorporation of a backbone architecture. Additionally, DeepLabV3+ was used in the comparison due to its distinct architectural characteristics.

2.3.1 U-Net

U-Net is a convolutional neural network architecture that was originally introduced for biomedical image segmentation, specifically targeting the segmentation of neuronal structures in brain tissue (Ronneberger et al., 2015). It comprises an encoder pathway for capturing contextual information and a decoder pathway for precise localization, facilitated by skip connections that enable the fusion of low-level and high-level features. Due to its exceptional performance in segmentation tasks, U-Net has gained popularity and been successfully applied across diverse domains, extending beyond medical imaging.

Building on its foundational design, U-Net's encoder-decoder architecture is intricately designed with convolutional layers, interspersed with max pooling for down-sampling and transposed convolutional layers for up-sampling, facilitating detailed feature extraction and spatial information preservation. The integration of ReLU activation functions introduces necessary non-linearity. Skip connections, a hallmark of U-Net, bridge the gap between encoder and decoder, ensuring high-resolution features are directly propagated across the network, which is essential for accurate segmentation. Training U-Net often involves tailored loss functions, like the Dice Coefficient, to directly optimize for segmentation performance, alongside techniques like dropout and batch normalization to enhance model generalization.

Moreover, U-Net's pioneering approach to leveraging extensive data augmentation has notably expanded its utility in medical imaging, addressing the challenge of limited annotated datasets. Furthermore, its adaptable framework has spurred the development of numerous variants, each enhancing U-Net's foundational strengths to cater to specific segmentation needs or improve overall performance. This legacy of innovation underscores U-Net's significant impact on the field of image segmentation, propelling advancements across both medical and non-medical applications.

2.3.2 RIU-Net

RIU-Net, or Residual-Inception U-Net, is an advanced adaptation of the U-Net architecture, where the integration of the Residual and Inception modules serves to enhance feature extraction performance (He et al., 2016; Sariturk & Seker, 2022). Residual module contributes to gradient propagation, facilitating effective information flow (Alom et al., 2019), while Inception module facilitates the detection of multi-scale features. Both the encoder and decoder pathways of RIU-Net benefit from the application of these Residual and Inception modules.

Diving deeper into RIU-Net’s architecture, the Residual modules help mitigate the vanishing gradient problem, allowing for deeper network constructions without loss of performance. Meanwhile, the Inception modules, by processing data through multiple filter sizes simultaneously, capture a broader range of spatial information, enhancing the model's ability to recognize features at various scales. This combination not only improves segmentation accuracy but also increases the model's robustness to variations in input image sizes and shapes.

Additionally, RIU-Net's design optimizes computational efficiency, making it suitable for processing of large datasets in high-resolution imaging applications. By inheriting U-Net's flexible architecture and incorporating these sophisticated modules, RIU-Net represents a significant leap forward in the field of image segmentation, promising improvements in both precision and scalability.

2.3.3 U-Net ++

U-Net++ extends the U-Net architecture by introducing nested and dense skip connections to capture features at multiple scales and enhance information flow within the network (Zhou et al., 2018). By incorporating these connections, U-Net++ effectively leverages hierarchical features, enabling it to capture and utilize rich contextual information.

Further dissection of U-Net++'s design reveals its unique skip pathway architecture, which differs from the traditional U-Net by allowing for more flexible feature fusion across different levels of the network. This is achieved through its nested and dense skip connections that facilitate more effective integration of low-level detail with high-level semantic information, significantly improving the accuracy of segmentation tasks. The design of U-Net++ not only enhances feature extraction capabilities but also introduces redundancy reduction, which streamlines the model's learning process.

As a result, U-Net++ demonstrates improved performance on a variety of segmentation tasks, showing promise in areas requiring fine-grained detail recognition, such as satellite image analysis. The adaptability and improved efficiency of U-Net++ make it a valuable tool for researchers and practitioners seeking to push the boundaries of image segmentation accuracy and performance.

2.3.4 Res-U-Net

Res-U-Net entails the fusion of the fundamental architecture of U-Net with a pre-trained backbone network, exemplified by ResNet (employed as a case study in this research), as its encoder (Diakogiannis et al., 2020). This strategic amalgamation yields enhancements in the efficacy of feature extraction, facilitating the comprehensive capture of both low-level and high-level features. Significantly, this approach affords the flexibility to adapt and fine-tune the pre-trained weights and backbone, thereby mitigating the constraints imposed by the input characteristics during the learning process.

In the architectural innovation of Res-U-Net, the incorporation of ResNet as the encoder backbone marks a significant advancement, transcending traditional segmentation tasks. ResNet's deep residual learning framework, characterized by its identity shortcut connections, solves the depth dilemma, enabling the effective training of deeper neural networks by facilitating gradient flow. Within Res-U-Net, these residual mechanisms enhance the encoder's ability to maintain critical feature details through layers, crucial for capturing nuanced textures and structural variations across diverse imaging landscapes.

The strategic use of ResNet's pre-trained weights within Res-U-Net not only propels the model towards faster convergence but also equips it with a versatile, pre-learned feature repertoire, ready to be adapted and fine-tuned across a broad spectrum of segmentation challenges beyond the confines of any imagery. This adaptability, coupled with the model's fine-tuning capabilities, allows for precise customizations to the backbone, optimizing performance for specific tasks. Consequently, Res-U-Net stands as a paradigm of versatility and efficiency, setting new benchmarks in image segmentation across various domains, from environmental monitoring to urban planning.

2.3.5 DeepLabV3+

DeepLabV3+ is a state-of-the-art deep learning model for semantic segmentation tasks. It employs atrous convolution, also known as dilated convolution, to capture both local and global contextual information at multiple scales (Chen et al., 2018; Liu et al., 2021).

Furthermore, DeepLabV3+ utilizes a spatial pyramid pooling module to aggregate multi-scale features effectively.

Exploring its architecture further reveals that the integration of atrous convolution and spatial pyramid pooling is crucial for efficiently capturing contextual information at multiple scales. Expanding on this, atrous convolution, by varying dilation rates, expands the receptive field of convolutional filters, enabling the model to grasp both minute details and broader contextual cues. This method preserves the spatial resolution of feature maps, crucial for maintaining the fidelity of segmentation boundaries. Complementarily, pyramid pooling systematically aggregates features at multiple scales by applying pools at different resolutions, ensuring that global context is integrated across the entire scene. This layered approach mimics a multi-scale pyramid, where each level captures distinct spatial hierarchies, facilitating the model's ability to discern features from varying distances and sizes. Additionally, its encoder-decoder architecture, refined with depthwise separable convolution, optimizes boundary delineation while maintaining model compactness. These innovations position DeepLabV3+ as a versatile and efficient choice for semantic segmentation, applicable across a diverse range of visual understanding tasks.

In summary, in this investigation, we strove to provide a meticulous analysis and comparative assessment of U-Net, RIU-Net, U-Net++, Res-U-Net, and DeepLabV3+ models. By addressing the specific task of building extraction from targeted study regions, we aimed to discern the most effective model for this application domain.

2.4 Evaluation metrics

The evaluation techniques employed in this study encompassed a range of metrics, among which the Confusion Matrix (Kohavi & Provost, 1998) holds significant importance. It delineates crucial conditions for analysis, including:

• True Positive (TP): Corresponding to accurate predictions of building positions that align with the ground truth.

• True Negative (TN): Reflecting precise predictions of non-building positions, in accordance with the ground truth.

• False Positive (FP): Representing erroneous predictions of positions as buildings, despite being non-building areas.

• False Negative (FN): Indicating flawed predictions of positions as non-buildings, when they are indeed buildings.

These metrics play a pivotal role in the evaluation process, facilitating the utilization of equations (1) to (4).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

$$F1\ Score = \frac{2 * TP}{2 * TP + FP + FN} \tag{4}$$

Additionally, in this study, two similarity coefficient metrics, Intersection-over-Union (IoU) and Dice Coefficient, were employed to facilitate comparative analysis.

2.4.1 Intersection-over-Union (IoU)

IoU, also known as Jaccard Index, is a commonly used evaluation metric in semantic segmentation tasks. It was first proposed by Jaccard (1912) and has since become a standard measure for assessing accuracy of segmentation algorithms. IoU quantifies the overlap between the predicted segmentation mask and the ground truth mask by calculating the ratio of their intersection to their union. By comparing the predicted and actual building regions, IoU provides a quantitative measure of their spatial alignment and accuracy. The formula for the Jaccard Index is presented as equation (5).

$$IoU = \frac{TP}{TP + FP + FN} \tag{5}$$

2.4.2 Dice Similarity Coefficient (DSC)

The Dice Coefficient, also known as the Dice Similarity Coefficient (DSC), is a widely used similarity metric for evaluating semantic segmentation models. It was introduced by Dice (1945) and has since become a popular measure in medical image analysis (Sørensen, 1948) and later other applications. Dice Coefficient quantifies the overlap between the predicted and ground truth segmentation masks by calculating the intersection divided by the sum of the areas of both masks. It provides a robust measure of segmentation performance. The formula for the Dice Coefficient can be seen in equation (6).

$$DSC = \frac{2 * TP + smooth}{2 * TP + FP + FN + smooth} \tag{6}$$

In the training of deep learning models for segmentation tasks, a critical adaptation of their application is the introduction of a smoothing term, which is essential for ensuring the stability of the training process. This modification addresses a potential issue that arises when both the ground truth and the prediction entirely consist of zeros, indicating perfect agreement in cases of background dominance but leading to an undefined condition in the DSC formula due to division by zero.

Without the smoothing term, such scenarios would result in an infinite loss, disrupting the weight propagation and potentially derailing the training procedure. The smoothing term, therefore, serves a dual purpose: it prevents the occurrence of an infinite loss by avoiding division by zero and maintains the continuity of the gradient flow, thereby safeguarding the integrity of the model's learning trajectory. By incorporating this small constant, the training process becomes more robust, allowing the model to learn effectively even in the presence of challenging cases where the overlap between the predicted and actual segmentations might be minimal or non-existent.

3. Results and Discussion

In this study, the experiments covered the entire process from data preparation to result evaluation. Python code was utilized in conjunction with Tensorflow (Abadi et al., 2016) library, and the processing resources of Google Colab was employed (Google, n.d.).

Additionally, the processing resource used for deep learning model training was NVIDIA Tesla V100 SXM2 16 GB GPU backend. Satellite imagery data was imported and stored in Personal Google Drive before being processed in Google Colab. Both the training and testing sets of image data from the created dataset, comprising four sets in total (RGB, Enh-RGB, Gray, and Enh-Gray), were imported and underwent the Flatten process and the results were then stored as image files.

The deep learning models were trained with consistent learning configurations. The validation split was set at 70% for training and 30% for the validation set. The models processed 250 epochs of training with a batch size of 8. The RMSProp optimizer with a learning rate of 0.001 was used. To compare the models' performances, IoU and Dice Similarity Coefficient (DSC) were employed. The experiments generated results of 48 model combinations in total, based on the datasets and evaluation metrics, as shown in Table 1.

The Res-U-Net and DeepLabV3+ models utilized pre-trained weights from the ImageNet dataset, designed for 3-channel RGB images. However, the study also explored datasets comprising gray and enhanced gray images, characterized by 1-channel RGB representation. Each pixel in these images spans a grayscale intensity range of 0-255, diverging from the multichannel color information used in the initial pre-training.

Consequently, grayscale and enhanced grayscale image datasets could not be used for training of these specific models. Therefore, Res-U-Net and DeepLabV3+ models were trained exclusively using RGB and enhanced RGB image datasets.

The U-Net series models underwent architectural refinements in both Encoder and Decoder components to ensure their harmonious configuration. The selection of filters was customized to suit the input dimensions of the image data used for training. For Inception module in RIU-Net, there were four branches of processing. The initial three branches applied Convolutional Layers that varied in terms of parameter settings, including filters, kernel size, and the number of layers. The final branch incorporated a max pooling layer. As a result, the outputs of all four branches were concatenated before being subjected to an additional process, facilitating the fusion of the input and the inception module's output in Residual block. For U-Net++, Res-U-Net, and DeepLabV3+ architectures, adjustments were only applied for overall parameter configuration.

However, within the group of models that employed Res-U-Net and DeepLabV3+ backbones, the experiments were conducted using the original architectures in conjunction with a comparison of three backbone versions: ResNet-50, ResNet-101, and ResNet-152. The aim of this investigation was to ascertain which backbone configuration produced the most favorable outcomes when applied to the given input datasets. It was noteworthy that the architectural compositions of these three ResNet backbones exhibited varying levels of complexity, resulting in disparate learning durations and resource-intensive memory requirements.

Following the definition of the model's architecture within the specified architectural framework and the parameterization tailored to align with the characteristics of the four datasets, the subsequent step involved testing the trained models against a test dataset (test set). Each model, trained on a specific type of dataset, was evaluated using a test set possessing similar characteristics. For instance, the Res-U-Net model trained with an RGB dataset was subsequently tested against an RGB test set. Furthermore, the evaluation employed three metrics: Intersection over Union (IoU), Dice Similarity Coefficient (DSC), and an evaluation metric using the F1 Score as the primary criterion, considered alongside Precision and Recall. Notably, the accuracy score could also be included in the analysis to examine predictions categorized as True Negatives (TN), given that this was the only equation used to evaluate TN values, as per Section 3.4 (1).

Table 1. Comparing results obtained from each model by F1 score in descending order

Model	Dataset	Training Metric	Accuracy	Precision	Recall	F1	DSC	IoU
RUN w. RN-50	**RGB**	**DSC**	**0.9941**	**0.8492**	**0.8898**	**0.8543**	**0.8548**	**0.7812**
DLv3+ w. RN-50	Enh-RGB	DSC	0.9924	0.8820	0.8725	0.8518	0.8525	0.7855
DLv3+ w. RN-152	RGB	DSC	0.9942	0.8450	0.8842	0.8448	0.8455	0.7717
DLv3+ w. RN-101	RGB	DSC	0.9940	0.8366	0.8892	0.8424	0.8430	0.7700
DLv3+ w. RN-50	Enh-RGB	IoU	0.9917	0.8343	0.9076	0.8413	0.8419	0.7730
DLv3+ w. RN-101	Enh-RGB	DSC	0.9925	0.8539	0.8873	0.8408	0.8415	0.7742
DLv3+ w. RN-152	Enh-RGB	IoU	0.9926	0.8687	0.8690	0.8401	0.8409	0.7759
RUN w. RN-152	Enh-RGB	DSC	0.9926	0.8623	0.8739	0.8395	0.8450	0.7747
RUN w. RN-101	RGB	DSC	0.9941	0.8367	0.8849	0.8373	0.8408	0.7642
RUN w. RN-152	RGB	DSC	0.9942	0.8475	0.8724	0.8372	0.8389	0.7636
DLv3+ w. RN-152	Enh-RGB	DSC	0.9924	0.8600	0.8693	0.8355	0.8366	0.7705
RUN w. RN-152	Enh-RGB	IoU	0.9920	0.8429	0.8942	0.8354	0.8374	0.7691
RUN w. RN-101	Enh-RGB	IoU	0.9923	0.8310	0.9046	0.8336	0.8344	0.7680
RUN w. RN-101	Enh-RGB	DSC	0.9923	0.8310	0.9046	0.8336	0.8344	0.7680
RUN w. RN-50	Enh-RGB	DSC	0.9926	0.8583	0.8738	0.8323	0.8330	0.7687
U-Net++	RGB	DSC	0.9934	0.8208	0.8828	0.8277	0.8286	0.7515
DLv3+ w. RN-50	RGB	DSC	0.9941	0.8426	0.8628	0.8230	0.8248	0.7500
RUN w. RN-50	Enh-RGB	IoU	0.9927	0.8517	0.8705	0.8222	0.8267	0.7588
DLv3+ w. RN-101	Enh-RGB	IoU	0.9918	0.8026	0.9167	0.8201	0.8219	0.7516
U-Net++	Enh-RGB	IoU	0.9920	0.8380	0.8820	0.8183	0.8196	0.7520
DLv3+ w. RN-152	RGB	IoU	0.9935	0.7863	0.9121	0.8164	0.8229	0.7400
RUN w. RN-50	RGB	IoU	0.9925	0.7779	0.9153	0.8153	0.8160	0.7364
U-Net	RGB	IoU	0.9929	0.7979	0.8865	0.8143	0.8152	0.7361
U-Net	RGB	DSC	0.9933	0.8259	0.8595	0.8136	0.8147	0.7382
DLv3+ w. RN-50	RGB	IoU	0.9935	0.7910	0.9012	0.8085	0.8095	0.7321
U-Net++	Enh-RGB	DSC	0.9920	0.8320	0.8733	0.8050	0.8069	0.7375
DLv3+ w. RN-101	RGB	IoU	0.9932	0.7607	0.9194	0.8009	0.8096	0.7243
RIU-Net	RGB	DSC	0.9931	0.8374	0.8113	0.7956	0.8014	0.7170
U-Net	Enh-RGB	DSC	0.9914	0.7892	0.8766	0.7787	0.7820	0.7093
U-Net	Enh-RGB	IoU	0.9914	0.7892	0.8766	0.7787	0.7820	0.7093
U-Net++	RGB	IoU	0.9925	0.7386	0.9089	0.7779	0.7797	0.6983
RUN w. RN-101	RGB	IoU	0.9927	0.7420	0.9191	0.7756	0.7822	0.6990
U-Net++	Gray	DSC	0.9922	0.7730	0.7990	0.7373	0.7405	0.6628
RIU-Net	Enh-RGB	DSC	0.9907	0.7136	0.8590	0.7048	0.7206	0.6322
U-Net++	Gray	IoU	0.9883	0.6729	0.8876	0.7011	0.7032	0.6240
RIU-Net	Gray	DSC	0.9923	0.7805	0.7668	0.7010	0.7102	0.6300
RUN w. RN-152	RGB	IoU	0.9918	0.6284	0.9217	0.6770	0.6851	0.5992
U-Net	Gray	DSC	0.9911	0.7163	0.7875	0.6732	0.6799	0.5960
U-Net	Gray	IoU	0.9908	0.6659	0.7978	0.6569	0.6634	0.5788
RIU-Net	RGB	IoU	0.9908	0.6135	0.9009	0.6556	0.6695	0.5732
RIU-Net	Enh-RGB	IoU	0.9908	0.6592	0.8562	0.6450	0.6796	0.5732
RIU-Net	Gray	IoU	0.9904	0.5885	0.8278	0.5889	0.6060	0.5121
RIU-Net	Enh-Gray	DSC	0.9874	0.7270	0.6493	0.5193	0.5393	0.4573
U-Net++	Enh-Gray	DSC	0.9870	0.6017	0.6827	0.4907	0.4960	0.4230
U-Net	Enh-Gray	DSC	0.9864	0.5535	0.6876	0.4580	0.4650	0.3895
U-Net++	Enh-Gray	IoU	0.9852	0.4641	0.7484	0.4255	0.4308	0.3517
U-Net	Enh-Gray	IoU	0.9842	0.3980	0.7309	0.3767	0.3817	0.3060
RIU-Net	Enh-Gray	IoU	0.9856	0.4051	0.7438	0.3685	0.3837	0.2974

Note: In this Table, for example, 'RUN w. RN-50' refers to Res-U-Net with ResNet-50, and 'DLv3+ w. RN-50' denotes DeepLabV3+ with ResNet-50. Abbreviations are utilized to concisely represent model configurations.

An example involved the testing of the Res-U-Net model with a ResNet-50 backbone, trained on an RGB dataset, against a test set, as depicted in Figure 5. In this figure, Figure 5(a) represents the input image for the model to predict building structures, and Figure 5(b) denotes the Ground-truth identifying the locations of these structures. These images were utilized to evaluate the model's performance using the three metrics mentioned. Subplot (c) highlights the areas predicted by the model (in white) using the IoU metric during training, yielding an F1 Score of 73.13%, with IoU and DSC scores of 57.64% and 73.13%, respectively. Subplot (d) shows the error in predictions as deviations from the Ground-truth (b), similar to subplots (e) and (f), which represent the prediction outcomes and their corresponding errors, achieving evaluation scores of 76.17% for the F1 Score and 61.51% and 76.17% for IoU and DSC, respectively. Additionally, for the specific examples tested and the evaluation of the outcomes using the F1 and DSC metrics, the scores are identical due to the previously mentioned calculation formula.

Consequently, upon subjecting the entire cohort of 48 models to the testing phase using the designated test set and subsequently applying evaluation metrics to the predictive outcomes, the resultant test scores are presented as per Table 1. Additionally, the research design employed herein engendered an extensive array of models and corresponding outcomes. The ensuing discourse will methodically address each point of interest in a sequential manner.

The model that yielded the best predictive results when evaluated by the F1 score was Res-U-Net with ResNet-50, which utilized an RGB dataset and was assessed during training with DSC, achieving an F1 score of 85.43%. In comparison, when evaluated with the DSC and IoU metrics, the scores were 85.48% and 78.12%, respectively. However, when considering the balance between Precision and Recall, it was observed that DeepLabV3+ with ResNet-50 (Enh-RGB, DSC) performed better but the outcomes yielded F1 scores of 85.18%. Yet, the F1 calculation does not derive from a joint operation of precision and recall scores, as per Section 3.4 (4).

Therefore, to analyze the outcomes derived from the evaluation, it was imperative to conduct a comparative analysis of the predictive results between the two aforementioned models, as depicted in Figure 6. A key observation was that the white predicted areas by Res-U-Net with ResNet-50 (RGB, DSC) delineated the structures more distinctly than the predictions by DeepLabV3+ with ResNet-50 (Enh-RGB, DSC), which tended to merge structures into a single area, indicating a lack of clear separation. Furthermore, the surrounding predicted areas were larger, which was consistent with the average accuracy values that consider True Negatives in the calculation. Nonetheless, when compared to the Ground-truth, the DeepLabV3+ with ResNet-50 (Enh-RGB) model predicted a more comprehensive coverage of structures, highlighting its precision. However, considering the areas where the predicted structures were connected, this leads to result in a slightly lower average F1 score.

Furthermore, when comparing models of identical types and dataset features but employing the IoU metric throughout the training phase, the outcomes yielded F1 scores of 84.13% for DeepLabV3+ with ResNet-50 (Enh-RGB, IoU) and 81.53% for Res-U-Net with ResNet-50 (RGB, IoU). Upon examination, it became evident that the score gap between the two versions of DeepLabV3+ was not substantial when compared to Res-U-Net.

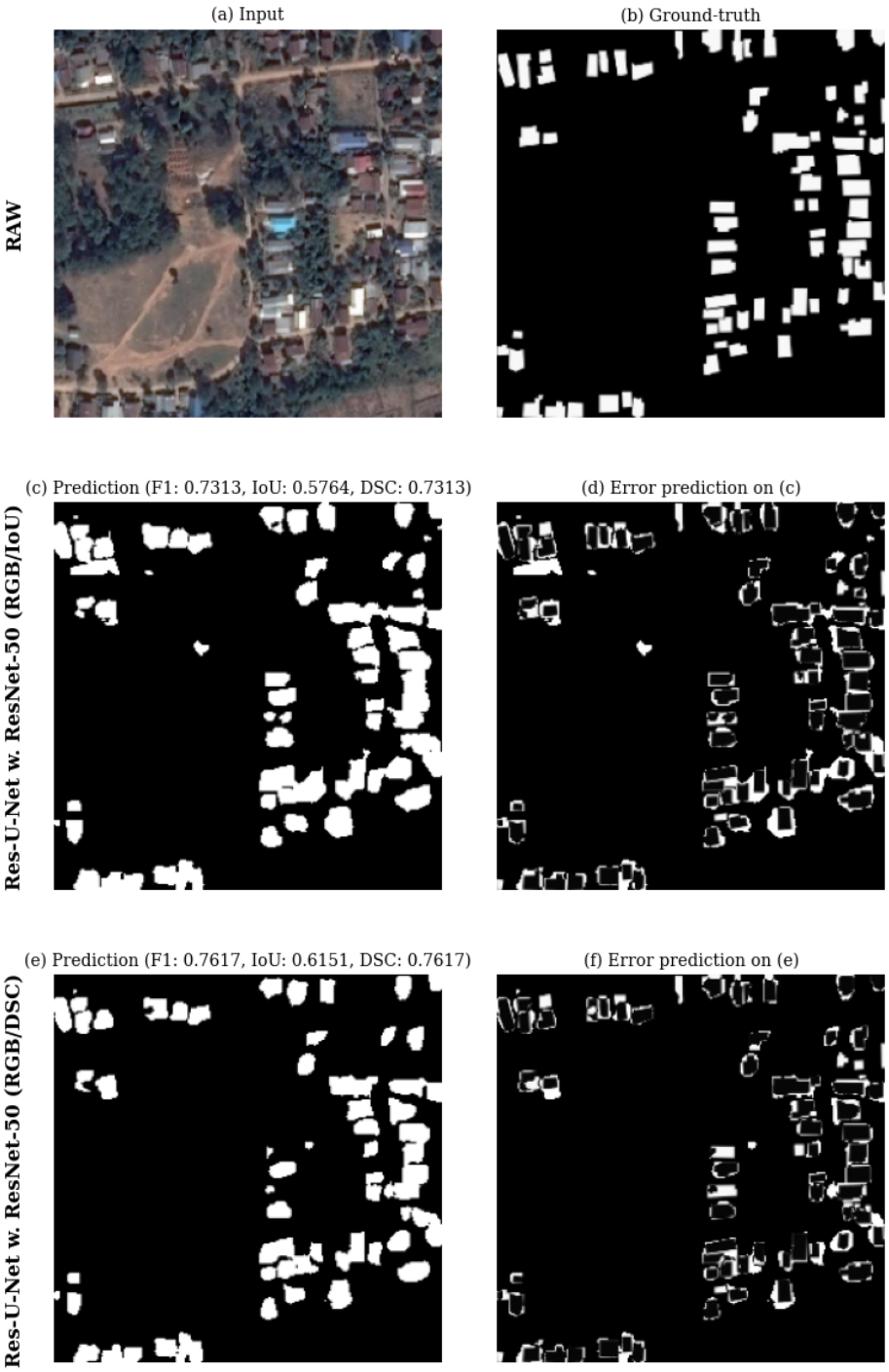

Figure 5. Comparative evaluation of Res-U-Net with ResNet-50 on RGB dataset: predictive accuracy through F1 score, IoU, and DSC

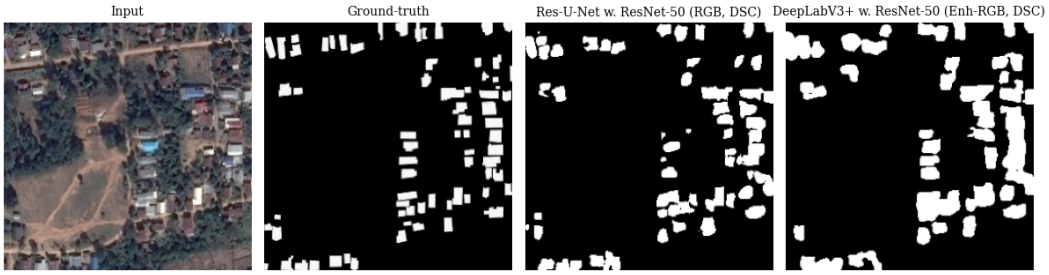

Figure 6. Comparative of structural segmentation by Res-U-Net with ResNet-50 (RGB, DSC) and DeepLabV3+ with ResNet-50 (Enh-RGB, DSC)

Subsequent to addressing models with comparably high performance, the evaluation scores unequivocally suggested that models incorporating a backbone architecture exhibited superior effectiveness. Models that were positioned at the higher end of the table, arranged in descending order of F1 scores, consistently incorporated backbone structures. Another noteworthy observation is that the models trained on the RGB image datasets consistently achieved higher evaluation scores compared to their counterparts trained on grayscale image datasets. Moreover, the deployment of the DSC metric during the training phase contributed to the enhanced performance of the models.

Therefore, the models trained and evaluated using DSC metric demonstrated slight variations in performance. Specifically, Res-U-Net model with ResNet-50 backbone exhibited a slightly better performance than DeepLabV3+ model with a ResNet-50 backbone, with DSC scores of 85.48% and 85.25%. It is important to note that Res-U-Net model was trained using RGB images, while DeepLabV3+ model utilized enhanced RGB images during training.

However, upon reviewing the sample prediction images (referred to Figure 7 in the DSC result comparison). The significance of color in the datasets and backbone architecture in segmentation models is important. The comparison revealed a consistent trend: color datasets (both RGB and Enhanced RGB) yielded superior segmentation results, markedly impacting F1 scores more than grayscale datasets (both Gray and Enhanced Gray), affirming the necessity of thoughtful dataset choice. Additionally, the use of backbone architectures, such as ResNet-50, ResNet-101, and ResNet-152, improved F1 scores and segmentation precision compared to models without backbones. This finding accentuates the critical role of both dataset selection and backbone inclusion in enhancing model performance.

Discrepancies in model predictions provide actionable insights for refining model parameters and informing preprocessing strategies. Specifically, our results indicated that models without sophisticated backbones, such as U-Net, RIU-Net and U-Net++, were prone to structural distortions in segmentation. In contrast, models with advanced backbones demonstrated fewer errors and more accurate delineation of structures. It was observed that Res-U-Net model provided more distinct segmentation of building structures than DeepLabV3+. Nonetheless, the Res-U-Net model also exhibited a higher occurrence of small white dots resembling noise.

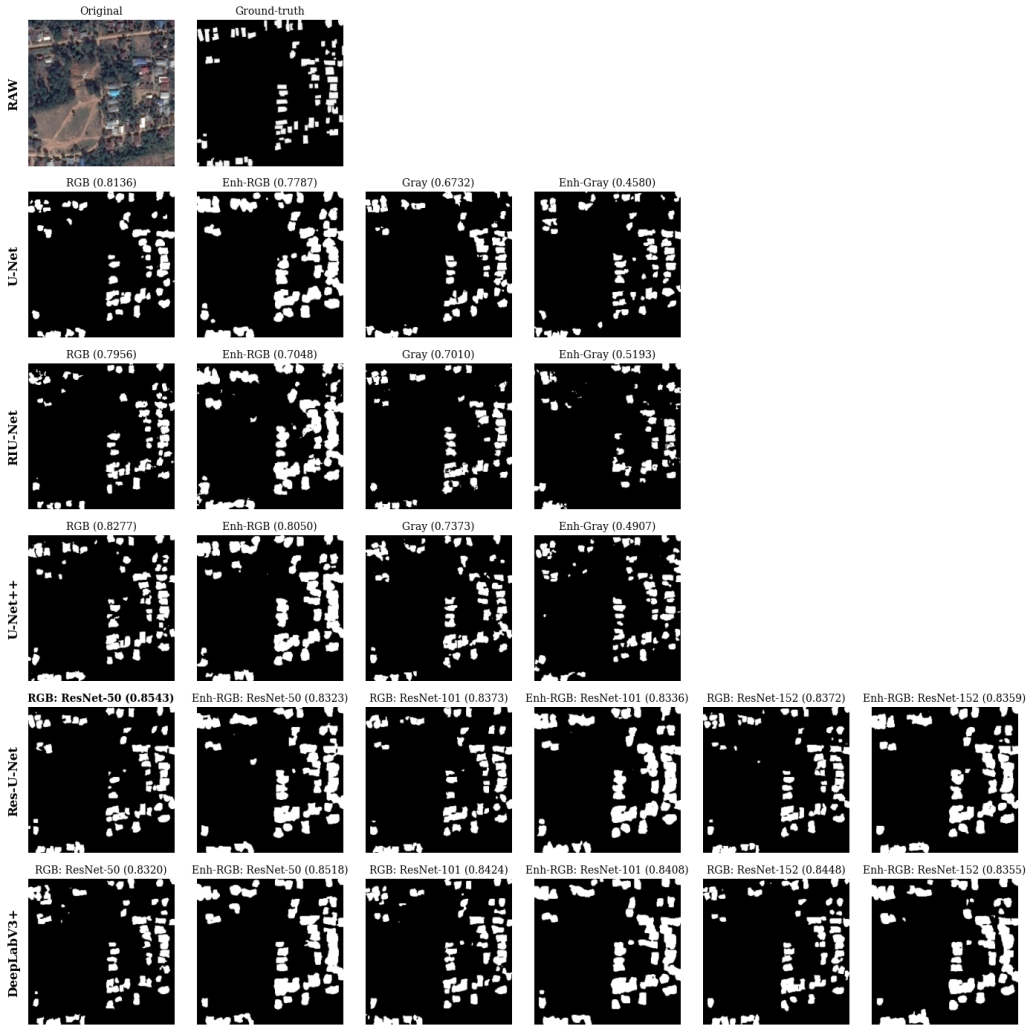

Figure 7. Examples of results obtained from each model based on DSC metric with F1 score

The models were evaluated using the IoU metric and tested on the designated test set. The experimental results indicated that DeepLabV3+ coupled with ResNet-152 and ResNet-50 backbones, trained on the RGB image dataset, achieved IoU scores of 77.59% and 77.30%, respectively. Similarly, the corresponding F1 scores were 84.01% and 84.13%.

Moreover, when examining the prediction outcomes for a specific image (referred to Figure 8 in the comparison of IoU results), it became apparent that models trained and evaluated with the IoU metric during the training process exhibited a lower overall performance compared to those assessed with the DSC metric. This observation suggests that while IoU provides a stringent measure of overlap between predicted and actual segments, the DSC metric, with its balanced account of precision and recall, was probably

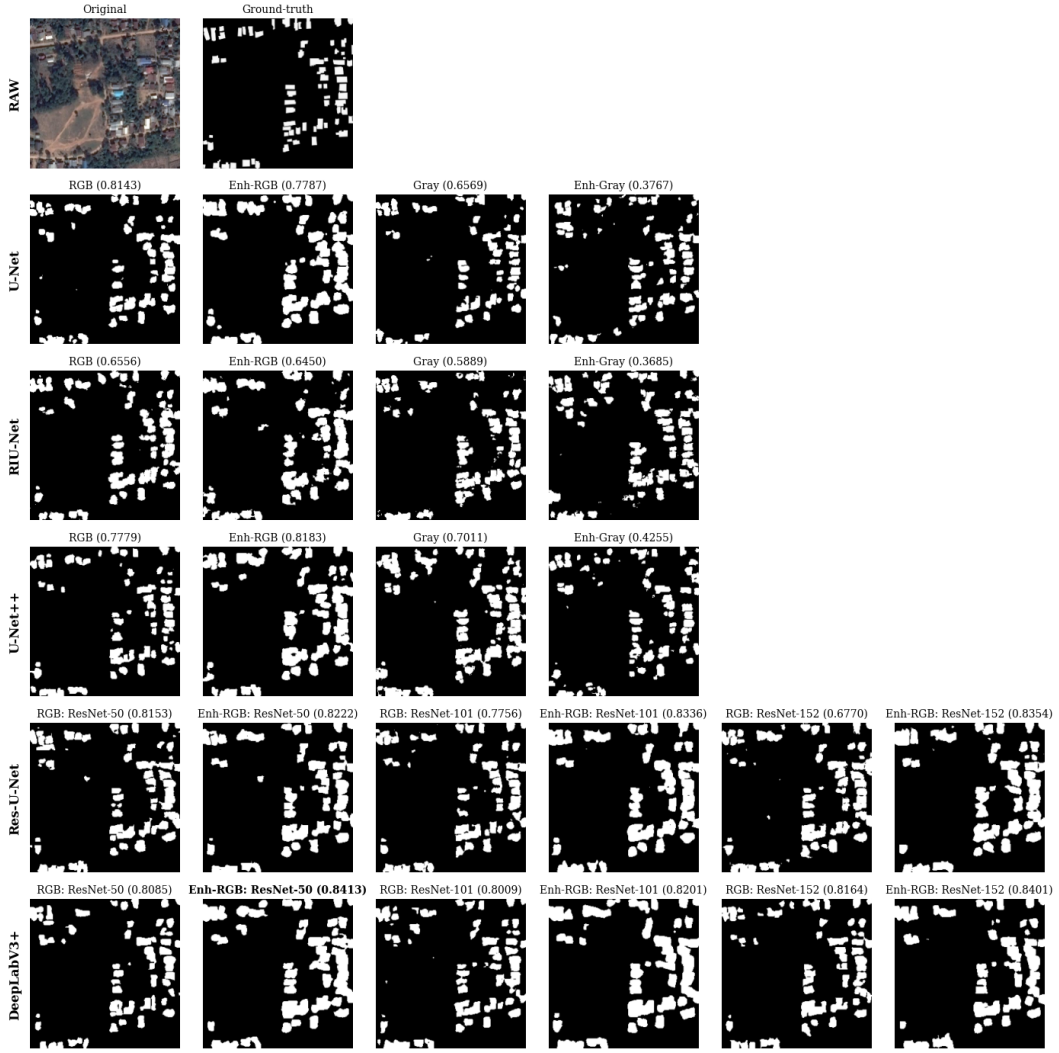

Figure 8. Examples of results obtained from each model based on IoU metric with F1 score

more aligned with the visual quality of segmentation, yielding higher scores in the evaluation of results. It was evident that the DeepLabV3+ models with ResNet-152 and ResNet-50 backbones performed promisingly in accurately identifying building structures, exhibiting minimal instances of insignificant white noise. Despite the interconnectedness of the building structures, the models' predictions remained remarkably accurate. Overall, RGB and enhanced RGB image dataset yielded superior performance across all models. On the other hand, the grayscale and enhanced grayscale image datasets showed considerably lower evaluation scores when compared to their RGB counterparts. Upon visual inspection of the predicted images, it became apparent that while the buildings were not connected, there were several false-positive predictions, indicating a low precision value.

The findings from this experiment demonstrated that models trained on different datasets and evaluated using the DSC metric generally exhibited better performance compared to those evaluated using IoU metric, as indicated by the F1 score. Specifically, the Res-U-Net model with ResNet-50 (RGB) achieved the highest F1 score of 85.43%, outperforming the DeepLabV3+ model with ResNet-50 (Enh-RGB), which scored 85.18%. According to the experiments, models utilizing specific backbone architectures and pre-trained weights yielded superior results. Furthermore, the results suggested that color plays a significant role in accurately extracting building structures from images. Models evaluated using DSC metric consistently performed better on RGB image datasets, while models evaluated using IoU metric exhibited higher performance on enhanced RGB image datasets.

In the assessment of building extraction accuracy from satellite imagery, utilizing both Intersection over Union (IoU) and Dice Coefficient metrics produced distinct F1 scores, despite the employment of identical models and datasets. The discrepancy in F1 scores, detailed in Table 2, underscores the concept that the choice of metric does not influence the model training duration. It was established that the architectural complexity of the models primarily dictated training times. The analysis indicates that the U-Net model, with its minimalistic architecture, needed lower training time relative to more complex models. Furthermore, models incorporating backbones, notably Res-U-Net and DeepLabV3+, exhibited extended training periods.

This insight emphasizes the importance of considering both performance effectiveness and training duration in model selection, advocating for a strategic balance to optimize accuracy and computational time expenditure.

Table 2. Comparing computational time of each model

Model	Dataset	Training Time (h)	Training Time (min) (multiply the time (h) by 60)
U-Net	RGB	1.264	75.84
	Enh-RGB	1.344	80.64
	Gray	1.344	80.64
	Enh-Gray	1.346	80.76
RIU-Net	RGB	2.082	124.92
	Enh-RGB	2.181	130.86
	Gray	1.977	118.62
	Enh-Gray	2.121	127.26
U-Net++	RGB	1.561	93.66
	Enh-RGB	1.584	95.04
	Gray	1.583	94.98
	Enh-Gray	1.579	94.74
Res-U-Net with ResNet-50	**RGB**	**2.416**	**144.96**
	Enh-RGB	2.135	128.1
Res-U-Net with ResNet-101	RGB	1.976	118.56
	Enh-RGB	2.211	132.66
Res-U-Net with ResNet-152	RGB	2.446	146.76
	Enh-RGB	2.352	141.12
DeepLabV3+ with ResNet-50	RGB	2.346	140.76
	Enh-RGB	**2.297**	**137.82**
DeepLabV3+ with ResNet-101	RGB	2.374	142.44
	Enh-RGB	2.352	141.12
DeepLabV3+ with ResNet-152	RGB	2.435	146.1
	Enh-RGB	2.451	147.06

Therefore, based on the findings of this experiment, it can be concluded that Res-U-Net model with ResNet-50 is a suitable choice for the given dataset. This model exhibited superior performance in terms of selected evaluation metrics while maintaining relatively lower complexity compared to other backbone architectures, such as ResNet-101 and ResNet-152. The adoption of ResNet-50 not only contributes to efficient resource utilization and time saving during the learning process but also allows for greater flexibility in further customization and development by independently exploring and fine-tuning the backbone architecture.

From this study, it is evident that the selected models show potential for image segmentation. However, each model still has limitations worth analyzing and considering. Starting with U-Net, despite its fast-training time and satisfactory results, its architecture has limitations in detecting features from the images. Despite the model captures local information well, it possibly misses contextual details from larger image regions due to its limited depth and field of view (Wu et al., 2022). Furthermore, it requires a considerable amount of data to enhance its performance further. RIU-Net, which integrates Residual and Inception modules to improve feature extraction, introduces increased complexity to the model architecture, demanding more time and computational resources for training. The experimental results indicate that despite this is the only model without a backbone application, it requires almost as much time as models that do use one. The added complexity also complicates hyperparameter tuning, making optimization more challenging.

U-Net++ features an architecture with nested skip pathways, increasing the number of parameters that need processing. According to experimental durations, it takes a longer time to train than the original U-Net architecture, and its dense skip connections might increase the risk of overfitting on smaller datasets. Res-U-Net, which scored the highest F1, encounters the double-edged sword of complexity and the need to rely on a backbone. The performance of the model can be limited if the backbone is not appropriately selected to match the data and task objectives. Additionally, the use of a backbone brings about transfer learning limitations, utilizing pre-trained weights. Hence, fine-tuning to specific segmentation tasks might present challenges. Lastly, DeepLabV3+, while slightly underperforming in F1 scores compared to Res-U-Net, exhibits a notable balance between precision and recall. However, the application of atrous convolution can increase processing complexity and affect training duration, especially with high-resolution images. The incorporation of spatial pyramid pooling, designed to capture multi-scale information, might still miss capturing spatial details or significant image features at the same level, potentially losing some details. The reliance on a backbone, as in the case of Res-U-Net's limitations, involves transfer learning challenges.

This comparative analysis of the models illuminates the intricate balance between their complexity, performance, and the resources they need. The choice of model hinges on the precise objectives and the computational resources available. Additionally, it is essential to consider the dataset's specific attributes and its volume. However, it is important to be mindful of the potential risks associated with these modifications (Namdeo & Bhadoriya, 2016; Zhang et al., 2018). Moreover, exploring alternative deep learning models and investigating the development of encoders, decoders, and the utilization of different backbones can be fruitful avenues for further research. In a previous study (Zhang et al., 2021), a two-stage framework called SRBuildingSeg was proposed to achieve super-resolution (SR) building extraction, which demonstrated superior prediction accuracy compared to the U-Net Series and DeepLabV3+ models.

4. Conclusions

This study presented a comprehensive approach for building extraction from high-resolution satellite imagery in Loei province. The presented approach covered various stages, including data collection, application of preprocessing techniques, and utilization of diverse models for result comparison. The experimental findings demonstrated that Res-U-Net model with ResNet-50 (RGB, DSC) achieved the highest accuracy in prediction results. The extracted buildings exhibited clearer characteristics compared to DeepLabV3+ model with ResNet-50 (Enh-RGB, DSC). These predictions displayed connected buildings and wider estimated areas, resembling an expansion of the building boundaries as compared to the ground truth.

Upon examining the experimental outcomes in conjunction with the performance scores derived from various metrics, it became evident that certain key elements were instrumental in enhancing the models' efficiency and effectiveness in making predictions. A primary factor was the integration of a backbone within the model's structural architecture. Evaluations consistently demonstrated that models equipped with a backbone, such as Res-U-Net and DeepLabV3+, markedly surpassed their counterparts in performance. This study specifically revealed that the ResNet-50 backbone secured the highest F1 Score for both Res-U-Net and DeepLabV3+, underscoring the backbone's critical role in achieving superior predictive accuracy.

Furthermore, the importance of color channels as a feature for building extraction that significantly impacting the model's performance was observed. The dataset categorization, based on color attributes discerned during the preprocessing phase, revealed that models trained on RGB and Enhanced RGB datasets consistently outperformed those trained on grayscale (Gray, Enh-Gray) images. This finding underscores the critical importance of color in the segmentation process. The impact of the chosen evaluation metric on the model predictive accuracy is also noteworthy. Among the metrics evaluated, the Dice Similarity Coefficient (DSC) demonstrated superior performance over Intersection-over-Union (IoU). However, from a research and development perspective, employing both metrics simultaneously is advisable when resource constraints are not a primary concern, as they generally align in indicating performance trends. In scenarios where resource limitations are a factor, prioritizing the DSC metric for image segmentation tasks has been shown to enhance model outcomes.

Considering the points and factors previously discussed, and upon analyzing the highest F1 scores and DSC, it can be concluded that Res-U-Net with ResNet-50 offers the most accurate predictions. However, when a balance between precision and recall was considered, DeepLabV3+ with ResNet-50 emerged as the superior performer. The study of this model, trained on an Enhanced RGB dataset, confirmed that preprocessing significantly improved predictive accuracy, leading to more effective image segmentation. Nonetheless, caution is warranted as preprocessing images from diverse areas beyond the initial dataset may introduce variability due to temporal, weather, and lighting conditions inherent in satellite imagery. The research area, characterized by rural and suburban settings, included extensive forested regions. The model's ability to precisely identify non-structural elements contributed to an elevated accuracy score.

In addition to dataset characteristics, it is imperative to consider the architectural complexity of models and their training duration. Notably, models with more intricate structures demand extended training times. Despite U-Net having the simplest architecture and the shortest training duration among the evaluated models, its performance remained competitively close to that of other model types. Therefore, U-Net is an excellent option for

conducting proof of concept on unfamiliar datasets to test specific hypotheses. Meanwhile, models that integrate backbone architectures, while requiring longer training periods, tend to achieve more effective results. For future development and advancements, there are several noteworthy aspects to consider for stages such as preprocessing and postprocessing, to enhance the completeness of images and predictions.

5. Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., Kudlur, M., Levenberg, J., Monga, R., Moore, S., Murray, D.G., Steiner, B., Tucker, P.A., Vasudevan, V., Warden, P., Wicke, M., Yu, Y., & Zhang, X. (2016). TensorFlow: A system for large-scale machine learning. In *Proceedings of the 12th USENIX symposium on operating systems design and implementation* (pp. 265-283). USENIX Association. https://doi.org/10.48550/arXiv.1605.08695

Alom, M. Z., Yakopcic, C., Hasan, M., Taha, T. M., & Asari, V. K. (2019). Recurrent residual U-Net for medical image segmentation. *Journal of Medical Imaging*, 6(1). Article 014006. https://doi.org/10.1117/1.JMI.6.1.014006

Alsabhan, W., & Alotaiby, T. (2022). Automatic building extraction on satellite images using Unet and ResNet50. *Computational Intelligence and Neuroscience*, 2022(1), Article 5008854. https://doi.org/10.1155/2022/5008854

Alsabhan, W., Alotaiby, T., & Dudin, B. (2022). Detecting buildings and nonbuildings from satellite images using U-Net. *Computational Intelligence and Neuroscience*, 2022(1), Article 4831223. https://doi.org/10.1155/2022/4831223

Aslantaş, N., Bayram, B., Bakirman, T. (2021). Building segmentation from VHR aerial imagery using DeepLabv3+ architecture. In *Proceedings of the 42nd Asian conference on remote sensing* (pp. 135-143). Asian Association on Remote Sensing.

Bakirman, T., Komurcu, I., & Sertel, E. (2022). Comparative analysis of deep learning based building extraction methods with the new VHR Istanbul dataset. *Expert Systems with Applications*, 202, Article 117346. https://doi.org/10.1016/j.eswa.2022.117346

Boyle, S. A., Kennedy, C. M., Torres, J. Colman, K., Pérez-Estigarribia, P. E., & de la Sancha, N. U. (2014). High-resolution satellite imagery is an important yet underutilized resource in conservation biology. *PLoS One*, 9(1), Article e86908. https://doi.org/10.1371/journal.pone.0086908

Chen, LC., Zhu, Y., Papandreou, G., Schroff, F., Adam, H. (2018). Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. In Proceedings editors Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.). In *Proceeding of the 15th European conference on computer vision* (pp. 833-851). Springer. https://doi.org/10.1007/978-3-030-01234-2_49

Chen, S., Ogawa, Y., Zhao, C., & Sekimoto, Y. (2023). Large-scale individual building extraction from open-source satellite imagery via super-resolution-based instance segmentation approach. *ISPRS Journal of Photogrammetry and Remote Sensing*. 195, 129-152. https://doi.org/10.1016/j.isprsjprs.2022.11.006

Chhor, G., Aramburu, C. B., & Bougdal-Lambert, I. (2017). *Satellite image segmentation for building detection using U-Net*. http://cs229.stanford.edu/proj2017/final-reports/5243715.pdf

Daranagama, S., & Witayangkurn, A. (2021). Automatic building detection with polygonizing and attribute extraction from high-resolution images. *ISPRS International Journal of Geo-Information*, 10(9), Article 606. https://doi.org/10.3390/ijgi10090606

Diakogiannis, F. I., Waldner, F., Caccetta, P., & Wu, C. (2020). ResUNet-a: A deep learning framework for semantic segmentation of remotely sensed data. *ISPRS Journal of Photogrammetry and Remote Sensing*, 162, 94-114. https://doi.org/10.1016/j.isprsjprs.2020.01.013

Dice, L. R. (1945). Measures of the amount of ecologic association between species. *Ecology*, 26(3), 297-302. https://doi.org/10.2307/1932409

Gavankar, N. L., & Ghosh, S. K. (2018). Automatic building footprint extraction from high-resolution satellite image using mathematical morphology. *European Journal of Remote Sensing*, 51(1), 182-193. https://doi.org/10.1080/22797254.2017.1416676

Gedraite, E. S., & Hadad, M. (2011). Investigation on the effect of a Gaussian Blur in image filtering and segmentation. In *Proceedings of the 53rd international symposium on electronics in marine* (pp. 393-396). IEEE.

GIS English. (2023). *SAS Planet*. https://gisenglish.geojamal.com/2018/06/download-sas-planet-nightly-all.html

Google. (2022). *Google maps/google earth additional terms of service*. https://www.google.com/help/terms_maps/?hl=en-US

Google. (n.d.). *Google colaboratory*. https://colab.google/

Han, J., Wang, Z., Wang, Y., & Hou, W. (2022). Building extraction algorithm from remote sensing images based on improved DeepLabv3+ network. *Journal of Physics: Conference Series*, 2303(1), Article 012010. https://doi.org/10.1088/1742-6596/2303/1/012010

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of 2016 IEEE conference on computer vision and pattern recognition* (pp. 770-778). IEEE. https://doi.ieeecomputersociety.org/10.1109/CVPR.2016.90

International Commission on Illumination. (2012). *CIE 15: Technical report: Colorimetry (3rd edition)*. https://archive.org/details/gov.law.cie.15.2004/page/n1/mode/2up

Ivanovsky, L., Khryashchev, V., Pavlov, V., & Ostrovskaya, A. (2019). Building detection on aerial images using U-NET neural networks. In *Proceedings of the 24th Conference of Open Innovations Association* (pp. 116-122). IEEE. https://doi.org/10.23919/FRUCT.2019.8711930

Jaccard, P. (1912). The distribution of the flora in the Alpine zone.1. *New Phytologist*, 11(2), 37-50. https://doi.org/10.1111/j.1469-8137.1912.tb05611.x

Kohavi, R. & Provost, F. (1998). Glossary of terms. Machine learning—special issue on applications of machine learning and the knowledge discovery process. *Machine Learning*, 30, 271-274. https://doi.org/10.1023/A:1017181826899

Rosentrater, K. A., & Evers, A. D. (2018). Flour treatments, applications, quality, storage and transport. In K. A. Rosentrater, & A. D. Evers (eds.). *Kent's technology of cereals* (5th ed., pp. 515-564). Woodhead Publishing. https://doi.org/10.1016/B978-0-08-100529-3.00007-4

Li, Z., Zheng, J., Zhu, Z., Yao, W., & Wu, S. (2014). Weighted guided image filtering. *IEEE Transactions on Image processing*, 24(1), 120-129. https://doi.org/10.1109/TIP.2014.2371234

Lin, B. S., Michael, K., Kalra, S., & Tizhoosh, H. R. (2017). Skin lesion segmentation: U-nets versus clustering. In *Proceeding of the 2017 IEEE symposium series on computational intelligence* (pp. 1-7). IEEE. https://doi.org/10.1109/SSCI.2017.8280804

Liu, M., Fu, B., Xie, S., He, H., Lan, F., Li, Y., Lou, P. & Fan, D. (2021). Comparison of multi-source satellite images for classifying marsh vegetation using DeepLabV3 Plus deep learning algorithm. *Ecological Indicators*, 125, Article 107562. https://doi.org/10.1016/j.ecolind.2021.107562

McCarthy, M. J., & Halls, J. N. (2014). Habitat mapping and change assessment of coastal environments: an examination of WorldView-2, QuickBird, and IKONOS satellite imagery and airborne LiDAR for mapping barrier island habitats. *ISPRS International Journal of Geo-Information*, 3(1), 297-325. https://doi.org/10.3390/ijgi3010297

Namdeo, A., & Bhadoriya, S. S. (2016). A review on image enhancement techniques with its advantages and disadvantages. *International Journal for Science and Advance Research in Technology*, 2(5), 171-182.

Ronneberger, O., Fischer, P., Brox, T. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. In N. Navab, J. Hornegger, W. Wells, & A. Frangi (eds.). In *Proceeding of 18th medical image computing and computer-assisted intervention* (pp. 234-241). Springer. https://doi.org/10.1007/978-3-319-24574-4_28

Sariturk, B., & Seker, D. Z. (2022). A residual-inception U-Net (RIU-Net) approach and comparisons with U-Shaped CNN and transformer models for building segmentation from high-resolution satellite images. *Sensors*, 22(19), Article 7624. https://doi.org/10.3390/s22197624

Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. *Journal of Big Data*, 6(1), 1-48. https://doi.org/10.1186/s40537-019-0197-0

Sørensen, T. (1948). A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on Danish commons. *Biologiske Skrifter*, 5, 1-34.

Vidhya, G. R., & Ramesh, H. (2017). Effectiveness of contrast limited adaptive histogram equalization technique on multispectral satellite imagery. In *Proceedings of the 17th international conference on video and image processing* (pp. 234-239). Association for Computing Machinery. https://doi.org/10.1145/3177404.3177409

Wen, Q., Jiang, K., Wang, W., Liu, Q., Guo, Q., Li, L., & Wang, P. (2019). Automatic building extraction from Google Earth images under complex backgrounds based on deep instance segmentation network. *Sensors*, 19(2), Article 333. https://doi.org/10.3390/s19020333

Wu, Y., Wang, G., Wang, Z., Wang, H., & Li, Y. (2022). DI-Unet: Dimensional interaction self-attention for medical image segmentation. *Biomedical Signal Processing and Control*, 78, Article 103896. https://doi.org/10.1016/j.bspc.2022.103896

Xu, Y., Wu, L., Xie, Z., & Chen, Z. (2018). Building extraction in very high resolution remote sensing imagery using deep learning and guided filters. *Remote Sensing*, 10(1), Article 144. https://doi.org/10.3390/rs10010144

Zhang, L., Dong, R., Yuan, S., Li, W., Zheng, J., & Fu, H. (2021). Making low-resolution satellite images reborn: a deep learning approach for super-resolution building extraction. *Remote Sensing*, 13(15), Article 2872. https://doi.org/10.3390/rs13152872

Zhang, Y., Chen, W., Chen, Y., & Tang, X. (2018). A post-processing method to improve the white matter hyperintensity segmentation accuracy for randomly-initialized U-net. In *Proceeding of the 23rd international conference on digital signal processing* (pp. 1-5). IEEE. https://doi.org/10.1109/ICDSP.2018.8631858

Zhou, Z., Siddiquee, M. M. R., Tajbakhsh, N., & Liang, J. (2018). UNet++: A Nested U-net architecture for medical image segmentation. In *Proceedings of deep learning in medical image analysis and multimodal learning for clinical decision support* (pp. 3-11). Springer. https://doi.org/10.1007/978-3-030-00889-5_1