Lecture Notes in Networks and Systems 1782

Nhu-Ngoc Dao
Quang-Dung Pham
Hong Anh Le
Tran Ngoc Thinh *Editors*

Intelligent Aerial Access and Applications Towards 6G and Beyond

International Conference, IAAA 2025, Hanoi, Vietnam, July 16–18, 2025, Proceedings

Springer

Nhu-Ngoc Dao · Quang-Dung Pham ·
Hong Anh Le · Tran Ngoc Thinh
Editors

Intelligent Aerial Access and Applications Towards 6G and Beyond

International Conference, IAAA 2025, Hanoi, Vietnam, July 16–18, 2025, Proceedings

Springer

Editors
Nhu-Ngoc Dao
Department of Computer Science
and Engineering
Sejong University
Seoul, Korea (Republic of)

Hong Anh Le
Faculty of Information Technology
Hanoi University of Mining and Geology
Hanoi, Vietnam

Quang-Dung Pham
Faculty of Information Technology
Vietnam National University of Agriculture
Hanoi, Vietnam

Tran Ngoc Thinh
Ho Chi Minh City University
of Technology (HCMUT)
Vietnam National University—Ho Chi
Minh City (VNU-HCM)
Ho Chi Minh City, Vietnam

ISSN 2367-3370 ISSN 2367-3389 (electronic)
Lecture Notes in Networks and Systems
ISBN 978-3-032-14934-3 ISBN 978-3-032-14935-0 (eBook)
https://doi.org/10.1007/978-3-032-14935-0

Contents

Improvement of Building Segmentation from Very High-Resolution Remote Sensing Images Through a Transfer Learning Approach with ResUnet Architecture

Pham Trung Dung, Pham Ngoc Hung, Nguyen Quang Minh, Le Duc Tinh, and Dung Nguyen

Abstract Geographical information about buildings and other spatial objects is crucial for urban management and development, especially in the context of rapid urbanization. In recent years, the use of very high-resolution remote sensing for building segmentation has attracted considerable attention, driven by advancements in deep learning techniques. However, achieving the required accuracy in segmentation tasks using deep learning requires a large, manually labeled dataset, which can vary in characteristics across different areas and regions. To address this challenge, we applied a transfer learning approach for segmentation from satellite images with a small training dataset. In this study, we examined the effectiveness of a pretrained ResUnet architecture, which integrates U-Net and ResNet models, for building segmentation using a limited number of training samples. The experimental results demonstrated that transfer learning consistently outperforms training from scratch in both accuracy and computational efficiency. Specifically, the pretrained ResNet-101 backbone led to an improvement of approximately 4.3% in Intersection over Union (IoU) and reduced execution time by half. With the ResNet-18 backbone, the model achieved a 3.3% increase in precision and a fivefold improvement in processing speed. These findings confirm that acceptable accuracy in segmenting

P. T. Dung (✉) · N. Q. Minh · L. D. Tinh · D. Nguyen
Hanoi University of Mining and Geology, Hanoi, Vietnam
e-mail: phamtrungdung@humg.edu.vn

N. Q. Minh
e-mail: nguyenquangminh@humg.edu.vn

L. D. Tinh
e-mail: leductinh@humg.edu.vn

D. Nguyen
e-mail: nguyenthimaidung@humg.edu.vn

P. N. Hung
Phenikaa University, Hanoi, Vietnam
e-mail: hung.phamngoc@phenikaa-uni.edu.vn

N.-N. Dao et al. (eds.), *Intelligent Aerial Access and Applications Towards 6G and Beyond*, Lecture Notes in Networks and Systems 1782,
https://doi.org/10.1007/978-3-032-14935-0_26

urban spatial features, such as buildings, can be achieved using transfer learning models pretrained on general-purpose datasets like ImageNet, even with a small set of training samples of very high-resolution remote sensing.

Keywords Transfer learning · Building segmentation · Unet · ResNet · Very high-resolution remote sensing images

1 Introduction

The segmentation of spatial urban objects, particularly buildings, from remote sensing (RS) data sources, including imagery captured by satellites, aircraft, and drones, has gained increasing attention in recent years. Automatically extracting these objects from RS imagery provides essential information for various applications such as urban planning, infrastructure management, disaster response, and population estimation [1]. Spatial objects such as buildings, roads, and vegetation are commonly segmented using convolutional neural network (CNN)-based architectures, including Fully Convolutional Networks (FCN), VGG, U-Net, and ResNet, with results consistently demonstrating high performance. Pixel-level object segmentation typically involves training neural networks with millions of parameters, which requires large-scale datasets comprising hundreds of thousands to millions of segmentation masks. However, compiling such extensive datasets is both time-consuming and costly. Conversely, training on insufficient data increases the risk of overfitting, where the model performs well on training data but poorly on unseen data. As a result, deep neural networks often encounter significant challenges when applied to tasks with limited training data [2].

To address the challenge of limited training data, transfer learning has emerged as an effective solution. In artificial intelligence, transfer learning is inspired by the human ability to apply knowledge acquired in one domain to related tasks in another. In the field of computer vision, convolutional neural networks (CNNs) such as VGG, DenseNet, EfficientNet, and ResNet are commonly pretrained on large-scale datasets like ImageNet, which contains millions of labeled images. Through this process, these models learn to extract rich and generalizable features. When applying a pretrained model to a new task, the early layers, responsible for capturing low-level features, are often retained, while the final layers are fine-tuned or replaced to meet the specific requirements of the target application. This approach significantly enhances performance, particularly in scenarios where annotated data is scarce [3].

The use of transfer learning models pretrained on the ImageNet dataset is widely applied across various domains. Gopalakrishnan, Khaitan [4] applied transfer learning to detect pavement cracks by modifying the VGG-16 model developed by the Visual Geometry Group. The model, originally trained on the ImageNet dataset, was adapted by removing its final fully connected classifier layers. In the medical field, transfer learning has also been widely utilized in medical imaging. For instance,

Khan and Abraham [2] fine-tuned the VGG model pretrained on ImageNet to classify a small dataset of magnetic resonance imaging (MRI) scans for the diagnosis of Alzheimer's disease. In the domain of topology segmentation, Dey, Prakash [3] proposed a transfer learning framework called UnetEdge, which integrates topological information into the feature maps. Their innovative Edge module propagates edge-level topological features alongside contextual spatial data to the final decoder layer. Experiments conducted on an Indian drone dataset reported an Intersection over Union (IoU) score of 0.702, demonstrating the effectiveness of the approach.

In transfer learning-based research for building segmentation, popular convolutional architectures such as ResNet and DenseNet are often adopted as encoder backbones due to their strong feature extraction capabilities. These encoders are typically integrated into segmentation frameworks like U-Net, where the decoder component may also incorporate similar or complementary architectures to reconstruct the segmented output. This transfer learning approach proves especially beneficial when working with limited datasets, as it enhances segmentation accuracy while significantly reducing the training time and computational resources required [5].

Building upon this approach, several studies have demonstrated notable improvements in segmentation performance through the integration of transfer learning strategies. For instance, Panboonyuen, Jitkajornwanich [6] applied a transfer learning method using a pre-trained model across various image resolutions. They developed a global convolutional network based on CNNs, incorporating additional layers and channel attention mechanisms, and achieved a 17.5% improvement in the F1 score for Landsat-8 images and a 2.5% improvement for the ISPRS dataset, outperforming the conventional U-Net architecture. Similarly, Xu, Zhang [5] introduced ResFAUnet, a building segmentation network that leverages transfer learning and multi-scale fusion, employing ResNeXt101 as the encoder backbone with pre-trained weights. This model demonstrated accuracy improvements by 1.5%, 2.1%, 2.3%, and 4.7% comparing to those obtained using SegNet, FCN, SuUNet, and U-Net, respectively. Furthermore, Cui, Chen [6] proposed DenseUNet, a transfer learning-based model built upon the U-Net architecture. It utilizes a DenseNet encoder pretrained on ImageNet and incorporates dense connections in the decoder to combine multi-scale features effectively. Their results indicated a 7.5% improvement in the kappa coefficient compared to several state-of-the-art models.

Regarding the improvement of accuracy, Panboonyuen, Jitkajornwanich [6] employed a transfer learning approach using a pre-trained model with various image resolutions. They developed a global convolutional network based on a convolutional neural network (CNN), incorporating additional layers and channel attention features. The study found that the global convolutional network outperformed the convolutional encoder-decoder network (Unet), achieving an improvement of 17.5% in the F1 score for Landsat-8 images and 2.5% for the ISPRS dataset. Xu, Zhang [7] also proposed ResFAUnet, a building segmentation network that utilizes transfer learning and multi-scale fusion to enhance accuracy for small sample sizes. In this network, ResNetXt101 has been used as the encoder backbone with pre-trained weights. This model enabled accuracy improvements of 1.5%, 2.1%, 2.3%, and 4.7% compared to those achieved using SegNet, FCN, SuUNet, and U-Net, respectively.

Furthermore, Cui, Chen [8] proposed a transfer learning model called DenseUNet, which is based on the UNet architecture. In that model, the encoder incorporates a DenseNet pre-trained on ImageNet, while the decoder utilizes dense connections to combine multiscale information at each layer. The results showed that the Dense-UNet achieves a kappa coefficient that is 7.5% higher than those of several other state-of-the-art models.

Regarding training efficiency, Neupane, Aryal [9] employed a modified U-Net model for transfer learning on a small building dataset from the City of Melbourne. Their experimental results demonstrated a reduction of 300 times and 2.5 times in the number of training parameters and training time by 2.5 times, respectively, while maintaining high segmentation precision.

Motivated by such findings, this study aims to apply pretrained models to the building segmentation task using a small dataset. Specifically, we adopt ResUNet, a hybrid architecture combining ResNet and U-Net, as the backbone of our transfer learning approach. In this model, the encoder utilizes a ResNet pretrained on the ImageNet dataset with 1000 classes, while the decoder is modified at the final layer to suit the segmentation objective. The main contributions of this work are as follows:

- A comprehensive comparison between pretrained models and models trained from scratch, with a focus on segmentation precision and computational efficiency.
- An in-depth evaluation of the effect of backbone depth in transfer learning, using various ResNet architectures, including ResNet-18, ResNet-34, ResNet-50, and ResNet-101.
- Empirical evidence demonstrating the feasibility and effectiveness of transferring knowledge from natural image domains to satellite image domains, thereby opening new possibilities for accessible and efficient transfer learning in remote sensing analysis.

The proposed model and its performance on remote sensing building segmentation samples are presented in detail in the following sections. This includes a description of the model architecture, the dataset used for evaluation, experimental results, and a discussion of the findings.

2 Methodology

2.1 Transfer Learning

Transfer learning is a widely adopted deep learning technique that enables a model trained on one task to be adapted for use on a different, but related, task. This approach is particularly effective when only a limited amount of training data is available for the target task, as it allows the model to leverage previously learned features and converge more quickly during training. As a result, models utilizing transfer learning

often achieve significantly higher accuracy compared to those trained from scratch using the same limited dataset [10].

In transfer learning, selecting an appropriate backbone for feature extraction is a critical step, as different tasks often benefit from specific architectural designs [11]. For instance, models such as MobileNet, YOLO, and SSD are better suited for object detection, while VGG, EfficientNet, and ResNet are widely used in image classification due to their proven effectiveness across various visual recognition tasks. VGG [12] is valued for its simplicity and consistent performance in classification problems, while EfficientNet [13] introduces a scalable and efficient design that balances accuracy and computational cost. ResNet [14], in particular, stands out for its deep residual learning framework, which effectively mitigates the vanishing gradient issue and enables learning from large-scale visual data. Given these advantages, this study evaluates several ResNet variants as encoder backbones, including ResNet-18 (11.7 million parameters), ResNet-34 (25.6 million), ResNet-50 (26 million), ResNet-101 (44.6 million), and ResNet-152 (230 million). These differences highlight the trade-offs between model complexity and performance—factors that are essential to consider in the context of building segmentation from remote sensing imagery.

In this study, architectures such as ResNet-18, ResNet-34, ResNet-50, and ResNet-101 were used as encoders, and their pretrained weights from ImageNet were employed instead of random initialization.

2.2 ResUnet Architecture

The proposed ResUNet models are built upon the conventional U-Net architecture [12] for image segmentation, with the *encoder-decoder* structure illustrated in Fig. 1. The encoder, commonly referred to as the backbone, is responsible for capturing the contextual information of the input image through feature extraction. It receives the input image and progressively reduces the spatial resolution (height and width) of the tensors while increasing their depth using convolutional and max pooling layers. The final layer of the encoder is known as the bottleneck, which has dimensions of $1 \times 1 \times d$, where d varies depending on the specific U-Net configuration. This bottleneck also serves as the starting point for the decoder.

The decoder is composed of convolutional and transposed convolutional layers that restore the spatial dimensions of the tensors. It incorporates skip connections by concatenating feature maps from the encoder to corresponding decoder blocks, preserving spatial details lost during downsampling. The final output layer of the decoder is a single-channel binary map generated using a sigmoid activation function, representing the presence of the targeted spatial objects.

As shown in Fig. 1, the encoder branch of the proposed model adopts one of the ResNet variants pretrained on the ImageNet dataset. The decoder consists of five blocks, each concatenated with the corresponding feature maps from the encoder through skip connections. The entire ResUNet model is trained on the building segmentation dataset to fine-tune the parameters of the pretrained ResNet encoder and

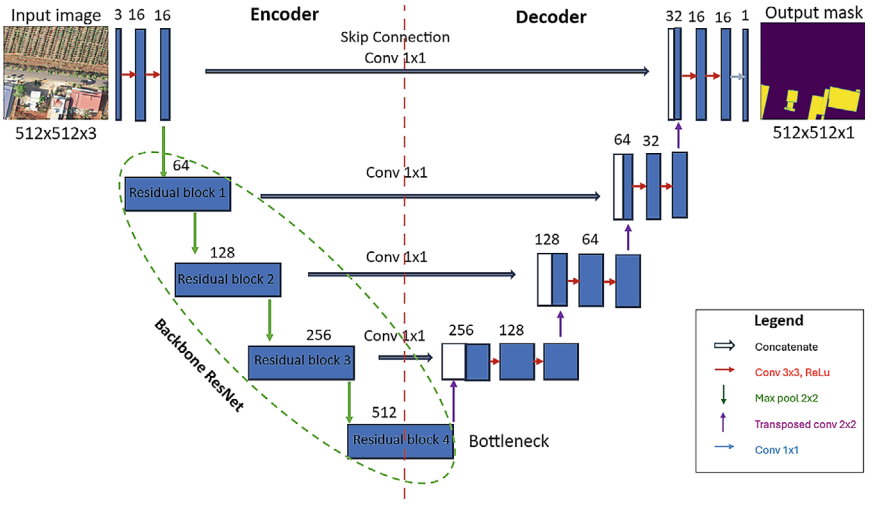


Fig. 1 The architecture of Unet with ResNet backbone

to learn the weights of the decoder layers for the segmentation task. To evaluate which ResNet backbone yields the best performance within the ResUNet architecture, different variants, including ResNet-18, ResNet-34, ResNet-50, and ResNet-101, were employed.

2.3 Architecture of Backbone ResNet

The depth of a neural network refers to the number of sequential layers in a convolutional neural network (CNN). Generally, deeper networks can achieve better performance. However, increasing the number of layers in a neural network can lead to the vanishing gradient problem. To address this issue, He, Zhang [14] proposed using shortcut connections between convolutional layers, which help overcome the vanishing gradient problem.

Formally, let (x) represent the input to the first layer, and let H(x) denote the desired underlying mapping. The key idea behind a residual neural network is the addition of a residual through a nonlinear function defined as:

$$F(x) = H(x) - x \tag{1}$$

This allows us to rewrite the original mapping as:

$$Y = H(x) = F(x) + x \tag{2}$$

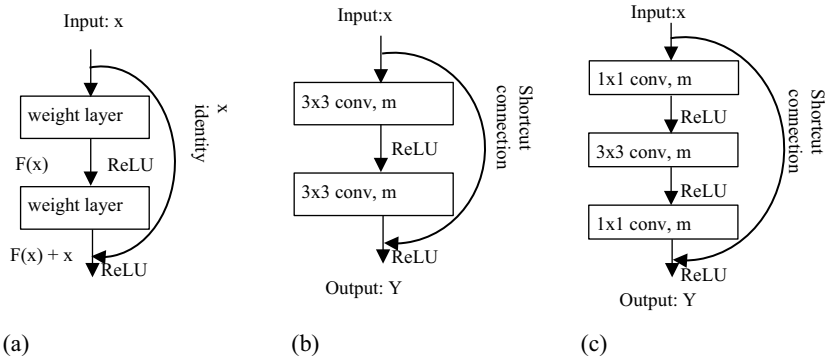


Fig. 2 Block types of the ResNet: **a** A basic building block with a shortcut connection; **b** basic residual blocks used for Resnet18 and Restnet34; **c** bottleneck residual block for Resnet-50, Resnet-101 models. m is the number of filters

where Y is the output feature map of the mapping. If the weights of F(x) are zero, then the output $Y = x$ and H(x) becomes an identity mapping (see Fig. 2a). In the context of a neural network, a shortcut connection is used to forward through identity mapping. The shortcut connection can add neither more parameters nor computational complexity.

There are five commonly used variants of the ResNet architecture, which are based on two types of residual blocks: the basic residual block (Fig. 2b) and the bottleneck residual block (Fig. 2c). ResNet-18 (18 layers) and ResNet-34 (34 layers) utilize basic residual blocks, while the deeper models, ResNet-50, ResNet-101, and ResNet-152, employ bottleneck residual blocks to improve computational efficiency and enable deeper network design.

Figure 3 illustrates the architecture of ResNet, which is designed to process input images with dimensions of 224×224 pixels. The initial convolutional block consists of four operations: a 7×7 convolution (stride 2, padding 3), followed by batch normalization, the ReLU activation function, and max pooling. The resulting feature map is then passed through four sequential ResNet blocks. These blocks are implemented as basic residual blocks in ResNet-18 and ResNet-34, and as bottleneck blocks in ResNet-50 and ResNet-101, as shown in Fig. 2b, c [15]. The network concludes with a fully connected classification head, which includes average pooling, flattening, and a linear layer to produce the final output.

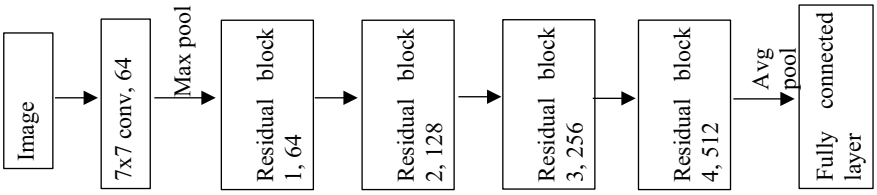


Fig. 3 The architecture of ResNet

3 Experiments

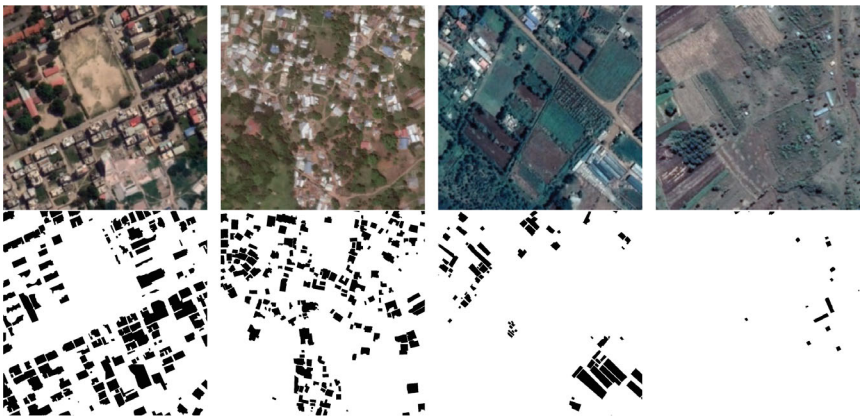
3.1 Data

The dataset used in this study is sourced from publicly available data on the IEEE DataPort platform, accessible at https://ieee-dataport.org/open-access/dataset-detecting-buildings-containers-and-cranes-satellite-images. It includes building annotations derived from satellite imagery provided through Google Earth. The annotations are encoded in binary format, where a value of 1 represents building pixels and a value of 0 denotes non-building areas. The annotated masks are stored in PNG format, while the corresponding image tiles are in JPG format, both of which are widely supported and compatible with popular machine learning and deep learning frameworks, facilitating model training and evaluation [16].

This study utilizes a building dataset from Kenya, comprising a total of 2,991 image tiles along with their corresponding annotations. The input image tiles are RGB (Red, Green, and Blue channels) photographs with a size of 1024 × 1024 pixels and a spatial resolution of 0.298 m. Accordingly, each tile covers an area of approximately 305 × 305 m, and the entire survey region spans about 280 km^2. The building annotations were manually created based on vector data obtained from a topographic database. For the deep learning task, the dataset was divided into training and testing subsets using an 80:20 ratio. The training set includes 2395 image tiles, while the test set consists of 598 tiles.

Figure 4 presents several sample image tiles and their corresponding annotations. The building samples in the study area are categorized based on their structural characteristics and built-up density. Specifically, Fig. 4a, b illustrate "high-density built-up areas" (with a built-up area greater than 20% [17]) across different settings. In urban areas (Fig. 4a), buildings are typically moderate in size, with square or L-shaped designs arranged in a well-structured manner. In suburban areas (Fig. 4b), the buildings are generally smaller and distributed more irregularly.

In rural regions, building density varies considerably. Figure 4c shows a "medium-density built-up area" (with a built-up ratio between 10 and 20% [17]), often characterized by residential houses situated near agricultural fields. Figure 4d depicts a "low-density built-up area" (with a built-up area below 10% [17]), consisting of scattered, detached houses. Notably, a large portion of the dataset falls within the medium- and low-density rural categories.



(a) high-density built-up urban area (b) high-density built-up suburban area (c) medium-density built-up rural area (d) low-density built-up rural area

Fig. 4 Categories of the buildings in urban, suburban and rural areas

3.2 Experiment Frameworks

The experiments were conducted using Google Colaboratory, which provide access to the CUDA parallel computing platform and graphical processing units (GPUs). The use of a GPU significantly reduced training time by enabling multiple computations to run in parallel. The specific GPU configuration utilized was the Tesla T4, a professional graphics card by NVIDIA equipped with 16 GB of GDDR6 memory and a 256-bit memory interface. The programming framework employed was PyTorch version 1.11.0 for deep learning.

In the experiment, two training strategies were carried out. In the first strategy, the U-Net model was trained from scratch. Due to limited memory provided by Colaboratory, batch size of 1 was selected and the networks were trained for 100 epochs using the cross-entropy loss function and stochastic gradient descent (SGD) optimizer with an initial learning rate of 0.01.

In the second strategy, the ResUnet is proposed with varying backbones ResNet. The performance in accuracy of different variants of backbones ResNet with that of the benchmark U-Net model presented in the study by Alsabhan and Alotaiby [18]. We focused on transferring knowledge from natural images in the ImageNet dataset to the Kenya building dataset. The encoder, implemented using different ResNet variants, ResNet-18, ResNet-34, ResNet-50, and ResNet-101, was initialized with pretrained weights from ImageNet and subsequently fine-tuned using the Kenya building dataset. Each new ResUnet variant was trained for 100 epochs with a batch size of 2.

3.3 *Evaluation Metrics*

Evaluation metrics play a critical role in assessing the performance of neural networks, especially in image segmentation tasks, where both classification accuracy and spatial localization must be considered. These metrics measure the similarity between the predicted segmentation outputs and the corresponding ground truth masks. Among them, the F-measure (or F1 score) is one of the most widely used evaluation metrics for segmentation performance [19].

Two key metrics based on the F measure are the Intersection over Union (IoU), also known as the Jaccard index [20], and the Dice similarity coefficient, or Sorensen Dice index [21]. These metrics are derived from the confusion matrix, which is essential for binary segmentation tasks. In a confusion matrix [22] (see Fig. 5), we find four key components: true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). TP represents the number of pixels that the model correctly predicts as ground truth. FP indicates the number of pixels incorrectly predicted as ground truth. TN indicates the number of background (negative) pixels that the model correctly predicts, and FN is the number of background pixels that the model incorrectly predicts.

The IoU is defined by the ratio of intersection and union between the predicted segmentation (P) and actual segmentation (A) as:

$$IoU = \frac{|P \cap A|}{|P \cup A|} = \frac{TP}{TP + FP + FN}. \tag{3}$$

Like the Jaccard index, the Dice similarity coefficient can be computed by:

$$Dice = \frac{2|P \cap A|}{|P| + |A|} = \frac{2TP}{2TP + FP + FN}. \tag{4}$$

The IoU and Dice range from 0 to 1. A value of "0" indicates no overlap, while a value of "1" indicates perfect overlap.

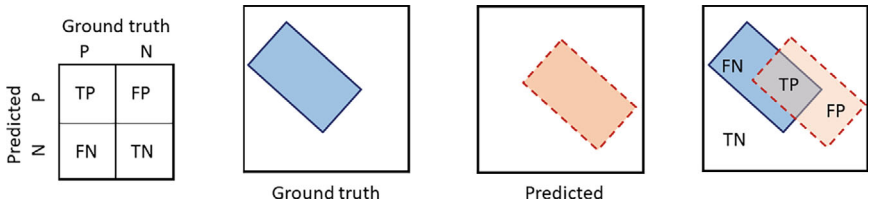


Fig. 5 Confusion matrix for image segmentation. P and N denote positive and negative, respectively

4 Results and Discussions

Table 1 presents the precision and training time for the different models. For the benchmark U-Net, the highest training accuracy reached 0.754 in terms of Intersection over Union (IoU) and 0.930 for the Dice score during the training phase. However, the accuracy dropped significantly when evaluated on the test data, with the IoU decreasing to 0.615 and the Dice score falling to 0.876.

Across all configurations, the proposed ResUNet models outperform the benchmark U-Net, particularly in terms of test accuracy. Although the IoU and Dice scores during training are relatively similar between benchmark U-net and transfer learning approaches, a significant improvement is observed in test performance with transfer learning. Specifically, the IoU scores for the transfer learning approach exceed those of the benchmark U-Net by 3.3%, 3.8%, 4.1%, and 4.3% when using ResNet-18, ResNet-34, ResNet-50, and ResNet-101 as backbones, respectively. Likewise, the Dice scores for the transfer learning models are approximately 2–3% higher than those of the benchmark U-Net across all ResNet variants. These results indicate that transfer learning effectively reduces overfitting, particularly when working with a limited number of training samples.

In terms of computational cost, transfer learning using pretrained ResNet models such as ResNet-18, ResNet-34, ResNet-50, and ResNet-101is approximately 2, 3, 4, and 5 times faster, respectively, than training the U-Net model from scratch. Transfer learning enables the model to converge more rapidly during the initial training epochs, as it begins with pretrained weights. In contrast, scratch learning starts from random initialization and often requires more training time to achieve comparable performance. This makes transfer learning particularly advantageous for tasks with limited datasets, as it enhances both training efficiency and model accuracy by leveraging knowledge from large-scale datasets.

The benefits of using the transfer learning approach are illustrated in Fig. 6, which highlights both model precision and training time. Among the ResUNet variants, the ResNet-101-based model achieves the highest precision; however, it requires the longest training time due to its depth of 101 layers. While deeper architectures improve feature extraction, they also increase computational cost. In contrast, ResNet-18 offers the shortest training time owing to its simpler structure with only 18

Table 1 Evaluation of performance and training time of the UNet and ResUnet models

Model	Backbone	Training		Test		Time
		IoU	Dice	IoU	Dice	1 epoch
Unet		0.754	0.930	0.615	0.876	38′ 00″
ResUnet	ResNet18	0.748	0.933	0.648	0.898	07′ 13″
ResUnet	ResNet34	0.752	0.934	0.653	0.898	08′ 58″
ResUnet	ResNet50	0.756	0.935	0.656	0.897	14′ 14″
ResUnet	ResNet101	0.761	0.934	0.658	0.902	19′ 15″

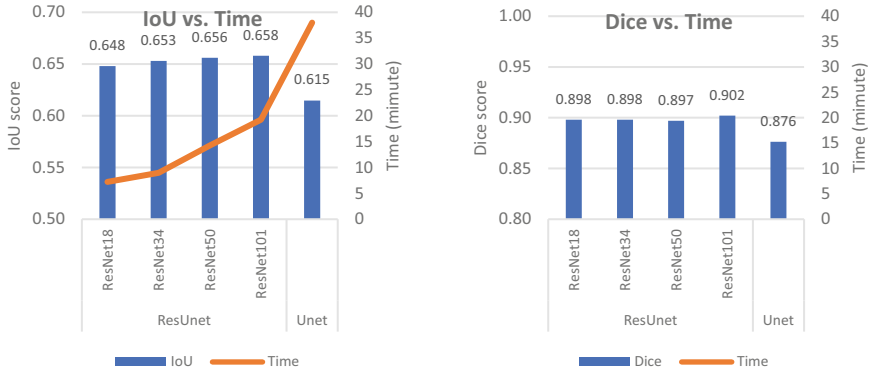


Fig. 6 The precision and execution time for scratch learning and transfer learning

layers, but it yields the lowest precision among the variants. ResNet-34 and ResNet-50 represent a balanced compromise, delivering a favorable trade-off between precision and training time. Notably, all ResUNet variants outperform the benchmark U-Net not only in terms of test accuracy but also in training time, highlighting the efficiency and effectiveness of transfer learning with pretrained ResNet backbones. These results are largely attributable to the depth of the backbone architecture, as deeper models enable the encoder subnetwork to extract more complex and abstract feature representations.

Figure 7, 8, 9, and 10 present visual evaluations of the models under four scenarios: high-density urban areas and high, medium, and low-density suburban areas. Precision is assessed by analyzing the differences between the predicted masks and the ground truth masks. The second column in each figure displays the ground truth, while the third through seventh columns show the predicted masks generated by the ResUNet models with ResNet-18, ResNet-34, ResNet-50, and ResNet-101 backbones, as well as the benchmark U-Net model, respectively. The visual assessment shows that the transfer learning approach yields more accurate segmentation results compared to the traditional method. Incorrect predictions, highlighted with red circles, are noticeably more frequent in the seventh column, which corresponds to the U-Net model, than in the ResUNet variants.

5 Conclusions

This study investigated the use of transfer learning with pre-trained ResNet models that were initially trained on general images of the ImageNet dataset to improve satellite image segmentation, especially for very high-resolution remote sensing imagery. The primary motivation for this approach is the ability to transfer knowledge from one domain to another. Our experiment results demonstrated that transfer learning

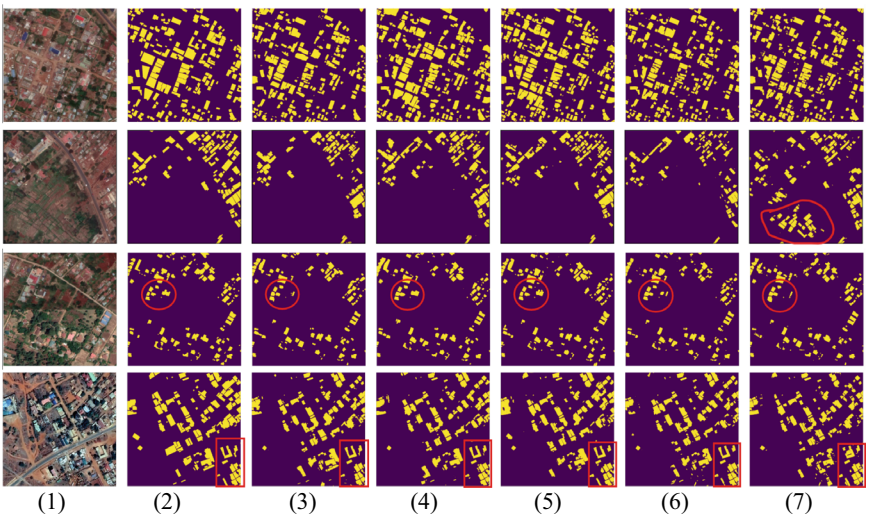


Fig. 7 Visual assessment for the high-density built-up urban area. (1) RGB ortho image; (2) the ground truth; (3), (4), (5), (6) the predicted mask by ResUnet with backbone ResNet 18, ResNet34, ResNet50, ResNet101, respectively; (7) the predicted mask by Unet without pre-trained weights

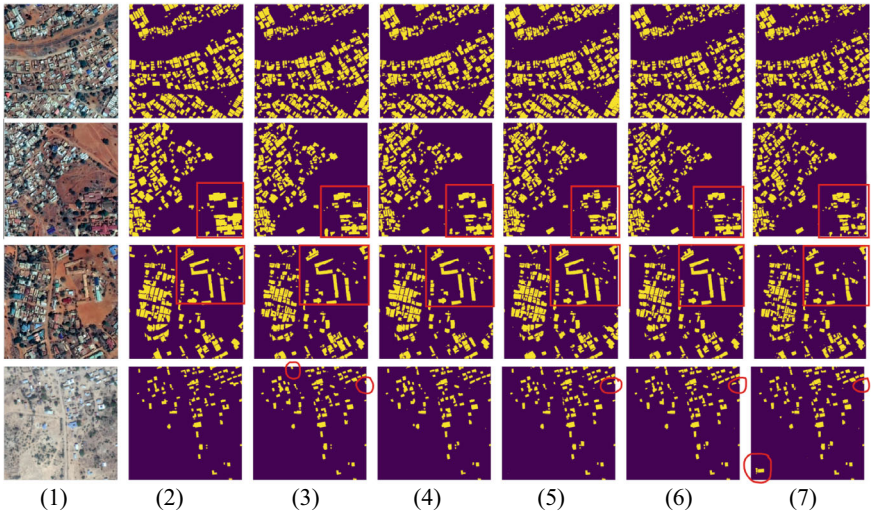


Fig. 8 Visual assessment for the high-density built-up suburban area. (1) RGB ortho image; (2) the ground truth; (3), (4), (5), (6) the predicted mask by ResUnet with backbone ResNet 18, ResNet34, ResNet50, ResNet101, respectively; (7) the predicted mask by Unet without pre-trained weights

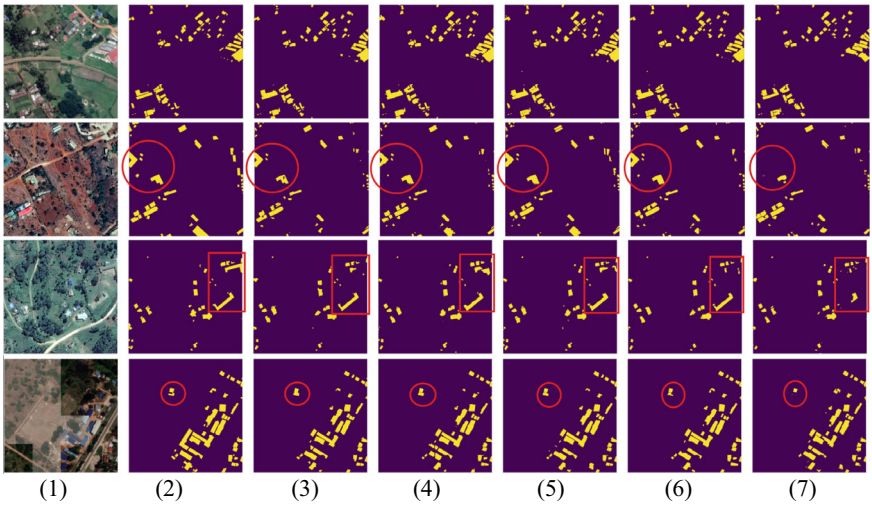


Fig. 9 Visual assessment for the medium-density built-up urban area. (1) RGB ortho image; (2) the ground truth; (3), (4), (5), (6) the predicted mask by ResUnet with backbone ResNet 18, ResNet34, ResNet50, ResNet101, respectively; (7) the predicted mask by Unet without pre-trained weights

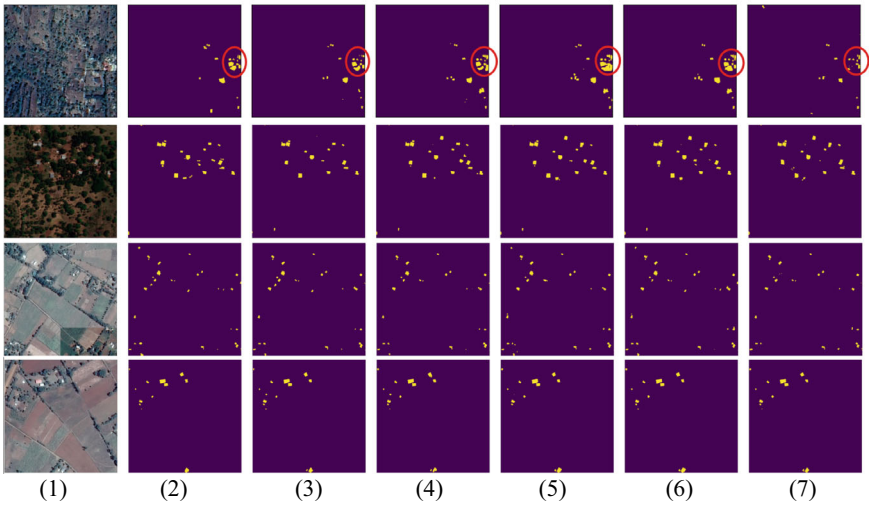


Fig. 10 Visual assessment for the low-density built-up urban area. (1) RGB ortho image; (2) the ground truth; (3), (4), (5), (6) the predicted masks by ResUnet with backbone ResNet 18, ResNet34, ResNet50, ResNet101, respectively; (7) the predicted mask by Unet without pre-trained weights

can significantly enhance both the accuracy and computational time of building segmentation from satellite imagery.

Transfer learning techniques have proven to be effective in remote sensing applications. In this study, the transfer learning approach outperformed the conventional U-Net model, achieving improvements of approximately 3–4% in Intersection over Union (IoU) and 2–3% in Dice coefficients. Additionally, transfer learning enables the model to converge more rapidly during the initial training epochs, significantly reducing the overall training time. By leveraging pretrained models from large-scale general image datasets, transfer learning not only accelerates the training process but also enhances segmentation accuracy, making it a more efficient and robust solution for tasks with limited labeled data.

The findings indicate that the depth of the backbone network significantly impacts both the precision and efficiency of the transfer learning approach. Deeper architectures generally achieve higher segmentation accuracy due to their enhanced capacity for learning complex features. However, the trade-off between improved accuracy and increased computational demands should be carefully considered, as deeper models require longer training times and greater processing power. Selecting an appropriate backbone, therefore, depends on the specific application requirements and the available computational resources.

This study used only deep learning models pretrained on widely used image datasets. Future research should explore the application of fine-tuning techniques, particularly targeting the initial or task-specific layers, to enhance the effectiveness of knowledge transfer. Moreover, while this study employed models pretrained on general image datasets, future work could benefit from leveraging models pretrained specifically on remote sensing datasets to further improve performance in domain-specific applications.

Funding This research was supported by the Vietnam Ministry of Training and Education with grant number B2024-MDA-09.

Conflicts of Interest The authors declare no conflicts of interest.

References

1. Neupane B, Horanont T, Aryal J (2021) Deep learning-based semantic segmentation of urban features in satellite images: a review and meta-analysis 13(4):808
2. Diez Y et al (2021) Deep learning in forestry using UAV-acquired RGB data: a practical review 13(14):2837
3. Ribani R, Marengoni M (2019) A survey of transfer learning for convolutional neural networks. In: 2019 32nd SIBGRAPI conference on graphics, patterns and images tutorials (SIBGRAPI-T), 2019. IEEE
4. Gopalakrishnan K et al (2017) Deep convolutional neural networks with transfer learning for computer vision-based data-driven pavement distress detection 157:322–330
5. Luo L, Li P, Yan X (2021) Deep learning-based building extraction from remote sensing images: a comprehensive review 14(23):7982

6. Panboonyuen T et al (2019) Semantic segmentation on remotely sensed images using an enhanced global convolutional network with channel attention and domain specific transfer learning 11(1):83
7. Xu X et al (2023) High-precision segmentation of buildings with small sample sizes based on transfer learning and multi-scale fusion 15(9):2436
8. Cui B, Chen X, Lu YJIA (2020) Semantic segmentation of remote sensing images using transfer learning and deep convolutional neural network with dense connection 8:116744–116755
9. Neupane B et al (2022) Building footprint segmentation using transfer learning: a case study of the city of Melbourne 10:173–179
10. Zhuang F et al (2020) A comprehensive survey on transfer learning 109(1):43–76
11. Elharrouss O et al (2024) Backbones-review: feature extractor networks for deep learning and deep reinforcement learning approaches in computer vision 53:100645
12. Simonyan K, Zisserman AJAPA (2014) Very deep convolutional networks for large-scale image recognition
13. Tan M, Le Q (2019) Efficientnet: rethinking model scaling for convolutional neural networks. In: International conference on machine learning, 2019. PMLR
14. He K et al (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition
15. Choi O et al (2020) Combustion instability monitoring through deep-learning-based classification of sequential high-speed flame images 9(5):848
16. Miyazaki H (2022) A dataset for detecting buildings, containers, and cranes in satellite images
17. Bossu A et al (2014) Effects of built landscape on taxonomic homogenization: two case studies of private gardens in the French Mediterranean 129:12–21
18. Alsabhan W, Alotaiby T (2022) Neuroscience, Automatic building extraction on satellite images using Unet and ResNet50. Comput Intell Neurosci 2022(1):5008854
19. Müller D, Soto-Rey I, Kramer FJBRN (2022) Towards a guideline for evaluation metrics in medical image segmentation 15(1):210
20. Real R, Vargas JM (1996) The probabilistic basis of Jaccard's index of similarity. Syst biol 45(3):380–385
21. Li X et al (2020) Generic sao similarity measure via extended sørensen-dice index 8:66538–66552
22. Liang JJPAC (2022) Confusion matrix: machine learning 3(4)