International Journal of Electrical and Computer Engineering (IJECE)
Vol. 14, No. 4, August 2024, pp. 4531~4541
ISSN: 2088-8708, DOI: 10.11591/ijece.v14i4.pp4531-4541

Building extraction from remote sensing imagery: advanced squeeze-and-excitation residual network based methodology

Smail Ait El Asri, Ismail Negabi, Samir El Adib, Naoufal Raissouni
Remote Sensing Systems and Telecommunications, UAE/U01 ENSATe National School for Applied Sciences, Abdelmalek Essaadi University, Tetuan, Morocco

Article Info

Article history:

Received Dec 17, 2023
Revised Mar 8, 2024
Accepted Mar 16, 2024

Keywords:

Building extraction
Deep learning
Remote sensing images
Squeeze-and-excitation residual network
Wuhan University building dataset

ABSTRACT

Extracting buildings from remote sensing imagery (RSI) is an essential task in a wide range of applications, such as urban and monitoring. Deep learning has emerged as a powerful tool for this purpose, and in this research, we propose an advanced building extraction method based on SE-ResNet18 and SE-ResNet34 architectures. These models were selected through a rigorous comparative analysis of various deep learning models, including variations of residual networks (ResNet), squeeze-and-excitation residual networks (SE-ResNet), and visual geometry group (VGG), for their high performance in all metrics and their computational efficiency. Our proposed methodology outperformed all other models under consideration by a significant margin, demonstrating its robustness and efficiency. It achieved superior results with less computational effort and time, a testament to its potential as a powerful tool for semantic segmentation tasks in remote sensing applications. An extensive comparative evaluation involving a wide range of state-of-the-art works further validated our method's effectiveness. Our method achieved an unparalleled intersection over union (IoU) score of 88.51%, indicative of its exceptional accuracy in identifying and segmenting buildings within the Wuhan University (WHU) building dataset. The overall performance of our method, which offers an excellent balance between high performance and computational efficiency, makes it a compelling choice for researchers and practitioners in the field.

Corresponding Author:

Smail Ait El Asri
Remote Sensing Systems and Telecommunications, UAE/U01 ENSATe National School for Applied Sciences, Abdelmalek Essaadi University
Tetuan, Morocco
Email: smail.aitelasri@etu.uae.ac.ma

1. INTRODUCTION

Accurate building footprint extraction from remote sensing imagery (RSI) has become increasingly critical for various applications, including urban planning, disaster management, and environmental monitoring [1]. Despite its significance, extracting building footprints from high-resolution imagery remains a complex task due to factors such as intricate backgrounds, diverse building structures, and limitations in image contrast [2]. While traditional methods offer some solutions, they often face computational limitations and struggle to adapt to the intricacies of diverse datasets. In contrast, deep learning techniques have recently emerged as promising alternatives due to their ability to automatically learn complex features from large datasets. However, deep learning techniques are not without limitations, such as the potential for overfitting, the requirement for large amounts of training data, and the inherent ambiguity in defining "true" building footprints based on real-world complexities. These limitations can hinder their ability to consistently achieve

optimal accuracy across diverse datasets, leaving room for further improvement in building footprint extraction. This study proposes a novel squeeze-and-excitation residual networks (SE-ResNet) and U-shaped convolutional neural network (U-Net) based approach aiming to address these limitations and achieve superior accuracy and efficiency compared to existing methods.

Modern smart cities make extensive use of high-resolution aerial imagery. Automatic building extraction ranks among the most prominent applications that intend to differentiate pixels belonging to buildings from those belonging to other objects in the input image. It may be considered as a classification challenge at the pixel level, which is also a semantic segmentation. In the field of remote sensing, semantic segmentation is a vital aspect of research, including tasks such as urban land cover classification [3], change detection [4], and road detection [5]. Currently, advanced remote sensing sensors have been producing a larger number of high-resolution aerial images with a much closer ground sampling distance than previously, so that the images often include an abundance of land cover information and complex environmental backgrounds (e.g., Diverse building structures and configurations, limited contrast between buildings and their surrounding object, and the existence of obstacles), making semantic segmentation tasks more difficult, particularly in urban areas. Consequently, classical methods, which are overly reliant on manually constructed features, are unable to tackle large-scale dataset challenges and fulfill the needs of today's practical applications [6].

The use of deep learning techniques and other artificial intelligence technologies has resulted in significant progress in the identification of objects within images using convolutional neural networks [7]. These methods, developed in the computer vision field, are generally employed to extract buildings using high-resolution remotely sensed images. They outperform traditional semantic segmentation techniques in the process of extracting high-dimensional features from images. Convolutional neural networks (CNNs) have undergone significant development and are now widely used in fields such as automatic road extraction [5], building extraction [8], and semantic segmentation [9]. One of the advantages of employing deep learning algorithms is that they can learn the correlative features between different ground components in the input RSI. This eliminates the human error present in conventional approaches. CNNs are able to extract hierarchical building features from RSI through automatic encoding, using the advantages of local perception and parameter sharing. This allows CNNs to learn and extract features automatically, enabling them to achieve high levels of accuracy and efficiency in tasks such as image classification and object detection. Currently, popular CNN models such as visual geometry group network (VGGNet) [10], residual network (ResNet) [11], and GoogLeNet (known also as Inception v1) [12] are widely used.

The fully convolutional network (FCN) has generated significant interest and led to the development of end-to-end deep convolutional neural networks (DCNNs). Wu *et al.* [13] introduced a multi-constrained FCN architecture that applies multiple constraints to enhance intermediate layer parameters and improve the acquisition of multi-scale features. Zhou *et al.* [14] employed mask region-based convolutional neural network (R-CNN) model to recognize buildings of various scales in their study and get significant improvements in building segmentation for the edge region. Accurately capturing spatial details in image segmentation remains a challenge. To address this problem, the process of mapping low-level features through a skip connection and decoding them at the decoding stage is utilized [15]. This helps improve the precision of the segmentation process. Skip connections bridge the gap between low-level and high-level features, enabling direct integration of low-level features into spatial resolution recovery, eliminating the need for extra parameters. In this context, Ji *et al.* [16] suggested a two-branch Siamese U-Net architecture using shared weights. The model takes both the original image and its down-sampled feature map as input. The extraction efficiency of large buildings increased dramatically after training on multisource datasets that included satellite and aerial images, raster, and vector labels. Pan *et al.* [17] used a U-Net network to detect buildings by combining spatial and channel attention methods instead of a simple connection. However, stacking a large number of convolution layers in the encoder can lead to slow convergence and reduced model performance, as well as to the gradient-vanishing problem. To address these issues, residual learning has been integrated into end-to-end DCNNs [11]. Residual learning not only speeds up model training but also enables the use of low-level features effectively [18]. In this regard, various studies have included the notion of residual learning in their architectures. Lin *et al.* [19] introduced a deep network architecture for building extraction that combined a residual block with expanded convolution. This architecture improved computational complexity but resulted in some loss of accuracy in building extraction. Xu *et al.* [20] combined two different neural network architectures, U-Net and ResNet, to extract building structures from high-resolution images obtained by remote sensing techniques. Subsequently, they employed guided filters to merge the extracted building structures to reduce noise, improve accuracy, and improve the overall performance of the results. Bai *et al.* [21] introduced an improved version of the faster R-CNN, which is a powerful object detection algorithm that has achieved leading results on several benchmarks. The authors of this last paper use faster R-CNN with dense residual blocks and region-of-interest (RoI) matching to improve building detection.

This research paper introduces a novel methodology for building extraction from RSI, leveraging the combined strengths of two highly efficient models, SE-ResNet18 and SE-ResNet34, within the robust U-Net architecture. The selection of these models was the result of a meticulous evaluation process involving 12 potential candidates, all assessed within the U-Net framework using the Wuhan University (WHU) building dataset. To further enhance the performance of our method, we employed extensive data augmentation techniques, significantly enriching the dataset, and thereby contributing to the overall improvement of our approach. By fusing multiple backbones within the U-Net architecture and incorporating comprehensive data augmentation, our method effectively circumvents the limitations of individual models, thereby achieving superior accuracy in building extraction. The practical application of this method can yield valuable insights for a range of critical areas, including urban planning, disaster management, and environmental monitoring. Ultimately, this paper presents a highly efficient and effective strategy for building extraction from RSI, marking a significant contribution to the advancement of remote sensing analysis and its multitude of practical applications.

2. METHOD

2.1. Dataset and pre-processing

In this paper, we utilize the WHU building dataset, a widely recognized dataset in the domain of building detection, which was developed by the research team at Wuhan University [16]. This dataset spans an area of over 450 square kilometers and includes high-accuracy building maps. The images in the dataset have a ground resolution of 0.3 m and feature approximately 22,000 distinct buildings in the Christchurch area. The dataset is provided in both shapefile format and rasterized data of buildings. The aerial image dataset comprises 8,188 high-resolution RSIs, each with dimensions of 512×512 pixels. Given the large size of the original images, we divided them into smaller segments of 256×256 pixels for compatibility with our deep learning model. This process yielded 22,928 training samples, 4,912 validation samples, and 4,912 testing samples.

The WHU building dataset is a challenging dataset that meets the training requirements of deep learning samples [22]. It has been compared with other datasets such as the Massachusetts Buildings Dataset (MBD), the Aerial image segmentation dataset (AISD), and the Wuhan University aerial remote sensing images dataset [23], and has proven to be a valuable resource for building detection research. Overall, the WHU building dataset plays a crucial role in advancing research and development in the field of building detection. An example from the samples of WHU aerial building dataset is presented in Figure 1.

Figure 1. Samples from WHU aerial building dataset

Deep convolutional neural networks require large amounts of training data, which are not always available during the learning phase. Data augmentation is essential to teach the network the desired invariance and robustness properties and to avoid over-fitting when only a few training samples are available [15]. To improve the diversity of our training dataset and boost our model generalization capabilities, we perform a series of augmentation operations using the Albumentations library. Our augmentation pipeline consisted of various transformations which are presented in Table 1.

2.2. Proposed method

In this research, we propose an enhanced approach for the identification of buildings from RSI. This approach employs a combination of ensemble learning techniques and convolutional neural networks, resulting in a robust and efficient model for our specific task. The architecture of our model integrates the squeeze-and-excitation (SE) mechanism, ResNet, and the UNet framework. These advanced methodologies collectively enhance the performance of our model.

Table 1. Image augmentation operations used in this study

Operation	Description	Probability
Horizontal flip	Flips the input image horizontally.	0.5
Shift-scale-rotate	Applies random shifting, scaling, and rotation to the input image.	1
Additive Gaussian noise	Adds additive Gaussian noise to the input image.	0.2
Perspective transform	Applies perspective transformation to the input image.	0.5
Image enhancements	Apply one of: - Sharpening, - Blurring, - Motion blur	0.9
Color transformations	Apply one of: - Contrast limited adaptive histogram equalization (CLAHE) - Random brightness adjustment - Random gamma adjustment - Random contrast adjustment	0.9

2.2.1. ResNet and SE-ResNet architectures

ResNet, also known as residual networks [11], represents a groundbreaking architecture in the realm of neural networks, specifically designed to tackle the complexities associated with training profoundly deep neural networks. The fundamental building block of this architecture is the residual block, typically composed of a series of convolutional layers and a shortcut or skip connection that circumvents these layers [18]. The input is merged with the output of the pre-activation layers to learn the residual function. A distinctive feature of ResNet, unlike traditional convolutional networks as shown in Figure 2(a), is the identity shortcut connection, depicted in Figure 2(b), which enables the unaltered input to a block to be directly forwarded to its output, thereby mitigating the vanishing gradient challenge and facilitating the training of profound neural networks [24]. In instances where the input and output dimensions are not identical, a convolutional shortcut connection is employed. This involves a convolutional layer in the shortcut connection, transforming the input to match the required dimensions. ResNet's architecture has been instrumental in enabling the training of extremely deep networks, with several variants such as ResNet-18, ResNet-34, ResNet-50, ResNet-101, and ResNet-152, where the numbers signify the depth of the network. Despite the heightened computational complexity, deeper networks typically deliver superior performance. ResNet has made a significant impact on the field of deep learning, with its architecture being widely adopted for a variety of tasks beyond image recognition, including object detection and semantic segmentation.

SE-ResNet, or squeeze-and-excitation residual networks, is an advanced variant of the original ResNet architecture, designed to enhance the representational power of the network. This architecture introduces a mechanism known as squeeze-and-excitation (SE) (depicted in Figure 2(c)) into the standard ResNet [25], which allows the network to perform dynamic channel-wise feature recalibration. The SE mechanism works by first 'squeezing' the spatial dimensions of the input to generate global descriptor statistics, and then 'exciting' or reweighting the channels based on these statistics. This process is achieved through two operations: global average pooling to produce channel-wise statistics, and a gating mechanism implemented by a sigmoid activation function to carry out the reweighting. The SE block is lightweight and can be integrated into the existing ResNet architecture with minimal computational overhead. The integration of the SE mechanism into ResNet allows the model to focus more on informative features and less on less relevant ones, thereby improving the model's performance. SE-ResNet has shown significant improvements in various computer vision tasks, including image classification, object detection, and semantic segmentation, outperforming the original ResNet architecture in many cases. Despite the increased complexity, the computational cost and the number of parameters of SE-ResNet remain comparable to those of the original ResNet, making it a powerful and efficient choice for deep learning tasks.

2.2.2. Proposed architecture

As depicted in Figure 3, the proposed architecture commences with the introduction of an input image, sized 256×256×3, into the system via a designated input layer. This image is simultaneously processed through two pre-trained models, UNet-SE-ResNet18 and UNet-SE-ResNet34, each of which generates a distinct mask from the input image. In parallel, the input image is subjected to two additional convolutional layers. The initial layer applies a 3×3 convolution, supplemented by a rectified linear unit (ReLU) activation function, while the subsequent layer implements a 1×1 convolution, also enhanced with a ReLU activation function. The masks produced by the pre-trained models and the convolutional layers are then combined along the channel dimension, resulting in a comprehensive mask representation. This combined mask is then processed through a UNet model, which employs SE-ResNet18 as the encoder, leading to the generation of the final output mask of the system.

In the architecture proposed herein, models 1 (M1) and 2 (M2), pre-trained on the WHU building dataset, are integrated in a frozen state. This implies that their weights remain unaltered during the training phase, thereby preserving their learned features. This strategy significantly expedites the training process and enhances efficiency. Within the entirety of the architecture, only model 3 (UNet-SE-ResNet 18) and two supplementary convolutional layers undergo training, further optimizing the learning process. The combined model is compiled using the Adam optimizer and the binary cross-entropy loss function.

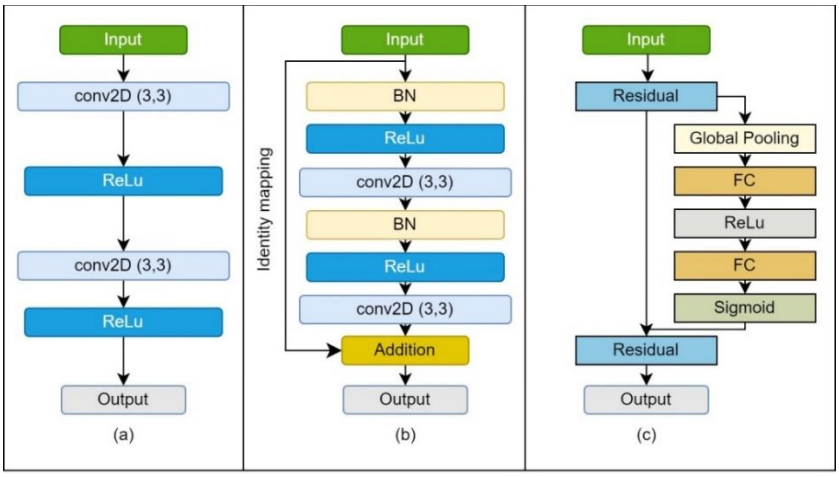

Figure 2. Comparison of building blocks: (a) regular convolutional building block; (b) residual block illustrating the skip connection for improved gradient flow; and (c) SE-ResNet block with squeeze-and-excitation mechanism for adaptive feature recalibration, enhancing representational capacity

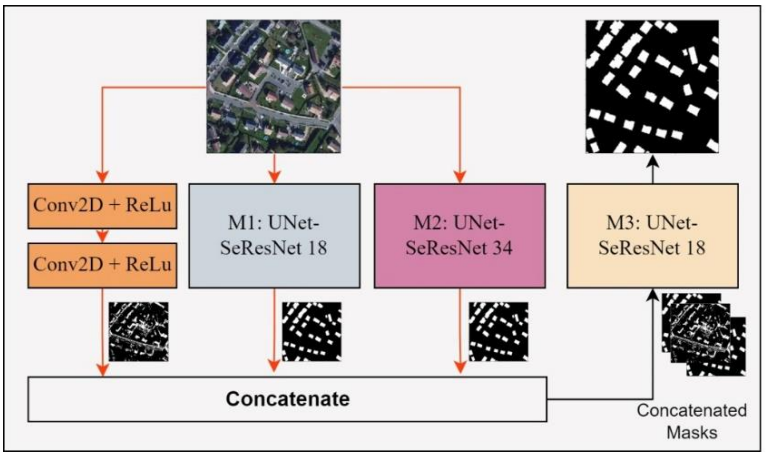

Figure 3. Schematic representation of our advanced building extraction methodology utilizing SE-ResNet18 and SE-ResNet34 architectures

3. RESULTS AND DISCUSSION

3.1. Assessment metrics

In our study, we conducted a thorough evaluation of the proposed methodology using five universally acknowledged evaluation measures: intersection over union (IoU) (1), accuracy (2), F1-score (3), precision (4), and recall (5). These metrics provided a comprehensive assessment of the performance of our approach, which leverages the power of deep learning for semantic segmentation in RSI [26].

$$IoU = \frac{target\ U\ prediction}{target\ \cap\ prediction} \quad (1)$$

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (2)$$

$$F1 - Score = \frac{2*(Recall*Precision)}{Recall+Precision} \quad (3)$$

$$Precision = \frac{TP}{TP+FP} \quad (4)$$

$$Recall = \frac{TP}{TP+FN} \quad (5)$$

where TP is the number of true positives (samples correctly identified as positive), TN is the number of true negatives (samples correctly identified as negative), FP is the number of false positives (samples incorrectly identified as positive), and FN is the number of false negatives (samples that are truly positive but have been inaccurately classified as negative).

3.2. Rationale behind the architecture design

The development of our advanced methodology for building extraction from remote sensing imagery was anchored on the critical step of identifying the most effective foundational sub-models. This pivotal selection was informed by an exhaustive comparative analysis of a spectrum of deep learning architectures, encompassing ResNet [11], SE-ResNet [25], and VGG variants [10]. Our approach entailed meticulously training these models using the WHU building dataset [16], a process designed to evaluate their performance comprehensively. The outcomes of this evaluation, which are pivotal to our model selection rationale, are systematically presented in Table 2. This table not only showcases the performance metrics of each model but also serves as the empirical basis for our subsequent selection of the optimal architectures for our proposed method.

Table 2. Results of testing baseline models on test set of WHU building dataset. Higher metric values indicate better performance, except for the number of parameters, where lower is preferable for efficiency

Model	Accuracy	Precision	Recall	F1-Score	IoU	Number of parameters
ResNet101	0.9777	0.9414	0.9096	0.9246	0.8608	51,605,466
ResNet152	0.9759	0.9228	0.9172	0.9192	0.8517	67,295,194
ResNet18	0.9752	0.9253	0.9113	0.9175	0.8488	14,340,570
ResNet34	0.9772	0.9310	0.9169	0.9229	0.8584	24,456,154
ResNet50	0.9765	0.9263	0.9182	0.9212	0.8557	32,561,114
SE-ResNet101	0.9799	0.9464	0.9188	0.9320	0.8734	56,398,273
SE-ResNet152	0.9771	0.9281	0.9199	0.9233	0.8587	73,939,329
SE-ResNet18	0.9796	0.9465	0.9173	0.9312	0.8721	14,429,650
SE-ResNet34	0.9791	0.9339	0.9271	0.9299	0.8702	24,617,350
SE-ResNet50	0.9788	0.9295	0.9296	0.9288	0.8683	35,107,201
VGG16	0.9766	0.9404	0.9029	0.9204	0.8540	23,752,273
VGG19	0.9726	0.9178	0.8992	0.9071	0.8321	29,061,969

In the evaluation of various deep learning models on the WHU-building dataset [16], a comparative analysis revealed distinct performance metrics across the board, with the SE-ResNet series consistently outperforming their ResNet counterparts. The SE-ResNet101 model demonstrated superior accuracy (0.9799), underscoring its overall effectiveness in correctly identifying building and non-building elements. Notably, the SE-ResNet18 model achieved the highest precision (0.9465), indicating its exceptional ability to correctly identify positive instances, while the SE-ResNet50 model exhibited the highest recall (0.9296), showcasing its capability to minimize missed buildings. Among these high achievers, SE-ResNet34 distinguished itself by achieving an impressive balance between precision and recall with an IoU of 0.8702, reflecting its robustness in accurately segmenting buildings. Crucially, among the high-performing SE-ResNet models, SE-ResNet18 and SE-ResNet34 stand out not only for their performance but also for their computational efficiency, boasting the lowest parameter counts. This characteristic underscore the significant advantage of these models in optimizing for both accuracy and operational efficiency, making them particularly appealing for applications where model size and computational resources are limiting factors. These findings are further corroborated by F1-Scores and IoU metrics, where SE-ResNet101 led with an F1-Score of 0.9320 and an IoU of 0.8734, indicating its proficiency in balancing precision and recall, as well as in accurately delineating building boundaries. This comparative analysis underscores the enhanced performance introduced by the squeeze-and-excitation (SE) mechanism, evident in the improved metrics of SE-ResNet models over traditional ResNet and VGG models.

The selection of SE-ResNet18 and SE-ResNet34 as the foundational architectures for our advanced building extraction methodology was justified through a systematic evaluation of performance versus

computational efficiency. Both models demonstrated exceptional balance across key performance indicators, with SE-ResNet18 notably achieving high precision (0.9465) and an impressive IoU (0.8721) with a relatively low parameter count (14,429,650). Similarly, SE-ResNet34 showcased commendable performance with an IoU of 0.8702, coupled with a moderate number of parameters (24,617,350), making these models optimal choices for applications requiring high accuracy and boundary precision without the prohibitive computational cost associated with larger models. This strategic selection leverages the benefits of the SE mechanism, optimizing for both computational efficiency and the ability to accurately segment buildings in remote sensing imagery, thus offering a compelling solution for semantic segmentation tasks in the field.

Our proposed method, primarily grounded on the SE-ResNet18 and SE-ResNet34 architectures, has exhibited notable performance, even with minimal training iterations. Specifically, it demonstrated superiority over all competing models after just a single epoch of training on the WHU building dataset (refer to Table 3). To facilitate a concise comparison of our method against others, we meticulously selected the three most proficient models based on Accuracy and IoU, recognized as the standard metrics in semantic segmentation [26]. The comparative analysis, depicted in Figure 4, succinctly illustrates the significant performance advantage of our method over the selected models.

Table 3. Results of testing our proposed method on test set of WHU-building dataset

Model	Accuracy	Precision	Recall	F1 Score	IoU	Number of parameters
Our method	0.9853	0.9490	0.9405	0.9420	0.8851	53,476,738

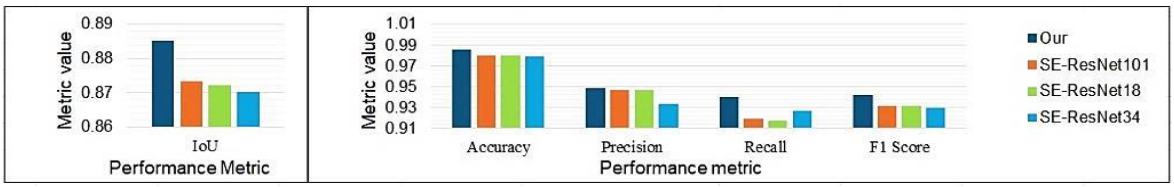

Figure 4. Comparison of our proposed method with the three most efficient models in terms of accuracy and IoU

In our rigorous evaluation of the proposed method, we meticulously compared it with the most proficient model within our trained set, SE-ResNet101. This comparison involved making predictions using both our method and SE-ResNet101 across a diverse range of demanding scenarios within the WHU building dataset. The outcomes of these predictions, meticulously presented in Figure 5, offer an in-depth and comprehensive comparison between the two models.

In Figure 5, we depict a series of demanding scenarios designed to thoroughly assess the robustness and adaptability of our method in comparison to the highly esteemed SE-ResNet101. In Figure 5(a), both models accurately predict the mask. However, the supremacy of our method becomes conspicuous when confronted with the shadow of a single wall (highlighted within the red box). While SE-ResNet101 erroneously interprets this shadow as a building, our method astutely avoids this misclassification, exemplifying its superior grasp of intricate environmental variables. Figure 5(b) introduces a more intricate scenario, featuring a diminutive building adjacent to a road, seemingly blending into the background. Within the confines of the yellow box lies a building of moderate complexity. Our method, equipped with advanced feature extraction capabilities, effectively identifies it, while SE-ResNet101 falls short. The red box signifies buildings eluding detection by both methods, underscoring the inherent challenges posed by semantic segmentation tasks. In Figure 5(c), we present buildings partially concealed by trees. The building enclosed in the yellow box, entirely overlooked by SE-ResNet101, is proficiently detected by our method, highlighting its superior handling of occlusions, a common challenge in building extraction from remote sensing data. Furthermore, our method successfully detects the building at the image's periphery (indicated by the red box), a nearly missed detail by SE-ResNet101.

In summary, our method surpasses all trained models in this paper, including the highly efficient SE-ResNet101, in building extraction from remote sensing images. Rigorous testing on the WHU building dataset demonstrates its superior performance, outperforming them in all metrics including IoU, F1-Score, and accuracy. Notably, our method achieves a fine balance between high performance and computational efficiency, making it an ideal solution for semantic segmentation tasks in remote sensing imagery and promising for real-world applications in urban planning, mapping, and environmental monitoring.

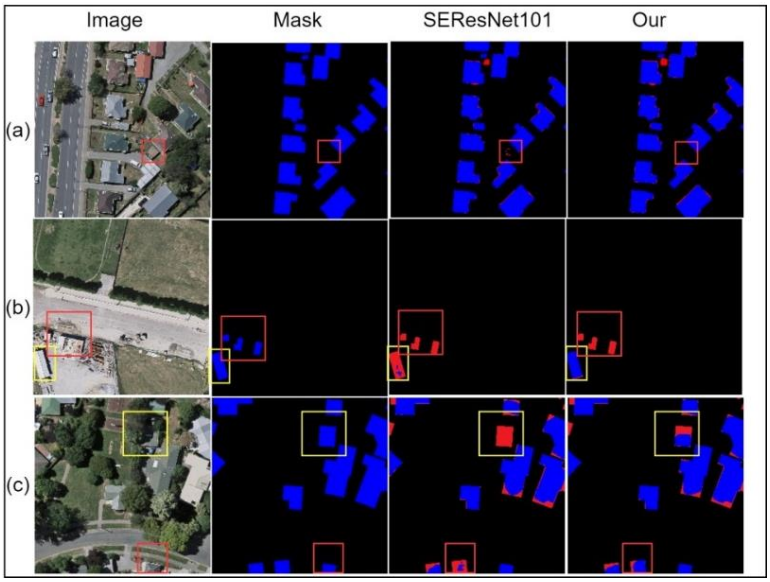

Figure 5. Comparative predictions in various scenarios: SE-ResNet101 vs. our proposed method. Error maps, indicated in red, delineate areas of discrepancy between the models' predictions. Images (a), (b), and (c) depict diverse scenarios of complexity sourced from the test set of the WHU-Building dataset

3.3. Benchmarking and future directions: our method vs state-of-the-art

Our proposed SE-ResNet based method for building footprint extraction demonstrates remarkable performance, achieving the highest intersection over union (IoU) score of 0.8851, surpassing most state-of-the-art models considered in this paper as shown in Table 4. This metric is crucial in building footprint extraction, as it measures the overlap between predicted and actual building footprints [26]. Additionally, our method achieves a recall of 0.9405, indicating its ability to identify a high proportion of actual buildings, and a precision of 0.9490, signifying a low rate of false positives (incorrectly identified buildings). These exceptional results highlight the effectiveness of our proposed architecture in accurately identifying and segmenting buildings (high IoU) while also ensuring a high proportion of true positives (high precision) and minimizing false positives (high recall) within the WHU building dataset.

Table 4. Performance of our proposed method compared to leading approaches in building footprint extraction (WHU building dataset)

Methods	Year of study	IoU	Recall	Precision
FCN [27] p[16]	2018	0.8540	0.8920	0.9530
SiU-Net p[16]		0.8840	0.9390	0.9380
U-Net [15] p[16]		0.8680	0.9450	0.9140
CU-Net [13] p[16]		0.8710	0.9170	0.9460
2-scale FCN [28] p[16]		0.7010	0.7580	0.9030
MLP [29] p[16]		0.7130	0.7850	0.8870
HRnet [30] p[23]	2023	0.7532	0.8520	0.8562
PSPNet [31] p[23]		0.6643	0.7722	0.8060
DeepLabv3+ [32] p[23]		0.7222	0.8230	0.8414
SegFormer [33] p[23]		0.7950	0.8831	0.8817
Proposed method in [23]		0.8169	0.8976	0.9246
Our method	2023	0.8851	0.9405	0.9490

p: refers to the paper where the method was implemented and tested.

Table 4 highlights our method's superior performance, achieving the highest IoU score at 0.8851, which is a critical measure of the spatial alignment between the predicted and ground truth segmentations [26]. This indicates that our method is highly accurate in identifying and segmenting buildings in the WHU dataset, even more so than the SiU-Net method, which is the second-best performer in this metric. Recall is a crucial metric in many applications as it measures the model's ability to find all the relevant cases within a dataset [26]. The fact that U-Net slightly outperforms our method in terms of recall indicates that it might be slightly better at identifying all buildings in the dataset. However, it is important to note that our method still

achieves a very high recall score and outperforms U-Net in terms of IoU and precision. FCN [16] achieves the highest precision score of 0.9530, slightly outperforming our method's precision score of 0.9490. Precision is an important metric as it measures the proportion of true positive predictions (in this case, correctly identified buildings) in all positive predictions [26]. The fact that FCN slightly outperforms our method in terms of Precision indicates that it may make fewer false positive predictions. However, it is important to note that our method still achieves a very high Precision score and outperforms FCN and all other methods in terms of IoU and Recall. This suggests that while FCN may make fewer false positive predictions, our method is better at identifying all relevant instances (higher Recall) and has a better overall match with the ground truth (higher IoU). Our proposed SE-ResNet based method's strong performance can likely be attributed to the incorporation of SE blocks, enabling focus on crucial building features, and the use of data augmentation techniques, potentially enhancing generalizability. While the creative design involving initial training of three individual models might contribute, further investigation is needed. However, limitations exist. Our evaluation is currently limited to the WHU building dataset, requiring further testing on diverse datasets to assess generalizability and robustness. Additionally, while the method demonstrates good computational efficiency, further optimization is crucial to reduce complexity and maintain high performance, especially for real-world applications. Notably, an unexpected rapid convergence to an optimal state within a single epoch emerged during training, followed by overfitting. Further investigation into potential causes, such as hyperparameter settings, data imbalance, or model complexity, and exploration of mitigation strategies like early stopping, learning rate decay, or tailored data augmentation, are crucial for future research addressing these limitations and optimizing the model for real-world scenarios.

This study aimed to develop and evaluate a novel SE-ResNet based method for building footprint extraction. Our proposed method achieved significant advancements, surpassing state-of-the-art models in terms of IoU. This highlights its effectiveness in accurately identifying and delineating buildings. While demonstrating good computational efficiency, further optimization is crucial for real-world applicability. This study contributes to the field of remote sensing image analysis, where accurate building footprint extraction holds immense significance for various applications like urban planning, disaster management, and resource management. However, limitations like evaluation on a single dataset and the unexpected rapid convergence during training warrant further investigation. Exploring these aspects and addressing potential limitations, such as improving computational efficiency and reducing model complexity, will be valuable avenues for future research.

4. CONCLUSION

This research presents a novel and efficient method for building extraction from RSI utilizing deep learning models. Through a rigorous comparative analysis of 12 state-of-the-art architectures, our method, built upon SE-ResNet18 and SE-ResNet34, demonstrably outperforms all considered models in terms of IoU. Compared to the best ResNet variant (ResNet101), our approach achieves a 2.43% higher IoU (88.51% vs 86.08%), showcasing its significant improvement in building extraction accuracy. Similarly, it surpasses the best VGG variant (VGG16) by 3.11% and even outperforms the best SE-ResNet model (SE-ResNet101) by 1.17%. This superior performance, coupled with the computational efficiency of SE-ResNet18 and SE-ResNet34 (having considerably fewer trainable parameters than SE-ResNet101), underscores the robustness and effectiveness of our proposed method, even with minimal training requirements. While SE-ResNet101 achieves slightly higher IoU (0.13% and 0.32% higher than SE-ResNet18 and SE-ResNet34, respectively), we opted for the latter two models due to their significantly lower computational complexity. The substantial difference in trainable parameters (56.4 million in SE-ResNet101 versus 14.4 million and 24.6 million in SE-ResNet18 and SE-ResNet34) translates to faster training times, lower resource requirements, and potentially wider deployment feasibility. This trade-off between minimal performance gain and substantial complexity reduction positions SE-ResNet18 and SE-ResNet34 as optimal choices for our practical and efficient building extraction solution. Furthermore, solidifying its effectiveness, our method demonstrably surpasses prominent existing works in building extraction accuracy on the WHU building dataset. Notably, it outperforms established approaches like 2-scale FCN by 18.41%, SegFormer by 9.01%, FCN by 3.11%, U-Net by 1.71%, CU-Net by 1.41%, and SiU-Net by 0.11%. This consistent and marked improvement positions our method as a valuable contribution to the field, enabling efficient and accurate building extraction for diverse applications in urban planning, land-use analysis, and environmental monitoring. Future work will concentrate on enhancing our method's performance, expanding its application to diverse remote sensing datasets and tasks, and optimizing its computational efficiency for real-world deployment. Additionally, we will investigate the unexpected rapid convergence observed during training to further enhance the model's performance.

REFERENCES

[1] S. He and W. Jiang, "Boundary-assisted learning for building extraction from optical remote sensing imagery," *Remote Sensing*, vol. 13, no. 4, pp. 1–18, 2021, doi: 10.3390/rs13040760.

[2] M. Dixit, K. Chaurasia, and V. Kumar Mishra, "Dilated-ResUnet: A novel deep learning architecture for building extraction from medium resolution multi-spectral satellite imagery," *Expert Systems with Applications*, vol. 184, 2021, doi: 10.1016/j.eswa.2021.115530.

[3] R. Goldblatt *et al.*, "Using Landsat and nighttime lights for supervised pixel-based image classification of urban land cover," *Remote Sensing of Environment*, vol. 205, pp. 253–275, 2018, doi: 10.1016/j.rse.2017.11.026.

[4] X. Lu, Y. Yuan, and X. Zheng, "Joint dictionary learning for multispectral change detection," *IEEE Transactions on Cybernetics*, vol. 47, no. 4, pp. 884–897, 2017, doi: 10.1109/TCYB.2016.2531179.

[5] G. Cheng, Y. Wang, S. Xu, H. Wang, S. Xiang, and C. Pan, "Automatic road detection and centerline extraction via cascaded end-to-end convolutional neural network," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 6, pp. 3322–3337, 2017, doi: 10.1109/TGRS.2017.2669341.

[6] Q. Hu, L. Zhen, Y. Mao, X. Zhou, and G. Zhou, "Automated building extraction using satellite remote sensing imagery," *Automation in Construction*, vol. 123, Mar. 2021, doi: 10.1016/j.autcon.2020.103509.

[7] M. Karthikeyan and T. S. Subashini, "Automated object detection of mechanical fasteners using faster region based convolutional neural networks," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 11, no. 6, pp. 5430–5437, Dec. 2021, doi: 10.11591/ijece.v11i6.pp5430-5437.

[8] S. Ait El Asri, I. Negabi, S. El Adib, and N. Raissouni, "Enhancing building extraction from remote sensing images through UNet and transfer learning," *International Journal of Computers and Applications*, vol. 45, no. 5, pp. 413–419, 2023, doi: 10.1080/1206212X.2023.2219117.

[9] A. Kherraki, S. S. Warraich, M. Maqbool, and R. El Ouazzani, "Residual balanced attention network for real-time traffic scene semantic segmentation," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 13, no. 3, pp. 3281–3289, 2023, doi: 10.11591/ijece.v13i3.pp3281-3289.

[10] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *3rd International Conference on Learning Representations*, Sep. 2014.

[11] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2016, pp. 770–778, doi: 10.1109/CVPR.2016.90.

[12] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Jun. 2015, pp. 1–9, doi: 10.1109/CVPR.2015.7298594.

[13] G. Wu *et al.*, "Automatic building segmentation of aerial imagery usingmulti-constraint fully convolutional networks," *Remote Sensing*, vol. 10, no. 3, 2018, doi: 10.3390/rs10030407.

[14] K. Zhou, Y. Chen, I. Smal, and R. Lindenbergh, "Building segmentation from airborne VHD images using mask R-CNN," in *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences - ISPRS Archives*, 2019, vol. 42, no. 2/W13, pp. 155–161, doi: 10.5194/isprs-archives-XLII-2-W13-155-2019.

[15] O. Ronneberger, P. Fischer, and T. Brox, "U-net: convolutional networks for biomedical image segmentation," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2015, vol. 9351, no. NA, pp. 234–241, doi: 10.1007/978-3-319-24574-4_28.

[16] S. Ji, S. Wei, and M. Lu, "Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 1, pp. 574–586, 2019, doi: 10.1109/TGRS.2018.2858817.

[17] X. Pan *et al.*, "Building extraction from high-resolution aerial imagery using a generative adversarial network with spatial and channel attention mechanisms," *Remote Sensing*, vol. 11, no. 8, 2019, doi: 10.3390/rs11080917.

[18] Z. Zhang, Q. Liu, and Y. Wang, "Road extraction by deep residual U-Net," *IEEE Geoscience and Remote Sensing Letters*, vol. 15, no. 5, pp. 749–753, 2018, doi: 10.1109/LGRS.2018.2802944.

[19] J. Lin, W. Jing, H. Song, and G. Chen, "Esfnet: efficient network for building extraction from high-resolution aerial images," *IEEE Access*, vol. 7, pp. 54285–54294, 2019, doi: 10.1109/ACCESS.2019.2912822.

[20] Y. Xu, L. Wu, Z. Xie, and Z. Chen, "Building extraction in very high resolution remote sensing imagery using deep learning and guided filters," *Remote Sensing*, vol. 10, no. 1, pp. 144-NA, 2018, doi: 10.3390/rs10010144.

[21] T. Bai *et al.*, "An optimized faster R-CNN method based on DRNet and RoI align for building detection in remote sensing images," *Remote Sensing*, vol. 12, no. 5, 2020, doi: 10.3390/rs12050762.

[22] J. Xue *et al.*, "Multi-feature enhanced building change detection based on semantic information guidance," *Remote Sensing*, vol. 13, no. 20, 2021, doi: 10.3390/rs13204171.

[23] M. Li *et al.*, "Method of building detection in optical remote sensing images based on SegFormer," *Sensors*, vol. 23, no. 3, Jan. 2023, doi: 10.3390/s23031258.

[24] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *European conference on computer vision*, 2016, pp. 630–645.

[25] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-excitation networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 8, pp. 2011–2023, 2020, doi: 10.1109/TPAMI.2019.2913372.

[26] S. Minaee, Y. Boykov, F. Porikli, A. Plaza, N. Kehtarnavaz, and D. Terzopoulos, "Image segmentation using deep learning: a survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 7, pp. 3523–3542, 2022, doi: 10.1109/TPAMI.2021.3059968.

[27] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2015, pp. 3431–3440, doi: 10.1109/CVPR.2015.7298965.

[28] E. Maggiori, Y. Tarabalka, G. Charpiat, and P. Alliez, "Convolutional neural networks for large-scale remote-sensing image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 2, pp. 645–657, Feb. 2017, doi: 10.1109/TGRS.2016.2612821.

[29] J. Yuan, "Learning building extraction in aerial scenes with convolutional networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 11, pp. 2793–2798, 2018, doi: 10.1109/TPAMI.2017.2750680.

[30] J. Wang *et al.*, "Deep high-resolution representation learning for visual recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 10, pp. 3349–3364, 2021, doi: 10.1109/TPAMI.2020.2983686.

[31] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, 2017, vol. 2017-Janua, no. NA, pp. 6230–6239, doi: 10.1109/CVPR.2017.660.

[32] L. C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2018, vol. 11211 LNCS, pp. 833–851, doi: 10.1007/978-3-030-01234-2_49.
[33] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "SegFormer: simple and efficient design for semantic segmentation with transformers," in *Advances in Neural Information Processing Systems*, 2021, vol. 15, pp. 12077–12090.

BIOGRAPHIES OF AUTHORS

Smail Ait El Asri received the bachelor's degree in electronics and industrial computer science from the University of Moulay Ismail (UMI), Meknes, Morocco. Holds a research master's degree in signal processing and machine learning from the National School of Applied Sciences, University of Abdelmalek Essaadi (UAE), Tetouan, Morocco, in 2017 and 2019 respectively. Currently, PhD student in mathematical-physical sciences and new technologies at the Remote Sensing Systems and Telecommunications (RSST) Laboratory, National School of Applied Sciences, University of Abdelmalek Essaadi, Tetouan, Morocco. His research interests include the design of an intelligent system for automatic detection of buildings on very high-resolution satellite remote sensing images. He can be contacted at the following email address: smail.aitelasri@etu.uae.ac.ma.

Ismail Negabi received the bachelor's degree in electronics from the University of Sidi Mohamed Ben Abdellah (USMBA), Fes, Morocco. Holds research master's degree in signal processing and machine learning from the National School of Applied Sciences, University of Abdelmalek Essaadi (UAE), Tetouan, Morocco, in 2017 and 2019, respectively. Currently, PhD student in mathematical-physical sciences and new technologies at the Remote Sensing Systems and Telecommunications (RSST) Laboratory, National School of Applied Sciences, University of Abdelmalek Essaadi, Tetouan, Morocco. His research interests include the design of intelligent crypto-systems based on deep learning modules. He can be contacted at email: ismail.negabi@etu.uae.ac.ma.

Samir El Adib received a degree in informatics, electronics, electrotechnics, and automatics (IEEA) and M.S. degree in automatic and data processing from University Abdelmalek Essaadi (UAE), Tetuan, Morocco, in 2004 and 2006 respectively. He has been a professor of physics and remote sensing at the National Engineering School for Applied Sciences of the UAE of Tetuan, since 2015. He is also heading the Remote Sensing Systems and Telecommunications (RSST) Lab at the UAE. His main research interests are FPGAs in custom-computing applications, and more concretely, applications of reconfigurable hardware to cryptography. He can be contacted at email: seladib@uae.ac.ma.

Naoufal Raissouni received a M.S., and a Ph.D. degree in physics from the University of Valencia, Spain, in 1997, and 1999, respectively. He has been a professor of physics and remote sensing at the National Engineering School for Applied Sciences of the University Abdelmalek Essaadi (UAE) of Tetuan, since 2003. He previously headed the Remote Sensing and GIS Lab in the UAE. His research interests include atmospheric correction in visible and infrared domains, the retrieval of emissivity and surface temperature from satellite image, huge remote sensing computations, mobile GIS, Adhoc networks and the development of remote sensing methods for land cover dynamic monitoring. He can be contacted at email: naoufal.raissouni.ensa@gmail.com.