Spectral and Spatial Attention Fusion for Building Segmentation in Remote Sensing Imagery

Marwa Chendeb El Rai[1], Muna Darweesh[2], Aicha Beya Far[3], Amjad Gawanmeh[2]

[1]American University in Dubai, Mathematics Division, Dubai, United Arab Emirates - melrai@aud.edu
[2]University of Dubai, College of Engineering and IT, Dubai - midarweesh@ud.ac.ae, agawanmeh@ud.ac.ae
[3] American University in Dubai, Department of Electrical and Computer Engineering, Dubai, United Arab Emirates - afar@aud.edu

Keywords: Building Segmentation, Remote sensing, Deep learning, Spatial Attention, Spectral Attention, Attention Mechanism Fusion

Abstract

The building segmentation in very high resolution remote sensing imagery presents challenges due to the need to delineate features accurately in a wide range of urban landscapes. Many existing building segmentation methods struggle in discerning complex structures and providing fine grained generalisation over different geographic regions. Additionally, these methods often require to extensive preprocessing and struggle to combine multispectral data. Addressing the different challenges, we introduce the Multi-Band Spectral-Spatial Fusion Attention Network (MBSSFA-Net), a novel method for semantic segmentation. MBSSFA-Net implements a dual encoder designed to exploit the complementary spectral information provided by the Near-Infrared and the RGB bands, to improve the feature representation and the segmentation accuracy. The approach incorporates multiscale spectral and spatial attention fusion blocks in the encoder to fuse the extracted features to enhance boundary delineation, and a spectral and spatial attention fusion blocks in the decoder to merge the spatial and abstract features. The proposed framework can extract buildings in different environment since it can process multispectral data. The experiments have been performed on GaoFen-7 and WHU-Satellite II datasets. The experiments prove that our method outperforms current state of the art Deep Learning segmentation techniques, demonstrating its potential for building segmentation in complex urban environments.

1. Introduction

Building Segmentation from remote sensing (RS) imagery is fundamental in urban planning, disaster management and environmental monitoring [J. and M., 2018] [Melgarejo and Lakes, 2014]. For instance, the accurate extraction of building lead first responders to the affected region more effectively, thus they will know where to send assistance and reallocate emergency funds towards the appropriate areas. Furthermore, in environmental monitoring, building segmentation is used to monitor urban expansion and to understand how it affects local ecosystems and resource demands [Yu and Wen, 2016]. Over the last few decades, numerous approaches based on hand crafted feature descriptors and Conventional Machine Learning classifiers have been proposed for building extraction [Turker and Koc-San, 2015] [Yang and Newsam, 2008]. As Deep Learning (DL) has advanced, automated building extraction has become popular, most notably using high resolution satellite imagery and advanced Neural Networks. However, Convolutional Neural Networks (CNNs) especially encoder decoder models such as U-Net are effective at learning hierarchical features that encode both semantic and spatial details needed to accurately delineate buildings in complex urban areas [He et al., 2021]. The availability of Very High Resolution (VHR) multispectral datasets has expanded the building segmentation capabilities. Datasets like GaoFen-7 [Chen et al., 2024] and WHU satellite-II (WHU-II) [Ji et al., 2019], which inlclude both RGB and Near-Infrared (NIR) bands, provide richer spectral information allowing for better differentiation between man-made structures and natural elements, with NIR band improving contrast for more precise segmentation. Nevertheless, integrating RGB and NIR bands is still challenging because each band has its spectral characteristics. Tailored fusion techniques are needed to exploit the joint information from different bands. To address the challenge of combining different bands or sources for building extraction, multi path encoder architectures are promising [Chen and all, 2021]. In this case, the model processes the channels separately, such that each encoder can learn the properties of its own channel before merging features. The feature extraction using this approach is enhanced with spectral specific details that lead to more accurate and robust segmentation [Ye et al., n.d.] [Liu et al., 2019]. Inspired by [Y. and all, 2018], we present a novel framework shown in Figure 1, called Multi-Band Spectral-Spatial Fusion Attention Network (MBSSFA-Net). The dual-channel encoder model is presented with integrated spectral and spatial attention mechanisms. Independent processing of RGB and NIR channels takes place through two separate encoders, due to the fact that the spectral characteristics of the object are captured effectively by each band. Feature fusion is performed at each encoder layer through Spatial and Spectral Attention Fusion Block Encoder (SSAFB-E), enhancing both spectral and spatial details. The fused features are transferred to the decoder as skip connection, where we implement a Spatial and Spectral Attention Fusion Block Decoder (SSAFB-D) to ensure the fusion of low-level spatial features with high-level abstract features. The model is evaluated against state-of-the-art segmentation models, on two benchmark building datasets GaoFen-7 and WHU satellite-II having RGB and near-infrared (NIR) bands. The Key contributions of this work include:

1. Dual Encoder Architecture for Multi-Band Data Integration: we propose a dual encoder structure, specialized for fusing multispectral data (RGB and Near-Infrared (NIR) bands) to improve feature representation and segmentation accuracy for building segmentation tasks in remote sensing imagery.

2. Incorporation of multiscale Attention Mechanisms: multiscale attention mechanisms are used to refine feature extraction, such as spectral and spatial attention. This

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-X-G-2025-181-2025

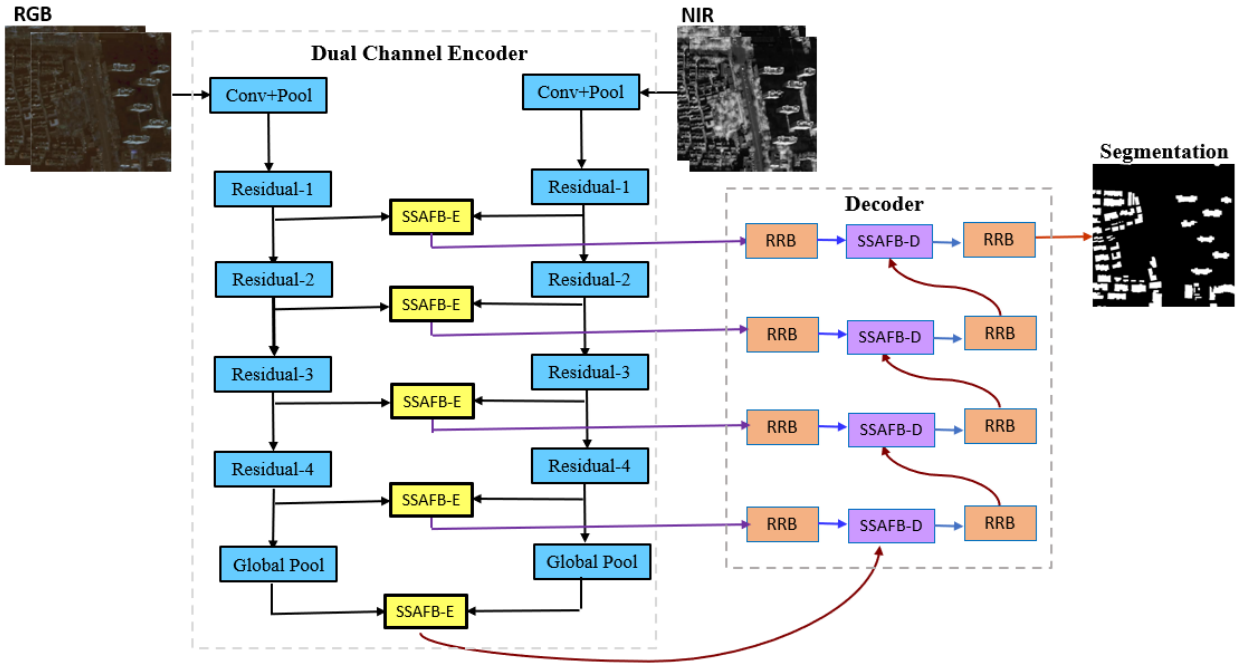

Figure 1. An overview of the Multi-Band Spectral-Spatial Fusion Attention Network (MBSSFA-Net).

approach addresses to the challenge of the boundary delineation in complex urban environments.

3. Thorough Evaluation Using Standard Metrics: The performance of the model was evaluated thoroughly on standard metrics, including Precision, Recall, Intersection over Union (IoU), F1-score on benchmark datasets GaoFen-7 and WHU-Satellite II.

This paper is organized as follows: Section 2 reviews related work, Section 3 presents the methodology, Section 4 discusses the results and analysis, and the final section concludes the paper.

2. Related Work

Conventional Machine Learning classifiers have been applied in building segmentation. [Turker and Koc-San, 2015] segment building from high-resolution optical spaceborne images using the integration of support vector machine classification, perceptual grouping and Hough transformation. [Yang and Newsam, 2008] used the scale invariant feature transform to identify objects in remote sensing images. While these methods achieved acceptable results, they rely heavily on manually crafted features and are labor-intensive to design. Moreover, these traditional techniques struggle to handle the varied and complex nature of high-resolution remote sensing imagery. With the advent of Convolutional Neural Networks (CNNs) and in particular encoder decoder architectures such as U-Net [Ronneberger et al., 2015], it became feasible and accurate for building segmentation. The efficiency of retaining spatial details in U-Net encoder-decoder structure with skip connection has been widely adopted. The authors in [Dixit et al., 2021] proposed a Dilated ResUnet, which was trained on a combined Sentinel 2 imagery and OpenStreetMap dataset. They found optimal band combinations and hyperparameters that improve building extraction accuracy.

In [Fu et al., 2019] The dual attention approach, incorporating both channel and position attention to refine feature representation and thus improve boundary delineations for buildings within complex urban settings for scene segmentation was presented. In [Hajjar et al., 2024], the authors proposed a multi-view U-Net deep model to enhance building segmentation that incorporates multiple views of the images. They used two pre-trained Convolutional Neural Network architectures, MobileNetV2 and ResNet50, to extract features representing two different views of the images. [Liu et al., 2020] used U-Net with ResNet encoder for building segmentation from satellite imagery. [Zhang et al., 2019] suggested a nested network with dense hierarchical connections for effective fusion of multi-level feature maps to recover structural details. [M. et al., 2019] used self-attention mechanisms to solve long range dependencies in building extraction from aerial imagery, [Liu et al., 2019] integrated a Pyramid Pooling Module (PSM) in UNet to aggregate multiscale contextual information effectively. Table 1 shows summary analysis of building segmentation methods in the literature. It includes the method used, type of data, the resolution of the input images, the highest achieved metrics, key features, and main limitations of each approach. In this work, we incorporate an attention fusion strategy that integrates spectral and spatial feature information. Additionally, spectral attention helps highlighting wavelength specific information for segmentation performance improvement and, spatial attention helps to refine the focus of the model on structural details. By combining spectral and spatial attention mechanisms, this method offers robust building segmentation across different remote sensing datasets, overcoming the high variability of urban environments.

3. Methodology

In this paper, we introduce a novel method to enhance the building segmentation in multispectral RS imagery. Inspired by the Discriminative Feature Network (DFN) architecture, introduced in [Y. and all, 2018], our approach operates within an

Table 1. Comprehensive Analysis of Building Segmentation Methods

Reference	Method	Data Type	Resolution	Best Metrics	Key Features	Main Limitations
[Hajjar et al., 2024]	Deep Multiview Classification	Optical	High	F1: 0.95	Sustainable urban dev. focus	High computational cost
[Liu et al., 2020]	U-Net with ResNet Encoder	Satellite Imagery	High	IoU: 0.87	Accurate segmentation	Requires large training data
[Ye et al., n.d.]	CT-UNet	Remote Sensing Images	Medium	F1: 0.90	Enhanced U-Net	Complex architecture
[Ji et al., 2019]	FCN	Multi-source Aerial	High	IoU: 0.88	Open dataset compatibility	Limited multi-source fusion
[J. and M., 2018]	Fuzzy Systems	LiDAR	Medium	Acc: 92%	Post-event detection	Sensitive to noise
[He et al., 2021]	U-Net + CRFs	Remote Sensing	High	IoU: 0.89	Enhanced boundary detection	Complex processing
[Melgarejo and Lakes, 2014]	Integrated Assessment	Public Infra.	Low	Qualitative	Climate adaptation focus	Limited data granularity
[Dixit et al., 2021]	Dilated-ResUNet	Multi-spectral Imagery	Medium	IoU: 0.86	Novel deep arch.	Limited to medium resolution
[Y. and all, 2018]	Discriminative Feature Network	Semantic Segmentation	High	F1: 0.91	Feature network	Requires high-quality data
[Chen and all, 2021]	Attention-Fused Network	VHR Imagery	Very High	IoU: 0.93	Attention mechanisms	Limited generalization
[Xu et al., 2021]	Attention Fusion Network	Multi-spectral	Medium	IoU: 0.88	Attention fusion	High memory usage
[Chen et al., 2024]	Benchmark Dataset	GaoFen-7	Very High	-	Building extraction benchmark	Dataset limitations
[Fu et al., 2019]	Dual Attention Network	Scene Segmentation	Medium	IoU: 0.92	Dual attention	Requires fine-tuning
[Yang and Newsam, 2008]	SIFT + Gabor Features	Remote Sensing Imagery	Medium	Acc: 89%	Texture-based features	Limited feature variety
[Turker and Koc-San, 2015]	SVM + Hough Transform	High-Res Imagery	High	Acc: 91%	Perceptual grouping	Limited to high-res
[Zhang et al., 2019]	Web-Net	Aerial Imageries	Very High	F1: 0.94	Ultra-hierarchical sampling	High training time
[M. et al., 2019]	Relation-Augmented FCN	Aerial Scenes	Very High	IoU: 0.90	Relation augmentation	Complex training
[Liu et al., 2019]	Encoder-Decoder with SPP	High-Res Imagery	Very High	F1: 0.92	Spatial pyramid pooling	GPU intensive
[Yu et al., 2025]	Shadow-Aware Edge Perception	Optical Imagery	High	Acc: 0.93	Shadow handling	Edge complexity
[Lee and Chang, 2024]	NPSFF-Net	High-Res Imagery	High	IoU: 0.91	Pseudo-Siamese feature fusion	Computationally intensive

encoder-decoder framework that preserves both spatial details and abstract features simultaneously. In the proposed method, a dual channel encoder is introduced to extract features from different multispectral sources to ensure the fusion of information from different spectral bands. To combine the multispectral features efficiently, we incorporate a fusion block located at the output of each layer called Spectral Spatial Attention Fusion Block Encoder (SSAFB-E) [Chen and all, 2021]. The purpose is to learn and enhance the spectral and spatial information crucial for segmentation. For the purpose of more fusion of high-level abstract features with low-level spatial details, the Spectral Spatial Attention Fusion Block Decoder (SSAFB-D) is implemented into the decoder.

The encoder gradually decreases the size of the feature maps, which reduces the computational load and increases the field of view, which enhances the abstraction of the network. As the encoder is deepened, the feature maps are less detailed and more abstract to capture more features. However, this abstraction process leads to the loss of fine spatial details which are very important when pixel-level classification is being carried out such as in the case of segmentation. The decoder is used to upsample the feature maps to the spatial dimension for pixel-level classification. This is done by integrating high abstract features with low spatial features, thus enabling the preservation of real spatial information for accurate boundary descriptions. The encoder-decoder architecture is especially useful for RS imagery acquired at high altitudes when some small structures can be missed. This architecture retains the positional and structural information of such structures, which makes it effective in detecting and segmenting even small target objects.

3.1 Discriminative Feature Network: The Baseline Architecture

In [Y. and all, 2018], the authors present a framework that is aimed to improve the semantic segmentation while focusing on intra-class inconsistency and inter-class indistinction. The

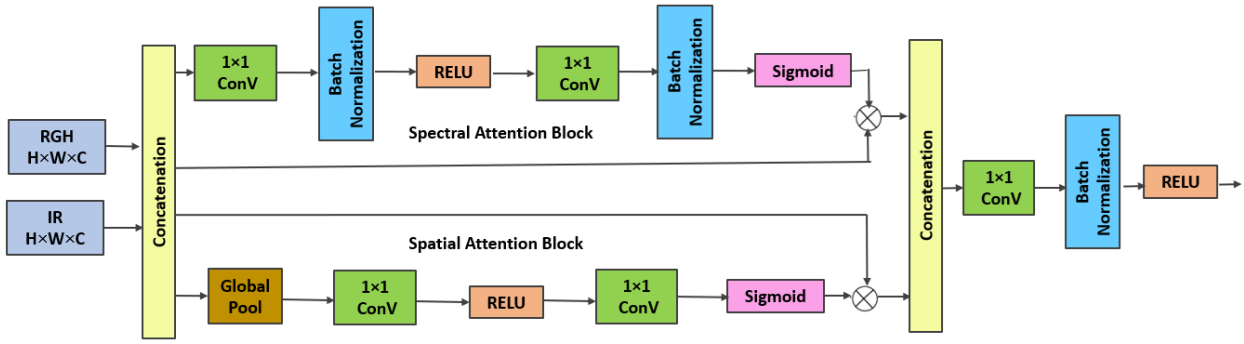

Figure 2. An overview of the Spectral Spatial Attention Fusion Block Encoder (SSAFB-E).

architecture comprises two primary components: the Smooth Network and the Border Network. Smooth Network aims at getting good consistency within the class by trying to improve the feature representation. It adopts the U-shaped encoder-decoder architecture with the consideration of multiscale contextual information. In this network, the Refinement Residual Block (RRB) [Y. and all, 2018] is used to refine the features by the residual connections which guarantee the crucial information retained while improving feature quality. Furthermore, the Channel Attention Block is integrated into the proposed network to adjust dynamic channel-wise feature response maps. Border Network tried to improve the inter-class discriminant ability by focusing on the semantically defined margins. It also uses deep supervision to control the learning process so that the features on the two sides of the boundary are becoming more salient. This approach assists the network to better define the boundaries of the objects being classified and minimizes confusion between two adjacent classes. As it is possible to improve feature representation and segment boundaries by incorporating attention mechanisms directly into the encoder-decoder network, a border network is not included in our proposed algorithm.

3.2 Dual Channel Encoder (DCE)

The original DFN model [Y. and all, 2018] was designed to take in RGB three-channel images as input. However, for certain remote sensing segmentation datasets, additional features like Near-Infrared (NIR) bands are also available, providing valuable information that can enhance segmentation performance. One of the main drawbacks of the initial DFN architecture is its inapplicability to the processing of multiple sources. To overcome this, we present DCE which is a modified encoder to replace the DFN encoder presented in [Y. and all, 2018]. The DCE uses different encoders for each input data source. In the case of building segmentation datasets, two branches are implemented: One branch is to process RGB bands, and the other for NIR bands. This architecture makes it possible to extract more discriminative features from each type of data thus improving segmentation. This implementation is able to minimize overfitting and enhance computational speed as noted by [Chen and all, 2021].

3.3 Spectral Spatial Attention Fusion Block Encoder (SSAFB-E)

For the purposes of building segmentation, it is crucial to fuse multi-spectral RGB and Near-Infrared (NIR) data to enhance the level of detail in the identification of building structures. RGB channels represent colour and texture, whereas NIR data improve the contrast of building materials and vegetation, which are significant in complex urban areas. In order to improve the feature fusion in this dual-encoder system, the Spectral-Spatial Attention Fusion Block – Encoder (SSAFB-E) is proposed in this work, as shown in Figure 4. The SSAFB-E can learn spectral and spatial weights in parallel, enabling the network to focus on the important features of the object in both RGB and NIR bands. With the help of such attention mechanisms, SSAFB-E improves the model’s performance in terms of perceiving spatial details and spectral differences crucial for the building segmentation task. In SSAFB-E, the spectral attention block aims emphasizes on the bands depending on their relation to building structures. At the same time, the spatial attention block enhances essential spatial features such as the edges and shapes of buildings, enhancing the model’s output of building-related features. Spectral attention guarantees that the multi-spectral data are well integrated within the network while spatial attention enhances building detection across different urban scenes.

3.4 Decoder Spectral Spatial Attention Fusion Block Decoder (SSAFB-D)

The Spectral-Spatial Attention Fusion Block Decoder intends to improve feature representation and fusion in multi-spectral building segmentation tasks. SSAFB-D operates on the high-level features that gains semantic information and the low-level features that preserves spatial details. Thereafter, the low-level and high-level features are combined along the channel dimension to form the fusion process. The concatenation of these two streams preserves more information from lower layers, where finer spatial details are encoded, and from deeper layers, where semantic features are extracted, which is crucial for identifying subtle patterns in RGB and NIR urban scenes. To enhance the feature fusion, SSAFB-D uses two forms of attention mechanisms in its structure. The spectral attention mechanism provides different amount of attention to each channel depending on its importance. By learning channel weight vector, SSAFB-D is able to highlight channels that contain important spectral information about building material, vegetation and other relevant components especially from NIR and RGB bands. The spatial attention module on the other hand addresses spatial patterns by assigning certain weights to the spatial locations in the feature maps. This mechanism makes it easier to reveal the spatial distribution and size of the buildings as well as the degree of isolation from other different areas [Xu et al., 2021].

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-X-G-2025-181-2025 | © Author(s) 2025. CC BY 4.0 License.

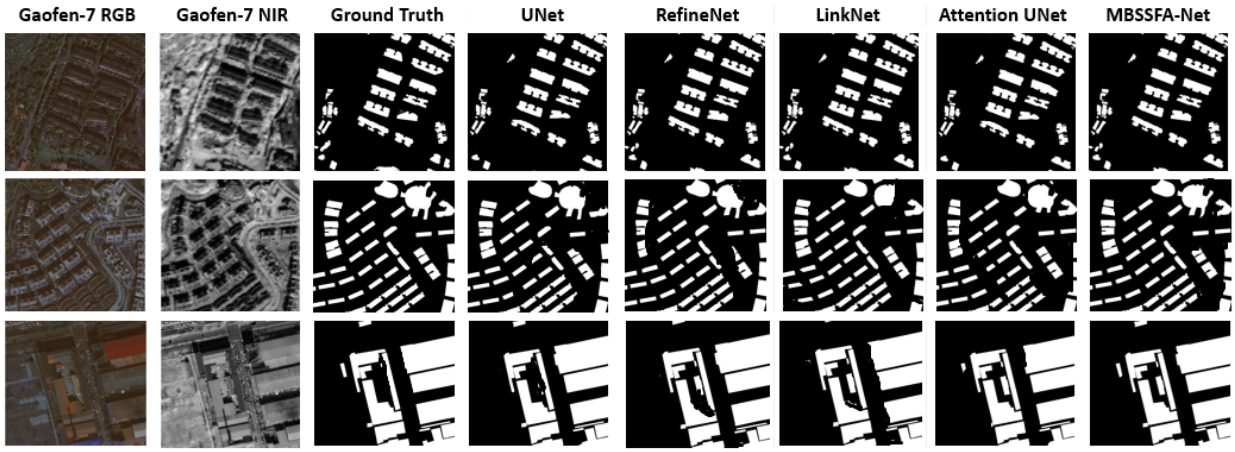

Figure 3. Semantic segmentation results of different DL methods on representative images from the RGB and NIR GaoFen-7 test set.

4. Results and Analysis

4.1 Training Settings and Evaluation Metrics

For training and testing the proposed model, two publicly datasets GaoFen-7 [Chen et al., 2024] and WHU satellite-II [Ji et al., 2019] were used. The training was done on an NVIDIA RTX 2070 GPU with 8 GB VRAM. The Adam optimizer was used with learning rate of 0.0001 with 150 epochs and batch size of 2. Model performance was assessed on the test set using four evaluation metrics: Precision, Recall, F1-Score and Intersection of Union (IoU) [Chen et al., 2024]. Precision measures the accuracy of the building class and it looks at the correct predictions out of all of all positive predictions. Recall refers to the ability of a model to correctly identify positive classes from a dataset. It is a measure of how well a model can find all the relevant classes. F1-score is the combination of Precision and Recall with equal weight or what could be referred to as the harmonic mean. IoU measures the overlap of the predicted and ground truth segments.

4.2 Datasets

4.2.1 GaoFen-7 Dataset Constructed by [Chen et al., 2024], GaoFen-7 Building dataset includes ground-truth building labels and four multispectral channels RGB and NIR, all with a resolution of 0.65 m. These images were derived from six GaoFen-7 acquisitions that encompass both urban and rural areas in Tianjin, Lanzhou, Chongqing, Ningbo, Guangzhou, and Shenzhen. These cities represent a mix of coastal and inland regions in China, showcasing a range of built environments, urban layouts, and topographical characteristics. For example, the building distribution in Tianjin, Lanzhou, Chongqing, and Ningbo tends to be more regular, whereas in Guangzhou and Shenzhen, the buildings are densely clustered with a variety of styles and heights. The GaoFen-7 Building dataset offers labels of excellent quality and accuracy, mainly created through manual digitizing. Furthermore, it has multiple building sizes, types and heights. The building sizing in GaoFen-7 Building dataset spans a range from 3 m^2. to 1,480,000 m^2 and are mainly concentrated within 1,000 m^2. The GaoFen-7 Building dataset comprises 5,175 pairs of 512 × 512 image tiles. To facilitate model training and evaluation, the dataset was methodically divided into training, validation, and test subsets using a 6:2:2 split ratio. These reflections are applied in a way that the subsets are formed intentionally and they contain a variety of buildings to consider multiple scenarios. Thus, the final dataset contains 3,106 image tile pairs for training set, 1,034 pairs for validation set and 1,035 pairs for testing set [Chen et al., 2024].

4.2.2 WHU Satellite Dataset II WHU Satellite Dataset II has a coverage area of 550 km² in East Asia with ground resolution of 2.7 m [Ji et al., 2019] [Chen et al., 2022]. The dataset contains 29,085 building instances, cropped into 17,388 patches of size 512 × 512 pixels. Of these, 13,662 patches are used for training, while the remaining 3,726 patches are used for testing. The data is acquired from multiple high-resolution satellite sensors, including QuickBird, WorldView series, IKONOS, and ZY-3. WHU-II offers a pool of images to evaluate the ability of Deep Learning Models in generalizing building extraction on large scale building types within a single region. The annotation is performed using ArcGIS [Ji et al., 2019].

4.3 Experimental Results

4.3.1 GaoFen-7 Dataset The GaoFen-7 Building dataset was tested with various state-of-the-art Deep Learning models including FCN 8S, SegNet, UNet, RefineNet, LinkNet, Attention UNet, and High Resolution Network (HRNet) [Chen et al., 2024]. Table 2 reports the results of different metrics: We compare all models in terms of recall, precision, IoU, and F1-Score. All models performed well, with precision values greater than 82%. HRNet with multiscale fusion and high resolution retention achieved the highest F1 score and IoU due to the ability to capture fine building details. Standard UNet was also surpassed in precision by attention UNet, by 6.47%, indicating the advantages of attention mechanism for focusing on features. The NIR band integration increased the precision, recall and IoU scores of several models such as UNet, RefineNet and Attention UNet, revealing that NIR spectral information helps to discriminate building materials from vegetation. However, our proposed MBSSFA Net performed better than all the other models and achieved the highest Precision (92.36%), F1 Score (86.23%) and IoU (75.79%) among the tested architectures. The multi-branch spectral spatial feature attention mechanism of MBSSFA-Net leads to superior performance, which effectively captures spatial details and spectral variations especially in complex building structures. With this advanced feature extraction capability, and more enhanced skip connections, MBSSFA-Net is able to preserve spatial details throughout the segmentation process, and hence produce more accurate building identification results.

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-X-G-2025-181-2025 | © Author(s) 2025. CC BY 4.0 License.

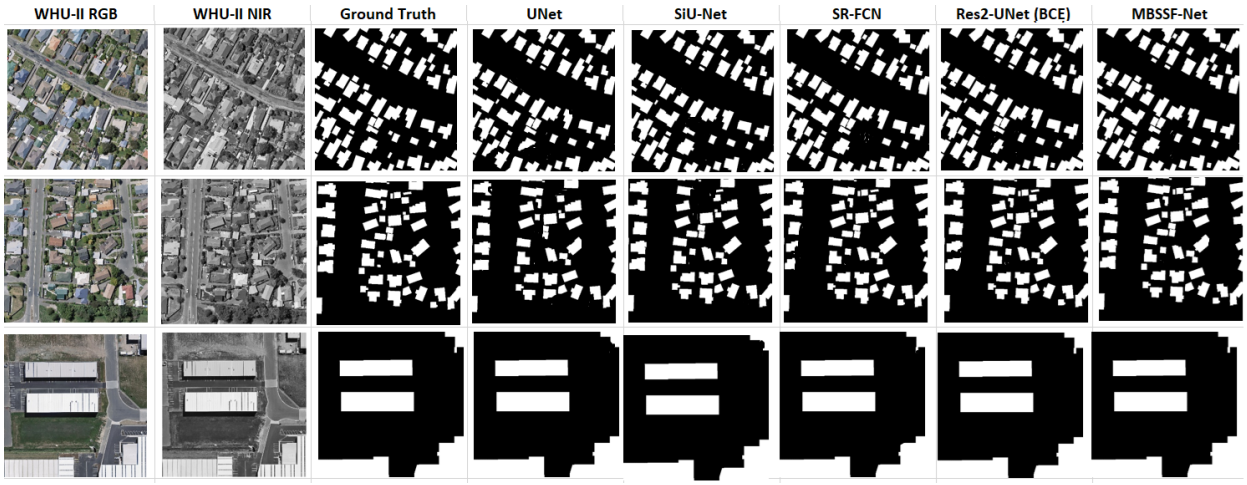

Figure 4. Semantic segmentation results of different DL methods on representative images from the RGB and NIR WHU-II test set.

Table 2. Performance metrics for different state-of-the-art segmentation methods and the proposed MBSSFA-Net on the test set of GaoFen-7 Building dataset.

Method	Precision	Recall	F1-Score	IoU
FCN 8S	0.8381	0.7661	0.8004	0.6665
SegNet	0.8470	0.8255	0.8369	0.7188
UNet	0.8059	**0.8764**	0.8447	0.7303
RefineNet	0.8371	0.8505	0.8508	0.7310
LinkNet	0.8675	0.7086	0.7801	0.6394
Attention UNet	0.9003	0.8146	0.8558	0.7475
HRNet	0.9144	0.8136	0.8603	0.7560
MBSSFA-Net	**0.9236**	0.8086	**0.8623**	**0.7579**

Table 3. Performance metrics for different state-of-the-art segmentation methods and the proposed MBSSFA-Net on the test set of WHU Satellite Dataset II dataset.

Method	**Precision**	**Recall**	**F1-score**	**IoU**
U-Net	0.6530	**0.8690**	0.7456	0.5940
SiU-Net	0.7250	0.7960	0.7588	0.6110
SR-FCN	0.7900	0.7700	0.7799	0.6400
Res2-UNet	0.8129	0.7864	0.7994	0.6659
Res2-UNet (BCE)	0.8158	0.7736	0.7942	0.6586
MBSSFA-Net	**0.8332**	**0.8076**	**0.8202**	**0.6952**

4.3.2 WHU Satellite Dataset II To evaluate the performance of our framework on the WHU-II dataset, we report results based on the benchmarks provided in [Chen et al., 2022]. SiU-Net [Ji et al., 2019], has a structural modifications to refine boundary details. SR-FCN [Shunping Ji and Lu, 2019] is developed to provide precise spatial accuracy, which is beneficial for pixel level segmentation where object boundaries are important. Res2-UNet [Chen et al., 2022] combines Res2Net modules with a boundary loss function to optimize for detailed edge delineation in complex or small building instances. Res2-UNet (BCE) [Chen et al., 2022] is only using the Binary Cross-Entropy (BCE). As shown in Table 3, our proposed model, MBSSFA-Net, leads to significant performance improvement on the WHU Satellite Dataset II, which contains training and testing images from different sources. MBSSFA-Net obtains an IoU of 0.6952, which is about 2.93% better than 0.6659 of Res2-UNet. The result demonstrates that MBSSFA-Net is capable of precise segmentation, which is important in applications where boundary accuracy in satellite images is critical. The proposed MBSSFA-Net also performs with a precision of 0.8332, which equates to a 1.74% improvement over the best performing previous model, Res2-UNet (BCE). The higher precision, means that MBSSFA-Net can do better than reducing false positives, which is highly desirable for satellite applications. Furthermore, the F1-score of MBSSFA-Net is 0.8202, which is 2.06% higher than that of Res2-UNet (BCE). The improvement proves that MBSSFA-Net is balanced on precision and recall and robust to detect the relevant feature in different dataset conditions.

5. Discussion

By combining spectral and spatial information, the proposed Multi-Band Spectral-Spatial Fusion Attention Network (MBSSFA-Net) shows enhacement in building segmentation within RS imagery. In order to improve the segmentation accuracy, we design the model with a dual encoder structure for RGB and NIR bands, which captures spectral specific features while preserving spatial details for separating building boundaries in varying urban environments. On the GaoFen-7 and WHU-II datasets, MBSSFA-Net achieves improvement on the standard metrics such as Precision, F1-score, and IoU, compared to existing models. In particular, the multi-branch attention mechanism, incorporating both channel and spatial attention, has been found particularly effective in improving feature extraction to refine the building edge delineation, a typical issue in urban remote sensing. The advantage of this approach is the ability to distinguish buildings from other surrounding vegetation and other elements. Compared to the best existing models, MBSSFA-Net shows a clear performance advantage, especially in terms of Precision and IoU, achieving 0.92% and 0.19% better Precision and IoU than HRNet respectively, proving robust and accurate for building segmentation tasks in complex urban environments. According to the results of the WHU Satellite Dataset II, MBSSFA-Net achieves the highest IoU of 69.52%, outperforming the best previous IoU result of 66.59% from Res2-UNet by 2.93%. The improvement in this shows that MBSSFA-Net is able to accurately segment building boundaries in complex urban environments. MBSSFA-Net paves the way for more feature representation techniques in RS imagery analysis. Future work will consider the optimization of this model for resource constrained settings to extend its applicability to real time and edge computing scenarios.

6. Conclusion

In this research, we introduced a new method called MBSSFA-Net which aims at improving the performance of building segmentation in VHR RS imagery. The proposed dual-encoder framework integrates RGB and Near-Infrared (NIR) information, which allows the model to capture both spectral and spatial features that are important for segmentation in the wide range of urban settings. The performance of MBSSFA-Net is then demonstrated using experimental results on the GaoFen-7 and WHU- II datasets and highlights the method as a new benchmark for building segmentation in the RS domain. The results are a new reference point for the extraction of buildings from satellite images, which will contribute to further advancement in this field. The datasets allow to assess the performance of the model in different building types and environments, providing a better understanding of the advantages and drawbacks of different Deep Learning architectures for the building segmentation.

References

Chen, F., Wang, N., Yu, B., Wang, L., 2022. Res2-Unet, a New Deep Architecture for Building Detection From High Spatial Resolution Images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 15, 1494-1501.

Chen, P., Huang, H., Ye, F., Liu, J., Li, W., Wang, J., Wang, Z., Liu, C., Zhang, N., 2024. A benchmark GaoFen-7 dataset for building extraction from satellite images. *Nature Scientific Data*, 11(1), 187.

Chen, X. Y. S. L. Z., all, 2021. An attention-fused network for semantic segmentation of very-high-resolution remote sensing imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 177.

Dixit, M., Chaurasia, K., Mishra, V. K., 2021. Dilated-ResUnet: A novel deep learning architecture for building extraction from medium resolution multi-spectral satellite imagery. *Expert Systems with Applications*, 184, 115530.

Fu, J., Liu, J., Tian, H., Li, Y., Bao, Y., Fang, Z., Lu, H., 2019. Dual attention network for scene segmentation. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 3141–3149.

Hajjar, S. E., Kassem, H., Abdallah, F., Omrani, H., 2024. Enhancing building segmentation by deep multiview classification for advancing sustainable urban development. *Journal of Building Engineering*, 83, 108421.

He, H., Wang, S., Zhao, Q., Lv, Z., Sun, D., 2021. Building extraction based on u-net and conditional random fields. *2021 6th International Conference on Image, Vision and Computing (ICIVC)*, 273–277.

J., M., M., A., 2018. Evaluation of effectiveness of three fuzzy systems and three texture extraction methods for building damage detection from post-event LiDAR data. *International journal of digital earth*, 11(12), 1241–1268.

Ji, S., Wei, S., Lu, M., 2019. Fully Convolutional Networks for Multisource Building Extraction From an Open Aerial and Satellite Imagery Data Set. *IEEE Transactions on Geoscience and Remote Sensing*, 57(1), 574-586.

Lee, C., Chang, L., 2024. NPSFF-Net: Enhanced Building Segmentation in Remote Sensing Images via Novel Pseudo-Siamese Feature Fusion. *Remote Sensing*.

Liu, Y., Gross, L., Li, Z., Li, X., Fan, X., Qi, W., 2019. Automatic Building Extraction on High-Resolution Remote Sensing Imagery Using Deep Convolutional Encoder-Decoder With Spatial Pyramid Pooling. *IEEE Access*, 7, 128774-128786.

Liu, Z., Chen, B., Zhang, A., 2020. Building segmentation from satellite imagery using u-net with resnet encoder. 1967–1971.

M., L., H., Y., Z., X. X., 2019. A relation-augmented fully convolutional network for semantic segmentation in aerial scenes. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 12408–12417.

Melgarejo, L.-F., Lakes, T., 2014. Urban adaptation planning and climate-related disasters: An integrated assessment of public infrastructure serving as temporary shelter during river floods in Colombia. *International journal of disaster risk reduction*, 9, 147–158.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, abs/1505.04597, 234–241.

Shunping Ji, S. W., Lu, M., 2019. A scale robust convolutional neural network for automatic building extraction from aerial and satellite imagery. *International Journal of Remote Sensing*, 40(9), 3308–3322.

Turker, M., Koc-San, D., 2015. Building extraction from high-resolution optical spaceborne images using the integration of support vector machine (SVM) classification, Hough transformation and perceptual grouping. *International Journal of Applied Earth Observation and Geoinformation*, 58-69.

Xu, J., Lu, K., Wang, H., 2021. Attention fusion network for multi-spectral semantic segmentation. *Pattern Recognition Letters*, 146, 179-184.

Y., Changqian; W., J. P. C., all, 2018. Learning a discriminative feature network for semantic segmentation. *IEEE conference on computer vision and pattern recognition*.

Yang, Y., Newsam, S., 2008. Comparing sift descriptors and gabor texture features for classification of remote sensed imagery. *2008 15th IEEE International Conference on Image Processing*, 1852–1855.

Ye, H., Liu, S., Jin, K., Cheng, H., n.d. *2020 25th International Conference on Pattern Recognition (ICPR)*, 166–172.

Yu, J., Wen, J., 2016. Multi-criteria satisfaction assessment of the spatial distribution of urban emergency shelters based on high-precision population estimation. *International Journal of Disaster Risk Science*, 7, 413–429.

Yu, Y., Wang, C., Kou, R., Wang, H., Yang, B., Xu, J., Fu, Q., 2025. Enhancing Building Segmentation With Shadow-Aware Edge Perception. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 18, 1-12.

Zhang, Y., Gong, W., Sun, J., Li, W., 2019. Web-Net: A Novel Nest Networks with Ultra-Hierarchical Sampling for Building Extraction from Aerial Imageries. *Remote Sensing*, 11(16).