

Article

A Residual U-Net Architecture for Built-Up Area Segmentation from Sentinel-2 Images

Mehtap Ülker 

Computer Engineering Department, Bitlis Eren University, Bitlis 13100, Turkey; m.ulker@beu.edu.tr

Abstract

Accurate and up-to-date mapping of built-up areas is of great importance for sustainable urban planning, disaster management, and the monitoring of environmental changes. In this study, a residual U-Net-based deep learning architecture named FiveBandTTA is proposed for built-up area segmentation from Sentinel-2 multispectral satellite imagery. The proposed model aims to simultaneously learn spatial and spectral features by jointly processing RGB, NIR (B8), and SWIR (B11) bands within the same encoder–decoder structure. The model incorporates standard residual blocks following the conventional residual learning principle, multi-level skip connection mechanisms, and TTA-based inference strategies. Within the scope of the study, a multi-temporal built-up area dataset was constructed from Sentinel-2 imagery acquired over Kocaeli Province. The performance of the proposed model was comparatively evaluated against RGB Baseline, FiveBand Single, DeepLabV3+, and SegFormer models. Experimental results demonstrated that the proposed model achieved the highest segmentation performance among all compared approaches, obtaining 0.8447 IoU, 0.9124 Dice, and 0.9249 Precision scores. It was observed that the use of multispectral bands together with the residual encoder–decoder structure may contribute to improved representation of small-scale built-up regions and complex boundary structures. Furthermore, the comparative experiments indicated that the NIR and SWIR bands provide complementary spectral information for distinguishing built-up areas, while the TTA-based inference strategy may contribute to improved segmentation stability and prediction consistency. Overall, the obtained results demonstrate that the proposed approach is an effective and robust method for built-up area segmentation from medium-resolution Sentinel-2 imagery.

Keywords: remote sensing; built-up area segmentation; Sentinel-2 imagery; residual U-Net; urban area extraction



Received: 20 May 2026

Revised: 18 June 2026

Accepted: 24 June 2026

Published: 26 June 2026

Copyright: © 2026 by the author.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and

conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

The accelerating urbanization process on a global scale has made accurate, up-to-date and comparable mapping of built-up areas a critical requirement for sustainable urban planning, disaster risk reduction and monitoring of environmental changes [1]. In this context, built-up area segmentation and building footprint extraction are among the fundamental research topics in remote sensing-based spatial analysis studies [2]. Developing methods that are applicable over large areas, automated, scalable and temporally repeatable is of great importance for effectively monitoring urban growth and strengthening decision support systems.

Very High Resolution (VHR) imagery enables the detailed delineation of building boundaries and has significantly expanded the potential for automated building footprint

extraction. However, practical challenges associated with large-scale and frequently updated applications may limit the operational sustainability of these data sources [3]. In contrast, Sentinel-2 imagery, provided within the scope of the Copernicus program, offers an important alternative due to its advantages of free accessibility, global coverage, and high temporal revisit capability [4]. Nevertheless, the 10 m spatial resolution of Sentinel-2 introduces several structural limitations in terms of the detailed representation of urban areas. At this resolution, a single pixel represents approximately 100 m², and in urban regions where different building and surface types coexist, multiple land-cover components may be present within the same pixel. Commonly referred to as the “mixed pixel” or sub-pixel problem, this phenomenon complicates the accurate and spectrally pure representation of built-up areas [5].

In recent years, deep learning-based semantic segmentation approaches have provided significant performance improvements in building extraction tasks from remote sensing imagery. In particular, encoder–decoder architectures and U-Net-based models have demonstrated superior accuracy compared with traditional pixel-based classification methods, owing to their ability to learn multi-scale features and utilize skip connection mechanisms [2,6,7]. Several recent studies have further improved encoder–decoder architectures through advanced contextual representation and feature refinement strategies. For example, RCFSNet [8] proposed an improved encoder–decoder architecture that incorporates multi-scale context extraction, full-stage feature fusion, and a coordinate dual-attention mechanism to strengthen road feature representation and improve the connectivity of extracted road labels. TransRoadNet [9] proposed a novel road extraction approach that combines high-level semantic features with foreground contextual information (FCI). Specifically, a Position Attention (PA) mechanism was designed to enhance the representation capability of road features, a Contextual Information Extraction Module (CIEM) was introduced to capture road contextual information, and a Foreground Context Information Supplement Module (FCISM) was developed to provide foreground contextual information at different stages of the decoder. CRNet [10] proposed a novel road extraction network to enhance the precision and topological connectivity of extracted road networks. Specifically, a Global–Local Context Decoupling Module (GLCDM) was introduced to model long-range contextual dependencies while preserving fine-grained local road features. In addition, a Semantic–Spatial Feature Refinement Module (SSFRM) was integrated into the skip connections to suppress background noise in shallow feature maps using deep semantic features. However, the vast majority of existing studies report their highest performance on high or very high-resolution imagery [3]. When these methods are applied to medium-resolution Sentinel-2 imagery, spectral heterogeneity and the limited spatial resolution make the detailed representation of urban areas more challenging. Furthermore, the mixed-pixel effect, where multiple land-cover components coexist within a single pixel, complicates the discrimination of built-up surfaces from spectrally similar classes. These limitations become particularly evident in heterogeneous urban environments and small-scale building detection tasks [5,11]. The multispectral configuration of Sentinel-2 provides important spectral information for distinguishing impervious surfaces, particularly through the NIR and SWIR bands [4]. However, simple early fusion strategies implemented together with RGB channels may reduce the distinctiveness of band-specific spectral features. To address this issue, recent studies have increasingly adopted multi-encoder architectures that process spectral band groups independently and integrate them through controlled fusion mechanisms [2,12].

These encoder-based architectures effectively reduce inter-band semantic interference by allowing each spectral group to learn features within its own representation space. On the other hand, the building segmentation problem requires not only the learning of

local edge and texture features but also the modeling of the spatial patterns of structures within their surrounding environmental context. CNN-based architectures demonstrate effective performance in learning local features. However, these architectures may remain limited in modeling global spatial dependencies [13]. To overcome this limitation, transformer-based and hybrid CNN–Transformer architectures have been proposed. In these architectures, transformer-based self-attention mechanisms are utilized to capture long-range spatial dependencies, while convolution operations are employed to preserve fine-grained local texture and boundary information [14]. Consequently, hybrid remote sensing architectures aim to improve segmentation performance by jointly utilizing global contextual representations and local spatial feature information [15].

The existing studies focus on very high-resolution imagery or employ traditional single-stream segmentation models that process multispectral information within standard encoder–decoder frameworks [3]. In contrast, only a limited number of studies comprehensively investigate the effectiveness of multispectral feature reconstruction strategies for medium-resolution Sentinel-2 imagery together with detailed ablation-based performance analyses. This limitation highlights the need for a clearer and more quantitative assessment of the contribution of multispectral feature integration to built-up area segmentation. The main contributions of this study can be summarized as follows:

- For built-up area segmentation from Sentinel-2 multispectral imagery, a simplified residual U-Net-based segmentation architecture is proposed, incorporating conventional residual blocks, multi-level skip connection mechanisms, and TTA-based inference while jointly processing RGB, NIR (B8), and SWIR (B11) bands.
- The performance of the proposed model was comparatively evaluated using a multi-temporal Sentinel-2 multispectral dataset constructed in this study for Kocaeli Province using satellite images. Furthermore, different multispectral configurations and segmentation architectures were comparatively evaluated through experimental analyses. In addition, the influence of the TTA-based inference strategy on the final segmentation performance was examined.

2. Related Works

In recent years, the rapid increase in urbanization has made the accurate and up-to-date mapping of built-up areas an important research topic for urban planning, disaster management, and environmental change analysis. Automatic built-up area detection based on satellite imagery has become one of the widely used approaches in this field, as it enables the fast, consistent, and repeatable analysis of large areas. Therefore, Sentinel-2 satellite imagery has emerged as an important data source for built-up area monitoring, owing to its global coverage, high temporal revisit frequency, and open data access [4].

In particular, its multispectral bands with a spatial resolution of 10 m enable large-scale urban analysis. However, the medium spatial resolution of Sentinel-2 data introduces several challenges in the detailed representation of urban areas. In heterogeneous urban environments, a single pixel often contains multiple land cover components. This phenomenon is referred to in the literature as the mixed pixel problem, which makes it difficult to spectrally distinguish built-up surfaces [5]. The mixed pixel effect can significantly limit the accuracy of built-up area detection, particularly in regions with small or scattered structures. Therefore, built-up area segmentation using medium-resolution Sentinel-2 imagery emerges as a challenging research problem in remote sensing, involving both spectral and spatial uncertainties. In the literature, a number of studies have addressed built-up area analysis using Sentinel-2 data. These studies can generally be grouped into three main approaches: (i) large-scale built-up or settlement mapping, (ii) building footprint extraction from medium-resolution data, and (iii) sub-pixel modeling approaches aimed at mitigating

the mixed pixel problem. In this context, representative examples of Sentinel-2-based built-up area studies are summarized in Table 1.

An examination of the studies summarized in Table 1 indicates that Sentinel-2 satellite imagery has been widely used for mapping built-up areas and conducting urban analysis. These studies can generally be grouped into three main approaches. The first approach focuses on large-scale built-up or settlement mapping. Built-up areas have been automatically extracted from Sentinel-2 imagery at a global scale using convolutional neural network-based approaches, leading to the generation of global built-up probability maps at 10 m spatial resolution [16]. Similarly, a deep learning-based framework called Sen2HSE was proposed using Sentinel-2 multispectral data, demonstrating that large-scale human settlement areas can be automatically mapped [17].

The second approach includes studies that aim to extract building footprints from medium-resolution Sentinel-2 data. A super-resolution-based semantic segmentation approach was developed to generate building footprint maps at a national scale from Sentinel-2 imagery [18]. In this study, building footprint maps at 2.5 m resolution were produced from 10 m Sentinel-2 imagery, and millions of buildings were detected. Such studies provide significant contributions by enabling the extraction of more detailed spatial information from medium-resolution data.

The third approach involves sub-pixel modeling techniques designed to reduce the mixed-pixel problem commonly encountered in medium-resolution imagery. In this context, a machine learning-based regression unmixing method was proposed to estimate the proportion of building area within individual pixels by jointly utilizing Sentinel-1 and Sentinel-2 data [19]. This approach enables a more accurate representation of built-up surfaces, particularly in heterogeneous urban environments.

In addition, several studies have explored super-resolution approaches to mitigate the loss of spatial detail in medium-resolution imagery. A convolutional autoencoder-based super-resolution model, named SEN-2-CAENET, was developed to enhance the spatial resolution and spatial detail of Sentinel-2 imagery [20]. Moreover, the literature also includes studies that integrate multiple data sources. A cross-fusion-based deep learning framework combining Sentinel-1 SAR and Sentinel-2 optical data was proposed, demonstrating the effectiveness of automatic built-up area extraction [21]. Furthermore, the IRUNet model, which integrates UNet and InceptionResNetV2 architectures for land-use classification using Sentinel-2 imagery, achieved high classification performance through multi-scale feature fusion [22].

However, a significant study in the literature processes multispectral bands within a single encoder stream. That is, although spectral differences between bands are utilized, these bands are not explicitly represented separately at the architectural level. However, the spectral information provided by different bands in Sentinel-2 images has the potential to improve building segmentation performance, especially in urban areas where the mixed-pixel effect is significant. Despite this, studies systematically examining the performance of different encoder–decoder-based architectures using multispectral band information are limited. Furthermore, studies evaluating the impact of different band combinations and architectural components on segmentation performance through detailed ablation experiments are also quite limited. Therefore, the development of deep learning-based architectural approaches that effectively utilize multispectral band information in medium-resolution Sentinel-2 data, and the investigation of the contribution of different band combinations to segmentation performance, represent a significant research gap in the literature.

Table 1. Comparative overview of studies on built-up area detection using Sentinel-2 data.

Ref.	Data Source	Spatial Resolution	Task	Model Type	Output	Key Findings
[16]	Sentinel-2	10 m	Built-up area mapping	CNN-based pixel classification	Global built-up probability map	A lightweight CNN using 5×5 pixel patches was employed to generate a global built-up map for 2018, validated across 277 regions using building footprint data.
[17]	Sentinel-2	10 m	Human settlement mapping	Sen2HSE (FCN-based deep learning model)	Settlement map (HSE)	Human settlement areas were automatically mapped from Sentinel-2 imagery.
[19]	Sentinel-1 Sentinel-2	10 m	Sub-pixel building estimation	ML regression-based unmixing	Building area fraction	The building area fraction within each pixel was estimated to reduce the mixed pixel effect.
[18]	Sentinel-2	10 m	Building footprint extraction	Super-resolution semantic segmentation	Building footprint map	A national-scale building footprint map at 2.5 m resolution was generated from 10 m Sentinel-2 imagery, detecting over 86 million buildings.
[20]	Sentinel-2	10 m	Super-resolution	CNN-based autoencoder	High-resolution imagery	The SEN-2_CAENET model was used to enhance the spatial detail of Sentinel-2 images.
[21]	Sentinel-1 Sentinel-2	10 m	Built-up area mapping	Cross-fusion neural network	Built-up map	Automatic built-up mapping was achieved by fusing Sentinel-1 SAR and Sentinel-2 optical data.
[22]	Sentinel-2	10 m	Land-use classification	IRUNet (UNet InceptionRes-NetV2)	Land-use map	High classification accuracy (98.21%) was achieved using multi-scale feature fusion and test-time augmentation.

3. Materials and Methods

3.1. Overview

The proposed model follows an encoder–decoder-based deep learning framework designed for multispectral built-up area segmentation. As illustrated in Figure 1, the model utilizes Sentinel-2 multispectral imagery consisting of RGB bands together with additional spectral bands, including B8 (NIR) and B11 (SWIR). All spectral bands are jointly processed within a unified single-branch encoder–decoder structure, enabling the extraction of complementary spatial and spectral feature representations. All spectral bands are jointly processed within a unified single-branch encoder–decoder structure, enabling the extraction of complementary spatial and spectral feature representations.

The encoder part consists of successive residual convolution blocks and downsampling operations for hierarchical feature extraction, while the decoder stage progressively reconstructs the spatial resolution through up-convolution and skip connection mecha-

nisms. The shallow encoder layers mainly preserve low-level spatial information such as local structures and boundary details, whereas deeper layers progressively capture higher-level contextual and spectral representations required for built-up area discrimination. As shown in Figure 1, the bottleneck stage connects the encoder and decoder representations through a deep residual feature transformation layer. Finally, a segmentation head composed of convolution operations generates the binary built-up area prediction map. In addition, TTA is applied during the inference stage to improve prediction robustness and segmentation stability.

The proposed model is based on a residual U-Net-inspired encoder–decoder framework designed for multispectral built-up area segmentation. Within this structure, the encoder hierarchically extracts spatial and spectral feature representations, while the decoder progressively reconstructs high-resolution segmentation maps through multi-level feature integration. The residual blocks employed in this study do not introduce a new residual block design; instead, they follow the conventional residual learning principle and are integrated into the encoder–decoder framework to support the processing of multispectral Sentinel-2 imagery. The architecture is constructed using successive conventional residual convolution blocks, where the encoder is responsible for hierarchical feature extraction and the decoder progressively reconstructs the segmentation map through up-convolution and skip connection mechanisms. Shallow encoder layers mainly preserve low-level spatial information such as local structures and boundary details, whereas deeper layers progressively capture higher-level contextual and spectral representations required for built-up area discrimination. The primary objective of this approach is to effectively utilize complementary information obtained from different multispectral bands within a unified deep learning framework rather than relying solely on RGB information. While RGB bands mainly capture spatial information such as object boundaries, texture patterns, and visual contrast, the additional spectral bands provide complementary information related to surface materials and land-cover characteristics. Consequently, the proposed model enables more effective discrimination of built-up areas by jointly learning spatial and spectral representations from multispectral Sentinel-2 imagery. By processing RGB, NIR (B8), and SWIR (B11) bands within a unified single-branch encoder–decoder framework, the model simultaneously exploits complementary spatial and spectral characteristics. The joint utilization of RGB, NIR, and SWIR information enables the network to generate more robust and discriminative feature representations, thereby improving built-up area segmentation performance. In addition, Test Time Augmentation (TTA) is applied only during the inference stage. Predictions obtained from the original image together with horizontally flipped, vertically flipped, and horizontally–vertically flipped inputs are averaged to generate the final segmentation map.

The encoder structure consists of conventional residual convolutional blocks integrated into the proposed residual U-Net framework. At each encoder stage, the spatial resolution is progressively reduced while higher-level spatial and spectral feature representations are learned. This hierarchical encoding strategy enables the network to effectively capture both low-level spatial details and high-level contextual information required for accurate built-up area segmentation. Although medium-resolution Sentinel-2 imagery may contain mixed pixels that introduce noise into shallow feature representations, deeper encoder layers provide more contextual information that helps alleviate the influence of such noise. Furthermore, the symmetric encoder–decoder structure together with multi-level skip connections helps preserve spatial information across different resolution levels and contributes to the reconstruction of built-up area boundaries. This design enables the decoder to progressively recover spatial details while combining low-level and

high-level feature representations during the segmentation process, thereby improving segmentation consistency.

The proposed model processes Sentinel-2 multispectral imagery consisting of RGB bands together with additional spectral bands, including B8 (NIR) and B11 (SWIR), through a single-branch encoder structure. Due to the use of multispectral input data, the first convolution layer of the encoder is configured to accept five-channel input representations. The feature maps extracted at different encoder stages are progressively transferred to the decoder structure through skip connections, as illustrated in Figure 1. The decoder combines low-level spatial information with high-level contextual and spectral representations to progressively reconstruct the final segmentation map.

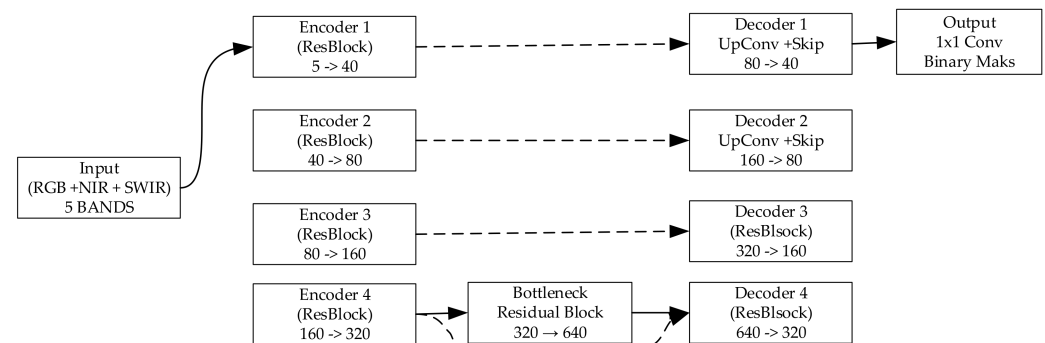


Figure 1. Overview of the proposed FiveBandTTA segmentation model.

TTA is incorporated during the inference stage to enhance prediction robustness and segmentation consistency. During inference, the model generates segmentation predictions using the original input image together with horizontally flipped, vertically flipped, and horizontally–vertically flipped versions of the same image. The prediction maps obtained from these augmented inputs are transformed back to their original orientations and averaged pixel-wise to produce the final segmentation output.

Following the hierarchical feature extraction process, the decoder progressively reconstructs the learned feature maps and generates the final segmentation representation for built-up area extraction. At the deepest encoder stage, the extracted feature maps contain high-level contextual and spectral information, which is subsequently transferred to the decoder stage. These operations facilitate more stable feature information transfer and support the learning of meaningful feature representations at different spatial scales. Finally, the network is terminated with a segmentation head performing pixel-level binary classification, and a binary segmentation map representing building and background classes is generated using a sigmoid activation function.

3.2. Multispectral Feature Reconstruction

In the proposed model, the multispectral features extracted from RGB bands together with additional spectral bands are jointly processed within a unified single-branch encoder–decoder framework. The encoder consists of four hierarchical stages ($i = 1, 2, 3, 4$), where spatial and spectral feature representations are progressively learned at different representation levels. The feature maps obtained at each encoder stage are progressively transferred to the decoder through skip connections to support multi-scale feature reconstruction and preservation of fine spatial details.

Feature reconstruction is further supported through standard residual blocks following the conventional residual learning principle and multi-level skip connections incorporated within the decoder structure, enabling more effective utilization of spatial and spectral feature information. These operations facilitate feature information transfer during the

reconstruction process and support segmentation consistency in complex built-up regions. This reconstruction process can be expressed as Equation (1).

$$S_i = F_i(E_i), \quad (1)$$

where S_i denotes the transformed feature representation transferred to the decoder stage, F_i represents the feature transformation operation performed using standard residual blocks following the conventional residual learning principle and E_i denotes the encoder feature representation obtained at the i -th encoder stage. The number of channels is progressively increased along the encoder hierarchy ($c_1 < c_2 < c_3 < c_4$), thereby enabling the network to learn more complex and informative feature patterns. This hierarchical channel expansion allows the model to capture both low-level spatial structures and high-level contextual and spectral feature representations across different multispectral feature spaces. Feature reconstruction is performed through multi-level skip connections and standard residual blocks incorporated within the decoder structure. The detailed architecture of the proposed encoder–decoder framework is illustrated in Figure 2.

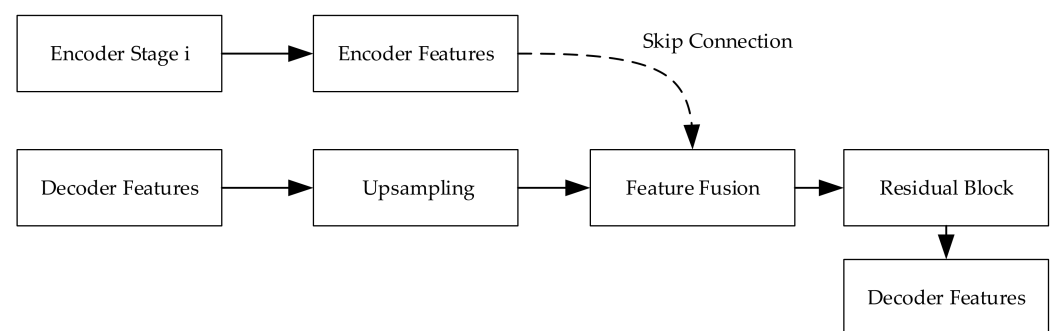


Figure 2. Multispectral encoder–decoder architecture with standard residual blocks and multi-level skip connections.

In the proposed model, multispectral feature representations obtained in the encoder layers are transferred to the decoder layers via a multi-level skip connection mechanism. In the decoder stage, features from deeper layers are progressively reconstructed through upsampling operations, while they are integrated with the corresponding encoder features during the feature fusion process. Subsequently, standard residual blocks following the conventional residual learning principle are employed within the decoder to support the learning of spatial and spectral feature representations, and the final segmentation feature representation is obtained. The encoder feature maps transferred through skip connections are combined with decoder representations to facilitate multi-scale feature integration and preservation of fine boundary details. This process is expressed in Equation (2) as follows:

$$\check{F}^i = R(F^i), \quad (2)$$

where $\check{F}^i$ denotes the transformed feature representation at the i -th stage, while R represents the feature transformation operation performed using standard residual blocks following the conventional residual learning principle during the decoding process. These standard residual blocks facilitate feature information transfer and support the learning of spatial and spectral feature representations during the decoding process. In particular, the transferred encoder features and decoder representations are jointly utilized through skip connections to support boundary preservation and multi-scale feature learning. In this way, multispectral feature representations are effectively utilized prior to the final segmentation reconstruction process.

In the second stage, the feature maps are progressively transferred to the decoder structure through skip connections and integrated with decoder feature representations during the decoding process. Following feature fusion, standard residual blocks following the conventional residual learning principle are employed within the decoder to support feature information transfer and the learning of spatial and spectral feature representations during multi-scale feature integration. This process is expressed in Equations (3)–(5) as follows:

$$F_{ref}^i = R(F^i), \quad (3)$$

$$D_{fuse}^i = \text{Concat}(U(D^{i+1}), F_{trans}^i), \quad (4)$$

$$S_i = \text{RELU}(\text{Conv}(D_{fuse}^i)), \quad (5)$$

where R denotes the standard residual block used in the decoder stage, U represents the upsampling operation, Concat corresponds to channel-wise concatenation, and S_i denotes the decoder feature representation obtained after the residual block at the i -th decoding stage. The resulting S_i representations are progressively propagated through the decoder hierarchy and integrated with encoder features via skip connections at the corresponding spatial levels. This structure enables the simultaneous utilization of low-level spatial details and high-level contextual feature information. At the deepest encoder stage, the extracted feature representations are transferred to the decoder as the bottleneck representation of the network. This bottleneck representation serves as the initial input to the decoder stage and contains compact high-level spatial and spectral information required for the reconstruction of the final segmentation map.

In the proposed architecture, since RGB bands together with NIR (B8) and SWIR (B11) bands are processed within a unified single-input tensor, spatial and spectral information are jointly learned throughout the encoder–decoder structure. High-resolution feature maps obtained in the early encoder stages enable the preservation and extraction of low-level spatial characteristics, including local texture information, edge responses, object boundaries, and fine-scale building structures. As the network depth increases, successive downsampling operations enlarge the receptive field and increase the channel capacity. This enables deeper encoder layers to represent broader contextual information and to extract higher-level spectral representations by exploiting the complementary information provided by multispectral bands. In particular, the reflectance differences captured by the NIR and SWIR bands facilitate the integration of spectral information at deeper representation levels, thereby contributing to the discrimination between built-up and non-built-up areas. Furthermore, the residual block structure preserves previously learned information by combining residual connections with transformed feature representations. Consequently, the loss of low-level spatial details and multispectral information across network depth is alleviated, while more discriminative feature transformations can be learned simultaneously. In addition, skip connections between the encoder and decoder stages fuse high-resolution spatial information from shallow layers with high-level contextual representations extracted by deeper layers, contributing to more accurate reconstruction of building boundaries and improved segmentation consistency. Through this multi-level feature integration mechanism, the proposed model effectively utilizes both fine spatial details and multispectral contextual information for built-up area segmentation.

3.3. Decoder Architecture

The bottleneck representation B obtained at the deepest encoder stage is transferred to the decoder structure. The decoder is designed with a hierarchical architecture symmetrical to the encoder and progressively restores the spatial resolution through successive up-

sampling operations performed at each decoding stage. This process is implemented using multi-scale feature reconstruction and skip connection-based feature integration, enabling low-resolution feature representations to be reconstructed at higher spatial resolutions. The decoder input at the i -th level is obtained by up-sampling the feature representation from the previous decoder stage, as expressed below. This process is formulated in Equation (6) as follows:

$$U_i = \text{Up}(D_{i+1}), \quad (6)$$

where D_{i+1} denotes the decoder feature representation obtained from the previous decoder stage, Up represents the learnable transposed convolution-based up-sampling operation, and U_i corresponds to the upsampled decoder feature representation at the i -th decoding level. As a result of this operation, the size of the feature map is increased from $H_{i+1} * W_{i+1}$ to $H_i * W_i$. Following the up-sampling stage, the upsampled feature map is fused along the channel dimension with the skip feature representation. S_i obtained from the corresponding encoder level through skip connections. This fusion process enables the decoder to jointly utilize high-level contextual information and fine-grained spatial details preserved within the encoder representations. The feature fusion process is expressed in Equation (7) as follows:

$$\hat{U}_i = \text{Concat}(U_i, S_i), \quad (7)$$

where U_i represents the upsampled decoder feature map, S_i denotes the encoder feature representation transferred through the skip connection, and Concat corresponds to the channel-wise feature concatenation operation. This fusion process reduces information loss by enabling low-level spatial details learned during the encoding stage to be directly transferred to the decoding stage through skip connections. The resulting fused feature representation is subsequently refined through residual decoder reconstruction operations consisting of convolution, normalization, and nonlinear activation functions. This refinement process improves feature consistency and facilitates the reconstruction of detailed building boundaries and small-scale urban structures. The overall process is expressed in Equation (8) as follows:

$$D_i = \text{ResBlock}(\hat{U}_i), \quad (8)$$

where $\hat{U}_i$ ResBlock denotes the standard residual block following the conventional residual learning principle and used within the decoder stage. The resulting D_i representation corresponds to the decoder feature map obtained at the i -th decoding stage. The standard residual block consists of successive convolution, normalization, and nonlinear activation operations, facilitating feature information transfer and supporting stable learning behavior. This structure allows low-level spatial details and high-level contextual feature representations to be processed simultaneously. Therefore, the overall processing flow at the i -th decoder stage is formulated in Equation (9) as follows:

$$D_i = \text{ResBlock}(\text{Concat}(\text{Up}(D_{i+1}), S_i)), \quad (9)$$

This multi-level integration strategy enables the preservation of spatial details through skip connections from the encoder while simultaneously allowing their integration with high-level semantic representations learned at the bottleneck stage. This structure particularly contributes to more accurate and consistent segmentation of small-scale built-up areas and complex boundary regions.

4. Experiment

4.1. Dataset

In this study, Kocaeli Province, Türkiye, was selected as the study area due to its intensive industrial activities and heterogeneous topographical characteristics. To perform multi-temporal analysis, Sentinel-2 multispectral satellite images acquired in 2015, 2020, and 2025 were utilized. The Sentinel-2 platform provided by the European Space Agency (ESA) served as the primary data source. A total of 5292 Sentinel-2 images were collected for dataset construction. After preprocessing and quality control procedures, 3482 image patches containing valid built-up area information were retained for model training and evaluation. Subsequently, all images were resized to 256×256 pixels and divided into patches for model training and evaluation. As a result, a total of 3482 image patches were generated from the multi-temporal Sentinel-2 dataset. Among these patches, 2772 were allocated for training, 357 for validation, and 353 for testing. Detailed patch distributions corresponding to different acquisition years and dataset partitions are presented in Table 2. This dataset configuration improved the reproducibility and transparency of the experimental setup.

Table 2. Distribution of image patches used in the experiments.

Year	Training	Validation	Test	Total
2015	904	126	119	1149
2020	926	114	114	1154
2025	942	117	120	1179
Total	2772	357	353	3482

The proposed model utilized both visible and additional spectral bands. Rather than relying solely on visible spectral bands, the study incorporated the B2 (Blue), B3 (Green), B4 (Red), B8 (Near-Infrared—NIR), and B11 (Shortwave Infrared—SWIR) bands. The NIR band facilitates the discrimination of vegetation from other surface types, while the SWIR band enhances the separation of spectrally similar surfaces such as concrete and bare soil. All bands were resampled to a spatial resolution of 10 m, and pixel values were normalized to the $[0, 1]$ range.

In the dataset creation process, automatically generated reference masks were generated using a three-stage elimination strategy based on spectral indices. In the first stage, open water surfaces were removed from the dataset using NDWI-based thresholding techniques. In the second stage, the NDVI index was computed, and pixels satisfying the condition $NDVI > 0.25$ were classified as natural areas and excluded from the building mask. These masks were used as reference labels for model training and evaluation rather than fully independent ground-truth annotations.

In the final stage, the Bare Soil Index (BSI) was employed to reduce the spectral interference problem frequently encountered in remote sensing studies. BSI enhances the spectral separability between bare soil and built-up surfaces, thereby contributing to a more effective discrimination of these surface types [23]. The BSI is computed as expressed in Equation (10).

$$BSI = \frac{(SWIR + Red) - (NIR + Blue)}{(SWIR + Red) + (NIR + Blue)}, \quad (10)$$

where SWIR represents the Shortwave Infrared band, Red denotes the red spectral band, NIR refers to the Near-Infrared band, and Blue corresponds to the blue spectral band. In this study, by applying a BSI threshold of 0.05, arid and mountainous regions spectrally similar to concrete surfaces were excluded from the mask. Since there is no universally accepted

fixed threshold value for BSI in the literature, the selected threshold may vary depending on the characteristics of the study area. Previous studies have reported BSI threshold values generally ranging between 0.02 and 0.10 [24,25]. Thus, this approach contributes to reducing false positive predictions by minimizing spectral interference between bare soil and built-up surfaces in semi-arid and topographically complex regions.

4.2. Evaluation Metrics

The uneven class distribution in satellite imagery and the dominance of background pixels may cause the overall accuracy metric to produce misleading performance evaluations. Therefore, the Intersection over Union (IoU) metric, which measures the overlap between predicted and ground-truth class regions, was adopted as the primary evaluation criterion [26,27]. The IoU formulation is given in Equation (11).

$$\text{IoU} = \frac{TP}{TP + FP + FN'} \quad (11)$$

where TP, FP, and FN denote the numbers of true positive, false positive, and false negative pixels, respectively. In addition, the Dice coefficient is computed as formulated in Equation (12).

$$\text{Dice} = \frac{2TP}{2TP + FP + FN'} \quad (12)$$

4.3. Implementation Details

All experiments were trained on an NVIDIA RTX 5070 Ti GPU. During the training process, the batch size was set to 32, and all input images were resized to 256×256 pixels before training. The AdamW optimizer was employed with hyperparameters of $\beta_1 = 0.9$, $\beta_2 = 0.999$, and weight decay = 1×10^{-4} . The initial learning rate was initialized to 1×10^{-4} . Instead of maintaining a fixed learning rate throughout training, the learning rate was dynamically adjusted using the ReduceLROnPlateau scheduling strategy (factor = 0.5, patience = 5) based on the validation IoU performance. This adaptive learning strategy contributed to more stable convergence and enabled a finer optimization process when the validation performance stagnated. The number of training epochs was set to 40. TTA was applied only during the inference phase and was not used during training. In addition, all comparative models were evaluated using standard inference settings without TTA. To ensure the reproducibility of the experiments, fixed random seed values of 42, 123, and 3407 were used during data shuffling, model initialization, and training. The detailed training hyperparameters and implementation settings are summarized in Table 3.

Table 3. Training hyperparameters and implementation details used in the experiments.

Parameter	Value
Input image size	256×256 pixels
Batch size	32
Optimizer	AdamW
Weight decay	1×10^{-4}
Initial learning rate	1×10^{-4}
Scheduler factor	0.5
Scheduler patience	5 epochs
Learning rate scheduler	ReduceLROnPlateau
Random seed	42, 123, 3407
Number of epochs	40

4.4. Ablation Experiments

Ablation experiments were conducted to comprehensively evaluate the performance of the proposed approach and the effectiveness of the proposed architectural design. In this context, five different models were compared by considering different band utilization strategies and architectural designs: (i) RGB Baseline, which utilizes only visible spectral bands; (ii) FiveBand Single, in which all spectral bands are processed within a single input stream; (iii) DeepLabV3+, a widely used encoder–decoder-based segmentation architecture; (iv) SegFormer, a transformer-based segmentation model. To improve the reliability and reproducibility of the experimental evaluation, all models were independently trained three times using different random seed values (42, 123, and 3407). The reported IoU, Dice, and Precision values correspond to the mean results obtained from these repeated experiments, and the associated standard deviations are also presented.

- RGB Baseline is based on the U-Net architecture [28] and utilizes only the visible spectral bands (B2, B3, B4).
- FiveBand Single is also based on the U-Net architecture [28] and incorporates the NIR (B8) and SWIR (B11) bands in addition to the visible bands. In this model, all spectral bands are concatenated along the channel dimension and processed through a single encoder stream. This structure enables the direct utilization of multispectral information.
- DeepLabV3+ [29] is a widely used encoder–decoder-based semantic segmentation architecture that aims to capture multi-scale contextual information using convolution operations.
- SegFormer [30] is a transformer-based semantic segmentation model that employs hierarchical transformer encoders and a lightweight MLP decoder to effectively capture both local and global contextual information.
- FiveBand Residual (No TTA) employs the same residual U-Net architecture as the proposed model but does not apply test-time augmentation during inference.

The RGB Baseline model is based on the U-Net architecture [25] and utilizes only the visible spectral bands (B2, B3, and B4). The FiveBand Single model extends this configuration by incorporating the NIR (B8) and SWIR (B11) bands into a single encoder stream through channel-wise concatenation. DeepLabV3+ [26] is an encoder–decoder-based semantic segmentation architecture designed to capture multi-scale contextual information using dilated convolution operations. SegFormer [27] is a transformer-based segmentation model that employs hierarchical transformer encoder structures and a lightweight MLP decoder to effectively learn both local and global contextual dependencies. In contrast, the FiveBandTTA model adopts a residual U-Net-based encoder–decoder model that processes RGB, NIR (B8), and SWIR (B11) bands together within a multispectral learning structure. Furthermore, TTA is integrated at the inference stage to improve prediction stability and segmentation consistency.

Examining the results presented in Table 4, it can be observed that the RGB_Baseline model, which utilizes only the visible spectral bands (B2, B3, and B4), achieved the lowest performance among all evaluated models, with 0.4211 IoU, 0.5764 Dice, and 0.6099 Precision values. This result indicates that RGB bands alone are insufficient for the accurate detection of built-up areas and highlights the critical importance of multispectral information in remote sensing-based segmentation tasks.

Table 4. The comparison of segmentation models on Sentinel-2 test data.

Model	Input Bands	Architecture	TTA	IoU (Mean $\pm$ Std)	Dice (Mean $\pm$ Std)	Precision (Mean $\pm$ Std)
RGB Baseline	B2, B3, B4	U-Net	No	0.4211 $\pm$ 0.0020	0.5764 $\pm$ 0.0019	0.6099 $\pm$ 0.0018
FiveBand Single	B2, B3, B4, B8, B11	U-Net	No	0.8230 $\pm$ 0.0016	0.9012 $\pm$ 0.0011	0.9166 $\pm$ 0.0010
DeepLabV3+	B2, B3, B4, B8, B11	CNN encoder–decoder	No	0.8019 $\pm$ 0.0018	0.8877 $\pm$ 0.0014	0.8709 $\pm$ 0.0012
SegFormer	B2, B3, B4, B8, B11	Transformer- based	No	0.8137 $\pm$ 0.0016	0.8950 $\pm$ 0.0015	0.9214 $\pm$ 0.0011
FiveBand Residual (No TTA)	B2, B3, B4, B8, B11	Residual U-Net	No	0.8172 $\pm$ 0.0015	0.8975 $\pm$ 0.0012	0.8882 $\pm$ 0.0008
FiveBandTTA (proposed)	B2, B3, B4, B8, B11	Residual U-Net with TTA	Yes	0.8447 $\pm$ 0.0013	0.9124 $\pm$ 0.0010	0.9249 $\pm$ 0.0009

The FiveBand Single model, which utilizes NIR (B8) and SWIR (B11) bands in addition to RGB bands, achieved significantly higher performance despite having the same U-Net-based encoder–decoder architecture as the RGB Baseline model. The model’s achievement of 0.8230 IoU and 0.9012 Dice values indicates that the performance improvement primarily stems from the contribution of multispectral information rather than architectural modifications. In particular, the approximately 95% increase in IoU demonstrates that the NIR and SWIR bands provide highly discriminative spectral features for distinguishing built-up areas. Furthermore, the 0.9166 Precision value indicates that the incorporation of multispectral bands significantly contributes to the reduction in false-positive predictions.

The transformer-based SegFormer model demonstrated strong segmentation performance, achieving 0.8137 IoU and 0.8950 Dice scores. The model also achieved a high Precision value of 0.9214. This result indicates that the transformer-based global attention mechanism is effective in reducing false-positive predictions. However, the slightly lower IoU performance compared to the FiveBand Single model may be associated with the limitations of transformer-based architectures in preserving fine-grained spatial details and boundary information in small-scale built-up regions.

The FiveBand Residual (No TTA) model, which shares the same residual U-Net architecture as the proposed FiveBandTTA model but does not employ test-time augmentation, achieved an IoU of 0.8172, a Dice score of 0.8975, and a Precision value of 0.8882. These results demonstrate that the residual encoder–decoder structure itself provides strong segmentation capability for multispectral built-up area extraction. However, compared with the proposed FiveBandTTA model, lower performance values were obtained across all evaluation metrics. Specifically, after applying TTA, the IoU, Dice, and Precision values increased from 0.8172, 0.8975, and 0.8882 to 0.8447, 0.9124, and 0.9249, respectively. This observation suggests that the application of test-time augmentation during the inference stage contributes to improved prediction robustness, segmentation consistency, and more accurate delineation of built-up area boundaries, thereby enhancing the overall reliability of the segmentation results.

The proposed FiveBandTTA model achieved the highest overall segmentation performance among all evaluated approaches, achieving 0.8447 IoU, 0.9124 Dice, and 0.9249 Precision values. From an architectural perspective, the superior performance of the proposed model can be attributed to its residual encoder–decoder design, which enables more stable feature learning and more effective integration of multispectral information throughout the reconstruction process. Unlike conventional U-Net-based structures, residual reconstruction blocks facilitate improved feature information transfer between encoder and decoder stages, more effectively preserving both low-level spatial details and high-level contextual

representations. Additionally, the combined processing of RGB, NIR (B8), and SWIR (B11) bands within a single multispectral learning framework contributes to stronger discrimination of built-up regions and boundary structures. Furthermore, the use of TTA during the inference stage improves prediction stability and segmentation consistency, contributing to a reduction in false-positive predictions and more accurate delineation of complex urban boundaries. Although medium-resolution Sentinel-2 imagery may contain mixed pixels that introduce noise into low-level feature representations, the obtained results indicate that the proposed encoder–decoder framework maintains stable segmentation performance. The symmetric encoder–decoder structure together with multi-level skip connections enables the integration of low-level spatial details and high-level contextual representations during feature reconstruction, contributing to more consistent segmentation outputs and boundary preservation.

To assess the reliability of the generated reference masks, an additional set of 85 randomly selected Sentinel-2 image patches was collected and manually annotated independently from the automatically generated reference masks used in the proposed dataset. The selected samples were obtained from Sentinel-2 images acquired in 2015, 2020, and 2025. These downloaded image patches were manually delineated based on visual interpretation and subsequently used to evaluate the proposed model independently. The corresponding performance results are presented in Table 5. As summarized in Table 5, the proposed model achieved an IoU of 84.12%, a Dice coefficient of 91.23%, a Precision of 93.81%, and a Recall of 89.02% on this manually annotated evaluation subset. These results were highly consistent with those obtained on the original test set, with only minor differences observed across the evaluation metrics. This close agreement indicates that the generated reference masks provide reliable annotations and that the proposed approach maintains stable segmentation performance when evaluated against manually annotated samples.

Table 5. Performance evaluation on the manually annotated reference dataset.

Evaluation Dataset	IoU	Dice	Precision	Recall
Proposed dataset	0.8447	0.9141	0.9249	0.9064
Human-Annotated Evaluation dataset	0.8412	0.9123	0.9381	0.8902

Examining the computational complexity and inference performance results presented in Table 6, it is observed that there are significant differences among the models in terms of parameter count, computational load (FLOPs), and inference speed (FPS). Since the RGB_Baseline and FiveBand_Single models employ the same U-Net-based encoder–decoder architecture, they exhibit similar parameter counts (~8.12M) and computational costs. In the FiveBand_Single model, the incorporation of multispectral bands (B8 and B11) resulted in only a slight increase in computational cost, while providing improved segmentation performance. This finding indicates that although multispectral information introduces additional computational cost, it provides a noticeable improvement in segmentation performance. Despite having a lower number of parameters (3.72M), the DeepLabV3+ model achieved the highest FLOPs value (32.42 G) due to its dilated convolution operations and ASPP-based multi-scale processing mechanism. DeepLabV3+ achieved 27.25 FPS, whereas the proposed model achieved 220.41 FPS without TTA and 54.62 FPS with TTA. The reduction in inference speed when TTA is enabled is expected, since the TTA strategy requires additional forward passes on multiple transformed versions of the input image. Despite this additional computational overhead, the proposed model

maintained a higher inference speed than all comparison models, even with TTA enabled, while simultaneously benefiting from the improved prediction stability provided by TTA.

The SegFormer model exhibited the lowest number of parameters (1.29M) among the compared methods and relatively low computational complexity because of its efficient transformer-based design. These results indicate that the SegFormer architecture provides an efficient balance between model size and computational cost.

Table 6. Comparison of computational complexity and inference performance of segmentation models.

Model	PARAMS	FLOPs (G)	FPS Without TTA	FPS with TTA
RGB_Baseline	8.12M	15.04 G	21.39	-
FiveBand_Single	8.12M	15.08 G	19.17	-
DeepLabV3+	3.72M	32.42 G	27.25	-
SegFormer	1.29M	9.66 G	20.14	-
Proposed_model	12.69M	23.50 G	220.41	54.62

The proposed model, despite having the highest number of parameters (12.69M), exhibited a moderate computational cost (23.50G FLOPs). The model achieved an inference speed of 220.41 FPS without TTA and 54.62 FPS with TTA. The reduction in inference speed when TTA was enabled is expected, since additional inference operations are performed on multiple transformed versions of the input image, and the resulting predictions are combined to obtain the final output. Despite this additional computational overhead, the proposed model maintained a higher inference speed than all comparison models, even when TTA was enabled. Nevertheless, the proposed model achieved the highest IoU, Dice, and Precision values, demonstrating that the architecture is highly effective for multispectral feature learning and the preservation of complex built-up area boundaries.

Figure 3 shows that the RGB Baseline model is insufficient in identifying the boundaries of built-up areas and produces incomplete and irregular segmentation results, especially in densely built-up regions. The FiveBand, DeepLabV3+, and SegFormer models, which utilize multispectral bands, detect built-up areas more successfully and better preserve the integrity of the segmented regions. In particular, the DeepLabV3+ and SegFormer models were observed to reconstruct complex built-up boundaries more consistently. However, the proposed model demonstrates improved segmentation performance in small-scale built-up areas, dense building clusters, and complex boundary structures. Furthermore, it is noteworthy that the proposed model reduces unnecessary segmented regions and generates more consistent segmentation maps that show a higher visual agreement with the reference masks.

Examining the training and validation curves of the RGB_Baseline model in Figure 4, it can be observed that both the training and validation loss values decrease consistently throughout the training process, indicating stable optimization behavior. However, the obtained IoU performance remains relatively limited compared to the other multispectral models. This finding suggests that relying solely on visible spectral bands provides insufficient spectral information for the accurate discrimination of built-up areas. Furthermore, the close correspondence between the training and validation IoU curves indicates that the model demonstrates satisfactory generalization capability without exhibiting a significant overfitting tendency.

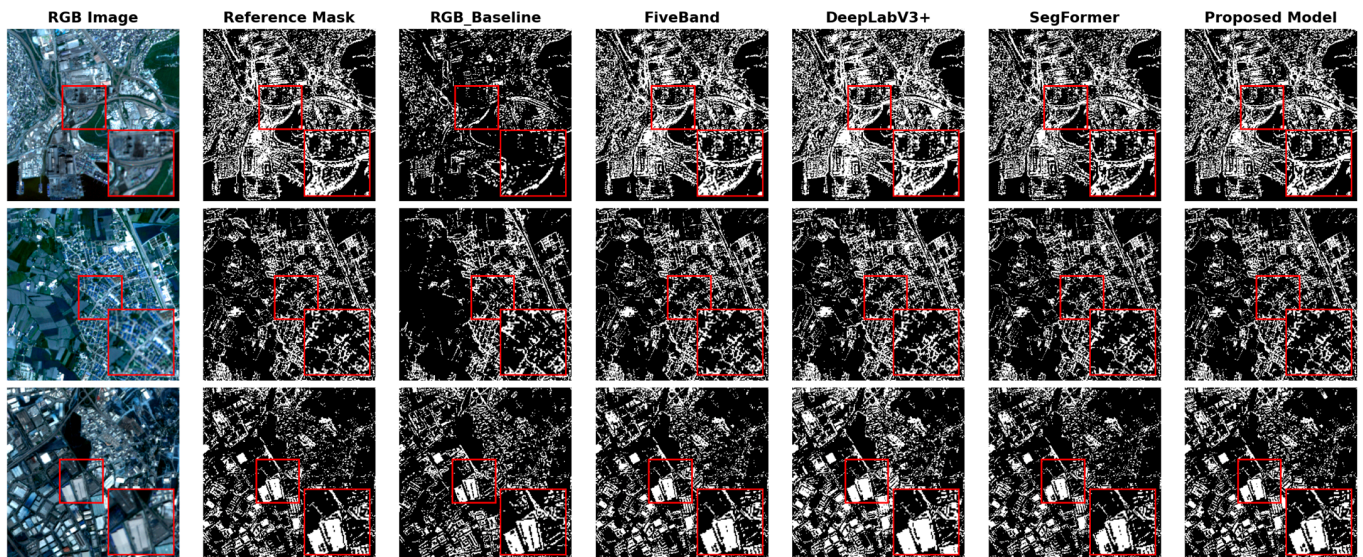


Figure 3. Comparison of built-up area segmentation results obtained from Sentinel-2 test images.

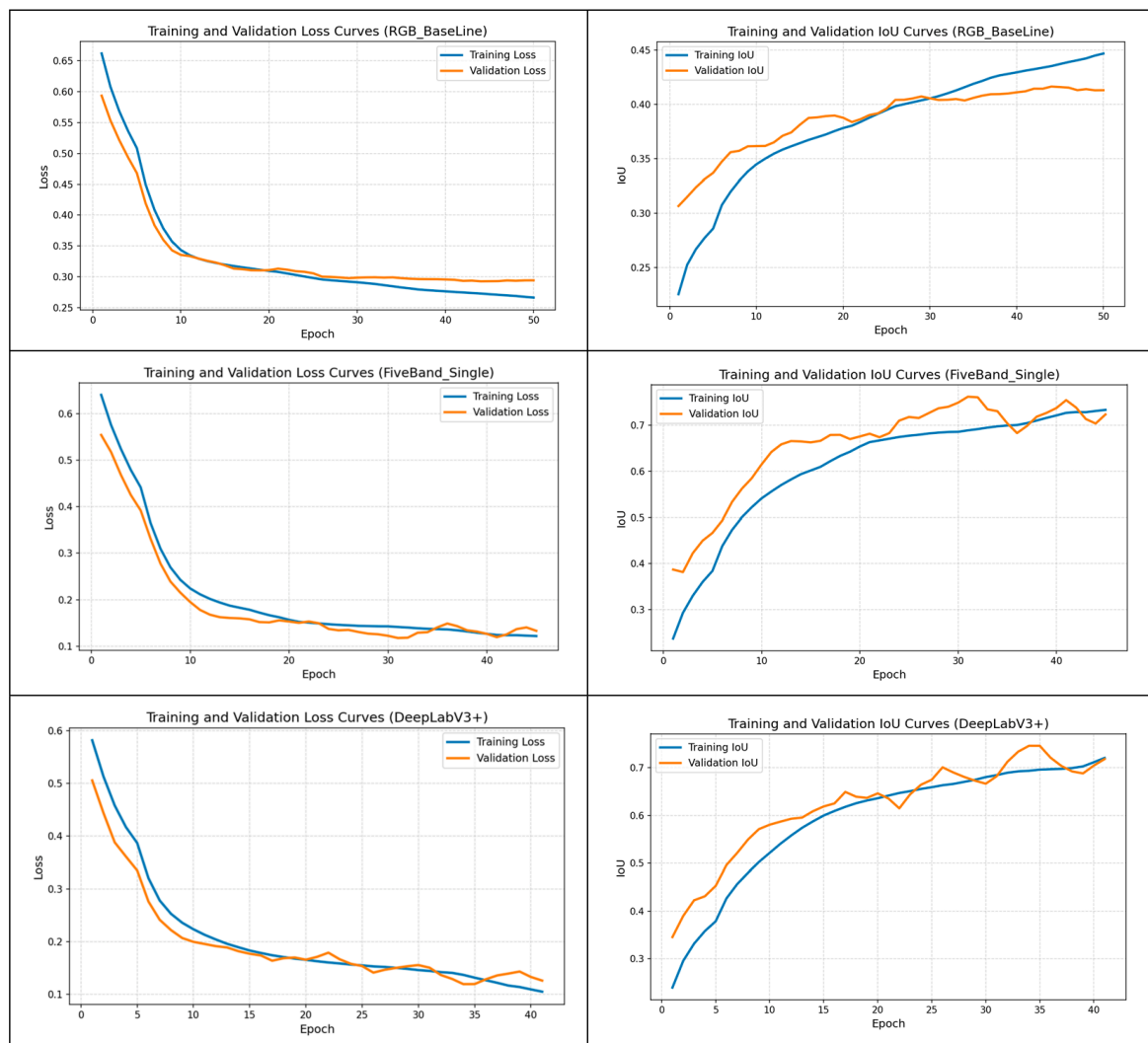


Figure 4. Cont.

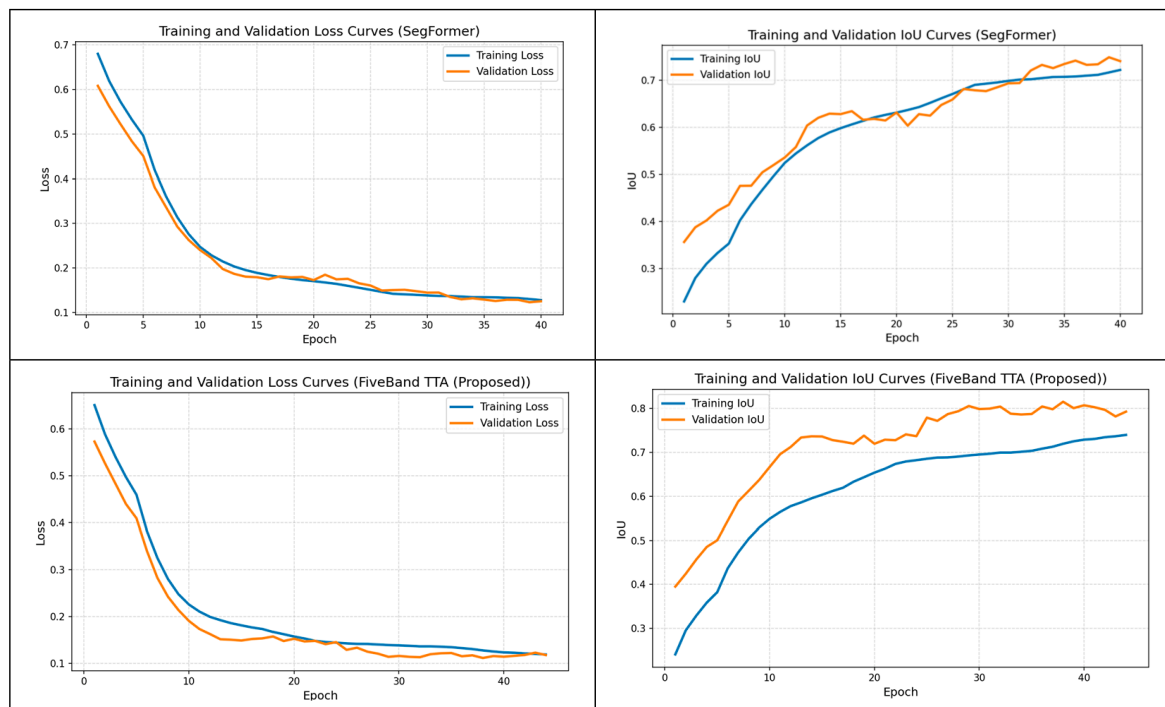


Figure 4. Training and validation loss and IoU curves of segmentation models.

In the FiveBand Single model shown in Figure 4, a significant improvement in learning performance was observed following the incorporation of the NIR and SWIR spectral bands. During the training process, the loss values decreased rapidly, while the IoU performance reached substantially higher levels compared to the RGB Baseline model. In particular, the stable increase observed in the validation IoU curve indicates that multispectral information provides a strong contribution to the discrimination of built-up areas. Furthermore, the close correspondence between the training and validation curves demonstrates stable convergence behavior and strong generalization capability without significant overfitting tendencies. Figure 4 shows the training curves of the DeepLabV3+ model, demonstrating that the model performs a stable optimization process. Both the training and validation loss values decreased regularly, exhibiting convergence behavior in the later stages of the training process. While the model appears to produce successful results in multi-scale contextual feature extraction, it is observed that the increase in IoU performance reaches saturation after a certain level. This suggests that although the model effectively learns contextual information, it may have some limitations in preserving fine spatial details.

The transformer-based SegFormer model shown in Figure 4 exhibited stable convergence behavior and strong IoU performance throughout the training process. The close correspondence between the training and validation IoU curves indicates that the model possesses strong generalization capability with limited overfitting behavior. Furthermore, the stable progression of the validation IoU values at relatively high levels suggests that the transformer-based global attention mechanism is capable of generating effective feature representations for the detection of built-up areas. However, the minor fluctuations observed during the later stages of training may be associated with the sensitivity of transformer-based architectures to small-scale spatial details and boundary regions.

Examining the training and validation curves given in Figure 4, it is seen that the proposed FiveBandTTA model exhibits stable convergence behavior throughout the training process. The regular decrease in both training and validation loss curves indicates that the model successfully learns the discriminative spatial and spectral features of the built-up areas. Furthermore, the validation loss being very close to the training loss shows

that there is no significant overfitting problem in the model and that strong generalization performance has been achieved.

Examining the IoU curves, it is seen that the validation IoU values are slightly higher than the training IoU values. The close correspondence between the training and validation IoU curves indicates that the proposed model exhibits strong generalization capability without significant overfitting. The stabilization of the curves, especially in later epochs, suggests that the model converges toward an optimal segmentation solution. Since TTA is applied only during the inference stage, it does not affect the training or validation curves. The validation IoU performance, reaching approximately 0.80, demonstrates that the proposed model can successfully detect built-up area boundaries by effectively utilizing information obtained from multispectral Sentinel-2 bands.

5. Discussion

- The results demonstrate that the use of multispectral bands significantly improves built-up area segmentation performance. In particular, the integration of NIR (B8) and SWIR (B11) bands contributed to a more accurate representation of small-scale built-up regions and complex boundary structures. The proposed FiveBandTTA model achieved the highest IoU, Dice, and Precision scores among all compared models. Furthermore, it was observed that the combination of multispectral information, the encoder–decoder architecture, and the TTA-based inference strategy improved segmentation stability and produced more consistent predictions.
- The results demonstrate that multispectral information is critical for structured area segmentation in medium-resolution images such as Sentinel-2. In particular, the poor performance of the model using only RGB bands revealed that visible bands alone do not provide sufficient discrimination. In contrast, the proposed residual U-Net-based structure learned multispectral features more effectively, producing more successful segmentation results in complex urban areas. Furthermore, it was observed that the TTA-based inference strategy increased prediction stability and generated more consistent segmentation maps.
- The proposed model's high IoU and Dice performance demonstrates its capability to successfully extract built-up areas from medium-resolution Sentinel-2 imagery. In particular, the encoder–decoder architecture together with multi-level skip connections contributed to the preservation of fine spatial details and boundary information. Experimental results indicate that multispectral bands enhance class separability, especially in complex urban regions. However, the relatively low FPS performance of the model suggests that future studies may focus on developing computationally more efficient architectures while preserving segmentation accuracy.
- The study results indicate that built-up area segmentation in medium-resolution Sentinel-2 imagery depends not only on the model architecture of the model but also on the selected spectral band combinations. In particular, the incorporation of B8 (NIR) and B11 (SWIR) bands contributed to reducing spectral confusion between built-up surfaces and bare soil regions. The stable progression of the training curves demonstrates that the proposed model achieved a balanced learning process and exhibited strong generalization capability. Furthermore, the proposed approach produced more consistent segmentation maps and more accurately delineated building boundaries in complex urban environments.
- The findings demonstrate that residual encoder–decoder architectures can achieve effective performance in multispectral feature learning. In particular, the proposed model generated more consistent segmentation outputs in dense built-up clusters and irregular boundary regions. Although transformer-based models successfully

captured global contextual information, the proposed model demonstrated superior performance in preserving fine spatial details and boundary structures. These results indicate that residual-based architectures provide a strong alternative for built-up area segmentation in multispectral Sentinel-2 imagery.

- The results indicate that high-capacity models alone are insufficient for achieving successful built-up area segmentation in medium-resolution Sentinel-2 imagery. In particular, the proposed model produced more stable segmentation outputs in low-resolution boundary regions. Furthermore, the combined use of multispectral bands contributed to reducing spectral interference between built-up areas and spectrally similar surfaces. These findings demonstrate that the proposed approach provides more reliable and consistent segmentation performance in complex urban environments.
- The fact that the proposed model produced more successful results, particularly in small-scale built-up regions and complex boundary structures, demonstrates the effectiveness of the combination of multispectral information, the encoder–decoder model, and multi-level skip connections. The close correspondence between the training and validation curves indicates that the model achieved stable convergence behavior without exhibiting significant overfitting. Furthermore, it was observed that the TTA-based inference strategy improved segmentation stability by combining predictions obtained from different spatial transformations.
- Furthermore, the study results demonstrate that the use of multispectral data significantly improves segmentation performance, particularly in heterogeneous urban regions. While models utilizing only RGB bands produced more irregular and incomplete built-up boundaries, the incorporation of NIR and SWIR bands resulted in more coherent and holistic segmentation maps. The more consistent performance of the proposed model, especially in dense built-up clusters, indicates that the combination of multispectral information and the encoder–decoder model enables more effective utilization of multi-level feature representations. In addition, the stable improvement observed in the model’s validation performance demonstrates the strong generalization capability of the proposed approach.

6. Conclusions

This study proposes a residual U-Net-based deep learning model for built-up area segmentation from Sentinel-2 multispectral satellite imagery. In the proposed model, RGB bands together with NIR (B8) and SWIR (B11) bands are jointly processed within the same encoder–decoder structure, while residual convolution blocks, multi-level skip connection mechanisms, and TTA-based inference strategies are incorporated into the model. Furthermore, a multi-temporal built-up area dataset was constructed for Kocaeli Province using Sentinel-2 imagery. Experimental results demonstrated that the use of multispectral bands significantly improves built-up area segmentation performance. In particular, compared with the RGB Baseline model utilizing only visible spectral bands, the incorporation of NIR (B8) and SWIR (B11) bands resulted in substantial improvements in IoU performance. The proposed FiveBandTTA model achieved the highest segmentation performance among all compared models, obtaining 0.8447 IoU, 0.9124 Dice, and 0.9249 Precision scores. Furthermore, it was observed that the combination of multispectral information and the encoder–decoder model contributed to a more accurate representation of small-scale built-up regions and complex boundary structures. The conducted ablation studies revealed that multispectral bands play a critical role, particularly in separating built-up areas from spectrally similar surfaces. Furthermore, it was determined that the TTA-based inference strategy improves segmentation stability and generates more consistent predictions. The close convergence between the training and validation curves indicates that the proposed

model possesses stable generalization capability. In conclusion, the proposed approach was found to be an effective method for built-up area segmentation in medium-resolution Sentinel-2 imagery. Future studies may focus on developing computationally more efficient architectures, integrating attention-based feature fusion strategies, and investigating the generalizability of the model across different geographic regions.

Funding: This study received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author.

Conflicts of Interest: The author declares no conflicts of interest.

References

1. Seto, K.C.; Güneralp, B.; Hutya, L.R. Global forecasts of urban expansion to 2030. *Proc. Natl. Acad. Sci. USA* **2014**, *109*, 16083–16088.
2. Zhu, X.X.; Tuia, D.; Mou, L.; Xia, G.-S.; Zhang, L.; Xu, F.; Fraundorfer, F. Deep learning in remote sensing: A comprehensive review and list of resources. *IEEE Geosci. Remote Sens. Mag.* **2017**, *5*, 8–36. [\[CrossRef\]](#)
3. Hakim, Y.F.; Tsai, F. Building Footprint Extraction for Large-Scale Basemaps Using Very-High-Resolution Satellite Imagery. *Buildings* **2026**, *16*, 675.
4. Fletcher, K. (Ed.) *ESA's Optical High-Resolution Mission for GMES Operational Services*; ESA Communications: Paris, France, 2012.
5. Foody, G.M. Status of land cover classification accuracy assessment. *Remote Sens. Environ.* **2002**, *80*, 185–201. [\[CrossRef\]](#)
6. Sun, Y.; Bi, F.; Gao, Y.; Chen, L.; Feng, S. A multi-attention UNet for semantic segmentation in remote sensing images. *Symmetry* **2022**, *14*, 906. [\[CrossRef\]](#)
7. Wu, Y.; Wang, F.; Zhao, P.; Zhou, M.; Geng, S.; Zhang, D. UNet with multibranch prior information encoding for building segmentation in remote sensing images. *Adv. Space Res.* **2025**, *76*, 4296–4313. [\[CrossRef\]](#)
8. Yang, Z.; Zhou, D.; Yang, Y.; Zhang, J.; Chen, Z. Road Extraction from Satellite Imagery by Road Context and Full-Stage Feature. *IEEE Geosci. Remote Sens. Lett.* **2023**, *20*, 1–5.
9. Yang, Z.; Zhou, D.; Yang, Y.; Zhang, J.; Chen, Z. TransRoadNet: A Novel Road Extraction Method for Remote Sensing Images via Combining High-Level Semantic Feature and Context. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 5506705.
10. Yang, Z.; Yao, H.; Li, Q.; Ni, W.; Wu, J.; Wang, Q. Semantic–Spatial Feature Refinement Network for Road Extraction from Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2026**, *64*, 1–10. [\[CrossRef\]](#)
11. Schug, F.; Frantz, D.; Okujeni, A.; van Der Linden, S.; Hostert, P. Mapping urban-rural gradients of settlements and vegetation at national scale using Sentinel-2 spectral-temporal metrics and regression-based unmixing with synthetic training data. *Remote Sens. Environ.* **2020**, *246*, 111810. [\[CrossRef\]](#) [\[PubMed\]](#)
12. Audebert, N.; Le Saux, B.; Lefèvre, S. Beyond RGB: Very high resolution urban remote sensing with multimodal deep networks. *ISPRS J. Photogramm. Remote Sens.* **2018**, *140*, 20–32. [\[CrossRef\]](#)
13. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16 × 16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
14. Wang, L.; Li, R.; Wang, D.; Duan, C.; Wang, T.; Meng, X. Transformer meets convolution: A bilateral awareness network for semantic segmentation of very fine resolution urban scene images. *Remote Sens.* **2021**, *13*, 3065. [\[CrossRef\]](#)
15. Wang, L.; Li, R.; Zhang, C.; Fang, S.; Duan, C.; Meng, X.; Atkinson, P.M. UNetFormer: A UNet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery. *ISPRS J. Photogramm. Remote Sens.* **2022**, *190*, 196–214. [\[CrossRef\]](#)
16. Corbane, C.; Syrris, V.; Sabo, F.; Politis, P.; Melchiorri, M.; Pesaresi, M.; Soille, P.; Kemper, T. Convolutional neural networks for global human settlements mapping from Sentinel-2 satellite imagery. *Neural Comput. Appl.* **2021**, *33*, 6697–6720.
17. Qiu, C.; Schmitt, M.; Geiß, C.; Chen, T.H.K.; Zhu, X.X. A framework for large-scale mapping of human settlement extent from Sentinel-2 images via fully convolutional neural networks. *ISPRS J. Photogramm. Remote Sens.* **2020**, *163*, 152–170. [\[CrossRef\]](#) [\[PubMed\]](#)
18. Feng, L.; Xu, P.; Tang, H.; Liu, Z.; Hou, P. National-scale mapping of building footprints using feature super-resolution semantic segmentation of Sentinel-2 images. *GISci. Remote Sens.* **2023**, *60*, 2196154.
19. Schug, F.; Frantz, D.; Okujeni, A.; Hostert, P. Sub-pixel building area mapping based on synthetic training data and regression-based unmixing using Sentinel-1 and-2 data. *Remote Sens. Lett.* **2022**, *13*, 822–832.

20. Arık, A.E.; Paşaoğlu, R.; Emrahaoğlu, N. Sentinel-2 uydu görüntüleri için evrişimli otokodlayıcı sinir ağı ile süper çözünürlük yaklaşımı. *Türk Uzak. Algılama CBS Derg.* **2023**, *4*, 231–241.
21. Li, Y.; Matgen, P.; Chini, M. Extraction of built-up areas using Sentinel-1 and Sentinel-2 data with automated training data sampling and label noise robust cross-fusion neural networks. *Int. J. Appl. Earth Obs. Geoinf.* **2025**, *139*, 104524.
22. Jagannathan, J.; Vadivel, M.T.; Divya, C. Land use classification using multi-year Sentinel-2 images with deep learning ensemble network. *Sci. Rep.* **2025**, *15*, 29047. [[CrossRef](#)] [[PubMed](#)]
23. Rikimaru, A.; Roy, P.S.; Miyatake, S. Tropical forest cover density mapping. *Trop. Ecol.* **2002**, *43*, 39–47.
24. Diek, S.; Fornallaz, F.; Schaepman, M.E.; De Jong, R. Barest pixel composite for agricultural areas using Landsat time series. *Remote Sens.* **2017**, *9*, 1245. [[CrossRef](#)]
25. Chinilin, A.V.; Lozbenov, N.I.; Shilov, P.M.; Fil, P.P.; Levchenko, E.A.; Kozlov, D.N. Synergetic use of bare soil composite imagery and multitemporal vegetation remote sensing for soil mapping (a case study from Samara region's upland). *Land* **2024**, *13*, 2229. [[CrossRef](#)]
26. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, San Francisco, CA, USA, 7–12 June 2015; pp. 3431–3440.
27. Rahman, M.A.; Wang, Y. Optimizing intersection-over-union in deep neural networks for image segmentation. In *Proceedings of the International Symposium on Visual Computing*; Springer: Cham, Switzerland, 2016; pp. 234–244.
28. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional networks for biomedical image segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), Munich, Germany, 5–9 October 2015; pp. 234–241.
29. Chen, L.-C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder–decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*; Springer: Cham, Switzerland, 2018; pp. 801–818.
30. Xie, E.; Wang, W.; Yu, Z.; Anandkumar, A.; Alvarez, J.M.; Luo, P. SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 12077–12090.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.