International Journal of Emerging Research in Engineering, Science, and Management
Vol. 5 | Issue 2 | pp. 96-118 | Apr-Jun 2026.
www.ijeresm.com | eISSN – 2583-4894
Received: 12 Mar 2026 | Revised: 29 May 2026 | Accepted: 10 Jun 2026 | Published: 16 Jun 2026

# Feedback-Guided Parallel Transformer Framework for Remote Sensing Image Semantic Segmentation

[1]K. S. Raghavendra Reddy, [2]B. Sudhakar, [3]K. Venkata Ramanaiah

[1]Research Scholar, Department of Electronics and Communication Engineering, Faculty of Engineering and Technology, Annamalai University, Tamil Nadu- 608002, India.
[2]Associate Professor, Department of Electronics and Communication Engineering, Faculty of Engineering and Technology, Annamalai University, Tamil Nadu- 608002, India.
[3]Department of Electronics and Communication Engineering, Y.S.R. Engineering College of Yogi Vemana University, Proddattur, Andhra Pradesh – 516360, India.

[1]ks.raghavendrareddy@gmail.com, ORCID iD: 0000-0001-7411-4594,
[2]balrajsudhakar@gmail.com, ORCID iD: 0000-0001-6320-8579,
[3]ramanaiahkota@gmail.com, ORCID iD: 0009-0007-5371-3628.

**Corresponding author: K. S. Raghavendra Reddy*

**Abstract:** Accurate semantic segmentation of remote sensing images is essential for applications such as urban planning, environmental monitoring, and land-use analysis. This study proposes a parallel-branch feedback-guided transformer framework for semantic segmentation of remote sensing images in complex scenes. Initially, images obtained from publicly available datasets are pre-processed using bilateral filtering to reduce noise while preserving important edge details. The pre-processed images are forwarded to a hierarchical transformer encoder, where feature representations are progressively extracted. Parallel processing is subsequently performed within the Atrous Spatial Pyramid Pooling-based Densely connected Residual (ASPP-DR) and Dual Attention Mechanism (DAM) modules to capture multi-scale contextual information together with spatial and channel-wise feature dependencies. A dual attention mechanism is incorporated to capture both spatial and channel-wise dependencies, improving contextual feature representation. The extracted multi-level feature maps are passed to a lightweight ASPP-DR module, which enhances contextual feature representation across multiple receptive fields. These enhanced features are subsequently forwarded to a multi-stage decoder that progressively reconstructs the spatial resolution. A feedback pyramid module is integrated within the decoding process to iteratively refine feature representations using previously generated outputs. In parallel, a multi-scale feature aggregation strategy combines features from different levels to produce a more discriminative representation. Finally, a cascaded upsampling decoder generates a high-resolution semantic segmentation map with accurate pixel-level classification. The proposed framework was evaluated on the LoveDA and WHU Building datasets. Experimental results achieved mIoU values of 95.4% and 96.1%, Dice scores of 97.7% and 98.0%, and precision values of 97.8% and 98.1% on the LoveDA and WHU datasets, respectively. The results indicate that the proposed framework effectively handles multi-scale variations while maintaining high segmentation accuracy.

**Keywords:** Remote sensing image segmentation, Vision transformer, Dual attention mechanism, Multi-scale feature learning, Feedback refinement, Semantic segmentation.

## 1 INTRODUCTION

Remote sensing imagery presents significant challenges for image annotation [1]. Complex scenes, large variations in object scales, and intricate spatial relationships make pixel-level annotation of remote sensing images a challenging task [2]. Furthermore, remote sensing image annotation requires substantial time and human effort. Several approaches, including semi-supervised segmentation [3], unsupervised segmentation [4], few-shot learning [5], self-supervised segmentation [6], and weakly supervised segmentation, have been developed to reduce annotation requirements; however, their performance often varies across different remote sensing scenarios.

Despite recent advances in remote sensing image segmentation, several challenges remain when these methods are applied to complex remote sensing imagery. Automated semantic segmentation has become an essential tool in computer vision, and convolutional neural networks (CNNs) have been widely adopted for feature extraction and image understanding tasks [7]. Earlier studies utilized CNN-based architectures to extract features from multiple modalities and perform segmentation through feature fusion strategies [8]. Existing research has also explored multimodal fusion-based semantic segmentation frameworks that integrate high-resolution multispectral imagery and Digital Surface Model data to improve segmentation performance in remote sensing environments [9].

International Journal of Emerging Research in Engineering, Science, and Management
Vol. 5 | Issue 2 | pp. 96-118 | Apr-Jun 2026.
www.ijeresm.com | eISSN – 2583-4894
Received: 12 Mar 2026 | Revised: 29 May 2026 | Accepted: 10 Jun 2026 | Published: 16 Jun 2026

First, many existing segmentation networks struggle to simultaneously capture global contextual information and preserve fine boundary details, often resulting in either semantic inconsistency or inaccurate object delineation. Second, conventional techniques often lack sufficient flexibility to handle the diverse characteristics of different remote sensing datasets [10]. Compared with natural images, remote sensing images present unique feature extraction challenges due to large variations in object scale, complex scene structures, and irregular object boundaries, all of which can negatively affect segmentation accuracy. Furthermore, the development of accurate segmentation models relies heavily on high-quality annotated data, which is often expensive and labor-intensive to obtain [11]. In recent years, several advanced deep learning methods have been proposed for remote sensing image segmentation. To address the above limitations, this study proposes a parallel-branch feedback-guided transformer framework for remote sensing image semantic segmentation.

### 1.1. Motivation

Numerous methods mainly emphasize multi-scale feature learning while frequently overlooking the effective preservation of edge information. This limitation creates difficulties in balancing multi-scale contextual information and edge features, often leading to inadequate feature fusion and inaccurate object boundaries. Models trained on specific datasets often exhibit poor performance when applied to alternative datasets, revealing limited transferability and generalization capability. To address these challenges, this study proposes a parallel-branch feedback-guided transformer framework for remote sensing image semantic segmentation. Traditional encoder-decoder models often fail to capture fine but important semantic details that are spread across different channels. As a result, the semantic output may appear broken, especially for small or thin structures like power lines. These shortcomings highlight the need for methods that can adaptively refine feature representations and concentrate more on essential regions. To address these issues, attention mechanisms have been introduced as a powerful addition to deep learning models for semantic segmentation. This motivates the proposed framework to employ an efficient encoder-decoder architecture with a DAM that assigns adaptive importance to informative features.

### 1.2. Major Contributions

Based on the identified research gaps and challenges, this study proposes a Feedback-Guided Parallel Transformer Framework for remote sensing image semantic segmentation. The main contributions of this work are summarized as follows:

- A parallel-branch feedback-guided transformer framework is proposed for remote sensing image semantic segmentation, enabling effective integration of global contextual information and local spatial details.
- A lightweight Atrous Spatial Pyramid Pooling-based Densely Connected Residual module is introduced to enhance multi-scale feature representation while maintaining computational efficiency.
- A Feedback Pyramid Module (FPM) is incorporated into the decoder to iteratively refine feature representations using previously generated outputs.
- A multi-scale feature aggregation strategy combined with a cascaded upsampling decoder is designed to effectively fuse semantic and spatial information for accurate pixel-level classification.
- Extensive experiments conducted on the LoveDA and WHU Building datasets demonstrate the effectiveness of the proposed framework in terms of segmentation accuracy, boundary preservation, and multi-scale feature representation.

The remainder of this paper is organized as follows. Section 2 reviews the related literature. Section 3 presents the proposed methodology. Section 4 discusses the experimental results and comparative analysis. Finally, Section 5 concludes the paper and outlines future research directions.

## 2 Literature Survey

Semantic segmentation of remote sensing images has received increasing attention due to its importance in land-cover mapping, urban planning, and environmental monitoring. To improve segmentation performance under limited annotation conditions, Chen et al. [12] proposed a semi-supervised boundary segmentation network that emphasizes boundary information during feature learning. Yang et al. [13] introduced EasySeg, which combines active learning and interactive learning strategies to improve domain adaptation performance in remote sensing imagery. An et al. [14] further investigated contextual feature learning and reported that interactions between dataset-level and image-level contextual information can improve segmentation accuracy. Although these methods improve feature representation, accurately preserving fine structural details and object boundaries remains challenging in complex remote sensing scenes.

Interactive segmentation techniques have also been explored to improve segmentation quality. Marinov et al. [15] presented a comprehensive review of deep interactive segmentation methods and highlighted the growing role of user-guided learning in segmentation systems. Huang et al. [16] developed an efficient click-based segmentation framework using an improved Vision Transformer architecture, while Du et al. [17] proposed SegVol for universal interactive volumetric image segmentation. These methods achieve promising results; however, they rely on user interactions during the segmentation process, which limits their applicability to fully automated remote sensing image analysis.

International Journal of Emerging Research in Engineering, Science, and Management
Vol. 5 | Issue 2 | pp. 96-118 | Apr-Jun 2026.
www.ijeresm.com | eISSN – 2583-4894
Received: 12 Mar 2026 | Revised: 29 May 2026 | Accepted: 10 Jun 2026 | Published: 16 Jun 2026

Transformer-based segmentation models have recently shown strong capability in learning contextual relationships across complex scenes. Liu et al. [18] proposed MiLNet to improve RGB-T semantic segmentation through multiplex interactive learning and effective feature interaction. Lin et al. [19] introduced AdaptiveClick, which incorporates a click-aware transformer and adaptive focal loss to improve segmentation performance. Xing et al. [20] combined graph deep learning with laser point cloud information for hand segmentation and demonstrated the effectiveness of integrating complementary feature representations. While these approaches improve segmentation accuracy, challenges still remain in balancing multi-scale feature learning, boundary preservation, and computational efficiency. These limitations motivate the development of the proposed Feedback-Guided Parallel Transformer Framework.

Wang et al. [21] developed a plug-and-play Mamba-based decoder that incorporates Dense Spatial Pyramid Pooling for effective multi-scale semantic feature encoding and Pyramid Fusion Mamba (PFM) for reducing semantic redundancy during feature fusion. The proposed decoder achieved mIoU values of 70.8%, 84.8%, 88.0%, and 54.8% on the OpenEarthMap, ISPRS Vaihingen, Potsdam, and LoveDA datasets, respectively. Despite its effective multi-scale feature encoding capability, preserving fine boundary details remains challenging, which may lead to less accurate object delineation. Wang et al. [22] introduced BEMS-UNetFormer, a boundary-enhanced multi-scale semantic segmentation network based on UNetFormer for improving segmentation accuracy in remote sensing images containing blurred object boundaries and scale variations. Experimental results on the Potsdam and Vaihingen datasets achieved mIoU values of 86.12% and 83.10%, respectively. Although the boundary enhancement strategy improves edge representation, maintaining global semantic consistency in complex scenes remains challenging.

Wang et al. [23] presented a diffusion-guided feature modeling network (DiffMamba) based on diffusion modeling and State Space Models for remote sensing image segmentation. DiffMamba employs a hybrid architecture that combines CNNs and Transformers for its encoder structure. It is also integrated with an efficient phase sensing module (EPSM), a multi-view transformer module (MVTrans), a semantic diffusion alignment module (SDAM), and a coordinate state space model. Experiments carried out on the ISPRS Vaihingen, ISPRS Potsdam, and GID-15 datasets indicate that the DiffMamba method significantly enhances semantic segmentation accuracy when compared to current benchmark techniques. However, the integration of multiple modules increases architectural complexity and computational cost, which may limit its suitability for real-time applications.

Zhu et al. [24] designed dual-branch network named global–local feature fusion network for semantic segmentation of high-resolution RS images. Initially residual network used in main branch as local feature extractor. Especially this research introduces VMamba as an auxiliary branch encoder designed to supply global information for the main branch. In parallel, multi-scale feature refinement module is developed to effectively leverage global information for reducing detail loss during extraction of global features. Moreover, semantic bridging fusion module implemented to integrate global and local features, enhancing refined feature representations. Evaluation results show that GLFFNet attains mIoU scores of 84.01% on ISPRS Vaihingen, 87.54% on ISPRS Potsdam, and 54.73% on LoveDA, along with mF1 scores of 91.11%, 93.23%, and 70.07% on these respective datasets. However, the dual-branch design increases model complexity and may introduce challenges in effectively integrating global and local feature representations.

Li et al. [25] designed RS image semantic segmentation network called CSNet, which is based on coordinate attention and skip connections, enhancing precision of segmentation and aiding in restoration of spatial configurations. Compared with traditional models, CSNet achieved excellent results of 81.4%, 70.3% and 90.5% of Dice coefficient, mIoU, and overall accuracy respectively. Nevertheless, handling large-scale object variations and complex scene diversity remains challenging. Table 1 illustrates the existing research analysis of segmentation for remote sensing images.

Table 1. Comparative Analysis of Existing Remote Sensing Image Segmentation Methods

| Method | Performance | Limitations |
|---|---|---|
| DSPP by Wang et al. [21] | Efficient results compare to state-of-the-art methods in terms of mIoU of 70.8% | Capability of the model to preserve minute boundary details is limited |
| UNetFormer by Wang et al. [22] | Results are evaluated using two datasets Potsdam and Vaihingen, attaining 86.12% and 83.10% MIoU. | However, the method fails to keep the global semantic consistency in scenes with complex layouts. |
| MVTrans by Wang et al. [23] | DiffMamba method significantly enhances semantic segmentation accuracy when compared to current benchmark techniques. | The system architecture is highly complex and computationally expensive, which limits its suitability for real-time applications. |
| GLFFNet by Zhu et al. [24] | Evaluation results show that GLFFNet attains mIoU scores of 84.01% on ISPRS Vaihingen, 87.54% on ISPRS Potsdam. | The proposed method increases model complexity. |
| CSNet by Li et al. [25] | CSNet achieved excellent results of 81.4% and 70.3% Dice coefficient and mIoU | Less robust in tackling large-scale variations and complex scene diversity. |

International Journal of Emerging Research in Engineering, Science, and Management
Vol. 5 | Issue 2 | pp. 96-118 | Apr-Jun 2026.
www.ijeresm.com | eISSN – 2583-4894
Received: 12 Mar 2026 | Revised: 29 May 2026 | Accepted: 10 Jun 2026 | Published: 16 Jun 2026

### 2.1. Problem Statement

Semantic segmentation plays an essential role in remote sensing, and Deep Convolutional Neural Networks have become a common choice for developing automated segmentation systems. These models are effective at capturing layered visual patterns and can generate accurate segmentation outputs. However, conventional CNN architectures exhibit several limitations when dealing with high-resolution remote sensing images, where segmentation performance may degrade due to scene complexity and scale variations. For example, the DSPP model proposed by Wang et al. [21] demonstrated effective multi-scale feature extraction; however, its ability to retain fine-grained boundary details remains limited.

Similarly, the UNetFormer-based framework developed by Wang et al. [22] improves boundary representation but faces challenges in maintaining global semantic consistency in complex scenes. Likewise, the MVTrans-based framework presented by Wang et al. [23] employs multiple processing modules, resulting in increased architectural complexity and computational cost. GLFFNet proposed by Zhu et al. [24] achieved competitive segmentation performance; however, the dual-branch design increases model complexity. Similarly, the CSNet model developed by Li et al. [25] faces challenges in handling large-scale object variations and complex scene diversity.

However, most of the existing methods utilize independent multi-scale feature extraction and decoding strategies without exploiting iterative feedback refinement. As a result, challenges such as small-object segmentation, boundary ambiguity, and inadequate preservation of fine structural details remain unresolved. To address these challenges, this study proposes a parallel-branch feedback-guided transformer framework that integrates multi-scale feature enhancement, feedback-based feature refinement, and cascaded upsampling reconstruction within a unified architecture. The proposed framework is designed to improve contextual feature representation, boundary preservation, and segmentation accuracy while maintaining computational efficiency.

## 3 Proposed Methodology

This section presents the proposed parallel-branch feedback-guided transformer framework for remote sensing image semantic segmentation. Initially, the input images are collected from publicly available datasets. Fig. 1 illustrates the workflow of the proposed parallel-branch feedback-guided transformer framework for remote sensing image semantic segmentation.

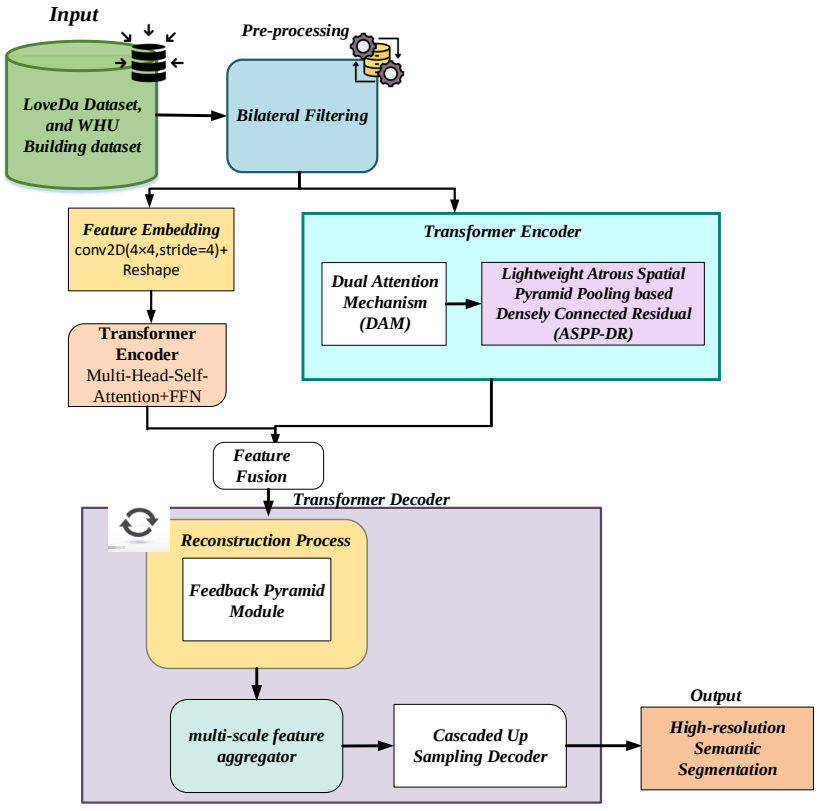



Fig. 1. Workflow of the Proposed Feedback-Guided Parallel Transformer Framework

International Journal of Emerging Research in Engineering, Science, and Management
Vol. 5 | Issue 2 | pp. 96-118 | Apr-Jun 2026.
www.ijeresm.com | eISSN – 2583-4894
Received: 12 Mar 2026 | Revised: 29 May 2026 | Accepted: 10 Jun 2026 | Published: 16 Jun 2026

The proposed architecture combines complementary modules to effectively capture contextual information, preserve boundary details, and enhance multi-scale feature representation. The input images are pre-processed using bilateral filtering to preserve edge information while reducing image noise. The pre-processed image is first converted into patch embeddings using a Conv2D layer with a kernel size of 4×4 and a stride of 4. The resulting feature representations are processed by a hierarchical transformer encoder comprising MHSA and FFN blocks to capture long-range semantic dependencies and hierarchical contextual information.

The extracted feature maps are subsequently enhanced through the ASPP-DR and DAM, where parallel processing is performed to learn multi-scale contextual information together with complementary spatial and channel-wise attention features. DAM selectively emphasizes informative spatial regions and feature channels. The encoder produces multilevel feature maps, which are subsequently passed to the ASPP-DR module. The features are enhanced to capture rich contextual information. The enhanced features are passed to a multi-stage decoder that progressively reconstructs the spatial resolution of the image. During this stage, the feedback module utilizes previously generated outputs to iteratively refine feature representations and improve segmentation quality. Finally, multi-scale feature aggregation and cascaded upsampling facilitate the effective fusion of semantic and spatial information, enabling accurate segmentation of objects with varying scales and complex structures.

### 3.1. Preprocessing through Bilateral Filtering

The raw input images are pre-processed using Bilateral Filtering [26] to remove noise while preserving important edge details, which are critical for accurate boundary detection during segmentation. Unlike traditional filtering methods, BF computes each pixel value by considering both the spatial proximity of neighboring pixels and their intensity similarity. This ensures that pixels that are both nearby and similar in appearance contribute more to the filtering process, thereby maintaining sharp boundaries. It preserves edge information while removing noise from remote sensing images, thereby improving the quality of subsequent feature extraction. In this formulation, the spatial closeness function is modeled using a Gaussian distribution, which assigns higher weights to pixels that are closer in distance. The closeness function is formulated as:

$$C(\varepsilon, y) = e^{-\frac{1}{2}\left(\frac{g(\varepsilon,y)}{\sigma_g}\right)^2} \tag{1}$$

$$g(\varepsilon, y) = g(\varepsilon - y) = \|\varepsilon - y\| \tag{2}$$

$$t(\varphi, h) = \delta(\varphi - h) = e^{-\frac{1}{2}\left(\frac{\delta(\varphi,h)}{\sigma_h}\right)^2} \tag{3}$$

Here, $\varepsilon$denotes position of the neighboring pixel, while $y$ represents the position of the target pixel. The terms $g(\varepsilon, y)$ and $\|\varepsilon - y\|$ denote the Euclidean distance between the neighboring pixel and the target pixel, representing spatial distance between pixels. Moreover, $\sigma_g$represents the spatial standard deviation, and $C(\varepsilon, y)$ denotes the corresponding spatial closeness weight.

$$\delta(\varphi, h) = \delta(\varphi - h) = \|\varphi - h\| \tag{4}$$

The above formulation measures the distance between the two intensity values $\varphi$ and $h$. Finally, this stage produces noise-reduced and edge-preserved images for subsequent feature extraction.

### 3.2. Feature Learning

The pre-processed image is simultaneously forwarded to two complementary feature extraction branches. The first branch consists of a feature embedding layer followed by a transformer encoder incorporating MHSA and FFN blocks that capture global semantic dependencies and long-range contextual information. In parallel, the second branch performs hierarchical feature extraction, where the image is divided into patches and processed to generate multi-level feature representations. The hierarchical design employs transformer blocks at multiple stages to progressively learn feature representations at different spatial scales. This process extracts hierarchical multi-scale features and captures long-range contextual dependencies in complex remote sensing scenes. In each block, the DAM [27] analyzes relationships among different image regions, allowing the model to capture global contextual dependencies while preserving important spatial details. It selectively emphasizes important spatial regions and informative feature channels. DAM consists of a Channel Attention Module (CAM) and a Spatial Attention Module (SAM). Fig. 2 illustrates Workflow of Channel Attention Module.

International Journal of Emerging Research in Engineering, Science, and Management
Vol. 5 | Issue 2 | pp. 96-118 | Apr-Jun 2026.
www.ijeresm.com | eISSN – 2583-4894
Received: 12 Mar 2026 | Revised: 29 May 2026 | Accepted: 10 Jun 2026 | Published: 16 Jun 2026

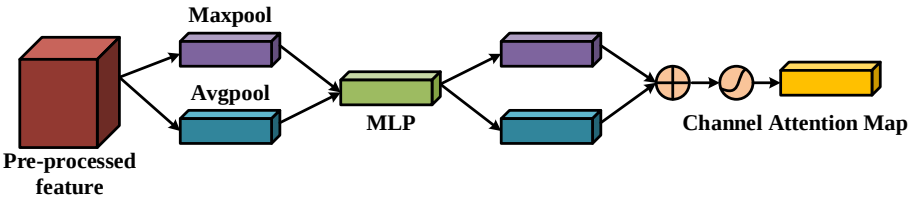



Fig. 2. Workflow of the Channel Attention Module

CAM treats each channel of the feature map as a separate detector and learns to emphasize the most informative ones. It learns the importance of each feature channel and strengthens relevant semantic features while removing irrelevant information. To capture channel-wise importance, the model compresses spatial information using both Global Average Pooling and Global Max Pooling. These pooled representations are then passed through a shared two-layer neural network to learn nonlinear relationships among channels. The outputs from both paths are combined and activated using a sigmoid function to generate channel-wise weights, which are applied to the original feature map to enhance relevant features and suppress less informative responses.

$$C_{CAM} = \sigma\left(MLP\left(avgpool(H)\right) + MLP\left(max\,pool(H)\right)\right) = \sigma\left(M_1\left(M_0\left(H_{avg}^{CAM}\right)\right) + M_1\left(M_0\left(H_{max}^{CAM}\right)\right)\right) \quad (5)$$

Here, $MLP$ denotes the multilayer perceptron, $\sigma$ is the sigmoid activation function, $H$ is the feature map of size $M \times CAM$, $CAM$ representing the number of channels in the CAM module. $C_{CAM}$ is the resulting channel attention map. Workflow of Spatial Attention Module is illustrated in Fig. 3.

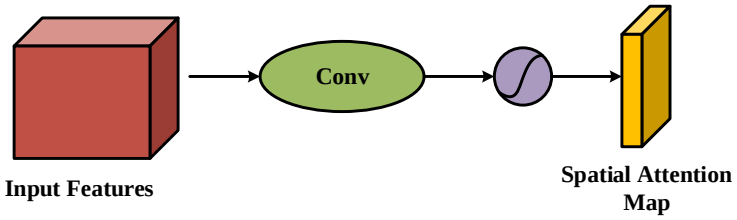



Fig. 3. Workflow of the Spatial Attention Module

SAM identifies significant spatial locations and object regions, allowing the network to focus on meaningful areas within the image. Given an input feature map, spatial information is compressed along the channel dimension using both average pooling and max pooling, as these operations capture complementary cues. Average pooling reflects the general spatial distribution, whereas max pooling highlights the most prominent responses. The resulting two spatial maps are concatenated and passed through a convolution layer to learn spatial dependencies. A sigmoid activation is then applied to generate a spatial attention map, which is multiplied with the refined feature map to highlight significant regions and suppress irrelevant background information.

$$C_S(H) = \sigma\left(f\left([avgpool(H); maxpool(H)]\right)\right) = \sigma\left(f\left(\left[H_{avg}^S; H_{max}^S\right]\right)\right) \quad (6)$$

Here, f denotes the convolution operation, and [·] represents feature concatenation along the channel dimension. Finally, the DAM generates multi-level feature representations that capture both local and global contextual dependencies.

### 3.3. Feature Enhancement Through ASPP-DR

Unlike conventional ASPP architectures, the proposed ASPP-DR incorporates dense residual feature propagation within atrous pyramid learning, enabling efficient multi-scale contextual representation while reducing information loss. The informative feature maps generated by the previous stage are passed to the Lightweight ASPP-DR module. It captures contextual information at multiple scales and improves feature reuse through dense residual connections, facilitating the detection of objects with varying sizes. Fig. 4 illustrates the Architecture of Parallel-Branch Feedback-Guided Transformer Framework for segmentation of remote sensing images. This module is positioned between the hierarchical transformer encoder and the decoder to strengthen feature representation across multiple scales. Within the ASPP-DR module, parallel atrous convolution branches capture contextual information at multiple receptive fields, while the DAM simultaneously learns complementary spatial and channel attention, enabling effective feature enhancement before decoding. Instead of using the standard parallel ASPP structure, it adopts a cascaded arrangement in which atrous convolution layers are applied sequentially with increasing dilation rates r={3,6,12}.

International Journal of Emerging Research in Engineering, Science, and Management
Vol. 5 | Issue 2 | pp. 96-118 | Apr-Jun 2026.
www.ijeresm.com | eISSN – 2583-4894
Received: 12 Mar 2026 | Revised: 29 May 2026 | Accepted: 10 Jun 2026 | Published: 16 Jun 2026

Each layer utilizes a 3×3 convolution to capture contextual information at different receptive fields. Before each atrous convolution, a 1×1 convolution is applied to reduce the number of channels, thereby lowering computational cost while retaining important features. Fig. 5 demonstrates the Workflow of the Lightweight Atrous Spatial Pyramid Pooling Model.

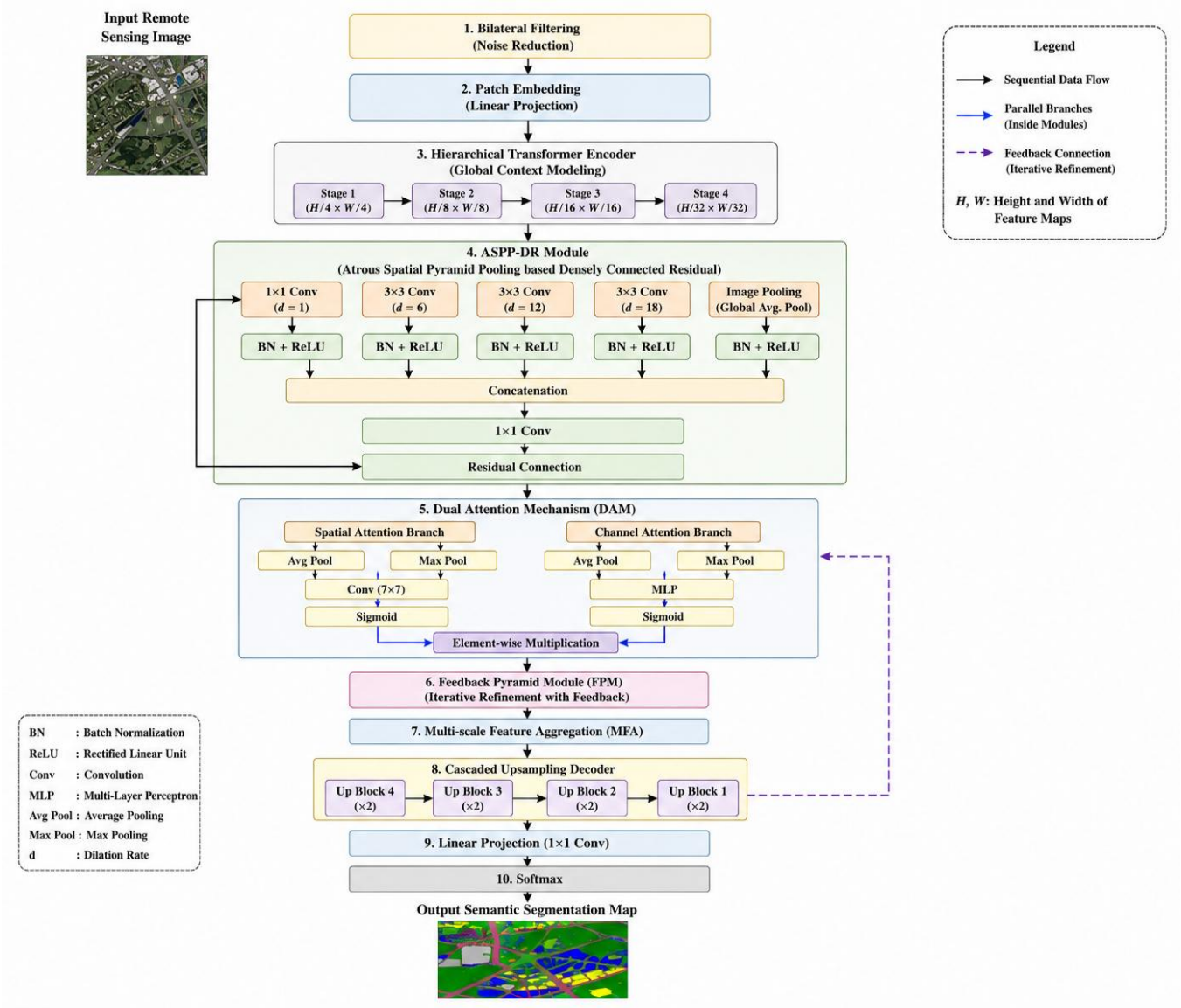



Fig. 4. Architecture of the Proposed Feedback-Guided Parallel Transformer Framework

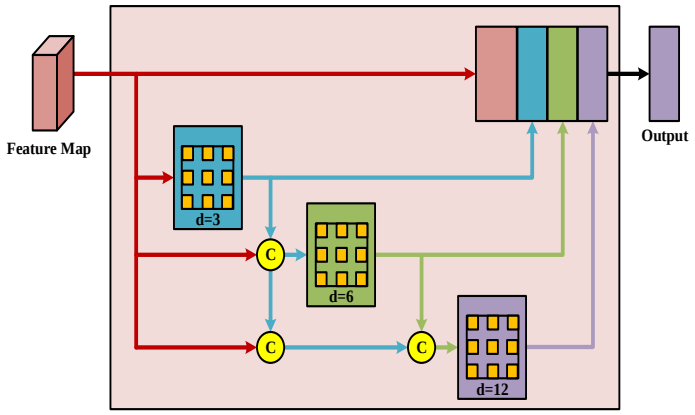



Fig. 5. Workflow of the Lightweight ASPP-DR Module

International Journal of Emerging Research in Engineering, Science, and Management
Vol. 5 | Issue 2 | pp. 96-118 | Apr-Jun 2026.
www.ijeresm.com | eISSN – 2583-4894
Received: 12 Mar 2026 | Revised: 29 May 2026 | Accepted: 10 Jun 2026 | Published: 16 Jun 2026

The key idea is the dense connection strategy, where the output of each stage is concatenated with all preceding feature maps, forming a progressively enriched feature pyramid. This allows the network to combine local details from smaller dilation rates with broader contextual information from larger dilation rates, improving segmentation performance for objects of varying sizes.

$$A_{ASPP}^{l} = D_{3,d_1}\left(Conca(A_{in}, A_{ASPP}^{1}, .., A_{ASPP}^{l-1})\right) \quad (7)$$

Here, $A_{in}$ represents the input feature map, while $A_{ASPP}^{l}$ and $A_{ASPP}^{l-1}$ denote the feature map at the l-th layer and the aggregated feature maps from the preceding layers, respectively. $D_{3,d_1}$ denotes a 3 × 3 atrous convolution with dilation rate d, which controls the receptive field size at each layer. This design ensures effective fusion of multi-scale contextual information while maintaining computational efficiency. The workflow of the proposed Densely Connected Residual Network (DCRN) is illustrated in Fig. 6.

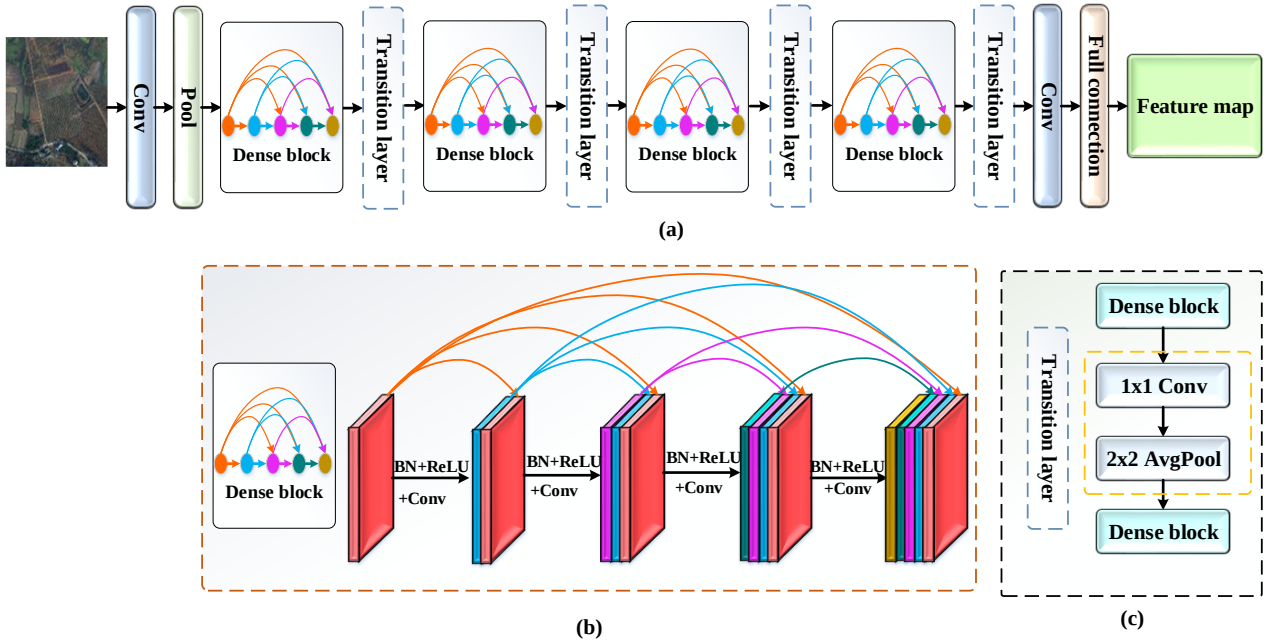



Fig. 6. Workflow of the Densely Connected Residual Module in ASPP-DR

A DCRN-based feature extraction strategy is employed to obtain robust and discriminative feature representations from remote sensing images. By leveraging multi-scale feature learning, feature maps at different levels are extracted, where deeper layers capture richer semantic representations. To further strengthen feature representation, skip connections are introduced between deeper dense blocks, enabling the model to effectively utilize high-level semantic features that are generally more stable and informative than low-level features. Each dense block consists of various densely connected units, where the result of every unit is concatenated with the output of all preceding units.

This dense connectivity ensures feature reuse, strengthens information flow, and reduces gradient degradation during training. Unlike the conventional DenseNet architecture, a residual learning block is introduced between the dense block and the transition layer. The residual connection enables the network to preserve important feature information from earlier layers, while facilitating the learning of residual mappings, thereby improving feature representation and training stability. Subsequently, the transition layer, composed of a 1 × 1 convolution followed by pooling, compresses the feature dimensions and reduces computational complexity. By integrating dense connectivity with residual learning, the proposed DCRN effectively captures both local details and contextual information in remote sensing scenes.

### 3.4. Feedback Pyramid Module

In this research, the Feedback Pyramid Module continuously refines feature representations using previous decoder outputs, thereby enhancing boundary accuracy and semantic consistency. Within each decoder stage, the output feature maps generated in the first stage are reused as auxiliary inputs to the corresponding layers in the second stage. The feedback paths enable high-level semantic representations extracted during the first stage to guide the refinement of lower-level features in the second stage. Consequently, semantic information is propagated iteratively across decoder layers, improving boundary preservation and segmentation consistency. The workflow of the proposed Feedback Pyramid Module is illustrated in Fig. 7. The FPM is trained jointly with the entire segmentation framework. During backpropagation, gradients are propagated through both the feed-forward and feedback paths, allowing the feedback connections to be optimized simultaneously with the encoder and decoder parameters.

International Journal of Emerging Research in Engineering, Science, and Management
Vol. 5 | Issue 2 | pp. 96-118 | Apr-Jun 2026.
www.ijeresm.com | eISSN – 2583-4894
Received: 12 Mar 2026 | Revised: 29 May 2026 | Accepted: 10 Jun 2026 | Published: 16 Jun 2026

The input feature from the ASPP-DR stage is first passed through a feedback connection structure, which reuses and refines earlier outputs to improve feature quality. This refined feature is then processed by a pyramid non-local block that captures long-range dependencies at multiple scales, enabling a better understanding of both local and global contexts.

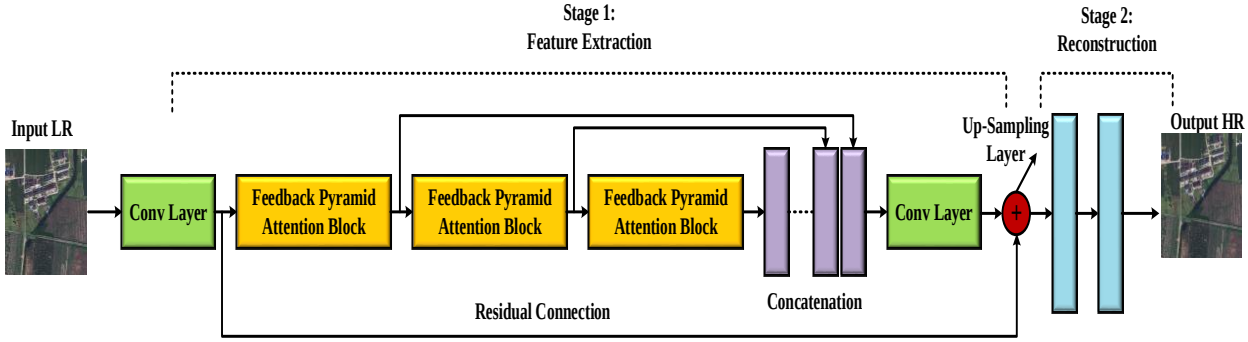



Fig. 7. Workflow of the Proposed Feedback Pyramid Module

Finally, local residual learning is applied by adding the processed feature back to the original input, ensuring that important information is preserved while enhancing discriminative details. The operation is defined as:

$$FB_g = FB_{g-1} + f_{PN}\left(f_{FBC}\left(FB_{g-1}\right)\right) \tag{9}$$

Here, $f_{FBC}(.)$ and $f_{PN}(.)$ denote the feedback connection structure and pyramid non-local block, respectively. $FB_g$ represents the refined feature map at the current stage, while $FB_{g-1}$ denotes the feature map obtained from the previous stage. The addition operation implements residual learning, enabling the model to preserve previously learned feature information while incorporating refined contextual representations for improved segmentation performance. Moreover, the feedback connection structure is designed to iteratively refine feature representations by reusing high-level semantic information to improve lower-level features. It preserves the spatial dimensions while refining the feature representation.

Therefore, the outputs of the feedback connection structure and pyramid non-local block maintain the original feature dimensions. Initially, the input feature map is propagated through feed-forward connections to initialize the refinement process. In the first stage, each layer receives information from its previous two layers, enabling richer feature interactions and progressive feature extraction. In the second stage, the model refines these features by linking each layer with its corresponding output from the first stage, effectively introducing feedback that enhances feature quality and semantic consistency. The outputs from both stages are then forwarded to subsequent layers, improving information flow across the network. The initialization stage is defined as:

$$Y_0 = \sigma\left(M_0 * FB_{g-1}\right) \tag{10}$$

The first refinement stage is defined as:

$$Y_1^j = \sigma\left(M_1^j * \left[Y_1^{j-1}, Y_1^{j-2}\right]\right) \tag{11}$$

The second refinement stage is defined as:

$$Y_2^j = \sigma\left(M_2^j * \left[Y_2^{j-1}, Y_1^j\right]\right) \tag{12}$$

Here, $Y_i^j$ represents the feature map of the $i^{th}$ convolutional layer in the $j^{th}$ stage. The initialization conditions define the starting feature representations, while $Y_0^2 = Y_1^1$ indicates that the second refinement stage utilizes the output generated during the first stage. $Y_0^1 = Y_0^1$ are the initial conditions, while $Y_0^2 = Y_1^1$ indicates that the second stage starts using the output of the first stage. $M_0$, $M_i^1, M_i^2$ denote the convolutional weight parameters associated with different layers. The symbol $*$ denotes convolution operation, whereas $[\cdot]$ denotes channel wise concatenation of feature maps. The residual learning strategy incorporated within the Feedback pyramid Module facilitates stable gradient propagation and mitigates gradient degradation during training. Furthermore, the feedback mechanism promotes iterative feature refinement while preserving semantic consistency, resulting in faster convergence and improved training stability. Experimental observations indicate that the proposed architecture achieves stable convergence. The pyramid non-local mechanism is introduced to effectively capture long-range dependencies and contextual relationships among image regions.

International Journal of Emerging Research in Engineering, Science, and Management
Vol. 5 | Issue 2 | pp. 96-118 | Apr-Jun 2026.
www.ijeresm.com | eISSN – 2583-4894
Received: 12 Mar 2026 | Revised: 29 May 2026 | Accepted: 10 Jun 2026 | Published: 16 Jun 2026

Instead of computing attention at a single scale, this approach builds a pyramid structure where features are processed at different resolutions, allowing the model to understand both fine local details and broader global context. At each scale, a non-local operation computes the similarity between features and aggregates contextual information. The outputs from all scales are then concatenated and fused. This multi-scale aggregation enhances feature representation while maintaining computational efficiency, enabling the model to better handle complex scenes. The formulation is given as:

$$\begin{cases} W = Y + PN_{LB}(Y) \\ PN_{LB}(Y) = \delta\left(Concat(PA_1, PA_2, PA_3)\right) \\ PA_j = \sum_{i=1}^{U} b_i^j y_i \\ b^j = soft\,max\left(M_l^j Y^j\right) \end{cases} \tag{13}$$

Here, $PN_{LB}(Y)$ denotes the pyramid non-local operation, $j$ represents the scale parameter, $PA_j$ denotes the context modeling module at the $j^{th}$ scale, $concat(\cdot)$ represents the feature concatenation operation [28]-[29].

### 3.5. Multi-scale Feature Aggregator

The Multiscale Fusion (MF) Module [30] is designed to effectively combine fine-grained local features from the encoder with high-level semantic information from earlier decoder stages. This fusion ensures that both detailed spatial information and abstract contextual knowledge are preserved during reconstruction. The workflow of the Multiscale Feature Aggregator Module is illustrated in Fig. 8.

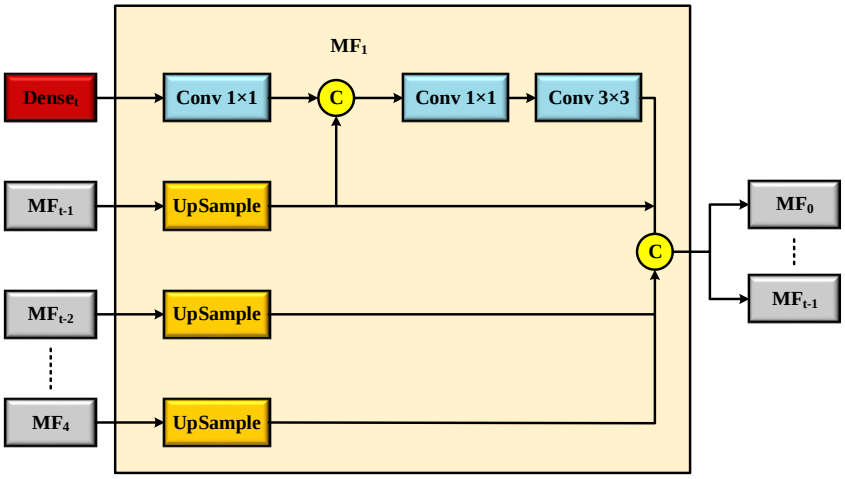



Fig. 8. Workflow of the Multi-Scale Feature Aggregation Module

The decoder is composed of multiple MF modules connected through dense links, allowing features from earlier stages to be reused and improving information flow across the network. This dense connectivity strengthens feature propagation and helps the model learn richer feature representations. Each MF module processes inputs by first reducing channel dimensions using $1 \times 1$ convolutions, which lowers computational cost while retaining essential information. The features are then upsampled to match spatial dimensions and concatenated with features from other stages, enabling effective multi-scale integration. A $3 \times 3$ convolution is subsequently applied to capture contextual relationships within the fused features. Every convolution operation in the module follows a sequence of Batch Normalization (BN), convolution, and ReLU activation, ensuring stable training, efficient feature extraction, and nonlinear transformation. The output of each MF module is a refined feature map that combines both local details and global semantic context, contributing to more accurate segmentation results.

The two concatenation steps within the MF modules are used to combine features from different sources and scales. In the first concatenation, a skip connection is applied to merge the feature map from the corresponding encoder layer with the output of the previous MF module. Before this fusion, the decoder feature is upsampled so that both inputs share the same spatial resolution, enabling proper alignment. This step ensures that fine spatial details from the encoder are effectively integrated with the progressively decoded features. In the second concatenation, features from all earlier MF modules are combined following a dense connection strategy to promote feature reuse and richer representation learning. Since concatenation requires matching spatial sizes, all lower-resolution feature maps from previous MF modules are first upsampled to the highest resolution among them.

International Journal of Emerging Research in Engineering, Science, and Management
Vol. 5 | Issue 2 | pp. 96-118 | Apr-Jun 2026.
www.ijeresm.com | eISSN – 2583-4894
Received: 12 Mar 2026 | Revised: 29 May 2026 | Accepted: 10 Jun 2026 | Published: 16 Jun 2026

After resizing, these multi-scale features are concatenated to form the output of the current MF module, which is then passed to subsequent modules for further refinement. The MF4 module is an exception, as it only receives input from the encoder and does not perform multi-scale fusion due to the absence of prior MF outputs.

### 3.6. Cascaded Upsampling Decoder

In the cascaded decoder blocks, the feature maps are progressively upsampled until they match the original image resolution. Fig. 9 illustrates the workflow of the Cascaded Upsampling Decoder.

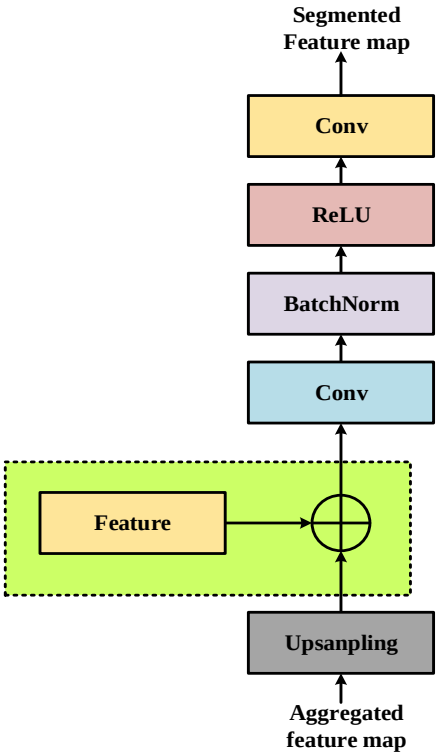



Fig. 9. Workflow of the Cascaded Upsampling Decoder

Here, Cascaded Upsampling Decoder (CUD) [31] is used to improve computational efficiency while maintaining high segmentation accuracy. This module progressively reconstructs the spatial resolution of feature maps through multiple upsampling stages, allowing the decoder to produce an output with the same spatial resolution as the input image. At each stage, the upsampled decoder feature is fused with the corresponding encoder feature through concatenation, ensuring that both high-level semantic information and low-level spatial details are preserved. The combined features are then passed through a convolution layer to adjust the number of channels, followed by normalization and a nonlinear activation to stabilize training and improve feature representation. Finally, a depthwise convolution is applied to efficiently capture fine local patterns and texture information with fewer parameters compared to standard convolution. The process is expressed as:

$$Z = conv_{CUD}\left(\sigma\left(norm\left(conv\left(cat\left(feature, Up(y)\right)\right)\right)\right)\right) \tag{14}$$

Here, $y$ denotes the aggregated feature map from the previous decoder stage, $Up(\cdot)$ represents bilinear interpolation-based upsampling, $\sigma$ denotes the activation function, $norm$ represents batch normalization, and $Z$ denotes the refined high-resolution feature map with enhanced spatial and semantic information. The decoder is composed of four CUD blocks, of which only the first three perform feature fusion with the corresponding encoder outputs, as the encoder provides three levels of feature representations. In these initial stages, the upsampled decoder features are concatenated with encoder features to combine high-level semantic information with low-level spatial details.

The remaining CUD block operates without feature fusion and focuses solely on further upsampling and feature refinement. Each CUD block simultaneously performs resolution enhancement, feature merging, and feature processing, making the decoding process efficient and compact. The inclusion of convolution improves the model's ability to recover fine details and texture information while keeping the number of parameters and computational cost low.

International Journal of Emerging Research in Engineering, Science, and Management
Vol. 5 | Issue 2 | pp. 96-118 | Apr-Jun 2026.
www.ijeresm.com | eISSN – 2583-4894
Received: 12 Mar 2026 | Revised: 29 May 2026 | Accepted: 10 Jun 2026 | Published: 16 Jun 2026

Overall, the encoder extracts rich semantic features, skip connections preserve detailed spatial information, and the decoder progressively integrates both to produce an accurate high-resolution segmentation map. Algorithm 1 presents the overall proposed transformer-based framework for remote sensing image semantic segmentation.

**Algorithm 1:** Overall Proposed Feedback-Guided Parallel Transformer Framework

**Input:** Collect input images from the dataset $d$ =LoveDa and WHU building dataset
**1.** Pre-processing: $P = Bi\_filter(d)$
Result edge preserved images $P$
**2.** Hierarchical transformer encoding
**3.** $C$ =divide patches $(p)$, Generate patch embeddings
**4.** Feed embeddings into Hierarchical transformer encoder
**5.** Apply $E = DAM(C)$
**6.** Result multilevel feature maps
**7.** Feature enhancement $Fea = ASPP - DR(E)$
**8.** Perform atrous convolutions with Multiple dilation rates
**9.** Apply dense feature concatenation and residual learning $Fea$
**10.** Generate enhanced feature map
**11.** Feedback Pyramid Refinement $A = FPM(E, Fea)$
**12.** Feed$Fea$ through feedback connection structure
**13.** Perform stage-1 forward feature propagation
**14.** Perform stage-2 feedback guided refinement
**15.** Apply Pyramid Non-Local Attention
**16.** Generate refined feature map $Fea$
**17.** Perform Multiscale feature aggregation$N = (A, E, Fea)$
**18.** Collect encoder and decoder features
**19.** Concatenate Multiscale feature representations
**20.** Generate aggregated feature map $N$
**21.** Cascaded Upsampling Decoder
**22.** for each decoder stage do
**23.** Upsampled feature map
**24.** Fuse with corresponding encoder features
**25.** Apply convolution +batch normalization+ ReLU
**26.** Apply depthwise convolution
**27.** end for
**28.** Generate high resolution decoded feature map $M$
**29.** Compute pixel wise class probabilities
**30.** Generate Segmentation map $S$
**31.** Return $S$
Output: Segmented map $S$

### 3.7. Rationale and Implication of the Proposed Methodology

The rationale behind the proposed framework is to address the limitations of existing semantic segmentation methods in handling complex remote sensing scenes and multi-scale objects. Bilateral filtering is used to preserve edge information and suppress noise before feature extraction. The Hierarchical Transformer Encoder with the DAM effectively captures both local spatial details and long-range contextual dependencies. The ASPP-DR module strengthens multi-scale contextual learning through dense and residual feature propagation. The FPM continuously refines feature representations using high-level semantic guidance, while the multi-scale feature aggregator enhances feature fusion across different resolutions. The Cascaded Upsampling Decoder reconstructs detailed segmentation maps by progressively restoring spatial information. As a result, the proposed framework improves segmentation accuracy, boundary delineation, and robustness, making it suitable for remote sensing applications such as urban planning and environmental assessment.

International Journal of Emerging Research in Engineering, Science, and Management
Vol. 5 | Issue 2 | pp. 96-118 | Apr-Jun 2026.
www.ijeresm.com | eISSN – 2583-4894
Received: 12 Mar 2026 | Revised: 29 May 2026 | Accepted: 10 Jun 2026 | Published: 16 Jun 2026

## 4 Results and Discussion

This section presents the results obtained using the proposed semantic segmentation framework. The dataset description is presented first, followed by the experimental results and performance analysis. Additionally, the hardware and software specifications used in this research are provided. The experiments were conducted using an Intel i7-6700 processor to handle the computational requirements of model training and evaluation. The system was equipped with 16 GB RAM to support model execution and data processing. A 64-bit operating system was used to efficiently utilize the available hardware resources and memory capacity. Tables 2–5 present the hyperparameter settings used for the proposed framework, including the ASPP-DR module, DAM, Feedback Pyramid Module, and transformer components. Hyperparameter values were selected empirically based on preliminary experiments and validated using the training subset to achieve a balance between segmentation accuracy and computational efficiency.

Table 2. Hyperparameter Values for the Proposed Framework

| **Hyperparameter** | **Value** |
|---|---|
| **Input Shape** | (128, 128, 3) |
| **Batch Size** | 16 |
| **Epochs** | 300 |
| **Optimizer** | Adam |
| **Learning Rate** | 1.00E-04 |
| **Loss Function** | Categorical CE |
| **Normalization** | 255.0 |
| **Encoder Hyperparameters** | |
| **Layer Stage** | **Filters** |
| **Conv Block 1** | 32 |
| **Conv Block 2** | 32 |
| **Conv Block 3** | 64 |
| **Conv Block 4** | 64 |
| **Conv Block 5** | 128 |
| **Conv Block 6** | 128 |
| **Conv Block 7** | 256 |
| **Conv Block 8** | 256 |
| **Conv Block 9** | 512 |
| **Conv Block 10** | 512 |

Table 3. Hyperparameter Values for the ASPP-DR and Dual Attention Modules

| **Parameter** | **Value** |
|---|---|
| **Filters** | 64 |
| **Reduction Ratio** | 4 |
| **Dilation Rates** | 3, 6, 12 |
| **Kernel Sizes** | 1 × 1 and 3 × 3 |
| **SE Reduction** | 16 |
| **Position Attention** | Enabled |
| **Channel Attention** | Enabled |
| **Key channels** | 8 |
| **Output Fusion** | Addition |
| **Final Conv** | 3×3 |

Table 4. Hyperparameter Values for the Feedback Pyramid and Multi-Scale Feature Aggregation Modules

| **Parameter** | **Value** |
|---|---|
| **Filters** | 64 |
| **Stages** | 5 |
| **Upsampling** | Bilinear |
| **Block Type** | SRM (Residual) |
| **Feedback Type** | Additive |
| **Filters** | 64 |
| **Resize Method** | Bilinear |
| **Fusion Type** | Concatenation |
| **Final Conv** | 3×3 |

International Journal of Emerging Research in Engineering, Science, and Management
Vol. 5 | Issue 2 | pp. 96-118 | Apr-Jun 2026.
www.ijeresm.com | eISSN – 2583-4894
Received: 12 Mar 2026 | Revised: 29 May 2026 | Accepted: 10 Jun 2026 | Published: 16 Jun 2026

### 4.1. Dataset Description

This research utilized two datasets, namely LoveDa [32] and the WHU Building Dataset [33]. The LoveDa dataset consists of 4,191 high-resolution remote sensing images with a spatial resolution of 0.3 m, collected from cities such as Nanjing and Wuhan. Since the official LoveDA test set does not provide publicly available ground-truth masks, direct quantitative validation is not feasible. To address this limitation, the original training and validation subsets were combined, randomly shuffled, and re-divided into new training and testing sets. In this study, 80% of the LoveDA images were used for training and the remaining 20% were reserved for testing. The WHU Building Dataset contains a total of 8,188 image samples, of which 6,550 images are used for training and 1,638 images are reserved for testing, corresponding to 80% and 20% of the dataset, respectively.

Table 5. Hyperparameter Values for the Hierarchical Transformer Encoder

| Parameter | Value |
|---|---|
| **Up 1** | 128 |
| **Up 2** | 64 |
| **Up 3** | 32 |
| **Final** | 16 |
| **Output Layer** | num_classes |
| **Kernel Size** | 1×1 |
| **Activation** | Softmax |
| **Channels** | num_classes |
| **Patch size** | $4 \times 4$ |
| **Embedding dimension** | 64 |
| **Number of transformer layers** | 1 Transformer encoder layer |
| **Number of attention heads** | 4 |
| **Key dimension** | 64 |
| **Feed forward network dimension** | 256 |
| **Dropout** | 0.1 |
| **Weight decay** | $(1/10^{-4})$ (L2 Regularization) |
| **Normalization layer** | Layer normalization |
| **Attention type** | MHSA |
| **Activation function** | GELU |
| **Patch Embedding method** | Conv2D with kernel size=4 and stride=4 |
| **Transformer output shape** | Reshaped to $(h, w, 64)$ |

### 4.2. Performance Metrics

This section presents the performance metrics used to evaluate the effectiveness of the proposed model, including Pixel Accuracy (PA), Mean Intersection over Union (mIoU), Precision, Recall, and Dice coefficient.

$$PA = \frac{\sum_{j=1}^{L} p^{i,i}}{\sum_{i=1}^{L}\sum_{j=1}^{L} p^{i,j}} \tag{15}$$

$$mIoU = \frac{1}{L}\sum_{i=1}^{L} \frac{p^{i,i}}{\sum_{j} p^{i,j} + \sum_{j=1}^{L} p^{j,i} - p^{i,i}} \tag{16}$$

$$Precesion = \frac{TP}{TP + FP} \tag{17}$$

$$Recall = \frac{TP}{TP + FN} \tag{18}$$

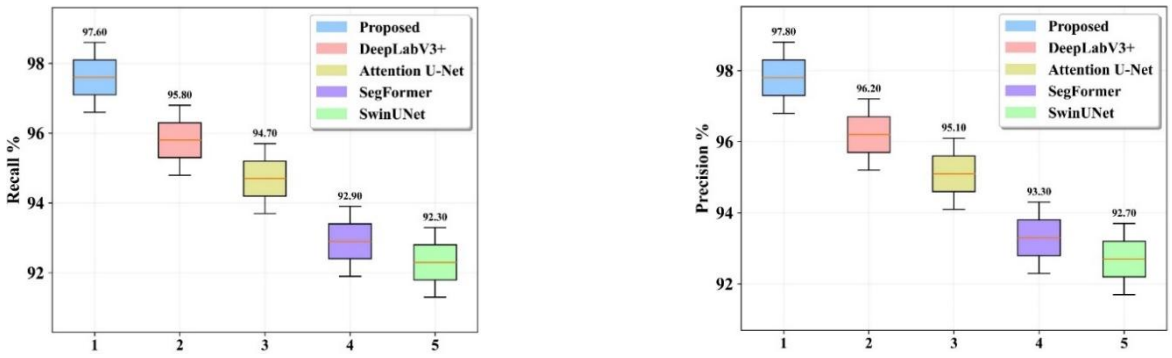
$$Dice = \frac{2 * Pr\,e\,cision * Re\,c\,all}{Pr\,e\,cision + Re\,c\,all} \tag{19}$$

Here, $p^{i,j}$ denotes the number of pixels classified from class $i$ to $j$, while $p^{i,i}$ denotes the number of correctly classified pixels belonging to class $i$. TP, FP, and FN denote True Positives, False Positives, and False Negatives, respectively.

International Journal of Emerging Research in Engineering, Science, and Management
Vol. 5 | Issue 2 | pp. 96-118 | Apr-Jun 2026.
www.ijeresm.com | eISSN – 2583-4894
Received: 12 Mar 2026 | Revised: 29 May 2026 | Accepted: 10 Jun 2026 | Published: 16 Jun 2026

### 4.2.1 Performance Evaluation on the LoveDA Dataset

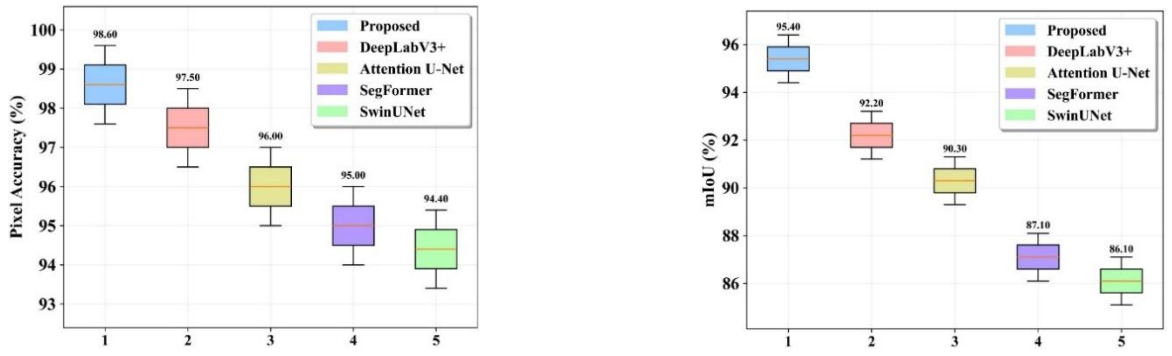
This section presents the performance evaluation results of the proposed semantic segmentation framework on the LoveDA dataset. The proposed framework was compared with four existing models, namely DeepLabV3+ [34], Attention U-Net [35], SegFormer [36], and SwinUNet [37]. These models were selected as benchmark methods for semantic segmentation of remote sensing images. Fig. 10(a) presents the recall comparison among the evaluated models.




(a) (b)

Fig. 10. Recall and Precision Comparison on the LoveDA Dataset

SwinUNet achieved the lowest recall of 92.30%. SegFormer achieved a recall of 92.90%. Attention U-Net achieved a recall of 94.70%. DeepLabV3+ achieved a recall of 95.80%. The proposed framework achieved the highest recall of 97.60%. This improvement can be attributed to the DAM and FPM modules, which enhance feature representation by emphasizing informative spatial and channel features while suppressing irrelevant background information. Fig. 10(b) presents the precision comparison among the evaluated models. The ASPP-DR module enhances multi-scale contextual learning, improving the detection of small and complex objects. SwinUNet achieved the lowest precision of 92.70%. SegFormer achieved a precision of 93.30%. Attention U-Net achieved a precision of 95.10%. DeepLabV3+ achieved a precision of 96.20%. The proposed framework achieved the highest precision of 97.80%. The consistent improvement in precision indicates that the proposed feature enhancement and attention mechanisms effectively suppress false positive predictions while preserving object boundaries.




(a) (b)

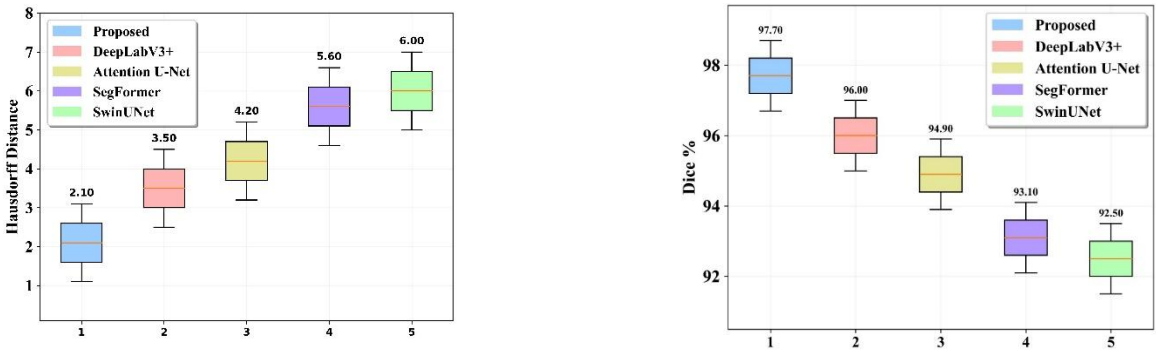
Fig. 11. Pixel Accuracy and mIoU Comparison on the LoveDA Dataset

Fig. 11(a) presents the pixel accuracy comparison among the evaluated models. The SwinUNet has the lowest value of 94.40%. SegFormer achieved 95.00%. The proposed framework achieved the highest pixel accuracy of 98.60%, which can be attributed to the combined effect of the hierarchical transformer encoder, DAM, and FPM in enhancing feature representation, semantic consistency, and boundary refinement. The improvement in pixel accuracy indicates that the hierarchical transformer encoder and feedback-guided refinement enable more reliable pixel-level classification across heterogeneous land-cover regions. Fig. 11(b) presents the mIoU comparison among the evaluated models. SegFormer achieved 87.10%. Attention U-Net achieved 90.30%. The ASPP-DR and Multiscale feature aggregator effectively capture multi-scale contextual information and preserve object details. The proposed framework achieved the highest mIoU of 95.40%. This enables better overlap between the predicted segmentation regions and the ground-truth regions, particularly in complex scene structures. The higher mIoU also suggests improved overlap between predicted and ground-truth regions, particularly around irregular object boundaries and complex scene structures.

Fig. 12(a) presents the Hausdorff distance comparison among the evaluated models. The FPM and Cascaded Upsampling Decoder help refine object boundaries and preserve structural details. As a result, the proposed framework achieved the lowest Hausdorff distance of 2.1%. This reduction indicates more accurate boundary localization and fewer extreme boundary deviations compared with the baseline methods. The Attention U-Net has the value of 4.20%. The deepLabV3+ has the value of 5.60%.

International Journal of Emerging Research in Engineering, Science, and Management
Vol. 5 | Issue 2 | pp. 96-118 | Apr-Jun 2026.
www.ijeresm.com | eISSN – 2583-4894
Received: 12 Mar 2026 | Revised: 29 May 2026 | Accepted: 10 Jun 2026 | Published: 16 Jun 2026

Fig. 12(b) presents the Dice score comparison among the evaluated models. SwinUNet achieved the lowest Dice score of 92.50%. The SegFormer has the next level value of 93.10% and then the Attention U-Net has the value of 94.90%. The deepLabV3+ has the dice score of 96.00%. The higher Dice score demonstrates that the proposed framework achieves better agreement between the predicted segmentation maps and the ground-truth annotations, reflecting improved overall segmentation quality.




(a) (b)

Fig. 12. Hausdorff Distance and Dice Score Comparison on the LoveDA Dataset

Table 6 presents the quantitative performance evaluation results on the LoveDA dataset. Statistical significance was evaluated using an independent two-sample t-test over five repeated experimental runs at a significance level of 0.05.

Table 6. Performance Evaluation on the LoveDA Dataset

| Model | mIoU | Dice | Pixel Accuracy (PA) | Precision | Recall | Hausdorff Distance |
|---|---|---|---|---|---|---|
| **DeepLabV3+ [34]** | 92.20 | 96.00 | 97.50 | 96.20 | 95.80 | 3.5 |
| **Attention U-Net [35]** | 90.30 | 94.90 | 96.00 | 95.10 | 94.70 | 4.2 |
| **SegFormer [36]** | 87.10 | 93.10 | 95.00 | 93.30 | 92.90 | 5.6 |
| **SwinUNet [37]** | 86.10 | 92.50 | 94.40 | 92.70 | 92.30 | 6 |
| **Proposed** | **95.40** | **97.70** | **98.60** | **97.80** | **97.60** | **2.1** |

Tables 7–9 present the ablation study, model complexity analysis, real-time inference performance, and statistical analysis results for the LoveDA dataset. The ablation study demonstrates that each module contributes to the overall segmentation performance. Removing the bilateral filter results in the largest reduction in pixel accuracy and the highest Hausdorff distance, indicating the importance of effective edge-preserving preprocessing. Progressive improvements are observed after incorporating DAM, ASPP-DR, FPM, MSA, and the cascaded decoder. The complete framework achieves the highest pixel accuracy and the lowest Hausdorff distance, confirming that the proposed components complement each other in improving feature representation and boundary refinement.

Table 7. Ablation Study on the LoveDA Dataset

| Model | Pixel Accuracy | Hausdorff Distance |
|---|---|---|
| **Without BF** | 0.951 | 5.8 |
| **Without DAM** | 0.958 | 5.1 |
| **Without ASPP-DR** | 0.962 | 4.6 |
| **Without FPM** | 0.968 | 4 |
| **Without MSA** | 0.973 | 3.5 |
| **Without CD** | 0.978 | 2.9 |
| **Proposed** | **0.982** | **2.1** |

The ablation study results demonstrate that each component contributes positively to segmentation performance. The removal of DAM, ASPP-DR, FPM, MSA, or the cascaded decoder resulted in measurable reductions in pixel accuracy and boundary preservation, confirming the effectiveness of the proposed design choices. The combined integration of these modules enables the framework to simultaneously capture contextual information, preserve object boundaries, and improve multi-scale feature learning.

International Journal of Emerging Research in Engineering, Science, and Management
Vol. 5 | Issue 2 | pp. 96-118 | Apr-Jun 2026.
www.ijeresm.com | eISSN – 2583-4894
Received: 12 Mar 2026 | Revised: 29 May 2026 | Accepted: 10 Jun 2026 | Published: 16 Jun 2026

Table 8. Comparative Evaluation of Model Complexity and Real-Time Inference Performance

| Model | Parameters (M) | FLOPs (G) | GPU Memory (GB) | Training time (s) | Inference Speed (ms) | FPS |
|---|---|---|---|---|---|---|
| **DeepLabV3+ [34]** | 5.6 | 9.8 | 2.8 | 20.5 | 9.8 | 102 |
| **Attention U-Net [35]** | 34.5 | 16.9 | 4.7 | 31.6 | 15.4 | 64.9 |
| **SegFormer [36]** | 13.2 | 11.4 | 3.1 | 22.7 | 10.6 | 94.3 |
| **SwinUNet [37]** | 27.8 | 18.3 | 5.2 | 35.1 | 17.1 | 58.5 |
| **Proposed** | **8.7** | **12.6** | **3.4** | **24.8** | **11.2** | **89.3** |

Although the proposed framework requires slightly higher computational resources than DeepLabV3+, it achieves substantially better segmentation performance while remaining considerably lighter than Attention U-Net and SwinUNet. These results indicate a favorable balance between segmentation accuracy and computational efficiency.

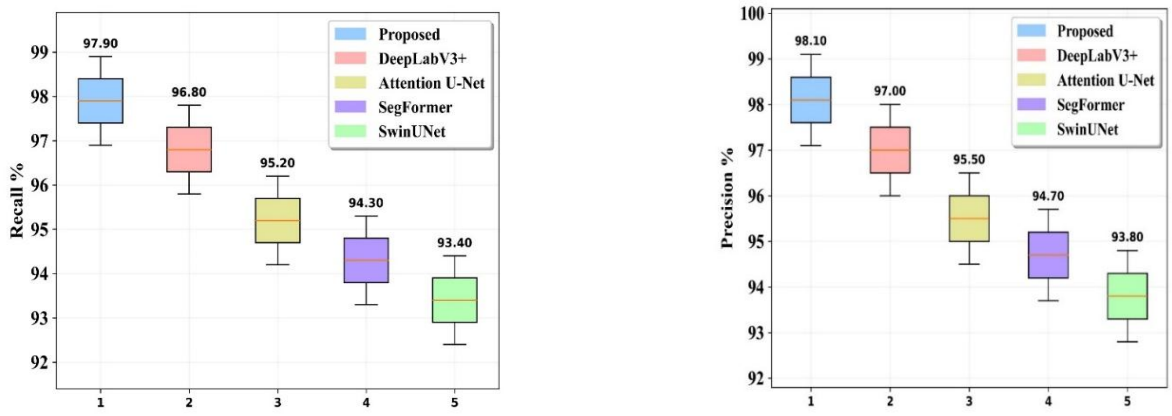
Table 9. Statistical Analysis on the LoveDA Dataset

| Model | Mean Accuracy | Variance | Standard deviation | t-value | p-value |
|---|---|---|---|---|---|
| **DeepLabV3+ [34]** | 0.969 | 0.000009 | 0.003 | 2.1845 | 0.0412 |
| **Attention U-Net [35]** | 0.954 | 0.000025 | 0.005 | 4.9261 | 0.0027 |
| **SegFormer [36]** | 0.945 | 0.000036 | 0.006 | 6.3184 | 0.0008 |
| **SwinUNet [37]** | 0.936 | 0.000049 | 0.007 | 7.5243 | 0.0003 |
| **Proposed** | **0.977** | **0.000001** | **0.001** | **N/A** | **N/A** |

The statistical analysis demonstrates the consistency of the proposed framework across repeated experiments. The lower variance and standard deviation together with statistically significant p-values indicate that the observed performance improvements are reliable rather than resulting from random experimental variation.

### 4.2.2 Performance evaluation on WHU Building Dataset

This section presents the performance evaluation of the proposed framework in comparison with four existing models on the WHU Building Dataset. Fig. 13(a-b) illustrates the recall and precision comparison between the proposed framework and the benchmark models.




(a) (b)

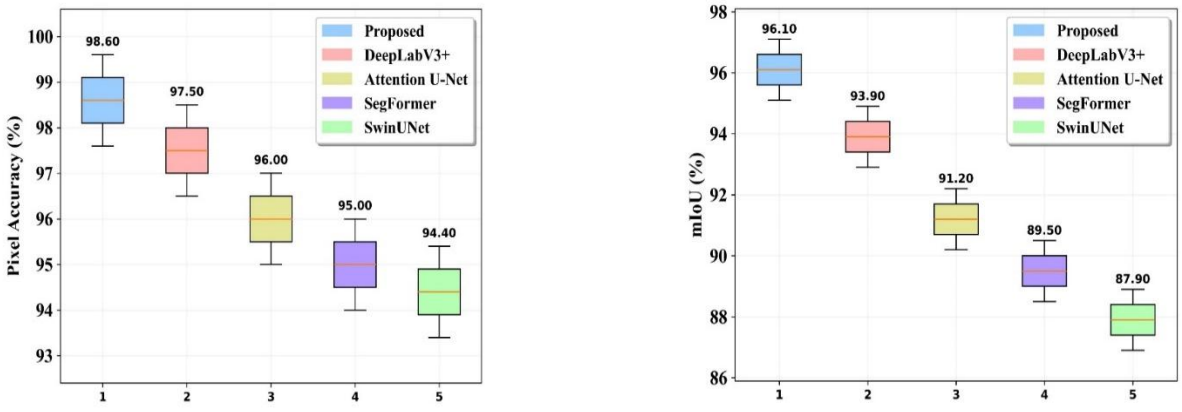
Fig. 13. Recall and Precision Analysis on the WHU Dataset

Fig. 13 presents the recall and precision results of the proposed framework and the benchmark models. SwinUNet achieved a recall of 93.40% and a precision of 93.80%. SegFormer achieved a recall of 94.30% and a precision of 94.70%. Attention U-Net achieved a recall of 95.20% and a precision of 95.50%. DeepLabV3+ achieved a recall of 96.80% and a precision of 97.00%. The proposed framework achieved the highest recall and precision values of 97.90% and 98.10%, respectively. These results indicate that the proposed framework consistently improves object detection performance while maintaining a balanced trade-off between precision and recall.

Fig. 14 presents the pixel accuracy and mIoU results obtained by the evaluated models. SwinUNet achieved a pixel accuracy of 94.40% and an mIoU of 87.90%. SegFormer achieved a pixel accuracy of 95.00% and an mIoU of 89.50%. Attention U-Net achieved a pixel accuracy of 96.00% and an mIoU of 91.20%. DeepLabV3+ achieved a pixel accuracy of 97.50% and an mIoU of 93.90%.

International Journal of Emerging Research in Engineering, Science, and Management
Vol. 5 | Issue 2 | pp. 96-118 | Apr-Jun 2026.
www.ijeresm.com | eISSN – 2583-4894
Received: 12 Mar 2026 | Revised: 29 May 2026 | Accepted: 10 Jun 2026 | Published: 16 Jun 2026

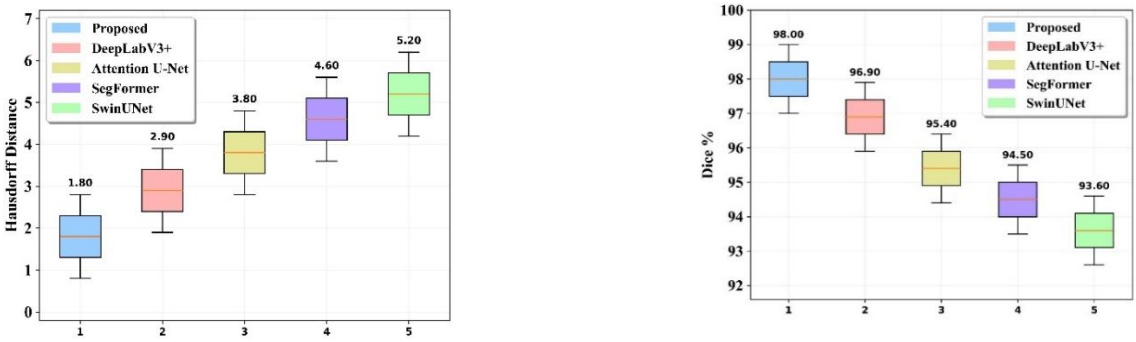
The proposed framework achieved the highest pixel accuracy and mIoU values of 98.60% and 96.10%, respectively. The improvement in both pixel accuracy and mIoU indicates that the proposed framework produces more reliable semantic predictions while preserving spatial consistency across building regions.




(a) (b)

Fig. 14. Pixel Accuracy and mIoU Comparison on the WHU Dataset

Fig. 15(a-b) presents the Hausdorff distance and Dice score comparison among the evaluated models. SwinUNet achieved a Hausdorff distance of 5.20 and a Dice score of 93.60%. SegFormer achieved a Hausdorff distance of 4.60 and a Dice score of 94.50%. Attention U-Net achieved a Hausdorff distance of 3.80 and a Dice score of 95.40%. DeepLabV3+ achieved a Hausdorff distance of 2.90 and a Dice score of 96.90%. The proposed framework achieved the lowest Hausdorff distance of 1.80 and the highest Dice score of 98.00%. The simultaneous reduction in Hausdorff distance and improvement in Dice score demonstrate that the proposed framework accurately preserves building boundaries while maintaining high segmentation overlap. Table 10 presents the quantitative performance evaluation results on the WHU Building Dataset.




(a) (b)

Fig. 15. Hausdorff Distance and Dice Score Comparison on the WHU Dataset

Table 10. Performance Evaluation on the WHU Building Dataset

| Model | mIoU | Dice | Pixel Accuracy (OA) | Precision | Recall | Hausdorff Distance |
|---|---|---|---|---|---|---|
| **DeepLabV3+ [34]** | 93.9 | 96.9 | 97.5 | 97 | 96.8 | 2.9 |
| **Attention U-Net [35]** | 91.2 | 95.4 | 96 | 95.5 | 95.2 | 3.80 |
| **SegFormer [36]** | 89.5 | 94.5 | 95 | 94.7 | 94.3 | 4.60 |
| **SwinUNet [37]** | 87.9 | 93.6 | 94.4 | 93.8 | 93.4 | 5.20 |
| **Proposed** | **96.1** | **98** | **98.6** | **98.1** | **97.9** | **1.80** |

Tables 11–13 present the ablation study, model complexity analysis, real-time inference performance, and statistical analysis results for the WHU Building Dataset. A similar trend is observed on the WHU Building dataset. Progressive improvements are achieved as each proposed component is incorporated into the framework. The complete model attains the highest pixel accuracy and the lowest Hausdorff distance, demonstrating that the proposed modules consistently enhance segmentation performance across different datasets.

International Journal of Emerging Research in Engineering, Science, and Management
Vol. 5 | Issue 2 | pp. 96-118 | Apr-Jun 2026.
www.ijeresm.com | eISSN – 2583-4894
Received: 12 Mar 2026 | Revised: 29 May 2026 | Accepted: 10 Jun 2026 | Published: 16 Jun 2026

Table 11. Ablation Study on the WHU Building Dataset

| Model | Pixel Accuracy | Hausdorff Distance |
|---|---|---|
| **Without BF** | 0.955 | 5 |
| **Without DAM** | 0.962 | 4.4 |
| **Without ASPP-DR** | 0.966 | 3.9 |
| **Without FPM** | 0.971 | 3.3 |
| **Without MSA** | 0.976 | 2.7 |
| **Without CD** | 0.981 | 2.2 |
| **Proposed** | **0.986** | **1.8** |

Table 12. Comparative Evaluation of Model Complexity and Real-Time Inference Performance

| Model | Parameters (M) | FLOPs (G) | GPU Memory (GB) | Training time (s) | Inference Speed (ms) | FPS |
|---|---|---|---|---|---|---|
| **DeepLabV3+ [34]** | 5.6 | 9.8 | 2.9 | 2.5 | 10 | 100 |
| **Attention U-Net [35]** | 34.5 | 16.9 | 4.9 | 4 | 15.8 | 63.3 |
| **SegFormer [36]** | 13.2 | 11.4 | 3.3 | 2.8 | 10.9 | 91.7 |
| **SwinUNet [37]** | 27.8 | 18.3 | 5.4 | 4.4 | 17.6 | 56.8 |
| **Proposed** | **8.7** | **12.6** | **3.6** | **3.1** | **11.5** | **87** |

Although the proposed framework contains more parameters than DeepLabV3+, it provides substantially better segmentation accuracy while maintaining competitive computational complexity and inference speed, indicating a favorable balance between accuracy and efficiency.

Table 13. Statistical Analysis on the WHU Building Dataset

| Model | Mean Accuracy | Variance | Standard deviation | t-value | p-value |
|---|---|---|---|---|---|
| **DeepLabV3+ [34]** | 0.969 | 0.000005 | 0.004 | 2.8647 | 0.0214 |
| **Attention U-Net [35]** | 0.954 | 0.000031 | 0.005 | 5.4182 | 0.0016 |
| **SegFormer [36]** | 0.945 | 0.00004 | 0.008 | 6.8953 | 0.0005 |
| **SwinUNet [37]** | 0.936 | 0.000049 | 0.009 | 8.1476 | 0.0002 |
| **Proposed** | **0.98** | **0.000001** | **0.002** | **N/A** | **N/A** |

The statistical analysis further supports the reliability of the proposed framework. The lowest variance and standard deviation indicate stable performance across repeated experiments, while the reported p-values demonstrate statistically significant improvements over the compared methods. Output samples and masked images of segmentation model of both LoveDa Dataset and WHU dataset is illustrated in Fig. 16 and 17 respectively.

### 4.3. Discussion

The ablation study results demonstrate that each component contributes to the overall segmentation performance. In particular, the removal of DAM, ASPP-DR, FPM, or multi-scale aggregation results in noticeable reductions in pixel accuracy and boundary preservation, confirming the effectiveness of the proposed design choices. Most existing semantic segmentation approaches focus on learning richer semantic representations by capturing long-range dependencies and global contextual information, which facilitate object identification and discrimination. However, these approaches often overlook important prior cues such as edge details and spatial information. This limitation can result in the loss of fine structural information, negatively affecting object localization and leading to unclear boundaries and incomplete segmentation results.

Wang et al. [21] proposed the DSPP model, which achieved an mIoU of 70.8%. However, preserving fine boundary information remained a challenge. The proposed framework addresses this limitation through the integration of the DAM and FPM, which enhance edge preservation and spatial feature representation. Similarly, Wang et al. [22] developed UNetFormer, which achieved mIoU values of 86.12% and 83.10% on the Potsdam and Vaihingen datasets, respectively. However, maintaining global semantic consistency across complex scenes remains challenging. The proposed framework addresses this limitation through the hierarchical transformer encoder and dual attention mechanism, which effectively capture long-range contextual dependencies. Wang et al. [23] introduced MVTrans, which demonstrated improved semantic segmentation performance compared with several benchmark methods. However, the architecture is relatively complex. In contrast, the proposed framework employs lightweight modules such as ASPP-DR and efficient attention mechanisms to achieve a balance between segmentation accuracy and computational efficiency.

International Journal of Emerging Research in Engineering, Science, and Management
Vol. 5 | Issue 2 | pp. 96-118 | Apr-Jun 2026.
www.ijeresm.com | eISSN – 2583-4894
Received: 12 Mar 2026 | Revised: 29 May 2026 | Accepted: 10 Jun 2026 | Published: 16 Jun 2026

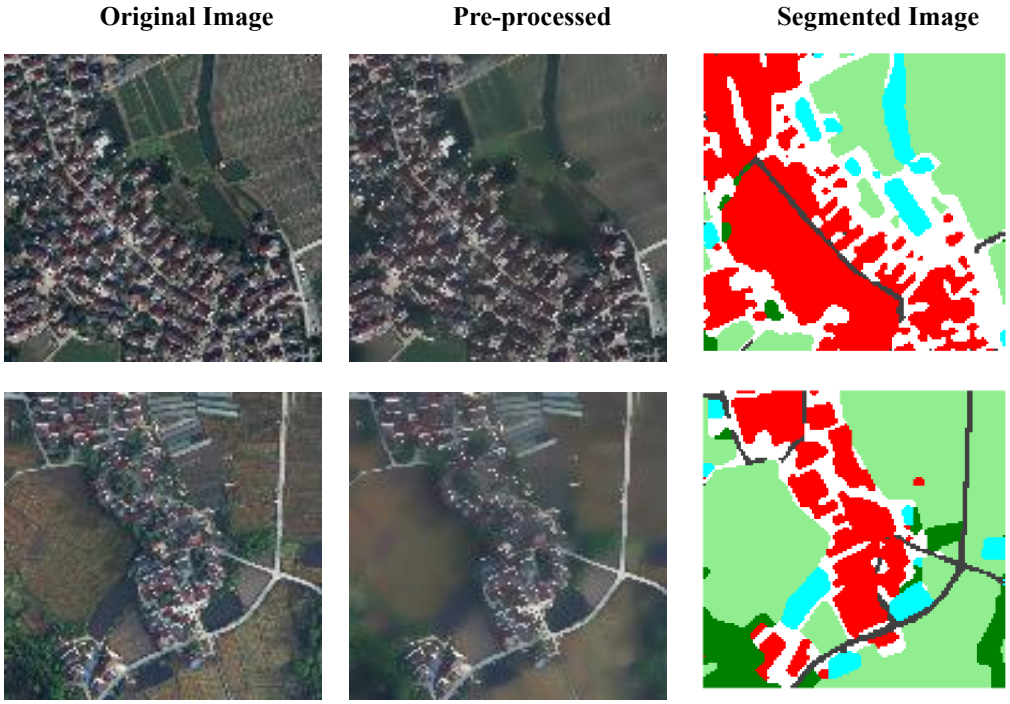



Fig. 16. Qualitative Segmentation Results on the LoveDA Dataset


Fig. 17. Qualitative Segmentation Results on the WHU Building Dataset

Zhu et al. [24] developed GLFFNet, which achieved an mIoU of 84.01%. However, the increased model complexity may limit its practical deployment. The proposed framework mitigates this issue by incorporating lightweight modules and efficient attention mechanisms that optimize feature learning while reducing computational overhead. Li et al. [25] introduced CSNet, which achieved mIoU scores of 81.4% and 70.3% on the evaluated datasets. The proposed framework addresses these limitations through multi-scale feature extraction and dual attention mechanisms, enabling effective handling of large-scale variations and complex scene structures. Table 14 presents a comparative analysis between the proposed framework and existing methods.

International Journal of Emerging Research in Engineering, Science, and Management
Vol. 5 | Issue 2 | pp. 96-118 | Apr-Jun 2026.
www.ijeresm.com | eISSN – 2583-4894
Received: 12 Mar 2026 | Revised: 29 May 2026 | Accepted: 10 Jun 2026 | Published: 16 Jun 2026

Table 14. Comparative Performance Analysis of the Proposed Framework and Existing Methods

| Method | LoveDA mIoU (%) | WHU mIoU (%) |
|---|---|---|
| **DSPP [21]** | 70.8 | Not Reported |
| **UNetFormer [22]** | 83.10 | Not Reported |
| **MVTrans [23]** | 85.75 | Not Reported |
| **GLFFNet [24]** | 84.01 | Not Reported |
| **CSNet [25]** | 70.3 | Not Reported |
| **Proposed** | **95.4** | **96.1** |

## 5 CONCLUSIONS

The proposed methodology introduces a parallel-branch, feedback-guided transformer framework for accurate semantic segmentation of remote sensing images. Input images are pre-processed using bilateral filtering to remove noise while preserving important edge information. A hierarchical transformer encoder extracts multi-scale feature representations by dividing images into patches. The dual attention mechanism enhances feature learning by capturing both spatial and channel-wise dependencies. The ASPP-DR module further enriches features by integrating multi-scale contextual information with efficient dense connections. A multi-stage decoder progressively reconstructs spatial resolution while preserving semantic consistency. The feedback pyramid module iteratively refines features using previous outputs to improve segmentation quality. Finally, multi-scale aggregation and cascaded upsampling generate precise, high-resolution pixel-wise segmentation maps. The proposed framework achieved pixel accuracies of 98.2% and 98.6% on the LoveDA and WHU datasets, respectively. The experimental results, ablation studies, and statistical analyses demonstrate the effectiveness and robustness of the proposed framework across different remote sensing datasets. The experimental results, ablation studies, and statistical analyses demonstrate the effectiveness and robustness of the proposed framework across different remote sensing datasets. Although the proposed framework achieves superior segmentation performance, the feedback-guided architecture introduces additional computational overhead compared with lightweight segmentation models. Future work will investigate model compression and cross-dataset adaptation strategies to improve deployment efficiency and generalization capability. Further improvements may include integrating lightweight architectures and extending the model to handle cross-domain generalization and multimodal data.

## FUNDING INFORMATION

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

## ETHICS STATEMENT

This study did not involve human or animal subjects and, therefore, did not require ethical approval.

## STATEMENT OF CONFLICT OF INTERESTS

The authors declare no conflicts of interest related to this study.



## REFERENCES

[1] J. Li, Y. Cai, Q. Li, M. Kou, and T. Zhang, “A review of remote sensing image segmentation by deep learning methods,” *International Journal of Digital Earth*, vol. 17, no. 1, Mar. 2024, doi: 10.1080/17538947.2024.2328827.

[2] G. Vivone et al., "Deep Learning in Remote Sensing Image Fusion: Methods, protocols, data, and future perspectives," in *IEEE Geoscience and Remote Sensing Magazine*, vol. 13, no. 1, pp. 269-310, March 2025, doi: 10.1109/MGRS.2024.3495516.

[3] R. Liu, T. Luo, S. Huang, Y. Wu, Z. Jiang and H. Zhang, "CrossMatch: Cross-View Matching for Semi-Supervised Remote Sensing Image Segmentation," in *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1-15, 2024, Art no. 5650515, doi: 10.1109/TGRS.2024.3507050.

[4] X. Ma, X. Zhang, X. Ding, M. -O. Pun and S. Ma, "Decomposition-Based Unsupervised Domain Adaptation for Remote Sensing Image Semantic Segmentation," in *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1-18, 2024, Art no. 5645118, doi: 10.1109/TGRS.2024.3483283.

[5] X. He et al., "Hierarchical Relation Learning for Few-Shot Semantic Segmentation in Remote Sensing Images," in *IEEE Transactions on Geoscience and Remote Sensing*, vol. 63, pp. 1-15, 2025, Art no. 4410615, doi: 10.1109/TGRS.2025.3571738.

[6] H. Xu, C. Zhang, P. Yue, and K. Wang, “SDCluster: A clustering based self-supervised pre-training method for semantic segmentation of remote sensing images,” *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 223, pp. 1–14, Mar. 2025, doi: 10.1016/j.isprsjprs.2025.02.021.

International Journal of Emerging Research in Engineering, Science, and Management
Vol. 5 | Issue 2 | pp. 96-118 | Apr-Jun 2026.
www.ijeresm.com | eISSN – 2583-4894
Received: 12 Mar 2026 | Revised: 29 May 2026 | Accepted: 10 Jun 2026 | Published: 16 Jun 2026

[7] W. Wang et al., "Multi-dimension Transformer with Attention-based Filtering for Medical Image Segmentation," *2024 IEEE 36th International Conference on Tools with Artificial Intelligence (ICTAI)*, Herndon, VA, USA, 2024, pp. 632-639, doi: 10.1109/ICTAI62512.2024.00095.

[8] K. Chen, J. Zhang, C. Liu, Z. Zou and Z. Shi, "RSRefSeg: Referring Remote Sensing Image Segmentation with Foundation Models," *IGARSS 2025 - 2025 IEEE International Geoscience and Remote Sensing Symposium*, Brisbane, Australia, 2025, pp. 1070-1074, doi: 10.1109/IGARSS55030.2025.11243338.

[9] T. Wang et al., "LMFNet: Lightweight Multimodal Fusion Network for high-resolution remote sensing image segmentation," *Pattern Recognition*, vol. 164, p. 111579, Mar. 2025, doi: 10.1016/j.patcog.2025.111579.

[10] S. Gui, S. Song, R. Qin, and Y. Tang, "Remote Sensing Object Detection in the Deep Learning Era—A Review," *Remote Sensing*, vol. 16, no. 2, p. 327, Jan. 2024, doi: 10.3390/rs16020327.

[11] G. Vivone et al., "Deep Learning in Remote Sensing Image Fusion: Methods, protocols, data, and future perspectives," in *IEEE Geoscience and Remote Sensing Magazine*, vol. 13, no. 1, pp. 269-310, March 2025, doi: 10.1109/MGRS.2024.3495516.

[12] Y. Chen, Z. Yang, L. Zhang, and W. Cai, "A semi-supervised boundary segmentation network for remote sensing images," *Scientific Reports*, vol. 15, no. 1, p. 2007, Jan. 2025, doi: 10.1038/s41598-025-85125-9.

[13] L. Yang, H. Chen, A. Yang and J. Li, "EasySeg: An Error-Aware Domain Adaptation Framework for Remote Sensing Imagery Semantic Segmentation via Interactive Learning and Active Learning," in *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1-18, 2024, Art no. 4407518, doi: 10.1109/TGRS.2024.3399260.

[14] K. An, Y. Wang and L. Chen, "Encouraging the Mutual Interact Between Dataset-Level and Image-Level Context for Semantic Segmentation of Remote Sensing Image," in *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1-16, 2024, Art no. 5606116, doi: 10.1109/TGRS.2024.3352582.

[15] Z. Marinov, P. F. Jäger, J. Egger, J. Kleesiek and R. Stiefelhagen, "Deep Interactive Segmentation of Medical Images: A Systematic Review and Taxonomy," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 12, pp. 10998-11018, Dec. 2024, doi: 10.1109/TPAMI.2024.3452629.

[16] M. Huang, J. Zou, Y. Zhang, U. A. Bhatti and J. Chen, "Efficient Click-Based Interactive Segmentation for Medical Image With Improved Plain-ViT," in *IEEE Journal of Biomedical and Health Informatics*, vol. 29, no. 12, pp. 8904-8916, Dec. 2025, doi: 10.1109/JBHI.2024.3392893.

[17] Y. Du, Fan Bai, Tiejun Huang, Bo Zhao, "Segvol: Universal and interactive volumetric medical image segmentation." *38th Conference on Neural Information Processing Systems*, pp. 110746-110783, 2024, doi: 10.52202/079017-3516.

[18] J. Liu, H. Liu, X. Li, J. Ren and X. Xu, "MiLNet: Multiplex Interactive Learning Network for RGB-T Semantic Segmentation," in *IEEE Transactions on Image Processing*, vol. 34, pp. 1686-1699, 2025, doi: 10.1109/TIP.2025.3544484.

[19] J. Lin et al., "AdaptiveClick: Click-Aware Transformer With Adaptive Focal Loss for Interactive Image Segmentation," in *IEEE Transactions on Neural Networks and Learning Systems*, vol. 36, no. 3, pp. 5759-5773, March 2025, doi: 10.1109/TNNLS.2024.3378295.

[20] Z. Xing, G. Ma, L. Wang, L. Yang, X. Guo and S. Chen, "Toward Visual Interaction: Hand Segmentation by Combining 3-D Graph Deep Learning and Laser Point Cloud for Intelligent Rehabilitation," in *IEEE Internet of Things Journal*, vol. 12, no. 12, pp. 21328-21338, 15 June15, 2025, doi: 10.1109/JIOT.2025.3546874.

[21] L. Wang, D. Li, S. Dong, X. Meng, X. Zhang, and D. Hong, "PyramidMamba: Rethinking pyramid feature fusion with selective space state model for semantic segmentation of remote sensing imagery," *International Journal of Applied Earth Observation and Geoinformation*, vol. 144, p. 104884, Oct. 2025, doi: 10.1016/j.jag.2025.104884.

[22] J. Wang et al., "A multi-scale remote sensing semantic segmentation model with boundary enhancement based on UNetFormer," *Scientific Reports*, vol. 15, no. 1, p. 14737, Apr. 2025, doi: 10.1038/s41598-025-99663-9.

[23] Z. Wang, N. Xu, Z. You, and S. Zhang, "DiffMamba: semantic diffusion guided feature modeling network for semantic segmentation of remote sensing images," *GIScience & Remote Sensing*, vol. 62, no. 1, Apr. 2025, doi: 10.1080/15481603.2025.2484829.

[24] S. Zhu, L. Zhao, Q. Xiao, J. Ding, and X. Li, "GLFFNET: Global–Local Feature Fusion Network for High-Resolution Remote Sensing Image Semantic Segmentation," *Remote Sensing*, vol. 17, no. 6, p. 1019, Mar. 2025, doi: 10.3390/rs17061019.

[25] J. Li, H. Zhang, L. Chen, B. He, and H. Chen, "CSNET: a remote sensing image semantic segmentation network based on coordinate attention and skip connections," *Remote Sensing*, vol. 17, no. 12, p. 2048, Jun. 2025, doi: 10.3390/rs17122048.

[26] L. Wu, L. Fang, J. Yue, B. Zhang, P. Ghamisi and M. He, "Deep Bilateral Filtering Network for Point-Supervised Semantic Segmentation in Remote Sensing Images," in *IEEE Transactions on Image Processing*, vol. 31, pp. 7419-7434, 2022, doi: 10.1109/TIP.2022.3222904.

[27] H. Zeng, S. Peng, and D. Li, "Deeplabv3+ semantic segmentation model based on feature cross attention mechanism," *Journal of Physics Conference Series*, vol. 1678, no. 1, p. 012106, Nov. 2020, doi: 10.1088/1742-6596/1678/1/012106.

International Journal of Emerging Research in Engineering, Science, and Management
Vol. 5 | Issue 2 | pp. 96-118 | Apr-Jun 2026.
www.ijeresm.com | eISSN – 2583-4894
Received: 12 Mar 2026 | Revised: 29 May 2026 | Accepted: 10 Jun 2026 | Published: 16 Jun 2026

[28] Y. Li, J. Gao, Y. Du, Y. Xiao, Z. Gao and H. Huang, "HiTrans-SAM: Hierarchical Transformer Encoder and SAM-Augmented Inputs for Multi-Scale Remote Sensing Image Segmentation," in *IEEE Access*, vol. 13, pp. 177957-177969, 2025, doi: 10.1109/ACCESS.2025.3617388.

[29] C. Gong, J. Liu, M. Gong, J. Li, U. A. Bhatti, and J. Ma, "Robust medical zero-watermarking algorithm based on Residual-DenseNet," *IET Biometrics*, vol. 11, no. 6, pp. 547–556, Sep. 2022, doi: 10.1049/bme2.12100.

[30] H. Wu, J. Gui, J. Zhang, J. T. Kwok and Z. Wei, "Feedback Pyramid Attention Networks for Single Image Super-Resolution," in *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 9, pp. 4881-4892, Sept. 2023, doi: 10.1109/TCSVT.2023.3250657.

[31] Y. Zhao, Y. Jiang, L. Huang, and K. Xia, "SEF-UNet: advancing abdominal multi-organ segmentation with SEFormer and depthwise cascaded upsampling," *PeerJ Computer Science*, vol. 10, p. e2238, Aug. 2024, doi: 10.7717/peerj-cs.2238.

[32] J. Wang, Z. Zheng, A. Ma, X. Lu, and Y. Zhong, "LoveDA: A Remote Sensing Land-Cover Dataset for Domain Adaptive Semantic Segmentation," in *Proc. NeurIPS Datasets Benchmarks Track*, 2021, doi: 10.5281/zenodo.5706578.

[33] S. Ji, S. Wei and M. Lu, "Fully Convolutional Networks for Multisource Building Extraction From an Open Aerial and Satellite Imagery Data Set," in *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 1, pp. 574-586, Jan. 2019, doi: 10.1109/TGRS.2018.2858817.

[34] Y. Wang, L. Yang, X. Liu, and P. Yan, "An improved semantic segmentation algorithm for high-resolution remote sensing images based on DeepLabv3+," *Scientific Reports*, vol. 14, no. 1, p. 9716, Apr. 2024, doi: 10.1038/s41598-024-60375-1.

[35] N. S. Jonnala et al., "DSIA U-Net: deep shallow interaction with attention mechanism UNet for remote sensing satellite images," *Scientific Reports*, vol. 15, no. 1, p. 549, Jan. 2025, doi: 10.1038/s41598-024-84134-4.

[36] S. Peng, H. Xie, N. Liu, and Y. Zeng, "Semantic Segmentation of Multispectral Remote Sensing Imagery for Coastal Wetlands with SegFormer," *Remote Sensing*, vol. 18, no. 5, p. 745, Feb. 2026, doi: 10.3390/rs18050745.

[37] Z. Chang, M. Xu, Y. Wei, and J. Lian, "CW-SwinUNet: a novel semantic segmentation approach for very-high-resolution remote sensing imagery," *International Journal of Remote Sensing*, vol. 46, no. 22, pp. 8614–8639, Oct. 2025, doi: 10.1080/01431161.2025.2571233.