Computing and Informatics, Vol. 45, 2026, 550–579, doi: 10.31577/cai_2026_3_550

ENHANCED CHANGE DETECTION IN REMOTE SENSING USING NAVIT WITH UNET

Satish Pawar*

Department of Computer Science and Engineering
Samrat Ashok Technological Institute Vidisha
Civil line, Vidisha, Madhyapradesh, 464001, India
e-mail: Satishpawar.cse@satiengg.in

Nishchol Mishra

School of Information Technology, RGPV
Bhopal, Madhya Pradesh, 462033, India
e-mail: nishchol@rgpv.ac.in

Neelesh Mehra

Department of Electronics Engineering
Samrat Ashok Technological Institute Vidisha
Civil line, Vidisha, Madhyapradesh, 464001, India
e-mail: neeleshmehra.ec@satiengg.in

Abstract. Change detection (CD) in remote sensing (RS) images has become an essential tool for land resource planning, disaster monitoring and urban growth analysis. Despite significant progress with deep learning (DL) methods, traditional approaches struggle to effectively extract multi-scale features and balance local and global information while maintaining computational efficiency. This research introduces NAViT-UNet a novel framework leveraging Neighbourhood Attention Vision Transformers (NAViT) and Multi-Temporal Fusion (MTF) Networks for accurate and efficient CD in multi-temporal RS imagery. The proposed model integrates

* Corresponding author

NAViT modules for local feature extraction, reducing computational overhead while enhancing performance. A robust Multi-Temporal (MT) Fusion Network dynamically aggregates temporal feature maps, optimizing the utilization of complementary MT information. The bottleneck layer incorporates Cascaded Atrous Spatial Pyramid Pooling (ASPP) to broaden the feature receptive field, enabling the precise segmentation of small and complex regions. To address imbalanced segmentation challenges, a Mixed Loss Function combining MS-SSIM Loss, Tversky Loss and Focal Loss ensures fine-grained boundary delineation and segmentation accuracy. The model also adopts the Swish activation function which enhances performance over traditional ReLU. Experiments demonstrated the proposed system achieved best performance in terms of 99.8 % of accuracy, 98.81 % of precision, 99.32 % of Recall and 99.07 % of F1 score. This framework establishes a new benchmark for multi-temporal CD, significantly improving the identification and segmentation of change regions with minimal computational costs. The proposed methodology offers a transformative solution for RS applications, unlocking new possibilities in land monitoring, disaster management and urban planning.

Keywords: Remote sensing, change detection, neighborhood attention vision transformer (NAViT), multi-temporal fusion network, temporal feature aggregation, land resource planning, mixed loss function

Mathematics Subject Classification 2010: 68T05

1 INTRODUCTION

Using RS imagery for CD has become a crucial research field in recent years because of its many uses in urban development, disaster response, and land resource management [1]. CD detects and tracks changes in the Earth's surface by examining MT photographs, offering vital information on human activity, environmental dynamics, and natural disasters [2]. As high-resolution satellite photos become more widely available and processing power increases, more advanced CD techniques are becoming possible [3]. While tackling the inherent difficulties brought on by the complexity of RS data, these techniques seek to increase accuracy and efficiency [4].

The process of CD entails comparing images taken at various times to find areas of change. RS images frequently create issues due to their different spatial resolutions, illumination fluctuations, and the intricacy of land-cover elements [5]. Furthermore, small-scale regions, fuzzy boundaries, and overlapping features in the images can result in imprecise segmentation, reducing overall detection accuracy [6]. Addressing these difficulties requires models that can extract and analyze fine-grained characteristics while remaining computationally efficient [7].

In recent years, DL approaches have transformed RS by providing powerful tools for feature extraction, image segmentation, and pattern detection. These ap-

proaches, particularly those based on convolutional neural networks (CNNs), have shown exceptional success in gathering spatial and temporal data [8]. However, the high-dimensional structure of RS data and the need to interpret MT information present additional challenges, particularly in terms of computational costs and resource utilization [9].

Traditional CD approaches frequently rely on handcrafted features and predetermined criteria, limiting their ability to generalize across multiple datasets. These methods are generally insufficient to manage the diversity and complexity of RS imagery, such as varied landscapes or rapidly changing urban areas [10]. While recent advances in machine learning (ML) have enhanced the automation of feature extraction, many existing models struggle to successfully integrate MT data or exploit all spatial and contextual information [11].

Another significant problem in RS CD involves achieving a balance between global and local feature extraction. Models must capture tiny local characteristics while preserving a broader contextual understanding to ensure accurate segmentation [12]. Simultaneously, computing performance must be optimized to allow for the processing of high-resolution images and large-scale datasets, both of which are becoming more common in RS applications. Efficient feature fusion and multi-scale (MS) analysis are required to obtain useful insights from RS imaging [13].

This paper explores improved approaches for overcoming the obstacles inherent in MT RS image analysis. This work attempts to enhance the efficiency and usability of CD algorithms by emphasizing rapid feature extraction, effective temporal data fusion, and robust segmentation strategies. The findings have ramifications for a wide range of applications, including urban planning, emergency management, and environmental monitoring, all of which require precise and timely information to make informed decisions.

This study's primary insight is as follows:

- Introduced a dynamic fusion mechanism that effectively aggregates and optimizes complementary information from MT RS images, ensuring accurate and efficient CD.
- Incorporated a transformer based NAViT module to extract local and global features with reduced computational overhead, enhancing feature representation for complex RS imagery.
- Designed a bottleneck layer using cascaded ASPP to expand the receptive field, enabling precise segmentation of small and complex regions in RS images.
- Introduced a multi-headed neighborhood attention (MHNA) mechanisms to dynamically learn weighted mappings from MT features enhancing temporal information integration.
- Developed a hybrid loss function combining MS-SSIM, Tversky and focal losses addressing imbalanced segmentation and ensuring accurate boundary delineation.

- Optimised the framework for reduced resource consumption while maintaining high detection accuracy, making it suitable for largescale and high-resolution RS dataset.

Section 2 outlines the relevant literature, Section 3 introduces the NAViT approach using the proposed framework from UNet, and Section 4 discloses the system evaluation and talks about the outcome. The conclusion and future directions of predictive work in the CD domain are finally covered in Section 5.

2 LITERATURE REVIEW

The goal of CD is to locate locations that have changed in the remotely sensed images obtained by MT satellite imaging, which can identify human or natural spatial changes. It is crucial for tracking the environment and identifying shifts in land cover and use. CD has been implemented using a variety of parameters between two identical network topologies that are contained inside it. The customized network module determines how similar the two inputs are, and the double-branch structure can take in a variety of inputs. Differentiating between changed and unaltered pixels in multi-temporal images is the aim of CD tasks. The two branches of the Siamese network are used to extract the information from multi-temporal images. Using the similarity between the two properties, the network then generates the change maps. The Siamese network is thus a leading alternative for applications of change detection. Using RGB and multi-spectral images, Daudt et al. developed two Siamese extensions of fully convolutional networks, which produced the best results on open change detection datasets. Significant improvements in data processing for large-scale Earth observation systems were demonstrated by their system, which trained from scratch using annotated photos, surpassing earlier techniques and operating 500 times quicker than similar systems. Based on deep networks and slow feature analysis theory Deep Slow Feature Analysis for CD in MT RS images was created by Du et al. [14]. The model projects bi-temporal vision using two symmetric deep networks, with the Slow Feature Analysis module emphasizing altered elements and suppressing unmodified ones. Chi-square distance is used to determine the change intensity, and Change Vector Analysis pre-detection finds pixels that have not changed. To increase the precision of encoded bitemporal visuals, Chen and Shi [15] established a novel Siamese-based spatial-temporal attention neural network. For more discriminative features, they integrated attention weights across time and location using a CD self-attention (SA) mechanism to describe spatial-temporal interactions. More reliable representations of a variety of items were made possible by the network's division into MS subregions by using LEVIR-CD. Chen et al. [16] addressed the limits of resisting pseudo changes by introducing dual attentive fully convolutional Siamese networks for CD in high-resolution images. By capturing long-range dependencies, the dual attention mechanism enhances recognition performance and feature representations. By paying attention to feature pairs that remained unchanged and increasing attention to feature pairs that changed,

they also addressed imbalanced samples and reduced pseudo-changes. SUNNet-CD, a closely linked Siamese network for CD tasks, was presented by Fang et al. [17]. It has deep supervision of network training through the aggregation and refinement of semantic information at multiple levels. An edge-guided recurrent CNN, which combines discriminative information with edge structure prior in a single framework, was presented by Bai et al. [18] in order to enhance CD and generate more accurate building boundaries. It is a Siamese convolutional neural network that simultaneously extracts the main multilayer features from multi-temporal images. A dual-branch network was presented by Ma et al. [19] for CD. To extract the depth-variant semantic features of MT picture pairs and the corresponding features of each image, the network is split into two branches. Additionally, a high-frequency attention block was included in the study by Zheng et al. [20], High-Frequency Attention-Guided Siamese Network. The HFAB uses high-frequency augmentation and spatial-wise attention to improve the buildings' high-frequency information. The HFA-Net enhances building CD performance by efficiently identifying changes in building edges. Three public datasets – WHU-CD, LEVIR-CD, and the Google dataset – were used to test the approach. Yang et al. [21] combined DL and dictionary learning approaches to create a CD method for RS photos. The technique ensured that reconstruction coefficients were transferable between blocks by achieving spatial-temporal modelling and correlation of multi-temporal images. A new loss function for discriminative feature extraction was presented, along with the development of a differential feature discriminant network. To enhance ultrahigh-resolution RS image CD tasks, Xu et al. [22] created the Feature-Enhanced Residual Attention Network. With a feature-enhanced skip connection module serving as the decoder and a Siamese network as the encoder, the network has a U-shaped encoder–decoder architecture. Using the WHU-CD and LEVIR-CD datasets, the network creates anticipated building change maps based on building change and building edge information.

As a result of the irregular boundaries, the MT RSI images are highly hazy, according to the above mentioned existing methods. An MT also struggles to provide the necessary information because of the complicated nature of the RSI. Additionally, feature extraction is a weakness of existing segmentation techniques. The feature resolution is also decreased by repeated pooling and striding procedures, which results in the loss of detailed information. In addition, the size of changes varies among regions. The receptive fields of convolutions may therefore be exceeded by large alterations, while slight changes may be overlooked. To overcome the above limitations, need to create a new model to increase the accuracy.

3 PROPOSED METHODOLOGY

The rapid development of RS technology has led to the widespread application of CD techniques based on RS images in urban expansion, catastrophe monitoring, and land resource planning. However, conventional building CD techniques are in-

capable of extracting MS building features effectively or using feature maps' local and global information at reduced computational cost. This affects the accuracy of the detection and limits the model's potential uses. To solve these difficulties, a NAViT with UNet method is proposed for MT RS images (Figure 1). The NAViT-UNet framework is a cutting-edge architecture developed to handle the issues of detecting change in MT RS imagery. It combines advanced feature extraction, temporal fusion, and segmentation approaches to provide precise and quick analysis of complex RS imagery. The framework is based on an encoder–decoder design, with the encoder developed using NAViT blocks. These blocks excel in capturing local and global properties while remaining computationally efficient using MHNA, group normalization (GN), and skip connections. To address the temporal component of RS images, the system includes an MTF Network, which dynamically merges information from multiple temporal images using weighted mapping techniques.

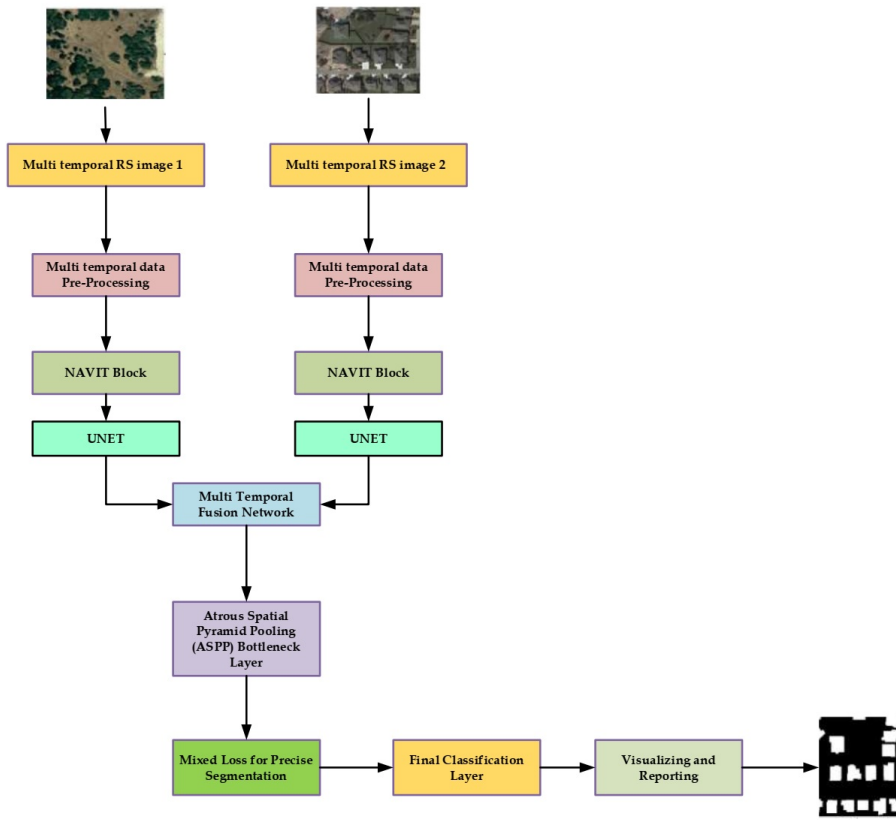

Figure 1. Workflow of proposed work

A crucial innovation is the addition of a bottleneck layer using Cascaded ASPP, which broadens the receptive field and allows for MS feature extraction, resulting in precise segmentation of fine and complicated structures. The decoder uses upsampling layers and skip connections to recreate high-resolution feature maps for segmentation. To improve performance even more, the system uses a mixed loss that combines MS-SSIM, Tversky, and focal loss to provide exact border identification and handling of imbalanced segmentation. NAViT-UNet offers a reliable and effective solution for MT RS image analysis by skillfully integrating these elements, which qualifies it for a variety of uses such as disaster response, urban development study, and land monitoring.

3.1 Preprocessing

This step is critical to maintaining the quality and consistency of input data for accurate CD in RS images. This study begins with data normalization, which standardizes the pixel intensity values of the RS images, limiting the influence of fluctuations caused by differences in illumination or sensor properties. To improve feature contrast and highlight regions of interest, contrast enhancement techniques are employed. These techniques enhance the ability to discern between areas that have changed and those that have not by altering the intensity distribution of the image. Denoising filters, which assist maintain fine details while eliminating artifacts that may negatively affect model performance, are also used to reduce the noise in RS images.

The high-resolution nature of RS images necessitates patch-wise division to partition them into manageable patches without losing spatial information, enhancing computational efficiency and allowing the model to concentrate on localized details, while temporal alignment ensures images from different time periods align. Finally, to improve model generalization and diversify training data, data augmentation techniques including rotation, flipping, and scaling are used. By carefully applying these processing steps, the input data is prepared to facilitate effective feature extraction, MTF and precise segmentation in the subsequent stages of NAViT-UNet framework.

3.2 Neighborhood Attention Vision Transformer

The proposed NAViT-UNet system relies heavily on the Neighborhood Attention Vision Transformer (NAViT), which is intended to minimize computational overhead and effectively extract useful features from RS images. Building on the idea of attention mechanisms, NAViT improves the extraction of local and global characteristics by concentrating on the neighborhood context inside feature maps. In contrast to conventional transformers that use full attention, which can be computationally costly, NAViT uses MHNA, which limits the computation of attention to a specific area [23]. The capacity to describe contextual information and local relationships

is maintained while computing performance is increased by this mechanism. The structure of NAViT Block is illustrated in Figure 2.

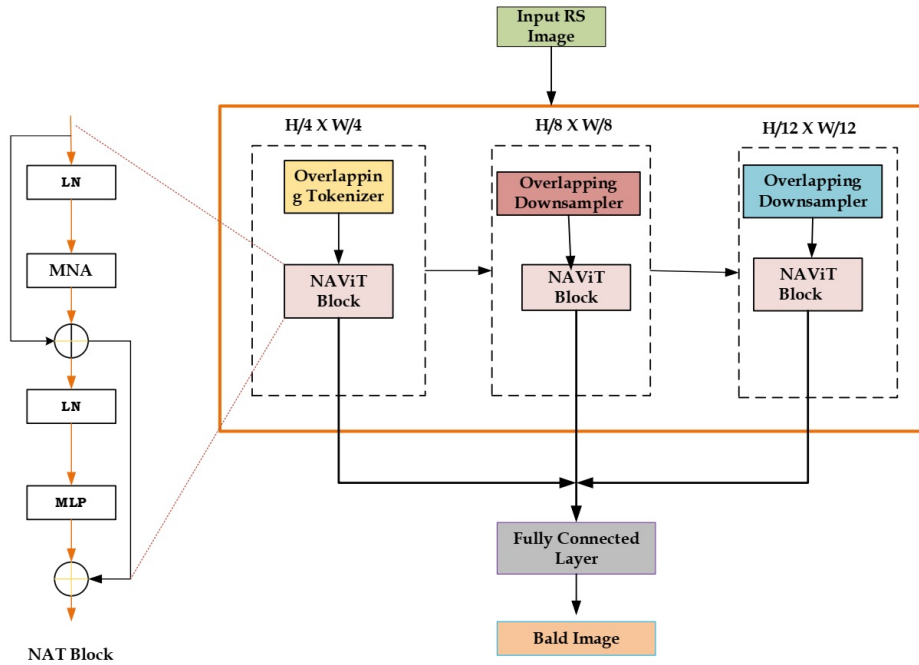

Figure 2. NAVIT block structure

MHNA, a multi-layer perceptron (MLP), GN, and skip connections are the four primary parts of each NAViT block [24]. By assigning attention weights to each neighborhood, the MHNA mechanism divides the feature map into overlapping areas. While lowering processing costs, this concentrated attention improves the representation of fine-grained spatial patterns. Because the attention heads work in simultaneously, the model can capture a variety of geographical and semantic data at different scales.

The MLP within each NAViT block offers extra non-linearity and feature transformation capabilities. The MLP enables the model to capture intricate feature interactions by projecting the attention-enhanced features into and back into higher-dimensional environments. By normalizing the feature maps in groups, the GN layer guarantees numerical stability and speeds up convergence. This method works especially well for processing high-resolution RS data. Every NAViT block incorporates skip connections to alleviate the vanishing gradient issue and maintain spatial information. The model may use both shallow and deep features for precise feature extraction thanks to these links, which make it easier for information to move between layers. These elements work together to give NAViT blocks a good balance

between accuracy and efficiency, which makes them ideal for managing the complex and high-dimensional properties of RS pictures.

The down sampler reduces the spatial size by half while doubling the number of channels as shown in Figure 2. The down sampler reduces the spatial dimension by half while increasing the number of channels. This facilitates the computation of subsequent levels. Our model creates feature maps with dimensions of maps of size $\frac{H}{4} \times \frac{W}{4} \times \frac{H}{8} \times \frac{W}{8}$ and $\frac{H}{12} \times \frac{W}{12}$. It is calculated as follows:

$$\mathrm{MNA}(X_{ij}) = \mathrm{softmax}\left(\frac{Q_{ij}K^T_{\rho(i,j)} + B_{ij}}{Scale}\right)V_{\rho(i,j)}. \tag{1}$$

where $\rho(i,j)$ denotes the neighborhood of a pixel at $(i,j) \parallel \rho(i,j) \parallel = S^2$ (S is the neighborhood size), Q, K, and V are linear projections of input X, $B_{i,j}$ is the relative position bias. Encoder and Decoder are two integration steps involved in UNet. The NAViT module replaces the UNet encoder's traditional convolutional layers, which employ attention methods to collect local information, resulting in the encoded feature representation. The encoded feature representation from NAViT is connected to the UNet decoder, which gradually up samples and refines the features to create the final segmentation map.

MNA block plays a crucial role in the NAViT block because it allows for the automated learning of changes to the contribution of feature mapping from each temporal image. This means that MNA enables the model to dynamically modify the relevance or focus given on different elements or areas of an image stream. In other words, MNA improves the network's capacity to focus on particular information across frames by offering a method for adaptively weighting individual frames' contributions to the overall learning process.

It can help the model understand how certain regions change over time or identify specific events or phenomena across multiple temporal frames. It is a hierarchical transformer design that uses Neighborhood Attention (NA) [25] to improve image classification. Figure 3 illustrates the query-key-value structure of NA vs. SA for a single pixel. The NA and SA Block in the context of the "NAViT Transformer" are important components utilized in this unique transformer-based architecture, developed for activities requiring the capture of both global and local dependencies in data. The "NAViT Transformer" concept expands the Transformer model by incorporating neighborhood attention to capture local context. It is expressed by

$$\mathrm{Attention}(Q,K,V) = \mathrm{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V. \tag{2}$$

Dot product attention, where d is the embedding dimension, is used for SA over linear projections of the same input as the query and key-value pairs.

NA allows a single pixel to attend to a local neighborhood of neighboring pixels, computing attention weights and considering relationships between the central pixel and its neighbors. A sliding pane that is specifically Attention restricts each pixel's

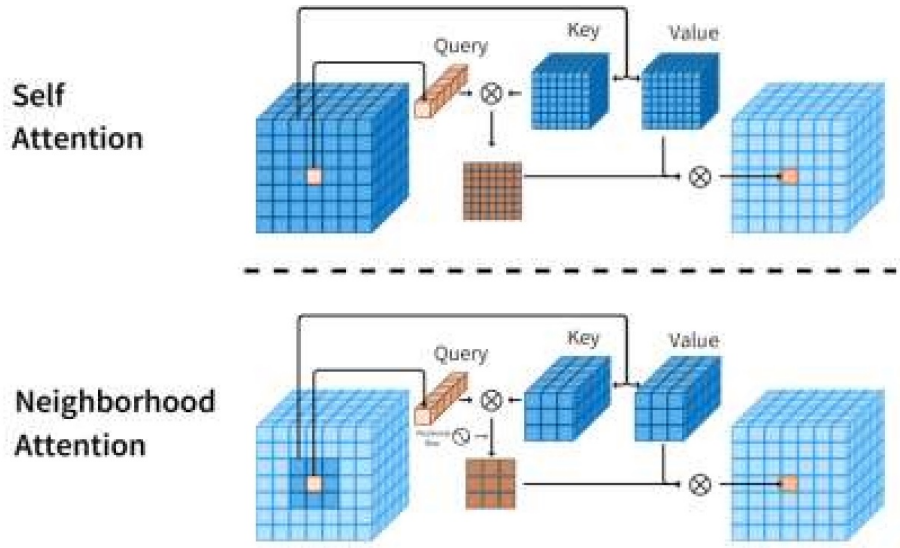

Figure 3. NA block

focus to its nearest neighbour while preserving linear equivariance, approaching SA as the span grows. SA is a mechanism that computes attention weights based on the relationships between different features within a single pixel. It is often used in tasks like image denoising, where the pixel's features are used to estimate its value. Prior to applying the attention weights to the data, the SA mechanism normalizes them using Softmax and calculates them using the dot product of the query and key.

By using these blocks, the framework's NAViT-based encoder creates hierarchical feature representations that gradually gather data at coarse and fine scales. In MT RS pictures, where both local details and global context are important, this skill is essential for detecting changes.

3.3 Multi Temporal Fusion Network

This is designed to dynamically adjust and combine feature contributions from different temporal images, enabling more effective analysis of multi-temporal RS data. By properly using correlation weight to aggregate the complimentary data, the MTF performs better than the conventional fusion algorithms currently in use. The characteristics concatenated from each down-sampling layer are fed into the up-sampling layers with two output channels via the MTF. This component is vital for capturing complementary information from temporal images, as it adapts to the varying quality and characteristics of the input data while ensuring accurate segmentation. Figure 4 depicts the structure of MTF. It integrates the MTF network into the de-

sign, allowing it to change the effect of different temporal feature maps dynamically. The fusion process begins by generating a weight matrix used a sigmoid activation function, which provides normalized weights for the input feature maps from multiple temporal images. The sigmoid function is particularly suited for this task as it outputs value between 0 and 1, effectively scaling the contributions of feature maps while maintaining numerical stability during the fusion process.

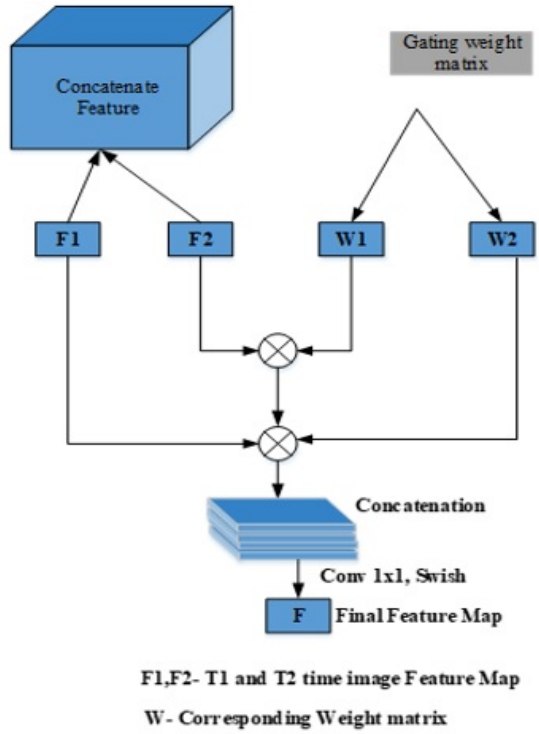

Figure 4. Structure of MTF network

Following its creation, the weight matrix is divided into two distinct maps that match the feature distributions of the two concurrent images. These impacts acts as the dynamic scaling factors for the respective feature maps allowing the network to focus more on ROIs in each temporal image while suppressing less relevant or noise regions. Each feature map is then multiplied by its corresponding weight map, ensuring that the relative importance of each temporal feature is preserved and enhanced during this step. Following the weighting process, the temporal images are aggregated to perform feature fusion. This aggregation integrates the refined feature representations into a single, unified collection of feature maps, effectively incorporating MS and complimentary information from temporal images. The MTF network surpasses existing fusion algorithms, which sometimes rely on static or pre-

defined weights that may fail to capture the subtle interactions between temporal images.

Mathematically MTF Network Can be expressed as

- Let $F_1(X)$ is the feature extracted by the UNet encoder from the input X.
- $F_2(X)$ is the features extracted by NAViT blocks capturing temporal dependencies in the input (assumed to be a sequence of images).
- $F'(X)$ is the fused features combining information from both UNet and NAViT blocks.
- Y_{UNet} is the output segmentation mask generated by the UNet decoder.

Combining the features of UNet and NAViT using concatenation the expression becomes

$$F'(X) = F_1(X) \oplus F_2(X) \tag{3}$$

(or)

$$F'(X) = F_1(X) + F_2(X). \tag{4}$$

The combined features $F'(X)$ are passed through the UNet Decoder to generate the final Segmentation Y_{UNet}. Hence overall mathematical expression is mentioned in Equation (5):

$$Y = \text{UNet Decoder}(F'(X)). \tag{5}$$

Incorporating this fusion mechanism within the NAViT layers ensures that MS features are efficiently fused throughout the encoder. This dynamic modification and fusing procedure considerably improves the network's ability to handle complicated temporal variations, allowing for more accurate segmentation and CD in MT RS images. The resulting fused features preserve valuable contextual information, which is critical for identifying and distinguishing fine-grained details and minor changes in the spatial landscape.

3.4 Bottleneck Layer with ASPP

The bottle neck layer serves as a crucial intermediary between the encoder and decoder ensuring the preservation and enhancement of semantic features while expanding the receptive field. This layer addresses the challenge of effectively capturing multi scale contextual information which is essential for segmentation fine-grained details in RS images, especially when dealing with complex or small object regions.

3.4.1 Cascaded Atrous Convolution

The ASPP module aids in the collection of MS contextual information, which can increase network performance in tasks such as object segmentation. Even if ASPP can handle items of different sizes to some extent, the practical scenarios' actual complexity is far more than ASPP's [26]. All of the items from which ASPP can

gather data have sizes that only differ by four distinct sizes. However, sampling enough data to provide thorough and accurate features is challenging in open settings due to unpredictable object sizes and distributions. There are two distinct manifestations of this problem. First, ASPP with limited sample ranges cannot fully gather the features of object components that are dispersed and disconnected over a wide scope and create a global understanding of target objects due to the arbitrary sizes of the objects in practice [27]. The structure of a fully convolutional network combining encoder–decoder structures with ASPP is shown in Figure 5. The approach intelligently includes characteristics from distinct receptive fields in the proposed design with four cascaded distributed atrous convolutions (r = 1, 4, 8, 12) by deploying a series of atrous convolutions with varied dilation rates. The dilation rate of 1 focuses on local features, ensuring that fine details in small regions are captured. Larger dilation rates such as 4, 8 and 12 progressively broaden the receptive field, allowing the network to incorporate global and contextual spatial features. By gradually merging data from multiple layers, the model improves its capacity to capture hierarchical representations and hence improves its ability to recognize patterns at different sizes. This subtle integration allows for a more thorough comprehension of the input data, which might lead to increased performance in tasks like image analysis or semantic segmentation. The existing study [24] employed four parallel distributed atrous convolutions with dilation rates of (r = 1, 6, 12, 18). While both approaches aim to combine features from different receptive fields, the proposed architecture with cascaded convolutions may provide advantages in terms of computational efficiency and more flexible feature extraction because each layer is connected sequentially, allowing for a continuous flow of information across the network. In comparison to the parallel method, the cascaded design enables a progressive and organized integration of elements, perhaps resulting in a more nuanced representation of the incoming data.

This cascading structure ensures that features from multiple receptive fields are aggregated seamlessly. By incorporating MS information, the bottleneck enhances the detection and segmentation of small objects and subtle changes in RS images which are often challenging to identify using standard convolution.

Atrous convolution is a potent method for altering the resolution of features identified by DCNNs and modifying the response field used to capture MS information. CNNs capture MS information by altering the receptive field. The application of atrous convolution to the input x for each pixel i on the output y and filter w is demonstrated in Equation (6) [28].

$$y[i] = \sum_{k} x[i + r.k]w[k]. \tag{6}$$

Four parallel atrous convolutions with varying atrous rates make up the ASSP module, was created to enhance the UNet technique. An image-level feature produced by global average pooling, three parallel 3×3 strides with rates of 1, 4, 8, and 12, respectively, and one 1×1 convolution comprises the ASPP module. After bilinearly

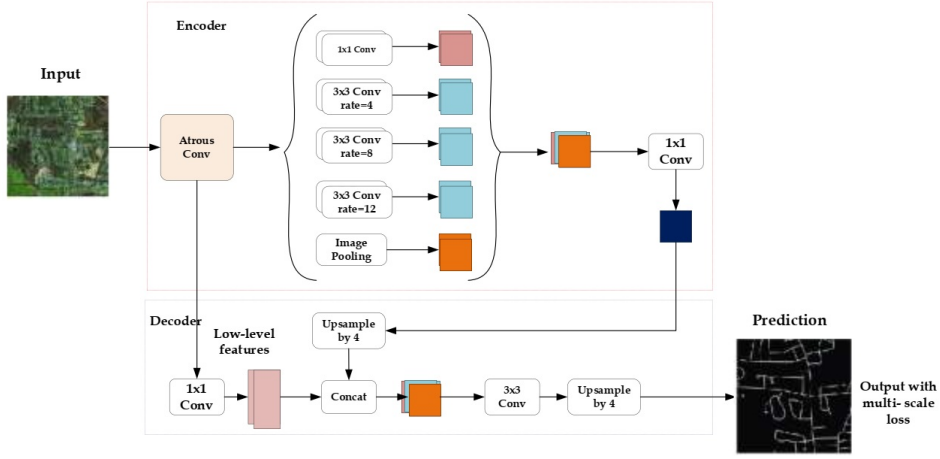

Figure 5. The architecture of a fully convolutional network with the fusion of ASPP and encoder–decoder structures

upsampling each branch's output features to the input size, they are concatenated and fed into a second 1×1 convolution. As seen in Figure 6, ASPP is applied to the feature map of the encoder portion, and the resultant feature map is then sent into the decoder portion [29].

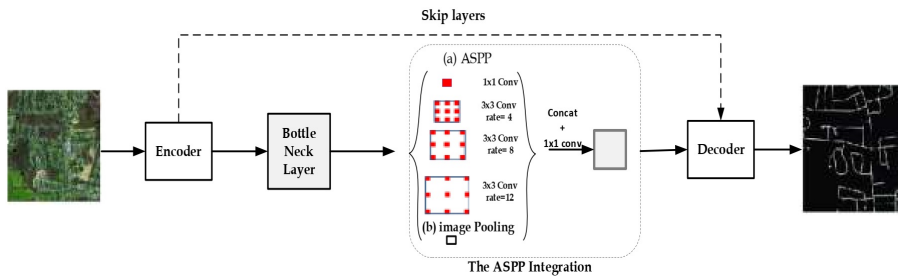

Figure 6. Proposed architecture of UNet with ASPP

3.4.2 Spatial Pyramid Pooling

To complement the atrous convolution, the bottleneck integrates spatial pyramid pooling which processes features extracted at different scales. The pooling mechanism partitions the feature maps into multiple regions and extracts representative features from each region, creating a richer multilevel representation. The network maintains both localized details and global context by concatenating these pooled features before passing them through later levels. The bottleneck efficiently con-

nects the encoder and decoder by combining spatial pyramid pooling and cascaded atrous convolution, offering improved feature representations for segmentation applications.

3.4.3 Decoder with Upsampling Layers

Each upsampling layer has two output channels and processes concatenated features received from the corresponding downsampling layers in the encoder. These layers ensure that high-resolution spatial features are integrated with the abstract features learned during encoding, preserving critical details for accurate segmentation. The upsampling operation effectively recovers spatial information lost during the downsampling process ensuring a seamless flow of information across the network

The final segmentation map is generated by concatenating processed features and applying the Swish activation function which is used in a neural network architecture for exact segmentation, incorporating several loss functions for performing various aspects of the task. It employs the sigmoid function to create smooth, non-monotonic non-linearity, allowing for stable optimization and the capture of complicated data patterns. Swish outperforms ReLU in terms of training convergence and model generalization, making it an attractive function for DNN activation functions. However, the efficacy of Swish is dependent on the dataset, network design, and task characteristics.

For every input element with values below zero, the ReLU activation function establishes a threshold, which is provided by

$$f(x) = \max(0, x) = f(x) = \begin{cases} x_i, & x_i \geq 0, \\ 0, & x_i < 0. \end{cases} \tag{7}$$

A hybrid function that combines the sigmoid and ReLU activation functions is called a self-gated activation function (Swish).

$$f(x) = x.\text{sigmoid}(\beta.x) = \frac{x}{1 + e^{\beta x}}; -\infty < x < +\infty, \tag{8}$$

where β is the parameter that can be predefined, Swish is more efficient than ReLU in most circumstances, and its key advantage is its smoothness property. The following relationship gives the equation Swish activation function:

$$f(x) = \begin{cases} \alpha.x.\text{sigmoid}(\delta.x), & x \leq \beta, \\ \max(\beta, x), & x > \beta, \end{cases} \tag{9}$$

where α, β, δ are the three parameters.

By employing the swish function and leveraging the precise feature refinement of decoder, the proposed method ensures high quality segmentation, particularly in

challenging scenarios with complex and small object regions. The combined power of bottleneck and decoder guarantees robust and accurate segmentation outcomes for CD in multi-temporal RS images.

To tackle the challenges of uneven segmentation and fuzzy boundaries in RS images, the proposed method incorporates a mixed loss function. This innovative loss design combines three complementary loss functions, each addressing specific challenges in segmentation tasks. Together, they enable precise and balanced learning, even in scenarios with complex boundaries, small object regions and class imbalances.

MS-SSIM Loss.

The anticipated and ground truth segmentation maps' similarity across several scales is measured by the Multi-Scale Structural Similarity Index (MS-SSIM) loss. This loss function ensures that the model captures fine-grained details and preserves structural integrity in segmented regions.

Equation (10) depicts the binary cross-entropy (BCE) [30] loss function of a broad semantic segmentation, which is used as the basic loss function in the proposed network.

$$L_{BCE} = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log \hat{y}_i + (1-y_i)\left(1-\log \hat{y}_i\right)\right], \tag{10}$$

where N is the total number of pixels, y_i is the i^{th} pixel's ground truth, and $\hat{y}_i$ is its prediction.

The SSIM [31] compares the luminance, contrast, and structure of two images to determine their resemblance.

The combination of the Swish activation function, mixed loss functions, and a suitable neural network architecture provides a comprehensive method for exact image segmentation. This approach should be proven by extensive testing and assessment of individual datasets, taking into account the unique characteristics and the needs of the segmentation task.

To begin, the brightness is compared by calculating the mean intensity.

$$\mu_x = \frac{1}{n}\sum_{i=1}^{N} x_i. \tag{11}$$

As specified in Equation (11), the function that combines the three elements is the ultimate similarity metric.

$$S(x,y) = f(l(x,y), c(x,y), s(x,y)). \tag{12}$$

The luminance comparison of three functions is defined in Equation (13):

$$l(x,y) = \frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1}, \tag{13}$$

where C_1 is a constant used to prevent instability as $\mu_x^2 + \mu_y^2$ approaches zero. Then the contrast comparison is defined as

$$c(x,y) = \frac{2\sigma_x\sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2}, \tag{14}$$

C_2 is also a constant. After normalizing the signals, the structural comparison is estimated. The correlation is a straightforward and useful metric for assessing structural similarity, as defined in Equation (15):

$$s(x,y) = \frac{\sigma_{xy} + C_3}{\sigma_x\sigma_y + C_3}. \tag{15}$$

σ_{xy} is calculated by

$$\sigma_{xy} = \frac{1}{N-1}\sum_{i=1}^{max}(x_i - \mu_x)(y_i - \mu_y). \tag{16}$$

By comparing Equations (12), (13), (14) and (15), the final SSIM measurement will get, which is defined in Equation (17):

$$\mathrm{SSIM}(x,y) = \left[l(x,y)^{\alpha}.c(x,y)^{\beta}.s(x,y)^{\gamma}\right]. \tag{17}$$

where $\alpha > 0, \beta > 0, \gamma > 0$ are the parameters, for assigning various weights to the three components. To simplify the parameters, set the value as $\alpha = \beta = \gamma = 1$ and $C_3 = C_2/2$ [31].

This configuration is a commonly used initial setting that makes future comparisons with related works easier in addition to providing a succinct description of the SSIM.

$$\mathrm{SSIM}(x,y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}. \tag{18}$$

When training the network, the measurements must be condensed into a single function. As a result, a mean SSIM (MSSIM) loss is employed to assess overall image quality is defined in Equation (19):

$$\mathrm{SSIM}(X,Y)\frac{1}{M}\sum_{j=1}^{M}\mathrm{SSIM}(x_y, y_j). \tag{19}$$

The final loss function of the SSIM is given:

$$L_{SSIM}(X,Y) = -\frac{1}{M}\frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}. \tag{20}$$

The network was trained by combining the SSIM and BCE losses, and Equation (21) was used to determine the final loss.

$$L = L_{BCE} + L_{SSIM}. \tag{21}$$

By incorporating MS analysis, MS-SSIM Loss is a MS analysis framework that effectively segments intricate boundaries and small details in RS images, prioritizing local structures preservation for detecting subtle changes in MT datasets.

Tversky Loss (TL).

This loss is specifically designed to address class imbalance by modifying the standard Dice similarity coefficient, a common issue in RS image segmentation where certain classes may be underrepresented. It introduces tunable parameters to control the weighting false positives and false negatives enabling the model to focus on small or underrepresented regions in segmentation tasks [32].

The Tversky Index (TI) is defined as

$$\mathrm{TI} = \frac{|\mathrm{TP}|}{|\mathrm{TP}| + \alpha|\mathrm{FP}| + \beta|\mathrm{FN}|}, \tag{22}$$

where α and β are the weights controlling the tradeoff between FP and FN. The Tversky loss is calculated as

$$L_{Tversky} = 1 - \mathrm{TI}. \tag{23}$$

This formulation allows for flexible handing of imbalanced datasets by adjusting α and β.

This makes it particularly effective for capturing minor details and reducing errors in highly imbalanced datasets ensuring that small yet significant regions are not overlooked.

Focal Loss (FL).

This loss enhances the model's performance in challenging regions by addressing the imbalance between easy to classify and hard to classify examples. It achieves this by downweighting the contribution of easily classified pixels while emphasizing harder examples such as fuzzy boundaries or regions with significant overlap between classes [32]. This focused strategy improves segmentation accuracy in challenging situations by assisting the model in learning more efficiently in regions that are prone to ambiguity or misclassification. The FL is defined as

$$L_{Focal} = -\sum_{i} \alpha(1 - p_i)^{\gamma} \cdot \log(p_i), \tag{24}$$

where p_i is the predicted probability for the correct class, α is a weighted factor to balance the importance of positive and negative samples, γ is the focusing parameter to adjust the contribution of hard to classify samples.

The term $(1 - p_i)^\gamma$ reduces the weight of well classified examples p_i close to 1 and emphasizes harder examples p_i close to 0. A higher γ places greater focus on hard to classify pixels while α balances the importance of different classes.

By integrating these three loss functions, the mixed loss function provides a balanced framework the critical challenges inherent in RS image segmentation. The MS-SSIM ensures the preservation of structural details by capturing MS similarity which is vital for maintaining the integrity of fine-grained features and complex boundaries. The TL effectively mitigates class imbalance, particularly for small and underrepresented regions by dynamically adjusting the tradeoff between FPs and FNs, thereby ensuring that even minor yet significant areas are accurately segmented. Meanwhile, FL enhances the model's focus on difficult-to-classify regions such as fuzzy boundaries or ambiguous transitions by prioritizing harder examples while downweighting the influence of easily segmented areas.

This synergistic combination empowers the network to capture global and local features, enabling precise segmentation across uneven, small, and complex regions of RS images. The mixed loss Function holistically addresses segmentation challenges, ensuring robust learning, improved boundary delineation, and enhanced model performance for CD tasks in MT RS images.

4 RESULTS AND DISCUSSION

The model of an improved NAViT using the UNet method with two enhancements serves as the foundation for the suggested strategy. SSIM Loss is used to enhance the quality, focus the fuzzy border, and accomplish accurate segmentation. ASSP is used to capture MS features and to extend the feature receptive field. This approach conducted trials with three models to validate our approach: the baseline NVT, a bottleneck layer using ASPP-integrated UNet with normal loss, and the ASPP-UNet with SSIM loss. Except for the network design and loss function alterations indicated in Table 1, the convolutional layer, activation, and optimizer hyperparameters of the three models are invariant to track the impact of suggested improvements.

Model Name	**ASSP Integration**	**Loss Function**
UNet	No	L_{BCE}
ASSP-UNet	Yes	L_{BCE}
ASSP-UNet + SSIM	Yes	$L_{BCE} + L_{SSIM}$

Table 1. Network configuration of the three models

4.1 System and Tool Configuration

Python 3.9 was used as the major programming language for the implementation. Python's broad ecosystem of tools and frameworks provides a flexible and efficient

environment for developing deep learning models and performing image processing tasks. This implementation runs on the 64-bit version of Windows 10 with an Intel Pentium CPU and 16 GB of RAM.

4.2 Data Set Description

A brand-new, extensive building CD dataset based on RS is called LEVIR-CD. There are 637 extremely high-resolution, 0.5 m/pixel Patch pairs of Google Earth images that are 1 024 × 1 024 pixels in size [15]. The proposed work would serve as a new standard for evaluating CD methods, particularly those based on DL. Model training uses 4 923 pairs of samples in total, of which 10 % are randomly selected as validation samples. These samples are used just to assess the network training procedure and are not involved in model training. For a total of 10 iterations, each of the 32 sample sets in each training batch are iterated once for 100 Epochs. The learning rate is continuously modified based on the training procedure, with a beginning training learning rate of 0.001. In order to experiment sequentially with several networks with the same parameters for this dataset, UNet, AUNet, SiameseNet, Siamese_AUNet, and NAViT-UNet, where the network model employed, a Python script is used to train numerous networks one at a time. In this process, the dataset is split into training (144 samples), testing (20 samples), and validation (36 samples). The overall percentage of training, testing, and validation is 72 %, 18 %, and 10 %, respectively.

4.3 Performance Metrics

The following four evaluation metrics are used to assess the accuracy of the outcomes.

Precision is a metric that calculates the ratio of correctly predicted positive samples to all anticipated positive samples. It is calculated as follows:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}. \tag{25}$$

The ratio of correctly predicted positive samples to all true positives is known as Recall.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \tag{26}$$

The total of the precision and recall rates is the F1-score value, which is represented as

$$\text{F1} = \frac{2}{\frac{1}{\text{precision}} + \frac{1}{\text{recall}}} = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}. \tag{27}$$

The Dice is commonly used to compare two samples' similarity:

$$\text{Dice} = \frac{2 \times \text{TP}}{2 \times \text{TP} + \text{FP} + \text{FN}}. \tag{28}$$

The capacity of a CD technique to accurately recognize and categorize changes in RS images is known as Accuracy. Overall accuracy can be calculated.

$$\text{Accuracy} = \frac{(\text{TP} + \text{TN})}{(\text{TP} + \text{TN} + \text{FP} + \text{FN})}. \tag{29}$$

4.4 Experimental Results

The proposed NAViT with UNet method was evaluated on multi-temporal RS datasets and demonstrated superior performance in CD. The output of the proposed NAViT with UNet approach is shown in Figure 7, which also includes the source images, the matching ground truth masks, and the anticipated segmentation outcomes. The model's ability to precisely identify and segment ROI is demonstrated by the anticipated outputs closely match with the ground truth masks. The method's robustness is demonstrated by the exact detection of changes and the clear delineation of boundaries, even in small and complex areas. The model's capacity to successfully manage the difficulties of MT RS image segmentation is demonstrated by this depiction.

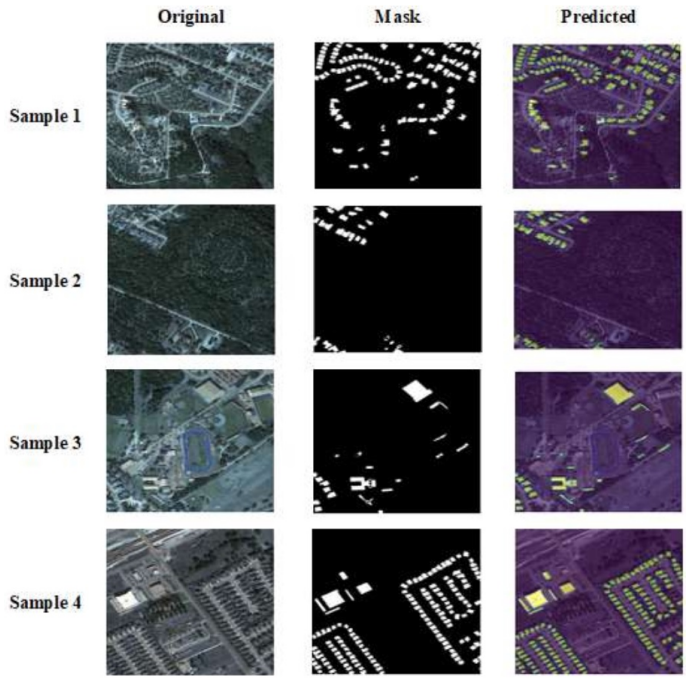

Figure 7. Experimental results of CD using the LEVIR dataset

The performance assessment of the proposed model is shown in Table 2 and is based on important metrics. The method's remarkable accuracy of 0.9988 shows how reliable it is in accurately detecting changes in MT RS images. The model's ability to reduce false positives and successfully identify real changes is demonstrated by its precision of 0.9881 and recall of 0.9932, respectively. The model's outstanding segmentation performance is further supported by the F1-score of 0.9907 and the Dice coefficient of 0.9593, especially when handling small and complex regions. The loss function's ability to balance structural preservation, class imbalance, and challenging region segmentation is demonstrated by the mixed loss value of 0.2902. Together, these measures show how reliable, accurate, and robust the suggested method is, which makes it ideal for CD tasks in RS image segmentation.

Performance Metrics	**Values**
Accuracy	0.9988
Precision	0.9881
Recall	0.9932
F1	0.9907
Dice	0.9593
Mixed Loss	0.2902

Table 2. Performance results of the proposed method

Also, this research would analyse the performance with three different ratios of testing and training

1. 15 % testing and 85 % training,
2. 20 % testing and 80 % training, and
3. 10 % testing and 90 %.

The overall performance measures of the parameters for three ratios which shown in Figure 8 and the obtained values are represented in Table 3. Thus the result of the proposed technique, while implementing the training and testing ratio of 90–10 % whose accuracy of 99 % has been obtained. Our study found that NAViT with UNet exhibits stable metrics for high-resolution images, LEVIR dataset demonstrates that our proposed NAViT with UNet is more robust.

Training Ratios	**Precision**	**Recall**	**F1**	**Dice**	**Accuracy**	**Loss**
90–10 %	0.9881	0.9932	0.9907	0.9593	0.9988	0.2902
80–20 %	0.8889	0.9270	0.9070	0.8680	0.9885	0.4747
85–15 %	0.7343	0.8006	0.7660	0.7349	0.9699	0.5508

Table 3. Proposed performance results for different training testing ratios

The model's great generalization when trained on a bigger dataset is demonstrated by the 90–10 % ratio, which exhibits the best performance, attaining nearly

flawless scores across all criteria. Despite a minor drop in metrics, the 85–15 % ratio has strong dependability and a notably consistent precision and recall. The 80–20 % ratio performs the worst out of the three, especially in recall and the dice coefficient. This is most likely because the absence of training data affects the model's ability to handle small or complex regions. Overall, the findings demonstrate how training data size affects model performance, highlighting the fact that greater segmentation accuracy and resilience may be achieved by the model with bigger training datasets such as those in the 90–10 % ratio.

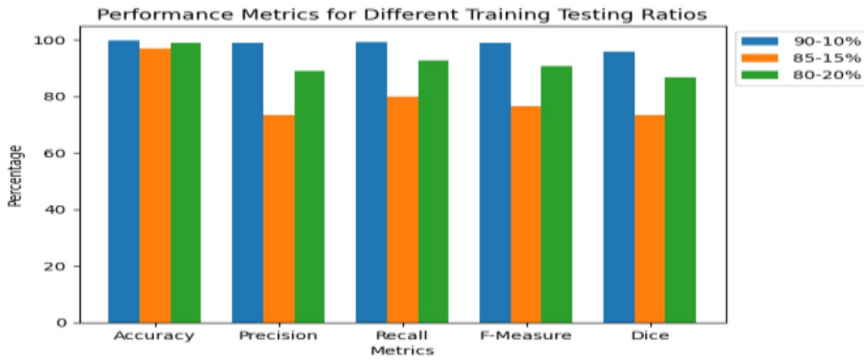

Figure 8. Overall performance results of proposed method in different ratios

4.5 Comparative Analysis

For this analysis, the proposed method compared with the existing method such as DASNet, UNet, AUNet, SiameseNet and Siamese_AUNet across the performance metrics such as precision, recall, dice and F1. Table 4 represents the comparative performances, and the visualization graph is shown in Figure 9.

Methods	Precision	Recall	F1	Dice
DASNet [25]	0.6942	0.8901	0.7661	0.7604
UNet [25]	0.8558	0.7355	0.7888	0.7911
AUNet [25]	0.8743	0.7802	0.8206	0.8248
SiameseNet [25]	0.8216	0.8283	0.8248	0.8329
Siamese_AUNet [25]	0.8582	0.8702	0.8557	0.8565
Proposed NAViT-UNet	0.9881	0.9932	0.9907	0.9593

Table 4. Comparative performance results

The proposed NAViT-UNet demonstrates its better segmentation skills by outperforming all other approaches in every statistic. Its highest precision (0.9881), recall (0.9932), F1-score (0.9907), and Dice coefficient (0.9593) demonstrate how

well it captures changes and preserves segmentation resilience. The model's ability to reduce false positives is shown in its high precision, while its ability to detect minor changes without overlooking important features is demonstrated by its great recall. Its accurate and balanced performance in MT RS image segmentation is further supported by the F1-score and Dice coefficient. DASNet demonstrates poor performance, particularly in precision 0.6942 and dice coefficient 0.7604, which is probably because it cannot handle small and complicated regions. Although UNet outperforms DASNet, it still falls behind in terms of total segmentation accuracy and recall (0.7355). Even with better recall and Dice coefficients, AUNet and SiameseNet still perform poorly when compared to the NAViT-UNet approach. With a balance between recall (0.8702) and precision (0.8582), Siamese_AUNet performs admirably. Its measures, however, fall short of the NAViT-UNet's precision and recall.

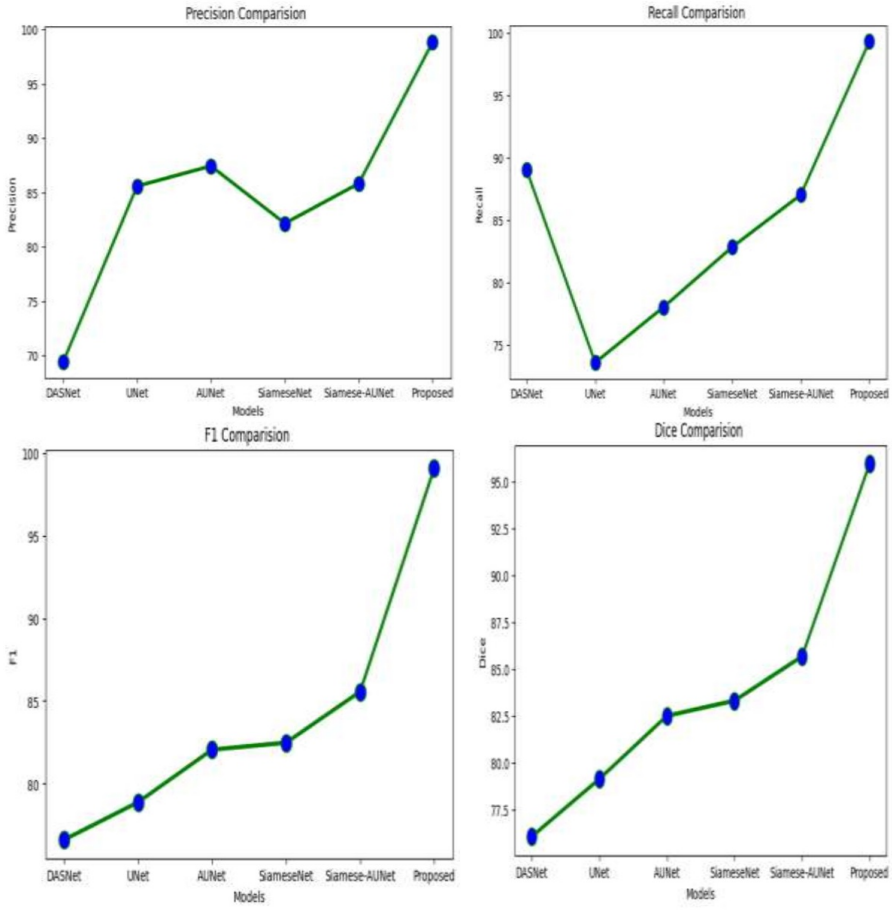

Figure 9. Comparative results of Precision, Recall, F1 and Dice performances

Figure 9 highlights the significant performance gap between the proposed NAViT-UNet and other methods. Although the segmentation accuracy of each method varies, the NAViT-UNet is the most dependable and efficient way for MT RS picture segmentation due to its superior performance across all parameters.

This comparative analysis confirms that the advanced architectural features including the NAViT, MTF and ASPP contribute to the NAViT-UNet's superior performance which ensure precise detection of changes and accurate segmentation even in complex RS images.

4.6 Discussion

The intrinsic difficulties of uneven segmentation, fuzzy borders, and class imbalance are addressed by the proposed NAViT-UNet, which exhibits remarkable performance in MT RS image segmentation. Both local, extremely fine features and global, context-specific data are successfully captured by the model by combining a NAViT with a novel MTF mechanism. This allows for precise segmentation of small and complicated locations. The feature receptive field is further improved by the addition of the Bottleneck Layer with ASPP, which guarantees reliable processing of MS data. With exceptional performance metrics – achieving an accuracy of 99.88 %, a precision of 98.81 %, and a recall of 99.32 % – the trial results confirm the efficacy of NAViT-UNet. The resilience of NAViT-UNet in capturing subtle and complex changes is demonstrated by its strong outperformance in all major measures, including F1-score and Dice coefficient, when compared to existing approaches such as DASNet, UNet, AUNet, SiameseNet, and Siamese_AUNet.

Additionally, the Mixed Loss Function guarantees that the model tackles issues like unbalanced datasets and difficult-to-classify regions by integrating MS-SSIM Loss, TL, and FL. By enabling accurate delineation of fuzzy borders and enhancing segmentation accuracy, this synergy strengthens the model's resilience. The NAViT-UNet's adaptability to different training-testing ratios further demonstrates its scalability and suitability for a range of RS scenarios. All things considered, this study demonstrates NAViT-UNet's promise as a state-of-the-art approach to RS image segmentation. It is a useful tool for CD and segmentation in RS applications because of its capacity to incorporate loss functions and sophisticated architectural characteristics, setting a new standard for accuracy and resilience in the industry. Future studies might look into applying this strategy to different fields or adding further improvements to achieve even greater performance improvements.

4.6.1 Ablation Studies

In order to assess the contribution of each module in NAViT-UNet framework, several ablation experiments were conducted. These experiments were designed to isolate and evaluate the individual impact of NAViT, the cascaded ASPP module and mixed loss function. By systematically removing or replacing each component, access the impact on performance metrics such as accuracy, precision, recall and F1

score. The result reveal the critical role of each module in enhancing the model's performance.

NAViT Module: Removing the NAViT module resulted in a significant drop in accuracy and recall which indicating that local feature extraction through attention mechanisms is crucial for capturing multi-scale details in RS imagery.

ASPP Module: Replacing the ASPP module with a simple conv layer caused a slight decrease in segmentation accuracy particularly in complex boundary regions which highlight the importance of broad receptive field.

Mixed Loss Function: Removing the Mixed loss function which integrates the MS-SSIM, Tversky loss and Focal loss led to poorer segmentation performance, especially in cases with class imbalance or small, hard-to-classify regions.

These findings confirm the significance of each module and provide insights into their contribution to the overall success of NAViT-UNet framework.

5 CONCLUSION

This research introduces NAViT-UNet, a unique approach for MT RS image segmentation, combining the strengths of NAViT with UNet. The model delivers improved segmentation accuracy by combining novel elements such as the Neighbourhood Attention mechanism, MTF, and Bottleneck Layer with ASPP particularly in handling complex boundaries and small objects. By combining MS-SSIM Loss, TL, and FL, the Mixed Loss Function guarantees balanced optimization while tackling issues like class imbalance and hard-to-classify areas. Extensive experiments demonstrate that NAViT-UNet significantly outperforms existing methods obtaining remarkable metrics including an accuracy of 99.88 %, a precision of 98.81 % and a recall 99.32 %. Its scalability across different training testing ratios and robustness in diverse scenarios further highlight its applicability and effectiveness. Thus, NAViT-UNet sets a new benchmark for RS image by delivering accurate, robust and efficient solutions for CD tasks. This research provides a strong foundation for future advancements in RS images with potential applications extending to various domains requiring precise image analysis. Future work could explore extending NAViT-UNet for RS applications and integrating it with self-supervised learning techniques to reduce dependency and enhance its performance across diverse environments and domains.

REFERENCES

[1] Asokan, A.—Anitha, J.: Change Detection Techniques for Remote Sensing Applications: A Survey. Earth Science Informatics, Vol. 12, 2019, No. 2, pp. 143–160, doi: 10.1007/s12145-019-00380-5.

[2] Wang, S. W.—Gebru, B. M.—Lamchin, M.—Kayastha, R. B.—Lee, W. K.: Land Use and Land Cover Change Detection and Prediction in the Kathmandu District of Nepal Using Remote Sensing and GIS. Sustainability, Vol. 12, 2020, No. 9, Art. No. 3925, doi: 10.3390/su12093925.

[3] Shi, W.—Zhang, M.—Zhang, R.—Chen, S.—Zhan, Z.: Change Detection Based on Artificial Intelligence: State-of-the-Art and Challenges. Remote Sensing, Vol. 12, 2020, No. 10, Art. No. 1688, doi: 10.3390/rs12101688.

[4] Khelifi, L.—Mignotte, M.: Deep Learning for Change Detection in Remote Sensing Images: Comprehensive Review and Meta-Analysis. IEEE Access, Vol. 8, 2020, pp. 126385–126400, doi: 10.1109/ACCESS.2020.3008036.

[5] Lalitha, V.—Latha, B.: A Review on Remote Sensing Imagery Augmentation Using Deep Learning. Materials Today: Proceedings, Vol. 62, 2022, pp. 4772–4778, doi: 10.1016/j.matpr.2022.03.341.

[6] You, Y.—Cao, J.—Zhou, W.: A Survey of Change Detection Methods Based on Remote Sensing Images for Multi-Source and Multi-Objective Scenarios. Remote Sensing, Vol. 12, 2020, No. 15, Art. No. 2460, doi: 10.3390/rs12152460.

[7] El Mendili, L.—Puissant, A.—Chougrad, M.—Sebari, I.: Towards a Multi-Temporal Deep Learning Approach for Mapping Urban Fabric Using Sentinel 2 Images. Remote Sensing, Vol. 12, 2020, No. 3, Art. No. 423, doi: 10.3390/rs12030423.

[8] Zhang, M.—Liu, Z.—Feng, J.—Liu, L.—Jiao, L.: Remote Sensing Image Change Detection Based on Deep Multi-Scale Multi-Attention Siamese Transformer Network. Remote Sensing, Vol. 15, 2023, No. 3, Art. No. 842, doi: 10.3390/rs15030842.

[9] Wang, Z.—Zhang, Y.—Luo, L.—Wang, N.: TransCD: Scene Change Detection via Transformer-Based Architecture. Optics Express, Vol. 29, 2021, No. 25, pp. 41409–41427, doi: 10.1364/OE.440720.

[10] Zhang, C.—Yue, P.—Tapete, D.—Jiang, L.—Shangguan, B.—Huang, L.—Liu, G.: A Deeply Supervised Image Fusion Network for Change Detection in High-Resolution Bi-Temporal Remote Sensing Images. ISPRS Journal of Photogrammetry and Remote Sensing, Vol. 166, 2020, pp. 183–200, doi: 10.1016/j.isprsjprs.2020.06.003.

[11] Tan, K.—Zhang, Y.—Wang, X.—Chen, Y.: Object-Based Change Detection Using Multiple Classifiers and Multi-Scale Uncertainty Analysis. Remote Sensing, Vol. 11, 2019, No. 3, Art. No. 359, doi: 10.3390/rs11030359.

[12] Zhang, J.—Pan, B.—Zhang, Y.—Liu, Z.—Zheng, X.: Building Change Detection in Remote Sensing Images Based on Dual Multi-Scale Attention. Remote Sensing, Vol. 14, 2022, No. 21, Art. No. 5405, doi: 10.3390/rs14215405.

[13] Caye Daudt, R.—Le Saux, B.—Boulch, A.: Fully Convolutional Siamese Networks for Change Detection. 2018 25th IEEE International Conference on Image Processing (ICIP), 2018, pp. 4063–4067, doi: 10.1109/ICIP.2018.8451652.

[14] Du, B.—Ru, L.—Wu, C.—Zhang, L.: Unsupervised Deep Slow Feature Analysis for Change Detection in Multi-Temporal Remote Sensing Images. IEEE Transactions on Geoscience and Remote Sensing, Vol. 57, 2019, No. 12, pp. 9976–9992, doi: 10.1109/TGRS.2019.2930682.

[15] Chen, H.—Shi, Z.: A Spatial-Temporal Attention-Based Method and a New Dataset for Remote Sensing Image Change Detection. Remote Sensing, Vol. 12, 2020, No. 10, Art. No. 1662, doi: 10.3390/rs12101662.

[16] Chen, J.—Yuan, Z.—Peng, J.—Chen, L.—Huang, H.—Zhu, J.—Liu, Y.—Li, H.: DASNet: Dual Attentive Fully Convolutional Siamese Networks for Change Detection in High-Resolution Satellite Images. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, Vol. 14, 2021, pp. 1194–1206, doi: 10.1109/JSTARS.2020.3037893.

[17] Fang, S.—Li, K.—Shao, J.—Li, Z.: SNUNet-CD: A Densely Connected Siamese Network for Change Detection of VHR Images. IEEE Geoscience and Remote Sensing Letters, Vol. 19, 2022, pp. 1–5, doi: 10.1109/LGRS.2021.3056416.

[18] Bai, B.—Fu, W.—Lu, T.—Li, S.: Edge-Guided Recurrent Convolutional Neural Network for Multitemporal Remote Sensing Image Building Change Detection. IEEE Transactions on Geoscience and Remote Sensing, Vol. 60, 2022, pp. 1–13, doi: 10.1109/TGRS.2021.3106697.

[19] Ma, C.—Weng, L.—Xia, M.—Lin, H.—Qian, M.—Zhang, Y.: Dual-Branch Network for Change Detection of Remote Sensing Image. Engineering Applications of Artificial Intelligence, Vol. 123, 2023, Art. No. 106324, doi: 10.1016/j.engappai.2023.106324.

[20] Zheng, H.—Gong, M.—Liu, T.—Jiang, F.—Zhan, T.—Lu, D.—Zhang, M.: HFA-Net: High-Frequency Attention Siamese Network for Building Change Detection in VHR Remote Sensing Images. Pattern Recognition, Vol. 129, 2022, Art. No. 108717, doi: 10.1016/j.patcog.2022.108717.

[21] Yang, W.—Song, H.—Du, L.—Dai, S.—Xu, Y.: A Change Detection Method for Remote Sensing Images Based on Coupled Dictionary and Deep Learning. Computational Intelligence and Neuroscience, Vol. 2022, 2022, Art. No. 3404858, doi: 10.1155/2022/3404858.

[22] Xu, X.—Zhou, Y.—Lu, X.—Chen, Z.: FERA-Net: A Building Change Detection Method for High-Resolution Remote Sensing Imagery Based on Residual Attention and High-Frequency Features. Remote Sensing, Vol. 15, 2023, No. 2, Art. No. 395, doi: 10.3390/rs15020395.

[23] Hassani, A.—Walton, S.—Li, J.—Li, S.—Shi, H.: Neighborhood Attention Transformer. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 6185–6194, doi: 10.1109/CVPR52729.2023.00599.

[24] Li, B.—Yang, T.—Zhao, X.: NVTrans-UNet: Neighborhood Vision Transformer Based U-Net for Multi-Modal Cardiac MR Image Segmentation. Journal of Applied Clinical Medical Physics, Vol. 24, 2023, No. 3, Art. No. e13908, doi: 10.1002/acm2.13908.

[25] Chen, T.—Lu, Z.—Yang, Y.—Zhang, Y.—Du, B.—Plaza, A.: A Siamese Network-Based U-Net for Change Detection in High Resolution Remote Sensing Images. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, Vol. 15, 2022, pp. 2357–2369, doi: 10.1109/JSTARS.2022.3157648.

[26] Wang, Y.—Liang, B.—Ding, M.—Li, J.: Dense Semantic Labeling with Atrous Spatial Pyramid Pooling and Decoder for High-Resolution Remote Sensing Imagery.

Remote Sensing, Vol. 11, 2019, No. 1, Art. No. 20, doi: 10.3390/rs11010020.

[27] Lian, X.—Pang, Y.—Han, J.—Pan, J.: Cascaded Hierarchical Atrous Spatial Pyramid Pooling Module for Semantic Segmentation. Pattern Recognition, Vol. 110, 2021, Art. No. 107622, doi: 10.1016/j.patcog.2020.107622.

[28] Chen, L. C.—Zhu, Y.—Papandreou, G.—Schroff, F.—Adam, H.: Encoder–Decoder with Atrous Separable Convolution for Semantic Image Segmentation. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (Eds.): Computer Vision – ECCV 2018. Springer, Cham, Lecture Notes in Computer Science, Vol. 11211, 2018, pp. 833–851, doi: 10.1007/978-3-030-01234-2_49.

[29] He, H.—Yang, D.—Wang, S.—Wang, S.—Li, Y.: Road Extraction by Using Atrous Spatial Pyramid Pooling Integrated Encoder–Decoder Network and Structural Similarity Loss. Remote Sensing, Vol. 11, 2019, No. 9, Art. No. 1015, doi: 10.3390/rs11091015.

[30] Goodfellow, I.—Bengio, Y.—Courville, A.: Deep Learning. The MIT Press, 2016.

[31] Wang, Z.—Bovik, A. C.—Sheikh, H. R.—Simoncelli, E. P.: Image Quality Assessment: From Error Visibility to Structural Similarity. IEEE Transactions on Image Processing, Vol. 13, 2004, No. 4, pp. 600–612, doi: 10.1109/TIP.2003.819861.

[32] Jadon, S.: A Survey of Loss Functions for Semantic Segmentation. 2020 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), 2020, pp. 1–7, doi: 10.1109/CIBCB48159.2020.9277638.

Satish Pawar is currently pursuing his Doctoral Programme in computer science and engineering from the Rajiv Gandhi Proudyogiki Vishwavidyalaya, Bhopal. He has received his M.Tech. degree in computer science engineering from the Samrat Ashok Technological Institute Vidisha, and his Bachelor's degree in computer science and engineering from RGPV Bhopal University. His research interests include image data analysis, data mining, deep learning, image processing and machine learning. He has published more than 10 articles in reputed peer-reviewed national and international journals. He has attended/presented his research papers in various seminars, conferences and workshops at the national and international levels.

Nishchol Mishra is working as Associate Professor at the School of Information Technology in the Rajiv Gandhi Proudyogiki Vishwavidyalaya, Bhopal. He has done his M.Tech. and Ph.D. in computer science and engineering. He has published over 40 papers in various reputed international journals. He has more than 20 years of research experience. His area of research includes multimedia, data mining, and image processing and data analytics.

Neelesh Mehra is working as Assistant Professor in the Electronics and Communication Engineering Department, Samrat Ashok Technological Institute Vidisha. He received his M.Tech. and Ph.D. degrees from the Electronics and Communication Department, MANIT Bhopal. His research interests include image processing, information security, wireless communication, application of machine learning, AI and IoT in the field of communication. He has notable publications in various national and international conferences, journals and as book chapters. He is a life member of ISTE, a senior member of ACM, and a member of IEEE.