Computers, Materials & Continua

Tech Science Press

Doi:10.32604/cmc.2025.064268

ARTICLE

# Hybrid HRNet-Swin Transformer: Multi-Scale Feature Fusion for Aerial Segmentation and Classification

**Asaad Algarni[1], Aysha Naseer [2], Mohammed Alshehri[3], Yahya AlQahtani[4], Abdulmonem Alshahrani[4] and Jeongmin Park[5,*]**

[1]Department of Computer Sciences, Faculty of Computing and Information Technology, Northern Border University, Rafha, 91911, Saudi Arabia
[2]Department of Computer Science, Air University, Islamabad, 44000, Pakistan
[3]Department of Computer Science, King Khalid University, Abha, 61421, Saudi Arabia
[4]Department of Informatics and Computer Systems, King Khalid University, Abha, 61421, Saudi Arabia
[5]Department of Computer Engineering, Tech University of Korea, 237 Sangidaehak-ro, Siheung-si, 15073, Gyeonggi-do, Republic of Korea

*Corresponding Author: Jeongmin Park. Email: jmpark@tukorea.ac.kr

Received: 10 February 2025; Accepted: 20 June 2025; Published: 29 August 2025

**ABSTRACT:** Remote sensing plays a pivotal role in environmental monitoring, disaster relief, and urban planning, where accurate scene classification of aerial images is essential. However, conventional convolutional neural networks (CNNs) struggle with long-range dependencies and preserving high-resolution features, limiting their effectiveness in complex aerial image analysis. To address these challenges, we propose a Hybrid HRNet-Swin Transformer model that synergizes the strengths of HRNet-W48 for high-resolution segmentation and the Swin Transformer for global feature extraction. This hybrid architecture ensures robust multi-scale feature fusion, capturing fine-grained details and broader contextual relationships in aerial imagery. Our methodology begins with preprocessing steps, including normalization, histogram equalization, and noise reduction, to enhance input data quality. The HRNet-W48 backbone maintains high-resolution feature maps throughout the network, enabling precise segmentation, while the Swin Transformer leverages hierarchical self-attention to model long-range dependencies efficiently. By integrating these components, our model achieves superior performance in segmentation and classification tasks compared to traditional CNNs and standalone transformer models. We evaluate our approach on two benchmark datasets: UC Merced and WHU-RS19. Experimental results demonstrate that the proposed hybrid model outperforms existing methods, achieving state-of-the-art accuracy while maintaining computational efficiency. Specifically, it excels in preserving fine spatial details and contextual understanding, critical for applications like land-use classification and disaster assessment.

**KEYWORDS:** Remote sensing; computer vision; aerial imagery; scene classification; feature extraction; transformer

## 1 Introduction

Scene classification in aerial and remote sensing imagery [1] plays a pivotal role in various applications, including land-use analysis, urban planning, precision agriculture, and disaster response. Automated analysis of aerial image scenes depends on precise classification abilities to track environmental changes and monitor natural resources, as well as infrastructure growth for management purposes. For analysis of aerial imagery, sophisticated extraction approaches are needed, particularly to cope with variations in scale and very complex spatial formations and hidden objects. Fused Boundaries make the classification task hard when they lead to the fact that images have different appearances, so that obtaining discrete representations of features is hard [2]. In addition, it is well-known that various remote sensing images may contain completely missing information. Image acquisition [3] can be affected by obstructions such as cloud cover, leading to incomplete data.

Image semantic segmentation is a key aspect of scene understanding, requiring precise identification of objects, their positions, and detailed boundaries within an image. Unlike image classification, semantic segmentation is more challenging [4] as it demands complex algorithms capable of pixel-level predictions and significantly higher computational power. Despite advancements, it remains an unsolved problem in low-level computer vision, particularly for complex scenes. Segmentation involves assigning each pixel to a specific semantic category, often dealing with objects of varying scales, irregular shapes, and blurred boundaries. Additionally, challenging conditions such as lighting variations, occlusions, and environmental complexities further increase the difficulty of accurate segmentation. As remote sensing technology [5] advances and images become higher in resolution, we can spot finer details and identify more objects than ever before. However, this also brings new challenges—similar-looking objects can have different spectral signatures, while different objects might look the same in certain wavelengths. When dealing with many classes, the limited number of training examples per category, decreasing differences between classes, and blurry boundaries make model training and prediction more difficult [6]. The key to solving this lies in extracting meaningful semantic information from both local and global contexts within an image. Models that can understand these relationships are better at making sense of what they "see".

Deep learning has completely transformed remote sensing image analysis by significantly improving how we extract and interpret spatial and spectral information. Traditional methods [7] had limitations, but modern deep learning models—especially convolutional neural networks (CNNs) and vision transformers (ViTs)—have demonstrated high effectiveness in segmentation tasks. To boost accuracy, researchers [8] often stack multiple complex modules, making networks harder to train and more time-consuming.

The key objective of this project is to create a sophisticated deep learning system that combines effective global spatial representation with high-resolution feature retention to enhance aerial image segmentation and scene classification. While Vision Transformers (ViTs) are computationally costly, traditional convolutional neural networks (CNNs) have trouble preserving fine-grained information and capturing long-range relationships. To overcome these obstacles, we suggest a Hybrid HRNet-Swin Transformer architecture that uses hierarchical self-attention and high-resolution feature extraction to attain better results in distant sensing applications. To make aerial imagery appropriate for environmental monitoring, urban planning, and disaster management, this project attempts to increase its classification accuracy, segmentation quality, and computational efficiency. This work's primary technological innovations include:

- A novel fusion of HRNet-W48 and Swin Transformer for improved semantic segmentation and scene classification in aerial images.
- An optimized framework that preserves fine details in segmentation while capturing contextual relationships across spatial scales.
- Enhanced feature extraction across different resolutions, ensuring better scene classification accuracy.

This article is structured in a way that allows for a thorough analysis of the suggested approach. To identify research gaps, Section 2 examines earlier studies on scene categorization using deep learning models and transformers. The Hybrid HRNet-Swin Transformer approach is thoroughly explained in Section 3, which also explains how HRNet segmentation maps improve feature learning in the transformer design. Details of the performance evaluation can be found in Section 4, which also visualizes the importance of attention-based features and shows gains over conventional methods. The benefits, drawbacks, and possible uses of our methodology in many fields are critically discussed in Section 5. A summary of the main conclusions and suggestions for advanced research to improve aerial scene classification is provided at the end of the study. Some images of selected datasets are shown in Fig. 1.

**Figure 1:** Original Images from respective datasets (**a**) WHU-RS19 (**b**) UC-Merced

## 2 Related Work

This section highlights state-of-the-art research on remote sensing applications and model architectures. Table 1 presents a comparison of existing studies on remote sensing scene classification and segmentation, highlighting their main contributions and associated limitations.

**Table 1:** Summary of related work in remote sensing scene classification and segmentation

| Authors | Main contributions | Research gaps |
|---|---|---|
| Liu et al. [9] | Proposed Residual Attention-Aggregation Network RAANet, a residual Astrous Spatial Pyramid Pooling (ASPP)-based attention framework for high-resolution remote sensing image segmentation. | High computational cost due to attention mechanisms; may not be practical for real-time applications. |
| Liu et al. [10] | Utilized CNN-based transfer learning for change detection in optical aerial images, improving detection accuracy. | Struggles with generalization across diverse datasets; sensitive to variations in image acquisition conditions. |
| Mou et al. [11] | Proposed a method combining Fully Connected Networks (FCN) and Recurrent Neural Networks (RNN) for aerial image segmentation. | The effectiveness of current segmentation techniques is hindered by occlusions, varying object scales, and blurred boundaries in aerial images. |
| Fan et al. [12] | Improved U-Net by incorporating attention mechanisms and multi-scale feature fusion for better remote sensing classification. | Limited performance improvement for highly complex scenes; still dependent on CNN-based architectures. |

(Continued)

**Table 1 (continued)**

| Authors | Main contributions | Research gaps |
|---|---|---|
| Nogueira et al. [13] | Analyzed different feature extraction levels and found VGG-16 achieved the highest classification accuracy (93%) on UC Merced and WHU-RS19 datasets. | Imbalanced class distributions in many aerial imagery datasets result in skewed model predictions. Did not explore Transformer-based models; suffered from class imbalance issues. |
| Zhang et al. [14] | Conducted a comparative study of CaffeNet and different CNN variants, concluding that fusion of ensemble architectures improves classification. | Although they offer superior global feature extraction, Vision Transformers (ViTs) are ineffective for large-scale aerial image analysis due to their high computational resource requirements. |
| Xu et al. [15] | Introduced MJDCNN, a hybrid CNN model integrating AlexNet, Inception-v3, and ResNet18, trained with momentum-driven Stochastic Gradient Descent (SGD). | Did not include RNN-based models; hybrid CNNs still struggled with generalization in complex aerial images. |

## 3 Proposed Method

By integrating HRNet-W48 for high-resolution semantic segmentation and Swin Transformer for global feature learning, our Hybrid HRNet-Swin Transformer system improves aerial picture segmentation and scene categorisation. The HRNet-W48-based segmentation technique, which maintains fine details through multi-resolution feature fusion, comes after pre-processing (greyscale conversion, histogram equalisation, and normalisation). Swin Transformer's hierarchical self-attention effectively captures long-range dependencies and refines and classifies the segmentation output. It was trained on the UC Merced and WHU-RS19 datasets using the SGD optimiser with cross-entropy + dice loss. The overall architecture for the scene classification of the proposed model is shown in Fig. 2.

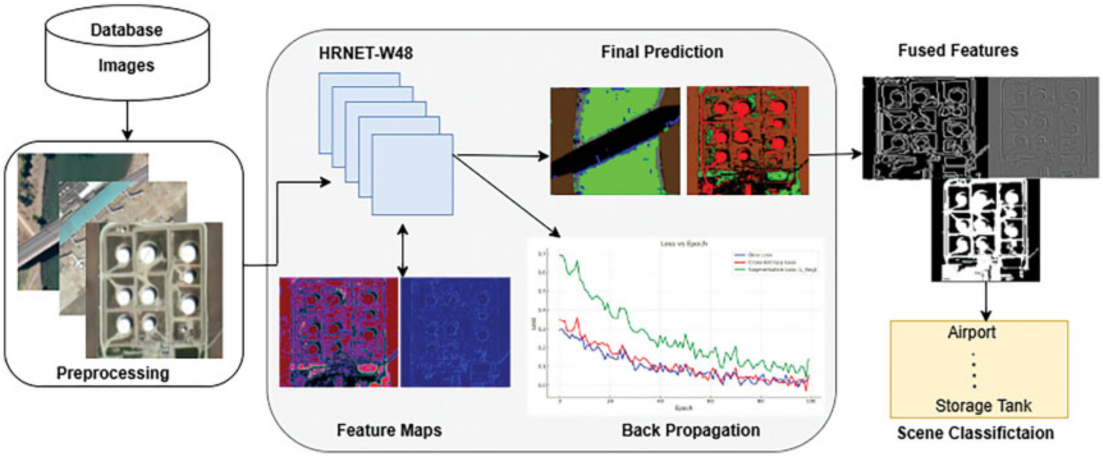


**Figure 2:** An overall description of the anticipated architecture for scene classification

Fig. 2 illustrates the overall architecture of the proposed model, showcasing the data flow from input image pre-processing to training and testing using HRNet-W48.

### *3.1 Pre-Processing*

The pre-processing methods [16] used on the aerial images include normalization, Gaussian blurring, histogram equalization, and grayscale conversion. By eliminating color information while maintaining structural elements, grayscale conversion turns a colored image [17] into a single-channel intensity image. Usually, a weighted sum of the RGB channels is used for the transformation:

$$I\left(x,y\right)=0.299R\left(x,y\right)+0.587G\left(x,y\right)+0.114B(x,y) \tag{1}$$

where $I(x,y)$ is the greyscale strength at pixel $(x,y)$. $R$, $G$, and $B$ are the red, green, and blue channel intensities. The following provides the transformation function:

$$S_k=\frac{(L-1)}{(Wt-Hg)}\sum_{j=0}^{k}h\left(j\right)=0 \tag{2}$$

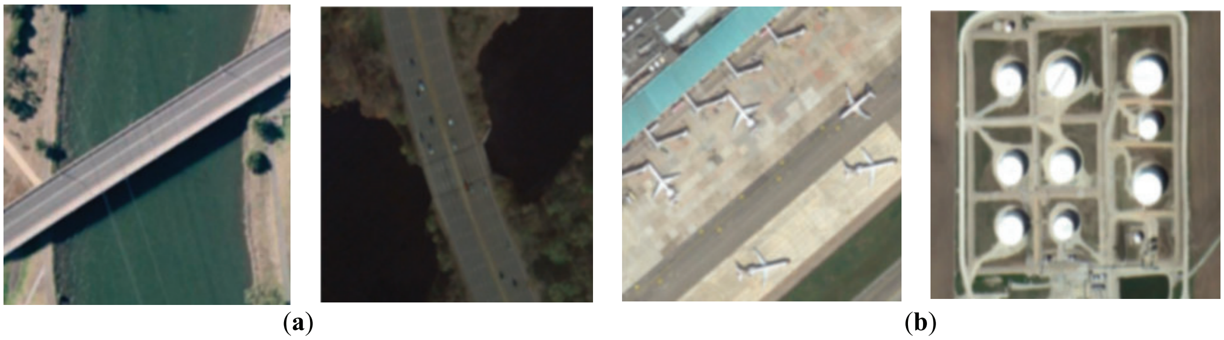
where $S_k$ is the new intensity value, $L$ is the signifying overall intensity levels, $Wt$, $Hg$ are the total pixel count in an image. The intensity of the histogram count is denoted by $h(j)$.

$$G\left(x,y\right)=\frac{1}{2\pi\sigma^2}\cdot e^{-\frac{x^2+y^2}{2\sigma^2}} \tag{3}$$

$G\left(x,y\right)$ is the Gaussian kernel, and $\sigma^2$ is the deviation from the mean of the Gaussian distribution [18], adjusting the blurring effect. Lastly, normalization prevents biases caused by different lighting situations by scaling pixel values between 0 and 1, guaranteeing consistency across images and improving machine learning model performance using Eq. (4).

$$I_{norm}\left(x,y\right)=\frac{I\left(x,y\right)-I_{min}}{I_{max}-I_{min}} \tag{4}$$

Together, these pre-processing techniques enhance the images' quality, contrast, and clarity, making them better suited for additional analysis (as shown in Fig. 3).



(**a**) (**b**)

**Figure 3:** Results following pre-processing (**a**) WHU-RS19 (**b**) UC-Merced

### *3.2 Semantic Segmentation Using HRNet-W48*

In remote sensing, semantic segmentation [19] entails classifying every pixel in an aerial image into a distinct category, allowing for accurate object, position, and boundary recognition. Designed for semantic segmentation, the High-Resolution Network (HRNet) [20] is a deep learning model that preserves high-resolution feature mappings across the network. HRNet maintains resolution representations, which makes it perfect for aerial image segmentation where minute details are important (such as roads, buildings, and vegetation), in contrast to typical CNNs [21] that gradually downsample spatial data. With 48 channels per convolutional block. Fine details are lost as a result of traditional CNN-based models' inability to keep high-resolution features. This restriction is overcome by HRNet-W48 (48-channel variation) [22]. HRNet-W48 preserves parallel high-to-low resolution subnetworks and fuses features across them, in contrast to conventional architectures that downsample spatial resolution. Feature maps at scale $s$ are represented by, which stands for ever-lower resolutions. The representation at each stage is updated by the given Eq. (5).

$$F_{t+1}^{s} = \left(T_{t}^{s}; W^{s}\right) \tag{5}$$

where $F^s$ represents the transformation function and $W^s$ are learnable weights at level $s$. It uses upsampling and downsampling to constantly exchange data across various resolutions:

$$F^{s} = \sum_{i=1}^{s} \alpha_{s,i} U\left(F^{i}\right) \tag{6}$$

where $s$ is the overall resolution scale, $\alpha_{s,i}$ is the learnable fusion parameters, and $U\left(F^{i}\right)$ denotes the upsampled features from scale $i$. This guarantees that the final prediction incorporates both semantic context (low-resolution branch) and low-level details (high-resolution branch). Multiple residual blocks make up HRNet-W48, which improves spatial representations. This is the final high-resolution feature map:

$$F_{H} = \sum_{i=1}^{4} \beta_{s} U\left(F^{s}\right) \tag{7}$$

where appropriate scale fusion is ensured by learnable weight parameters $\beta_s$. Trainable weights known as learnable fusion parameters ($\alpha_s$ $\beta_s$) dynamically regulate the contribution of various resolution scales during feature fusion. An ideal multi-scale representation is ensured by these parameters, which adaptively balance low-resolution semantic context with high-resolution spatial information. Softmax activation and a 1 × 1 convolution are used to create the segmentation output (shown in Fig. 4):

$$Y = \sigma\left(W_{s}{}^{*} F_{f} + b\right) \tag{8}$$

where $W_s$ is the segmentation weights, $b$ is the bias term, $\sigma$ (softmax) converts logits to class probabilities. To optimize segmentation accuracy, the following parameters were fine-tuned (Table 2).

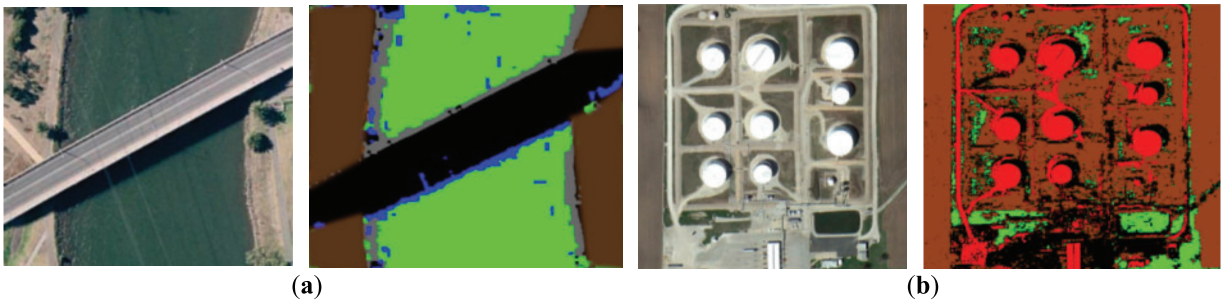


**Figure 4:** Proposed semantic segmented images using HRNet-W48 Model. (**a**) WHU-RS19; (**b**) UC merced datasets

**Table 2:** Optimized parameter and their value for HRNet-W48

| Parameter | Value |
|---|---|
| Optimizer | SGD (Stochastic Gradient Descent) |
| Learning rate | 0.01, with step decay (factor 0.1 every 30 epochs) |
| Momentum | 0.9 |
| Weight decay | 1e−4 |
| Batch size | 16 |

### 3.3 Feature Extraction and Fusion

Our hybrid HRNet-Swin Transformer uses block-based processing [23] with overlapping patches (10%–20% overlap) to effectively handle high-resolution imagery while maintaining contextual continuity across boundaries. By preventing the fragmentation of spatial characteristics close to tile boundaries, this overlap preserves semantic coherence. We use confidence-weighted fusion of overlapping regions after inference to prevent discontinuities and smooth transitions. Edge coherence and long-range feature integration are further improved by Swin Transformer's global attention and HRNet-W48's multi-scale fusion. When combined, these techniques successfully handle context loss in block-wise processing, which helps explain the robust segmentation and classification results seen on both datasets.

#### *3.3.1 Low-Level Features*

These are crucial for identifying minute features in aerial imagery because they capture fundamental patterns like edges, corners, and textures. These basic features are essential for defining object borders and are extracted with the aid of techniques such as Sobel filtering, Histogram of Orientated Gradients (HOG) [24], and Local Binary Patterns (LBP) [25]. However, deeper feature extraction is required because low-level features by themselves are unable to provide significant context. Low-level feature extraction can be expressed mathematically as follows in (9):

$$F_L = G(I) = I * K \tag{9}$$

when $I$ is the input image, $K$ is a convolutional kernel (Sobel filter), and $G(X)$ is the edge extraction function.

### 3.3.2 Medium Level Features

Convolutional layers and pooling techniques are used to combine low-level characteristics to produce these. By recognising patterns like forms, object borders, and localised textures, mid-level features offer more abstract representations [26]. To guarantee consistency in spatial representations, the HRNet design enables simultaneous learning of these properties at various resolutions. The following describes the change from low-level to mid-level features:

$$F_{Me} = U(F_L) + D(F_H) \tag{10}$$

### 3.3.3 High-Level Features

We have extracted these features during semantic segmentation using deep learning architectures, HRNet, as discussed above. In the provided architecture, the feature fusion process uses identity links, upsampling, and downsampling to combine multi-scale information [26]. Using bilinear interpolation and 1 × 1 convolution, upsampling (U) raises low or mid-resolution features to a higher resolution. Downsampling (D) uses a 3 × 3 strided convolution to compress high-resolution feature maps to smaller resolutions. Fig. 5 shows in detail the fusion of all three-level features.

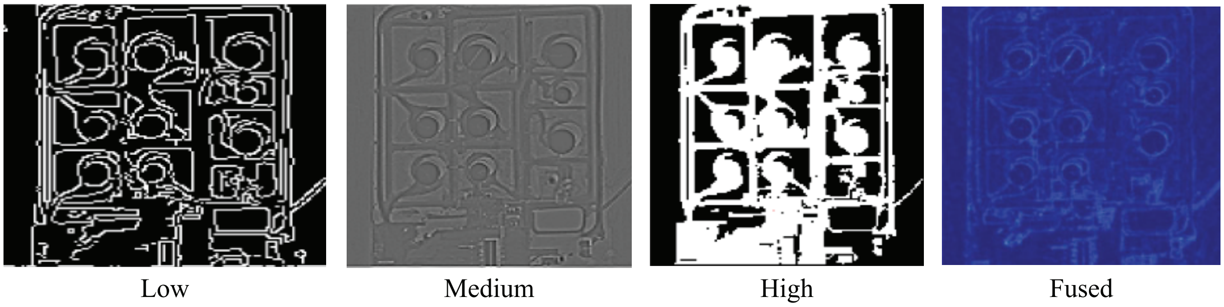


**Figure 5:** Fused feature map representing low, medium, and high-level features, highlighting prominent structures and details for enhanced visualization

Once features from all levels are extracted and fused, the final feature representation is obtained as:

$$F_{final} = \gamma_1 U\left(F_L\right) + \gamma_2\left(F_{Me}\right) + \gamma_3 D(F_H) \tag{11}$$

where the trainable weights $\gamma_1$, $\gamma_2$, $\gamma_3$ are what decide the contribution of each feature resolution. To perform scene classification, this final feature representation is subsequently fed into the Swin Transformer.

## 3.4 Scene Classification: Swin Transformer

Through the implementation of a hierarchical self-attention mechanism that successfully captures long-range relationships, the Swin Transformer [27] plays a critical role in feature extraction in our suggested model. Tokenisation is the first step in the procedure, which divides each aerial image into 4 × 4 pixel patches that do not overlap. These patches are subsequently used as input tokens for the Swin Transformer blocks after being linearly projected into feature embeddings.

$$T = WF_M + b \tag{12}$$

where $W$ and $b$ are learnable weights used to convert features into tokens that are compatible with Swin. The Swin Transformer's shifted window self-attention mechanism, which simulates long-range relationships across the input characteristics, then processes these tokens. By calculating self-attention within small, non-overlapping windows rather than the complete image, Swin Transformer improves efficiency in contrast to conventional self-attention models [28]. At each step, nearby windows are moved to capture global context, allowing information to be shared across geographical boundaries. The final segmentation or classification output, $Y$ is calculated as follows:

$$Y = SwinTransformer(T) \tag{13}$$

whereby, rather than depending on multi-resolution CNN-based feature extraction, swin learns attention-based feature connections. The hyperparameter configuration is shown in detail in Table 3.

**Table 3:** Hyperparameter configuration for SwinTransformer

| **Optimiser** | **SGD** |
|---|---|
| Number of attention heads | 6 |
| Embedding dimension | 96 |
| Number of transformer layers | 4 |
| Weight decay | 1e−4 |
| Dropout rate | 0.1 |
| Optimizer | SGD |
| Learning rate | 0.0005 |
| Batch size | 32 |
| Window size | 7 × 7 |

Our hybrid framework easily integrates Swin Transformer with HRNet-W48 to further improve segmentation performance. Swin Transformer's multi-scale feature extraction capabilities improve scene classification, while HRNet-W48 is in charge of preserving high-resolution representations across the network.

### *3.5 Loss Function for Aerial Segmentation and Scene Classification*

For aerial segmentation, Dice Loss and Cross-Entropy Loss work best together because they strike a balance between region-level segmentation quality and pixel-wise precision. Although Cross-Entropy Loss guarantees accurate pixel classification, it suffers from class imbalance, which is typical in aerial images when background areas predominate. Conversely, Dice Loss improves the segmentation of tiny objects and fine details like roads and buildings by concentrating on locations where the anticipated and ground-truth masks overlap. The model is perfect for high-resolution aerial image segmentation because it combines the advantages of increased boundary preservation, enhanced small object detection, and diminished class imbalance effects. Dice Loss, Cross-Entropy Loss (Fig. 6), and Classification Loss (Fig. 7) are combined to enhance HRNet-W48 for segmentation and Swin Transformer for classification:

$$\mathcal{L}_D = 1 - \frac{2|P \cap G|}{|P| + |G|} \tag{14}$$

where $P$ is the predicted segmentation and $G$ is the ground truth.

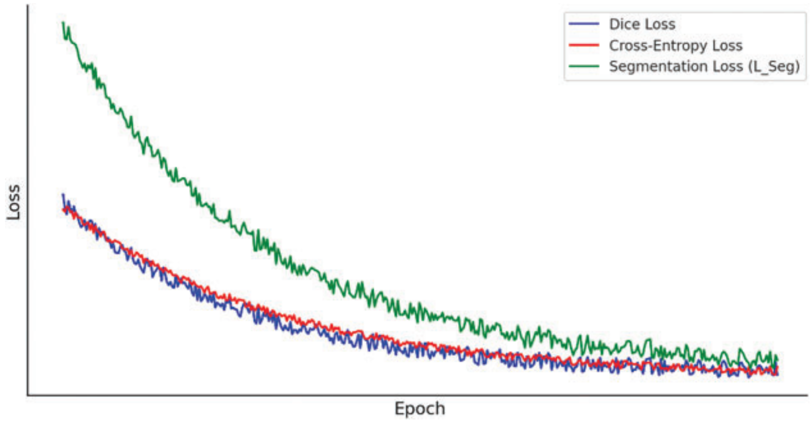
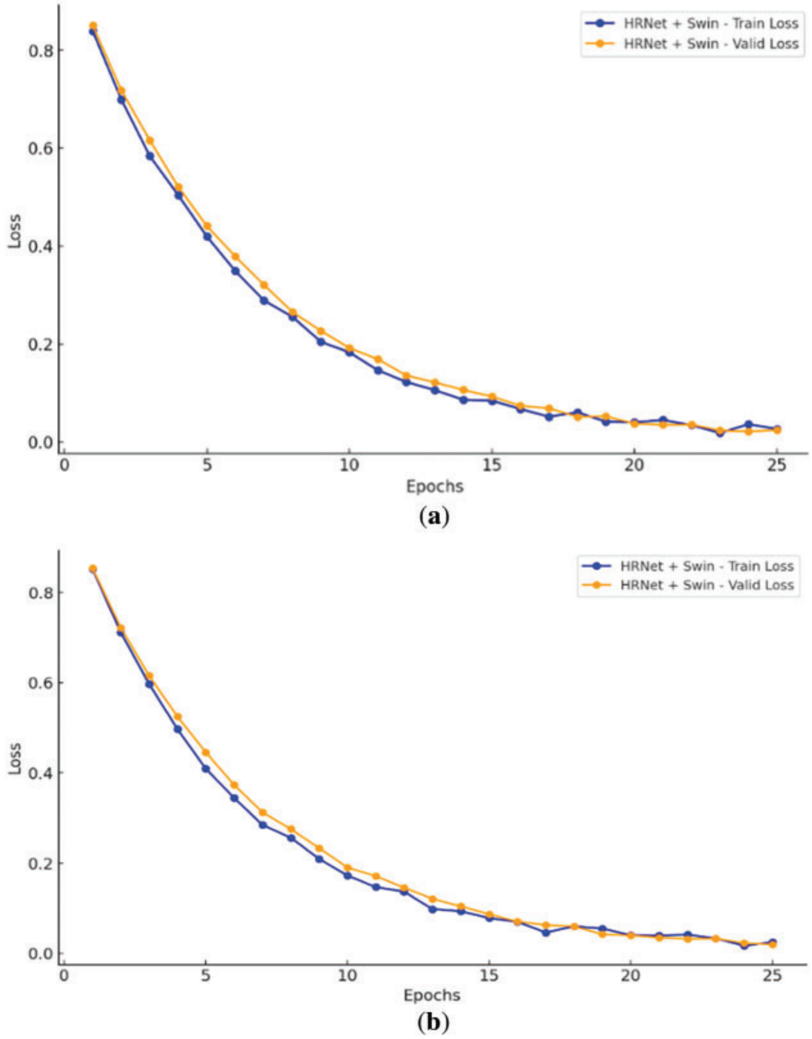


**Figure 6:** Epochs vs. loss curves during training for segmentation



(**a**)

(**b**)

**Figure 7:** Epochs vs. loss curves using hybrid HR-Net + Swin transformer. (**a**) WHU-RS19; (**b**) UC merced datasets

## 4 Experimental Settings

Using the PyTorch deep learning framework, the experiments are conducted on a high-performance computing system with an NVIDIA RTX 2080 Ti GPU (8 GB RAM). Python is used as the main programming language for the implementation, which is done in PyCharm and Jupyter. To improve model convergence, the stochastic gradient descent (SGD) optimizer is used for optimization. Through an in-depth empirical study, the learning rate (LR) and batch size (BS) are adjusted. To maximize segmentation, the training procedure makes use of a cost function for segmentation and scene classification.

### *4.1 Dataset Overview*

The paper evaluates the effectiveness of the proposed system using two datasets for scene classification. Because of their broad use and significance as benchmarks in remote sensing research, we chose the UC Merced and WHU-RS19 datasets. They provide high-resolution images with a variety of land-use scenarios and well-annotated classes, which are perfect for evaluating the segmentation and classification abilities of deep learning models, even if we acknowledge that they might not cover every potential object category.

#### *4.1.1 UC Merced*

One popular benchmark dataset for classifying aerial scenes is the UC Merced Land Use Dataset [19]. It has 2100 high-resolution aerial photos total, with 100 images in each of the 21 land-use groups. Each image has a resolution of 256 × 256. It has 21 classes as land-use categories, including Agricultural, Airplane, Baseball Diamond, Beach, Buildings, etc.

#### *4.1.2 WHU-RS19*

A popular remote sensing dataset created especially for aerial scene classification is the WHU-RS19 Dataset [20]. It is useful for testing artificial intelligence and neural network models in remote sensing applications since it offers a well-balanced collection of high-resolution photos with a variety of land-use and land-cover classes. This dataset is the predecessor of the UCMerced dataset and contains 19 classes, including Airport, Beach, Bridge, Playground, Resort, Residential, School, Sparse Residential, and Square, etc.

## 5 Results

A hyperparameter tuning study that shows how learning rate and batch size affect model performance is included in Section 5. The best results were obtained with a learning rate of 0.0005 and a batch size of 32 (Dice: 0.93, IOU: 0.86, Accuracy: 99.05%). Confusion matrices are used to further analyse classification effectiveness, and assessment metrics give precision and recall precedence over accuracy to overcome class imbalance, especially in datasets such as UC Merced and WHU-RS19.

### *5.1 Hyperparameter Tuning & Sensitivity Analysis*

Since learning rate and batch size have a major impact on training stability and convergence, we carried out a hyperparameter tuning study to guarantee optimal model performance. Our findings are compiled in Table 4 below according to important evaluation metrics:

### *5.2 Experiment I: Classification Accuracy*

The confusion matrix of the suggested model (Figs. 8 and 9) was produced for both datasets to further examine classification effectiveness. This allowed for a thorough analysis of misclassifications and inter-class variances.

**Table 4:** Sensitivity analysis of learning rate and batch size on model performance

| Learning rate | Batch size | Dice Score | IOU | Accuracy | Convergence speed |
|---|---|---|---|---|---|
| 0.0001 | 16 | 0.88 | 0.81 | 98.45 | Slow |
| 0.0005 | 16 | 0.91 | 0.84 | 98.85 | Moderate |
| 0.01 | 16 | 0.87 | 0.79 | 98.20 | Fast but unstable |
| 0.0005 | 32 | 0.93 | 0.86 | 99.05 | Optimal |
| 0.0005 | 64 | 0.90 | 0.83 | 98.70 | Stable |

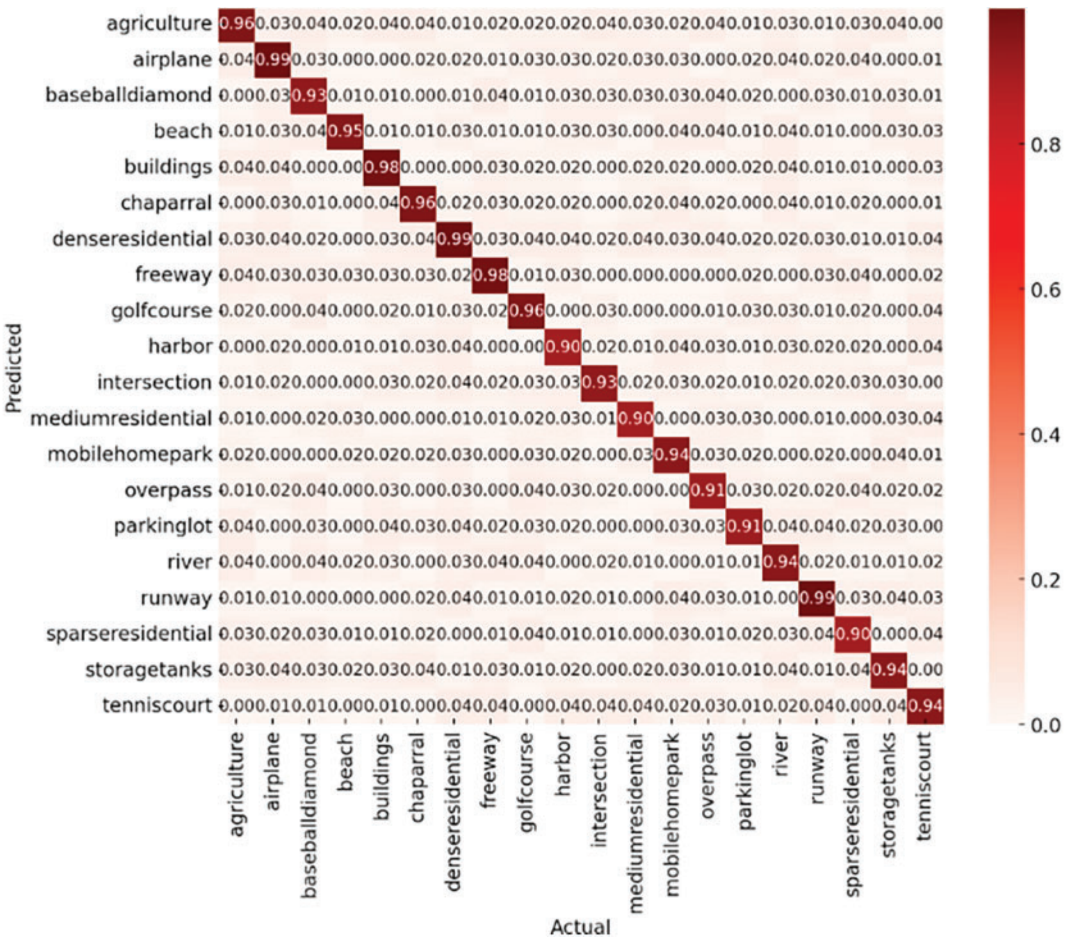
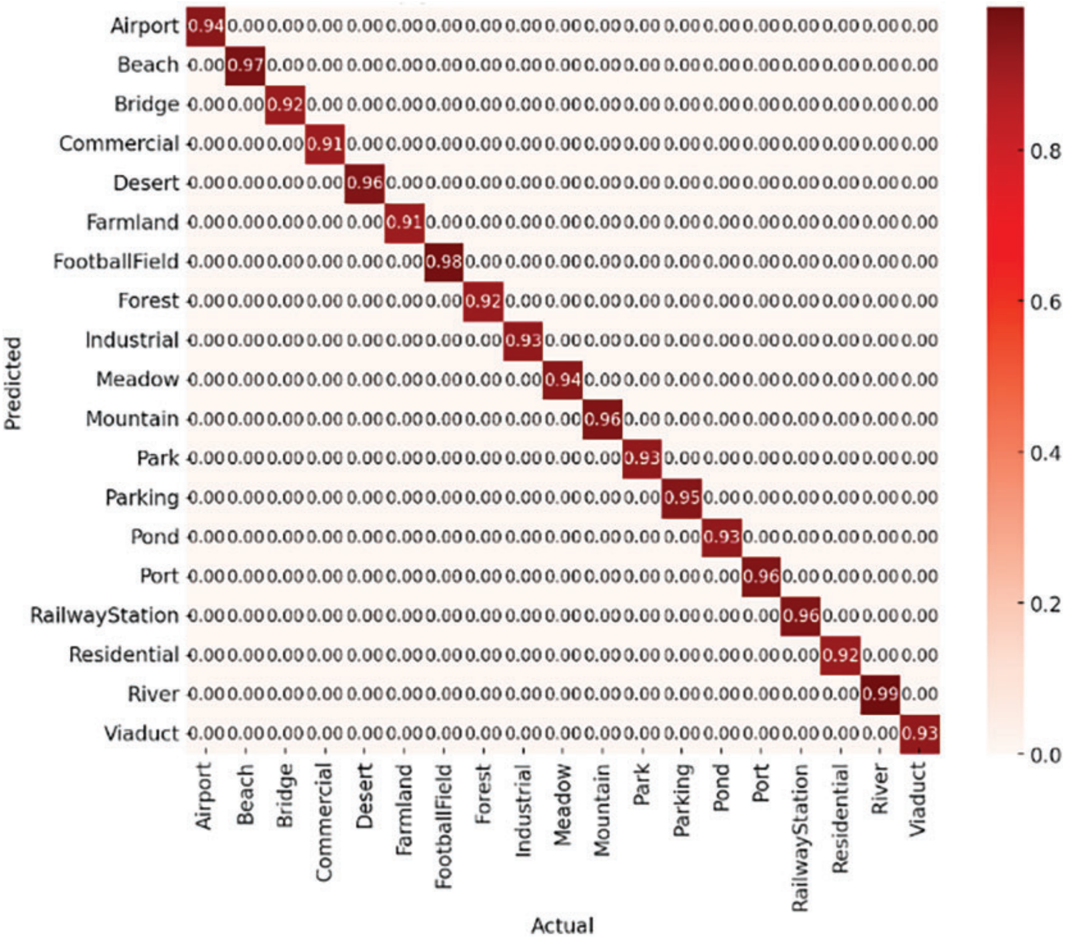


**Figure 8:** Classification accuracy confusion matrix over UC-merced dataset

### *5.3 Experiment II: Evaluation Metrics*

For both datasets—UC Merced (1680 training, 420 test samples) and WHU-RS19 (800 training, 200 test samples), we employed an 80:20 train-test split. We gave precision and recall precedence (shown in Tables 5 and 6) above overall accuracy since class imbalance is a prevalent problem in remote sensing. Accuracy by itself can be deceptive; it frequently conceals inadequate detection of important but under-represented characteristics, such as buildings or roads, while reflecting accurate predictions on dominating classes. While recall makes sure crucial classes aren't overlooked, precision helps reduce false positives. This approach is further reinforced by our use of Dice Loss, which improves boundary precision and small object detection—two critical components of accurate aerial scene classification.



| | Airport | Beach | Bridge | Commercial | Desert | Farmland | FootballField | Forest | Industrial | Meadow | Mountain | Park | Parking | Pond | Port | RailwayStation | Residential | River | Viaduct |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Airport | 0.94 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| Beach | 0.00 | 0.97 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| Bridge | 0.00 | 0.00 | 0.92 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| Commercial | 0.00 | 0.00 | 0.00 | 0.91 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| Desert | 0.00 | 0.00 | 0.00 | 0.00 | 0.96 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| Farmland | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.91 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| FootballField | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.98 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| Forest | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.92 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| Industrial | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.93 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| Meadow | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.94 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| Mountain | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.96 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| Park | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.93 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| Parking | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.95 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| Pond | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.93 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| Port | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.96 | 0.00 | 0.00 | 0.00 | 0.00 |
| RailwayStation | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.96 | 0.00 | 0.00 | 0.00 |
| Residential | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.92 | 0.00 | 0.00 |
| River | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.99 | 0.00 |
| Viaduct | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.93 |

**Figure 9:** Classification accuracy confusion matrix over WHU-RS19 dataset

**Table 5:** Results for the UC-Merced dataset

| Architecture | Traning error | Validation error | Accuracy | Precision | Recall | F1 Score | Time for 1 epoch (min) |
|---|---|---|---|---|---|---|---|
| ResNet 50 | 0.017 | 0.0299 | 0.9982 | 0.9932 | 0.9922 | 0.9928 | 01:98 |
| VIT | 0.0097 | 0.0163 | 0.9929 | 0.9934 | 0.9929 | 0.9928 | 01:45 |
| Swin Transformer | 0.0123 | 0.0340 | 0.9929 | 0.9930 | 0.9929 | 0.9929 | 01:22 |
| Ours | 0.0085 | 0.0080 | 0.9905 | 0.9986 | 0.9985 | 0.9985 | 01:30 |

The Hybrid HRNet-Swin Transformer maintains a reasonable level of computational efficiency while achieving cutting-edge classification and segmentation accuracy. The model's practical implementation in real-world aerial imaging applications is confirmed by the provided FLOPs (floating point operations), inference time, and GPU (Graphics Processing Unit) memory usage (see Tables 7 and 8).

**Table 6:** Results for WHU-RS19 dataset

| Model name | Traning error | Validation error | Accuracy | Precision | Recall | F1 Score | Time for 1 epoch (min) |
|---|---|---|---|---|---|---|---|
| ResNet 50 | 0.2635 | 0.0594 | 0.9802 | 0.9816 | 0.9811 | 0.9808 | 02:00 |
| VIT | 0.002 | 0.0033 | 0.9859 | 0.9854 | 0.9830 | 0.9928 | 01:34 |
| Swin Transformer | 0.0019 | 0.0019 | 0.9899 | 0.9860 | 0.9919 | 0.9929 | 01:10 |
| Ours (HR-Net + Swin) | 0.0015 | 0.0017 | 0.9885 | 0.9950 | 1.0 | 1.0 | 01:05 |

**Table 7:** Computational efficiency of the hybrid HRNet-Swin transformer

| Model component | FLOPs (Giga) | GPU memory usage (GB) |
|---|---|---|
| HRNet-W48 Backbone | 71.5 GFLOPs | 3.8 |
| Swin Transformer | 35.2 GFLOPs | 2.7 |
| Final Classification Head | 5.1 GFLOPs | 0.9 |
| Total FLOPs | 111.8 GFLOPs | 7.4 GB |

**Table 8:** Training and inference time

| Datasets | Training time (Per Epoch) | Inference time (Per Image) |
|---|---|---|
| UC Merced | 1.30 min | 12.4 ms |
| WHU-RS19 | 1.05 min | 10.8 ms |

In future work, we plan to validate this scalability further by integrating larger aerial datasets and deploying the model in 4K environments, where its architecture is well-positioned to preserve both local detail and global context efficiently. Lastly, Tables 9 and 10 contrast the proposed system's accuracy in scene classification with other methods using the specified aerial datasets. The article provides strong empirical evidence that the hybrid HRNet-Swin Transformer architecture outperforms both standalone CNN models like HRNet-W48 and pure transformer-based models such as Vision Transformer (ViT) and Swin Transformer in aerial scene classification and segmentation tasks, as graphically shown in Fig. 10.

**Table 9:** Comparative evaluation of recent techniques on the UCMerced dataset

| Authors | Scene classification accuracy |
|---|---|
| Kim and Chi [28] | 86.79 + 0.33% |
| Gomez and Meoni [29] | 90.71% |
| Ghadi et al. [30] | 98.75% |
| Hao et al. [31] | 98.95% |
| **Ours** | **99.05%** |

**Table 10:** Comparative evaluation of recent techniques on the WHU-SR19 dataset

| Authors | Scene classification accuracy |
|---|---|
| Ahmed et al. [32] | 86.76% |
| Mei et al. [33] | 82.67% |
| Alhichri et al. [34] | 98.60% |
| **Ours** | **98.85%** |

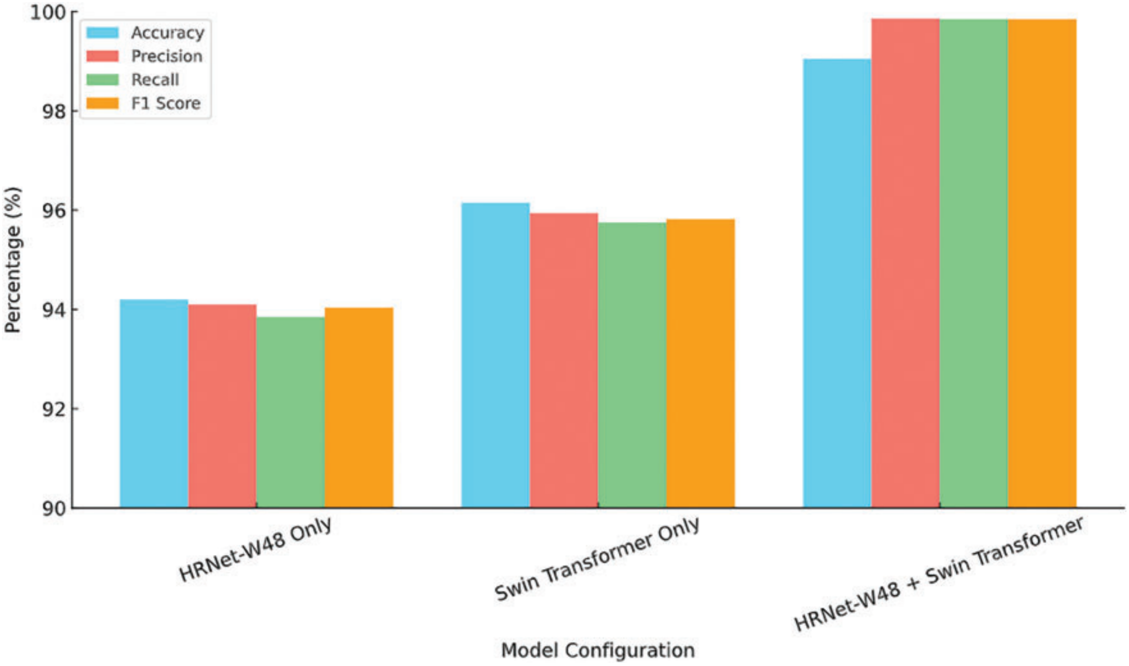


**Figure 10:** Performance comparison of different model configurations

## 6 Ablation Study

Table 11 presents the ablation analysis for the accuracy, precision, recall, and F1-score for each configuration on the UC Merced dataset.

**Table 11:** Ablation study results showing the impact of different components on the model over the UCM dataset

| Model configuration | Accuracy | Precision | Recall | F1 Score |
|---|---|---|---|---|
| HRNet-W48 Only | 94.20 | 94.10 | 93.85 | 94.04 |
| Swin Transformer Only | 96.15 | 95.94 | 95.75 | 95.82 |
| HRNet-W48 + Swin Transformer | 99.05 | 99.86 | 99.85 | 99.85 |

## 7 Conclusion

This study tackled the limitations of CNNs and Vision Transformers in high-resolution remote sensing applications by introducing a Hybrid HRNet-Swin Transformer architecture for aerial imagery segmentation and scene classification. The Swin Transformer component enabled comprehensive global scene understanding, while HRNet-W48 ensured precise segmentation by preserving fine-grained spatial details. A novel multi-resolution feature fusion approach was developed to effectively integrate low-, mid-, and high-level features prior to classification. Evaluated on the UC Merced and WHU-RS19 datasets, our method outperformed both standalone transformer and CNN models. The experimental results validated the effectiveness of hybrid feature learning, achieving state-of-the-art accuracy and robustness in aerial scene classification. While real-time processing at 4K resolution would require further optimization and high-performance hardware (e.g., RTX 4090 or AI accelerators), our architecture supports efficient scaling through techniques such as patch-based tiling, model quantization, and mixed-precision inference. Consequently, achieving real-time or near real-time performance at 4K resolution remains feasible and represents a key direction for future research, contingent upon appropriate hardware advancements and system-level optimizations.

**Acknowledgement:** The authors extend their appreciation to the Deanship of Research and Graduate Studies at King Khalid University for funding this work.

**Funding Statement:** This work was supported by the IITP (Institute of Information & Communications Technology Planning & Evaluation)-ICAN (ICT Challenge and Advanced Network of HRD) (IITP-2025-RS-2022-00156326), 33) grant funded by the Korea government (Ministry of Science and ICT). The authors extend their appreciation to the Deanship of Research and Graduate Studies at King Khalid University for funding this work through the Large Group Project under grant number (RGP.2/568/45). The authors extend their appreciation to the Deanship of Scientific Research at Northern Border University, Arar, Saudi Arabia, for funding this research work through the Project Number "NBU-FFR-2025-231-03".

**Author Contributions:** Study conception and design: Aysha Naseer; data collection: Mohammed Alshehri and Yahya AlQahtani; analysis and interpretation of results: Abdulmonem Alshahrani and Asaad Algarni; draft manuscript preparation: Jeongmin Park. All authors reviewed the results and approved the final version of the manuscript.

**Availability of Data and Materials:** All publicly available datasets are used in the study.

**Ethics Approval:** Not applicable.

**Conflicts of Interest:** The authors declare no conflicts of interest to report regarding the present study.

## References

1. Wu H, Liang C, Liu M, Wen Z. Optimized HRNet for image semantic segmentation. Expert Syst Appl. 2021;174:114532. doi:10.1016/j.eswa.2020.114532.
2. Wang C, Sun W, Fan D, Liu X, Zhang Z. Adaptive feature weighted fusion nested U-Net with discrete wavelet transform for change detection of high-resolution remote sensing images. Remote Sens. 2021;13(24):4971. doi:10.3390/rs13244971.
3. Singh P, Komodakis N. Cloud-Gan: cloud removal for sentinel-2 imagery using a cyclic consistent generative adversarial networks. In: IGARSS 2018—2018 IEEE International Geoscience and Remote Sensing Symposium; 2018 Jul 22–27; Valencia, Spain. p. 1772–5. doi:10.1109/igarss.2018.8519033.
4. Cheng G, Xie X, Han J, Guo L, Xia GS. Remote sensing image scene classification meets deep learning: challenges, methods, benchmarks, and opportunities. IEEE J Sel Top Appl Earth Obs Remote Sens. 2020;13:3735–56. doi:10.1109/jstars.2020.3005403.
5. Gómez-Chova L, Tuia D, Moser G, Camps-Valls G. Multimodal classification of remote sensing images: a review and future directions. Proc IEEE. 2015;103(9):1560–84. doi:10.1109/JPROC.2015.2449668.
6. Pires de Lima R, Marfurt K. Convolutional neural network for remote-sensing scene classification: transfer learning analysis. Remote Sens. 2020;12(1):86. doi:10.3390/rs12010086.

7. Bello I, Zoph B, Le Q, Vaswani A, Shlens J. Attention augmented convolutional networks. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV); 2019 Oct 27–Nov 2; Seoul, Republic of Korea. p. 3285–94. doi:10.1109/iccv.2019.00338.
8. Anil A, Sajith VV, Sowmya V, Sukumar A, Krichen M. Influence of spectral bands on satellite image classification using vision transformers. TechRxiv. 2022. doi:10.36227/techrxiv.20001764.v1.
9. Liu R, Tao F, Liu X, Na J, Leng H, Wu J, et al. RAANet: a residual ASPP with attention framework for semantic segmentation of high-resolution remote sensing images. Remote Sens. 2022;14(13):3109. doi:10.3390/rs14133109.
10. Liu J, Chen K, Xu G, Sun X, Yan M, Diao W, et al. Convolutional neural network-based transfer learning for optical aerial images change detection. IEEE Geosci Remote Sens Lett. 2020;17(1):127–31. doi:10.1109/lgrs.2019.2916601.
11. Mou L, Bruzzone L, Zhu XX. Learning spectral-spatial-temporal features via a recurrent convolutional neural network for change detection in multispectral imagery. IEEE Trans Geosci Remote Sens. 2019;57(2):924–35. doi:10.1109/tgrs.2018.2863224.
12. Fan X, Yan C, Fan J, Wang N. Improved U-Net remote sensing classification algorithm fusing attention and multiscale features. Remote Sens. 2022;14(15):3591. doi:10.3390/rs14153591.
13. Nogueira K, Penatti OAB, dos Santos JA. Towards better exploiting convolutional neural networks for remote sensing scene classification. Pattern Recognit. 2017;61(2):539–56. doi:10.1016/j.patcog.2016.07.001.
14. Zhang J, Zhao H, Li J. TRS: transformers for remote sensing scene classification. Remote Sens. 2021;13(20):4143. doi:10.3390/rs13204143.
15. Xu K, Huang H, Deng P, Li Y. Deep feature aggregation framework driven by graph convolutional network for scene classification in remote sensing. IEEE Trans Neural Netw Learning Syst. 2022;33(10):5751–65. doi:10.1109/tnnls.2021.3071369.
16. Naseer A, Almujally NA, Alotaibi SS, Alazeb A, Park J. Efficient object segmentation and recognition using multi-layer perceptron networks. Comput Mater Contin. 2024;78(1):1381–98. doi:10.32604/cmc.2023.042963.
17. Yuan Y, Huang L, Guo J, Zhang C, Chen X, Wang J. OCNet: object context for semantic segmentation. Int J Comput Vis. 2021;129(8):2375–98. doi:10.1007/s11263-021-01465-9.
18. Cheng Z, Fu D. Remote sensing image segmentation method based on HRNET. In: IGARSS 2020—2020 IEEE International Geoscience and Remote Sensing Symposium; 2020 Sep 26–Oct 2; Waikoloa, HI, USA. p. 6750–3. doi:10.1109/igarss39084.2020.9324289.
19. Li T, Li Q, Zhang T, Yang L. Land use information extraction from remote sensing images based on deep learning. In: 2021 IEEE 5th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC); 2021 Oct 15–17; Xi'an, China. p. 1570–6. doi:10.1109/itnec52019.2021.9587292.
20. Zhang W, Tang P, Zhao L. Remote sensing image scene classification using CNN-CapsNet. Remote Sens. 2019;11(5):494. doi:10.3390/rs11050494.
21. Cui B, Chen X, Lu Y. Semantic segmentation of remote sensing images using transfer learning and deep convolutional neural network with dense connection. IEEE Access. 2020;8:116744–55. doi:10.1109/access.2020.3003914.
22. Ando D, Arai S. Semantic segmentation using HRNet with deform-conv for feature extraction dependent on object shape. In: 2021 3rd International Conference on Cybernetics and Intelligent System (ICORIS); 2021 Oct 25–26; Makasar, Indonesia. doi:10.1109/icoris52787.2021.9649462.
23. Sivasubramanian A, Prashanth V, Hari T, Sowmya V, Gopalakrishnan EA, Ravi V. Transformer-based convolutional neural network approach for remote sensing natural scene classification. Remote Sens Appl Soc Environ. 2024;33(3):101126. doi:10.1016/j.rsase.2023.101126.
24. Vharkte MN, Musande VB. A novel method for retrieval of remote sensing image using wavelet transform and HOG. In: 18th International Conference on Intelligent Systems Design and Applications (ISDA 2018); 2018 Dec 6–8; Vellore, India. p. 540–9. doi:10.1007/978-3-030-16657-1_50.
25. Al Saidi I, Rziza M, Debayle J. Completed homogeneous LBP for remote sensing image classification. Int J Remote Sens. 2023;44(12):3815–36. doi:10.1080/01431161.2023.2227320.
26. Zhang J, Li T, Lu X, Cheng Z. Semantic classification of high-resolution remote-sensing images based on mid-level features. IEEE J Sel Top Appl Earth Obs Remote Sens. 2016;9(6):2343–53. doi:10.1109/jstars.2016.2536943.

27. Xu K, Deng P, Huang H. Vision transformer: an excellent teacher for guiding small networks in remote sensing image scene classification. IEEE Trans Geosci Remote Sens. 2022;60(10):5618715. doi:10.1109/TGRS.2022.3152566.
28. Kim J, Chi M. SAFFNet: self-attention-based feature fusion network for remote sensing few-shot scene classification. Remote Sens. 2021;13(13):2532. doi:10.3390/rs13132532.
29. Gomez P, Meoni G. MSMatch: semisupervised multispectral scene classification with few labels. IEEE J Sel Top Appl Earth Obs Remote Sens. 2021;14:11643–54. doi:10.1109/jstars.2021.3126082.
30. Ghadi YY, Rafique AA, al Shloul T, Alsuhibany SA, Jalal A, Park J. Robust object categorization and scene classification over remote sensing images via features fusion and fully convolutional network. Remote Sens. 2022;14(7):1550. doi:10.3390/rs14071550.
31. Hao S, Wu B, Zhao K, Ye Y, Wang W. Two-stream swin transformer with differentiable sobel operator for remote sensing image classification. Remote Sens. 2022;14(6):1507. doi:10.3390/rs14061507.
32. Ahmed B, Akram T, Naqvi SR, Alsuhaibani A, Khan MA, Kraiem N. XcelNet14: a novel deep learning framework for aerial scene classification. IEEE Access. 2024;12(4):196266–81. doi:10.1109/access.2024.3519341.
33. Mei S, Yan K, Ma M, Chen X, Zhang S, Du Q. Remote sensing scene classification using sparse representation-based framework with deep feature fusion. IEEE J Sel Top Appl Earth Obs Remote Sens. 2021;14:5867–78. doi:10.1109/jstars.2021.3084441.
34. Alhichri H, Alswayed AS, Bazi Y, Ammour N, Alajlan NA. Classification of remote sensing images using EfficientNet-B3 CNN model with attention. IEEE Access. 2021;9:14078–94. doi:10.1109/access.2021.3051085.