A Novel Multi-Scale Feature Fusion with Adaptive Scale-Space Pyramid Network for Aerial Scene Recognition using Remote Sensing Images

Muhammad John Abbas, Student *Member IEEE*, Muhammad Attique Khan, *Member IEEE*, Waqas Ahmed, Ameer Hamza, *Member IEEE*, Nejib Ben Hadj-Alouane, Shrooq Alsenan, M. Turki-Hadj Alouane, Yunyoung Nam, *Member IEEE*

Abstract— Remote Sensing is an area anthropogenic study undertaken worldwide. It has succeeded significantly in important applications such as climate monitoring, disaster prediction and land use planning. However, due to the diversity of scales, intra-class similarities, and complex scenes, the accurate recognition process remains challenging. Transformers' global attention mechanism helps them to overcome the limitations of CNNs' local receptive fields; however, they have drawback of increased computing complexity. To overcome such challenges, this work proposes an Adaptive Scale-Space Pyramid Network (ASSPN) for improved remote sensing image classification. The ASSPN architecture contains a learnable Gaussian pyramid module for multi-scale feature representation, a scale selection attention mechanism for dynamically weighing feature relevance, a cross-feature propagation module for fusion guided by uncertainty, and a complexity-aware adaptive pooling module for preserving semantic discriminative features. Experiments are performed three benchmark datasets such as EuroSAT, NWPU-RESISC-45, and MLRSNet. On these datasets, the ASSPN achieves state-of-the-art results with accuracies of 96.14%, 94.73%, and 95.42%, respectively. The obtained accuracy is outperforming previous CNN and transformer-based systems with significant margins. Furthermore, ASSPN is noise perturbation-resistant and shows generalization capability across a wide range of land-cover categories. Ablation studies established the complementary benefits of the core modules, while LIME-based explainability analysis confirmed the predicative trustworthiness of the model.

***Index Terms*— Remote sensing; Image classification; Adaptive Scale-Space Pyramid Network (ASSPN); Complexity-aware pooling; Multi-scale feature fusion; Explainability; Gaussian pyramid; Attention Mechanisms**

I. INTRODUCTION

Remote sensing can be defined as the method of gathering data and information about an object without making physical contact with it [1]. It involves the use of specific sensors that can detect the wavelengths of emitted and reflected radiation from an object [2]. As remote sensing provides detailed information about local and global regions, its applications include climate monitoring, resource management, disaster prediction, weather forecasting, precision agriculture, forest management, and many others [3-5]. Technological advancements in satellite imagery result in large amounts of RS data, which become a gain and a strain at the same time. It helps improve predictions and observations, but handling this large amount of data is a challenging task [6]. In early times, manual approaches were used to handle and analyze RS data, which required a lot of time and human labor. Despite these efforts, high error rates impaired the reliability of this approach, and with the surge in RS data, this option is no longer available [7]. With the rise of Industry 5.0, the use of AI systems has become highly popular, and researchers have started to explore their effectiveness in remote sensing image classification [8].

Remote sensing images are challenging to classify due to irregular regions, multiple scene associations, high inter-similarity, and complex scenic backgrounds [9, 10]. Therefore, the traditional Machine learning algorithms like Support Vector Machine [11], Random Forest [12], Decision trees [13, 14] and K-Nearest Neighbors [15, 16] are not suitable for this purpose due to their reliance on hand-crafted features. This drawback of ML algorithms limits their generalization and restricts their ability to adapt to changing environmental

Funding: This work was supported by Princess Nourah bint Abdulrahman University Researchers Supporting Project number (PNURSP2025R506), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia in the part by National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2023-00218176) and the Soonchunhyang University Research Fund in the part of Deanship of Research and Graduate Studies at King Khalid University for funding this work through small group research under grant number RGP1/47/46.

Corresponding author e-mail: attique.khan@ieee.org, ynam@sch.ac.kr).

Muhammad John Abbas and Muhammad Attique Khan are with Department of AI, Prince Mohammad bin Fahd University, Al-Khobar, KSA. (johnabbas@ieee.org; attique.khan@ieee.org)

Waqas Ahmed is with department of computer science, HITEC University, Taxila, Pakistan (waqas.ahmed@hitecuni.edu.pk)

Ameer Hamza is with Centre of Real Time Computer Systems, Kaunas University of Technology, Lithuania (ameerhamza@ieee.org).

Nejib Ben Hadj-Alouane is with Electrical and Computer Engineering Department, American University in Dubai, Dubai P. O. Box 28282, United Arab Emirates (Email: nalouane@aud.edu)

Shrooq Alsenan is with Information Systems Department, College of Computer and Information Sciences, Princess Nourah bint Abdulrahman University, P.O. Box 84428, Riyadh 11671, Saudi Arabia. (shaalsenan@pnu.edu.sa).

M. Turki-Hadj Alouane is with College of Computer Science, King Khalid University, Abha 62529, Saudi Arabia (malouane@kku.edu.sa)

Yunyoung Nam is with Department of ICR Convergence, Soonchunhyang University, South Korea. (ynam@sch.ac.kr)

conditions and data variations [17, 18]. These challenges were addressed by Deep Learning (DL), a branch of ML that supports automatic feature extraction to handle high-dimensional data such as images, audio, videos, etc. A Convolutional Neural Network (CNN) is a fundamental deep learning algorithm that shows promising results in image classification tasks; however, a traditional CNN is not suitable for RS image classification due to its inability to capture global spatial relations, which are highly important in RS images [19]. Moreover, the diverse scales and high variations in RS data can pose a challenge for traditional CNNs due to their fixed-scale feature extraction [20, 21]. A few sample RS images can be shown in Figure 1. In this figure, it is observed that the each image include different objects and patterns. Also, due to some images contains low resolution make the classification process using traditional techniques complex and harder [22].

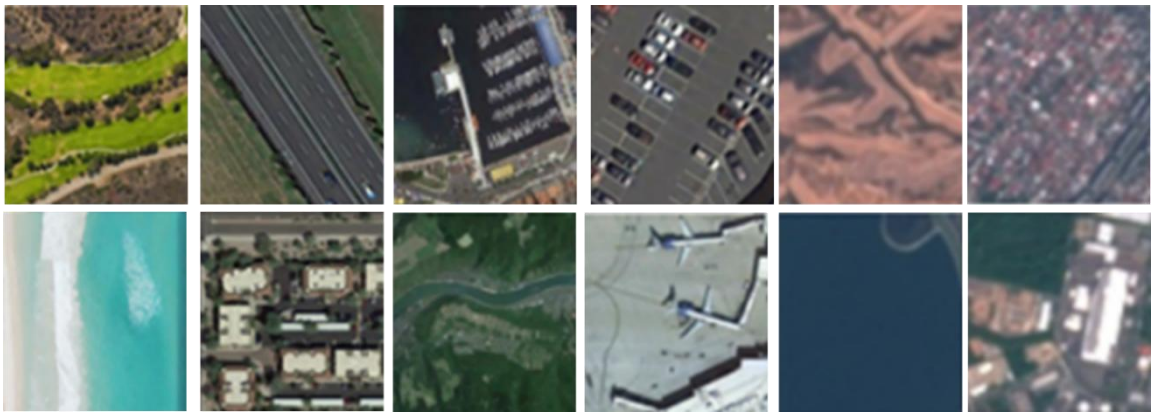


Figure 1: Sample Remote Sensing Images

To resolve these issues, various deep learning techniques have been introduced [23-25]. For instance, Alsubaei et al. [20] presented a novel block scrambling-based encryption with a privacy-preserving optimal DL technique (BSBE-PPODLC) for RS image classification. The presented technique incorporates a two-stage process, which includes RS image encryption via BSBE and classification through DenseNet feature extraction. XGBoost is used as a classifier, and the artificial Gorilla Troops Optimizer (AGTO) is utilized for hyperparameter optimization. This model is applied to the UCM dataset and produces a peak signal-to-noise ratio equal to 55.22dB demonstrating high quality of encryption and an overall accuracy equal to 99.06% showing good quality of classification performance. However, because it is specifically developed on aerial images, its applications are limited. Firat et al. [26] proposed a new three-dimensional residual spatial spectral convolutional network (3D-RSSCN) aimed at hyperspectral image classification. The approach performs principal component analysis (PCA) to perform spectral band reduction of HSRI, which are then segmented into 3D non-overlapping patches and fed into the network. ResNet18 processes the 3D patches to retrieve spatial-spectral features while maintaining a beneficial gradient flow. The proposed model is evaluated on three datasets, and the results show 3D-RSSCN achieved classification accuracy scores of 99.93% on Indian Pines, 99.99% on Pavia University, and 100% on the Salinas dataset. However, the added computational complexity of the model may limit its use in practice. Esmaeili et al. [27] developed a ResMorCNN model for improving HSRI classification accuracy, leveraging a 3D Convolutional Neural Network that included, as morphological features, spatial-spectral features extracted by skip connections. The approach is to combine the 3DCNN with a spatial-spectral morphology box (SSMB). The SSMB consists of a parallel network of four morphological operators with convolutional layers. Residual connections are used to inject morphological features into 3DCNN, thereby enhancing feature extraction and improving overall feature classification. The presented model is evaluated on four datasets, achieving an overall classification accuracy of 97.81% on the Indian Pines dataset, 99.33% on the Pavia University dataset, 98.67% on the Houston University dataset, and 99.71% on the Salinas dataset. However, the problem of increased computational complexity remains the same. To improve the classification accuracy of RS images with small training samples, Zhang et al. [28] presented a novel hybrid classical-quantum transfer learning CNN model that combines a 4-qubit tensor quantum circuit with a pre-trained ResNet34, utilizing quantum encoding and variational quantum circuits. The presented model is evaluated on two datasets, and experimental results reveal that the model achieves classification accuracies of 95.81% (90/10) and 96.62% (80/20) on EuroSAT, and 97.33% (20/80) and 98.82% (50/50) on the AID dataset, respectively. However, in addition to the high computational cost, the limited number of samples and the less diverse dataset question the generalization of the model.

Sitaula et al. [29] presented a novel deep learning technique that incorporates an enhanced attention module (EAM) to improve the classification of high-resolution remote sensing images. For this purpose, features are extracted from the ResNet-50 architecture, and EAM is applied with an improved convolutional block attention module (ICBAM) to extract multi-scale information. An atrous spatial pyramid pooling (ASPP) and global average pooling (GAP) are used to fuse the features, which are then used for classification. Three datasets are used to evaluate the model, and the results indicate that the presented model outperforms most state-of-the-art (SOTA)

architectures, achieving classification accuracies of 95.39%, 93.04%, and 98.61% on the AID, NWPU, and UCM datasets, respectively, with a low standard deviation of 0.001. However, this approach cannot effectively handle spatial similarities and also requires substantial computational resources. In [30], the authors presented a deep learning approach to detect large-scale forest disturbances by combining DL time series classification with prior knowledge constraints. The method involves a two-stage process, in which the first stage consists of preprocessing data using cloud masking and calculating the NBR index. This preprocessed data then enters the first stage of the model, where a self-attention module classifies the Landsat data to detect disturbance patches. In the second stage, a skip disturbance recovery index using prior knowledge is used to pinpoint disturbance years. Experimental findings reveal that the presented model outperforms LandTrendr and Global Forest Change, achieving an overall accuracy of 87.8% with lower emission rates (10.0% to 67.4%). However, the model struggles to detect low-magnitude changes and periods due to the limited training data. Pushpalatha et al. [31] presented a CNN-based deep learning technique for LULC classification and change detection. For this purpose, LISS-III satellite image data is used and preprocessed for radiometric, geometric, and atmospheric corrections. This preprocessed data is further processed to form a composite image from relevant bands, which is then used for feature extraction. A CNN model is trained with 60% training data, 20% validation data, and 20% test data. Experimental findings indicate that the presented model achieved classification accuracies of 94.08% and 96.30%, with kappa values of 0.926 and 0.934 for the 2010 and 2020 data, respectively. For land change detection, the study found an increase of 8.34 square kilometers in built-up areas, 2.21 square kilometers in agricultural lands, and a decrease of 1.49 square kilometers in forest cover. However, the accuracy of the model can be further improved by increasing the resolution of the images.

Van et al. [32] investigated the potential of deep learning models to detect natural climate disasters. In this study, two types of disasters —flooding and desertification —are considered. A customized Climate Change dataset is also considered, which consists of 6334 images collected from open-source datasets. Four DL models, namely DenseNet201, VGG16, ResNet50, and a Convolutional Neural Network, are trained on a customized dataset. Results indicate that all the models have potential in image classification and detection; however, DenseNet201 and ResNet50 achieved higher accuracy, reaching 99.37% and 99.21%, respectively. However, the small size and limited diversity of the dataset can be improved in the future for better results. Vaghela et al. [33] analyzed the performance of different versions of YoloV8 for the classification of agricultural lands. The dataset selected for this research contains a diverse range of classes, including forest, highway, Sealake, pasture, river, residential, etc. Three different versions of YoloV8 models are employed, including medium, small, and nano versions, and are trained on the selected dataset. The performance of all versions for different numbers of epochs, momentum, learning rates, optimizers, and weight decays is observed to analyze the impact of hyperparameters. Experimental results indicate the potential of YoloV8 in the image classification task; however, the comparison reveals that the medium version of YoloV8 is the most effective, achieving a classification accuracy of 99% at 50 epochs, while the other two models achieved 98.60% (nano) and 98.50% (small), respectively. Haider et al. [34] assessed the performance of a total of ten pre-trained convolutional neural networks from the deep learning family on three separate datasets: EuroSAT, NWPU, and Earth Hazard. The authors aimed to address the story of the accuracy and computational efficiency trade-off by implementing five separate neural network classifiers across each architecture. All models were examined on the chosen datasets and results show DenseNet201 was again shown to outperform all models in terms of accuracy, with classification accuracies of 97%, 99.40%, and 97.80%, respectively on EuroSAT, NWPU, and Earth Hazard. However, MobileNetv2, while not performing slightly lower accuracy than DenseNet201, Managed to outperform all models, achieving the highest computational efficiency measuring the time taken for each model to predict that stood at 39.943s on EURSAT, 27.482s on NWPU, and 2.8986s on earth hazard. The factor of classifier was an observed influence as well where it seemed that the Wide NN classifier was the better suited for more diverse datasets, whereas the Medium NN classified held more optimization towards speed.

In summary, the methods discussed above have emphasized pre-trained models, focusing on which one provides the best computational efficiency, accuracy, and precision rate. However, these methods have not built on the most recent approaches and models such as ViT architectures and fused models at the network level, so they do not tackle challenges associated with transferability and generalizability of the model. The prior challenges can be associated with limited generalization capabilities across datasets, poor performance due to limited data, noise sensitivity, inefficient modeling of multi-scale features, and static extraction of meaningful features. In this work, we propose a novel ASSPN network that incorporates multiple modules for the better feature learning. The modules are a learnable Gaussian Pyramid, scale selection attention module, a cross-feature propagation network, and complexity-aware adaptive pooling for classification. Following are our key contributions in this work.

- We proposed a novel deep learning architecture that integrates a learnable Gaussian pyramid, scale selection attention, cross-feature propagation, and complexity-aware pooling with named ASSPN, effectively addressing challenges of fixed-scale processing, inter-class similarity, and noise sensitivity in remote sensing image classification.
- We employed complexity-aware adaptive pooling mechanism that intelligently balances max-pooled and average-pooled features based on image complexity scores. This module enables the network to adapt to both simple and highly complex scenes flexibly.

- Tree Structured Parzen Estimator optimization is employed for selecting the best hyperparameters for the proposed ASSPN model during the training process.
- A detailed comparison and ablation studies has been conducted to validate the proposed ASSPN model.

II. Proposed Methodology

Remote sensing environmental scenes frequently contain objects that exist in a great diversity of scales, show pronounced inter-class similarity, and include confounding complex backgrounds; thus, employing a fixed scale (or fixed set of scales) for feature extraction will not produce meaningful and robust scene classification performance. To overcome these limitations with static, fixed-scale feature extractors, The ASSPN framework integrates four innovative and deliberately designed, complementary modules identified to address particular deficiencies in existing multi-scale fusion networks. The first module provides a learnable Gaussian Pyramid Module for generating adaptive scale-space representations, allowing the network to model fine-scale, medium-scale and coarse-scale spatial patterns that can frequently occur in aerial imagery, but are underrepresented in conventional CNN feature hierarchies. Following, because redundant or noisy scale contributions might still appear in the reasoning of the network, a Scale Selection Attention Module (SSAM) is implemented, forcing the network to evaluate each scale's relevance for a particular scene, thus allowing the network to concentrate on scale-specific information. However, even scale relevance cannot eliminate ambiguity in uncertain or visually overlapping regions; therefore, a Cross-Propagation Feature Module (CPFM) is added to allow the distinctive contribution of benign or useful features to be propagated across scales while minimizing unreliable features. Finally, since remote sensing scene complexity can vary substantially from a simple homogeneous field to a densely cluttered cityscape, Complexity-Aware Adaptive Pooling Module (CAPM) is incorporated to manage the complexity for balancing the dual pooling descriptor: max-pooled descriptors to represent semantic structure and average pooled to conserve fine-grained details. All of these modules are designed to work together for a coherent design continuum utilizing adaptive scale generation, selective attention, uncertainty-based refinement, and complexity aware pooling features into one system for robust classifying of remote sensing scenes. The detailed proposed ASSPN network is demonstrated in Figure 2.

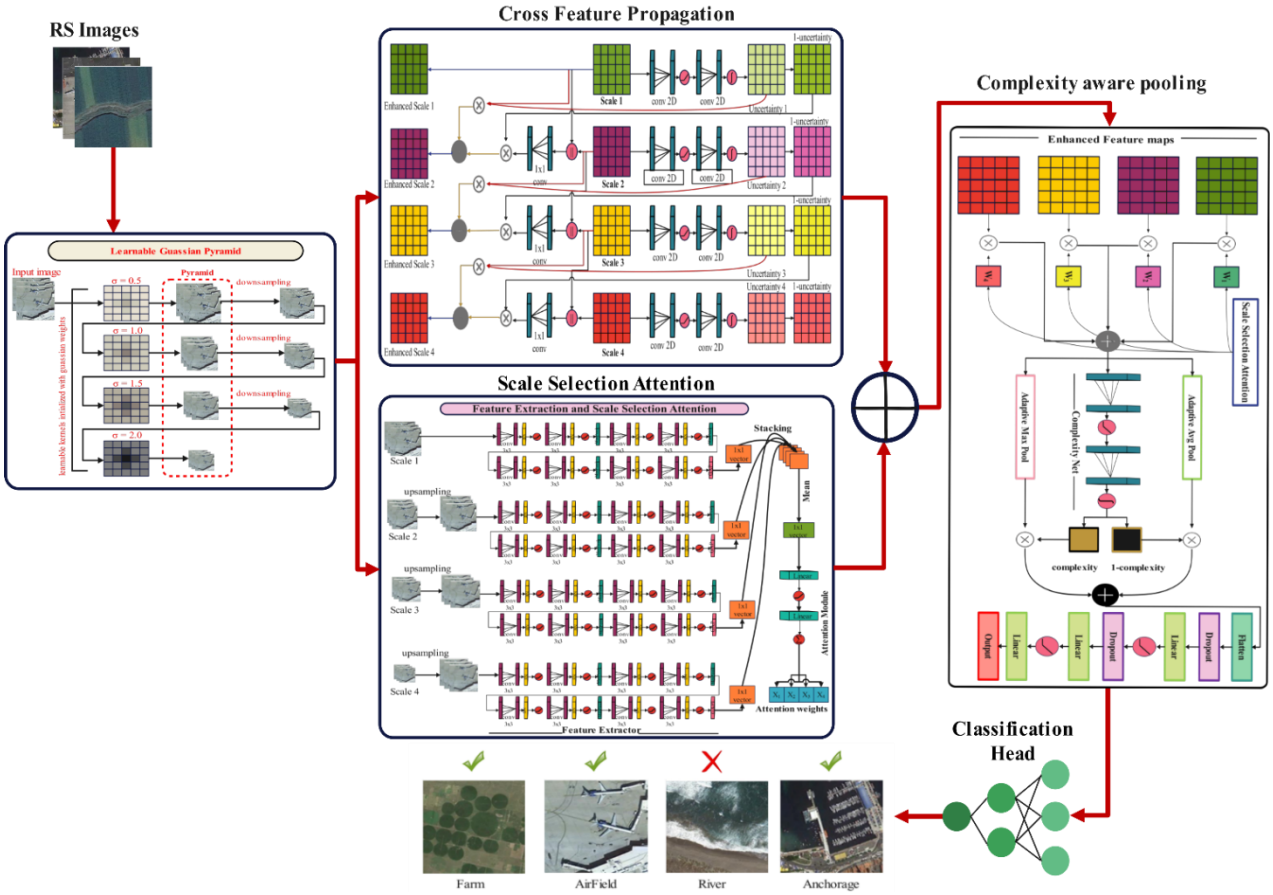


Figure 2: the proposed ASSPN framework that support RS scene classification through adaptive multi-scale generation, scale selection attention, cross-scale feature propagation and complexity-aware pooling.

To enhance architectural consistency and reduce confusion relating to components, we present a preliminary overview of the fundamental modules present in ASSPN. The Gaussian Pyramid Module (GPM) produces adaptive multi-scale

representations for capturing fine, medium, and coarse spatial structures. The Scale Selection Attention Module (SSAM) generates dynamic relevance weights for each individual scale to ensure that the most discriminative resolutions are optimally represented. The Cross-Propagation Feature Module (CPFM) is responsible for uncertainty-guided cross-scale fusion mechanisms that will allow reliable features to be propagated while at the same time suppressing uncertain or ambiguous information. The Complexity-Aware Pooling Module (CAPM) adaptively balances max- and average-pooled descriptors based on estimated scene complexity, allowing for the preservation of global semantic cues as well as fine scale detail. Collectively, these modular components provide a coherent multi-scale learning pipeline that improves feature representation and improves robustness across a variety of remote sensing applications.

A. *Gaussian Pyramid Module*

The proposed model accepts the input of $224 \times 224 \times 3$ and passed it to learnable Gaussian Pyramid Module (GPM). This module generates multiple feature representations of input image, represented as $V = \{V_o, V_1, \ldots V_{s-1}\}$, where V is the version of image and s denoted the number of scales. The Guassian Pyramid Module has four convolutional kernels initialized with Gaussian weights and increasing blur strengths such as $[0.5, 1.0, 1.5, 2]$ to generate slightly blurred, moderately blurred, strongly blurred, and super blurred versions of image. For a scale j, the 2D Gaussian kernel $K_j(x, y)$ is mathematically formulated as follows:

$$K_j(x,y) = e^{-\frac{(x-\lambda)^2+(y-\lambda)^2}{2\sigma_j^2}} \quad (1)$$

Where $\lambda = \frac{K_S-1}{2}$ and K_S denotes kernel size and σ is the blurring strength and defined as $\sigma_j = 0.5 + j \times 0.5$. Each version is down sampled to half of its precursor, resulting in 224×224,112×112 ,56×56, and 28×28 feature representations, which are defined by the following mathematical formulation.

$$\left.\begin{aligned} V_o &= K_o * Z \\ V_1 &= K_1 * \partial(V_o, 0.5) \\ V_j &= K_j * \partial(V_{j-1}, 0.5) \end{aligned}\right\} \quad (2)$$

Where ∂ denoted the downsampling and 0.5 is the rate for each version down samples. The learnable Gaussian module is visually presented in Figure 3.

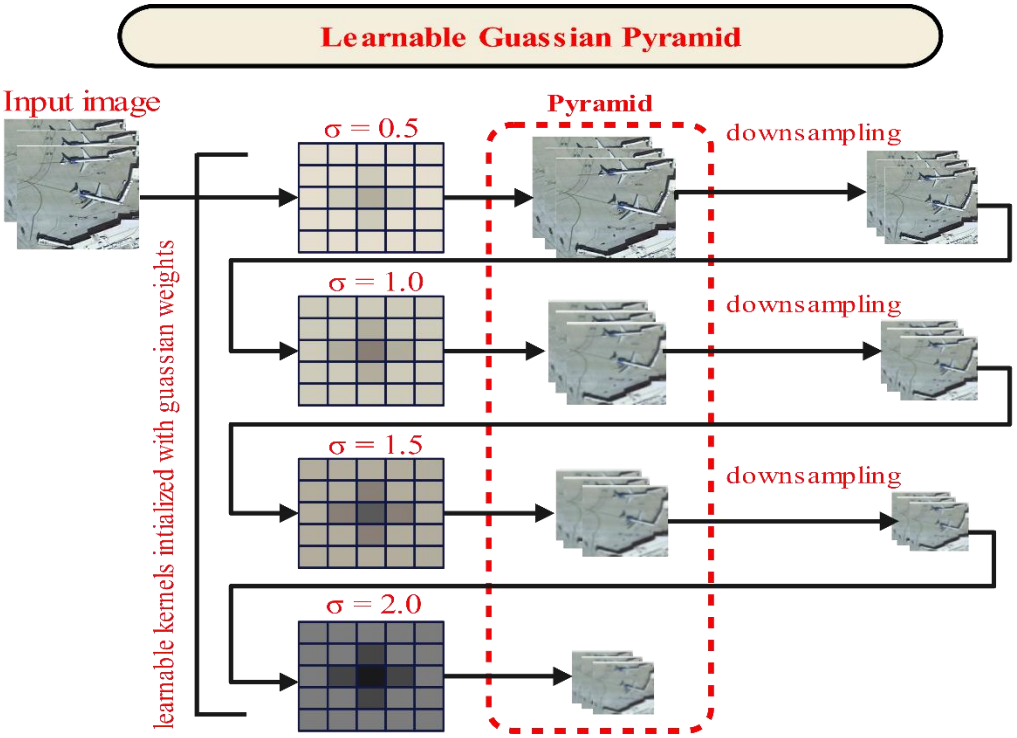


Figure 3: Gaussian Pyramid Module builds an adaptive multi-scale representation of the input image using adjustable Gaussian kernels with increasing standard deviations. Each blurred response will be down-sampled to establish four-resolution scale pyramids to consistently extract fine-to-coarse spatial structures for remote sensing scene classification.

B. *Scale Selection Attention Module (SSAM)*

At this phase, the coarsest three versions are upsampled to match the dimensions of the finest version, ensuring that the feature extraction module has all versions of the same dimensions. The feature extractor is composed of four convolutional blocks, and each block consists of two 3×3 convolutional layers followed by a batch normalization layer. In the first three blocks, a MaxPooling layer is employed to downsample the spatial dimensions. The feature extractor module is mathematically formulated as:

$$F = f_{pool}\left(f_3\left(f_2(f_1(V))\right)\right) \quad (3)$$

$$f_i = \delta\left(\beta\left(C_{3x3}\left(\delta\left(\beta(C_{3x3}(V))\right)\right)\right)\right) \quad for\ i = 1,2,3 \tag{4}$$

$$f_{pool} = M_{pool}(f_i) \tag{5}$$

Where V is the input version, β denotes the batch normalization operation, δ denotes the ReLU activation layer, $C_{3\times3}$ is the convolutional operation, and M_{pool} denotes the Max pooling operation. Each input version is passed to a separate feature extractor; however, all the feature extractors are identical in architecture. All the feature maps are passed through their corresponding global average pooling layers to convert the feature maps into $1D$ dimensions.

$$F_i = F(V_i) \quad for\ i = \{0,1,\dots 3\} \tag{6}$$

$$g_i = GAP(F_i) = \frac{1}{H\times W}\sum F_i(h,w) \tag{7}$$

Where H and W denote the height and width of the feature map, all four vectors are stacked and averaged across scales to form a single 1×1 vector. This single vector then passed through the attention module, which consists of two linear layers, with ReLU applied to the first layer and SoftMax applied to the second layer. This attention module generates four attention weights for corresponding scales. Mathematically, it is formulated as follows:

$$G = \frac{1}{s}\sum_{i=0}^{s-1} g_i \tag{8}$$

$$\alpha = \psi(W''.\delta(W'.G + b') + b'') \tag{9}$$

Where α denotes attention weights, ψ presents SoftMax activation, G represents the combined feature map, W and b denote weights and bias terms of the corresponding convolutional layers. These weights indicate the relative importance of each scale, helping the network focus on the most informative regions. The architecture of the feature extractor and the Scale selection attention module is shown in Figure 4.

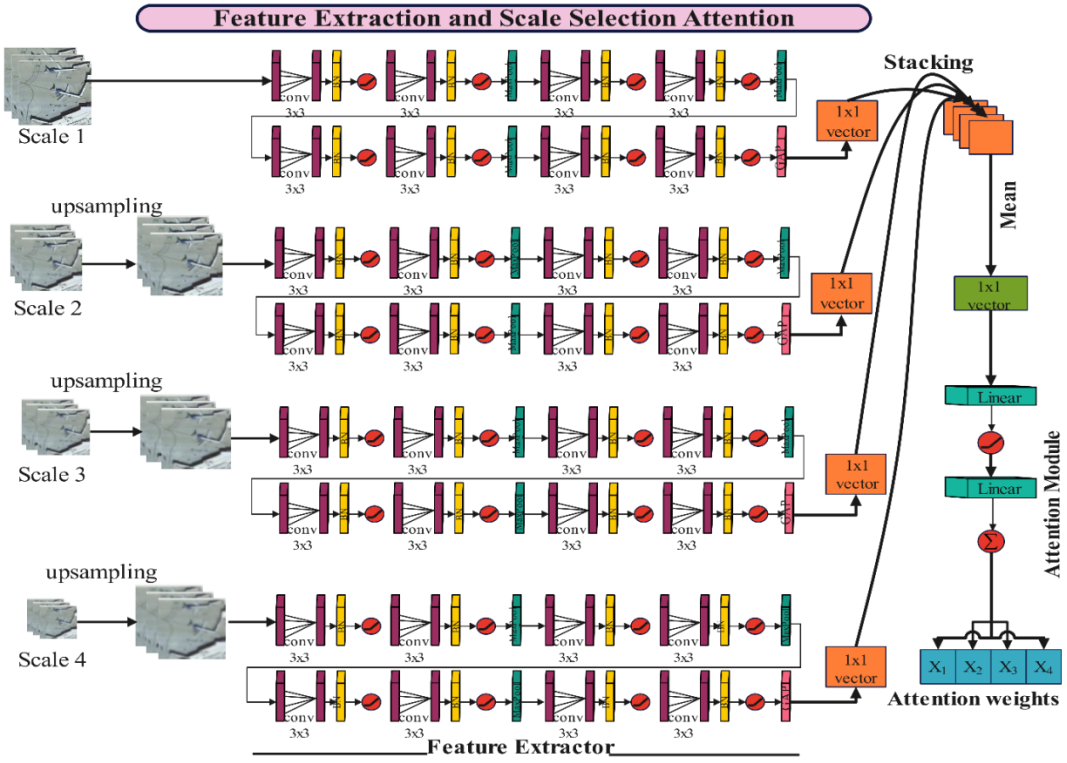


Figure 4: Proposed Feature Extractor and Scale Selection Attention Module (SSAM)

C. Cross Propagation Feature Module (CPFM)

The extracted feature maps are passed to the cross-feature propagation module, in which each scale is processed to generate an enhanced version. Each scale is passed through an uncertainty network to generate a corresponding uncertainty map. This uncertainty map contains a score from {0-1} for each pixel value, where higher scores mean that the model is uncertain about this pixel, while a lower score implies certainty. The uncertainty network is composed of two convolutional layers, where ReLU δ, follows the first layer, and the second layer is followed by sigmoid activation μ. Mathematically defined as:

$$U_i = \mu\left(C_{1\times1}\left(\delta(C_{1\times1}(F_i))\right)\right) \tag{10}$$

Where U_i denoted the uncertainty map of i^{th} scale. After creating uncertainty maps, the original scales are concatenated with each other according to these maps to generate enhanced versions. For this purpose, the finest version remains the same, while scale two is concatenated with scale one and passes through a 1×1 convolutional layer to create a fused representation. This fused feature map is multiplied by the 1-uncertainty map, while the original finer scale is multiplied by the uncertainty map. The outputs of these operations are then added together to generate enhanced scale 2. Mathematical expression for this process is defined as follows:

$$F_{i\_enh} = [\{C_{1\times1}(F_i \parallel F_{i-1})\} \otimes 1 - U_i] + [F_{i-1} \otimes U_i] \quad (11)$$

Where F_{i_enh} denoted enhanced version of i^{th} scale and $\otimes$ denoted element-wise multiplication. This enhanced version contains features of the original finer scale when the uncertainty score is high and prioritizes features of the fused scale when the uncertainty score is low. In this way, it preserved the original scale features while also adding the fused scale features. The same mechanism is applied to the other two scales. In this last of module, we have four enhanced scales, which used for further processing. The Cross Propagation Feature Module (CPFM) is shown in Figure 5.

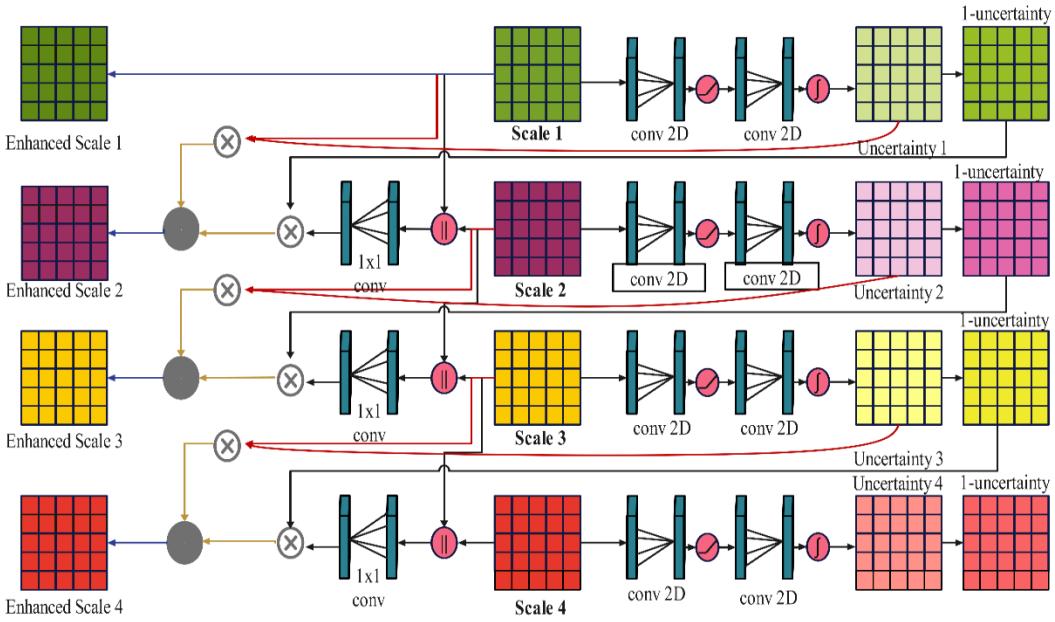


Figure 5: Proposed Cross Feature Propagation Module (CPFM) for cross-feature propagation

D. Complexity Aware Pooling Module and Classification Head

These enhanced scales are then multiplied with their corresponding attention weights generated by the scale selection attention module to create a single feature representation which can be defined as:

$$F = \sum_{i=0}^{s-1}(\alpha_i \times F_{i\_enh}) \quad (12)$$

A complexity-aware adaptive pooling is then applied to this feature map to balance the features according to their complexity level. For this purpose, the feature is passed through a complexity net, an average pooling layer, and a max pooling layer simultaneously. The complexity net, which consists of two convolutional layers followed by ReLU and sigmoid activation, generates a complexity score for each image. This complexity score is multiplied by the Maximum pooled features, while the average pooled features are multiplied by 1 minus the complexity score. The outputs of both operations are added to create a single pooled feature map. Mathematically, this process can be defined as follows:

$$\mathbb{C}_N = \mu\left(C_{1\times1}\left(\delta\left(C_{3\times3}(F)\right)\right)\right) \quad (13)$$

$$F_{adap} = (\mathbb{C}_N \otimes F_{max}) + \left(1 - \mathbb{C}_N \otimes F_{avg}\right) \quad (14)$$

Here, F_{max} represents Max pooled features and F_{avg} represents average pooled features. It means the model will prioritize max-pooled features when the complexity score for an image is high and average-pooled features when the complexity score is low. As a result, the model will be able to handle a diverse range of data accordingly. This pooled feature map is flattened to a 25088-dimensional vector and fed to the classifier. The classifier consists of a dropout layer with a dropout rate of 0.5, which is followed by a linear layer and a ReLU activation function. Another set of dropout, linear, and ReLU is incorporated after that. Finally, a linear layer is integrated that outputs the logit scores for each class, and the class with the highest score is considered the final prediction. A pictorial representation of this process is shown in Figure 6.

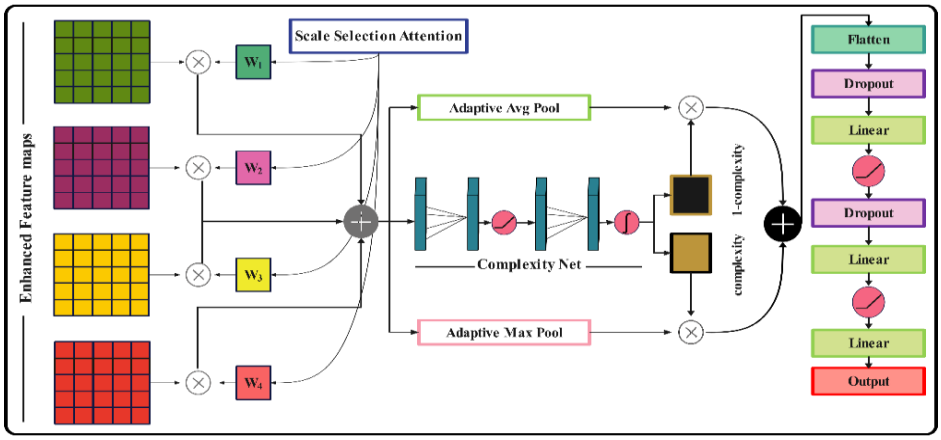


Figure 6: Complexity Aware Adaptive Pooling Module for RS image classification model

E. Hyperparameter Tuning Using Tree Structured Parzen Estimator

The proposed ASSPN model is trained on selected datasets using several hyperparameters. During the training process, the hyperparameters are optimized using the Tree Structured Parzen Estimator [35]. The TPE is a type of Bayesian optimization in which probabilistic model is built to search hyperparameter space. Unlike BO, where an objective function is modeled, TPE models the probability distribution of the objective function. It evaluates the hyperparameters and categorizes them into two separate groups: good-performing hyperparameters and bad-performing hyperparameters. The performance of hyperparameter is evaluated based on quantile γ. The distribution of hyperparameters can be defined as:

$$I(x) = P(x|y) < \gamma \quad \because \quad good\ performing\ hyperparameter \tag{16}$$

$$g(x) = P(x|y) > \gamma \quad \because \quad bad\ performing\ hyperparameter \tag{17}$$

Where $P(x|y)$ presented the probability of hyperparameter x given objective value y. The acquisition function is $\frac{I(x)}{g(x)}$ and the hyperparameters that maximize this acquisition function are likely to perform well. The stopping criteria is set at 50 epochs and 0.15 is the value of γ. The hyperparameter space that is given to optimization technique and the selected hyperparameters are listed in Table 1.

Table 1: Hyperparameter space and selected values for the proposed model

Hyperparameter	Search Space	Selected Value
Number of Epochs	50-150	100
Learning rate	0.0001- 0.001	0.001
Optimizer	AdamW, Adam, SGD	AdamW
Batch size	32, 64, 256, 512	512
Weight decay	0.001-0.01	0.01

III. Experiments and Discussion

A. Datasets

For the experimental process, three different datasets are selected such as EuroSAT, NWPU and MLRSNet.

EuroSAT [36]: EuroSAT is an open-access dataset containing sentinel-2 RGB images. The pixel size of each image is 64×64 and group-sampling distance is 10m and it has 10 different classes with 27000 samples.

NWPU [37]: NWPU is an open-source uni-label dataset where each image has pixel size of 256×256 and dpi of 96×96. This dataset consist of 12 classes and 10500 total samples.

MLRSNet [38]: An open-source multi-label dataset where each image has pixel size of 256×256 and resolution range of $10m - 0.1m$. This dataset is annotated with 60 labels where each image can have labels between 1 to 13. There are 46 number of classes with 109161 samples. A few sample images of each dataset are shown in Figure 7.

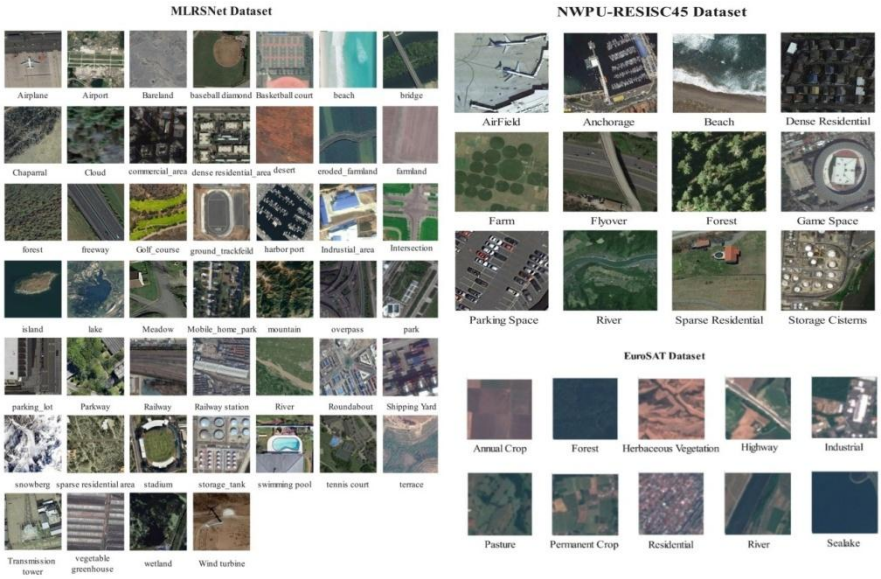


Figure 7: A few samples RS images of each selected dataset

B. Experimental Setup

The experimental process is discussed in this section. The proposed architecture is evaluated using three datasets, as mentioned in Section 3.1. Each dataset has been divided into a 70:30 approach, which means 70% of the images are used for training, and the remaining 30% are utilized for the testing phase. The hyperparameters, such as epochs, learning rate, optimizer, batch size, and weight decay, are selected using the TPE method during the training process. These hyperparameter values are given in Table 2. In testing, several performance measures are computed, including precision, recall, F1-score, support, ROC, PRC, and AUC curves. All the experiments are conducted using Python with Pytorch as a framework on the workstation with 1.5TB RAM, 16 Tesla V100 GPU, each of 32 GB.

1) Results of Proposed ASSPN on MLRSNet

In this section, the proposed model classification results for the MLRSNet dataset are discussed in the form of numerical values and a confusion matrix. Table 2 presents the performance of the proposed ASSPN model on the MLRSNet dataset. The overall test accuracy of the model was 95.42%. The macro and weighted averages for precision, recall, and F1-score were approximately 0.95, indicating balanced performance across all classes without a significant bias towards the majority of categories. The class-wise results demonstrate that the proposed ASSPN model has maintained high degrees of discriminative power across a wide range of categories. Classes that have distinct spatial and semantic features, like a swimming pool, have a precision value of 0.99, a recall rate of 0.99, and an F1-score value of 0.99. The shipping yard class, for example, had a recall score of 0.99, a precision score of 0.99, and an F1 score of 0.99. The vegetable greenhouse class has a precision score of 0.9903, a recall score of 0.98, and an F1 score of 0.98. These types of classes are successfully distinguished with a higher precision score. This performance shows the model is effectively capturing the fine-grained texture patterns and structure features that arise from the natural environment. Similarly, the others, such as the beach, cloud, and wind turbine categories, had an F1 score above a score of 0.98, again demonstrating the model's strength in producing structured landscapes that have distinctive visual cue attributes. Conversely, there were a few categories, such as railway, railway station, and stadium classes, that did not demonstrate as high a recognition performance and inter-class similarity and shared visual cue attributes. The railway class had an F1 score of 0.87, railway station classes a score of 0.83, and stadium classes had a score of 0.91, which are suspectable to misclassified outcomes. The wetland class had an F1 score of 0.93, while the park class had an F1 score of 0.90.F1-score had relatively lower precision and recall, which could be explained by overlapping vegetation patterns and heterogeneous land-cover textures. Overall, the proposed model consistently yielded strong results above 0.90 across these more challenging classes, emphasizing the model's generalization capacity.

Table 2: Classification report of proposed ASSPN architecture for MLRSNet dataset

Class	Precision	Recall	F1-Score	Support	Class	Precision	Recall	F1-Score	Support
airplane	0.96	0.96	0.96	527	**meadow**	0.94	0.93	0.93	749
airport	0.94	0.95	0.94	655	**lake**	0.96	0.97	0.96	729
bareland	0.91	0.92	0.91	472	**island**	0.98	0.98	0.98	748
Baseball diamond	0.97	0.94	0.97	573	**Industrial area**	0.94	0.92	0.93	623
Basketball court	0.94	0.92	0.93	912	**intersection**	0.97	0.96	0.96	746
beach	0.97	0.98	0.98	759	**mountain**	0.93	0.91	0.92	755
bridge	0.94	0.94	0.94	749	**overpass**	0.90	0.90	0.90	765
chaparral	0.97	0.98	0.98	776	**park**	0.93	0.88	0.90	502
cloud	0.97	0.98	0.98	535	**Parking lot**	0.97	0.97	0.97	752
Commercial area	0.90	0.94	0.92	737	**Harbor port**	0.98	0.98	0.98	741
Dense residential area	0.97	0.98	0.97	830	**Ground track field**	0.94	0.92	0.93	753
desert	0.95	0.96	0.95	751	**parkway**	0.93	0.93	0.93	775
Eroded farmland	0.91	0.93	0.92	766	**Mobile home park**	0.98	0.97	0.97	744
farmland	0.94	0.97	0.96	696	**railway**	0.89	0.86	0.87	723
forest	0.96	0.97	0.96	761	**Railway station**	0.86	0.81	0.83	653
freeway	0.94	0.97	0.96	717	**river**	0.96	0.95	0.96	741
Golf course	0.97	0.95	0.96	759	**roundabout**	0.95	0.97	0.96	585
Shipping yard	0.99	0.99	0.99	734	**Storage tank**	0.97	0.96	0.96	759
snowberg	0.94	0.97	0.95	748	**stadium**	0.91	0.91	0.91	743
Sparse Residential area	0.96	0.97	0.96	567	**Swimming pool**	0.99	0.99	0.99	629
Tennis court	0.93	0.95	0.94	763	**terrace**	0.95	0.97	0.96	758
Transmission tower	0.98	0.97	0.97	750	**Vegetable greenhouse**	0.99	0.98	0.98	827
wetland	0.93	0.92	0.93	788	**Wind turbine**	0.99	0.98	0.98	624
Overall performance									
Test Accuracy					0.9542				
Macro Avg	0.95	0.95	0.95		**Weighted Avg**	0.95	0.95	0.95	

Figure 8 illustrates the confusion matrix of the proposed model for this dataset, which can be used to confirm the overall performance, including the macro average test accuracy and weighted average test accuracy. The majority of predicted actions fall within the diagonal range, as expected from the model's ability to discriminate between classes strongly. Only minor off-diagonal elements were noted, illustrating that the degree of confusion about visually related categories was limited. For example, a minor misclassification would occur between a park and a mountain, or a tennis court and a basketball court, which was predictable as they share structural layouts in addition to a shared background context in high-spatial resolution remote sensing imagery. Importantly, the confusion matrix further endorses that these misclassifications occur at low rates, often below 5%.

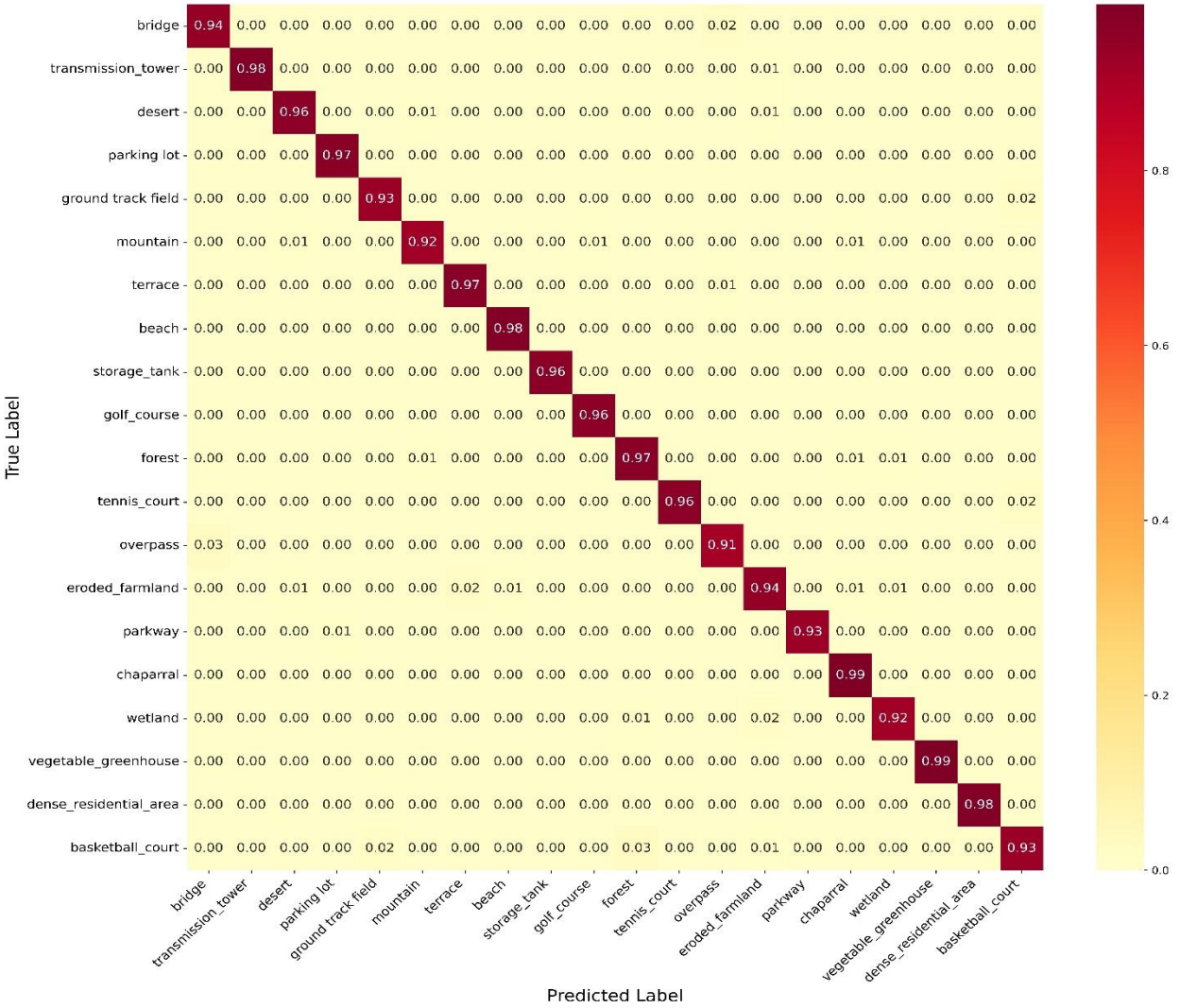


Figure 8: Confusion matrix of ASSPN for MLSRNet dataset

2) Results of Proposed ASSPN on NWPU-RESISC45

In this section, the proposed ASSPN model results are discussed on the NWPU dataset (see Table 4). In Table 3, the overall test accuracy is 94.73% with a macro-averaged F1-score of 0.94 and a weighted F1-score of 0.94, respectively. All of which is indicative of the model's stability across all categories, even with a dataset that is disreputably difficult due to high intra-class variation and inter-class similarity. In this table, class-wise accuracy is also discussed, which shows a good discrimination ability across all categories. As noted, there were distinct categories, including Anchorage with a precision rate of 0.99, Farm with a precision rate of 0.96, and Forest with a precision rate of 0.97, all of which showed high recognition. These performance results show that ASSPN can systematically incorporate both the spectral and structural patterns prevalent in naturally occurring land cover visible over the dataset. Besides these successful categories, other categories with similar strong structural regularities, such as Parking Space (precision rate of 0.92) and Dense Residential (precision rate of 0.94), are also classified with high reliability, showing that ASSPN took advantage of fine-grained spatial features. However, some classes performed relatively lower due to visual similarities with other categories, such as Game Space, which has an F1-score of 0.93, and Storage Cisterns, which has an F1-score of 0.93. Both classes exhibited slightly lower accuracy; their respective dimensionalities were likely rendered similarly to urban categories with shared structures nearby, such as Parking Space and Flyover. Also, Sparse Residential (F1-score) reflected the more difficult nature of distinguishing categories, as it sometimes overlaps with Dense Residential. These challenges are highlighted and present a challenge when it comes to high-resolution imagery, as there can be varying levels of density.

Figure 9 presents the confusion matrix of the proposed model for this dataset, showing that the majority of the predictions cluster along the diagonal, indicating a strong discriminative ability. A few cases have some degree of confusion, such as Beach and Game Space, or River and Sparse Residential classes. These classes have overlapping textural and contextual information. Misclassification occurred among Airfield and Game Space, respectively, which have high recall

rates (0.91 and 0.90). However, misclassification occurred at slightly higher rates, likely reflecting some background structures in the Flyover and Stadium areas as well.

Table 3: Classification report of proposed ASSPN model using NWPU remote sensing dataset

Class	Precision	Recall	F1-Score	Support	Class	Precision	Recall	F1-Score	Support
Airfield	0.91	0.91	0.91	390	**Forest**	0.97	0.97	0.97	207
Anchorage	0.99	0.96	0.98	198	**Game Space**	0.95	0.90	0.93	436
Beach	0.96	0.95	0.95	218	**Parking Space**	0.92	0.98	0.95	207
Dense Residential	0.94	0.98	0.96	205	**River**	0.92	0.93	0.93	232
Farm	0.96	0.97	0.96	425	**Sparse Residential**	0.93	0.94	0.93	202
Flyover	0.95	0.94	0.94	228	**Storage Cisterns**	0.93	0.93	0.93	203
Overall Performance									
Testing Accuracy					0.94				
Macro Avg	**0.94**	**0.95**	**0.94**		**Weighted Avg**	**0.94**	**0.94**	**0.94**	

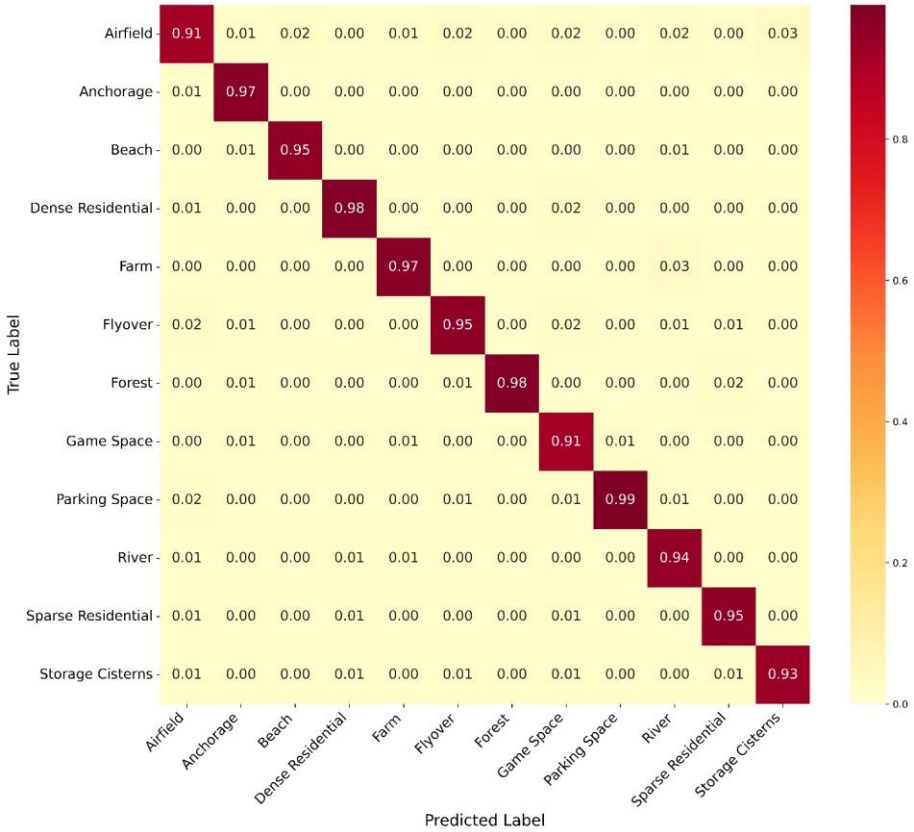


Figure 9: Confusion Matrix of proposed ASSPN model for NWPU dataset

3) Results of Proposed ASSPN on EuroSAT

The classification performance of ASSPN on the EuroSAT dataset further demonstrates its generalization capability, achieving an overall accuracy of 96.14% and a macro-averaged F1-score of 0.96. Assessment of performance on EuroSAT yielded slightly higher results than for the MLRSNet and NWPU datasets. The class-wise evaluation in Table 4 confirms that ASSPN achieves high precision and recall scores for the different scene categories. Classes such as Forest (0.99 precision, 0.98 recall, 0.98 F1-score), Residential (0.98 precision, 0.98 recall, 0.9867 F1-score), and SeaLake (0.97 precision, 0.98 recall, 0.98 F1-score) were classified with higher performance. Similar distinctions can be made regarding these classes based on their unique spectral signatures and spatial structuring, which allowed ASSPN to more successfully separate them from other classes. Industrial

class (0.97 F1-score) and Annual Crop (0.95 F1-score) also showed strong recognition of these land-use patterns.

In contrast, some classes still demonstrated inferior yet competitive performance. Permanent Crop (0.93 F1-score) and Herbaceous Vegetation (0.93 F1-score) were the classes that were most prone to confusion, as demonstrated in the confusion matrix. The spectral signature and similar structures overlap across cropland-related categories and are attributed to contextually based interpretations of the similar land-use categories. The overlap between seasonal, perennial, and herbaceous vegetation is often subtle and context-dependent. River (0.94 F1-score) also demonstrated a small decrement in performance, with a narrow confusion with SeaLake and Pasture. This is partly because, under different conditions, the spectral signatures of waterbodies may sometimes be indistinguishable. However, all classes maintained F1-scores above 0.93.

In Figure 10, the confusion matrix supports the above findings, and the dominance along the diagonal indicates the overall accuracy of the model in assessing land-use. Misclassifications for most accepted cases were generally concentrated among visually similar categories, specifically among either vegetation-related classes, such as Herbaceous Vegetation, Annual Crop, and Permanent Crop, or water-related land-use categories (River and Sea/Lake). Most importantly, the percentage of misclassifications is small overall, remaining essentially below 5%, and this is further indicative of the utility of ASSPN in decreasing ambiguity among some of the more challenging land-use categories.

Table 4: Classification report of proposed ASSPN model for EuroSAT dataset

Class	Precision	Recall	F1-Score	Support	Class	Precision	Recall	F1-Score	Support
Annual Crop	0.95	0.95	0.95	93	**Pasture**	0.93	0.96	0.94	608
Forest	0.99	0.98	0.98	873	**Permanent Crop**	0.94	0.92	0.93	767
Herbaceous Vegetation	0.93	0.93	0.93	903	**Residential**	0.98	0.98	0.98	904
Highway	0.95	0.95	0.9	772	**River**	0.94	0.93	0.94	757
Industrial	0.97	0.97	0.97	730	**Sealake**	0.97	0.98	0.98	848
Accuracy				**0.96**					
Macro Avg	**0.96**	**0.96**	**0.96**		**Weighted Avg**	**0.96**	**0.96**	**0.96**	

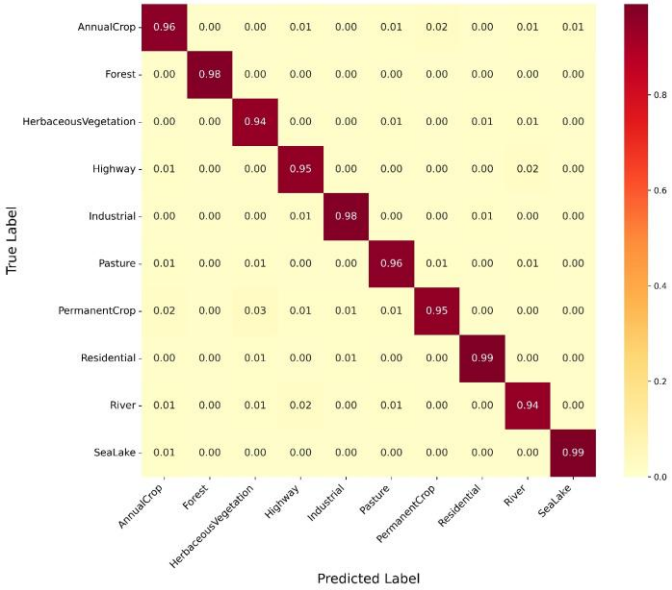


Figure 10: Confusion Matrix of proposed ASSPN for EuroSAT dataset

IV. Discussion

A. Training Performance

This section contains an elaborate discussion of the proposed model. First, we have described the training performance of the proposed model using the training curves. The training curves of the proposed model are illustrated in Figure 11 on the selected datasets. Each dataset has four plots as shown in this figure. The top left plot in each dataset shows validation loss and validation accuracy over epochs; the top right plot shows training accuracy and training loss. The bottom left plot shows change in learning rate over time, and the bottom right plot illustrates validation loss and validation accuracy. After few epochs, the validation accuracy is high and is represented by yellow point. The plots in this figure shows that training accuracy and validation accuracy are consistently increasing with epochs, while training loss and validation loss are also consistently decreasing.

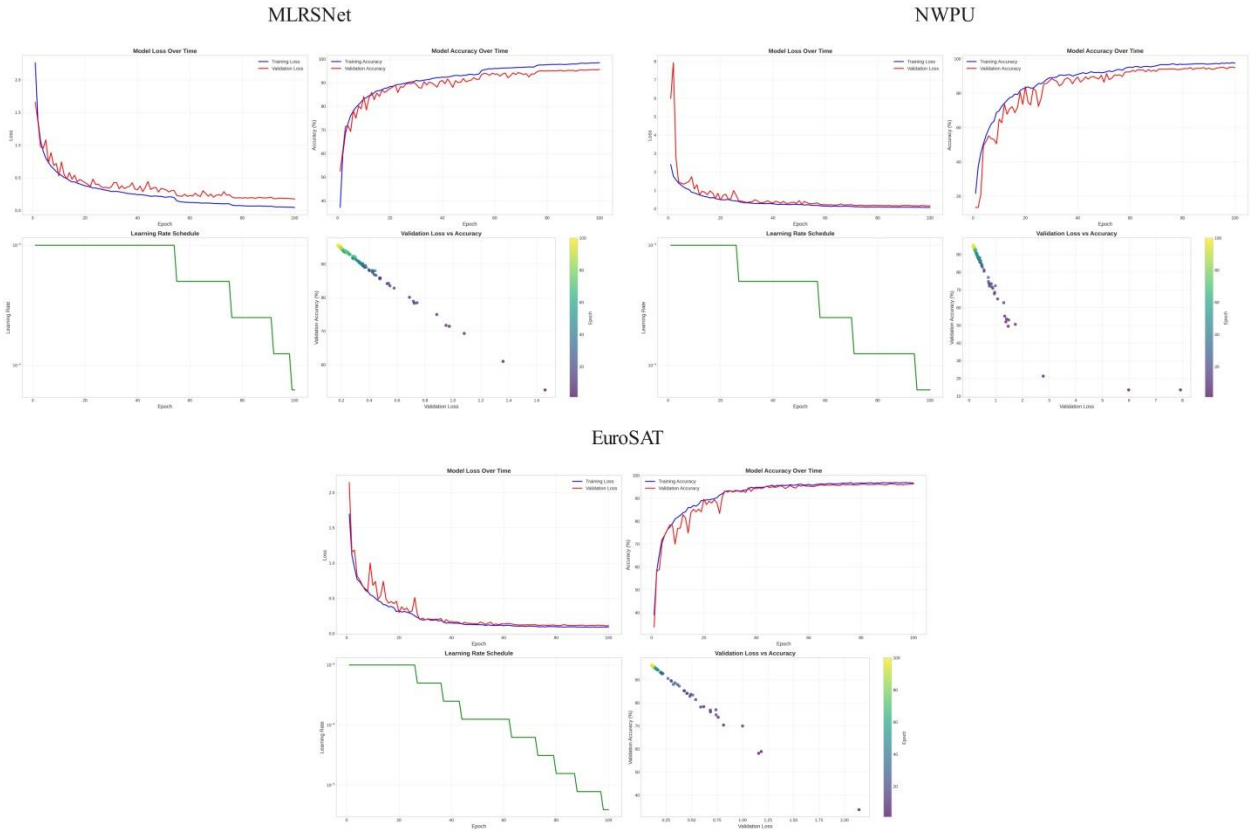


Figure 11: Training curves of the proposed ASSPN model using selected datasets

B. tSNE Visualization

T-Distributed Stochastic Neighbor Embedding (t-SNE) is a dimensionality reduction technique designed to preserve local relationships, reveal cluster structure, and emphasize separation. It plots 5,000 random data samples and separates similar data points into clusters. It helps us to understand the model's classification abilities while preserving the local relationships. Figure 12 illustrates the t-SNE visualization of the proposed model on selected datasets. The tightly packed clusters represent less intra-class variability, while the scattered points of the same class represent high variability. The separate and distinct clusters of each class in the EuroSAT and NWPU datasets suggest strong classification abilities of the model, while the overlapped clusters in MLRSNet point towards the model's struggle in differentiating among classes.

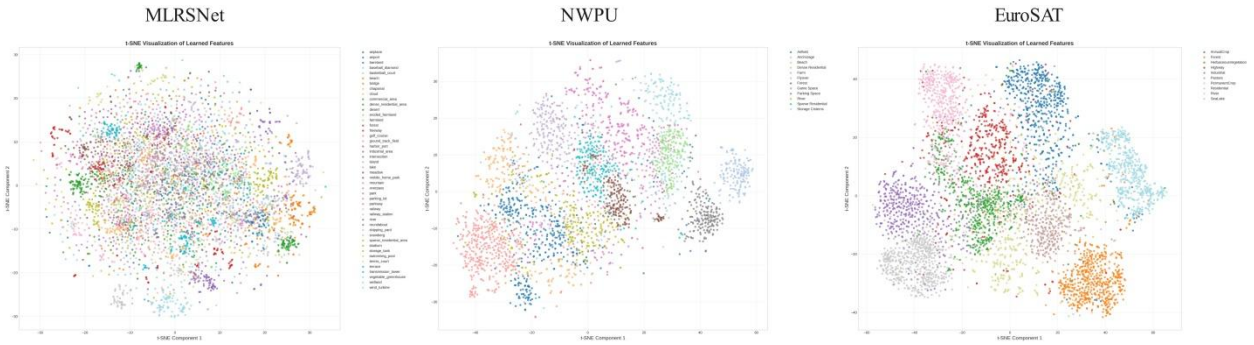


Figure 12: tSNE visualization of ASSPN model for selected datasets

C. PCA Visualization

PCA Visualization helps us understand the separability and structure of learned features. It helps in data visualization by reducing a high-dimensional feature space into 2D while preserving the most critical information. In Figure 13, two plots are shown for each dataset, where the left plot shows the clusters of different points. Here, each point represents a datapoint, and the colour of the point represents a specific class. These clusters represent how well a model can differentiate among different classes. Tightly packed separate clusters show high classification abilities, while overlapped clusters suggest that the model is struggling to distinguish among classes. Points spread out suggest higher variability in the class, while points tightly clustered together indicate consistent features. From a figure we see the model has strong

classification abilities for EuroSAT for distinguishing each class utilizing clusters.However, for NWPU and MLRSNet, overlapping between clusters is evident. PCA can also identify the most critical directions in the data and project it onto two axes, as shown in the right plots of the figure. This plot represents two bar graphs, where each bar graph illustrates the amount of information contained in each principal component. The left bar shows the individual variance captured by each element, and the right plot shows the total variance captured. From the figure, it can be seen that for MLRSNet, PC1 captures only 2.31% of the variance, and PC2 captures 2.05% of the variance, which means that less than 5% of the information is captured. For NWPU, PC1 captures 7.03% of the variance, and PC2 captures 5.74% of the variance, which is better than MLSNet. For EuroSAT, PC1 captures 10.98% of the variance, and PC2 captures 7.41% of the variance, which is the best among the others.

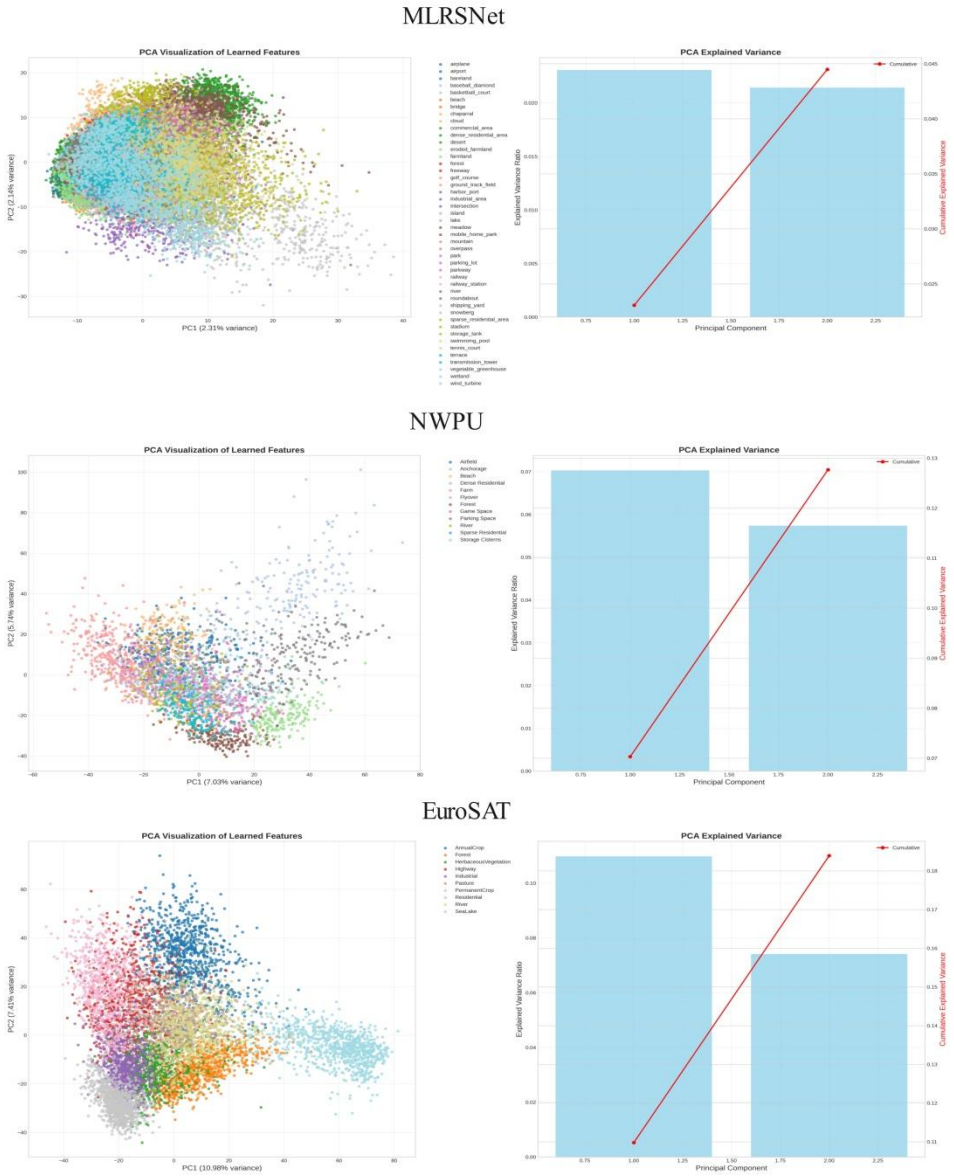


Figure 13: PCA visualization of ASSPN model for selected datasets

D. Ablation Studies

Ablation Study 1: In the first ablation study, the submodules of the proposed model are evaluated using the MLSRNet dataset, as shown in Figure 14. From this figure, the base model + SSAM reached a final accuracy of 89.3% after 100 epochs, indicating that spatial attention may be effective in optimizing feature maps. However, the benefit relative to other configurations was relatively small. The combined base model and CPFM improved to 93.5%, with the quickest gains occurring early in training, between epochs 10 and 30, demonstrating the beneficial capabilities of channel–pixel fusion for capturing discriminative information at the beginning stage of training. Performance with the base model + CPFM began to saturate more quickly than with the other submodules, suggesting it may not optimally capture discriminability in the long term when implemented in isolation. The combination of SSAM + CP improved performance to 91.6%, which demonstrates that spatial

attention combined with context preservation functions additively, leading to distinctive improvements to the model's performance by attending to both local and global dependencies more effectively. In particular, the entire design of SSAM + CPEM + CAPM achieved an optimal performance value of 95.3% accuracy, clearly demonstrating throughout the training epochs between configurations 5 and 7 that the arrangement consistently outperformed all other configurations. These results provide strong evidence indicating that the modules are mutually reinforcing features: SSAM contributes significant value by improving spatial performance; CPFM effectively consolidates the learned signal to achieve improved performance; CPEM preserves contextual signals; and CAPM provides a flexible assignment of adaptive attentional value. Descriptively, the analyses demonstrated that individual modularity provides significant additivity simply due to the individual contributions, yet, the joint implementation of SSAM, CPEM, and CAPM approaches leads to, in general, more consistent variances and more stable representations for a further improvement in convergence properties toward reducing saturation and higher performance accuracy. The primitive numerical responses clearly provided validation of the assumption that the proposed model design executed both learning-efficient means for retrieving the dynamics of the learning outcome iteratively, while providing for the predictive conformance properties of the design, which clearly demonstrated the additive value of the whole model design..

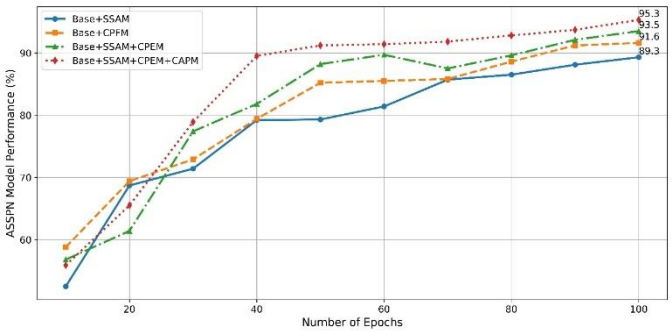


Figure 14: ablation study on submodules of proposed ASSPN model

Ablation Study 2: Figure 15 shows that an ablation study was conducted on the proposed ASSPN model with respect to an array of benchmark deep learning models: InceptionV3, ResNet50, MobileNetV2, VGG19, and DenseNet201 in high-level and low-level education... The visual analysis consists of the layers, parameters, and model size of each of the models. In the area of network depth, VGG19 has 19 layers reflecting the traditional fully sequential structure of its model grades. Similarly, DenseNet201 also had considerable depth with 201 layers. ResNet50 is the best network, with weight layers of 50. InceptionV3 had a weight depth model of 48 layers while the least amount of weight layers was MobileNetV2, but it also had the least amount of models to train in a constrained environment with just 53 weight layers. In terms of depth, we utilized relatively simple methods and utilized a total of 60 to ensure enough structure for the network to learn features while working with limited computation. When we compare depth with the parameter number and other structural configurations, it would be reasonable to conclude that the only layers that were significant were VGG19 with 143 million parameter count, and DenseNet201 with 20 million parameters, which also indicates their memory size. In regard to the workload of computational capabilities of a good model, which also provides more representative power than the ease of use, InceptionV3 had a parament term of 24 million learnable parameters, and ResNet50 had a little more than 25.6 million. The MobileNetV2 was still very efficient, but also had a lesser count of parameters of 3.4 million that were limited. Again, based on the introduction of previous models and an evaluation of only 5.2 million learnable parameters, once more, for the proposed model ASSPN; we also provide a sufficient balance of complexity, primarily to ensure a stronger performance to apply to and from different educational contexts. Then when we get to measure only the overall model weight data again, VGG19 had the greatest model weight of about 549 MB, DenseNet2 had (77 MB), ResNet50 (98 MB), InceptionV3 (92 MB). The MobileNetV2 was by far the compact model taking on only about 14 MB weight. On the contrary, the model identified as the proposed ASSPN was fair at only 21 MB which is significantly less weight with respect to these traditional architectures and it also had all of the mentioned representations properly, while also having far more layers than the model of MobileNetV2 alone.

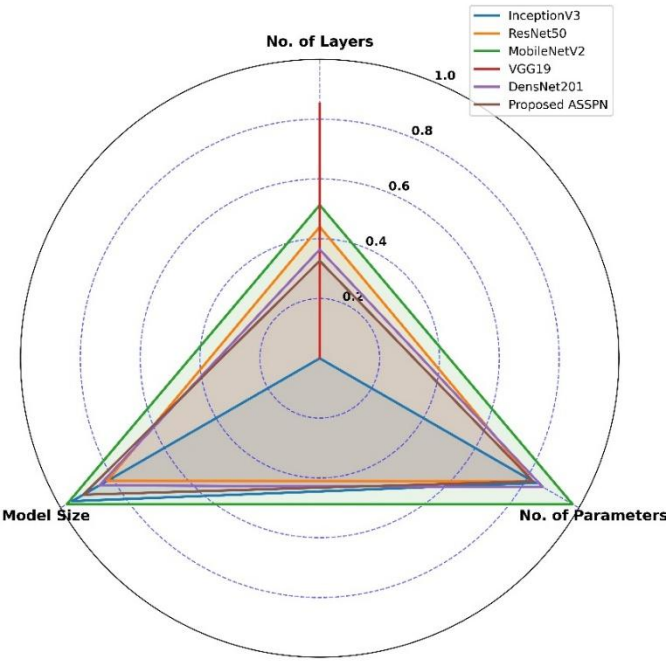


Figure 15: Comparison of proposed ASSPN model with pre-trained models in terms of depth, parameters, and size

Ablation Study 3: In the third ablation study, the ROC of the proposed model and deep learning models on MLRSNet dataset are compared in Figure 16. According to this figure, the proposed ASSSPN model achieved the highest AUC of 0.94, outperforming all baseline models significantly. BEiT-Base and CrossViT-15 achieved AUC values of 0.93, while InceptionV3 attained a value of 0.92. The transformer-based ViT-B/32 and CNN-based ResNet50 achieved AUC values of

0.89 and 0.88, respectively. In contrast, the proposed ASSSPN model's ROC curve steeply rises into the top-left corner, demonstrating stability in terms of sensitivity across multiple specificity thresholds, resulting in consistent true positive rates against false positive rates. This improved performance illustrates that improvements made to the ASSSPN model architecture not only improved class discreteness but also did so in areas of the curves where false positives become extremely costly to the bottom line. Since the difference represents true predictions around a vital decision boundary, the proposed model not only outperforms previous traditional CNN-based models, such as ResNet50 and InceptionV3, but also achieves improvements than state-of-the-art transformer-based models, demonstrating a level of robustness and generalization ability. Finally, the ROC analysis provides evidence that the proposed ASSSPN offers the most balanced trade-off between sensitivity and specificity, making it arguably the most reliable when applied to real-world situations.

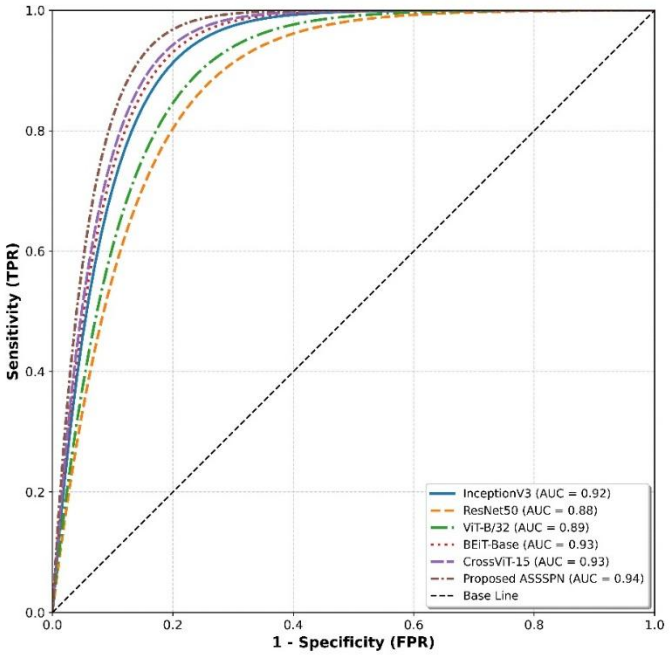


Figure 16: Comparison of ROC's of proposed ASSPN and state-of-the-art models

Ablation Study 4: This experiment presents the sensitivity analysis of the proposed model to variable noise across the three datasets, as illustrated in Figure 17. The analysis involves gradually increasing the level of noise by asking the models to classify the images while varying the amount of noise inclusively from 0.1 to 0.7 in the image, and subsequently measuring the classification performance. For the MLSRNet dataset, the proposed model attained the highest overall accuracy (95%) at the lowest noise level (0.1). The model's performance decreased to 92% at the next noise input level (0.3) and remained stable at 91% for the next two noise levels (0.5 and 0.7 noise inputs). This indicates that the model displayed a very good level of resilience even at high levels of distortion. A similar trend was noted with the NWPU dataset, where, at first, the accuracy dropped from 93% with low input noise to 91% with a moderate level of noise (0.3-0.5) and continued downward to the lowest accuracy of 87% when the maximum noise input (0.7) applied. Likewise, compared to the other datasets, with similar overall resilience, EuroSAT accuracy results ranged from 95%, 92%, 90%, and 88% at 0.1, 0.3, 0.5, and 0.7 noise levels. In general, the heatmap does confirm that the proposed model consistently exhibits similar performance resilience under noisy conditions, as demonstrated by accurate findings that were very close across levels of noise even after perturbation. Additionally, as noted in the results across these different datasets, the results suggest a level of generalizing another strength of the proposed model as it ultimately will need to be functional in real-world operations with input images that will, likely, be cluttered, distorted, or imperfect due to noise..

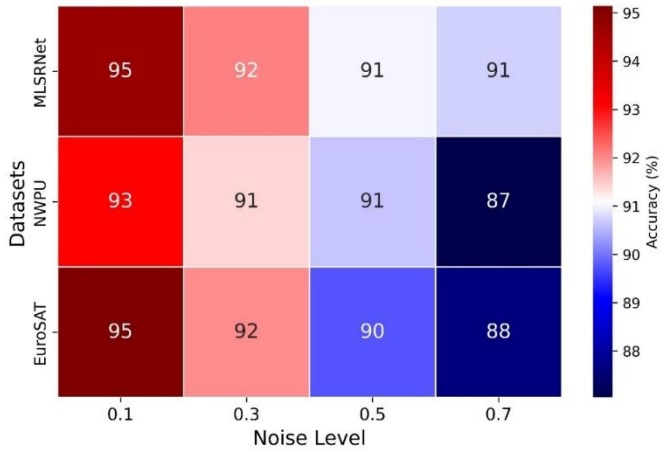


Figure 17: Ablation study conducted on proposed model effectiveness using the noisy dataset.

E. Proposed Model Explainability

The LIME method is used to comprehend the model's predictions and ensure these are interpretable and trustworthy. LIME allows us to understand the behavior of our intricate model, due to its ability to learn a simple, local, and interpretable model. LIME works by making perturbed samples around a selected target datapoint, and observing how perturbing the input datapoint changes the model's output is what points us toward the LIME utility and regions of the image that affect our prediction in positive and negative conditions. Figure 18 shows three samples from each data set, illustrating the model's predicted class, confidence score, and LIME-based explanation. For instance, in the first sample, the model is 100 percent certain that the image represents a tennis court, and the highlighted areas provide an adequate reason to warrant this prediction. The same behavior is observed across all classes, indicating that the proposed model has teeth and learned meaningful and discriminative patterns. LIME explanations not only provide visual justifications but suggest insights into the model's feature learning behavior. The highlighted superpixels align in a strong way with semantically meaningful structures—the presence of spatial boundaries, object contours, texture rich-patches, and high-frequency areas suggest that the Gaussian pyramid and SSAM components of the ASSPN led the model to foster informative

multi-scale cues. Moreover, the reduced emphasis on ambiguous or cluttered regions aligns with the intended functioning of the CPFM, which suppresses uncertain scale contributions. Collectively, these observations confirm that ASSPN attends to scale-consistent, context-aware, and semantically coherent image regions rather than relying on spurious correlations, thereby validating the interpretability and reliability of its predictions.

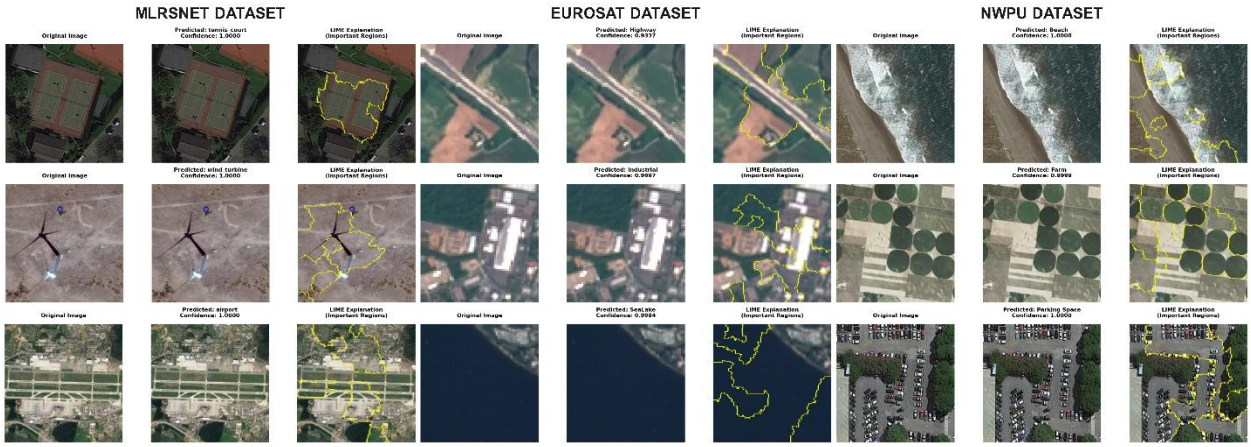


Figure 18: Proposed model explainability results using LIME

F. Comparative Analysis with Pre-trained Models

The proposed model surpasses every previous model build on all pre-trained models on each dataset with a significant gap. Performance of the suggested model compared with several pre-trained models using selected datasets is illustrated in Table 5. In this table, the MLRSNET dataset achieves 95.42% accuracy, exceeding the best baseline ResNet101 model by 8.11%. For the NWPU dataset, the proposed model achieves 94.73% and shows an improvement of 5.68% over the top baseline model ResNet101 (89.05%). The EUROSAT dataset yields the best results, with the proposed model achieving 96.14% accuracy, which is 6.14% better than ResNet101 and displays excellent generalization ability. On all datasets, the deeper networks, ResNet101 and ResNet50, continually outperformed the older pre-trained architectures, such as AlexNet and VGG, which highlights the importance of architectural depth for remote sensing applications.

Table 5: Comparative analysis with pre-trained models

Models	**Accuracy**	**Precision**	**Recall**	**F1-score**
MLRSNET Dataset				
Alexnet[39]	63.46	60.21	62.50	61.95
VGG16[40]	70.13	72.67	71.34	71.56
VGG19[40]	65.37	67.56	65.21	66.45
GoogleNet[41]	82.57	81.89	82.00	82.00
ResNet50[42]	84.21	83.45	81.50	83.45
ResNet101[42]	87.31	89.90	87.95	88.50
Proposed Model	**95.42**	**0.9541**	**0.9542**	**0.9540**
NWPU Dataset				
Alexnet[39]	71.00	72.65	71.89	71.89
VGG16[40]	74.94	75.85	74.90	75.00
VGG19[40]	74.97	75.00	74.85	74.50
GoogleNet[41]	83.54	83.98	83.50	83.00
ResNet50[42]	88.95	89.05	88.75	89.00
ResNet101[42]	89.05	88.00	88.59	89.00
Proposed Model	**94.73%**	**0.9475**	**0.9473**	**0.9472**
EUROSTAT Dataset				
Alexnet[39]	80.00	81.54	81.00	80.98
VGG16[40]	83.00	84.56	85.97	84.00
VGG19[40]	83.59	82.90	83.95	82.00
GoogleNet[41]	87.54	87.00	86.98	87.23
ResNet50[42]	88.95	89.00	89.05	88.89
ResNet101[42]	90.00	90.50	89.90	89.90
Proposed Model	**96.14%**	**0.9614**	**0.9614**	**0.9614**

G. Comparative Analysis with SOTA

Table 6 presents a comparison of the proposed model with state-of-the-art techniques on the selected datasets used in this work. On MLRSNet benchmark dataset, the proposed approach achieves 95.42% accuracy, showcasing an improvement of 3.91% over the previous best method, such as AMEGRF-Net. On NWPU dataset, the proposed model achieves an accuracy of 94.73%, representing an improvement of 1.43% over the previous best result of 93.3%, and establishing a new state-of-the-art on this complex dataset. The EUROSAT dataset demonstrates the most remarkable difference, where the proposed model achieved 96.14% accuracy as compared to the previous best Global Optimal structured loss method (88.68%). Furthermore, the proposed model outperforms more complex ensemble methods, such as WSADAN-ResNet50, EAM, and attention-based models, without incurring any additional computational cost. Another critical finding is that older CNN architectures (DenseNet201, MobileNet variants) with spatial pyramid pooling, multi-head channel attention mechanisms, and other augmentations do not perform well as compared to efficient architectures, such as EfficientNet and InceptionV1.

Table 6: Comparative analysis with SOTA models

Architecture	Accuracy	mF1
MLRSNet Dataset		
FMANet [43]	91.0	
AMEGRF-Net [44]	91.51	
MobileNetV3 + Channel Attention + Spatial pyramid pooling [45]	82.59	
DenseNet201 [46]	-	86.17
DenseNet201 + SSM [46]	-	86.56
DenseNet201 + SRBM [46]	-	86.26
Proposed	**95.42%**	
NWPU Dataset		
Architecture	**Accuracy**	
Global Optimal structured loss[47]	90.30	
DBOW feature based [48]	82.10	
DELF + VLAD [49]	85.70	
IBNR-65 + Densenet-64 [50]	91.70	
Khan, J.A., et al. [10]	93.3	
EAM [51]	93.04	
WSADAN-ResNet50 [52]	92.63	
Proposed	**94.73%**	
EUROSAT Dataset		
Global Optimal structured loss [47]	88.68	
EfficientNet [53]	85.23	
MobileNetV2 [54]	87.52	
InceptionV1 [41]	88.51	
Proposed	**96.14%**	

V. Conclusion

The task of remote sensing image classification presents several challenges, including heterogeneous land cover types, variability in scale, and inter-class similarities that pose difficulties for discrimination. Deep learning models, such as CNNs and transformers, have limitations in processing images at a fixed scale, resulting in a high computational burden and low interpretability. This paper proposes a novel ASSPN model based on four key modules. In the first module, a learnable Gaussian pyramid for multi-scale feature extraction is employed, which is followed by a multi-scale selection attention block. This attention block provides each input a weight based on the scale. After that, a cross-feature propagation that utilizes uncertainty to guide the propagation of sources of uncertainty follows the complexity-aware pooling, producing balanced representations of features. The proposed model was evaluated on several benchmarks, including EuroSAT, NWPU-RESISC-45, and MLRSNet, achieving overall accuracies of 96.14%, 94.73%, and 95.42%, respectively. Comparisons with SOTA techniques, pre-trained models, and several analyses, we conclude the following points:

- Proposed ASSPN architecture attained improved accuracy on the selected datasets for remote sensing image classification. Also, the proposed model required less learnable parameters compared to the existing SOTA models
- Adding the scale selection attention module alone increased accuracy to 89.3%, whereas the cross feature propagation module only attained 93.5%.
- Both SSAM and CPFM together achieved 91.6% accuracy; however, including complexity aware pooling, the overall model achieved 95.3% accuracy on MLRSNet.
- LIME-based explainability shows the proposed model is appropriate for attending to semantically coherent regions of the image that are justified its predictions.

Despite these achievements, the proposed model still has limitations, including confusion with visually similar categories such as croplands and residential areas, as well as challenges in applying it beyond the laboratory in resource-constrained situations. Although ASSPN exhibits strong robustness across datasets and noise perturbations, certain limitations remain. Its performance on hyperspectral or very low-resolution imagery is not yet established, as the current modules are optimized for RGB spatial patterns. Additionally, the model may face challenges under domain shifts, such as variations in sensors, regions, or environmental conditions. Future work could extend ASSPN with spectral-aware modules, domain-generalization strategies, and lightweight variants for broader adaptability. Also, the future work would involve obtaining a more lightweight response for split-second operational uses of the technologies, conducting large-scale testing across domains, and reducing reliance on annotated labeled data while conducting semi-supervised learning with massive unannotated data, thereby positioning ASSPN as a robust solution to scalable technologies in operational remote sensing classification.

Conflict of Interest

All authors declared no conflict of interest.

Dataset Availability

The datasets of this work are publically available for the research purposes.

XI. REFERENCES

[1] A. Hamza *et al.*, "An integrated parallel inner deep learning models information fusion with Bayesian optimization for land scene classification in satellite images," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 16, pp. 9888-9903, 2023.

[2] B. Ahmed, T. Akram, S. R. Naqvi, A. Alsuhaibani, M. A. Khan, and N. Kraiem, "XcelNet14: A Novel Deep Learning Framework for Aerial Scene Classification," *IEEE Access*, 2024.

[3] M. Weiss, F. Jacob, and G. Duveiller, "Remote sensing for agricultural applications: A meta-review," *Remote sensing of environment*, vol. 236, p. 111402, 2020.

[4] M. A. Shafaey, M. A.-M. Salem, H. M. Ebied, M. N. Al-Berry, and M. F. Tolba, "Deep learning for satellite image classification," in *International Conference on Advanced Intelligent Systems and Informatics*, 2018: Springer, pp. 383-391.

[5] M. A. Moharram and D. M. Sundaram, "Dimensionality reduction strategies for land use land cover classification based on airborne hyperspectral imagery: a survey," *Environmental Science and Pollution Research*, vol. 30, no. 3, pp. 5580-5602, 2023.

[6] E. Dahan, I. Aviv, and T. Diskin, "Aerial Imagery Redefined: Next-Generation Approach to Object Classification," *Information*, vol. 16, no. 2, p. 134, 2025.

[7] J. P. Shim, M. Warkentin, J. F. Courtney, D. J. Power, R. Sharda, and C. Carlsson, "Past, present, and future of decision support technology," *Decision support systems*, vol. 33, no. 2, pp. 111-126, 2002.

[8] M. A. Khan *et al.*, "Coastal and Land Use Land Cover Area Recognition from High-Resolution Remote Sensing Images using a Novel Multimodal Attention Inception Residual Deep Network," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2025.

[9] T. Xu *et al.*, "Evaluation of twelve evapotranspiration products from machine learning, remote sensing and land surface models over conterminous United States," *Journal of Hydrology*, vol. 578, p. 124105, 2019.

[10] J. A. Khan *et al.*, "Design of Super Resolution and Fuzzy Deep Learning Architecture for the Classification of Land Cover and Landsliding using Aerial Remote Sensing Data," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2024.

[11] G. Mountrakis, J. Im, and C. Ogole, "Support vector machines in remote sensing: A review," *ISPRS journal of photogrammetry and remote sensing*, vol. 66, no. 3, pp. 247-259, 2011.

[12] M. Belgiu and L. Drăguț, "Random forest in remote sensing: A review of applications and future directions," *ISPRS journal of photogrammetry and remote sensing*, vol. 114, pp. 24-31, 2016.

[13] M. Pal and P. M. Mather, "Decision tree based classification of remotely sensed data," in *22nd Asian conference on remote Sensing*, 2001, vol. 5, p. 9.

[14] M. Pal and P. M. Mather, "An assessment of the effectiveness of decision tree methods for land cover classification," *Remote sensing of environment*, vol. 86, no. 4, pp. 554-565, 2003.

[15] L. Samaniego, A. Bárdossy, and K. Schulz, "Supervised classification of remotely sensed imagery using a modified $ k $-NN technique," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 46, no. 7, pp. 2112-2125, 2008.

[16] X. Chao and Y. Li, "Semisupervised few-shot remote sensing image classification based on KNN distance entropy," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 15, pp. 8798-8805, 2022.

[17] M. J. Cracknell and A. M. Reading, "Geological mapping using remote sensing data: A comparison of five machine learning algorithms, their response to variations in the spatial distribution of training data and the use of explicit spatial information," *Computers & Geosciences*, vol. 63, pp. 22-33, 2014.

[18] A. E. Maxwell, T. A. Warner, and F. Fang, "Implementation of machine-learning classification in remote sensing: An applied review," *International journal of remote sensing*, vol. 39, no. 9, pp. 2784-2817, 2018.

[19] M. J. Abbas *et al.*, "SEMSF-Net: Explainable Squeeze-Excitation Multi-Scale Fusion Network for Aerial Scene and Coastal Area Recognition using Remote Sensing Images," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2025.

[20] F. S. Alsubaei, A. A. Alneil, A. Mohamed, and A. Mustafa Hilal, "Block-scrambling-based encryption with deep-learning-driven remote sensing image classification," *Remote Sensing*, vol. 15, no. 4, p. 1022, 2023.

[21] F. Gao, X. Jin, X. Zhou, J. Dong, and Q. Du, "MSFMamba: Multi-scale feature fusion state space model for multi-source remote sensing image classification," *IEEE Transactions on Geoscience and Remote Sensing*, 2025.

[22] S. Han *et al.*, "High-Accuracy Mapping of Coastal and Wetland Areas Using Multi-Sensor Data Fusion and Deep Feature Learning," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2025.

[23] M. Stoimchev, J. Levatić, D. Kocev, and S. Džeroski, "SSL-MAE: Adaptive Semi-Supervised Learning Framework for Multi-Label Classification of Remote Sensing Images Using Masked Autoencoders," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2025.

[24] K. El Khoury *et al.*, "Enhancing remote sensing vision-language models for zero-shot scene classification," in *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025: IEEE, pp. 1-5.

[25] Y. Li *et al.*, "Meet: A million-scale dataset for fine-grained geospatial scene classification with zoom-free remote sensing imagery," *IEEE/CAA Journal of Automatica Sinica*, vol. 12, no. 5, pp. 1004-1023, 2025.

[26] H. Firat, M. E. Asker, M. I. Bayindir, and D. Hanbay, "3D residual spatial–spectral convolution network for hyperspectral remote sensing image classification," *Neural Computing and Applications*, vol. 35, no. 6, pp. 4479-4497, 2023.

[27] M. Esmaeili, D. Abbasi-Moghadam, A. Sharifi, A. Tariq, and Q. Li, "ResMorCNN model: hyperspectral images classification using residual-injection morphological features and 3DCNN layers," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 17, pp. 219-243, 2023.

[28] Z. Zhang *et al.*, "Remote sensing image scene classification in hybrid classical–quantum transferring CNN with small samples," *Sensors*, vol. 23, no. 18, p. 8010, 2023.

[29] C. Sitaula, S. KC, and J. Aryal, "Enhanced multi-level features for very high resolution remote sensing scene classification," *Neural Computing and Applications*, vol. 36, no. 13, pp. 7071-7083, 2024.

[30] B. Du, Z. Yuan, Y. Bo, and Y. Zhang, "A combined deep learning and prior knowledge constraint approach for large-scale forest disturbance detection using time series remote sensing data," *Remote Sensing*, vol. 15, no. 12, p. 2963, 2023.

[31] V. Pushpalatha, P. Mallikarjuna, H. Mahendra, S. R. Subramoniam, and S. Mallikarjunaswamy, "Land use and land cover classification for change detection studies using convolutional neural network," *Applied Computing and Geosciences*, vol. 25, p. 100227, 2025.

[32] K. VanExel, S. Sherchan, and S. Liu, "Optimizing Deep Learning Models for Climate-Related Natural Disaster Detection from UAV Images and Remote Sensing Data," *Journal of Imaging*, vol. 11, no. 2, p. 32, 2025.

[33] R. Vaghela *et al.*, "Land cover classification for identifying the agriculture fields using versions of yolo v8," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2025.

[34] I. Haider *et al.*, "Performance of pre-trained deep learning models for land use land cover classification using remote sensing imaging datasets," *Environmental Earth Sciences*, vol. 84, no. 11, pp. 1-36, 2025.

[35] J. Bergstra, R. Bardenet, Y. Bengio, and B. Kégl, "Algorithms for hyper-parameter optimization," *Advances in neural information processing systems*, vol. 24, 2011.

[36] P. Helber, B. Bischke, A. Dengel, and D. Borth, "Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 12, no. 7, pp. 2217-2226, 2019.

[37] H. Haikel, "NWPU-RESISC45 Dataset with 12 classes," *Figshare: London, UK*, 2021.

[38] X. Qi *et al.*, "MLRSNet: A multi-label high spatial resolution remote sensing dataset for semantic scene understanding," *ISPRS*

Journal of Photogrammetry and Remote Sensing, vol. 169, pp. 337-350, 2020.

[39] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, 2012.

[40] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[41] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1-9.

[42] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770-778.

[43] F. Rauf *et al.*, "FMANet: Super Resolution Inverted Bottleneck Fused Self-Attention Architecture for Remote Sensing Satellite Image Recognition," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2024.

[44] Z. Li, J. Hu, K. Wu, J. Miao, and J. Wu, "Adjacent-Atrous Mechanism for Expanding Global Receptive Fields: An End-to-End Network for Multi-Attribute Scene Analysis in Remote Sensing Imagery," *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

[45] X. Yang *et al.*, "An Efficient Lightweight Satellite Image Classification Model with Improved MobileNetV3," in *IEEE INFOCOM 2024-IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, 2024: IEEE, pp. 1-6.

[46] X. Tan, Z. Xiao, J. Zhu, Q. Wan, K. Wang, and D. Li, "Transformer-driven semantic relation inference for multilabel classification of high-resolution remote sensing images," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 15, pp. 1884-1901, 2022.

[47] P. Liu, G. Gou, X. Shan, D. Tao, and Q. Zhou, "Global optimal structured embedding learning for remote sensing image retrieval," *Sensors*, vol. 20, no. 1, p. 291, 2020.

[48] X. Tang, X. Zhang, F. Liu, and L. Jiao, "Unsupervised deep feature learning for remote sensing image retrieval," *Remote Sensing*, vol. 10, no. 8, p. 1243, 2018.

[49] R. Imbriaco, C. Sebastian, E. Bondarev, and P. H. de With, "Aggregated deep local features for remote sensing image retrieval," *Remote Sensing*, vol. 11, no. 5, p. 493, 2019.

[50] H. M. Albarakati *et al.*, "A novel deep learning architecture for agriculture land cover and land use classification from remote sensing images based on network-level fusion of self-attention architecture," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2024.

[51] C. Sitaula, S. KC, and J. Aryal, "Enhanced Multi-level Features for Very High Resolution Remote Sensing Scene Classification. arXiv 2023," *arXiv preprint arXiv:2305.00679*.

[52] W. Liming, Q. Kunlun, Y. Chao, and W. Huayi, "Weakly supervised scale adaptation data augmentation for scene classification of high-resolution remote sensing images," *National Remote Sensing Bulletin*, vol. 27, no. 12, pp. 2815-2830, 2024.

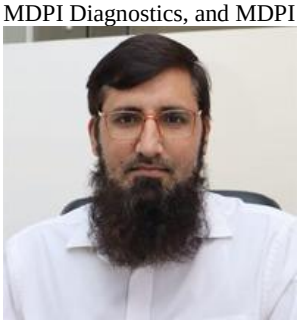
[53] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International conference on machine learning*, 2019: PMLR, pp. 6105-6114.

[54] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4510-4520.

Authors Biography

Muhammad John Abbas received the bachelor's degree in 2025 from HITEC University, Rawalpindi, Pakistan. Currently, he is research associate at Prince Mohammad bin Fahd Univeristy, KSA under the Center of AI. He is a highly skilled data scientist and machine learning expert with a passion for remote sensing and biomedical engineering. With a strong background in computer science and mathematics, he has extensive experience in developing and deploying complex models for a variety of applications. John major expertise includes a variety of machine learning techniques such as supervised and unsupervised learning, deep learning and computer vision.

Muhammad Attique Khan (Member IEEE) received the master's and Ph.D. degrees in human activity recognition for application of video surveillance and skin lesion classification using deep learning from COMSATS University Islamabad, Islamabad, Pakistan, in 2018 and 2022, respectively. He is currently an Assistant Professor with AI Department, Prince Mohammad Bin Fahd, Al-Khobar, Saudi Arabia. His primary research focus in recent years is medical imaging, COVID-19, MRI analysis, video surveillance, human gait recognition, and agriculture plants using deep learning. He has above 350 publications that have more than 16 000+ citations and an impact factor of 1050+ with h-index 74 and i-index 230. He is the Reviewer of several reputed journals, such as the IEEE Transaction on Industrial Informatics, IEEE Transaction of Neural Networks, Pattern Recognition Letters, Multimedia Tools and Application, Computers and Electronics in Agriculture, IET Image Processing, Biomedical Signal Processing Control, IET Computer Vision, EURASIP Journal of Image and Video Processing, IEEE Access, MDPI Sensors, MDPI Electronics, MDPI Applied Sciences, MDPI Diagnostics, and MDPI Cancers.



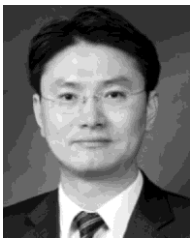
Waqas Ahmed is currently working at department of computer science at HITEC University, Taxila, Pakistan. His PhD in Computer vision, major electrical engineering from UEt Taxila, Pakistan. He is working in the area of computer vision and deep learning for target detection and remote sensing. He is also a reviewer of several prestigious journals including ieee access, pattern analysis and applications, international journal of remote sensing and remote sensing (mdpi).

Ameer Hamza is currently working toward the Ph.D. degree in computer science with KTU University, Kaunas, Lithuania. His major interests include object detection and recognition, video surveillance, medical, and agriculture using deep learning and machine learning. He has published 20 impact factor papers to date.

Nejib Ben Hadj-Alouane joined the American University in Dubai in 2023 as a Professor of Computer Engineering, specializing in AI, and currently serves as Director of the MSAI program. He previously held faculty positions at the National Engineering School of Tunis (ENIT) and the National School of Computer Sciences (ENSI), contributing to curriculum modernization and establishing advanced Master's and PhD programs. Before academia, he gained extensive industry experience, including R&D consulting at Dow Chemical in the USA, developing automated control software for chemical plants, and designing traffic simulation systems, as well as serving as Information Systems Manager for a major industrial and hospitality group in Tunisia, implementing ERP solutions.
Prof. Ben Hadj-Alouane has authored over 100 publications in AI, healthcare, agile automation, security, Web services, Cloud/Fog gaming, and smart agriculture. He has supervised more than 20 PhD and Master's students and co-founded a software start-up, mentoring engineers and architecting HR management solutions for remote mining operations in Canada.

Shrooq Alsenan received the Ph.D. degree in information systems' sciences from King Saud University, Riyadh, Saudi Arabia.,She is an academic and a researcher of artificial intelligence, and currently directs the AI Center with Princess Nourah bint Abdulrahman University, Riyadh, Saudia Arabia. She has received a prestigious postdoctoral fellowship with CSAIL and Jameel Clinic, MIT. Her research expertise spans AI in healthcare, remote sensing, bioinformatics, and hyperspectral images.

M. Turki-Hadj Alouane received the Senior Electrical Engineering Diploma from the National Engineering School of Tunis (ENIT) in 1989, the Master of Science in Systems Analysis and Signal Processing in 1991, and the Ph.D. Degree in Electrical Engineering from ENIT, in 1997. She is currently a Professor at the College of Computer Science, King Khalid University, KSA. In September 1997, she was recruited as an Assistant Professor of electrical engineering at ENIT. In June 2007, she received the National Tenure Diploma in Telecommunications from ENIT. In December 2007, she was promoted to Associate Professor of telecommunications at ENIT. From 2010 to 2012, she was a Visiting Associate Professor at the Electricity Department, Polytechnic School of Tunisia (EPT). Since 2012, she has been a Full Professor of telecommunications at the Information and Communication Technologies (ICT) Department, ENIT. She has coordinated internationally sponsored research projects. Since 1997, she has led more than 20 research master theses and 8 Ph.D. theses. She published more than 70 papers in impact journals and conferences. Her research interests include Signal Processing and Artificial intelligence.



Yunyoung Nam (Member, IEEE) received the B.S., M.S., and Ph.D. degrees in computer engineering from Ajou University, South Korea, in 2001, 2003, and 2007, respectively. He was a Senior Researcher with the Center of Excellence in Ubiquitous System, Stony Brook University, Stony Brook, NY, USA, from 2007 to 2010, where he was a Postdoctoral Researcher, from 2009 to 2013. He was a Research Professor with Ajou University, from 2010 to 2011. He was a Postdoctoral Fellow with Worcester Polytechnic Institute, Worcester, MA, USA, from 2013 to 2014. He was the Director of the ICT Convergence Rehabilitation Engineering Research Center, Soonchunhyang University, from 2017 to 2020. He has been the Director of the ICT Convergence Research Center, Soonchunhyang University, since 2020, where he is currently an Assistant Professor with the Department of Computer Science and Engineering. His research interests include multimedia database, ubiquitous computing, image processing, pattern recognition, context-awareness, conflict resolution, wearable computing, intelligent video surveillance, cloud computing, biomedical signal processing, rehabilitation, and healthcare systems.