DFSNet-VLM: A Hybrid Frequency-Aware and Vision-Language Framework for Remote Sensing Scene Classification and Semantic Image Explanation

Muhammad John Abbas, *Member IEEE*, **Muhammad Attique Khan,** *Member IEEE*, **Ameer Hamza,** ***Member IEEE,*** **Shrooq Alsenan, Areej Alasiry, Mehrez Marzougui, Jungpil Shin,** ***Senior Member IEEE,*** **Yunyoung Nam,** ***Member IEEE***

Abstract— Remote sensing has always been an area of interest for researchers due to its significance in Earth monitoring, which supports proper future planning for agriculture, construction, reforestation, and climate change. Transformer architecture achieves significant performance in remote sensing image classification; however, they come with the trade-off of higher computational complexity. In this paper, we propose a novel deep learning framework, DFSNet-VLM—Cross Domain Fusion based Texture-Sensitive Dual Stream Network — for high-precision remote sensing scene understanding. The proposed framework includes a classification model, "DFSNet," that improves feature representation by employing both spatial and frequency-domain features, which ultimately help detect global patterns and textures alongside local features. The model also promotes information exchange between both streams to complement one type of features with respect to the other by integrating cross-domain fusion blocks at multiple stages. Additionally, a pretrained VLM model, "BLIP-2," is integrated to provide semantic descriptions of classified images. Bayesian optimization is applied to fine-tune hyperparameters, reducing overfitting and improving model performance. The proposed model is evaluated on six diverse publicly available datasets and achieves improved accuracies of 97.13% on MLRSNet, 94.67% on NWPU-RESISC-45, 98.00% on EuroSAT, 92.25% on GeoSceneNet16k, 98.25% on cloud, and 96.03% on the Bijie-landslide dataset, respectively. Detailed ablation studies, comparative analysis, and Grad-CAM++-based model explainability demonstrate that the proposed model is generalizable and scalable, and that it achieves improved accuracy. In addition, the proposed model can be easily implemented in a real-time environment for diverse applications. The trained model's links are available in the data availability section.

Index Terms— Remote sensing; VLM; land use; land cover; vegetation; deep learning; spatial information; explainable AI

Funding: This work is supported by Princess Nourah bint Abdulrahman University Researchers Supporting Project number (PNURSP2026R506), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia. This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government (MSIT) (No. RS-2023-00218176) and the Soonchunhyang University Research Fund. The authors extend their appreciation to the Deanship of Research and Graduate Studies at King Khalid University for funding this work through small Research Project under grant number RGP1/290/46.

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

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

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

Areej Alasiry, Mehrez Marzougui are with College of Computer Science, King Khalid University, Abha 61413, Saudi Arabia (areej.alasiry@kku.edu.sa; mhrez@kku.edu.sa)

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).

Jungpil Shin is with School of Computer Science and Engineering, The University of Aizu, Aizuwakamatsu 965-8580, Japan (jpshin@u-aizu.ac.jp).

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

I. INTRODUCTION

Aerial remote sensing plays a significant role in earth monitoring as it helps in understanding environmental changes, predicting calamities, land change detection, and resource management [1, 2]. Technical evolutions in aerial sensing led to an increase in remote sensing (RS) data production, which is now used in various fields such as geomorphology [3], precision agriculture [4], forest ecology [5], climate monitoring [6], and many more. This enriched RS data and statistics result in a much cheaper and accurate analysis of Earth's observations and geographic importance [7]. However, data proliferation in aerial imagery gave rise to some significant challenges, including data management and analysis [8]. In the early days, data was analyzed manually, which required a lot of time and human power but still lacked accuracy. Moreover, the size, variations, and diversity of data make analysis almost impossible and highlight the need for computer-automated systems [9]. However, the advancement in computer-aided technology gave birth to Industry 5.0 that addresses the challenges related to numerous domains of science, engineering, and technology, including agriculture, medical imaging, classification, segmentation, scene detection, and so on [10].

Land Use Land Cover classification (LULC) involves a lot of complexities, mainly due to noisy background, inter-class similarity, and intra-class variations that can be shown in Figure 1. Apart from this, irregular regional data, a large number of extracted features, and multiple scene classification in a single image make LULC classification even harder [11]. Accurate classification of an aerial image by a computer involves data preprocessing, defining boundaries, efficient

feature extraction, dimensionality reduction, and, most importantly, a good classification technique [12, 13]. The accuracy and efficiency of a machine learning (ML) system depend on the nature of the dataset, the quality of the images, and the complexity of the problem [14]. In this context, supervised ML algorithms are practical when we have labelled data, unsupervised learning helps find hidden patterns through clustering and dimensionality reduction, and reinforcement learning will work best in decision-making problems [15].

Figure 1: Sample images of diverse data collected from RS datasets [16]

Deep learning (DL) is a subdomain of ML that was introduced to handle high-dimensional data such as images, audio, videos, etc. Various deep learning algorithms, especially Convolutional Neural Networks (CNN), show promising results in image classification tasks [16]. The CNN shows the improved performance for RS image classification to a whole new level by enabling automatic feature extraction [17]. Despite the good performance of various pre-trained models, namely EfficientNet, ResNet, MobileNet, and VGG, in other classification problems, their failure in RS applications suggests that simple pre-trained models are unable to capture complex relationships, mainly due to varying scales and spectral differences [18].

Recent advancements in deep learning have introduced Large Language Models (LLMs) and Visual Language Models (VLMs) that can understand both text and images [19, 20]. Models such as CLIP and BLIP-2 have shown great potential for providing semantic explanations of RS data by generating natural-language descriptions of images [21]. These abilities are further extended by GenAI, which allows models to give human-readable explanations, taking image classification to a whole new level [22, 23].

Remote sensing has been a field of interest for computer vision researchers in the last decade due to advances in technology and the abundance of data [17]. Due to the large amount of remote sensing data, the computational power, models' generalizability, scalability, and accurate prediction of each class remain a challenge [24]. With the advent of Industry 5.0, researchers are shifting towards AI systems and proposing models that can analyze and classify RS data at low cost. Many techniques are introduced in the literature for RS aerial scene classification, LULC classification, and rapid earth region classification from remote sensing images [16, 18].

Junaid et al. [25] designed a customized CNN model for the classification of Land cover areas through remote sensing images. The developed model consists of 40 convolutional layers, four bottleneck blocks, and four inverted bottleneck blocks to handle complex data while reducing computational cost. A fuzzy layer is also introduced to handle unclear information, which results in better feature extraction. The image quality is improved by introducing a separate 5-block CNN, which converts low-resolution images into high-resolution ones, and a chaotic particle swarm optimization (C-PSO) technique is incorporated to select the most essential features. The model is evaluated on three datasets, and experimental results show that the designed model outperforms pre-trained and SOTA models by obtaining an accuracy of 93.5%, 93.30% and 93.34% on the Bijie landslide dataset, EuroSAT dataset, and NWPU-RESISC45 dataset, respectively. Fatima et al. [26] presented a novel CNN-based deep learning model, FMANet, for the classification of remote sensing images into relevant classes such as land cover, landslide, and earthquake regions. The presented technique contains a 20-layered CNN architecture with residual connections to improve the resolution of the original images. High resolution images are passed through the FMANet which further consists of three sub-models: A bottleneck 22, bottleneck 33, and inverted bottleneck 33. Features from all three models are concatenated and passes through a classifier to make predictions which are interpreted by local Interpretable Model agnostic Explanations (LIME). Experiments are conducted on three datasets, and results demonstrate that FMANet surpassed all other models on the Bijie landslide and the Turkey Earthquake dataset by achieving an accuracy of 92.8% and 99.4%, respectively. However, on MLRSNet dataset, the accuracy was slightly lower than some compared studies which needs to be improved. Albarakati et al. [27] presented a new super-resolution (SR) technique using deep learning for accurate LULC classification through RS data. The SR technique is employed for dataset augmentation that is later passed to designed deep networks such as ResSAN6 and RS-IRSAN. The ResSAN6 model consists of 91 layers, whereas the RS-IRSAN includes 115 layers. The information of both models has been fused with the help of Mutual Information-based Serial Fusion (MIbSF), which was later optimized using the Arithmetic Optimization Algorithm (AOA). Three datasets are used for the experimental findings, and achieved an accuracy of 95.7% on the RSI-CB128 dataset, 97.5% on the RSISC dataset, and 92.0% on the NWPU-RESISC45 dataset, respectively. Fayaz et al. [28] compared the performance of three pre-trained models, such as InceptionV3, DenseNet121, and ResNet-50, for land area classification. All the models are fine-tuned to adapt to remote sensing data and then evaluated on the UCMerced_LandUse dataset. Data augmentation is also applied to prevent the model from overfitting. Comparative analysis and ablation studies demonstrate that InceptionV3 outperforms DenseNet121 and ResNet-50 by obtaining an accuracy of 92%. However, the dataset used in this research is minimal, which limits the generalization of the model.

Vaghela et al. [29] investigated the performance of different versions of YoloV8 for agricultural fields identification using remote sensing imagery. RS images of various classes are classified using different versions of YoloV8, including medium, nano, and small. The effects of different

hyperparameters —namely, epoch size, learning rate, momentum, weight decay, and optimizer, are observed across all versions of YoloV8. Experimental results show that the medium version outperformed the other two versions, achieving an accuracy of 99% at 50 epochs. The accuracies obtained by the small and nano versions are 98.50% and 98.60%, respectively. Van et al. [30] addressed the potential of ML and DL algorithms in natural disaster detection through UAV images and RS data. For this purpose, a customized climate change dataset is created to target only two calamities: flooding and desertification. Different open-access datasets are used as sources, and a new dataset of 6334 images is created. Four DL models, namely VGG-16, ResNet-50, optimized DenseNet201, and a custom-made CNN, are evaluated on the proposed dataset. Experimental findings reveal that all the ML models have great potential for RS image classification; however, DenseNet201 and ResNet-50 gain top positions by obtaining an accuracy of 99.37% and 99.21%, respectively.

Madala et al. [31] presented a deep learning framework for crop mapping using remote sensing data. The framework involves data mining techniques, including normalization and cleaning, to preprocess the data, followed by Bidirectional Gated Auto Encoders (BiGAE) for feature extraction. An opposition learning-based mud ring algorithm (Opp-MR) is used for feature selection, and an adaptive Kernel fuzzy clustering (AkFC) technique is used for feature clustering. After that, crops are mapped by Goshawk Integrated Convolutional Attention Efficient Net (GICANet) and obtained an overall accuracy of 97.74%. In [32], the authors introduced a novel deep learning approach for LULC classification using remote sensing images. At first, the original RS images are preprocessed, and classes are grouped into superclasses and subclasses. The main features are extracted by a Relational Autoencoder that establishes relationships between data points across different classes. In the next step, a Relevance Vector Machine (RVM) and a Support Vector Machine (SVM) are employed for the final classification. The presented model is evaluated on three datasets, and experimental results reveal that the framework obtained an overall classification accuracy of 96.70% on the UCM dataset, 95.35% on PatternNet, and 96.26% on the NWPU-RESISC-45 dataset, respectively. Aljebreen et al. [33] presented a novel River Formation Dynamics Algorithm (RFDA) for LULC classification using RS imagery. In the presented LULCC-RFDADL technique, the authors used a Dense EfficientNet for feature extraction and the RFDA technique for fine-tuning of hyperparameters. Multi-Scale Convolutional Autoencoder (MSCAE) is implemented for the classification process, and the Seeker Optimization Algorithm (SOA) is utilized to select parameters for the MSCAE system. The experimental process of the presented model is evaluated on the EuroSAT dataset, with results indicating an overall classification accuracy of 98.12% on the test set and 98.15% on the training set.

Recent years have also witnessed significant progress in applying VLMs and LLMs to remote sensing data. Li et al. [22] provided a comprehensive review of VLMs in remote sensing, which demonstrates that the models combining both visual and textual information perform better than visual-only models in terms of classification, retrieval, and captioning tasks. Hu et al. [34] proposed RSGPT, a remote sensing vision language model for image captioning and visual question answering with reduced hallucinations, whereas Wang et al. [35] introduced LLM4HRS, an LLM-based spatio-temporal imputation model for highly sparse remote sensing data, indicating the performance of LLMs in complex RS observations. Muhtar et al. proposed SeaMo, a season-aware multimodal foundational model that incorporates temporal and seasonal information into remote sensing scene understanding while Kuckreja et al. developed GeoChat, first grounded VLM that combines visual question answering, object detection and regional level image captioning in a single model.

Despite these advancements, most VLM-based approaches focus on retrieval or captioning tasks and do not integrate classification with natural-language explanation in a unified framework. Also, the classification-based approaches discussed above focused on pre-trained models, fine-tuning, feature selection, and data augmentation. These methods are still facing some challenges, such as limited generalization, less attention towards texture details, data imbalance, high computational cost, and sensitivity to noise and spatial variations. To address these challenges, this proposed a unified framework that combines DFSNet for accurate scene classification with BLIP-2 for natural language explanation of remote sensing scenes. Also, the dual stream architecture of DFSNet captures both local features and global patterns due to its spatial and frequency domains. These patterns make it robust to spatial transformations and data variability. Moreover, the incorporated texture-aware convolutional block pays attention to image texture, and the Adaptive feature recalibration blocks improve the overall generalization of the model by enhancing discriminative features dynamically. Instead of manual selection, the Bayesian Optimization technique is applied to optimize the model hyperparameters, which helps in more accurate and robust classification. The main contributions of the proposed DFSNet are as follows:

- In the proposed network, a frequency feature extraction block is incorporated, which converts spatial features into the frequency domain, thus providing robustness, translation invariance, and rich feature representation.
- A texture-aware convolutional block is proposed to handle texture details in RS data, which promotes multi-scale robust feature extraction and improves model generalization while balancing computational complexity.
- Cross-domain fusion blocks are integrated to promote information exchange between streams that ultimately lead to enhanced feature representation, robustness to spatial variations, and improved performance on unseen data.
- Adaptive feature recalibration is performed to emphasize the discriminative features, reduce computational overhead, and help to improve generalization and texture handling.
- A pretrained BLIP-2 vision language model is incorporated into the proposed framework to generate the natural language explanations of RS images, improving the interpretability of the model.

- Proposed DFSNet and BLIP-2 are integrated into a unified pipeline, making it the first work to combine frequency-domain dual stream classification with a pretrained VLM for explainable remote sensing scene understanding.

II. DATASETS

A. For Classification

In this work, six datasets are employed that are publicly available for research purposes. The selected datasets are MLRSNet, NWPU-RESISC45, EuroSAT, GeoSceneNet16k, Cloud dataset, and Bijie-landslide. These datasets are divided into training and a testing set for the detailed experimental process of the proposed architecture. Each dataset description is provided in Table 1, along with references and sample images.

Table 1: Description and samples of selected datasets

Dataset	Characteristics	Sample Images
MLRSNet [36]	An open-source dataset comprised of 109161 images across 46 classes, where the number of images varies from 1500 to 3000 for each class. Images have a fixed pixel size of 256x256 and varying resolution from 10m to 0.1m. The dataset is annotated with 60 labels, where each image can have labels from 1 to 13.	
NWPU [37]	An open-source dataset consists of 10500 images, where each image has a dpi of 96x96 and pixel size of 256x256. The dataset is separated into 12 classes, such as Airfield, Beach, Farm, Parking space, Overpass, Game space, River, Forest, Harbor, Sparse Residential, Dense Residential, and Storage tanks.	
EuroSAT [38]	It is also an open-access dataset consisting of Sentinel-2 RGB images, where each image has a pixel size of 64x64 and 10m as the group sampling distance. The dataset contains almost 27000 images, which are grouped into 10 classes, such as Forest, Annual Crop, Herbaceous vegetation, Industrial, Pasture, highway, Permanent crop, River, Residential, and Sealake.	

GeoSceneNet16k	It is an open-source multi-label dataset containing 16000 images of geospatial data divided across seven classes, namely Building and structures, desert, forest area, Hill or Mountain, Ice glacier, Sea or Ocean, and Street view. Each image is associated with more than one geographic scene (https://www.kaggle.com/datasets/prithivsakthiur/multilabel-geoscenenet-16k).	
Bijie-landslide [39]	It is also an open-source dataset containing satellite images of landslides, which are classified into two groups: landslide and non-landslide. Out of 2773 images, 770 images belong to the landslide class, while the rest belong to the non-landslide class.	
Cloud dataset [40]	Cloud dataset is an open-access dataset containing satellite images of cloud and non-cloud areas, which can help in cloud detection. About 1342 images are present in the cloud class, while 1134 images belong to the non-cloud class. The dataset is sourced from "Hurricane Lan" open repository.	

B. For VLM-based Image Explanation

To generate natural language explanations of RS images, we used the publicly available dataset "RSVLM-QA" [41], which contains aerial imagery of various geographic scenes and land-use types, including residential areas, densely populated urban regions, industrial areas, transportation networks, water bodies, and forest areas, with proper semantic annotations. The original dataset was in JSONL format, containing one entry per image, along with all corresponding semantic information: scene-level tags, object-relation pairs, and visual question-answer pairs. Data preprocessing began with consistency checks between metadata and samples to exclude unavailable samples. A hierarchical approach was used for text generation, with scene tags serving as the source for textual descriptions. When scene tags were absent, object-relation annotations were used. A generic description using RS language was assigned in the absence of both scene tags and object relation annotations. All these generated descriptions were then normalized by converting them to lowercase, removing special characters, and limiting the maximum length to 20 words. After preprocessing, 180 samples remained, split into a 60:15:25 train-val-test ratio. This clean, normalized, and processed dataset was used for training and testing the robust VLM.

III. Proposed Methodology

Remote sensing images involve a diverse range of structural and textual variations that require a fine-grained, high-resolution architecture for accurate classification and interpretability. In this section, we presented our proposed novel deep learning framework known as DFSNet-VLM. The proposed framework integrates "DFSNet" for image classification and pretrained BLIP-2 (OPT) for natural language text description. The proposed DFSNet-VLM framework addresses two fundamental problems in remote sensing scene understanding: accurate scene classification and semantic scene explanation. Given a remote sensing image dataset $D = \{(I_i, y_i, T_i)\}_{i=1}^{N}$, where $I_i \in \mathbb{R}^{H\times W\times 3}$ is the input image, $y_i \in \{1,2,\dots C\}$ is the ground truth class label from the C total scene categories, and T_i is the natural language description of the scene. The goal is to learn two functions: first, a classification function $f_{cls}: I \to \hat{y}$ that maps the input image to a predicted class label, and second a description generation function $f_{desc}: I \to \hat{T}$ that maps the input image to a natural language explanation. So, the classification task is formulated as supervised learning problem where objective is to minimize the loss function. On the other hand, the description generation task is formulated as a language modeling problem where the objective is to maximize the

likelihood of generating the correct description given the image. So, our proposed framework "DFSNet-VLM" optimizes both tasks independently on the same input image. A visual representation of the overall framework is shown in Figure 2.

A. Proposed DFSNet Architecture

The proposed DFSNet model involves both spatial and frequency domain features to capture local as well as global repetitive patterns and textures. This technique comprises four main steps, which involve Fast Fourier Transform feature extraction, Texture-aware convolutional blocks, Cross-domain feature fusion, and adaptive feature recalibration before final classification. These steps allow the model to differentiate among various classes of remote sensing despite their inter-class similarity and intra-class variability. Unlike existing dual-stream networks that typically operate in either spatial domain only or combine two spatial streams, this network combines a spatial stream with a frequency-domain stream, processed through FFT. Also, the integration of texture-aware convolutional blocks and attention-based cross-domain fusion at multiple stages instead of the final layer only makes the proposed DFSNet more effective at capturing both local and global patterns. Moreover, the proposed model accurately performed explainability using explainable AI techniques. Figure 2 shows the complete architecture of the proposed DFSNet.

1) Fast Fourier Transform Feature Extraction

As shown in Figure 2, the Fast Fourier transform (FFT) [42] is applied on input images at the initial step to convert the spatial features into the frequency domain to extract the information about texture, edges, and overall patterns. Unlike the spatial domain, which represents the image in terms of pixel intensities, the frequency domain computes the magnitude spectrum of the image. It means it is in the form of sinusoidal components of varying amplitudes, frequency, and phases. RS images often contain repetitive global patterns, such as agricultural fields, urban grids, and forest textures, that are difficult to capture with spatial convolutions alone. The frequency domain, obtained through FFT, naturally reveals that RS images often contain repetitive global patterns, such as agricultural fields, urban grids, and forest textures, which are difficult to capture. However, the model treats these repetitive structures as dominant frequency components, making it easier for it to distinguish between scene categories. Moreover, these frequency representations are robust to certain types of noise and slight spatial variations in the image, enabling the model to detect an object regardless of its position. In the proposed architecture, a FFT feature extraction function is defined, which converts an image from the spatial to the frequency domain by applying a 2D FFT on each channel separately. For an image of size M×N×3, frequency component can be computed using Equation (1).

$$F\ (u,v) = \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} I(x,y)\ .\ e^{-j2\pi(\frac{ux}{M},\frac{vy}{N})} \quad (1)$$

Here, x and y represents the coordinates of spatial domain while u and v represents the coordinates of frequency domain. The variable $I(x,y)$ represents pixel intensities at coordinates (x,y) while $F\ (u,v)$ represents the corresponding complex frequency component. The DFT operation places the DC component also known as zero frequency component at $F(0,0)$ which causes inconvenience in visualization and analysis, as low frequencies are present at corners of output.

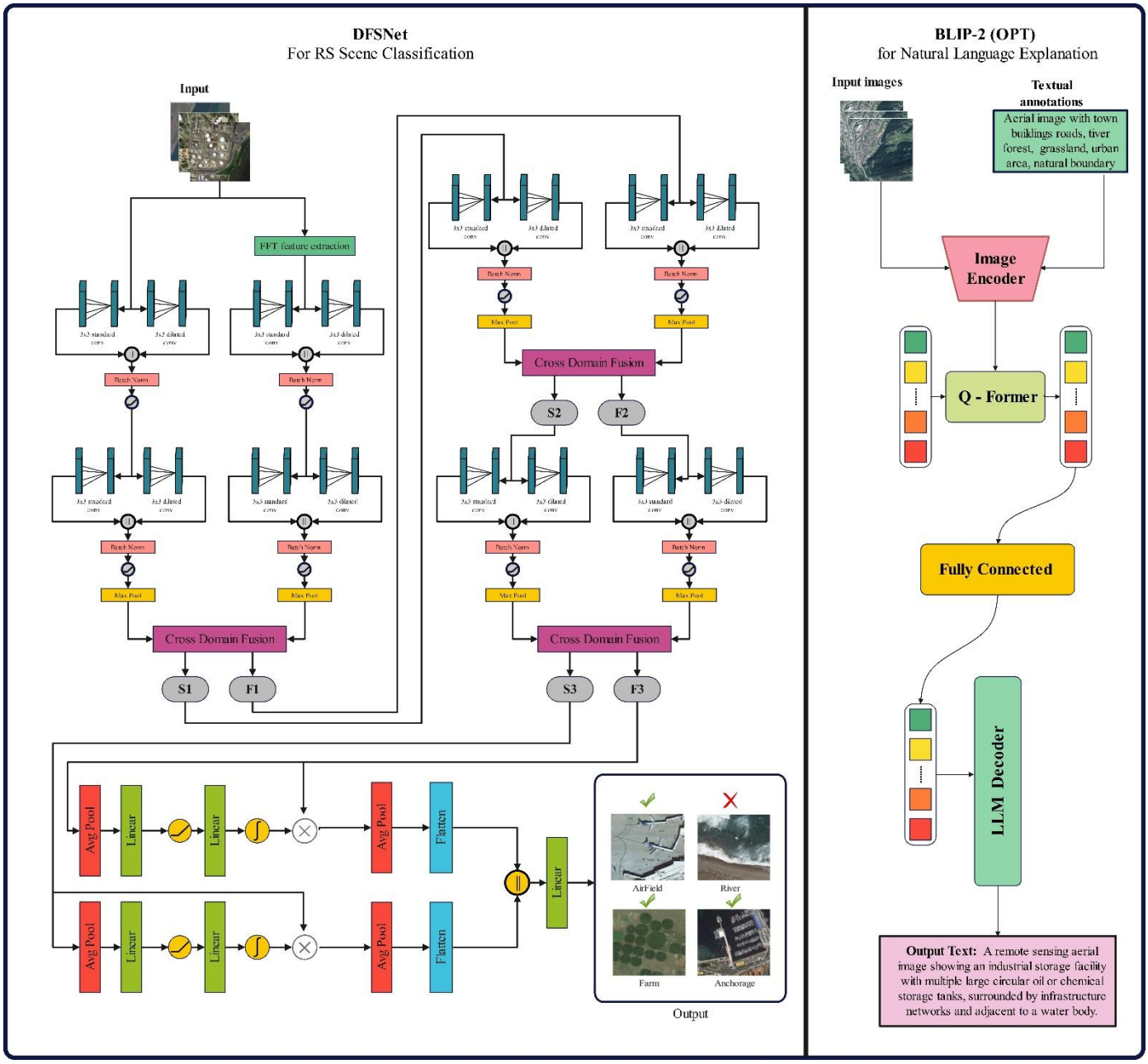

Figure 2: Proposed DFSNet-VLM model for remote sensing image classification

To handle this, a shift operation is performed which shuffles the indices in such a way that DC component is placed at the center of output. For $M \times N$ array, all the frequencies are centered around (M/2, N/2) after shift operation. These shuffled indices can be calculated using Equations (2-3).

$$u' = u + \left(\frac{M}{2}\right) \quad (2)$$

$$v' = v + \left(\frac{N}{2}\right) \quad (3)$$

Where u' and v' represent indices after shift operation. The frequency components contain information about both magnitude and phase; however, our main focus is on amplitudes only. While phase information in FFT encodes structural layout details, it is highly sensitive to small spatial shifts and misalignments that are commonly present in remote sensing imagery due to varying acquisition angles and sensor positions. This sensitivity makes phase features unreliable and inconsistent across different RS images of the same scene category. On the other hand, the magnitude spectrum is able to represent the strength of each frequency component regardless of their spatial position, providing more stable and consistent representations. Also, these magnitudes vary in a diverse range which makes it difficult to process in later stages. For this purpose, magnitude and log scaling is performed which discards the phase of frequency components as well as compress the magnitude range. Mathematically, it can be defined by Equation (4).

$$F(u', v') = \log(|F(u, v)| + 10^{-10}) \quad (4)$$

All these operations are performed on each channel separately which are then combined together to make a $M \times N \times 3$ output

tensor. A pictorial representation of FFT feature extraction block is also illustrated in Figure 3. The frequency and spatial features extracted through FFT are then passed to their corresponding texture-aware convolutional blocks for further refinement.

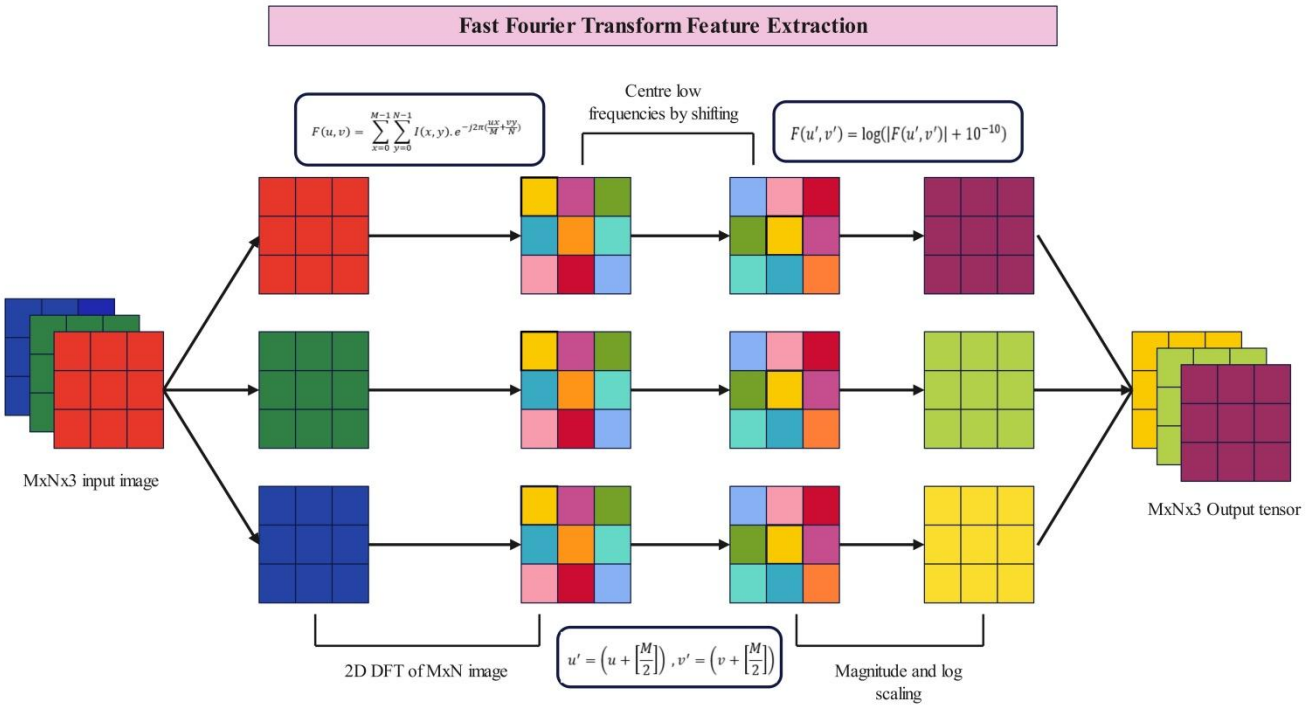

Figure 3: Pictorial representation of Fast Fourier Transform (FFT) feature extraction

2) Texture Aware Convolutional Block

In this section, the extracted features such as spatial domain features and frequency domain features are passed to their corresponding texture aware convolutional blocks (TACB). The TACB contains a standard 3×3 convolutional layer along with a dilated 3×3 convolutional layer to extract local features as well as capture large-scale textures. A standard convolutional layer with kernel size of 3×3 and padding of one capture local information by performing simple convolutional operation. In dilated convolutional layer, receptive field is dilated to capture larger area without increasing computational cost. The receptive field of dilated convolutional layer depends upon dilation rate ∂ and can be calculated by Equation (5):

$$C(k \times k) = 2\partial + 1, 2\partial + 1 \quad (5)$$

Here $k \times k$ represents the receptive field. In this case, ∂ is 2 so $k \times k$ is 5×5. The kernel weights are applied only at positions $(m.\partial\ , n.\partial)$ where $m, n \in \{-1, 0, 1\}$. At these dilated positions, input pixels are sampled, and the convolution is computed using Equation (6).

$$Y'(x,y) = \sum_{m,n} W(m,n).\ (x + m.\partial,\ \ y + n.\partial) \quad (6)$$

Where $W(m,n)$ represents convolutional weights and $Y'(x,y)$ represents output at x, y coordinates. The extracted features of both convolutional layers are then concatenated and passed to the batch normalization layer to stabilize the training. It is then followed by a ReLU activation function to introduce non-linearity which enables the model to capture complex patterns. The proposed texture aware convolutional block is shown in Figure 4. In the last, another set of texture aware convolutional block is incorporated in the architecture to enhance feature representation before passing the output of spatial and frequency stream to cross domain fusion block.

The refined spatial and frequency features are subsequently fed into the cross-domain fusion block, where complementary information from both streams is integrated.

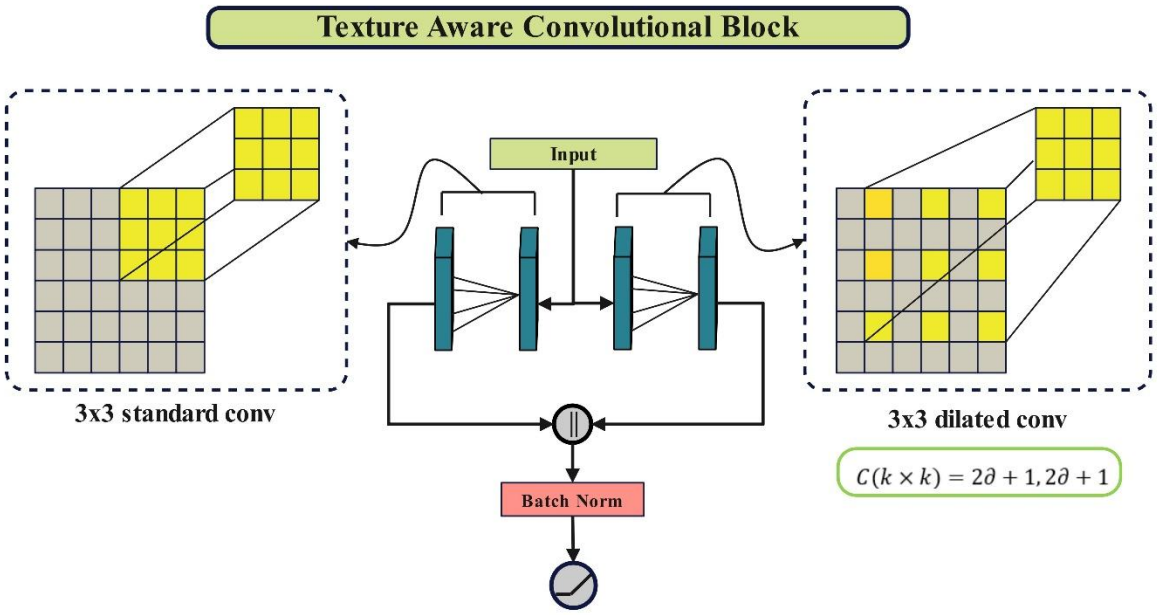

Figure 4: Proposed Texture Aware Convolutional Block (TACB) for RS image classification

3) Cross-Domain Feature Fusion

In this section, the proposed cross-domain features fusion block (CDFF) is presented. A cross-domain fusion block integrates complementary information from both streams to enhance the model's overall classification abilities. It works on a simple principle that spatial features know where things are, while frequency features know what patterns and textures exist globally. By allowing these two streams to guide each other through attention maps, the fusion block helps each stream focus on the most relevant information from the other. In simpler terms, spatial features highlight important frequency regions, and frequency features highlight important spatial regions, generating a more informed feature representation. It comprises four main components: spatial to frequency, frequency to spatial, spatial attention, and frequency attention. The CDFF block accepts spatial and frequency features as input and enhances them by performing a series of operations. First of all, input spatial features are passed through an 1x1-layer convolutional layer to align them with the frequency representation. After that, a batch normalization layer β is applied to stabilize the training process which is followed by ReLU activation σ to enable non-linear transformations. A similar procedure is applied to frequency features to make them compatible with spatial representations for information exchange. Mathematically, the CDFF is defined by Equations (7-8).

$$S_F = \sigma\big(\beta(W_s * S + b_s)\big) \quad (7)$$

$$F_s = \sigma\left(\beta\big(W_f * F + b_f\big)\right) \quad (8)$$

Where, S represents spatial features, F represents frequency features, S_F represents Spatial to frequency, and F_s represents frequency to spatial. The variable W_s and W_f represents corresponding convolutional weights and b_s and b_f represents bias terms. The spatial-to-frequency features are concatenated with the original frequency features and passed through a 3×3 convolutional layer, which processes these features to generate a spatial attention map. This attention map highlights important spatial regions complemented by overall frequency representations. After the convolutional layer, a batch normalization layer is applied to normalize the outputs, which are followed by the sigmoid activation function that scales the features in the range of [0,1]. The generated spatial attention map is then multiplied by the original spatial features to emphasize the essential features and suppress the less important ones. Mathematically, it can be represented as:

$$S_A = \psi(\beta((W_s' * (S_F \parallel F)) + b_s')) \quad (9)$$

$$A_s = S_A \times S \quad (10)$$

Where, S_A represents spatial attention map and A_s represents spatially emphasized feature map. In the same way, frequency attention map is also generated which is multiplied with original frequency features to enhance important frequency features guided by cross domain information. It can also be represented as:

$$F_A = \psi(\beta((W_f' * (F_S \parallel S)) + b_f')) \quad (11)$$

$$A_f = F_A \times F \quad (12)$$

Where, F_A represents frequency attention map and A_f represents frequency emphasized feature map. These spatially emphasized feature maps are combined with frequency and spatial features to generate an overall enhanced spatial feature map. Similarly, frequency emphasized feature maps are added to spatial to frequency features to produce enhanced frequency feature representation, which can be defined as:

$$S_{Enh} = A_s + F_s \quad (13)$$

$$F_{Enh} = A_f + S_F \quad (14)$$

Here, S_{Enh} represents spatially enhanced feature map and F_{Enh} represents frequency enhanced feature map. A pictorial representation of proposed cross domain fusion block is shown in Figure 5.

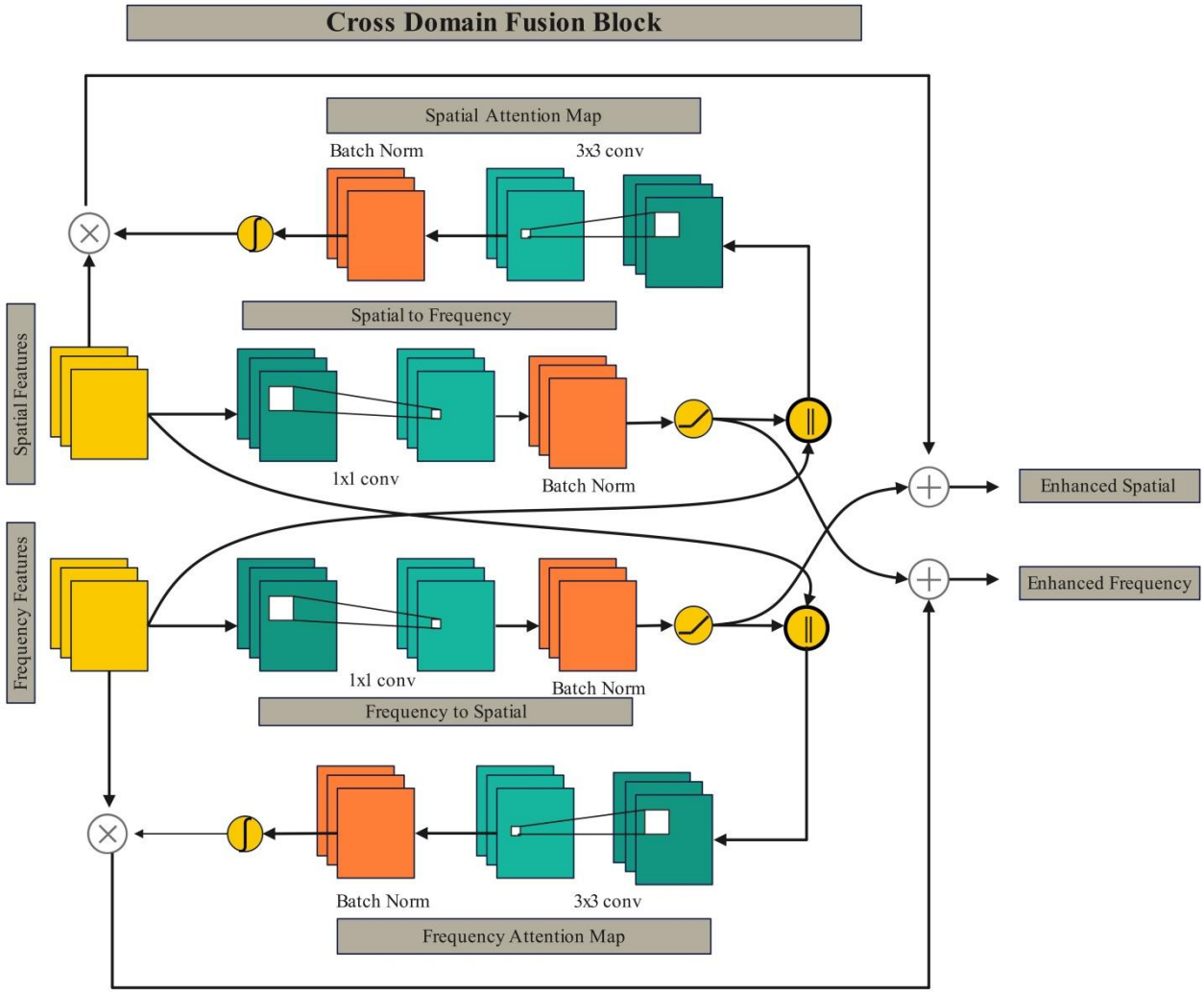

Figure 5: Proposed cross domain feature fusion block RS features

After the first cross-domain fusion block, another set of texture-aware convolutional blocks is incorporated in the architecture to process these enhanced spatial and frequency features. Another cross-domain fusion block follows it to integrate information at higher levels. After that, another texture-aware convolutional block is followed by a cross-domain fusion block integrated into the architecture for higher-level feature extraction and integration. These emphasized features are then passed to corresponding adaptive feature recalibration blocks to recalibrate the feature map based on their channel-wise importance.

4) Adaptive Feature Recalibration and Classification

In this phase, we presented an adaptive feature recalibration (AFR) block that is incorporated in the architecture for parallel processing of spatial domain and frequency domain features. Each block is composed of an adaptive average pooling layer, which summarizes the spatial information to compute channel-wise importance. It is then followed by a linear layer, which reduces the channel dimensions by a factor of 16 to reduce computational cost and learn spatial interconnections. A ReLU activation is applied to introduce non-linearity to handle complex relationships. Another linear layer is used to restore the original channel dimensions and produce weights that are later scaled by a sigmoid activation function. These scaled weights are then multiplied by the feature map, which was passed as an input to the recalibration block to emphasize important channels and features. After channel and feature recalibration, the feature maps of both streams are passed to their corresponding average pooling layers, followed by a flatten layer to reduce spatial dimensions to 1x1 and flatten the pooled features. The outputs of both streams are concatenated to generate a final feature representation containing both spatial and frequency features. This representation is then

passed through the fully connected layer to produce classification results.

B. BLIP-2 based Image Description Generation

To provide a semantic explanation of remote sensing scenes along with simple classification, a pretrained BLIP-2 (Bootstrapped Language Image Pretraining) model is integrated into the proposed framework. BLIP-2 takes the same images as input and generates their corresponding natural language description, working in parallel with DFSNet. The BLIP-3 model consists of three main components: a frozen image encoder, a lightweight Query Transformer (Q-Former), and a frozen Large Language Model (LLM) decoder based on OPT, as shown in Figure 2.

1) Image Encoding:

Given an input remote sensing image $I \in \mathbb{R}^{H\times W\times 3}$, the frozen image encoder $E_\theta(.)$ based on VIT, processes the image and extract visual feature representations. Here, the image is first divided into fixed-size patches, where each patch is then projected to a D-dimensional embedding space. The extracted features are defined as:

$$V = E_\theta(I) \in \mathbb{R}^{N_p\times D} \quad (15)$$

Where, N_p denotes the total number of image patches and D denotes the feature embedding dimension. The image encoder weights are kept frozen during the training to preserve the learned visual representations.

2) Query Transformer (Q-Former):

The Q-Former acts as a lightweight bridge between the frozen image encoder and the frozen LLM decoder. It contains a set of N_q learnable query vectors $Q \in \mathbb{R}^{N_q\times D_q}$ that interact with frozen visual features V through a cross-attention layers to extract the most relevant visual information for text generation. The cross-attention operation between the learned queries and visual features is mathematically defined as:

$$A = SoftMax\left(\frac{QW_q(VW_k)^T}{\sqrt{D_q}}\right)VW_v \quad (16)$$

Where, $W_q \in \mathbb{R}^{D_q\times D_q}$, $W_k \in \mathbb{R}^{D\times D_q}$ and $W_v \in \mathbb{R}^{D\times D_q}$ are learnable projection metrices for queries, keys and values respectively. The output $A \in \mathbb{R}^{N_q\times D_q}$ represents the visually grounded query features that summarize the most relevant scene information from the remote sensing image.

The Q-Former output is then passed through a fully connected projection layer to align the visual query features with the input dimension of the LLM decoder:

$$Z = W_{proj}.A + b_{proj} \quad (17)$$

Where $W_{proj}$and $b_{proj}$are the learnable weight matrix and bias of the projection layer respectively.

3) LLM Decoder for Description Generation

The projected features are passed as visual prompts to the frozen OPT based LLM decoder which generates a natural language description of RS scene. At each generation step t, the LLM decoder predicts the next token w_t based on visual prompt Z and all previously generated tokens $w_{<t}$:

$$P(w_t|Z, w_{<t}) = LLM_{dec}(Z, w_1, w_2, \dots . w_{t-1}) \quad (18)$$

The complete scene description $\hat{T} = \{w_1, w_2, \dots w_L\}$ is generated by sampling tokens sequentially until an end token is produced or the maximum length L is reached. The language modeling loss used to train the Q-Former while keeping both the image encoder and LLM decoder frozen is defined as:

$$L_{LM} = -\sum_{t=1}^{L} logP(w_t|Z, w_{<t}) \quad (19)$$

This loss encourages the Q-Former to extract visual features from the remote sensing image that are most useful for generating accurate and semantically rich scene descriptions.

C. Training of DFSNet

In this section, we discussed about the training process of the proposed DFSNet architecture selected six datasets, which are discussed in Table 1. Each dataset is divided into training and testing sets. In order to train the proposed model, the training set of dataset has been employed. For the DFSNet classification task, a combined loss function is used that consists of Focal loss and Soft Dice loss, where each component is weighted by 0.5. Focal loss is selected to address the class imbalance problem by focusing on hard samples, while Soft Dice loss is included to enhance the model's ability to handle multi-class classification. The combined loss function is defined as:

$$\mathcal{L}_{Total} = 0.5 \times \mathcal{L}_{focal} + 0.5 \times \mathcal{L}_{dice} \quad (20)$$

where $L_{focal} = -\alpha(1-p_t)^\gamma \log(p_t)$with focusing parameter $\gamma = 2.0$and $L_{dice} = 1 - \frac{2\sum p_i g_i}{\sum p_i + \sum g_i}$where $p_i$is the predicted probability and $g_i$is the ground truth.

In order to select hyperparameters during the training phase, we applied Bayesian Optimization (BO) [43] which can dynamically select the most optimal parameters at limited resources.

Bayesian optimization is a hyperparameter optimization technique which is very helpful when objective function is expensive, and the resources are very limited. Suppose we have a function $f(z)$ which needs to be maximized, and z is the input vector. For this purpose, we will built a surrogate model such as Gaussian process which models the objective function $f(z)$ as a probability distribution and provides with the predicted value $\mu(z)$ and uncertainty in the prediction $\sigma^2(z)$ for any value of z. The predictions made by Gaussian process follows a normal distribution and can be defined as:

$$f(z_1), \dots, f(z_n) = \mathbb{N}(\mu, \mathbb{K}) \quad (21)$$

Where $\mu = [\mu(z_1), \dots . \mu(z_n)]$ represents the mean vector and $\mathbb{K}$ represents internal covariate shift where $\mathbb{K}_{x,y} = \kappa(z_x, z_y)$ and,

$$\kappa(z_x, z_y) = \exp\left(-\frac{\|z_x - z_y\|^2}{2\iota^2}\right) \quad (22)$$

Where, ι represents distance between nearby points. Based on previous predictions $\{(z_j, f(z_j))\}_{j=1}^{n}$, GP now predicts the value at any new point z' by providing the prediction $\mu(z')$ and its uncertainty $\sigma^2(z')$ which can be defined as:

$$\mu(z') = \kappa(z', Z)[K(Z, Z) + \sigma_n^2 I]^{-1} v \quad (23)$$
$$\sigma^2(z') = \kappa(z', z') - \kappa(z', Z)[K(Z, Z) + \sigma_n^2 I]^{-1}\kappa(Z, z')$$

Here, I represents identity matrix, $= [z_1, \dots . z_n]$, $v = [f(z_1), \dots, f(z_n)]$ and σ_n^2 represents noise in observations. However, the point to be evaluate next is selected by

acquisition function, Expected Improvement in this case, by balancing the exploration and exploitation; hence, if the best observation so far is $f(z^*)$, then for any new point z, the improvement will be:

$$I(z) = \max\big(0, f(z) - f(z^*)\big) \quad (24)$$

And, as the function is modeled by GP, the expected improvement will be:

$$EI(z) = E[I(z)] = \big(\mu(z) - f(z^*)\big)\Phi\left(\frac{\mu(z)-f(z^*)}{\sigma(z)}\right) + \sigma(z)\phi\left(\frac{\mu(z)-f(z^*)}{\sigma(z)}\right) \quad (25)$$

Where Φ represents cumulative distribution function and ϕ represents probability density function of standard normal. In the next step, new points are selected by maximizing this Expected Improvement. The GP will again predict the value at this new point based on previous predictions, evaluate it, update the function, and check for new points. This procedure continues till we run out of budget, and the point that gives the best values is selected as the hyperparameter. In this case, Bayesian optimization has been performed up to 30 iterations in order to identify the best hyperparameter combinations in the defined search space. The Gaussian process surrogate model converged at around 20 iterations (there was little to no improvement in validation accuracy after that point). The expected improvement acquisition function was used throughout the optimization procedure to provide a balance between exploration of new hyperparameters and exploitation of currently available effective hyperparameters. The given hyperparameter range and hyperparameters selected by the BO are shown in Table 2. Based on these selected hyperparameters, the trained models are finally utilized in the testing phase.

Table 2: Hyperparameter range and selected value using BO

Hyperparameter	Given Range	Selected Value
Number of Epochs	50-150	Early Stopping
Optimizer	Adam, SGD	Adam
Learning rate	0.0001-0.01	0.001
Batch size	16, 32, 64	32

D. Integration of BLIP-2 with DFSNet:

The overall proposed DFSNet-VLM framework performs two tasks simultaneously on remote sensing images as shown in Figure 6. DFSNet processes its input images through its spatial and frequency domains to produce a scene classification label $\hat{y}$, while BLIP-2 processes the respective input image to generate a natural language description $\hat{T}$. The two tasks are independent of each other and do not share parameters, allowing each model to be optimized for its specific task. The final output of the framework for a given input image I for BLIP-2 and X for DFSNet can be summarized as:

$$Output = \{DFSNet(X) \rightarrow \hat{y}, BLIP\_2(I) \rightarrow \hat{T}\} \quad (26)$$

The visual testing framework of proposed DFSNet-VLM is shown in Figure 6.

IV. Experimental Setup and Performance Metrics

All experiments were conducted in Python using the PyTorch library and processed on a workstation with a high-performance server consisting of 2.0 TB of RAM and 16 NVIDIA Tesla V100 GPUs, each with 32 GB of memory. For DFSNet classification, the model was trained using the Adam optimizer with a learning rate of 0.001, batch size of 32, and early stopping patience of 10 epochs. For BLIP-2 description generation, only the Q-Former was fine-tuned while the image encoder and LLM decoder remained frozen, using a learning rate of 1e-5 and batch size of 16. The classification performance is evaluated using the Accuracy (the percentage of correctly classified samples), Precision (the ratio of true positive predictions to total positive predictions for each class), Recall (the ratio of true positive predictions to total ground truth positives for each class), F1-score (the harmonic mean of precision and recall) and AUC scores (the area under the ROC curve measuring the model's ability to distinguish between classes). For description generation, qualitative evaluation is performed by comparing BLIP-2-generated descriptions with ground truth annotations to assess semantic accuracy on RS data.

V. Results and Analysis

The proposed model results are presented in this section. As mentioned under section 2, the proposed model is evaluated on six diverse datasets, including MLRSNet, NWPU-RESISC-45, EuroSAT, Cloud dataset, GeoSceneNet16k, and Bijie-Landslide. A train-test split of 70-30 is performed on all the datasets, and the model is trained and tested on each dataset separately. All the testing results are thoroughly examined and discussed in this section in the form of tables, confusion matrices, ablation studies, and explainability visualization..

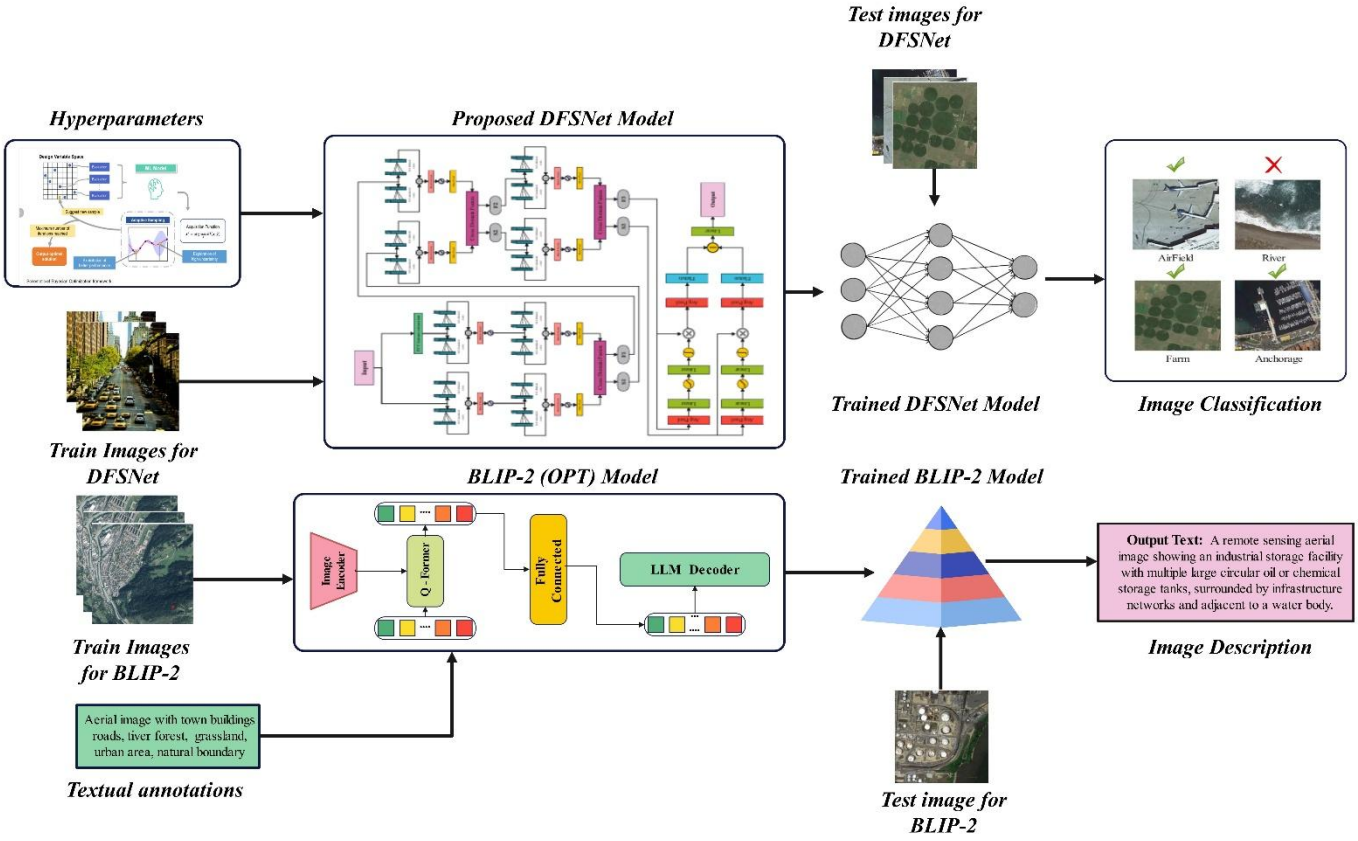

Figure 6: Training and testing framework of the proposed model using selected RS datasets

A. Classification Results

The proposed model results are presented in this section with detailed numerical values and confusion matrices. Results are computed on all six diverse datasets. Table 3 describes the results of the proposed model on the MLRSNet dataset. In this table, the class-wise results are presented, and for each class, precision, recall, and F1-score are computed. The proposed model obtained an average accuracy of 97%, a macro average precision rate of 97%, a recall rate of 97%, and an F1-score of 97%. Also, the weighted precision rate, recall rate, and F1-Score are 97% which shows the proposed model is robust and scalable.

Table 4 presents the proposed model results using the EuroSAT dataset. The obtained average accuracy of this dataset is 98%. The macro average precision rate of this dataset is 98.0, the macro recall rate is 98.0, and the macro F1-score value is 98%, respectively. The same value is obtained for the weighted precision, recall, and F1-score of this dataset. Further analysis of this table shows that each class precision rate is above 95%.

The proposed model results using modified NWPU dataset (contains 12 classes) are presented in Table 5. The average accuracy of this dataset is 95%. The macro average and precision, recall, and F1-Score values are 95.0, 95.0, and 95.0, respectively. Further analysis of this table show that precision rate of Game Space and River class is 91% that is less as compared to other classes. Also, the precision rate of Airfield class is 92% that is second less value after game Space and Rive classes. These values show that how the proposed model shows some difficulty in the correct classification of such classes.

Table 6 describes the proposed model results using the aerial scene recognition dataset GeoSceneNet16K. This dataset contains geospatial classes such as buildings, forest areas, ice glaciers, and the sea. The average accuracy of this dataset is 92%, while the macro precision and weighted precision rates are both 92.0. From this table, it is observed that the Ice Glacier and Hill or Mountain classes have a lower precision rate of 88% compared to other classes in this dataset. The highest precision rate is 98% for the Desert class.

Table 3: Proposed aerial scene recognition results using MLRSNet dataset

Class	Precision	Recall	F1-Score	Support	Class	Precision	Recall	F1-Score	Support
airplane	0.98	0.98	0.98	529	meadow	0.97	0.98	0.97	751
airport	0.96	0.96	0.96	657	mobile_home_park	0.98	0.99	0.98	750
bareland	0.96	0.97	0.97	450	mountain	0.97	0.97	0.97	750

baseball_diamond	0.97	0.99	0.98	601	overpass	0.93	0.94	0.94	750
basketball_court	0.96	0.95	0.95	869	park	0.96	0.95	0.95	505
beach	0.99	0.99	0.99	750	parking_lot	0.99	0.99	0.99	751
bridge	0.96	0.92	0.94	750	parkway	0.98	0.97	0.97	761
chaparral	0.99	0.99	0.99	750	railway	0.95	0.92	0.93	750
cloud	0.99	1.00	0.99	539	railway_station	0.86	0.88	0.87	656
commercial_area	0.96	0.97	0.96	750	river	0.97	0.98	0.97	750
dense_residential_area	0.97	0.99	0.98	835	roundabout	0.96	0.95	0.96	612
desert	0.98	0.97	0.98	762	shipping_yard	0.99	1.00	1.00	750
eroded_farmland	0.97	0.98	0.97	750	snowberg	0.98	0.99	0.99	766
farmland	0.99	0.99	0.99	750	sparse_residential_area	0.97	0.98	0.98	549
forest	0.98	0.99	0.98	750	stadium	0.95	0.92	0.94	739
freeway	0.96	0.98	0.97	750	storage_tank	0.99	0.99	0.99	750
golf_course	0.99	0.98	0.98	754	swimming_pool	1.00	1.00	1.00	601
ground_track_field	0.97	0.95	0.96	750	tennis_court	0.97	0.96	0.97	750
harbor_port	0.98	0.99	0.98	751	terrace	0.98	0.98	0.98	750
industrial_area	0.96	0.95	0.96	633	transmission_tower	0.98	0.98	0.98	750
intersection	0.98	0.97	0.97	749	vegetable_greenhouse	0.99	1.00	1.00	780
island	0.98	0.99	0.98	750	wetland	0.96	0.96	0.96	784
lake	0.97	0.98	0.97	750	wind_turbine	0.98	1.00	0.99	615
Accuracy		**0.97**		**32749**					
Macro Avg	**0.97**	**0.97**	**0.97**	**32749**					
Weighted Avg	**0.97**	**0.97**	**0.97**	**32749**					

Table 4: Proposed aerial scene recognition results using EuroSAT dataset

Class	**Precision**	**Recall**	**F1-Score**	**Support**
AnnualCrop	0.96	0.98	0.97	900
Forest	0.99	1.00	1.00	900
HerbaceousVegetation	0.98	0.98	0.98	900
Highway	0.97	0.96	0.96	750
Industrial	0.99	0.98	0.99	750
Pasture	0.98	0.98	0.98	600
PermanentCrop	0.97	0.96	0.96	750
Residential	0.99	1.00	0.99	900
River	0.98	0.97	0.97	750
SeaLake	0.99	1.00	0.99	900
Accuracy		**0.98**		8100
Macro Avg	0.98	0.98	0.98	8100
Weighted Avg	0.98	0.98	0.98	8100

Table 5: Proposed remote sensing classification results using NWPU (12 classes) dataset

Class	**Precision**	**Recall**	**F1-Score**	**Support**
Airfield	0.92	0.92	0.92	420
Anchorage	0.99	0.98	0.98	210
Beach	0.97	0.96	0.96	210
Dense Residential	0.94	0.98	0.96	210
Farm	0.95	0.97	0.96	420
Flyover	0.95	0.97	0.96	210
Forest	0.98	0.98	0.98	210
Game Space	0.91	0.92	0.92	420
Parking Space	0.98	0.96	0.97	210
River	0.91	0.87	0.89	210
Sparse Residential	0.95	0.96	0.96	210
Storage Cisterns	0.97	0.92	0.94	210
Accuracy			**0.95**	3150
Macro Avg	0.95	0.95	0.95	3150
Weighted Avg	0.95	0.95	0.95	3150

Table 6: Proposed model aerial scene recognition results using GeoSceneNet16K dataset

Class	**Precision**	**Recall**	**F1-Score**	**Support**
Buildings and Structures	0.91	0.88	0.90	657
Desert	0.98	0.98	0.98	600
Forest Area	0.97	0.98	0.98	681
Hill or Mountain	0.88	0.88	0.88	754
Ice Glacier	0.88	0.88	0.88	721
Sea or Ocean	0.94	0.93	0.93	682
Street View	0.90	0.94	0.92	715
Accuracy			**0.92**	4810
Macro Avg	0.92	0.92	0.92	4810
Weighted Avg	0.92	0.92	0.92	4810

Another experiment was conducted on the proposed model using the Bijie-Landslide dataset, which captures information about landslide regions. Table 7 presents the recognition results of this dataset, which achieved an average accuracy of 96.03%. The macro average precision rate, recall rate, and F1-Score value are all 95%, respectively. Also, the weighted precision rate, recall value, and F1-Score value are all 96%, respectively. In this table, the landslide class precision rate is 94%, whereas the recall rate is 92%. This high precision rate and accuracy indicate strong generalization and reliable classification performance of the proposed model.

Table 8 presents the results of cloud scene recognition using remote sensing data. For this purpose, a Cloud dataset is employed, yielding an average accuracy of 98%. Results are presented in the form of class-wise recognition, with the cloud class achieving macro precision of 99% and recall of 98%. These results show that the cloud scene images are identified with higher accuracy and recall. Overall, the results in this table indicate the ability to accurately identify cloud and non-cloud regions for integration into RS pipelines and weather monitoring systems.

Figure 7 illustrates the confusion matrices of all these selected datasets. These confusion matrices can be used to confirm the precision, recall, and F1-score across all chosen datasets. The diagonal values in these confusion matrices show the correct prediction rate of each class. Overall, the model's accuracy is high, and it is highly scalable and generalizable across the selected remote sensing datasets.

Table 7: Proposed model landslide regions classification using Bijie-landslide dataset

Class	**Precision**	**Recall**	**F1-Score**	**Support**
Landslide	0.94	0.92	0.93	231
Non_landslide	0.97	0.98	0.97	601
Accuracy			**0.9603**	832
Macro Avg	0.95	0.95	0.95	832
Weighted Avg	0.96	0.96	0.96	832

Table 8: Proposed model cloud scene recognition using Cloud dataset

Class	**Precision**	**Recall**	**F1-Score**	**Support**
cloud	0.99	0.98	0.98	403
not_cloud	0.97	0.99	0.98	340
Accuracy			**0.98**	743
Macro Avg	0.98	0.98	0.98	743
Weighted Avg	0.98	0.98	0.98	743

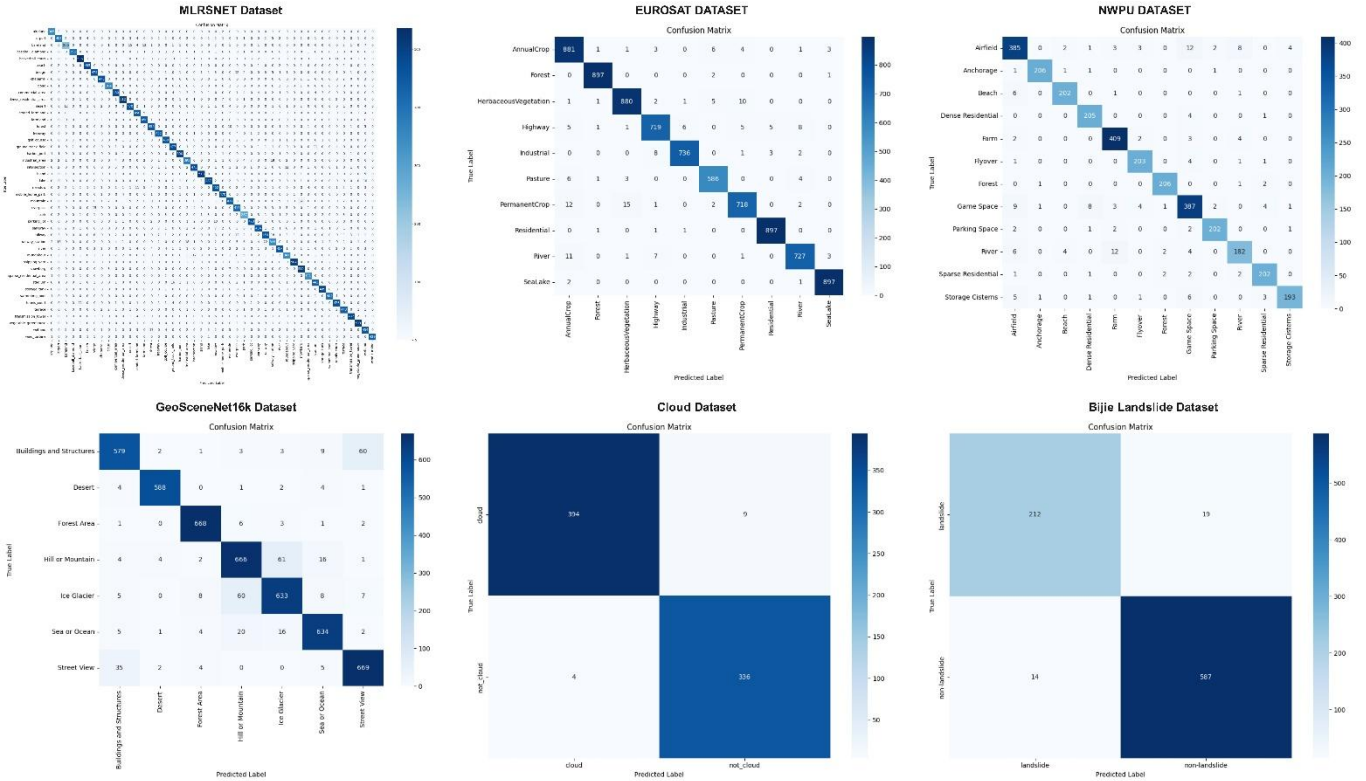

Figure 7: Confusion matrices across all RS datasets for proposed model performance

B. BLIP-2 Description Generation Results

1) Performance Evaluation of BLIP-2 for Remote Sensing Scene Description

The performance of BLIP-2 model for remote sensing scene description generation is presented in Table 9. The model demonstrates strong captioning performance across standard natural language generation metrics. The BLEU-1 score of 0.7824 indicates that approximately 78% of unigrams in the generated descriptions matched with the reference descriptions, showing a significant overlap. The BLEU-4 score of 0.6157 further reflects the model's ability to generate longer descriptions that are well-aligned with ground truth. The METEOR score of 0.6542 and ROUGE-L score of 0.7231 further confirms that generated descriptions have semantic similarity with reference descriptions. The CIDEr score of 1.9745 demonstrates that BLIP-2 effectively generates textual descriptions that are relevant to actual RS scenes. As for the computational efficiency, the model achieves an inference time of 2.34 seconds per image with a throughput of 0.4274 samples per second, which is acceptable for offline analysis tasks. The model complexity shows that only the Q-Former module (188.42M parameters) is trainable while the image encoder and LLM decoder remain frozen, resulting in a total model size of 2.89 GB.

Table 9: Performance Evaluation of BLIP-2 for Remote Sensing Scene Description

BLIP-2 Task	**BLEU-1**	**BLEU-4**	**METEOR**	**CIDEr**	**ROUGE-L**
Scene Description Generation	0.7824	0.6157	0.6542	1.9745	0.6231
Inference Efficiency					
Inference Time (s)			2.34(s)		
Throughput (samples/s)			0.4274(s/s)		
Model Complexity					
Trainable Parameters (Q-Former Only)			188.42 M		
Total Model Size (with frozen encoders)			2.89 GB		

2) Ablation Study on BLIP-2 Fine-tuning Strategies

To analyze the impact of different fine-tuning strategies on BLIP-2 performance for remote sensing scene description, an ablation study was conducted as shown in Table 10. For the zero-shot baseline (pretrained BLIP-2 model is used without any fine-tuning), a BLEU-4 score of 0.2847 and CIDEr score of 1.0234 was observed, which demonstrates the limited effectiveness for domain-specific RS scenes. When only the Q-Former module was fine-tuned while keeping the image encoder and LLM decoder frozen, a substantial improvement was observed in both BLEU-4 (0.6157) and CIDEr (1.9745). This strategy proves most effective because here Q-Former acts as a lightweight bridge that learns to extract visual features from frozen pretrained image encoder while frozen LLM decoder maintains its strong language generation capabilities without overfitting to small dataset. However,

unfreezing the image encoder results in performance degradation (BLEU-4 score : 0.5689, CIDEr :1.8524). This occurs because fine-tuning the large image encoder on a small dataset causes overfitting and disrupts the rich visual representations. Full model fine-tuning, where all components including the LLM decoder are updated, achieves the similar results (BLEU-4: 0.5356, CIDEr: 1.7987) due to poor generalization caused by updating all the parameters which destroys the pretrained knowledge.

The ablation study also examined the learning rate to find the optimal rate that provides the best balance between speed and performance; 1e-5 was found to be the optimum rate because lower rates of 1e-6 slowed the convergence of the evaluations and provided suboptimal results, whereas the higher rates of 1e-4 led to instabilities and reduction in the quality of the descriptions.

Table 10: Ablation Study on BLIP-2 Fine-tuning Strategies

Fine-tuning Strategy	BLEU-1	BLEU-4	METEOR	CIDEr	ROUGE-L
No fine-tuning (zero-shot)	0.5124	0.2847	0.3315	1.0234	0.4192
Q-Former only (selected)	**0.7824**	**0.6157**	**0.6542**	**1.9745**	**0.6231**
Q-Former + Image Encoder	0.7537	0.5689	0.6187	1.8524	0.5954
Full model fine-tuning	0.7312	0.5356	0.5921	1.7987	0.5621
Lower learning rate (1e-6)	0.7021	0.5932	0.5787	1.7734	0.5487
Higher learning rate (1e-4)	0.6734	0.4654	0.5124	1.6042	0.5223

C. Proposed DFSNet Training Graphs

Figure 8 shows the graphs of training accuracy, training loss, validation accuracy, and validation loss over epochs for all the datasets. These graphs show the DFSNet model's trend over epochs and help us understand how epochs affect the proposed model's performance. There are four plots in this figure, each showing accuracy and loss. There are six curves in each plot, each representing the performance of the corresponding dataset using the proposed model during training. These training curves show smooth training losses across all selected datasets, with a constant decrease until 35 epochs. A similar trend with the opposite direction is observed in the graph of training accuracy. However, the graphs of validation accuracy and validation loss exhibit significant fluctuations, indicating the model's struggle with unseen data.

D. Discussion

In this section, the discussion of the proposed DFSNet has focused on AUC Scores, ablation studies, comparisons with SOTA, and explainability results. The proposed model's visual architecture is presented in Figure 2 and was evaluated on six diverse publicly available datasets. Results of each dataset are presented in Tables 3-8. To support the results presented in these tables, confusion matrices are shown in Figure 7. From these results, it is clear that the proposed model's performance across all datasets has been outstanding. However, it is essential to validate the performance of this work through several steps. Therefore, the first step was to calculate AUC Scores across all datasets using different pre-trained models and compare them with the proposed model.

1) AUC Scores

First, the AUC-based analysis has been performed for the proposed model and several highly popular pre-trained deep architectures, including AlexNet, VGG16, VGG19, GoogleNet, ResNet50, and ResNet101, across all selected datasets. The AUC score is the area under the ROC curve, which represents the model's classification performance by plotting True Positive Rate (TPR) against False Positive Rate (FPR). An AUC score of 0.5 indicates random guessing, while an AUC score of 1.00 indicates perfect classification. From the Figure 9, it is observed that the proposed model outperforms all pre-trained models, achieving AUC scores above 0.97 across all datasets, demonstrating its strong classification performance. Specifically, the performance on MLRSNet is highly remarkable: the highest AUC score achieved by a pre-trained model was 0.920, which the proposed model significantly improved to 0.990.

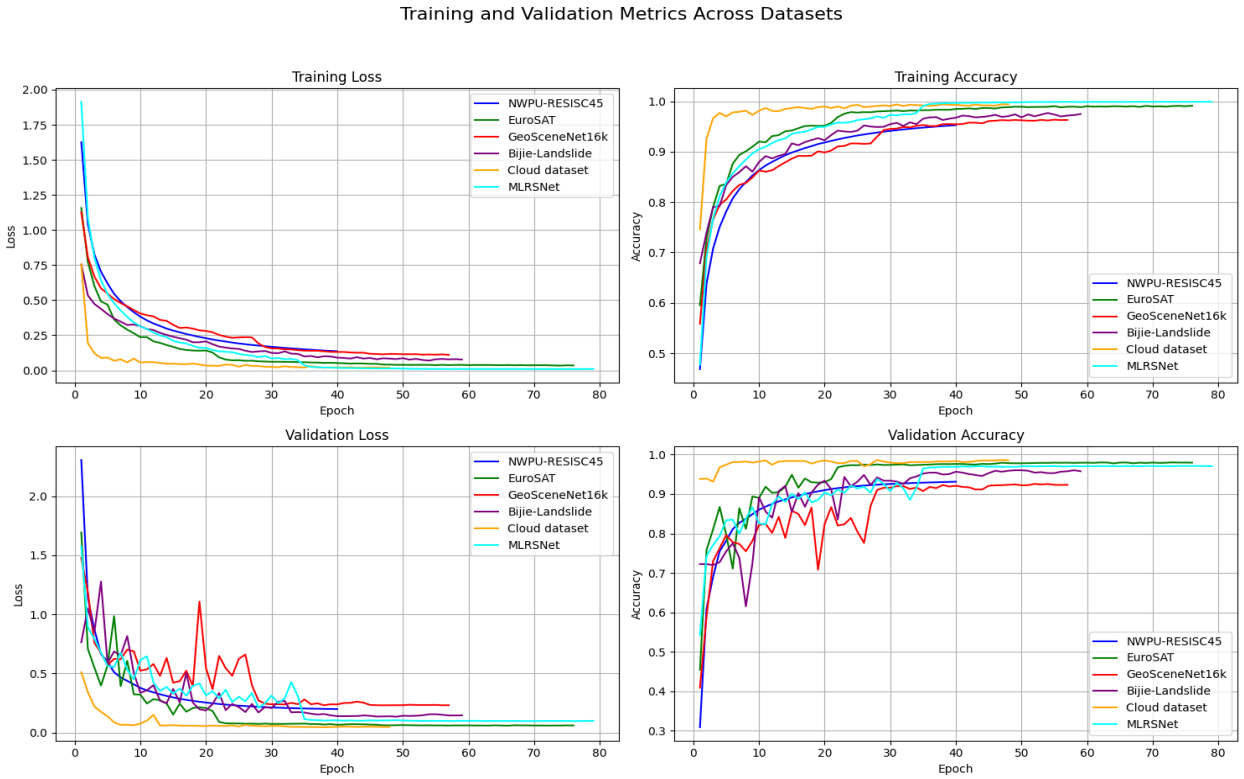

Figure 8: Training graphs of DFSNet architecture across all selected datasets

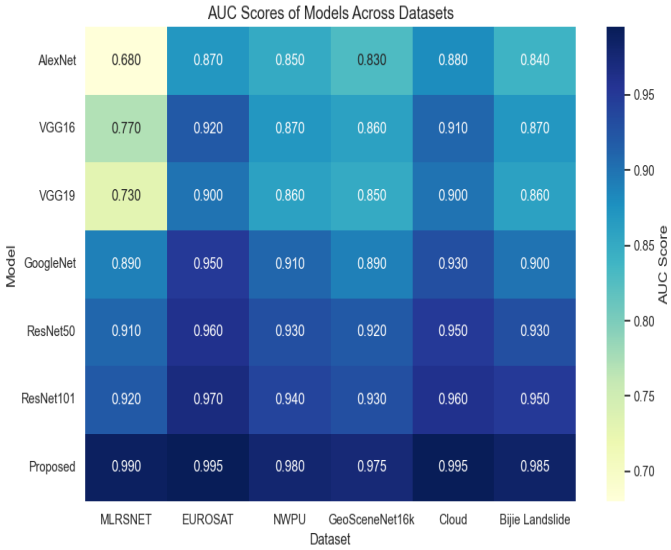

Figure 9: Analysis of proposed model with pre-trained models based on AUC score across all datasets

2) Ablation Studies on DFSNet architecture

During the further analysis of the proposed architecture, we conducted several ablation studies. The purpose of these ablation studies is to analyze the impact of each component in the proposed DFSNet. Ablation studies include removing the frequency-domain stream, replacing the log-scaled FFT with raw magnitudes, simplifying and removing fusion modules and feature recalibration blocks, modifying texture-aware convolutional blocks, and replacing the FFT with DCT (see Table 11). The model is evaluated across all datasets to analyze changes in performance metrics, including accuracy, precision, recall, and F1-Score. Figure 10 illustrates the results of all ablation studies across all datasets. From this figure, it is observed that the proposed model (Base) achieved the highest accuracy across all datasets compared to other components. In addition, it is noted that each element affects precision, especially cross-domain fusion and Recalibration: Simplified Pooling.

Table 11: Ablation studies of the proposed DFSNet using selected components

ID	Ablated Component	Modification Description
Base	—	Full DFS-Net: Spatial + Frequency streams, Fusion, Recalibration, Texture-Aware Conv
A1	Frequency Domain Stream	Removed FFT-based stream entirely
A2	FFT (Magnitude)	Replaced log-scaled FFT with raw magnitude
B1	Cross-Domain Fusion Block	Removed all fusion modules; no interaction between streams

B2	Fusion Simplification	Replaced attention-based fusion with simple addition
C1	Adaptive Feature Recalibration	Removed channel-wise recalibration block
C2	Recalibration: Simplified Pooling	Used global avg pooling instead of MLP-based recalibration
D1	Texture-Aware Conv Block	Replaced with standard 3×3 convolutions
D2	Dilated Convs Only	Used only dilated convolutions in texture-aware blocks
E1	Dual-Stream Design	Removed frequency stream (spatial only)
E2	Early Fusion Only	Merged spatial and frequency features at input only
F1	Single-Level Fusion	Fusion applied only at final layer
I1	FFT → DCT Replacement	Replaced FFT with Discrete Cosine Transform

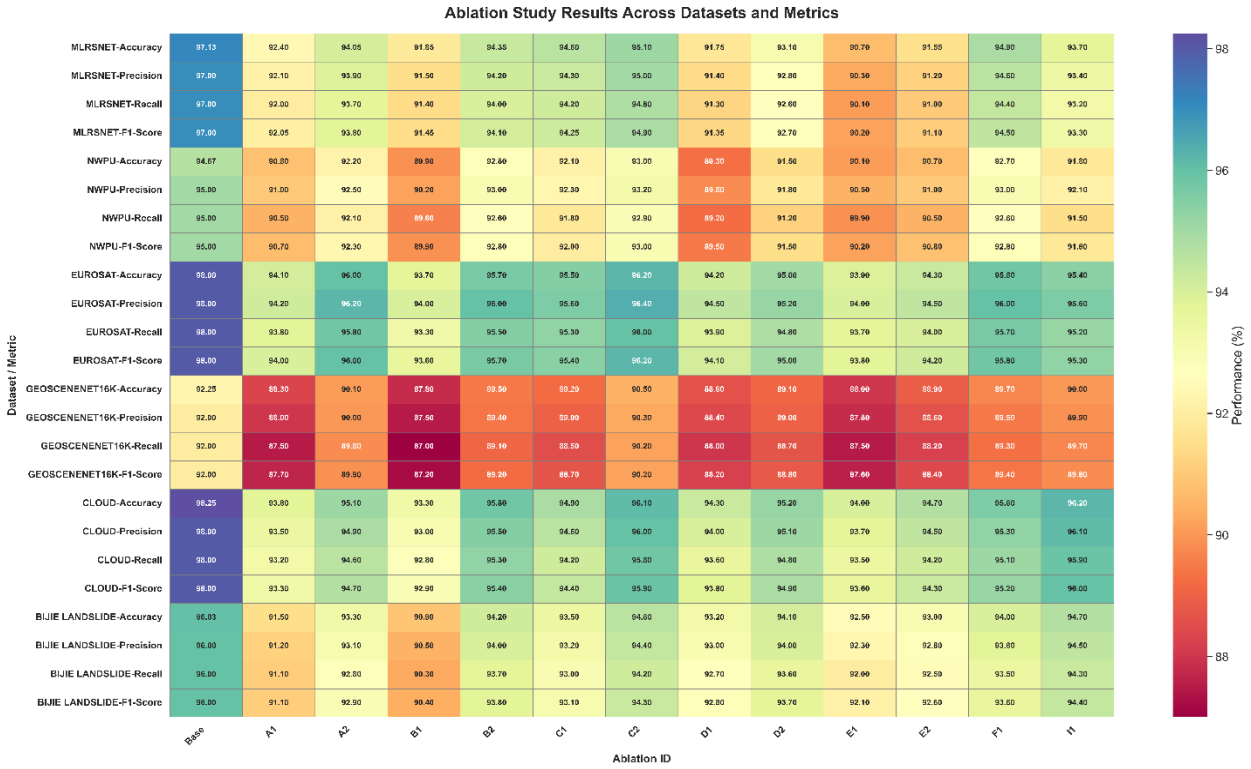

Figure 10: Ablation studies results across all RS datasets

3) Comparative Analysis with Pretrained Models

In this section, further analysis of the proposed model using popular pre-trained architectures, including AlexNet, VGG16, VGG19, GoogleNet, ResNet50, and ResNet101, is conducted across all the considered datasets. Figure 11 shows the analysis of this ablation study. In this figure, four plots show performance metrics—accuracy, precision, recall, and F1-score—for all selected pre-trained and proposed models. Based on these plots, the proposed model performs better than the other listed models across accuracy, precision, recall, and F1-score. The difference in peak heights indicates a significant improvement in performance compared to other models.

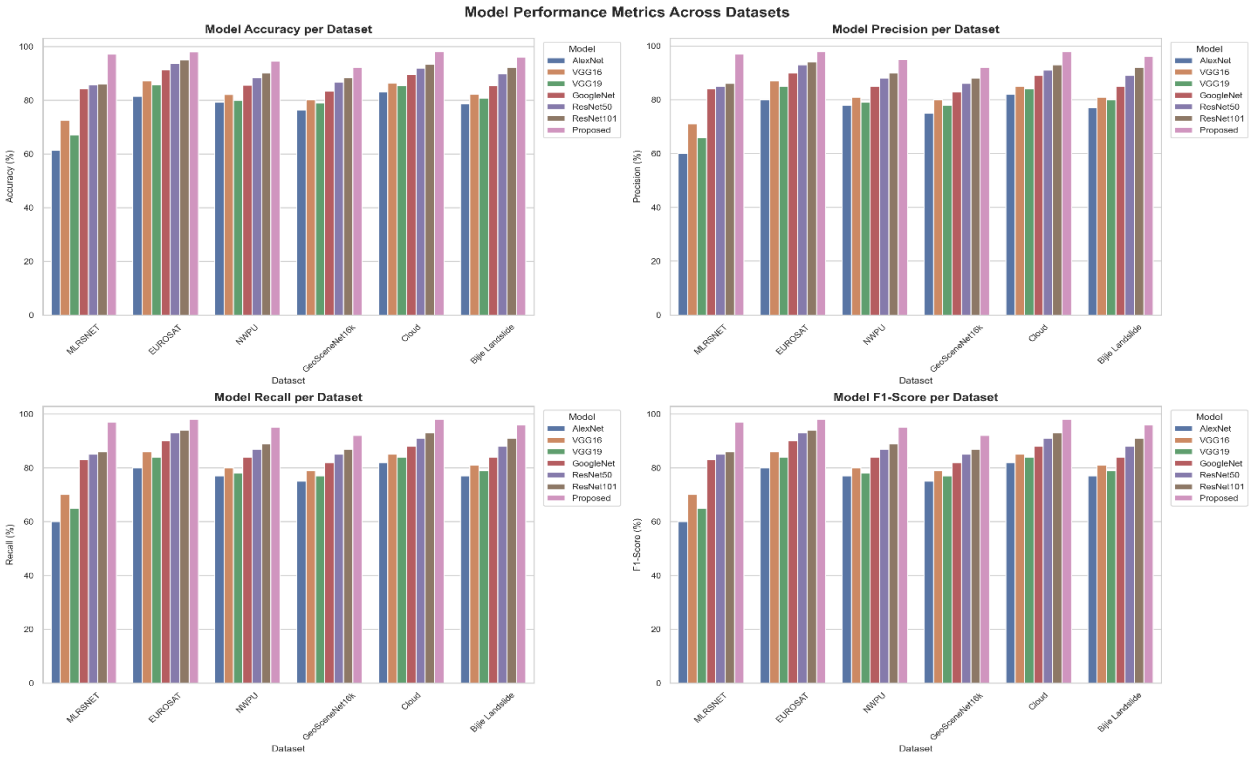

Figure 11: Comparison of proposed model performance with pre-trained deep architectures

4) Comparative Analysis with SOTA Models

The proposed model's performance is also compared with SOTA RS models on selected datasets (see Table 12). The comparison is conducted to assess its effectiveness relative to the average accuracy of other models. There are more than 23 recent models compared with DFSNet on the MLRSNet dataset, and the results show that DFSNet achieved a remarkable accuracy of 97.13% which is way higher than the highest accuracy achieved by existing models. The existing popular models such as FMA-Net [26] obtained an accuracy of 91.0, AMEGRF-Net [44] achieved 91.51, and Yang X. et al. [45] obtained 82.59%, respectively. Other models computed mF1, and overall, the proposed model shows improved performance. Similarly, DFSNet is compared with other existing techniques on the NWPU EuroSAT datasets. The proposed architecture shows significant improvement in the accuracy across these datasets.

Table 12: Comparative analysis with SOTA models using selected RS datasets

MLRSNet dataset comparison		
Architecture	**Accuracy**	**mF1**
FMA-Net [26]	91.0	
AMEGRF-Net [44]	91.51	
Yang X. et al. [45]	82.59	
VGG16 [46]	-	68.01
VGG16 + SSM [46]	-	72.44
VGG16 + SRBM [46]	-	71.26
VGG16 + SR-NET [46]	-	73.80
VGG19 [46]	-	67.10
VGG19 + SSM [46]	-	72.91
VGG19 + SRBM [46]	-	70.12
VGG19 + SR-Net [46]	-	73.33
ResNet50 [46]	-	85.68
ResNet50 + SSM [46]	-	86.58
ResNet50 + SRBM [46]	-	86.07
ResNet50 + SR-Net [46]	-	87.21
ResNet101 [46]	-	86.05
ResNet101 + SSM [46]	-	86.92
ResNet101 + SRBM [46]	-	87.71
ResNet101 + SR-Net [46]	-	87.55
DenseNet201 [46]	-	86.17
DenseNet201 + SSM [46]	-	86.56
DenseNet201 + SRBM [46]	-	86.26
DenseNet201 + SR-Net [46]	-	87.36
proposed	**97.13%**	

NWPU dataset comparison	
Architecture	**Accuracy**
Khan J.A et al. [25]	93.3
EAM [47]	93.04
WSADAN-ResNet50 [48]	92.63

BestC [49]	95.28
Albarakati, H.M., et al. [50]	91.7
Proposed	**98.00%**
EuroSAT dataset comparison	
Global Optimal structured loss [51]	88.68
EfficientNet [52]	85.23
MobileNetV2 [53]	87.52
InceptionV1 [54]	88.51
Proposed	**92.25%**

5) Generalizability Analysis

To demonstrate the generalizability of the proposed DFSNet-VLM framework across diverse remote sensing datasets and scene types, Table 13 summarizes the overall performance metrics. The consistent high performance across six diverse datasets with varying scene types, spatial resolutions, and sample sizes demonstrates the strong generalizability and robustness of the proposed framework. The model maintains accuracy above 92% across all datasets despite significant domain variations, confirming its scalability for real-world remote sensing applications.

6) Proposed Model Explainability

The proposed model explainability is evaluated under this section using Grad-CAM++ and Grad-CAM++ overlay. The trained models (different datasets) are applied to test images (unseen images), predictions are generated, and the results are then visualized using Grad-CAM techniques. Figure 12 illustrates the visual output of the proposed model and shows that the heatmap is generated effectively within the exact region of interest. Moreover, it is also noted that the Grad-CAM++ overlay performed well in terms of better explainability of visual features. For example, the heatmap is generated in a non-cloud region, identifies highways, identifies rivers, and identifies airplanes with higher probability scores. Based on these visualization results, the proposed model clearly demonstrated strong explainability.

Table 13: Generalizability Analysis of proposed DFSNet Model on RS data

Dataset	No. of Classes	DFSNet Accuracy (%)	Avg Precision (%)	Avg Recall (%)	Avg F1-Score (%)
MLRSNet	46	97.0	97.0	97.0	97.0
NWPU-RESISC45	12	95.0	95.0	95.0	95.0
EuroSAT	10	98.0	98.0	98.0	98.0
GeoSceneNet16k	7	92.0	92.0	92.0	92.0
Cloud	2	98.0	98.0	98.0	98.0
Bijie-landslide	2	96.03	96.0	96.0	96.0
Average across all datasets		96.0	96.0	96.0	96.0

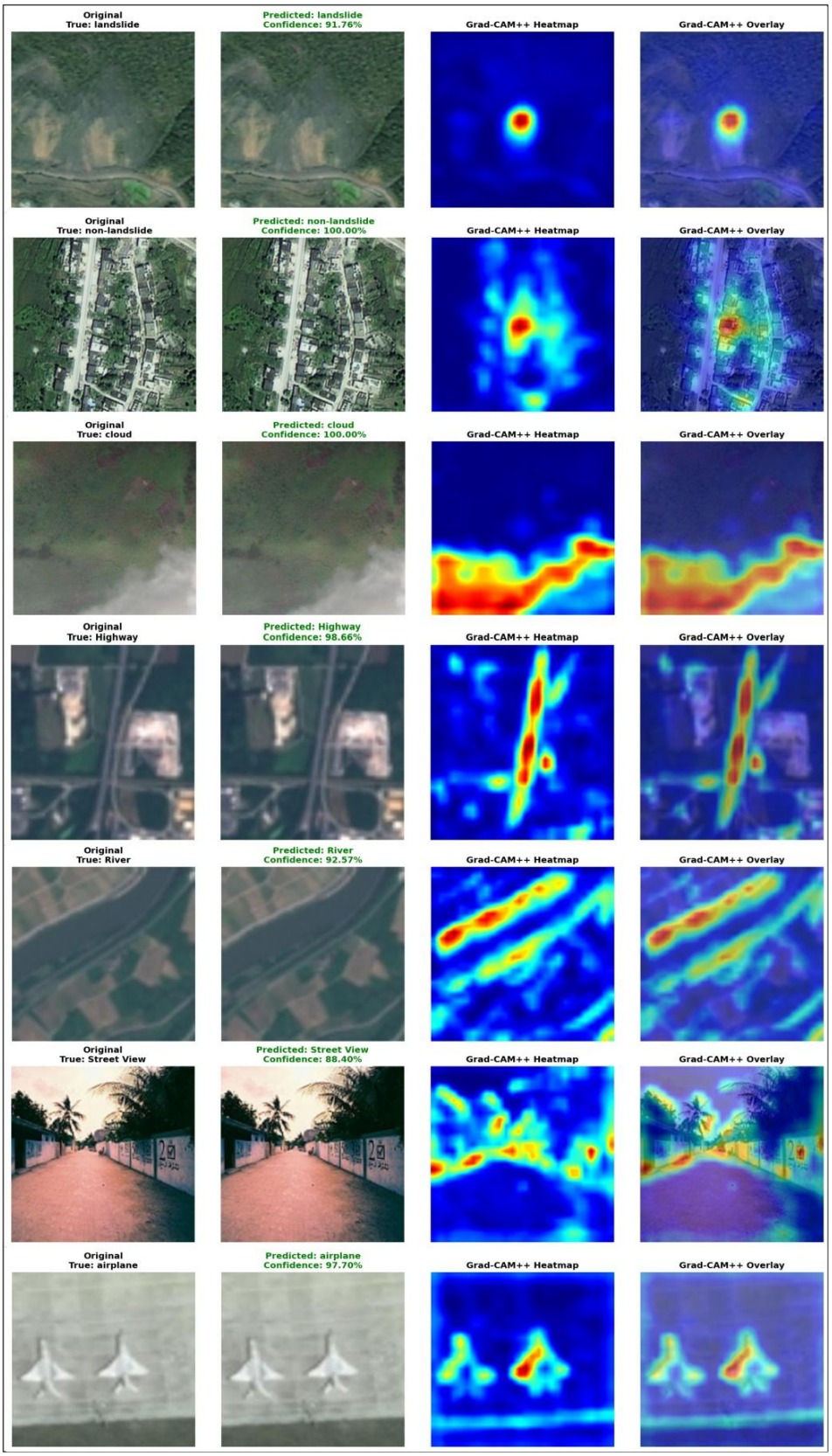

Figure 12: Proposed model explainability results on test images using Grad-CAM++

VI. Conclusion

This paper presented a novel unified framework, DFSNet-VLM, that combines a dual-stream deep learning architecture with vision-language modeling for comprehensive remote sensing scene understanding. The proposed DFSNet architecture leverages spatial and frequency-domain features to extract information about local and global patterns and textures in RS images. In contrast, the integrated BLIP-2 VLM generates natural-language descriptions to provide semantic explanations of classified scenes. The model also incorporated an attention-based cross-domain fusion block that exchanges information between the two DFSNet streams at multiple stages. During the training phase, the hyperparameters of the proposed DFSNet model are initialized using Bayesian Optimization (BO). In the experimental process of the proposed model, six diverse datasets are used for evaluation, achieving improved accuracy of 97.13% on MLRSNet, 94.67% on NWPU-RESISC-45, 98.00% on EuroSAT, 92.25% on GeoSceneNet16k, 98.25% on cloud, and 96.03% on the Bijie-landslide dataset, respectively. Detailed ablation studies, comparative analyses with pre-trained / SOTA models, and ROC plots demonstrate the generalization and scalability of the proposed model across selected datasets. In addition, the model's explainability on unseen images shows how well the learning process has been conducted on the training data. Overall, the proposed model shows improved performance across diverse publicly available datasets. The key findings of proposed work are:

- DFSNet's frequency-spatial domain fusion achieves 96.06% average accuracy across six diverse datasets, outperforming state-of-the-art CNN-based models for remote sensing scene classification.
- Fine-tuning only BLIP-2's Q-Former while freezing the image encoder and LLM decoder proves most effective for remote sensing description, achieving BLEU-4 of 0.6157 with 116% improvement over zero-shot performance.
- The unified DFSNet-VLM framework successfully combines accurate classification with natural language explanation, enhancing interpretability and semantic understanding in remote sensing applications.

Despite the promising results, certain limitations are present. The BLIP-2 component of the framework was evaluated on a limited dataset of 180 image-text pairs, which limits its generalization capability. Additionally, the current framework processes only RGB imagery and does not evaluate on multispectral or multitemporal data. Future work will focus on creating larger-scale multimodal remote sensing datasets with rich textual annotations and fine-tuning BLIP-2 with domain-specific remote sensing captions to improve description quality.

Conflict of Interest

All authors declared no conflict of interest.

Dataset Availability

The datasets of this work are publically available for the research purposes. Moreover, MLRSNET Trained Model: https://www.kaggle.com/models/jownabbas/dfsnet-mlrsnet-trained-model
NWPU Trained Model: https://www.kaggle.com/models/jownabbas/dfsnet-nwpu-trained-model
EUROSAT Trained Model: https://www.kaggle.com/models/jownabbas/dfsnet-eurosat-trained-model
GeoSceneNet16k Trained Model: https://www.kaggle.com/models/jownabbas/dfsnet-geoscenenet16k-trained-model
Cloud Trained Model: https://www.kaggle.com/models/jownabbas/dfsnet-cloud-trained-model
Bijie Trained Model: https://www.kaggle.com/models/jownabbas/dfsnet-bijielandslide-trained-model/

XI. References

[1] S. Chen, X. Wang, X. Wei, Y. Sun, and K. Yang, "Deeply understanding features to achieve efficient remote sensing image classification," *Expert Systems with Applications,* vol. 295, p. 128743, 2026.

[2] X. Cui and L. Zhang, "MTMixer: a hybrid Mamba-Transformer architecture for multimodal remote sensing image classification," *International Journal of Remote Sensing,* vol. 47, no. 2, pp. 770-799, 2026.

[3] M. Smith and C. Pain, "Applications of remote sensing in geomorphology," *Progress in Physical Geography,* vol. 33, no. 4, pp. 568-582, 2009.

[4] R. P. Sishodia, R. L. Ray, and S. K. Singh, "Applications of remote sensing in precision agriculture: A review," *Remote sensing,* vol. 12, no. 19, p. 3136, 2020.

[5] A. M. Lechner, G. M. Foody, and D. S. Boyd, "Applications in remote sensing to forest ecology and management," *One Earth,* vol. 2, no. 5, pp. 405-412, 2020.

[6] H. Han, Z. Liu, J. Li, and Z. Zeng, "Challenges in remote sensing based climate and crop monitoring: navigating the complexities using AI," *Journal of cloud computing,* vol. 13, no. 1, pp. 1-14, 2024.

[7] M. Heymann and A. Dahan Dalmedico, "Epistemology and politics in Earth system modeling: Historical perspectives," *Journal of Advances in Modeling Earth Systems,* vol. 11, no. 5, pp. 1139-1152, 2019.

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

[9] 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.

[10] R. Fan, L. Wang, R. Feng, and Y. Zhu, "Attention based residual network for high-resolution remote sensing imagery scene classification," in *IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium*, 2019: IEEE, pp. 1346-1349.

[11] 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.

[12] C. van Coller, "A decision support tool for implementing machine learning in SME manufacturing companies," *Diss. Reutlingen University, Germany,* 2024.

[13] F. Ma, Y. Feng, F. Zhang, and Y. Zhou, "Cloud adversarial example generation for remote sensing image classification," *IEEE Transactions on Geoscience and Remote Sensing,* 2025.

[14] A. Albahri *et al.*, "A systematic review of trustworthy artificial intelligence applications in natural disasters," *Computers and Electrical Engineering,* vol. 118, p. 109409, 2024.

[15] G. Obaido *et al.*, "Supervised machine learning in drug discovery and development: Algorithms, applications, challenges, and prospects," *Machine Learning with Applications,* vol. 17, p. 100576, 2024.

[16] J. Guo, S. Jiao, H. Sun, B. Song, and Y. Chi, "Cross-modal Compositional Learning for Multi-label Remote Sensing Image Classification," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing,* 2025.

[17] H. Song, J. Xie, Y. Duan, X. Xie, Y. Zhou, and W. Wang, "Cmkd-net: A cross-modal knowledge distillation method for remote sensing image classification," *Advances in Space Research,* 2025.

[18] Y. Liang, S. Cao, J. Zheng, X. Zhang, J. Huang, and H. Fu, "Low Saturation Confidence Distribution-based Test-Time Adaptation for cross-domain remote sensing image classification," *International Journal of Applied Earth Observation and Geoinformation,* vol. 139, p. 104463, 2025.

[19] A. Radford *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*, 2021: PmLR, pp. 8748-8763.

[20] L. Tao *et al.*, "Advancements in vision–language models for remote sensing: Datasets, capabilities, and enhancement techniques," *Remote Sensing,* vol. 17, no. 1, p. 162, 2025.

[21] J. Li, D. Li, S. Savarese, and S. Hoi, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in *International conference on machine learning*, 2023: PMLR, pp. 19730-19742.

[22] X. Li, C. Wen, Y. Hu, Z. Yuan, and X. X. Zhu, "Vision-language models in remote sensing: Current progress and future trends," *IEEE Geoscience and Remote Sensing Magazine,* vol. 12, no. 2, pp. 32-66, 2024.

[23] X. Weng, C. Pang, and G.-S. Xia, "Vision-language modeling meets remote sensing: Models, datasets, and perspectives," *IEEE Geoscience and Remote Sensing Magazine,* 2025.

[24] C. Kopidaki, G. Tsagkatakis, and P. Tsakalides, "Federated learning for remote sensing image classification using sparse image representations," *IEEE Geoscience and Remote Sensing Letters,* 2025.

[25] 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.

[26] 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.

[27] H. M. Albarakati *et al.*, "A Unified Super-Resolution Framework of Remote Sensing Satellite Images Classification based on Information Fusion of Novel Deep Convolutional Neural Network Architectures," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing,* 2024.

[28] M. Fayaz, J. Nam, L. M. Dang, H.-K. Song, and H. Moon, "Land-cover classification using deep learning with high-resolution remote-sensing imagery," *Applied Sciences,* vol. 14, no. 5, p. 1844, 2024.

[29] 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.

[30] 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.

[31] K. Madala and M. Siva Ganga Prasad, "Crop Mapping Using Advanced Deep Learning Framework in Remote Sensing Data," *International Journal of Image and Graphics,* p. 2750016, 2025.

[32] S. Sudha and S. Aji, "An Intelligent Two-Stage Hybrid Hierarchical Classification using Optimized Label Propagation for Remote Sensing Image Retrieval," *Journal of the Indian Society of Remote Sensing,* pp. 1-17, 2025.

[33] M. Aljebreen, H. A. Mengash, M. Alamgeer, S. S. Alotaibi, A. S. Salama, and M. A. Hamza, "Land use and land cover classification using river formation dynamics algorithm with deep learning on remote sensing images," *IEEE Access,* vol. 12, pp. 11147-11156, 2024.

[34] Y. Hu, J. Yuan, C. Wen, X. Lu, Y. Liu, and X. Li, "Rsgpt: A remote sensing vision language model and benchmark," *ISPRS Journal of Photogrammetry and Remote Sensing,* vol. 224, pp. 272-286, 2025.

[35] S. Wang *et al.*, "Llm4hrs: Llm-based spatio-temporal imputation model for highly-sparse remote sensing data," *IEEE Transactions on Geoscience and Remote Sensing,* 2025.

[36] 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.

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

[38] 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.

[39] S. Yang *et al.*, "Automatic identification of landslides based on deep learning," *Applied Sciences,* vol. 12, no. 16, p. 8153, 2022.

[40] M. A. Haque, R. H. Rifat, M. Kamal, R. George, K. D. Gupta, and K. Shujaee, "CDD & CloudNet: A Benchmark Dataset & Model for Object Detection Performance," in *2024 Fifth International Conference on Intelligent Data Science Technologies and Applications (IDSTA)*, 2024: IEEE, pp. 118-122.

[41] X. Zi *et al.*, "RSVLM-QA: A Benchmark Dataset for Remote Sensing Vision Language Model-based Question Answering," in *Proceedings of the 33rd ACM International Conference on Multimedia*, 2025, pp. 12905-12911.

[42] J. Zhang and Y. Tu, "SwinFR: Combining SwinIR and fast fourier for super-resolution reconstruction of remote sensing images," *Digital Signal Processing,* vol. 159, p. 105026, 2025.

[43] R. Abdelfattah, K. Abdelfatah, M. M. Fouda, Z. M. Fadlullah, M. Abouyoussef, and M. I. Ibrahem, "Bayesian Optimization-Aided Hybrid Deep Learning Model for Lightweight UAV-Based Smoke Detection," *IEEE Internet of Things Journal,* 2025.

[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] 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*.

[48] 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.

[49] W. Hu, C. Lan, T. Chen, S. Liu, L. Yin, and L. Wang, "Scene Classification of Remote Sensing Image Based on Multi-Path Reconfigurable Neural Network," *Land,* vol. 13, no. 10, p. 1718, 2024.

[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] 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.

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

[53] 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.

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

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.

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.

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.

Mehrez Marzougui was born in Kasserine, Tunisia, in 1972. He received the B.Sc. degree from the University of Tunis, Tunis, Tunisia, in 1996, and the M.Sc. and Ph.D. degrees from the University of Monastir, Monastir, Tunisia, in 1998 and 2005, respectively, all in electronics. From 2001 to 2005, he was a Research Assistant with Electronics and Micro-Electronics Laboratory. From 2006 to 2012, he was an Assistant Professor with Electronics Department, University of Monastir. Since 2013, he has been an Assistant Professor with Engineering Department, College of Computer Science, King Khalid University, Abha, Saudi Arabia. He is the author of more than 30 articles. His research interests include hardware/software cosimulation, image processing, and multiprocessor systems on chips.

Areej Alasiry received the B.Sc. degree in information systems from King Khalid University, Abha, Saudi Arabia, and the M.Sc. degree (Hons.) in advanced information systems and the Ph.D. degree in computer science and information systems from Birkbeck College, University of London, U.K., in 2010 and 2015, respectively. She is currently an Assistant Professor at the College of Computer Science, King Khalid University. She also holds the position of the College Vice Dean for Graduate Studies and Scientific Research. Her main research interests include machine learning and data science.

Jungpil Shin (Senior Member, IEEE) received the B.Sc. degree in computer science and statistics, and the M.Sc. degree in computer science from Pusan National University, Busan, South Korea, in 1990 and 1994, respectively, and the Ph.D. degree in computer science and communication engineering from Kyushu University, Fukuoka, Japan, in 1999, under a scholarship from the Japanese Government (MEXT). He was an Associate Professor, Senior Associate Professor, and Full Professor with the School of Computer Science and Engineering, The University of Aizu, Aizuwakamatsu, Japan, in 1999, 2004, and 2019, respectively. He has coauthored more than 500 published papers for widely cited journals and conferences. His research interests include pattern recognition, image processing, computer vision, machine learning,

human–computer interaction, non-touch interfaces, human gesture recognition, automatic control, Parkinson's disease diagnosis, ADHD diagnosis, user authentication, machine intelligence, bioinformatics, and handwriting analysis, recognition, and synthesis. He is a member of ACM, IEICE, IPSJ, KISS, and KIPS. He was an Editorial Board Member of Scientific Reports. He was included among the top 2% of scientists worldwide edition of Stanford University/Elsevier, in 2025, 2024, and the top 0.5% of AI researcher worldwide edition of ScholarGPS, in 2024. He was the General Chair, Program Chair, and Committee member of numerous international conferences. He was the Editor of IEEE journals, Springer, Sage, Taylor and Francis, Sensors (MDPI), Electronics (MDPI), and Tech Science. He was a Reviewer of several major IEEE and SCI journals.

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.