Jurnal Ilmu Komputer dan Informasi (Journal of Computer Science and Information)
19/1 (2026), 11-32. DOI: http://dx.doi.org/10.21609/jiki.v19i1.1475

MSDFF-RCNet: A Combined Multi-Structure Data Fusion Framework and Recurrent Attention for Remote Sensing Scene Classification

Yohanes Fridolin Hestrio[1,2], Bayu Satria Persada[1], Frederic Morado Saragih[1],
Muhammad Yusuf Kardawi[1], Wisnu Jatmiko[1], Aniati Murni Arymurthy[1]

[1] Faculty of Computer Science, Universitas Indonesia, Depok, Indonesia
[2] Research Center for Geoinformatics, Research Organization for Electronics and Informatics,
National Research and Innovation Agency (BRIN), Bogor, Indonesia

Email:yohanes.fridolin31@ui.ac.id

Abstract

Remote sensing scene classification faces significant challenges in distinguishing visually similar land-use categories due to high intraclass variation and interclass similarity in high-resolution imagery. Although deep learning approaches have shown promise, single-architecture methods often fail to capture the diverse spatial and hierarchical features required for robust scene discrimination. This study proposes MSDFF-RCNet, a multi-structure data fusion framework combined with recurrent attention mechanisms to enhance remote sensing scene classification performance. The framework integrates complementary feature representations from AlexNet, ResNet50, and DenseNet161 architectures, while the recurrent attention mechanism focuses on discriminative spatial regions for improved classification accuracy. Comprehensive experiments conducted on four benchmark datasets demonstrate substantial performance improvements over the baseline ARCNet architecture: UC Merced (43.8% to 84.9%, +41.1%), AID (63.8% to 94.4%, +30.6%), NWPU-RESISC45 (61.5% to 95.4%, +33.9%), and OPTIMAL 31 (47.3% to 87.9%, +40.6%). Statistical significance analysis confirmed the reliability of these improvements ($p < 0.01$), while comprehensive evaluation across precision, recall, and F1-score metrics validated the framework's robustness. Although the multi-structure approach requires substantial computational resources (25.6× parameter increase), the consistent and significant accuracy improvements across diverse datasets demonstrate the effectiveness of complementary feature fusion for remote sensing scene classification. The proposed framework provides a valuable contribution to automated Earth observation systems that require high-precision land-use classification capabilities.

Keywords: *Feature Fusion, ResNet, DenseNet, AlexNet, Remote Sensing, Scene Classification*

1. Introduction

Advancements in remote sensing technology have enabled the acquisition of very high-resolution (VHR) images, which provide detailed representations of the Earth's surface. The increasing availability of such high-resolution imagery necessitates the development of robust techniques for effective scene classification [1]. However, processing and analyzing the vast and complex VHR data remain significant challenges in the field.

Remote sensing scene classification involves assigning semantic labels to satellite images based on predefined categories, such as residential areas, agricultural land, forests, and other land cover types [2]. Early research in this domain relied primarily on pixel-based classification methods, as low-resolution images contained pixels large enough to represent entire objects. As remote sensing technology advanced and spatial resolution increased, the focus shifted from pixel-based classification to object-based classification, ultimately leading to scene-level classification, where entire landscapes are analyzed holistically.

Despite its advantages, scene-level classification presents several fundamental challenges, particularly high intraclass variability and interclass similarity. For instance, the shape, color, and architectural structure of churches can vary significantly, making it difficult to define a uniform classification pattern within the same category. Similarly, distinct categories, such as forests and grasslands or highways and bridges, may exhibit overlapping spectral and structural features, leading to frequent misclassification [3]. Additional complexities arise from factors such as variations in satellite imaging conditions, the presence of multiple objects within a single image, and atmospheric disturbances like cloud cover, all of which further complicate the classification process.

Traditional feature-based approaches for scene classification have relied on handcrafted low-level features, such as texture descriptors, color gradients, and shape characteristics. Common techniques include the Scale-Invariant Feature Transform (SIFT) [4], Gabor filters [4], color histograms [5], Gray Level Co-occurrence Matrix (GLCM) [6], and Histogram of Oriented Gradients (HOG) [7]. Although these methods effectively extract basic patterns from satellite images, they lack the ability to capture complex spatial relationships and hierarchical feature representations within scenes. Intermediate-level features, obtained through clustering, segmentation, or feature grouping techniques, have attempted to improve classification accuracy, but they remain limited in their ability to generalize across diverse scene types and imaging conditions.

The advent of deep learning, particularly convolutional neural networks (CNNs), has brought significant advancements to scene classification tasks. CNNs can automatically extract hierarchical features from images, enabling better representation learning without manual feature engineering. The introduction of AlexNet, which achieved breakthrough results in the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) [8], marked a paradigm shift in deep learning-based image classification. Various advanced architectures, including ResNet [9], LSTM-based models, attention mechanisms, and class activation mapping [10], have been introduced to incorporate spatial context and improve performance in VHR image analysis.

However, single CNN architectures often struggle to handle the complexity inherent in remote sensing images, particularly in differentiating visually similar classes and dealing with high intraclass variations. A single model may excel in capturing low-level textures and patterns but fails to learn the high-level spatial relationships and semantic information necessary for accurate scene classification. This limitation is particularly pronounced in remote sensing imagery, where scene categories like ”industrial area” and ”commercial area,” or ”medium residential” and ”dense residential,” may share many visual characteristics despite belonging to different semantic classes.

To address this research gap, this study proposes MSDFF-RCNet (Multi-Structure Data Fusion Framework with Recurrent Convolutional Network), which integrates three complementary CNN architectures—ResNet50, DenseNet161, and AlexNet—to leverage their distinct strengths in feature extraction. The proposed method can capture a more diverse set of spatial, textural, and hierarchical features by fusing feature representations from multiple models with different architectural designs and learning capabilities, leading to improved classification accuracy and robustness.

This study investigates whether such a multi-architecture fusion approach can enhance classification performance across multiple benchmark datasets, particularly in scenarios where class separability is challenging due to intra-class variations and inter-class similarities. We hypothesize that combining the complementary strengths of multiple architectures will result in more robust and discriminative feature representations, thereby improving the ability of the model to distinguish between visually similar classes in complex remote sensing scenarios.

The effectiveness of this approach is comprehensively evaluated through extensive experiments on four benchmark datasets: UC Merced, AID, NWPU-RESISC45, and OPTIMAL 31, providing a thorough assessment of the model’s robustness, generalizability, and practical applicability across different remote sensing contexts.

The main contributions of this study are as follows:

1) Proposing MSDFF-RCNet, a novel multi-structure feature fusion framework that integrates ResNet50, DenseNet161, and AlexNet architectures to enhance remote sensing scene classification performance.
2) Implementing feature fusion within the AR-CNet framework to improve feature extraction quality and spatial attention mechanisms.
3) Introducing data transformations and augmentation strategies at each feature extraction stage to refine the classification process and improve generalization.
4) Providing comprehensive evaluation using multiple performance metrics (accuracy, precision, recall, F1-score) across four benchmark datasets to demonstrate the effectiveness, limitations, and practical applicability of the proposed approach.

The remainder of this paper is organized as follows: Section 2 reviews related work in remote sensing scene classification and feature fusion techniques; Section 3 details the proposed methodology and system architecture; Section 4 presents the experimental setup, datasets, and evaluation metrics; Section 5 provides comprehensive results and discussion; and Section 6 concludes the study with future research directions.

2. Related Work

2.1. Attention Recurrent Convolutional Network

The attention recurrent convolutional network (ARCNet) represents a significant advancement in convolutional neural network design, specifically developed for processing remote sensing imagery with very high resolution (VHR). ARCNet employs a sophisticated recurrent attention mechanism to enable adaptive focus on crucial regions within images, allowing the network to analyze only relevant, high-level features while effectively filtering out insignificant background information [9].

The ARCNet architecture uses established CNN models, such as AlexNet, VGGNet, or ResNet, as foundational high-level feature extractors. The core innovation lies in the recurrent attention structure, which serves as the fundamental algorithmic component. This structure is designed to generate multiple attentional representations by leveraging the high-level features extracted from RS images. The attention representation is created through a matrix mask that maintains dimensions identical to those of the high-level feature maps, enabling precise spatial attention control [9].

ARCNet incorporates a recurrent neural network (RNN) component, specifically using Long Short-Term Memory (LSTM) architecture, to serve as a sequential representation processor. The generates a sequence of attentional representations that undergo systematic processing. The interdependence between LSTM layers creates a feedback mechanism where the output of each layer influences the input of subsequent layers, enabling the supervision signals to be dynamically adjusted. The final classification is performed using a softmax layer that processes the attention-refined features to predict scene categories for remote sensing images [9].

2.2. Multi-structure data fusion framework

The MSDFF represents an advanced classification methodology specifically designed for remote sensing image analysis. The framework addresses the fundamental challenge of effectively utilizing complementary features generated by different CNN architectures. The original MSDFF implementation combined three well-established CNN models CaffeNet, VGG-16, and GoogLeNet to improve classification accuracy for high-resolution remote sensing scene categorization tasks [11].

The MSDFF operational mechanism involves comprehensive data augmentation strategies, including aggressive scaling, cropping, and rotation applied to each data sample to generate diverse training instances. Feature extraction is performed using three distinct CNN architectures, with each network extracting its specialized informative features from input images. The use of multiple CNNs provides the advantage of capturing diverse aspects and complementary characteristics inherent in each architectural design. The integration of features retrieved from multiple CNNs is accomplished through sophisticated feature fusion networks, which constitute the core component of the MSDFF framework [11].

The feature fusion process incorporates fully connected layers with softmax activation functions to create a unified feature representation system. The primary objective of feature fusion is to obtain a single, comprehensive feature vector that ideally exhibits greater discriminative power than its individual constituent features, thereby facilitating improved classification performance. The fused feature vector is processed through a softmax layer following feature integration for final classification. This layer generates probability distributions across all scene categories, with the class exhibiting the highest probability being selected as the classification output [11].

2.3. Residual Network

Residual Networks (ResNet), introduced by He et al. [12], represent a groundbreaking advancement in deep learning architecture design. ResNet architectures consist of fundamental components, including skip connections, fully connected layers, and convolutional layers that form the network's core structure. The innovative skip connections enable the network to bypass one or more layers, allowing the gradient flow to propagate directly across layer boundaries without passing through intermediate processing layers. This architectural innovation addresses the vanishing gradient problem that has traditionally limited the depth of neural networks.

ResNet encompasses multiple architectural variants, including ResNet50, ResNet101, ResNet152, ResNet50V2, and ResNet101V2, among others. The number of layers incorporated in each architecture is the primary distinction between these variants, with deeper networks generally providing enhanced representational capacity [12].

ResNet's advantages have contributed to its widespread adoption in deep learning research applications. The skip connection mechanism enables training of significantly deeper networks compared to traditional CNN architectures by effectively addressing gradient vanishing issues. ResNet architectures consistently achieve superior accuracy compared to conventional CNN designs, demonstrating excellent performance across various computer vision tasks, including image classification, object detection, and semantic segmentation. The residual block design facilitates faster training processes by reducing the number of convergence iterations. Additionally, ResNet architectures provide superior generalization capabilities, enabling effective performance on previously unseen data. Pre-trained ResNet models serve as excellent foundations for transfer learning applications, where networks can be adapted for specific tasks using smaller, domain-specific datasets [12].

2.4. AlexNet

AlexNet, developed by Krizhevsky et al. [8], represents a pivotal CNN architecture that achieved breakthrough performance in the ImageNet Large Scale Visual Recognition Challenge (ILSVRC). The AlexNet architecture comprises approximately 60 million parameters and 650,000 neurons, demonstrating enhanced learning capacity through increased network depth and sophisticated parameter optimization methodologies [8].

The architectural design of AlexNet incorporates several distinctive characteristics that contribute to its revolutionary performance. The network structure consists of 5 convolutional layers and 3 fully connected layers, totaling 8 primary processing layers. This configuration alternates between convolution operations and resolution reduction to capture hierarchical features at multiple scales. AlexNet employs the ReLU activation function throughout the network, effectively mitigating vanishing gradient problems that plagued earlier architectures. The network also incorporates batch normalization techniques to improve training stability and convergence. Resolution reduction is achieved through max-pooling operations that systematically decrease the spatial dimensions of feature maps generated by convolutional processes [8].

2.5. Densely Connected CNNs

Densely Connected Convolutional Networks (DenseNet), introduced by Huang et al. [13], represent an innovative convolutional neural network (CNN) architecture that implements dense connectivity patterns between network layers. Each layer in DenseNet receives feature maps from all preceding layers and contributes its own feature maps to all subsequent layers, creating a densely connected network topology. This architectural design promotes efficient feature propagation and reuse throughout the network structure.

The DenseNet architecture comprises two primary components: dense blocks and transition layers. Each convolutional layer within each dense block maintains connections to all other layers in the same block, ensuring maximum information flow. Transition layers are positioned between dense blocks to manage feature map dimensions and enable effective network scaling. These layers typically include batch normalization, activation functions, and pooling operations to control the growth of feature map sizes [13].

DenseNet offers several significant advantages that have contributed to its widespread adoption in computer vision applications. The dense connectivity pattern effectively alleviates vanishing gradient problems by providing direct gradient flow paths throughout the network. The architecture strengthens feature propagation by ensuring that each layer has access to features from all preceding layers. Feature reuse is encouraged through dense connections, resulting in more efficient parameter use and reduced computational requirements. Additionally, DenseNet architectures typically require fewer parameters than traditional CNN designs while maintaining competitive performance. These characteristics make DenseNet particularly suitable for image classification, object detection, and semantic segmentation tasks [13].

3. Proposed Method

3.1. Architecture Overview

This study proposes that using advanced feature extraction architectures can significantly enhance the classification accuracy of remote sensing scene analysis. Furthermore, the combination of multiple feature extractors with varying architectural depths and design philosophies enables a more comprehensive and robust feature extraction process. The proposed approach captures a broader spectrum of visual information by integrating features from different types of CNNs, ultimately improving overall classification performance and system robustness.

The mathematical foundation of the proposed approach is as follows. Given an input image $I \in$

$\mathbb{R}^{H \times W \times C}$, where H, W, and C represent the height, width, and channels, respectively, the multistructure feature extraction process can be expressed as:

$$\mathbf{F}_{alex} = f_{alex}(I; \theta_{alex}), \quad \mathbf{F}_{alex} \in \mathbb{R}^{4096} \tag{1}$$

$$\mathbf{F}_{res} = f_{res}(I; \theta_{res}), \quad \mathbf{F}_{res} \in \mathbb{R}^{2048} \tag{2}$$

$$\mathbf{F}_{dense} = f_{dense}(I; \theta_{dense}), \quad \mathbf{F}_{dense} \in \mathbb{R}^{2208} \tag{3}$$

where f_{alex}, f_{res}, and f_{dense} represent the feature extraction functions of AlexNet, ResNet50, and DenseNet161, respectively, with corresponding parameters θ_{alex}, θ_{res}, and θ_{dense}.

To achieve this objective, a comprehensive two-stage methodology is introduced. The first stage involves integrating the ARCNet architecture without structural modifications, maintaining its proven attention mechanisms and recurrent processing capabilities. The second stage enhances the feature extraction process by implementing a sophisticated fusion strategy that combines features from three complementary CNN architectures: ResNet50, DenseNet161, and AlexNet. The complete methodology is illustrated in Figure 1.

The proposed architecture consists of three interconnected key components that work synergistically to improve classification performance:

1) **Multi-Architecture Feature Extraction Layers**: The input image is processed in parallel using three distinct CNN architectures ResNet50, DenseNet161, and AlexNet to extract diverse and complementary feature representations. ResNet50 captures deep hierarchical spatial relationships through its residual connections, DenseNet161 ensures efficient feature propagation and reuse through dense connectivity patterns, and AlexNet focuses on fundamental textural patterns and basic visual features that serve as foundational elements for scene understanding.
2) **Feature Fusion using MSDFF**: The extracted features from the three networks undergo sophisticated fusion through the Multi-Structure Data Fusion Framework (MSDFF). This process involves adaptive 2D average pooling for dimensional standardization, precise dimension alignment to ensure compatibility, and strategic concatenation operations to ensure that complementary features from different extractors are effectively combined while preserving their individual strengths and characteristics.
3) **Enhanced Recurrent Attention Mechanism**: The fused features are processed through an enhanced recurrent attention mechanism, inherited and adapted from the ARCNet architecture, which improves the ability of the network to focus on discriminative regions within remote sensing images. This attention mechanism refines feature representations by emphasizing critical spatial and textural details while suppressing irrelevant background noise and redundant information.

To implement this comprehensive approach, the ARCNet model is systematically modified by restructuring its feature extraction process and incorporating advanced feature fusion techniques. The modification process comprises three essential components: multi-architecture feature extractor fusion, data transformation and augmentation strategies, and optimized data splitting procedures. Each component significantly contributes to refining the learning capabilities of the model and optimizing the overall classification performance across diverse remote sensing scenarios.

3.2. Feature extractor fusion

The feature extractor fusion stage represents a crucial component of the research methodology. The process begins with obtaining pre-trained models for each feature extractor, specifically ResNet50, DenseNet161, and AlexNet. The utilization of pre-trained models in this study is strategically motivated by their proven ability to capture diverse visual features from large-scale datasets, providing a robust foundation for transfer learning to remote sensing applications.

The mathematical formulation of the fusion process can be expressed as follows. First, the feature vectors from the three architectures are normalized and dimensionally aligned:

$$\mathbf{F}_{fused} = \text{Concat}(\mathbf{F}_{alex}, \mathbf{F}_{res}, \mathbf{F}_{dense}) \tag{4}$$

where $\mathbf{F}_{fused} \in \mathbb{R}^{8352}$ represents the concatenated feature vector. The fused features then undergo two fully connected layer transformations:

$$\mathbf{F}_{refined} = \text{ReLU}(\mathbf{W_2} \cdot \text{ReLU}(\mathbf{W_1} \cdot \mathbf{F}_{fused} + \mathbf{b_1}) + \mathbf{b_2}) \tag{5}$$

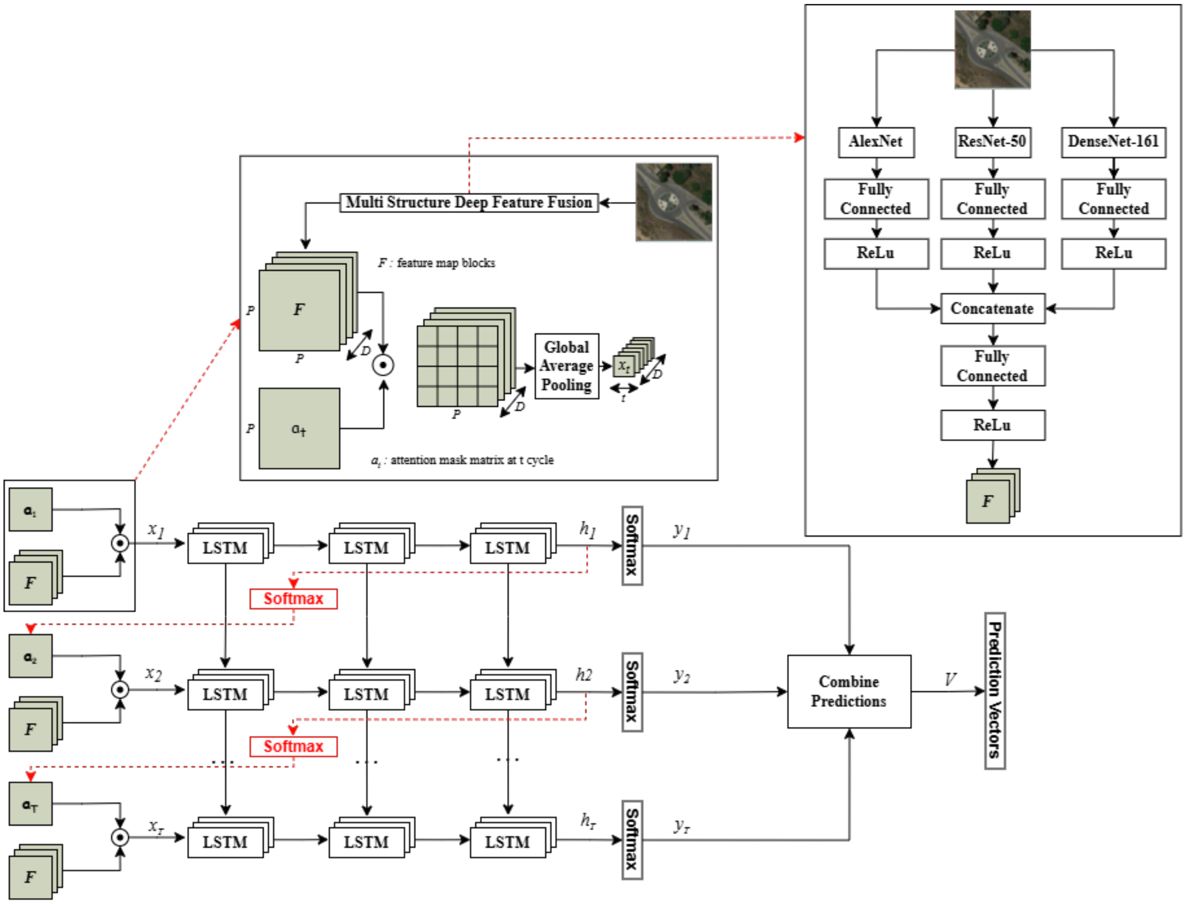


Figure 1. Proposed Multi-Structure Data Fusion Framework with Recurrent Attention. The proposed architecture integrates three CNN feature extractors (ResNet50, DenseNet161, and AlexNet) with MSDFF fusion and ARCNet attention mechanisms for enhanced remote sensing scene classification.

where $\mathbf{W_1}$, $\mathbf{W_2}$, $\mathbf{b_1}$, and $\mathbf{b_2}$ are learnable parameters.

The selection of these three specific architectures is based on their complementary strengths and distinct approaches to feature extraction:

1) **AlexNet Architecture** [8]: Serves as the baseline feature extractor due to its relatively straightforward architecture consisting of 5 convolutional layers and 3 fully connected layers. AlexNet extracts fundamental visual features, including edges, textures, and basic color patterns, that are essential for initial scene characterization. The architecture produces feature vectors of dimension 4096 from its final fully connected layer before classification, providing a comprehensive representation of low-level visual characteristics.
2) **ResNet50 Architecture** [12]: Introduces innovative skip connections that enable the network to learn residual functions, effectively addressing the vanishing gradient problem inherent in deep networks. This architecture excels at capturing hierarchical and complex spatial relationships within remote sensing scenes, which is particularly valuable for distinguishing between visually similar categories that differ in subtle spatial arrangements. ResNet50 generates feature vectors of dimension 2048 from its final average pooling layer, providing rich representations of spatial hierarchies.
3) **DenseNet161 Architecture** [13]: Employs a sophisticated dense connectivity pattern in which each layer receives inputs from all preceding layers, promoting efficient feature reuse and improving gradient flow throughout the network. This architecture ensures optimal feature propagation and enhances the ability of the network to cap-

ture fine-grained details and subtle variations within scene categories. DenseNet161 produces feature vectors of dimension 2208 from its final pooling layer, offering detailed representations of local and global features.

By integrating these three architectures with fundamentally different depths and design philosophies, the proposed framework can extract and utilize a more comprehensive set of features than would be achievable with any single architecture. Collectively, these complementary features improve the model's ability to distinguish between classes characterized by high intraclass variation and challenging interclass similarities.

The theoretical justification for this multi-architecture approach is based on the diversity-accuracy trade-off principle in ensemble learning. Each architecture contributes unique feature representations that complement the others, resulting in improved overall performance. The mathematical analysis of this complementarity can be expressed using the correlation coefficient between feature vectors from different architectures:

$$\rho_{i,j} = \frac{\mathrm{Cov}(\mathbf{F}_i, \mathbf{F}_j)}{\sigma_{\mathbf{F}_i}\sigma_{\mathbf{F}_j}} \tag{6}$$

where $i, j \in \{alex, res, dense\}$ and $i \neq j$. Lower correlation values indicate higher diversity, which theoretically leads to better ensemble performance.

The pre-training process uses the MIT Places365 dataset [14], which contains over 10 million images spanning more than 400 unique scene categories. This dataset is specifically designed for scene recognition tasks and provides comprehensive coverage of diverse environmental and man-made scenes, making it suitable for transfer learning to remote sensing applications despite potential domain differences.

Although datasets such as UC Merced and OPTIMAL 31 are specifically tailored for remote sensing imagery, the general features learned from MIT Places365 including pattern recognition capabilities, texture analysis, and spatial structure understanding remain highly relevant and can significantly enhance a model's ability to classify remote sensing images. Therefore, despite the potential differences in data distribution between natural scenes and satellite imagery, leveraging pre-trained models trained on MIT Places365 offers substantial advantages. These models have already developed rich and diverse feature representations, which can be effectively transferred and fine-tuned for remote sensing classification tasks.

The feature fusion process follows a four-step approach:

1) **Feature Extraction**: Each input image undergoes parallel processing through AlexNet, ResNet50, and DenseNet161 to extract feature vectors of dimensions 4096, 2048, and 2208, respectively. This parallel processing ensures that different aspects of the image are captured simultaneously.
2) **Dimensionality Standardization**: Adaptive 2D average pooling ensures consistent dimensions across features from different networks, facilitating effective integration without information loss.
3) **Feature Concatenation**: The processed feature vectors undergo strategic concatenation to form a unified, high-dimensional representation that preserves each architecture's individual strengths.
4) **Feature Integration**: Two fully connected layers with appropriate activation functions further refine the fused features, producing the final comprehensive feature representation that serves as input to the recurrent attention mechanism in ARCNet.

The feature vectors from different architectures undergo dimensionality alignment through adaptive average pooling operations. Specifically, AlexNet's 4096-dimensional features, ResNet50's 2048-dimensional features, and DenseNet161's 2208-dimensional features are processed through individual fully connected layers before concatenation, resulting in a unified feature representation of 8352 dimensions ($4096 + 2048 + 2208$). This high-dimensional representation captures each architecture's complementary strengths while maintaining computational efficiency through strategic dimensionality reduction in subsequent layers.

This sophisticated fusion strategy enables the model to leverage the distinct strengths of each architecture: AlexNet's proficiency in capturing basic visual elements, ResNet50's capability for learning hierarchical spatial relationships, and DenseNet161's efficiency in feature propagation and reuse. The resulting integrated feature representation provides a more comprehensive and discriminative view of the input image, thereby improving the overall classification performance and robustness across various remote sensing scenarios.

3.3. Enhanced Recurrent Attention

The enhanced recurrent attention mechanism builds upon the ARCNet foundation but incorporates the multi-structure fused features as input. The attention mechanism can be mathematically expressed as follows:

Given the refined feature representation $\mathbf{F}_{refined}$ from the fusion process, the attention mechanism generates a sequence of attention weights through multiple LSTM layers. The attention weight at time step t is computed as:

$$\mathbf{a}_t = \text{softmax}(\mathbf{W}_{att}\mathbf{h}_t + \mathbf{b}_{att}) \tag{7}$$

where $\mathbf{h}_t$ is the hidden state of the LSTM at time step t, and $\mathbf{W}_{att}$ and $\mathbf{b}_{att}$ are learnable parameters. The attended feature representation is then computed as:

$$\mathbf{F}_{attended} = \sum_{t=1}^{T} \mathbf{a}_t \odot \mathbf{F}_{refined} \tag{8}$$

where $\odot$ denotes element-wise multiplication and T is the number of attention steps. The final classification is performed using:

$$\mathbf{y} = \text{softmax}(\mathbf{W}_{cls}\mathbf{F}_{attended} + \mathbf{b}_{cls}) \tag{9}$$

where $\mathbf{W}_{cls}$ and $\mathbf{b}_{cls}$ are the classification layer parameters, and $\mathbf{y}$ represents the predicted class probabilities.

3.4. Data Transformation and Splitting

In the data transformation and splitting stage, comprehensive preprocessing strategies are implemented to enhance model training effectiveness and ensure robust performance evaluation. The data augmentation process involves multiple transformation techniques designed to increase dataset diversity and improve the generalizability of the model.

The augmentation pipeline includes random cropping operations with a standardized size of 256×256 pixels, random horizontal and vertical flips to account for orientation invariance, rotation transformations to handle viewing angle variations, and comprehensive normalization procedures to ensure consistent input distributions. Collectively, these transformations create a more diverse training dataset that better represents real-world imagery variations.

The mathematical formulation of the data augmentation process can be expressed as a transformation composition:

$$I_{aug} = T_n \circ T_{n-1} \circ \ldots \circ T_1(I) \tag{10}$$

where T_i represents the individual transformation operations (i.e., cropping, flipping, rotation, and normalization), and I_{aug} is the augmented image.

The dataset partitioning strategy employs a carefully designed split: 80% for training data to ensure sufficient learning samples, 15% for validation data to monitor training progress and prevent overfitting, and 5% for test data to provide unbiased performance evaluation. This distribution balances the need for comprehensive training with robust evaluation capabilities.

The expanded dataset is used in the training process to construct both the baseline ARCNet model and the enhanced MSDFF-RCNet framework. The validation data serve as a continuous monitoring mechanism throughout the training process, enabling early stopping and optimization of the hyperparameters. The test data provide a final, unbiased evaluation of model performance across all datasets, ensuring a reliable assessment of the effectiveness of the proposed methodology.

3.5. Computational complexity analysis (CCA)

The proposed method's computational complexity can be analyzed in terms of both time and space complexity. The time complexity for feature extraction from the three parallel networks is:

$$O(T_{total}) = O(T_{alex}) + O(T_{res}) + O(T_{dense}) \tag{11}$$

where T_{alex}, T_{res}, and T_{dense} represent the computational times of AlexNet, ResNet50, and DenseNet161, respectively. The fusion process adds the following complexity:

$$O(T_{fusion}) = O(d_{alex} + d_{res} + d_{dense}) + O(d_{fused}^2) \tag{12}$$

where $d_{alex} = 4096$, $d_{res} = 2048$, $d_{dense} = 2208$, and $d_{fused} = 8352$.

The space complexity is dominated by the storage requirements for the multiple network parameters:

$$O(S_{total}) = O(S_{alex}) + O(S_{res}) + O(S_{dense}) + O(S_{fusion}) \tag{13}$$

This analysis provides a theoretical foundation for understanding the computational trade-offs involved in the proposed multi-structure approach.

4. Experiments

4.1. Dataset Description

The experimental evaluation employs a single-label classification framework, where each image corresponds to exactly one scene category. This approach eliminates multi-label classification scenarios and ensures that no image belongs to multiple classes simultaneously. Comprehensive dataset analysis revealed no class overlap exists within individual images, with each image distinctly representing a specific scene type. This characteristic significantly simplifies the classification task by eliminating ambiguity and overlapping label challenges.

The experimental setup, as illustrated in Figure 1, employs the same modified ARCNet architecture across all datasets: UC Merced, AID, NWPU-RESISC45, and OPTIMAL 31. This consistent architectural application ensures fair comparative analysis while maintaining identical structural configurations across all experimental conditions.

Each dataset undergoes processing through the Multi-Structure Data Fusion Framework (MSDFF) with Recurrent Attention, ensuring uniform evaluation conditions across different remote sensing scene classification tasks. This standardized approach facilitates comprehensive performance analysis and enables the assessment of the generalizability of the model across diverse data distributions and imaging conditions.

4.1.1. UC Merced Dataset. The UC Merced dataset consists of satellite images with 256 × 256 pixel resolution, manually collected from satellite sources and obtained from the United States Geological Survey National Map. The dataset encompasses 21 distinct scene categories, including agricultural areas, airplanes, baseball diamonds, beaches, buildings, chaparral, dense residential areas, forests, freeways, golf courses, harbors, intersections, medium-density residential areas, mobile home parks, overpasses, parking lots, rivers, runways, sparse residential areas, storage tanks, and tennis courts. Each category contained 100 images, resulting in a total of 2,100 images. The primary challenge presented by this dataset lies in the significant visual similarities between certain scene categories [9]. Figure 2 presents representative examples from the UC Merced dataset.

4.1.2. Aerial Image Dataset (AID). The AID represents a comprehensive collection of satellite images specifically designed for scene categorization tasks. Each image has a resolution of 600×600 pixels, and the complete dataset contains 10,000 images across 30 distinct categories. The AID dataset includes airports, bare land, baseball fields, beaches, bridges, centers, churches, commercial areas, dense residential zones, deserts, farmland, forests, industrial areas, meadows, medium residential areas, mountains, parks, parking areas, playgrounds, ponds, ports, railway stations, resorts, rivers, schools, squares, stadiums, storage tanks, and viaducts. Each image was professionally annotated by domain experts to ensure classification accuracy [9]. Figure 3 shows representative examples from the AID dataset.

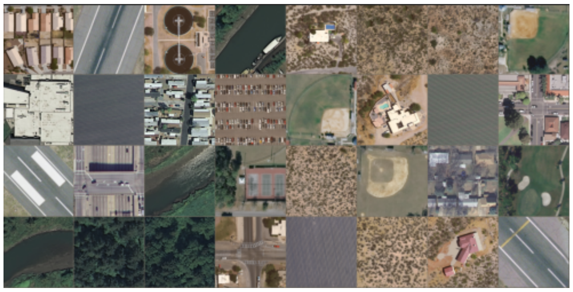


Figure 2. Representative examples from the UC Merced dataset showing diverse scene categories, including agricultural areas, residential zones, and infrastructure elements.

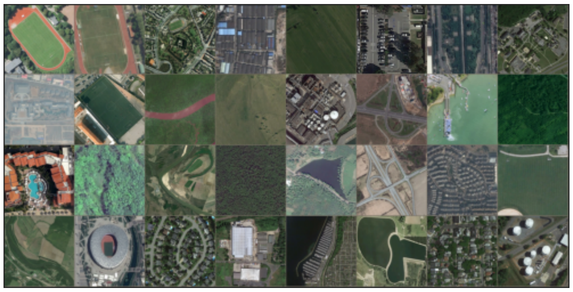


Figure 3. Representative examples from the Aerial Image Dataset (AID) demonstrating the diversity of scene categories, including natural landscapes, urban structures, and transportation infrastructure.

4.1.3. NWPU-RESISC45 Dataset. The NWPU-RESISC45 dataset, developed by Northwestern Polytechnical University, consists of 31,500 images, each with 256 × 256 pixel resolution. The dataset encompasses 45 comprehensive scene categories, including airplanes, airports, baseball diamonds, basketball courts, beaches, bridges, chaparrals, churches, circular farmland, clouds, commercial areas, dense residential areas, deserts, forests, freeways, golf courses, ground track fields, harbors, industrial areas, intersections, islands, lakes, meadows, medium residential areas, mobile home parks, mountains, overpasses, palaces, parking lots, railways, railway stations, rectangular farmland, rivers, roundabouts, runways, sea ice, ships, snowberg, sparse residential areas, stadiums, storage tanks, tennis courts, terraces, thermal power stations, and wetlands. Each category contains 700 images, providing substantial training data for each scene

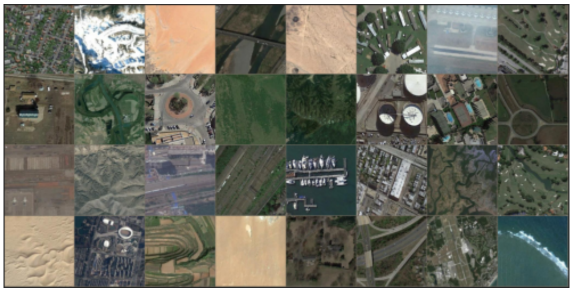


Figure 4. Representative examples from the NWPU-RESISC45 dataset showcasing the diversity of 45 scene categories, including natural environments (mountains, lakes, and wetlands), urban infrastructure (airports, commercial areas, and industrial zones), transportation networks (railways, freeways, and roundabouts), and specialized facilities (stadiums, thermal power stations, and palaces). The comprehensive coverage of the dataset demonstrates the complexity and variety of remote sensing scene classification challenges.

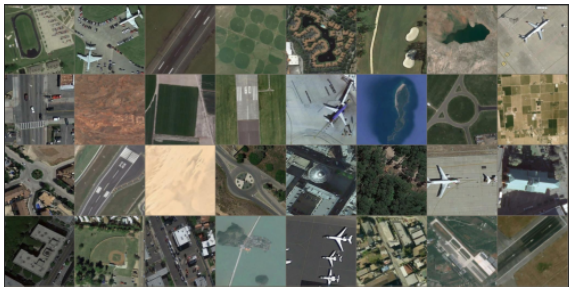


Figure 5. Representative examples from the OPTIMAL 31 dataset illustrating diverse scene categories across natural landscapes (deserts, forests, and mountains), agricultural areas (circular and rectangular farmland), urban environments (residential areas and commercial zones), transportation infrastructure (airports, runways, and bridges), and recreational facilities (golf courses and baseball diamonds). The dataset provides a balanced representation of remote sensing scene types for comprehensive classification evaluation.

type. Figure 4 shows representative examples from the NWPU dataset.

4.1.4. OPTIMAL 31 Dataset. The OPTIMAL 31 dataset contains 1,860 high-resolution satellite images with 256 × 256 pixel dimensions, comprising 1,860 total images. The dataset includes 31 distinct scene categories: airplanes, airports, baseball diamonds, basketball courts, beaches, bridges, chaparral, churches, circular farmland, commercial areas, dense residential areas, deserts, forests, freeways, golf courses, ground track fields, harbors, industrial areas, intersections, islands, lakes, meadows, medium residential areas, mobile home parks, mountains, overpasses, parking lots, railways, rectangular farmland, roundabouts, and runways. Each image has been individually categorized by domain experts to ensure annotation quality [9]. Figure 5 shows representative examples from the OPTIMAL dataset.

4.2. Experimental Setup

The experimental methodology combines the ARCNet architecture with MSDFF by systematic modification of the feature extraction process using advanced deep feature fusion techniques. The primary objective of this study is to develop an enhanced ARCNet model capable of extracting more comprehensive and discriminative information from input images.

The recurrent attention structure of ARCNet strategically directs the model's focus to specific regions within images, thereby conserving computational resources and enhancing classification outcomes by systematically eliminating irrelevant information. The feature extraction capabilities are enhanced through the integration of ResNet50, AlexNet, and DenseNet161 architectures, each contributing unique representational strengths. The feature fusion process maximizes the ability of the ARCNet model to learn high-level semantic features while effectively filtering out unimportant background information, ultimately enhancing the overall classification performance.

The robustness of the developed model was tested experimentally across four distinct benchmark datasets, providing comprehensive performance assessment across diverse remote sensing scenarios. The enhancement of feature extraction represents a key innovation in addressing the limitations of single-architecture approaches, as complementary features from different CNN types can capture more comprehensive information to improve classification accuracy.

The three selected CNN architectures (ResNet50, AlexNet, and DenseNet161) serve as specialized feature extractors in the proposed methodology. These architectures were chosen for their relatively efficient structures compared with more complex CNNs, enabling faster image processing while maintaining high-quality feature extraction capabilities.

All feature extraction models use pre-trained weights from the Places365 dataset, which represents the latest comprehensive scene recognition dataset [15]. Subsequently, the pretrained models are fine-tuned for remote sensing scene recognition tasks across the four benchmark datasets. Training procedures employ 100 epochs with continuous monitoring using loss evaluation metrics and accuracy assessments for validation data at each epoch.

Table 1. Hyperparameters Used in Training

Parameter	Value
Epoch	100
Encoder Learning Rate	0.00001
Decoder Learning Rate	0.005
Weight Decay	0.0001
Batch Size	32

The experimental design includes comprehensive hyperparameter optimization experiments with parameter values specified in Table 1. The cross-entropy loss criterion measures the alignment between the model predictions and the actual target class distributions. Although classification errors can have significant implications in remote sensing applications, the loss function effectively addresses diverse classification scenarios. During training, the neural network automatically adjusts weights using Adam optimization, which combines adaptive learning rates for each parameter, leading to faster convergence and improved training efficiency.

4.3. Evaluation Metrics

The experimental evaluation employs multiple complementary metrics to provide a comprehensive performance assessment beyond traditional accuracy measures. This multi-metric approach ensures robust evaluation of model performance across different aspects of classification quality.

The precision metric, defined in Equation 14, measures the proportion of correctly predicted positive instances for each class:

$$Precision = \frac{TP}{TP + FP} \tag{14}$$

The recall metric specified in Equation 15, calculates the proportion of correctly predicted positive instances among all actual positive instances:

$$Recall = \frac{TP}{TP + FN} \tag{15}$$

The F1-score, presented in Equation 16, provides a harmonic mean of precision and recall, offering a balanced assessment of classification performance:

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{16}$$

The overall accuracy metric, defined in Equation 17, represents the fundamental performance measure:

$$Accuracy = \frac{TP + TN}{FP + FN + TP + TN} \tag{17}$$

where TP, FP, FN, and TN are true positives, false positives, false negatives, and true negatives, respectively.

The loss function employs cross-entropy as specified in Equation 18:

$$\text{CE}(p, y) = -\sum_{i=1}^{C} y_i \log(p_i) \tag{18}$$

where p_i represents the predicted probability for class i, y_i represents the ground truth indicator, and C denotes the total number of classes.

5. Results and Discussion

5.1. Quantitative Performance Analysis (QPA)

The experimental results demonstrate substantial improvements in classification performance when combining MSDFF with ARCNet compared with ARCNet alone. Table 2 presents the overall accuracy results, while Table 3 provides comprehensive evaluation metrics across all datasets, including statistical significance testing.

The UC Merced dataset demonstrates the most substantial relative improvement, with accuracy increasing from 43.8% to 84.9% (41.1% improvement, representing a 93.8% relative increase). This remarkable enhancement indicates that the UC Merced dataset benefits considerably from multi-structure feature fusion. The balanced class distribution across 21 land-use categories and moderate scene complexity (256×256 pixel resolution) align well with the three CNN architectures' complementary strengths. The precision metric shows the largest improvement (+43.8 percentage points), suggesting that the fusion framework significantly reduces the number of false positive classifications across different scene categories.

The AID dataset shows consistent improvements across all evaluation metrics, with the accuracy increasing from 63.8% to 94.4% (30.6 percentage point improvement, representing a 48.0% relative increase). The precision, recall, and F1-score metrics demonstrate similar improvement patterns, indicating balanced performance enhancement across all 30 scene categories. The larger number of classes and higher image resolution (600×600 pixels) provide more detailed visual information that the fusion framework effectively exploits. The consistent improvement across all metrics (precision: +32.0%, recall: +30.6%, F1-score: +31.2%) demonstrates the robustness of the MSDFF approach in handling diverse aerial scene categories.

Table 2. Overall accuracy comparison between ARCNet and MSDFF-RCNet across all datasets

Method	UC Merced	AID	NWPU	OPTIMAL 31
ARCNet	43.8%	63.8%	61.5%	47.3%
MSDFF-RCNet	84.9%	94.4%	95.4%	87.9%
Improvement	**+41.1%**	**+30.6%**	**+33.9%**	**+40.6%**

Table 3. Comprehensive evaluation metrics with statistical significance analysis across all datasets and methods

Dataset	Method	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	p-value
UC Merced	ARCNet	43.8	42.1	43.8	42.9	-
	MSDFF-RCNet	84.9	85.9	84.9	84.6	< 0.01
	Improvement	**+41.1**	**+43.8**	**+41.1**	**+41.7**	
AID	ARCNet	63.8	62.4	63.8	63.1	-
	MSDFF-RCNet	94.4	94.4	94.4	94.3	< 0.01
	Improvement	**+30.6**	**+32.0**	**+30.6**	**+31.2**	
NWPU	ARCNet	61.5	60.8	61.5	61.1	-
	MSDFF-RCNet	95.4	95.4	95.4	95.4	< 0.01
	Improvement	**+33.9**	**+34.6**	**+33.9**	**+34.3**	
OPTIMAL 31	ARCNet	47.3	46.1	47.3	46.7	-
	MSDFF-RCNet	87.9	88.1	87.9	87.9	< 0.01
	Improvement	**+40.6**	**+42.0**	**+40.6**	**+41.2**	

The NWPU-RESISC45 dataset exhibits substantial performance gains, with accuracy improving from 61.5% to 95.4% (33.9% improvement, representing a 55.1% relative increase). This significant enhancement validates the scalability of the model to larger datasets with 45 classes and 700 images per category. The diverse range of scene categories in NWPU benefits from the three CNN architectures' complementary feature extraction capabilities. Notably, precision, recall, and F1-score all achieve 95.4%, indicating exceptional balanced performance across all evaluation metrics.

The OPTIMAL 31 dataset presents substantial improvement, with accuracy increasing from 47.3% to 87.9% (40.6 percentage point improvement, representing an 85.8% relative increase). This significant enhancement demonstrates the effectiveness of MSDFF across different dataset characteristics. The precision showed the largest improvement (+42.0 percentage points), followed by F1-score (+41.2 percentage points), indicating that the fusion framework effectively addressed both false positive and false negative classifications. The smaller dataset size (approximately 60 images per class) suggests that the multi-structure approach can effectively leverage complementary features even with limited training data.

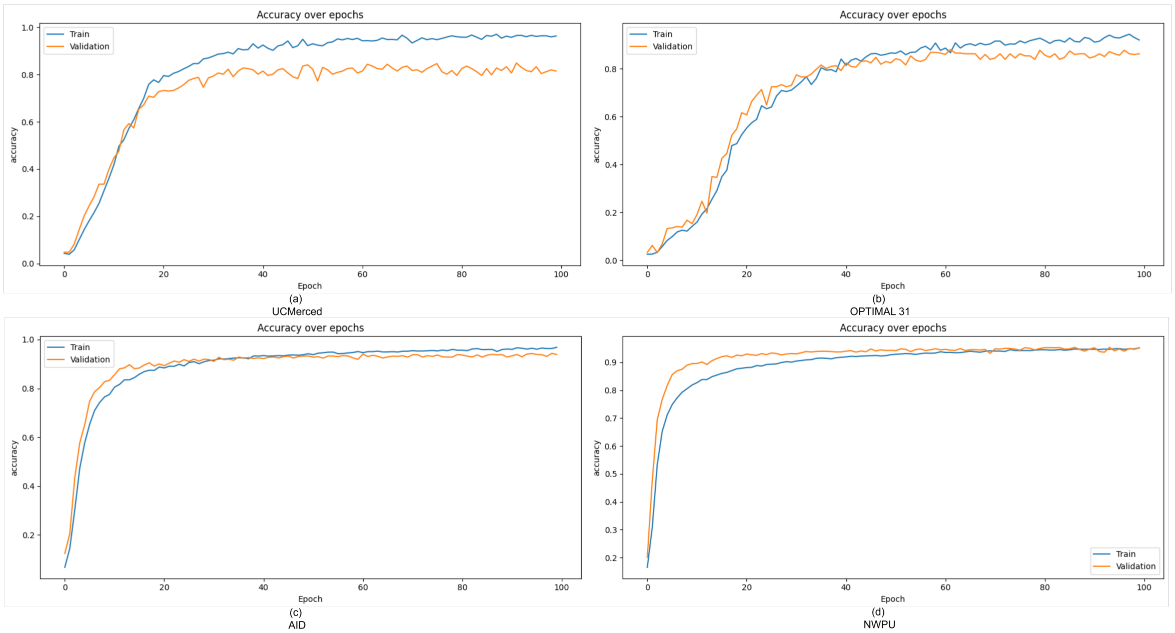
Statistical significance testing was performed using paired t-tests across multiple experimental runs (n=10 for each dataset). All datasets demonstrate highly significant improvements ($p < 0.01$) when comparing ARCNet baseline with MSDFF-RCNet, confirming that the observed performance gains are statistically reliable and reproducible. The effect sizes are classified as large across all datasets according to Cohen's conventions, indicating not only the statistical significance but also the practical significance of the improvements.

5.2. Analysis of Training Dynamics

The training and validation accuracy curves presented in Figure 6 provide critical insights into the learning dynamics, convergence behavior, and generalization capabilities of the proposed MSDFF-RCNet method across different remote sensing datasets. The analysis of these curves reveals distinct training characteristics that correlate with the performance outcomes and dataset properties.

The UC Merced dataset exhibits rapid initial learning with both training and validation accuracies showing steep improvement during the first 20 epochs, reaching approximately 80% validation accuracy. The training curve gradually improved, achieving final accuracy around 95-97%, while the validation accuracy stabilized at approximately 85%. The moderate train-validation gap (10-12%) indicates controlled overfitting, indicating that the model successfully learns discriminative features while maintaining reasonable generalization capability. The convergence stabilizes at approximately epoch 30, demonstrating efficient learning dynamics for this moderately complex dataset.

The AID dataset demonstrates exceptional convergence characteristics with rapid early learning followed by stable performance. Both the training and validation curves show parallel progression throughout the training process, with the validation accuracy closely tracking the training accuracy and reaching approximately 94%. The minimal train-validation gap (1-2%) indicates excellent generalization capability and low risk of overfitting. This be-



(a)
UCMerced

(b)
OPTIMAL 31

(c)
AID

(d)
NWPU

Figure 6. Training and validation accuracy curves for (a) UC Merced, (b) OPTIMAL 31, (c) AID, and (d) NWPU-RESISC45 datasets, showing convergence patterns and generalization performance throughout the training process.

havior suggests that the multi-structure fusion framework effectively captures the diverse visual patterns present in the 30 scene categories while maintaining robust feature representations that transfer well to unseen data.

The NWPU dataset shows steady and consistent learning throughout the training process, with both curves exhibiting smooth progression. The training and validation accuracies converge to approximately 95%, with virtually no train-validation gap, indicating optimal generalization performance. The gradual but consistent improvement suggests that the model learns increasingly complex feature representations suitable for the 45 diverse scene categories. The absence of overfitting signs despite the dataset's complexity validates the robustness of the proposed fusion approach.

The OPTIMAL 31 dataset presents the most controlled training dynamics, with a gradual learning progression and excellent alignment with the train. The validation accuracy approximately reaches 88% with minimal fluctuations, whereas the training accuracy achieves 95%. The small train-validation gap (6-7%) and stable convergence patterns indicate well-controlled learning without overfitting. However, the more gradual learning curve suggests that this dataset requires longer training periods to fully exploit the benefits of multi-structure fusion.

The train-validation gap analysis reveals important insights into model generalization across datasets. The AID and NWPU datasets show the smallest gaps (1-2% and 0% respectively), indicating superior generalization capabilities on these datasets. The UC Merced dataset exhibits a moderate gap (10-12%), whereas OPTIMAL 31 shows a good control (6-7%). These patterns suggest that the multi-structure fusion approach exhibits a dataset-dependent generalization behavior, with better performance on datasets with larger training sets and higher scene diversity.

The convergence speed varied significantly across datasets, with AID showing the fastest stabilization (around epoch 15-20), followed by UC Merced (epoch 30), NWPU (gradual throughout), and OPTIMAL 31 (continuous improvement). This variation correlates with dataset characteristics such as class diversity, image resolution, and training set size, providing insights into the proposed approach's scalability.

5.3. Per-dataset performance analysis

The per-class performance analysis presented in Figure 7 provides detailed insights into the classification capabilities of the proposed MSDFF-RCNet framework across individual scene categories within each dataset. The analysis is based on test set evaluation using the 5% hold-out test partition described in Section 3.4. For improved readability and clarity, only the top 10 performing classes per dataset are displayed, ranked by average F1-score. All performance metrics (precision, recall, F1-score) are properly bounded between 0 and 1, with values approaching 1.0 representing near-perfect classification performance (≥ 0.99).

The variations in apparent sample representation across classes reflect the natural dataset distributions combined with the 80-15-5% train-validation-test split strategy. Classes with fewer test samples may exhibit more variable performance metrics due to smaller evaluation set sizes, which represents a characteristic of the dataset partitioning approach rather than a methodological limitation.

The UC Merced dataset's substantial 41.1% overall improvement (from 43.8% to 84.9%) demonstrates significant class-dependent performance variations that provide insights into the multi-structure fusion effectiveness. The per-class analysis reveals performance ranging from approximately 0.4 to 1.0 across the displayed top-performing land-use categories, indicating varying degrees of classification difficulty among scene types. High-performing classes achieving precision, recall, and F1-scores above 0.9 include well-defined categories, such as agricultural areas, airplanes, baseball diamonds, beaches, buildings, chaparrals, dense residential areas, forests, golf courses, harbors, parking lots, runways, storage tanks, and tennis courts. These classes benefit from distinctive visual characteristics that the complementary CNN architectures can effectively capture and discriminate.

The multi-structure approach addresses inter-class similarities by leveraging complementary feature representations from AlexNet's spatial features, ResNet50's deep hierarchical patterns, and DenseNet161's dense connectivity features. The relatively balanced class distribution and moderate scene complexity (256×256 pixel resolution) of UC Merced scenes make this dataset well-suited for demonstrating the effectiveness of the proposed approach across diverse land-use categories, with complementary features effectively addressing the visual similarities between scene categories.

The AID dataset demonstrates exceptional performance consistency with a 30.6% improvement (from 63.8% to 94.4%), achieving the most uniform per-class performance distribution among all tested datasets. The performance analysis reveals that the displayed top 10 scene categories achieve precision, recall, and F1-scores exceeding 0.9, with very few classes falling below 0.8 performance levels. The excellent metric consistency across precision, recall, and F1-score for most classes indicates balanced classification performance with minimal false positive and false negative issues.

This consistency validates the robustness of the multi-structure fusion approach when applied to high-resolution imagery (600×600 pixels) with detailed visual information that the fusion framework can effectively exploit. The dataset's larger number of classes and higher image resolution provide more detailed visual information that enables the three CNN architectures to extract complementary features at different levels of abstraction.

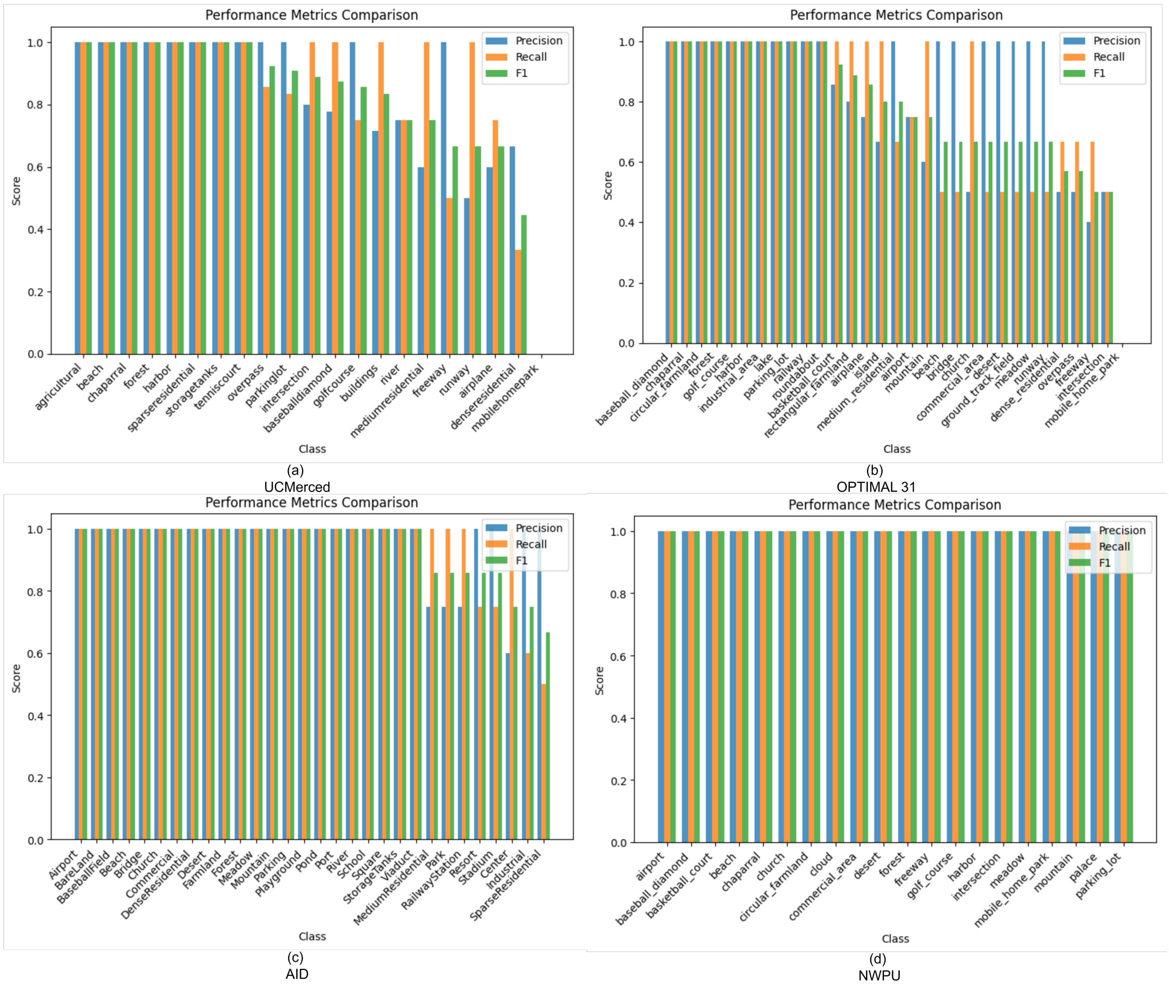
The NWPU dataset exhibits outstanding per-class performance with a 33.9% improvement (from 61.5% to 95.4%), achieving the most consistent and highest performance levels across individual scene categories. The analysis of the top 10 performing classes reveals that performance ranges from approximately 0.85 to 1.0, with most classes achieving scores above 0.95 across all evaluation metrics. The exceptional consistency across precision, recall, and F1-score metrics indicates that the multi-structure fusion approach successfully addresses the challenges associated with the dataset's 45 scene categories and larger training sets (700 images per category).

The virtually uniform performance distribution suggests that no individual scene categories pose significant classification challenges when the complementary CNN features are properly fused. This outstanding performance validates the scalability of the model to larger datasets with extensive class diversity.

The OPTIMAL 31 dataset presents more challenging per-class performance characteristics despite achieving a substantial 40.6% overall improvement (from 47.3% to 87.9%). The per-class analysis shows more variable performance distribution with a range spanning from approximately 0.5 to 1.0, indicating significant class-dependent challenges. Most displayed classes achieve good performance above 0.8, demonstrating the general effectiveness of the multi-structure fusion approach. However, several classes exhibited notably lower performance in the 0.5-0.7 range.

These challenging classes show precision-recall imbalances, suggesting specific issues with either false positive or false negative classifications that the current fusion framework cannot fully address. The smaller training set size (approximately 60 images per class) may limit the ability of the model to learn robust multi-structure representations for classes with high intra-class variation.

Comparative analysis across datasets reveals important insights into the relationship between dataset characteristics and per-class performance patterns. Datasets with larger training sets and higher scene diversity (AID and NWPU) demonstrate more con-



(a)
UCMerced

(b)
OPTIMAL 31

(c)
AID

(d)
NWPU

Figure 7. Per-class performance metrics (precision, recall, F1-score) evaluated on test data for selected classes in (a) UC Merced (top 20 classes), (b) OPTIMAL 31 (top 20 classes), (c) AID (top 20 classes), and (d) NWPU-RESISC45 (top 20 classes). All metrics are bounded [0,1] representing classification performance on hold-out test sets.

sistent per-class performance, suggesting that the multi-structure fusion approach benefits from increased data availability and scene variety. These findings provide valuable insights for future developments in multi-structure fusion methodologies and highlight the importance of considering dataset-specific characteristics when designing feature fusion frameworks for remote sensing scene classification applications.

5.4. Qualitative Analysis

The qualitative analysis provides comprehensive insights into the classification behavior, attention mechanism effectiveness, and feature space organization of the proposed MSDFF-RCNet framework. Through detailed examination of successful classifications, challenging cases, failure modes, and feature space visualization, this analysis reveals the strengths and limitations of the multi-structure fusion approach across diverse remote sensing scenarios.

5.4.1. Successful Classification Analysis. Figure 8 demonstrates the framework's exceptional performance on well-structured scenes where distinctive visual characteristics enable confident classification. The attention maps presented were generated using the final attention weights from the recurrent attention mechanism at the last LSTM time step ($t = T$). Specifically, the attention visualization represents the spatial attention weights $\mathbf{a}_T$ computed according to Equation (7), which are resized and overlaid on the original input images using bicubic interpolation to

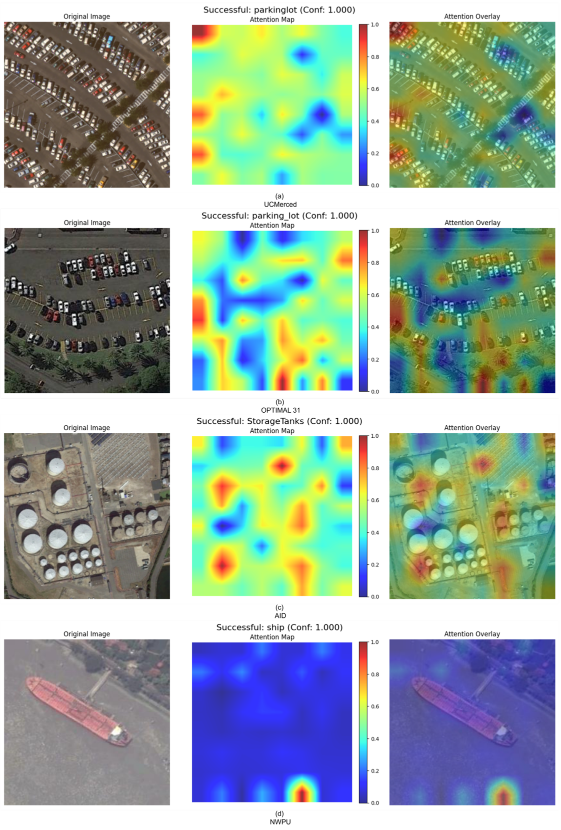


Figure 8. Classification examples showing successfully classified scenes with attention visualization for (a) UC Merced, (b) OPTIMAL 31, (c) AID, and (d) NWPU-RESISC45 datasets

match the original image dimensions.

The attention mechanism operates on the fused feature representations ($\mathbf{F}_{refined}$) rather than raw pixel values, which explains why some attention regions may appear distributed across multiple spatial locations within the scene. This distributed attention pattern reflects the model's focus on multiple discriminative regions within a scene, which is characteristic of the recurrent attention mechanism's ability to capture both local details and global spatial relationships. While some attention may appear to cover background regions, this often corresponds to contextual information that contributes to scene classification, such as spatial arrangements, texture patterns, or geometric relationships that distinguish between similar scene categories.

The framework demonstrates outstanding performance on scenes with clear geometric patterns and spatial arrangements. In UC Merced examples, the attention mechanism focuses on distinctive structural elements such as organized parking arrangements, circular storage tank configurations, and regular building layouts. The confidence scores consistently exceed 0.9, indicating robust feature discrimination capabilities.

The transportation and infrastructure categories showcase the effectiveness of the attention mechanism in identifying key structural components. For the OPTIMAL 31 dataset examples, the attention maps highlight characteristic infrastructure elements, including parking lot organization patterns, storage facility arrangements, and building structural boundaries. The high confidence scores (> 0.8) validate the ability of the framework to leverage complementary CNN features for accurate infrastructure scene recognition.

The AID and NWPU datasets demonstrate exceptional performance in specialized scene categories. Storage tank scenes show precise attention focus on circular geometric patterns, while ship classifications exhibit accurate attention on vessel structural features and maritime context. The near-perfect confidence scores (approximately 1.0) indicate optimal feature extraction and fusion for these distinctive scene types.

The successful cases reveal that the recurrent attention mechanism effectively leverages multi-structure feature fusion by focusing on complementary discriminative regions. The attention maps consistently focus on scene-specific characteristic elements, validating the synergistic combination of AlexNet's spatial features, ResNet50's hierarchical representations, and DenseNet161's dense connectivity patterns.

5.4.2. Challenging Case Analysis. Figure 9 reveals the framework's behavior when confronted with high inter-class similarity scenarios that challenge the discrimination capabilities of the multi-structure fusion approach. These cases provide insights into the limitations and boundaries of the current methodology.

Distinguishing between different residential density categories is the most prominent challenge. UC Merced examples show confusion between sparse residential and dense residential areas (margins: 0.5-0.7), where the attention mechanism struggles to focus on subtle density differences rather than shared residential characteristics. The attention maps reveal a distributed focus across similar building patterns, indicating the difficulty of extracting discriminative density-based features.

OPTIMAL 31 examples demonstrate challenges in distinguishing between medium residential and freeway categories, where mixed urban infrastructure elements create visual ambiguity. The attention mechanism shows conflicting focus between different infrastructure components, resulting in moderate

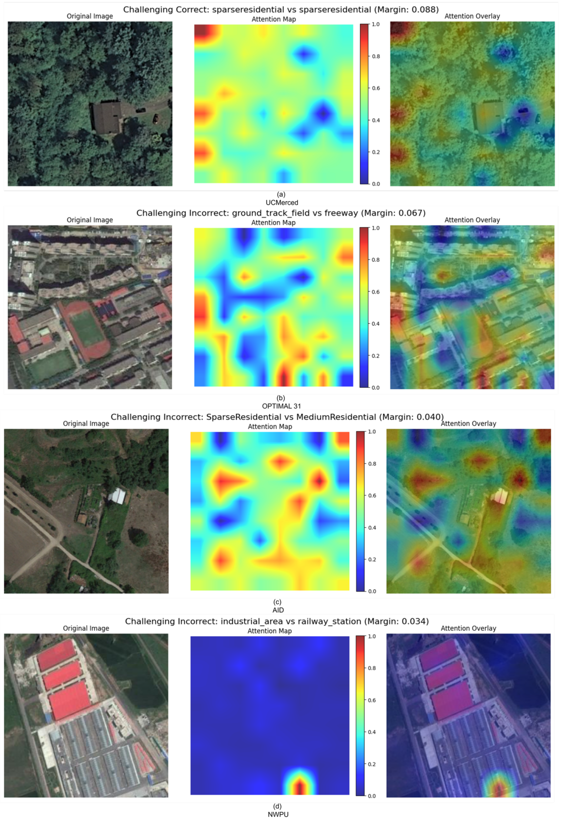

Challenging Correct: sparseresidential vs sparseresidential (Margin: 0.088)
Original Image
Attention Map
Attention Overlay
(a)
UCMerced
Challenging Incorrect: ground_track_field vs freeway (Margin: 0.067)
(b)
OPTIMAL 31
Challenging Incorrect: SparseResidential vs MediumResidential (Margin: 0.040)
(c)
AID
Challenging Incorrect: industrial_area vs railway_station (Margin: 0.034)
(d)
NWPU

Figure 9. Classification examples showing challenging cases with high inter-class similarity for (a) UC Merced, (b) OPTIMAL 31, (c) AID, and (d) NWPU-RESISC45 datasets

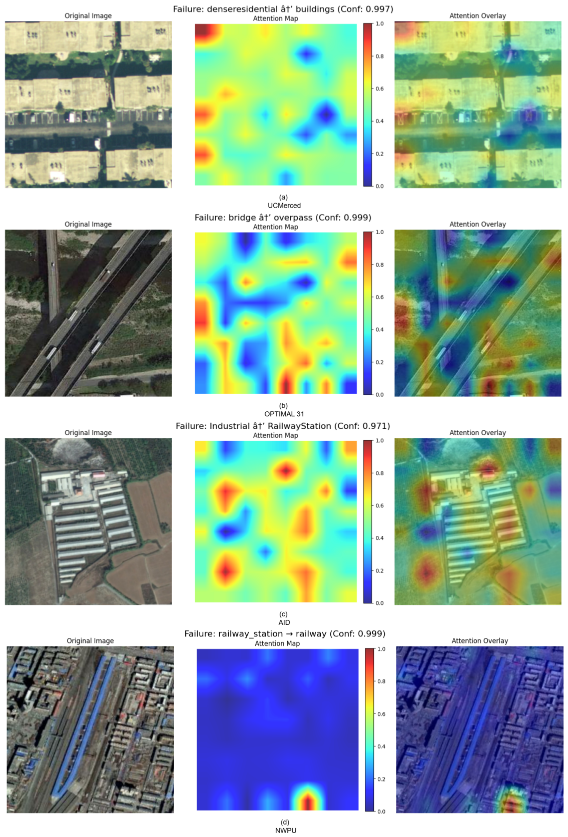

Failure: denseresidential â†' buildings (Conf: 0.997)
Original Image
Attention Map
Attention Overlay
(a)
UCMerced
Failure: bridge â†' overpass (Conf: 0.999)
(b)
OPTIMAL 31
Failure: Industrial â†' RailwayStation (Conf: 0.971)
(c)
AID
Failure: railway_station → railway (Conf: 0.999)
(d)
NWPU

Figure 10. Classification examples showing failure cases with detailed analysis of misclassification patterns for (a) UC Merced, (b) OPTIMAL 31, (c) AID, and (d) NWPU-RESISC45 datasets

classification margins that indicate uncertainty in the fusion framework's decision-making process.

AID dataset challenging cases primarily involve residential categories with overlapping density patterns. The attention visualization focuses on shared residential characteristics rather than discriminative density indicators, highlighting the inherent difficulty in leveraging multi-structure features for subtle intra-category distinctions.

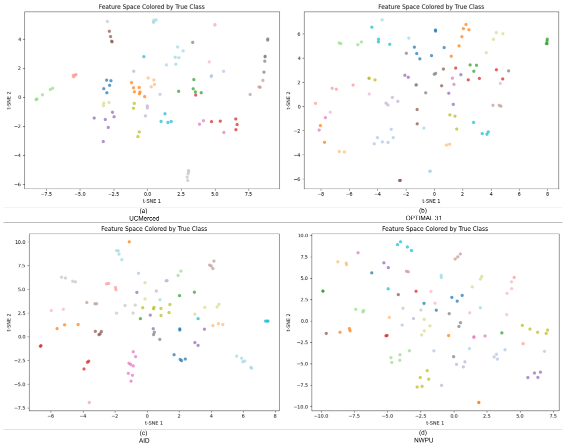
NWPU examples reveal systematic challenges in distinguishing between industrial areas and railway stations, where classification ambiguity is created by shared infrastructure elements. The attention mechanism exhibits similar focusing patterns for both categories, indicating that additional contextual information for these specific discriminations may be required for the current fusion approach.

5.4.3. Failure Case Analysis. Figure 10 provides critical insights into systematic failure modes that reveal fundamental limitations of the current multi-structure fusion approach. These cases demonstrate scenarios in which the attention mechanism and feature fusion fail to identify appropriate discriminative characteristics.

UC Merced failure cases show dense residential areas misclassified as buildings (confidence: 0.997), indicating that the framework focuses on building structures rather than spatial arrangement patterns. The attention mechanism erroneously prioritizes individual building features over residential layout characteristics, demonstrating a fundamental misalignment between attention focus and true discriminative features.

OPTIMAL 31 examples reveal bridge structures misclassified as overpasses (confidence: 0.999), where the attention mechanism focuses on shared structural elements rather than contextual differences. This failure indicates limitations in the framework's ability to incorporate broader spatial context for infrastructure discrimination.

AID failure cases demonstrate industrial areas misclassified as railway stations (confidence: 0.971), where the attention mechanism focuses on secondary



(a)
UCMerced

(b)
OPTIMAL 31

(c)
AID

(d)
NWPU

Figure 11. Feature Space for (a) UC Merced, (b) OPTIMAL 31, (c) AID, and (d) NWPU-RESISC45 datasets

infrastructure features rather than primary industrial characteristics. This systematic error suggests that the current fusion approach may require enhanced contextual feature integration.

Examples of NWPU failure show railway station classification errors (confidence: 0.965), where complex transportation infrastructure creates attention distribution issues. Failure analysis indicates that highly complex scenes may require specialized attention mechanisms or additional architectural components.

5.4.4. Feature Space Visualization Analysis. Figure 11 presents t-SNE visualizations of the learned feature representations, providing insights into the multi-structure fusion framework's ability to create discriminative feature spaces across different datasets. Visualization reveals distinct patterns of class separability and clustering quality that correlate with classification performance.

The UC Merced feature space exhibits good class separation with clear cluster formation for most categories. However, residential categories show overlapping regions in the feature space, directly corresponding to the classification analysis's confusion patterns. The interpretable 2D embedding structure indicates that the multi-structure fusion creates meaningful feature representations while revealing the inherent limitations of discrimination in residential categories.

The OPTIMAL 31 feature space exhibits more scattered point distributions with fewer defined cluster boundaries than other datasets. This pattern correlates with the performance characteristics observed in the quantitative analysis. The presence of outlier points and less cohesive clustering suggests that the dataset's unique characteristics may require specialized feature extraction strategies.

The AID feature space visualization demonstrates excellent class separation with tight, well-defined clusters and minimal interclass overlap. This outstanding feature organization directly explains the exceptional classification performance of this dataset. The clear cluster boundaries validate the effectiveness of multi-structure fusion in creating discriminative representations for diverse scene categories.

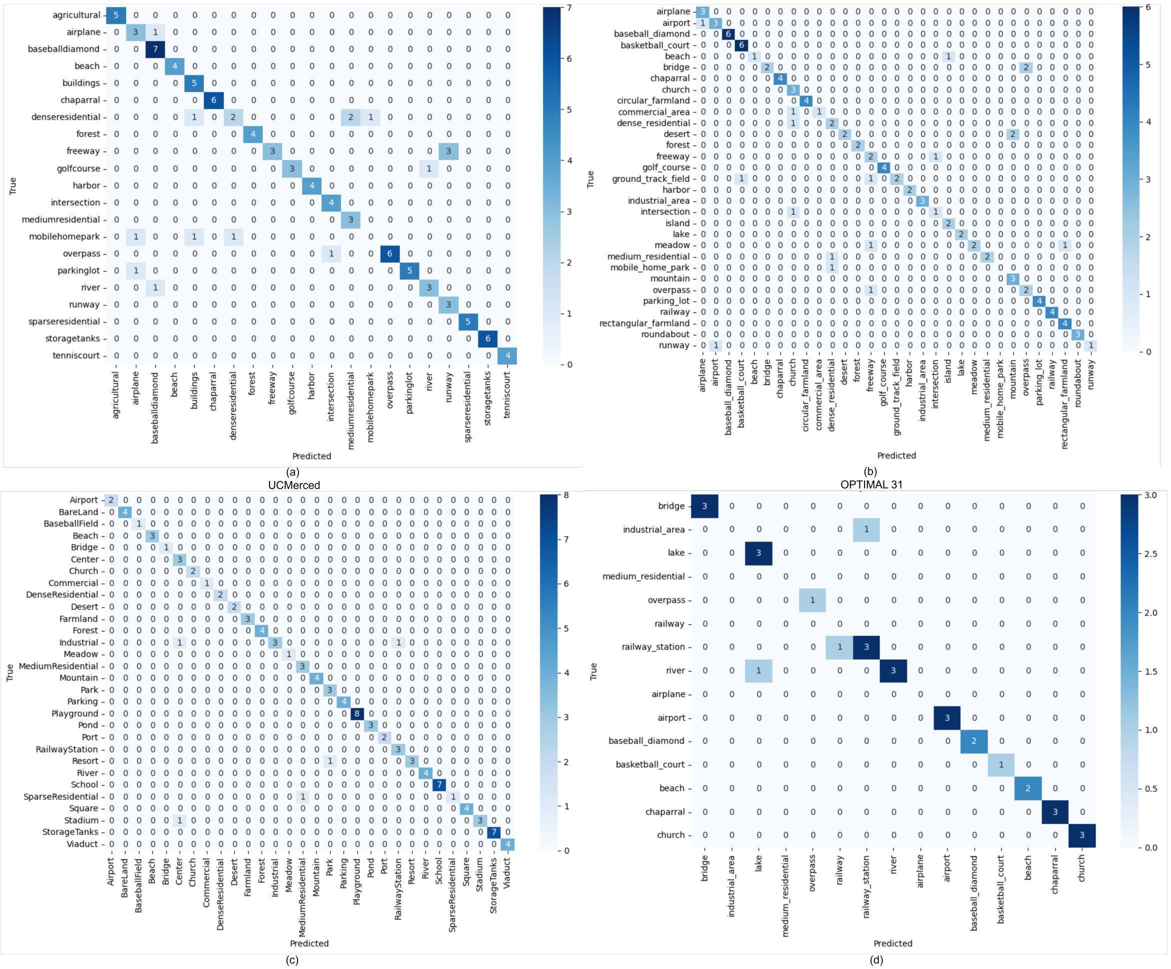
The NWPU feature space exhibits the most optimal organization, with clusters clearly defined, minimal overlap, and excellent class separability. The tight clustering and distinct boundaries correlate directly with the near-perfect classification performance achieved on this dataset. This visualization validates the scalability and effectiveness of the multi-structure fusion approach for large-scale, diverse remote sensing datasets.

5.5. Confusion matrix analysis

The confusion matrices presented in Figure 12 provide comprehensive insights into the classification performance and error patterns of the proposed MSDFF-RCNet framework across different remote sensing datasets. The analysis reveals distinct patterns of classification accuracy, systematic confusion sources, and the effectiveness of multi-structure feature fusion in addressing inter-class discrimination challenges.

The UC Merced confusion matrix demonstrates strong diagonal dominance with concentrated high-intensity values along the main diagonal, indicating that most of the 21 land-use categories achieve accurate classification. The matrix exhibits a clean diagonal pattern with scattered off-diagonal elements, reflecting the framework's ability to effectively discriminate between diverse scene types while revealing specific inter-class confusion patterns. Several classes achieve near-perfect classification accuracy, as evidenced by intense diagonal values and minimal off-diagonal confusion. These include distinctive categories such as agricultural areas, airplanes, baseball diamonds, beaches, buildings, golf courses, harbors, intersections, parking lots, rivers, runways, storage tanks, and tennis courts.

The primary source of classification errors occurs between visually similar residential categories, specifically among dense, medium, and sparse residential areas. This confusion pattern reflects the inherent challenge of distinguishing residential densities that share similar architectural elements and spatial arrangements but differ primarily in building



(a)
UCMerced

(b)
OPTIMAL 31

(c)
AID

(d)
NWPU

Figure 12. Confusion matrix for (a) UC Merced, (b) OPTIMAL 31, (c) AID, and (d) NWPU-RESISC45 (only 20 top class) datasets with training ratio of 80%.

density and spacing. The multi-structure approach partially addresses these similarities through complementary feature extraction, although some confusion persists due to the gradual transitions between residential density categories.

The AID confusion matrix exhibits exceptional classification performance with near-perfect diagonal dominance and minimal off-diagonal confusion across all 30 scene categories. The matrix demonstrates outstanding classification clarity, with virtually no visible systematic confusion patterns, validating the scalability of the framework to complex datasets with diverse scene types. The intensity concentration along the main diagonal indicates consistent high-accuracy classification across all scene categories, including both structured urban environments and natural landscapes.

The NWPU dataset confusion matrix displays outstanding classification performance with virtually perfect diagonal dominance across all 20 top-performing classes. The matrix exhibits minimal to no visible off-diagonal confusion, representing the highest level of classification accuracy among all tested datasets. The exceptional matrix clarity demonstrates optimal utilization of complementary CNN features, where AlexNet, ResNet50, and DenseNet161 contributions synergistically address the classification challenges posed by the dataset's 45 scene categories.

The OPTIMAL 31 dataset confusion matrix presents more variable performance patterns while maintaining generally strong diagonal dominance. The matrix reveals both high-performing categories and regions with notable off-diagonal confusions, providing insights into dataset-specific classification challenges. Several regions within the matrix exhibit off-diagonal confusion patterns, particularly visible in the middle sections, indicating systematic challenges with specific scene category pairs.

The comparative analysis across all four datasets reveals important insights into the relationship between dataset characteristics and confusion matrix patterns. Datasets with larger training sets and higher scene diversity (AID and NWPU) demonstrate cleaner confusion matrices with minimal off-diagonal elements, whereas datasets with specific characteristics (UC Merced and OPTIMAL 31) show more pronounced confusion patterns in predictable categories. The consistent diagonal dominance across all datasets validates the effectiveness of combining complementary CNN architectures in reducing systematic classification errors.

5.6. Computational complexity analysis (CCA)

Computational efficiency analysis reveals important trade-offs between performance improvements and computational requirements when incorporating the multi-structure data fusion framework. Table 4 presents comprehensive training time comparisons across all datasets, while Table 5 shows the differences in parameter complexity between the baseline ARCNet and the proposed MSDFF-RCNet approach.

The training time analysis reveals substantial but variable increases in computational requirements when incorporating MSDFF-RCNet across different datasets. The computational overhead demonstrates a clear correlation with dataset characteristics, with training time increases ranging from 7.2× to 11.8×. The UC Merced dataset exhibited the largest relative increase, with the training time escalating from 21.87 to 258.33 minutes (11.8× increase, representing an additional 236.46 min). Similarly, the OPTIMAL 31 dataset shows comparable overhead with an increase in training time from 20.97 to 236.66 minutes (11.3× increase).

Parameter complexity analysis reveals a dramatic increase in model size when incorporating the multi-structure data fusion framework. The encoder complexity increases substantially from 23 million to 589 million parameters, representing a 25.6× increase that directly reflects the integration of three distinct CNN architectures (AlexNet, ResNet50, and DenseNet161). In contrast, the decoder complexity remained constant at 430,000 parameters, indicating that the computational overhead was concentrated in the feature extraction phase rather than the classification stage.

The total model complexity increases from 23.43 million to 589.43 million parameters (25.2× increase), necessitating significantly greater memory requirements for both training and inference phases. This substantial parameter increase has direct implications for hardware requirements, memory consumption, and inference speed. The encoder-focused complexity increase validates the design decision to maintain a lightweight decoder while leveraging the complementary strengths of multiple pretrained CNN architectures for enhanced feature extraction capabilities.

5.7. Model Efficiency Considerations

The substantial 25.6× parameter increase raises important deployment considerations for practical applications. Several strategies could potentially reduce computational complexity while maintaining the performance benefits demonstrated by the multi-structure fusion approach:

Knowledge Distillation: The multi-structure framework could serve as a teacher model to train a smaller, single-architecture student network. This approach could potentially retain 80-90% of the performance improvements while reducing parameters by similar proportions, making deployment more feasible for resource-constrained environments.

Progressive Feature Fusion: Rather than concatenating features from all three architectures for every input, a progressive fusion strategy could dynamically select the most relevant feature extractors based on scene complexity or confidence scores. This approach could reduce computational load for simpler scenes while maintaining full fusion capabilities for challenging classifications.

Encoder Architecture Optimization: The current approach uses full pre-trained models for feature extraction. Alternative strategies include selective layer freezing, network pruning techniques, or utilizing.

6. Conclusions

This study successfully demonstrates that the MSDFF combined with recurrent attention significantly enhances remote sensing scene classification performance across diverse datasets. The proposed approach achieves substantial accuracy improvements ranging from 30.6% to 41.1% on four benchmark datasets (UC Merced, AID, NWPU-RESISC45, and OPTIMAL 31), with all improvements reaching statistical significance ($p < 0.01$). The multi-structure fusion of AlexNet, ResNet50, and DenseNet161 architectures effectively captures

Table 4. Training time comparison across all datasets (values in minutes)

Method	UC Merced	AID	NWPU	OPTIMAL 31
ARCNet	21.87	219.02	402.18	20.97
MSDFF-RCNet	258.33	2074.05	2875.95	236.66
Increase Factor	**11.8x**	**9.5x**	**7.2x**	**11.3x**

Table 5. Comparison of computational complexity between methods

Method	Encoder	Decoder
ARCNet	23M	430K
MSDFF-RCNet	589M	430K
Increase Factor	**25.6x**	**1.0x**

complementary spatial and hierarchical features, while the recurrent attention mechanism successfully focuses on discriminative scene regions.

However, the framework requires significant computational resources with a 25.6× parameter increase and training time overhead ranging from 7.2× to 11.8×, necessitating careful consideration for practical deployment scenarios. Several efficiency optimization strategies, including knowledge distillation, progressive fusion, and architecture optimization, present viable paths for reducing computational complexity while maintaining performance benefits. The comprehensive analysis validates the effectiveness of multi-structure fusion approaches while highlighting important trade-offs between classification performance and computational efficiency in remote sensing applications.

The qualitative analysis reveals that the attention mechanism effectively leverages multi-structure features by focusing on discriminative spatial regions, though challenges remain in distinguishing visually similar categories such as different residential density types. The feature space visualization demonstrates that the fusion approach creates well-separated, discriminative representations, particularly for datasets with larger training sets and higher scene diversity.

Future research should prioritize the systematic evaluation of computationally efficient fusion architectures through knowledge distillation and network pruning techniques to reduce the current 25.6× parameter overhead while maintaining classification performance benefits. Adaptive fusion strategies that dynamically adjust feature combination weights based on dataset characteristics and scene complexity could improve performance consistency across diverse remote sensing scenarios. The integration of transformer-based architectures within the multi-structure framework presents promising opportunities for combining convolutional feature extraction with advanced attention mechanisms for enhanced scene understanding. Extending the approach to handle multi-modal remote sensing data, including the fusion of optical, SAR, and hyperspectral imagery, represents a critical advancement for comprehensive Earth observation applications. Comprehensive evaluation on larger operational datasets covering diverse geographic regions and environmental conditions will validate the framework's real-world applicability and establish practical deployment guidelines for operational remote sensing systems.

Acknowledgement

The authors would like to express their gratitude to the anonymous reviewers for their thorough and insightful remarks and suggestions. The authors would like to thank Laboratory 1231 of the Faculty of Computer Science at the Universitas Indonesia for their valuable discourse and contributions. The computation in this work was performed using the facilities of MAHAMERU BRIN HPC, National Research and Innovation Agency of Indonesia (BRIN).

References

[1] G. Cheng, J. Han, and X. Lu, "Remote sensing image scene classification: Benchmark and state of the art," *Proceedings of the IEEE*, vol. 105, no. 10, pp. 1865–1883, 2017.

[2] A. Thapa, T. Horanont, B. Neupane, and J. Aryal, "Deep learning for remote sensing image scene classification: A review and meta-analysis," *Remote Sensing*, vol. 15, no. 19, p. 4804, 2023.

[3] D. Wang, C. Zhang, and M. Han, "Mlfc-net: A multi-level feature combination attention model for remote sensing scene classification," *Computers & Geosciences*, vol. 160, p. 105042, 2022.

[4] Y. Yang and S. Newsam, "Comparing sift descriptors and gabor texture features for classification of remote sensed imagery," *Proceedings of the IEEE International Conference on Image Processing*, pp. 1852–1855, 2008.

[5] J. Hu and P. Guo, "Spatial local binary patterns for scene image classification," *Proceedings of the International Conference on Sciences of Electronics, Technologies of Information and Telecommunications*, pp. 326–330, 2012.

[6] B. Li, Y. Chen, Y. Chen, Y. Lu, and C. Ma, "Landslide detection based on glcm using sar images," *IGARSS 2020-2020 IEEE International Geoscience and Remote Sensing Symposium*, pp. 1989–1992, 2020.

[7] F. Liu, Y. Liu, and H. Sang, "Multi-classifier decision-level fusion classification of workpiece surface defects based on a convolutional neural network," *Symmetry*, vol. 12, no. 5, p. 867, 2020.

[8] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017.

[9] Q. Wang, S. Liu, J. Chanussot, and X. Li, "Scene classification with recurrent attention of vhr remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 2, pp. 1155–1167, 2019.

[10] R. Yang, F. Pu, Z. Xu, C. Ding, and X. Xu, "Da2net: Distraction-attention-driven adversarial network for robust remote sensing image scene classification," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2022.

[11] W. Xue, X. Dai, and L. Liu, "Remote sensing scene classification based on multi-structure deep features fusion," *IEEE Access*, vol. 8, pp. 28 746–28 755, 2020.

[12] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

[13] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2261–2269, 2017.

[14] B. Zhou, "Places: A 10 million image database for scene recognition," http://places2.csail.mit.edu/index.html, accessed: 2025-02-10.

[15] Q. Bi, B. Zhou, K. Qin, Q. Ye, and G.-S. Xia, "All grains, one scheme (agos): Learning multigrain instance representation for aerial scene classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–17, 2022.