MSC-D3Net: A Resilient Multi-Scale Learning Framework for Adaptive Cross-Domain Scene Understanding in Remote Sensing

Muhammad John Abbas[1], Muhammad Attique Khan[1], Ameer Hamza[2], Ghassen Ben Brahim[1], Jehad Ali[3]

[1]Center of AI, Prince Mohammad Bin Fahd University, Al-Khobar, Saudi Arabia
[2]Centre of Real Time Computer Systems, Kaunas University of Technology, Kaunas, Lithuania
[3]Department of AI Convergence Network, Ajou University, 206 Worldcup-ro, Yeongtong-gu, Suwon, Gyeonggi-do, South Korea

johnabbas@ieee.org, mkhan3@pmu.edu.sa, ameer.hamza@ktu.lt, gbrahim@pmu.edu.sa, Jehadali@ajou.ac.kr

Abstract

Remote sensing scene classification is important for accurate earth observation, environmental monitoring, and geographic analysis. However, intra-class diversity and domain-specific-variations make this classification quite difficult. This study aims to design a model that can generalize well on cross-domain datasets without accessing target domain data. For this purpose, a novel deep learning model "MSC-D3Net" is proposed that combines CNN and ViT features within hierarchical domain disentanglement and cross-scale semantic alignment. An adversarial domain discriminator module, along with uncertainty calibration, is also integrated. The model is trained and tested on three publicly available datasets and achieved very high in-domain accuracies above 98% and cross-domain accuracies above 93% on all datasets. The model also demonstrates low-uncertainty errors and outperforms existing architectures by a large margin.

Introduction

Background and Motivation

Deep learning (DL) has emerged as the prevailing method of interpreting Remote Sensing (RS) images (Q. Liu et al., 2025). This is evidenced by the high level of performance achieved through the use of DL techniques for various RS applications (i.e., scene classification, land-cover mapping, and large-scale Earth observation analysis) (Huang et al., 2025). CNN-based models, as well as Vision Transformer (ViT), are highly capable of capturing both multi-scale spatial relationships and long-range contextual dependencies that exist in high-dimensional aerial imagery and satellite imagery (i.e., use of different ends to generate a single image) (Li et al., 2025). In the majority of cases, DL based methods have been developed under the closed-domain premise that training and test datasets are drawn from the same distribution; unfortunately, in reality, this premise does not hold, and therefore, DL based methods have never been validated in a representative manner within the context of their intended use cases (Górriz et al., 2023). In real-world scenarios, RS images generally reflect highly heterogeneous characteristics due to differences among RS sensing technologies, their various spatial resolutions, how the images were collected (i.e., observing platforms), types of geographical conditions (i.e., seasons), etc (Jafarbiglu & Pourreza, 2022). Due to these considerable cross-domain variations, a significant domain shift will occur, which will greatly impair the ability of DL models to generalize to data outside the source dataset used for training. This essentially creates a situation where the use of a source dataset is possibly neither scalable nor reliable with respect to future use of the model.

Problem Statement

Complex scene understanding represents one of the hardest problems in the context of cross-domain generalization due to high intra-class diversity and low inter-class reparability of semantic categories (D. Wang et al., 2025). The primary approach used by traditional domain adaptation methods when addressing this problem has been to align feature distributions between the source and target domains while relying on access to target-domain data during training (M. Xie et al., 2025). However, in many instances, target-domain data is not available or is too difficult to obtain in RS cases, thereby necessitating the use of domain-generalized models that can function without needing target-domain supervision (Ghazaei & Aptoula, 2025). Therefore, given a source domain $D_s = \{(x_i^s, y_i^s)\}_{i=1}^{N_s}$ with distribution $P_s(x, y)$

and an unknown target domain D_t with distribution $P_t(x, y) \neq P_s(x, y)$, the objective is to learn a classifier $f: X \rightarrow Y$ from D_s that generalizes to D_t without having any access to the target domain data during training.

Related Work

The focus of current research into RS image comprehension has been to enhance in-domain model performance using DL architectures, especially CNNs and recently, ViTs. The capabilities of the methodologies were significantly improved by using DL approaches to learn both spatial and contextual discriminating features. Despite their successes in identifying features, the application of such techniques on previously unseen domains results in large performance drops because of the inherent distributional shifts associated with RS data, such as (C. Liu et al., 2025) proposed, the HMPNet framework uses a hypergraph-based and multimodal prototype approach by constructing calibrated image-text prototypes to capture entity relationships that exist in RS scenes. Furthermore, HMPNet achieves state-of-the-art results on the task of generating RS scene graphs. An advanced MinkUNet structure is proposed by (Zhou et al., 2025) as an approach for automated highway roadside point cloud classification with the Highways dataset and occlusion-aware augmented data for semi-supervised training, resulting in substantially improved mIoU and increased reliability on sparse outdoor scenes at the same time. Similarly, (Duan et al., 2025) developed a Swin Transformer with Multi-scale Fusing (STMSF) for RS scene analysis to compensate for the limited multi-scale modelling capability of conventional ViTs by adding multi-scale feature fusion capabilities and a spatial attention pyramid network. This STMSF has shown to yield superior performance than traditional CNNs or Transformer-based techniques across benchmark datasets.

Domain Invariant Feature Learning (DIFL) approach, currently being researched, involves exploring methods for developing features that are independent of domain and therefore can be transferred from one domain to multiple other domains (W. Xie et al., 2025). Although this approach has provided some level of robustness against domain shifts, it has not provided a consistently reliable way of balancing the discriminative representation learning and domain-invariance requirements. Additionally, majority of the research to date evaluating research on domain-invariant feature learning has been conducted under a limited set of domain conditions; thus, only limited evidence exists as to how well these techniques would be able to generalize to other, more diverse or previously unseen RS domains.

Research Gaps

Although existing techniques provide some robustness, several critical gaps remain that need to be addressed. These gaps include the limited feature representation, as current approaches mostly learned a single shared feature space that is either too domain-specific or generic, which fail to separate the domain-specific and domain-invariant features. In addition, despite exploring the multi-scale feature extraction in RS imagery, current methods cannot align the semantic consistency across various scales for cross-domain robustness. Passive domain invariance and the lack of uncertainty quantification are also major limitations of existing architectures. So, the question arises: *how can we design a neural network that explicitly disentangles domain-specific, domain-invariant, and semantic-discriminative features while incorporating multi-scale semantic consistency and uncertainty quantification to achieve robust cross-domain generalization without accessing target-domain data?* This question encompasses several challenges, such as, how to separate features into orthogonal subspaces; how to maintain semantic consistency across scales; how to quantify the uncertainty of model predictions, and how to combine CNNs and ViTs effectively for cross-domain RS scene understanding.

Contributions

To overcome these challenges, this research introduces "MSC-D3Net" (Multi-Scale Cross-Domain Dynamic Disentanglement Network), a novel DL framework that enables cross-domain generalization of RS images, with an emphasis on creating robust and transferable feature representations for complex scenes. It is expected that by reducing domain dependence at the level of the feature representation and maintaining semantic consistency across domains, generalization can be enhanced without requiring a target domain dataset. Our specific contributions are:

- A hierarchical domain disentanglement module is proposed that explicitly disentangles the domain-specific, domain-invariant, and semantic-discriminative features into orthogonal subspaces.
- A Cross-Scale Semantic Alignment mechanism is introduced that aggregates patch-level, region-level, and global features and maintains semantic consistency across them through prototype-based contrastive learning.
- Unlike passive domain invariance approaches, an Adversarial Domain Diversification (ADD) Module is incorporated that generates synthetic domain variations and trains a domain discriminator to distinguish domains.
- An evidential deep learning framework based on Dirichlet distributions is designed to quantify uncertainty in model predictions.

Proposed Architecture

Figure 1 shows the complete architecture of the proposed MSC-D3Net, which comprises five key modules, including

a Dual-branch encoder that fuses a CNN with ViT to extract both local and global features.

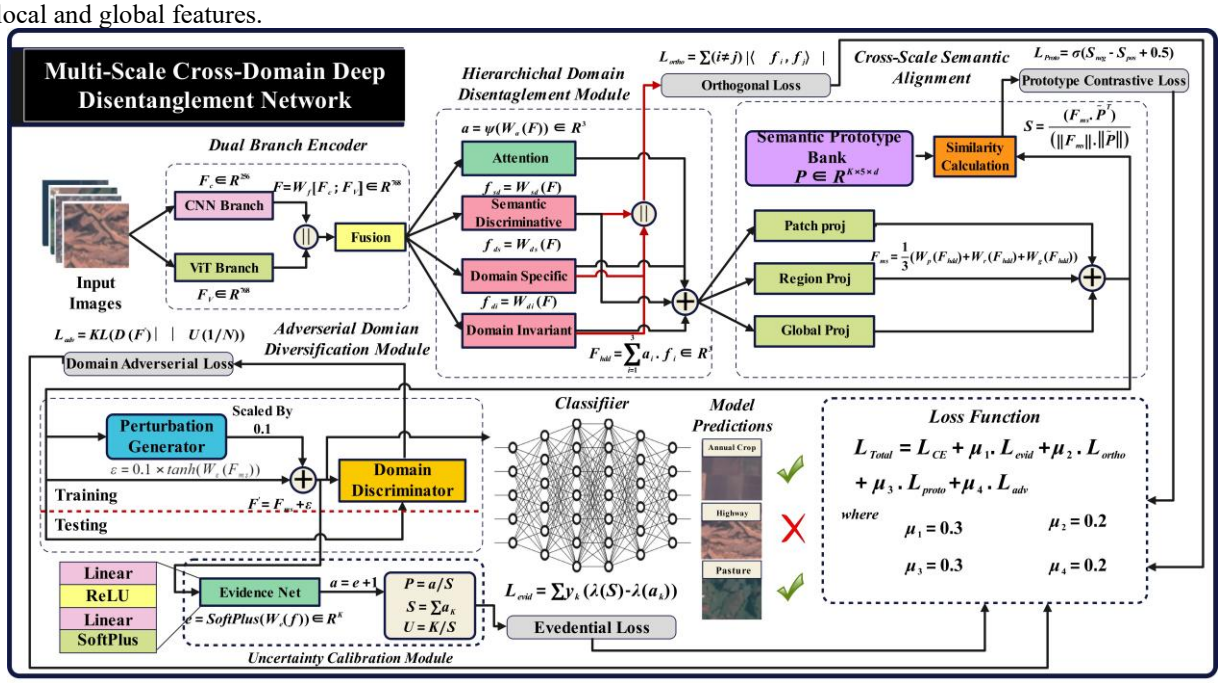

Figure 1: Proposed MSC-D3Net architecture for remote sensing scene classification

A hierarchical domain disentanglement module to separate orthogonal features into subspaces, a cross-scale semantic alignment mechanism to ensure semantic consistency via prototype contrastive learning, and an evidential uncertainty quantification algorithm based on Dirichlet distributions to quantify model predictions.

Data Flow

Given an input RS image $x \in \mathbb{R}^{224\times224\times3}$, the dual branch encoder produces fused features $F(x) \in \mathbb{R}^{768}$. These fused features passed through the HDD module, which processes these features into disentangled subspaces and then combines them in a single representation $F_{hdd}(x) \in \mathbb{R}^{512}$. The CSSA module takes this feature representation as input and projects it to different semantic scales. Output from all scales is aggregated $F_{ms}(x) \in \mathbb{R}^{512}$ and their similarities with learned prototypes are calculated. The aggregated multi-semantic feature representation is then passed to the ADD module, where domain perturbations are applied during the training stage to produce $F_{per}(x) \in \mathbb{R}^{512}$. The uncertainty calibration module takes these perturbed representations to estimate evidence for each class, providing both class predictions and uncertainty scores. Finally, a linear classifier produces logits for K semantic classes.

Dual Branch Encoder

RS images contain spatial textures, regional structures, and global information, which require both local and global feature extraction. For this purpose, a dual-branch encoder is introduced that combines CNN (extract local features) with ViT(excels at capturing global information) to produce a fused and compact feature representation. This module consists of two parallel branches: a CNN branch and a ViT branch. The CNN branch is composed of three sequential layers of Convolutional→BatchNorm→ReLU where the 1st convolutional layer maps 3 input channels to 64 output channels through a 7x7 kernel size with a stride of 2 and padding of 3. The 2nd convolutional layer, with a filter size of 3x3, maps these 64 channels to 128 channels, while the 3rd 3x3 convolutional layer maps them to a 256-channel feature representation. After the first sequential layer, a 3x3 MaxPooling layer with stride of 2 is added to down sample the feature representation while an adaptive average pooling is added after 3rd sequential layer. Mathematically, it can be defined as:

$$F_{CNN} = \Cap\left(h_3\left(h_2\left(\Cup\left(h_1(x)\right)\right)\right)\right) \quad (1)$$

Where

$$h_i(x) = \sigma\left(\prod(W_i \circledast x + b_i)\right),\ i \in \{1,2,3\} \quad (2)$$

Here, $\Cup$ and $\Cap$ represents MaxPooling and adaptive average pooling layers, σ represents ReLU activation, $\prod$ denotes batch normalization and $\circledast$ denotes a convolutional operation. W_i represents the convolutional weights and b_i represents bias terms of convolutional layers.

On the other hand, the ViT branch comprises a pretrained Swin Tiny transformer with 4 stages. In the 1st , 2nd , and 4th stages, 2 swin transformer blocks preceded by patch merging are present, while in the 3rd stage, 6 swin transformer blocks are integrated.

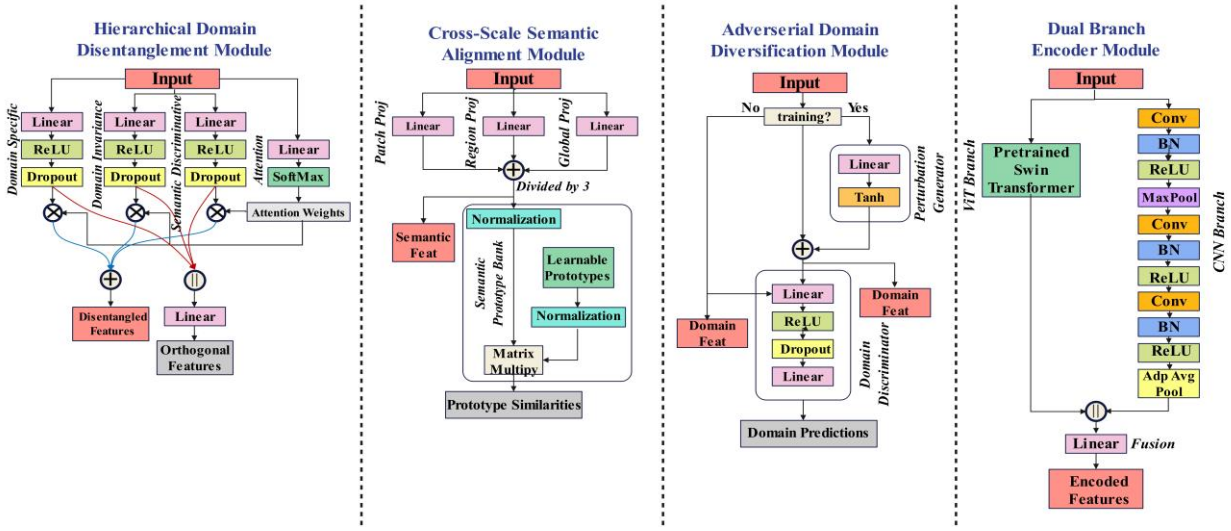

Figure 2: Proposed Hierarchical Domain Disentanglement Module, Cross-Scale Scene Alignment Module, Adversarial Domain Diversification Module and Dual Branch Encoder Module

$F_{vit} = SwinTinyTransformer(x)$ (3)

For efficiency, Swin Transformer computes self-attention within local windows rather than globally.

$$A(Q,K,V) = \psi\left(\frac{QK^T}{\sqrt{d_k}} + B\right)V \quad (4)$$

Here, ψ denotes SoftMax activation, B represents a learnable position bias, and attention is computed within 7x7 windows. Shifting windows in alternating blocks enables cross-window connections for global modeling. These CNN and ViT features are fused together through a linear transformation, which can be represented as:

$$F(x) = W_F[F_{CNN} \oplus F_{vit}] + b_F \quad (5)$$

Here, $\oplus$ denotes concatenation operation, W_F and b_F represents the weights and bias term of the linear layer. A visual representation of the dual-branch encoder is shown in Figure 2.

Hierarchical Domain Disentanglement Module

Traditional feature learning confuses domain-specific variations with semantic information, which results in poor performance on real-world data. The proposed HDD module disentangles the encoded features into domain-specific, domain-invariant, and semantic discriminative subspaces. Features are processed through dedicated projection layers simultaneously, where each layer consists of a linear layer followed by ReLU activation and a dropout layer (rate = 0.3). The original feature representation is also passed through the attention module, which is composed of a linear layer followed by a SoftMax activation to produce importance weights for all orthogonal features. The generated weights are multiplied by their corresponding feature representations, and the resulting weighted representations are then summed to produce the unified feature map. Mathematically:

$$F_{hdd} = w_1.f_{ds}(x) + w_2.f_{di}(x) + w_3.f_{sd}(x) \quad (6)$$

Where

$$f_{ds}(x) = \rho\big(\sigma(W_{ds}.x + b_{ds})\big) \quad (7)$$

$$f_{di}(x) = \rho\big(\sigma(W_{di}.x + b_{di})\big) \quad (8)$$

$$f_{sd}(x) = \rho\big(\sigma(W_{sd}.x + b_{sd})\big) \quad (9)$$

$$w_1, w_2, w_3 = \psi(W_A.x + b_A) \quad (10)$$

Here, ρ denotes dropout layer, $W_{ds}, W_{di}, W_{sd}, W_A$ represents the weights of linear layers while $b_{ds}, b_{di}, b_{sd}, b_A$ represents the corresponding bias terms.

Other than weighted addition, all the learned orthogonal features undergo concatenation followed by a linear layer to ensure subspace independence.

$$f_{ortho}(x) = W_o[f_{ds}(x) \oplus f_{di}(x) \oplus f_{sd}(x)] + b_o \quad (11)$$

After that, normalized inner products between all pairs of subspace features are computed to calculate orthogonality loss.

$$\mathcal{L}_{ortho} = \frac{1}{3}\left(\left|\left\langle \frac{f_{ds}}{\|f_{ds}\|}, \frac{f_{di}}{\|f_{di}\|}\right\rangle\right| + \left|\left\langle \frac{f_{ds}}{\|f_{ds}\|}, \frac{f_{sd}}{\|f_{sd}\|}\right\rangle\right| + \left|\left\langle \frac{f_{di}}{\|f_{di}\|}, \frac{f_{sd}}{\|f_{sd}\|}\right\rangle\right|\right) \quad (12)$$

This orthogonality loss is minimized to enforce pairwise orthogonality. A visual representation of this HDD Module is shown in Figure 2.

Cross-Scale Semantic Alignment Module

Remote sensing data has an inherent multi-scale nature, in which some classes are distinguished by texture, while others are distinguished by regional structure or global layout. To ensure semantic consistency across various scales, the

CSSA module is introduced, which extracts hierarchical features through three parallel projection layers. The patch projection layer captures low-level fine-grained patterns, the region projection layer captures mid-level regional patterns, and the global projection layer captures the overall layout. All these feature representations are aggregated and scaled by 1/3 to ensure equal contribution of each scale. Mathematically:

$$F_{ms} = \frac{1}{3}\left(W_p(x) + W_r(x) + W_g(x)\right) \quad (13)$$

Here, W_p, W_r and W_g represents the weight metrics of patch, region, and global projection layers. To handle the intra-class diversity of RS imagery, we introduced prototype contrastive learning, where we maintain $M = 5$ prototypes for each of the K classes. These prototypes $\mathbb{P} \in \mathbb{R}^{K\times M\times 512}$ are randomly initialized from $\mathcal{N}(0,0.02)$ where $\mathcal{N}$ represents the normal distribution. Multiple prototypes per class handle the diversity within each category, providing robustness against intra-class variations across domains.

For the feature vector F_{ms}, the normalized cosine similarity to all the prototypes is calculated. For a prototype $m \in M$, the similarity to class $k \in K$ is computed as:

$$s_{km}(x) = \frac{F_{ms}(x)^T P_{km}}{\|F_{ms}(x)\|_2 . \|P_{km}\|_2} \quad (14)$$

Where $P_{km} \in \mathbb{R}^{512}$ is the m-th prototype of class k. The prototypes for class k are aggregated together using Max-Pooling, which selects the most representative prototype for each class.

$$s_k(x) = \underbrace{max}_{m\in\{1,\dots,M\}} s_{km}(x) \quad (15)$$

To enforce that features are more similar to correct class-prototypes, prototype contrastive loss is introduced, which can be defined as:

$$\mathcal{L}_{proto}(x,y) = \mathbb{E}_{x,y}\left[\max\left(0, \underbrace{max}_{k\neq y} s_k(x) - s_y(x) + \gamma\right)\right] \quad (16)$$

Where, $\mathbb{E}_{x,y}$ represents the expected value over all (x,y) pairs, y denotes the true class label, $s_y(x)$ is the similarity to the correct class while $\underbrace{max}_{k\neq y} s_k(x)$ is the maximum similarity to any incorrect class and $\gamma = 0.5$ is the marginal hyperparameter. The objective is to increase the margin between correct and incorrect class similarities. This contrastive learning ensures semantic consistency by pulling the features towards their class prototypes. The visual representation of the CSSA module is shown in Figure 2.

Adversarial Domain Diversification Module

Training on limited data from a source domain might not be enough to model cross-domain variations. Therefore, the ADD module is incorporated in the network that generates perturbations in feature space to simulate domain variations. For this purpose, a perturbation generator is designed, which is composed of a linear layer followed by a tanh activation function, which bounds the perturbations to [-0.1, 0.1] to prevent extreme variations. The generated perturbations are then added to the original feature space.

$$\delta(x) = 0.1 \times \tanh\left(W_{pg}\left(F_{ms}(x)\right) + b_{pg}\right) \quad (17)$$

$$F_{per}(x) = F_{ms}(x) + \delta(x) \quad (18)$$

Here, W_{pg} and b_{pg} represents the

These perturbed features are then passed to the domain discriminator, which is designed to give adversarial feedback by trying to predict the domain of the features. This domain discriminator is composed of two linear layers, where 1st linear layer maps the feature space to 256 channels, and 2nd layer maps these 256 channels to the number of domains. A ReLU activation followed by the dropout layer (rate = 0.3) is present between both layers. If this discriminator successfully predicts the correct domain, it means that the feature extractor has not removed sufficient domain-specific patterns. An adversarial loss is computed from a module whose objective is to confuse the discriminator by making features indistinguishable across domains. It can be defined as:

$$\mathcal{L}_{adv} = D_{kL}\left(\psi\left(d\left(F_{per}\right)\right) \| \mathcal{U}(N_{domians})\right) \quad (19)$$

Here, $D_{kL}(.\|.)$ represents Kullback-Leibler divergence, $\mathcal{U}(N_{domians})$ represents uniform distribution over domains, and $d(.)$ represents the domain discriminator. Perturbations are applied only during training and model is penalized for producing domain-discriminative features.

During testing stage, perturbations are disabled and model used learned non-discriminative feature representations. A complete diagram of ADD module is shown in Figure 2.

Uncertainty Calibration Module

Standard classifiers produce overconfident predictions without quantifying their uncertainty, which may result in serious consequences in real-world scenarios. Therefore, we proposed evidence-based uncertainty quantification, which estimates evidence for each class instead of directly predicting class probabilities. Besides going to the classifier, the final feature representation also goes to the uncertainty calibration module, where it passes through Evidence-Net to produce an evidence vector for K classes. Evidence-Net is composed of two linear layers; the first one maps feature dimensions from 512 to 256 and 2nd one maps it further to the number of classes. A ReLU activation is present between both layers, and a SoftPlus activation is added after 2nd layer. This Evidence-Net produces an evidence vector $e = [e_1, \dots., e_k]$ where $e_k \geq 0$ represents the amount of evidence supporting class k. 1 is added to this evidence in order to satisfy the Dirichlet distribution. So total evidence is defined as:

$$S = \sum_{k=1}^{K} \alpha_k = \sum_{k=1}^{K} (e_k + 1) \quad (20)$$

In addition, the Dirichlet distribution is represented as:

$$p(\mathbb{P}|\alpha) = \frac{1}{B(\alpha)} \prod_{k=1}^{K} p_k^{\alpha_k - 1} \quad (21)$$

Here, $B(\alpha) = \frac{\prod_{k=1}^{K} \tau(\alpha_k)}{\tau(\sum_{k=1}^{K} \alpha_k)}$ is the multivariate beta function. This Dirichlet distribution models the uncertainty over the probability simplex, where high evidence concentrates the

distribution near a confident prediction and low evidence spreads it uniformly. Thus, uncertainty is computed as:

$u = \frac{K}{S} \in [0,1]$ (22)

Where $u \approx 0$ means high evidence and confident prediction, while $u \approx 1$ means low evidence and maximum uncertainty.

An evidential loss is computed from this module to encourage the model to place evidence in the correct class. It is defined as:

$\mathcal{L}_{evid} = \mathbb{E}_{(x,y)\sim D}[\sum_{k=1}^{K} \mathbb{I}(y = k)(\lambda(s) - \lambda(\alpha_k))]$ (23)

Here, $\lambda(.)$ represents the digamma function $\lambda(.) = \frac{d}{dx} log\Gamma(x)$ and $\mathbb{I}(y = k)$ is the indicator function, which means 1 if true and 0 otherwise. Unlike cross-entropy, which deals with probabilities, this evidential loss optimizes the evidence distribution.

Classification Head and Loss Function

A simple linear classifier maps the final feature space to class logits, where a SoftMax activation converts them into probabilities. Other than SoftMax prediction, evidence-based predictions are also computed to provide a dual calibration mechanism. The total loss function is composed of all the individual loss components. It can be defined as:

$\mathcal{L}_{Total} = \mathcal{L}_{CE} + \mu_1.\mathcal{L}_{evid} + \mu_2.\mathcal{L}_{ortho} + \mu_3.\mathcal{L}_{proto} + \mu_4.\mathcal{L}_{adv}$ (24)

Here, $\mathcal{L}_{CE}$ represents total cross-entropy loss, which can be defined as:

$\mathcal{L}_{CE} = -\sum_{k=1}^{K} y_k \log(\hat{y}_k)$ (25)

$\mathcal{L}_{evid}, \mathcal{L}_{ortho}, \mathcal{L}_{proto}$ *and* $\mathcal{L}_{adv}$ are defined in Eq (23), (12), (16) and (19) respectively and $\mu_1 = 0.3$, $\mu_2 = 0.2$, $\mu_3 = 0.3$, $\mu_4 = 0.2$ denotes the weighting parameters.

Experimental Setup

Three publicly available datasets are used to train and test the proposed architecture. These datasets include EuroSAT (Helber et al., 2019) (an open-source dataset with 27000 images categorized into 10 classes), NWPU (Haikel, 2021) (a uni-label dataset comprising 10500 images divided into 12 classes) and PatternNet (Zhou et al., 2018) (a high-resolution dataset with 38 classes, each containing 800 images). A train-val-test split of 80:10:10 was applied on all the datasets. Model was trained using AdamW optimizer with learning rate of 1e-4. Epoch size was set to 100 while batch size was set to 128. The entire experimental process for the proposed model is performed on a dedicated server with 512 GB of total GPU memory and 1.5 TB of RAM.

Results and Discussion

Overall Performance

Table 1 shows the overall performance of the proposed model on all three datasets. As shown, our proposed MSC-D3Net demonstrates outstanding performance, achieving an overall accuracy of 98.48%, 98.57%, and 99.73% on EuroSAT, NWPU, and PatternNet datasets, respectively.

Metric	EUROSAT	NWPU	Pattern-Net
ECE	0.0123	0.0138	0.0029
Brier Score	0.0256	0.0261	0.0052
Overall Accuracy (%)	98.4815	98.5714	99.7368
Precision (Micro)	0.9848	0.9857	0.9974
Recall (Micro)	0.9848	0.9857	0.9974
F1 Score (Micro)	0.9848	0.9857	0.9974
Precision (Macro)	0.9838	0.9836	0.9975
Recall (Macro)	0.9850	0.9849	0.9975
F1 Score (Macro)	0.9843	0.9839	0.9975
Precision (Weighted)	0.9850	0.9863	0.9974
Recall (Weighted)	0.9848	0.9857	0.9974
F1 Score (Weighted)	0.9848	0.9858	0.9974
Balanced Accuracy	0.9850	0.9849	0.9975
Average Specificity	0.9983	0.9987	0.9999
MCC	0.9831	0.9843	0.9973
Cohen's Kappa	0.9831	0.9843	0.9973
Failure Rate	0.0152	0.0143	0.0026
High-Confidence Failures	0.0097	0.0124	0.0020
Mean Uncertainty	0.0263	0.0791	0.0537
Avg. Inference Time (ms)	44.0279	12.6631	41.2230
Throughput (samples/sec)	2907.2473	10108.1309	3105.0590

Table 1: Overall Performance of proposed "MSC-D3Net" on EuroSAT, NWPU and PatternNet dataset

The micro, macro, and weighted precision, recall, and F1-score values for all the datasets exceed 98%, which indicates overall stable and balanced classification performance. Low ECE (Expected Calibration Error) and Brier score values suggest that model predictions are well calibrated and reliable. Moreover, the balanced accuracy, average specificity, MCC (Matthews Correlation Coefficient), and Cohen's Kappa score are close to 1, which indicates a highly reliable and unbiased classification across all classes. These scores confirm that the model can perform exceptionally well even in imbalanced conditions and maintains a strong agreement between actual and predicted labels. The failure rate and high-confidence failures are also very small, which means that the model is highly confident in its predictions and rarely misclassifies while giving high confidence scores. The average inference time of 44.02ms, 12.66ms, and 41.22 ms for EuroSAT, NWPU, and PatternNet datasets and high

throughput indicates that this model is well suited for real-world deployment under resource-constrained environments.

Cross-Dataset Evaluation

Table 2 shows the results of cross-domain evaluation results of proposed model.

Train Dataset	Test Dataset	Accuracy	Precision	Recall	F1-score
NWPU	EuroSAT	94.67	93.98	94.37	94.12
NWPU	PatternNet	95.13	94.98	95.01	94.99
EuroSAT	NWPU	93.20	93.20	93.20	93.20
EuroSAT	PatternNet	94.80	95.00	94.57	94.87
PatternNet	NWPU	95.16	95.16	95.00	95.43
PatternNet	EuroSAT	96.01	95.34	95.99	95.89

Table 2: Cross-domain evaluation results of proposed "MSC-D3Net" model

The cross-dataset evaluation results demonstrate that the proposed model can perform exceptionally well even on unseen domains. The strong cross-domain generalization is one of the main objectives of this model, and these results fully satisfied the rationale behind this architecture. As shown in Table 2, when the model was trained on one dataset and tested on another, the accuracy remained above 93% in all cases. The consistent accuracy, precision, recall, and F1-score indicates that model does not overfit to a single dataset. The best cross-domain performance was observed when the model was trained on PatternNet and EuroSAT, while the worst performance was observed when the training dataset was EuroSAT and the testing dataset was NWPU. Overall, high cross-domain generalization results proved the effectiveness of proposed domain-generalization strategy.

Ablation Study

Figure 3 shows the results of an ablation study on both in-domain and cross-domain configurations. The left graph shows the individual effect of each component on the in-domain model performance. As shown in graph, full model shows an overall accuracy of 98.48%, 98.57%, and 99.73% on EuroSAT, NWPU, and PatternNet datasets, respectively. This accuracy drops to ~3% on removing HDD module and ~3.5% on CSSA module, indicating the importance of these modules. On removal of ADD module, a decline of ~1.5% was observed and removing VIT branch from dual-branch encoder results in decline of ~4% while removal of ViT results in decline of 3.5%. These results indicates the individual impact of each module where CNN affects the most, followed by CSSA, ViT and HDD. Right graph shows cross-domain generalization performance for this ablation study. Graphs shows the average accuracy for all three datasets. A similar pattern is observed in cross-domain performance, but the accuracy drop is significantly larger than in-domain performance. These results validate the individual contribution of each component, necessitating the combined framework.

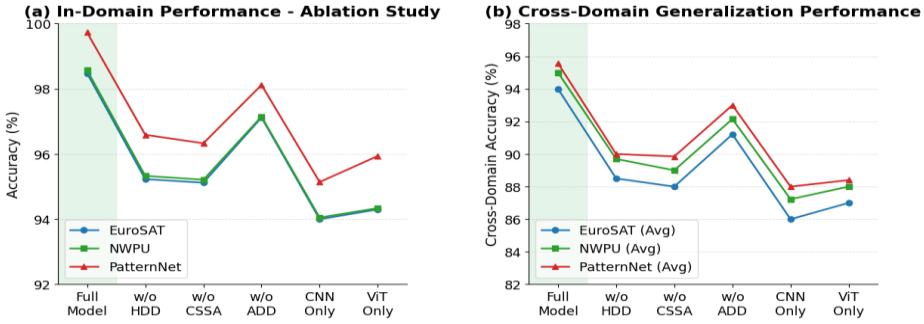

Figure 3: Results of ablation study on In-domain and cross-domain performance

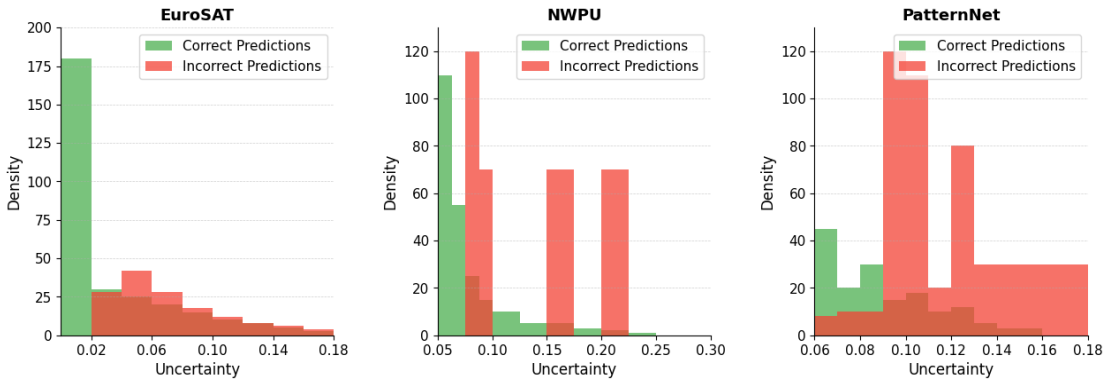

Figure 4: Uncertainty distribution of incorrect vs correct model predictions on EuroSAT, NWPU and PatternNet dataset

Model	Dataset	Accuracy
Global Optimal structured loss (Liu et al., 2020)	EuroSAT	88.68
InceptionV1 (Szegedy et al., 2015)		88.51
MobileNetV2 (Sandler et al., 2018)		87.52
EfficientNet (Tan & Le, 2019)		85.23
Proposed		**98.48**
Global Optimal structured loss(Liu et al., 2020)	NWPU	90.30
IBNR-65 + Densenet-64 (Albarakati et al., 2024)		91.70
WSADAN-ResNet50 (Liming et al., 2024)		92.63
Khan, J.A., et al. (Khan et al., 2024)		93.3
Proposed		**98.57**
FusionNet-RS (Aparanji et al., 2025)	PatternNet	95.98%
ResNet-50 + HAPF (Hu et al., 2025)		97.46%
HQSAN(R. Wang et al., 2025)		98.09%
VGG-16+HAPF (Hu et al., 2025)		98.19%
Proposed		**99.73**

Table 3: Comparative Analysis with SOTA models

Uncertainty Distribution

Figure 4 shows the uncertainty distribution of model for all three datasets. As shown in the graphs, most correct predictions have low uncertainty while the incorrect predictions have high certainty. Only a small number of confident predictions are incorrect which indicates that model is well aware of its confidence level and does not make random confident guesses. This type of behavior is extremely important for real-world applications of remote sensing, where incorrect confident predictions might result in serious consequences. The proposed evidential framework successfully discriminates between confident and uncertain predictions.

Comparative Analysis with SOTA

Table 3 shows the comparative analysis of proposed model with SOTA architectures on all three datasets. The statistics shown in table demonstrates the supremacy of proposed technique over existing models on all three datasets. Our proposed model achieved 98.48% accuracy on EuroSAT which is higher than best existing model (88.68%). Similarly, the 98.57% on the NWPU dataset easily surpassed the existing model with 93.3% accuracy, and 99.73% accuracy on PatternNet outperformed the compared best-performed model (98.19%). These results highlight the superiority of proposed architecture in both accuracy and robustness. Hence, MSC-D3Net sets a new performance benchmark for RS scene classification.

Conclusion

This paper presents “MSC-D3Net”, a dual-branch deep learning architecture that fuses CNN and ViT features with feature disentanglement and multi-scale semantic alignment. Additionally, modules like adversarial domain diversification and Evidence-based uncertainty calibration are integrated to enhance robustness and reliability. The proposed model achieved consistently high in-domain and cross-domain accuracies across the EuroSAT, NWPU, and PatternNet datasets. Experimental results also demonstrates better calibration, low failure rates and superior results compared to existing DL techniques. However, despite the exceptional performance, the proposed model exhibits certain limitations, which includes the high architectural complexity, which requires more computational resources during training. Future work should focus on making lightweight architectures and testing on diverse real-world remote sensing domains.

References

Albarakati, H. M., M. A. Khan, A. Hamza, F. Khan, N. Kraiem, L. Jamel, L. Almuqren and R. Alroobaea 2024. 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.

Aparanji, S., T. Nayana and S. Nikitha 2025. FusionNet-RS: A Deep Feature Fusion Model for Remote Sensing Image Classification on PatternNet. 2025 International Conference on Computing Technologies (ICOCT), IEEE.

Duan, Y., C. Song, Y. Zhang, P. Cheng and S. Mei 2025. STMSF: Swin Transformer with Multi-Scale Fusion for Remote Sensing Scene Classification. Remote Sensing **17**(4): 668.

Ghazaei, E. and E. Aptoula 2025. Text-conditioned State Space Model For Domain-generalized Change Detection Visual Question Answering. arXiv preprint arXiv:2508.08974.

Górriz, J. M., I. Álvarez-Illán, A. Álvarez-Marquina, J. E. Arco, M. Atzmueller, F. Ballarini, E. Barakova, G. Bologna, P. Bonomini and G. Castellanos-Dominguez 2023. Computational approaches to explainable artificial intelligence: advances in theory, applications and trends. Information Fusion **100**: 101945.

Haikel, H. 2021. NWPU-RESISC45 Dataset with 12 classes. Figshare: London, UK.

Helber, P., B. Bischke, A. Dengel and D. Borth 2019. 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 **12**(7): 2217–2226.

Hu, Z., M. Gong, Z. Dong, Y. Lu, J. Li and Y. Zhao 2025. Heterogeneity-aware pruning framework for personalized federated learning in remote sensing scene classification. Knowledge-Based Systems **311**: 113007.

Huang, D., Z. Zhou, Z. Zhang, Q. Dai, H. Lu, Y. Li and Y. Huang 2025. Land Use/Land Cover Remote Sensing Classification in Complex Subtropical Karst Environments: Challenges, Methodological Review, and Research Frontiers. Applied Sciences **15**(17): 9641.

Jafarbiglu, H. and A. Pourreza (2022). A comprehensive review of remote sensing platforms, sensors, and applications in nut crops. Computers and Electronics in Agriculture **197**: 106844.

Khan, J. A., M. A. Khan, M. Al-Khalidi, D. A. AlHammadi, A. Alasiry, M. Marzougui, Y. Zhang and F. Khan 2024. 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.

Li, Z., T. Shang, P. Xu and Z. Deng 2025. Place Recognition Meet Multiple Modalitie: A Comprehensive Review, Current Challenges and Future Directions. arXiv preprint arXiv:2505.14068.

Liming, W., Q. Kunlun, Y. Chao and W. Huayi 2024. Weakly supervised scale adaptation data augmentation for scene classification of high-resolution remote sensing images. National Remote Sensing Bulletin **27**(12): 2815–2830.

Liu, C., C. Deng, H. Yu, Q. Yan, L. Xu, T. Zhang, X. Sun and K. Fu 2025. Hypergraph-Guided Multimodal Prototype for Remote Sensing Scene Understanding." IEEE Transactions on Geoscience and Remote Sensing.

Liu, P., G. Gou, X. Shan, D. Tao and Q. Zhou 2020. Global optimal structured embedding learning for remote sensing image retrieval. Sensors **20**(1): 291.

Liu, Q., T. Huang, Y. Dong, J. Yang and W. Xiang 2025. From Pixels to Images: Deep Learning Advances in Remote Sensing Image Semantic Segmentation. arXiv preprint arXiv:2505.15147.

Sandler, M., A. Howard, M. Zhu, A. Zhmoginov and L.-C. Chen 2018. Mobilenetv2: Inverted residuals and linear bottlenecks. Proceedings of the IEEE conference on computer vision and pattern recognition.

Szegedy, C., W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke and A. Rabinovich 2015. Going deeper with convolutions. Proceedings of the IEEE conference on computer vision and pattern recognition.

Tan, M. and Q. Le 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. International conference on machine learning, PMLR.

Wang, D., Z. Yan and P. Liu 2025. Fine-Grained Interpretation of Remote Sensing Image: A Review. Remote Sensing **17**(23): 3887.

Wang, R., Z. Zhang, Y. Shen, Z. Sui and K. Lin 2025. Hqsan: a hybrid quantum self-attention network for remote sensing image scene classification: R. Wang et al. The Journal of Supercomputing **81**(15): 1422.

Xie, M., J. Liu, Y. Li and C. Yi 2025. An unsupervised domain adaptation method for intelligent fault diagnosis based on target feature enhancement and feature-boundary alignment. Journal of Intelligent Manufacturing: 1–15.

Xie, W., Z. Liu, L. Zhao, M. Wang, J. Tian and J. Liu 2025. DIFLF: A domain-invariant features learning framework for single-source domain generalization in mammogram classification. Computer Methods and Programs in Biomedicine **261**: 108592.

Zhou, D., Y. Yi, Y. Wang, Z. Shao, Y. Hao, Y. Yan, X. Zhao and J. Guo 2025. Enhancing Highway Scene Understanding: A Novel Data Augmentation Approach for Vehicle-Mounted LiDAR Point Cloud Segmentation. Remote Sensing **17**(13): 2147.

Zhou, W., S. Newsam, C. Li and Z. Shao 2018. PatternNet: A benchmark dataset for performance evaluation of remote sensing image retrieval. ISPRS journal of photogrammetry and remote sensing **145**: 197–209.

Zhu, Q., J. Xiao and L. Fan 2025. IndoorMS: A Multispectral Dataset for Semantic Segmentation in Indoor Scene Understanding. IEEE Sensors Journal.