Minimax Theory and its Applications

Volume 11 (2026), No. 4s, 76 to 100

An Adaptive Transformer Framework Integrating Scarcity-Guided Augmentation and Reliability-Aware Domain Alignment for Few-Shot Satellite Image Classification

Komatigunta Nagaraju[1], Dr.Malladi Ravisankar[2]

[1]Research Scholar, Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation,Vaddeswaram,Guntur 522302,A.P.,India, Email: komatigunta.nagaraju@gmail.com
[2]Professor in the Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation,Vaddeswaram,Guntur 522302,A.P.,India, Email: drmrsankar@gmail.com

Abstract
Satellite image classification is essential for land-use analysis, geographic monitoring, and environmental assessment; however, geographic domain shifts, class imbalance, and limited labeled target-domain samples remain significant challenges. Existing deep learning approaches often rely on abundant labeled data and may experience reduced classification performance when transferred across geographic regions, while conventional augmentation and domain adaptation methods may inadequately address class scarcity and synthetic-sample reliability. This study proposes an adaptive Transformer framework integrating scarcity-guided augmentation and reliability-aware domain alignment for few-shot satellite image classification. The framework uses the Domain Shift across Geographic Regions (DSGR) dataset with corresponding fMoW satellite imagery, covering 41 land-use classes across six geographic regions. Implementation is planned in Python using PyTorch, with a pretrained Swin Transformer-Tiny backbone. The proposed method estimates class scarcity, allocates adaptive augmentation, generates class-conditioned synthetic images using a generative adversarial network, and filters synthetic samples according to class consistency. Scarcity-Conditioned Token Adaptation (SCTA) refines token representations, while adaptive local–global feature fusion captures spatial details and contextual information. Reliability-aware class-conditional prototype alignment reduces geographic feature discrepancies, and class-balanced focal loss addresses imbalanced classification. Evaluation will consider 1-, 5-, 10-, and 20-shot target-support settings using accuracy, precision, recall, Macro-F1, and balanced accuracy. Preliminary illustrative results indicate 95.0% accuracy, 92.4% Macro-F1, 92.8% balanced accuracy, and 90.6% minority-class recall. Compared with the illustrative Inception-v3 accuracy of 92.0%, the proposed accuracy represents a 3.0-percentage-point improvement, equivalent to approximately 3.26% relative improvement. The framework aims to improve reliable cross-region classification under data scarcity, with potential applications in geographically diverse satellite-image analysis.

Keywords: Satellite Image Classification, Few-Shot Learning, Swin Transformer, Geographic Domain Shift, Scarcity-Guided Augmentation, Conditional Generative Adversarial Network

1.Introduction

Satellite image classification is essential for land use mapping, urban planning, environment surveillance, agriculture inspection, and infrastructure management[1], [2]. The deep learning techniques, especially Transformer models, have proved effective in terms of learning complicated visual representation and contextual relationship in satellite images[3], [4]. However, the reliable classification is hard due to the lack of enough training samples for specific land use classes and geographical differences between source and target domains. Insufficient samples make the learning biased towards common classes; while geographical domain shifts generate different buildings, vegetation, surfaces, and acquisition settings[5], [6]. The challenges are amplified in few-shot learning setting where few target domain samples are available for adaptation[7], [8]. Many previous works have explored conventional data augmentation, generative adversarial network (GAN), class balanced learning, Transformer feature extraction, and domain adaptation in improving the classification

performance[9], [10]. However, traditional augmentation can not provide enough diversity for seriously unbalanced classes, and its samples generated by GANs can have unreasonable structure and class property. Satellite imagery can vary significantly in terms of visual features depending on the geographic region, reflecting the diversity in these attributes. As a result, source-trained models are likely to perform poorly in areas where they are not originally trained on [11], [12]. Furthermore, class imbalance can skew the decision boundaries, as models tend to focus on the more common classes over the rare land use classes. The difficulties underscore the need to investigate classification techniques in the various geographic regions, especially when target domain samples are scarce and a robust recognition of underrepresented classes is required.

The proposed framework is based on DSGR benchmark and the corresponding fMoW satellite images to analyze the classification of various geographic areas and under-represented classes. First, class-wise scarcity values are computed from the training set to determine the augmentation needs for each category. Standard augmentations as well as conditional GAN generation, where feasible, are utilized for augmenting minority classes, and generated examples are verified for their quality and consistency. After that, discriminative visual features are extracted from the images using a pre-trained Swin Transformer. Then, reliability-guided domain alignment employs labeled samples from the target domain to determine class prototypes and considers the unreliable visual representations with low influence on the alignment process. Also, class-balanced focal loss function helps to minimize the effect of the majority class in the training procedure.

1.1 Problem Statement

Class imbalance and geographical differences pose a challenge for satellite imagery classification due to the lack of enough examples of underrepresented classes and geographical differences in buildings, vegetation, surfaces, and other factors that influence images and cause domain shifts which negatively affect generalizability of the classifier[13], [14]. Traditional data augmentation techniques might not be helpful enough in increasing the number of minority classes. Moreover, image synthesis techniques may generate unrealistic data and misclassified classes. Traditional techniques of domain adaptation do not take into account the importance of limited target domain data. These challenges are especially important in few-shot setting where only a small number of target domain samples is provided. Hence, it becomes crucial to have an adaptive classification approach to solve the problem of class imbalance, increase the quality of training samples, and ensure reliable cross-domain feature alignment.

1.2 Research Motivation

The motivation behind this work is the necessity to design accurate satellite image classification frameworks that function well using fewer labels and changes in geographical distribution of the data. Scarcity requires specific data augmentation, while cross-domain adaptation necessitates proper estimation of the reliability of target features. The integration of these abilities into a Transformer architecture can potentially help with improving the performance of minority class detection and decreasing performance drops with changing geography.

1.3 Significance of the Study

This research proposes an integrated learning approach that can be used to tackle issues like data scarcity, imbalanced data, and geographical domain shift during the process of satellite image classification. Augmentation guided by data scarcity will help in tackling scarcity of categories, and alignment guided by reliability will limit the effect of uncertain representation of the target domain. The analysis carried out in different few-shot scenarios and in different geographical locations will reveal how well the two aspects mentioned work.

1.4 Key Contributions

- Develops an adaptive Transformer framework addressing data scarcity, class imbalance, and geographic domain shifts in few-shot satellite image classification.
- Introduces scarcity-guided augmentation with quality-filtered conditional GAN synthesis to improve minority-class representation and training diversity.
- Employs pretrained Swin Transformer to extract discriminative hierarchical features and capture contextual information across diverse satellite-image categories.

- Proposes reliability-aware domain alignment that weights class prototypes using target-support availability and feature consistency to improve cross-domain adaptation.
- Establishes comprehensive evaluation across few-shot settings, imbalance levels, and geographic domains using Macro-F1, balanced accuracy, and ablation studies.

2. Related Works

Zhu et al. [15]Proposed Meta Visual Prompt Tuning (MVP) by combining parameter efficient prompt tuning and meta-learning for few-shot remote sensing scene classification. The pretrained Vision Transformer backbone is the same, and only the parameters of the prompt are updated, which can reduce the phenomenon of overfitting and the memory consumption. Data can be diversified even more by using patch embedding recombination.Recombination of patches further increases data diversity. These experiments show improved results in various few-shot scenarios and cross domain adaptation, but further experiments could be conducted with other datasets from the remote sensing field.

Qin et al. [16]proposed a subspace-based few-shot remote sensing scene classification approach based on multiscale feature learning and attention mechanisms. The multiple branches are designed for extracting features at different scales; the self-attention is for obtaining global information; the channel attention is for feature fusion. Subspace modeling finds common class attributes which eliminates irrelevant background noise. Experiments are conducted on public datasets and the performance is competitive, yet, computational complexity and cross-domain generalization needs to be further explored.

Deng et al. [17]Proposed Dual-Alignment CLIP (DA-CLIP) for few-shot remote sensing scene classification to address with the bias in text representation and uncertainty in image similarity. A novel method is proposed, which combines the pretrained knowledge and task-specific knowledge by integrating multilevel image–text alignment, image–image feature alignment and adaptive knowledge fusion. It has been tested on 12 remote sensing datasets and is proven to work, but more computational expense and robustness to significant shifts in the domain remain to be explored.

Du et al.[18]proposed MFE-NM, an unsupervised domain adaptation method for remote sensing scene classification. Middle-layer feature extraction eliminates intraclass variations, and nuclear norm maximization maximizes prediction diversity/discriminability in the presence of class imbalance. Three public datasets are used for experiments, which show competitive performance. Some variance in representations may be achieved by random sampling of intermediate features, but this will need to be studied in greater detail to determine computational efficiency and generalization across a wider range of domains.

Kim et al. [19]Proposed a self-attention-based feature fusion network (SAFFNet) for few-shot remote sensing scene classification. This approach captures multi-scale texture-semantic features and adaptively chooses and reweights informative representations with self-attention, where weights for features are optimized based on support samples. Results of experiments performed on three public data sets show substantial gains in classification accuracy compared to current techniques but may cause loss of robustness in other remote sensing applications due to reliance on a limited number of support samples.

Jiang et al. [20]proposed a Multi-scale Graph-based Feature Fusion (MGFF) approach to tackle few-shot remote sensing scene classification. The method is an approach based on graph representations that use traditional image features to encode spatial relations, and extends the approach by incorporating multiscale graph features to enhance sample discrimination. Experiments conducted on two public datasets show that the proposed classification is more accurate than other methods; however, the construction of the graph and multiscale fusion can increase the complexity of computation and dependency of the representation quality of the features.

Luo et al. [21]Proposed a two-stage domain adaptation framework (proposed DACST) for cross-spatiotemporal land-cover classification based on very-high-resolution remote sensing images. The method matches the appearance of the source and target images prior to minimizing differences in their distributions within the feature space. Better transferability and robustness shown in experiments with five large-scale datasets. However, adaptation methods at the feature level exhibit data-dependency, highlighting long-standing challenges with reliable cross domain generalization.

Li et al. [22]proposed Spatial-Topological-Semantic Alignment (STSA) for Cross-Domain Remote Sensing Scene Classification With Limited Source Labels. It simultaneously matches spatial information, topological information and semantic information and utilizes unlabeled target images. Experiments performed with multimodal datasets and optical datasets yield better results even if just one image is used as a data source. The scope of bigger domains and more complex content for which the program is scalable and efficient in computation should be tested.

3. Proposed Adaptive Transformer Framework for Few-Shot Cross-Domain Satellite Image Classification

The proposed approach proposes an adaptive Transformer framework that combines scarcity guided augmentation and reliability-aware domain alignment for few shot satellite image classification. The DSGR data is a set of 362,582 samples with 41 land-use classes and six geographic regions, which is merged with the same land-use class from the RGB fMoW data. First, image integrity verification, image resizing, normalization and split validation are used to set up consistent inputs. Then, the data scarcity is estimated within the domain-specific class statistics and the resources for augmentation are allocated. The conventional transformations and quality filtered conditional GAN synthesis boosts minority-class representation. The pretrained Swin Transformer-Tiny first extracts hierarchical visual features, followed by the refinement of these features by a Scarce Conditioned Token Adaptation (SCTA) and a local–global feature fusion for adaptive refinement. Then, class-conditional DA maps the corresponding class representations from source and target domains by computing the reliability weights based on the availability of the target support and the consistency of features. The classification of 41 different categories is supported by class-balanced focal optimization. Performance is judged based on Macro-F1, controlled ablation studies, balanced accuracy and minority class recall, across various few-shot scenarios and geographic domains. The proposed framework integrates DSGR preprocessing, scarcity estimation, adaptive augmentation, Transformer feature learning, and reliability-aware domain alignment. Source and target-support features guide prototype alignment, while a class-balanced focal-loss classifier predicts land-use categories across geographic domains under limited labeled data is shown in Figure 1.

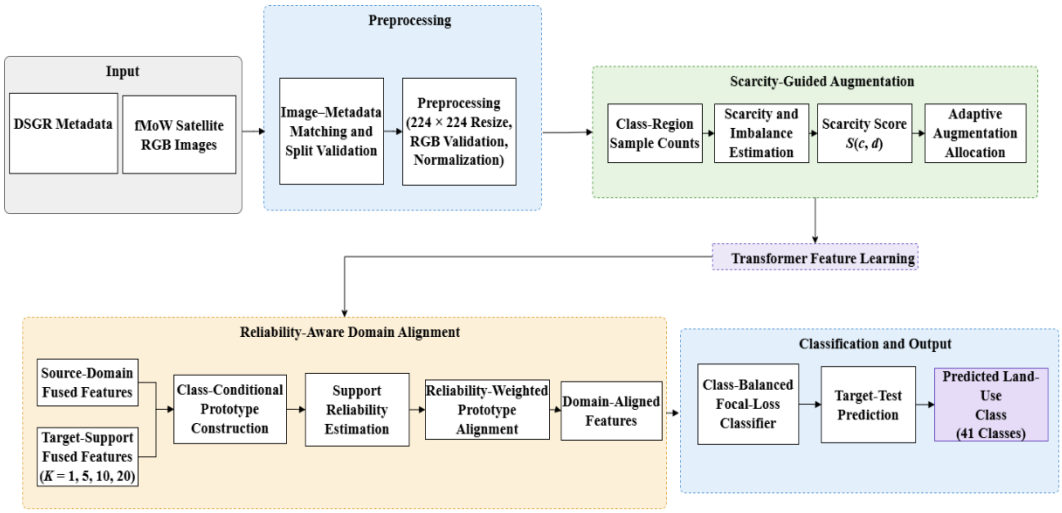


Figure 1. Proposed SCTA-RAST Framework for Few-Shot Satellite Image Classification

3.1 Data Collection

A dataset of DSGR was chosen for this study to study the problem of few-shot satellite image classification with the difficulty of class imbalance and geographic domain shift. It contains a total of 362,582 samples from 41 land use classes in 6 geographic regions: Asia, Africa, Oceania, Europe, Latin America and the Caribbean, and Northern America. The dataset includes metadata with paths to images, the class labels, geographic areas, country codes, image size, cloud cover, and official data-splits. DSGR is not storing the actual satellite images, but rather metadata, so the RGB imagery is acquired independently from the publicly available fMoW dataset[23]. The official splits are stored and a controlled few-shot protocol is set up by using a small number of labeled target-domain samples.

3.1 Data Preprocessing

Data preprocessing is an essential stage for preparing satellite images for reliable few-shot cross-domain classification. In this study, the DSGR metadata and corresponding fMoW RGB images are processed to ensure consistent image representation, accurate class labels, and valid geographic-domain assignments. Satellite images may have different spatial dimensions, whereas the selected pretrained Swin Transformer-Tiny backbone requires a consistent input size. Therefore, each valid RGB image is resized to 224×224 pixels before feature extraction. Let the original image be I_i, with height H_i and width W_i. The standardized image is expressed as in equation (1).

$$I_i^r = \text{Resize}(I_i, 224{,}224) \quad (1)$$

where I_i^r denotes the resized image and $\text{Resize}(\cdot)$ represents the selected interpolation operation. The same resizing procedure is applied across all dataset partitions. This standardization simplifies batch processing and ensures that the Transformer receives inputs with consistent dimensions.

3.2 Scarcity and Class-Imbalance Estimation

This stage is used to detect underrepresented types of land use in the DSGR data set and determine data scarcity of such categories prior to augmentation and training of the Transformer. All estimates are only based on allowed training information and some few-shot support information, which means that validation information and test information does not affect the proposed learning process. Let $n_c^{(d)}$ denote the number of available training samples belonging to class c in domain d. The maximum class count in that domain as in equation (2).

$$n_{\max}^{(d)} = \max_{c \in \mathcal{C}_d} n_c^{(d)} \quad (2)$$

where $n_{\max}^{(d)}$ is the largest class count, $\mathcal{C}_d$ is the set of classes represented in the permitted training data of domain d, c identifies a land-use category, and d denotes a geographic domain. The target-domain counts are calculated from the selected few-shot support set rather than from target validation or test labels. This produces a domain-specific representation of class availability.

3.2.1 Class-Imbalance Ratio Estimation

After calculating the class counts, the framework measures the imbalance between each class and the most represented class within the same domain. The imbalance ratio is defined as in equation (3).

$$IR_c^{(d)} = \frac{n_{\max}^{(d)}}{n_c^{(d)} + \epsilon} \quad (3)$$

where $IR_c^{(d)}$ is the imbalance ratio for class c in domain d, $n_{\max}^{(d)}$ is the maximum class count, $n_c^{(d)}$ is the count of class c, and ϵ is a small positive constant used for numerical stability. A ratio close to one indicates relatively balanced representation, whereas a larger ratio indicates stronger underrepresentation.

3.2.2 Scarcity Score Computation

The proposed framework normalizes classes whose counts are used to calculate severity scores to make them comparable across classes within a domain. The initial scarcity score is computed as in equation (4).

$$S_c^{(d)} = 1 - \frac{n_c^{(d)}}{n_{\max}^{(d)} + \epsilon} \quad (4)$$

where $S_c^{(d)}$ is the scarcity score for class c in domain d, $n_c^{(d)}$ is its available training count, and $n_{\max}^{(d)}$ is the largest class count in that domain. The score approaches zero for the most represented class and increases as the class count decreases.

3.2.3 Feature-Diversity-Aware Scarcity Estimation

Class counts alone cannot determine whether the available images contain sufficient visual diversity. The images used for this estimate must belong to the permitted training partition or target support set. Let h_i be the normalized feature vector of image i, and let $\bar{h}_c^{(d)}$ be the mean normalized feature vector for class c in domain d. A simple feature-dispersion estimate as in equation (5).

$$v_c^{(d)} = \frac{1}{n_c^{(d)}} \sum_{i:y_i=c} \left\| h_i - \bar{h}_c^{(d)} \right\|_2^2 \qquad (5)$$

where $v_c^{(d)}$ measures within-class feature dispersion, y_i is the land-use label of image i, and $\|\cdot\|_2$ denotes the Euclidean norm. Because dispersion is not automatically equivalent to useful diversity, $v_c^{(d)}$ is normalized across eligible classes to obtain $\tilde{v}_c^{(d)} \in [0,1]$. Higher normalized values indicate greater observed feature variation, not necessarily better coverage.

3.3 Scarcity-Guided Augmentation and Conditional GAN-Based Sample Generation

In this stage, the samples from the land-use categories identified in the scarcity estimation is normalized due to the scarcity of samples in some categories, and additional samples are created based on the needs of the class. The proposed framework does not use the same augmentation intensity for all the categories, instead it applies the scarcity scores calculated to distribute augmentation resources among geographic domains. These image transformations are conventional and offer a low cost way of increasing diversity, but for classes that are under-represented, candidate satellite images are generated using a conditional Generative Adversarial Network (GAN). Generated samples are assessed for visual quality, class consistency and diversity before added to the train set. The additional budget is called the augmentation budget which is stipulated in the following manner as in equation (6).

$$a_c^{(d)} = \mathrm{round}\left[a_{\min} + (a_{\max} - a_{\min})\bar{S}_c^{(d)}\right] \quad (6)$$

where $a_c^{(d)}$ denotes the number of candidate augmented samples per original image, $a_{\min}$ and $a_{\max}$ are the minimum and maximum augmentation budgets, $\bar{S}_c^{(d)}$ is the scarcity score for class c in domain d, and $\mathrm{round}(\cdot)$ converts the result into an integer. As an initial implementation, $a_{\min} = 1$ and $a_{\max} = 5$ may be tested, with the final values selected using validation performance and computational constraints. This allocation increases augmentation for underrepresented classes while limiting unnecessary sample generation for well-represented categories.

3.3.1 Conditional GAN Architecture and Training

For classes requiring additional diversity, the framework employs a conditional GAN that generates candidate satellite images based on a requested land-use category. The generator receives a random latent vector z and a class condition c, and produces a synthetic image as in equation (7).

$$x_{\mathrm{syn}} = G(z, e_c) \quad (7)$$

where x_{syn} is the generated image, $G(\cdot)$ denotes the generator, $z \sim \mathcal{N}(0, I)$ is a latent vector sampled from a standard normal distribution, and e_c is the embedding of the requested land-use class. The embedding represents one of the eligible DSGR categories, such as airport, factory or powerplant, crop field, or port. The generator is trained using real images from the permitted training data, and its discriminator learns to distinguish real training samples from generated candidates while considering the class condition.

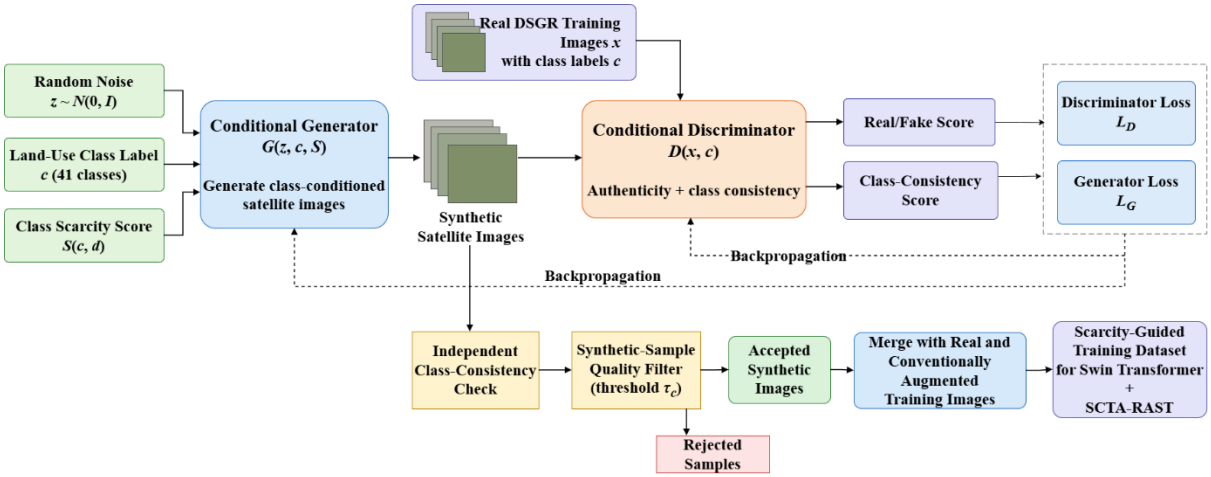


Figure 2. Architecture of the Proposed Scarcity-Guided Conditional Generative Adversarial Network (SG-cGAN)

Figure 2 shows the sgcGAN framework, where the generation of synthetic satellite images is conditioned on a random noise, land-use class labels and scarcity scores. A conditional discriminator checks authenticity and class consistency, and quality filtering removes samples that are rejected and adds the appropriate synthetic images to the SCTA-RAST training set.

3.3.2 Synthetic Image Quality and Class-Consistency Evaluation

GAN-generated images may contain distorted buildings, unrealistic surface patterns, or visual characteristics that do not match their requested class. Therefore, the proposed framework does not accept every generated image. Each candidate is evaluated using an image-quality and class-consistency procedure before entering the augmented training set. Let $p_\phi(c \mid x_{\text{syn}})$ denote the probability assigned to the requested class by an independently trained or appropriately validated classifier. A simple acceptance rule as in equation (8).

$$A(x_{\text{syn}}, c) = \begin{cases} 1, & p_\phi(c \mid x_{\text{syn}}) \geq \tau_c, \\ 0, & \text{otherwise}, \end{cases} \qquad (8)$$

where A is the acceptance indicator, $p_\phi(c \mid x_{\text{syn}})$ is the class-consistency probability, and τ_c is the class-specific acceptance threshold. A value of one means that the sample passes this particular filter, while zero means that it is rejected. The Fréchet Inception Distance (FID) can be used to compare the distributions of real and synthetic images.

3.3.3 Construction of the Augmented Training Dataset

After quality filtering, the framework combines the original training images, accepted conventional augmentations, and accepted synthetic samples. For each class c and domain d, the augmented training set is expressed as in equation (9).

$$\mathcal{D}_{\text{aug}}^{(d,c)} = \mathcal{D}_{\text{real}}^{(d,c)} \cup \mathcal{D}_{\text{trans}}^{(d,c)} \cup \mathcal{D}_{\text{syn}}^{(d,c)} \qquad (9)$$

where $\mathcal{D}_{\text{real}}^{(d,c)}$ contains original real training images, $\mathcal{D}_{\text{trans}}^{(d,c)}$ contains accepted transformed images, and $\mathcal{D}_{\text{syn}}^{(d,c)}$ contains accepted GAN-generated samples. The symbol $\cup$ denotes combining the sample sets; in implementation, duplicate files and repeated samples should be checked. The augmented dataset is then used to train the subsequent Transformer-based classifier.

3.4 Deep-Learning Patch Embedding and Transformer Feature Extraction

This stage converts the scarcity-aware satellite images generated into informative feature representations for subsequent reliability-aware domain alignment and classification. The proposed framework employs a pretrained Swin Transformer-Tiny backbone to extract visual features from the 41 land-use categories in DSGR.

The model divides every image into local regions, and learns hierarchical representations of structures like airport runways, industrial roofs, agricultural fields, and residential layouts, rather than directly from the raw pixel values of the image. The output of this component are the feature tokens that are used as input to the following adaptive learning components. This step is used to create the feature representation on which the proposed framework conducts domain adaptation and final classification. The dot product with the patch is represented as shown in equation (10).

$$Z^{(0)} = \text{PatchEmbed}(I) \quad (10)$$

where $I \in \mathbb{R}^{224\times224\times3}$ is the normalized input image, $\text{PatchEmbed}(\cdot)$ denotes the learned patch-projection operation, and $Z^{(0)} \in \mathbb{R}^{56\times56\times96}$ is the initial feature map. Each spatial position represents a learned description of a local image region rather than a raw pixel block. For example, an airport image may produce local representations associated with runway edges, terminals, and paved surfaces. These initial representations form the basis for subsequent Transformer feature learning.

3.4.1 Hierarchical Transformer Feature Learning

The feature map is then sent to the Swin Transformer-Tiny backbone, which works hierarchically. The attention operation within a window can be represented as in equation (11).

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^{\top}}{\sqrt{d_k}} + B\right)V \quad (11)$$

where Q, K, and V are the query, key, and value matrices projected from the input tokens; d_k is the query/key feature dimension; and B is the relative-position bias used by the Swin Transformer. The softmax function converts attention scores into normalized weights, determining how information is aggregated from other tokens within the window.

3.4.2 Extraction of Final Image Representations

The final hierarchical feature map is converted into a fixed-dimensional image representation for subsequent domain alignment and classification. Let the final feature map be $F_i \in \mathbb{R}^{H_f\times W_f\times C_f}$, where H_f and W_f denote its spatial dimensions and C_f denotes the number of feature channels. Spatial average pooling produces the image-level feature vector as in equation (12).

$$h_i = \frac{1}{H_f W_f}\sum_{u=1}^{H_f}\sum_{v=1}^{W_f} F_i\,(u, v, :) \quad (12)$$

where $h_i \in \mathbb{R}^{C_f}$ is the pooled representation of image i, and $F_i(u, v, :)$ denotes the feature vector at spatial position (u, v). For the standard Swin Transformer-Tiny configuration, the final stage has 768 channels, giving $h_i \in \mathbb{R}^{768}$.

3.4.3 Feature Preparation for Scarcity-Aware Adaptation

The retrieved image representations are linked to their respective names of land uses and geographic domains. Source features are used to learn general class characteristics while features from the small number of labelled target support set are used to estimate target class representations. Most importantly, the scores of scarcity are not derived from test features, but from data allowed for training and information from the support set. Prior to computing the prototypes, the resulting representations can be normalized to minimize differences in feature size. In image i, the feature normalization can be expressed as in equation (13).

$$\hat{h}_i = \frac{h_i}{\|h_i\|_2 + \epsilon} \quad (13)$$

where h_i is the pooled feature vector, $\hat{h}_i$ is the normalized vector, $\|\cdot\|_2$ is the Euclidean norm, and ϵ is a small positive constant for numerical stability. The normalized representations support consistent distance calculations in later alignment operations.

3.5 Scarcity-Conditioned Token Adaptation (SCTA)

This stage is used to adapt the representations of the Transformer using the number and variety of examples for each land-use type. The Swin Transformer-Tiny backbone extracts the visual features and image-level representation in a hierarchical fashion from each satellite image in the previous step. But the learning focus of learning process is not adjusted by the standard feature extraction explicitly based on class scarcity. Let the token representation of image ias in equation (14).

$$X_i = [x_{i,1}, x_{i,2}, \dots, x_{i,N}] \quad (14)$$

where $x_{i,j} \in \mathbb{R}^D$ represents the feature vector of token j, N is the number of tokens at the selected feature stage, and D is the feature dimension. The class-specific scarcity score is denoted by $\bar{S}_c^{(d)}$, where c represents the land-use category and d represents the geographic domain. Class scarcity alone does not indicate which regions of a satellite image are useful for classification. SCTA therefore estimates the relevance of each token from its learned visual representation. A token scoring function produces a relevance value for each spatial position as in equation (15).

$$r_{i,j} = w_r^{\top} \tanh\left(W_r x_{i,j} + b_r\right) \quad (15)$$

where $r_{i,j}$ is the relevance score for token j in image i; W_r is a learnable projection matrix; b_r is its bias; w_r is a learnable vector that maps the projected representation to a scalar; and $\tanh(\cdot)$ is a nonlinear activation function. These parameters are trained using the classification objective and the SCTA adaptation process.

3.6 Local–Global Feature Learning for Geographic-Robust Satellite Image Representation

This stage builds a shared representation of local visual structures and global context of the scene with the features yielded by the SCTA under the scarcity condition. In addition to adjusting the representations of tokens based on their scarcity in the respective classes and their relevance within a token, a model has to be able to separate fine-grained patterns that are local in nature from the overall spatial organization of a scene for reliable satellite image classification. The input is the obtained token sequence by using SCTA as in equation (16).

$$\tilde{X}_i = [\tilde{x}_{i,1}, \tilde{x}_{i,2}, \dots, \tilde{x}_{i,N}] \quad (16)$$

where $\tilde{X}_i$ denotes the SCTA-adapted features of image i, $\tilde{x}_{i,j} \in \mathbb{R}^D$ is the feature vector of token j, N is the number of tokens at the selected feature stage, and D is the feature dimension. For implementation, the tokens are arranged into their original spatial grid, denoted as in equation (17).

$$F_i^{\mathrm{SCTA}} \in \mathbb{R}^{H_f \times W_f \times D} \quad (17)$$

where H_f and W_f represent the feature-map height and width, respectively. For example, using the final 7×7 feature map of the standard Swin Transformer-Tiny backbone gives $N = 49$ spatial tokens and $D = 768$. The precise feature-map dimensions must match the selected backbone output and the SCTA implementation.

3.6.1 Local Feature Learning

The local pathway emphasizes structures that are localized in space and can differentiate among the same category of land use type that have similar visual appearance. The types of these structures that can be seen in satellite images range from runway edges and road intersections, to building boundaries, agricultural field patterns and outlines of storage tanks. Such patterns can be present in just a fraction of the image, so it is necessary to retain some local information, besides the single image level feature vector. Local representation is calculated with the aid of a convolutional projection of neighbouring feature positions as in equation (18).

$$F_i^L = \phi\left(\mathrm{Conv}_{3\times3}\left(F_i^{\mathrm{SCTA}}\right)\right) \quad (18)$$

where F_i^L is the local feature map, $\mathrm{Conv}_{3\times3}$ is a convolution with a 3×3 spatial kernel, and $\phi(\cdot)$ denotes a nonlinear activation function such as GELU. The convolution learns combinations of nearby token features,

allowing the representation to capture local edges, shapes, textures, and arrangements. Padding can be set to one spatial position to preserve the feature-map dimensions.

3.6.2 Global Scene Representation Learning

The global pathway captures the information at the scene level which is not necessarily present in local spaces. Two different categories of land use can present much the same building texture yet be organized, grouped and/or related differently to other buildings. A Global context vector is calculated via spatial average pooling as in equation (19).

$$g_i = \frac{1}{H_f W_f} \sum_{u=1}^{H_f} \sum_{v=1}^{W_f} F_i^{\mathrm{SCTA}}(u, v, :) \qquad (19)$$

where $g_i \in \mathbb{R}^D$ is the global representation of image i, $F_i^{\mathrm{SCTA}}(u, v, :)$ is the feature vector at spatial position (u, v), and $H_f W_f$ is the number of spatial positions. Average pooling provides a compact summary of the scene. To allow the global representation to adapt to different image contents, a learnable transformation is applied:

3.6.3 Adaptive Local–Global Feature Fusion

There is complementary information in the local and global pathways. But the category and image-specific relative importance might differ. A local feature map, that is, a map created by pooling features in the local neighborhood, is created first as in equation (20).

$$l_i = \frac{1}{H_f W_f} \sum_{u=1}^{H_f} \sum_{v=1}^{W_f} F_i^{L}(u, v, :) \qquad (20)$$

where $l_i \in \mathbb{R}^D$ is the pooled local representation. Next, a local–global gate is calculated.

3.7 Class-Conditional Domain Adaptation with Reliability-Aware Feature Alignment

This step matches the acquired representations of the source and target geographical domains while maintaining the differences among 41 categories of land use. The previous steps yield locally and globally informed image representations and scarcity-aware. The inputs to this stage are the normalized local–global representations from the previous stage as in equation (21).

$$\hat{h}_i^{LG} \in \mathbb{R}^D \qquad (21)$$

where $\hat{h}_i^{LG}$ is the normalized feature vector of image i, and $D = 768$ for the standard final-stage representation of Swin Transformer-Tiny. The labeled source training set is denoted as in equation (22).

$$\mathcal{D}_s = \{(x_i^s, y_i^s, d_i^s)\}_{i=1}^{N_s} \qquad (22)$$

and the labeled target support set as in equation (23).

$$\mathcal{S}_t = \{(x_j^t, y_j^t)\}_{j=1}^{N_t} \qquad (23)$$

where x_i^s and x_j^t are source and target images; y_i^s and y_j^t are their land-use labels; d_i^s is the source geographic domain; and N_s and N_t are the corresponding sample counts. The target support set contains only the permitted K-shot examples per eligible class.

3.7.1 Construction of Class-Conditional Prototypes

The class prototype is a summary of the learned feature representation of a land-use category in a specific domain. The proposed method constructs source prototypes from the source training representations and target prototypes from the limited labeled support representations, instead of comparing individual images. The source prototype is given as in equation (24).

$$\mu_{s,c}^{(d)} = \frac{1}{n_{s,c}^{(d)}} \sum_{i \in \mathcal{I}_{s,c}^{(d)}} \hat{h}_i^{LG} \qquad (24)$$

where $\mu_{s,c}^{(d)} \in \mathbb{R}^D$ is the source prototype; $\mathcal{J}_{s,c}^{(d)}$ is the set of source training samples belonging to class c in domain d; and $n_{s,c}^{(d)}$ is the number of those samples. The target prototype is calculated from the labeled target support set as in equation (25).

$$\mu_{t,c} = \frac{1}{n_{t,c}} \sum_{j \in \mathcal{J}_{t,c}} \hat{h}_j^{LG} \qquad (25)$$

where $\mu_{t,c} \in \mathbb{R}^D$ is the target prototype, $\mathcal{J}_{t,c}$ contains the permitted target support samples of class c, and $n_{t,c}$ is the number of such samples. In a balanced K-shot experiment, $n_{t,c} = K$ for each eligible class.

3.7.2 Estimation of Target-Prototype Reliability

The target prototypes in few-shot learning can be unreliable due to insufficient available labeled images. Five similar images, on which a prototype is computed, may not be sufficiently representative of the images in the target region. Thus, the proposed approach estimates the reliability of the target prototype based on the quantity of available support prototypes as well as the consistency of features of these support prototypes. First, the dispersion of the target support features for class c are calculated as in equation (26).

$$v_{t,c} = \frac{1}{n_{t,c}} \sum_{j \in \mathcal{J}_{t,c}} \left\| \hat{h}_j^{LG} - \mu_{t,c} \right\|_2^2 \qquad (26)$$

where $v_{t,c}$ denotes within-class feature dispersion, $\hat{h}_j^{LG}$ is a normalized target support feature, $\mu_{t,c}$ is its class prototype, and $n_{t,c}$ is the number of target support examples. A support-count reliability term is defined as in equation (27).

$$r_{n,c} = \frac{n_{t,c}}{n_{t,c} + \kappa} \qquad (27)$$

where $r_{n,c} \in [0,1)$ represents reliability associated with sample availability and $\kappa > 0$ controls how quickly the score increases as more support samples become available. A starting value of $\kappa = 5$ can be investigated. With five support examples, for instance, the count-based reliability is 0.5; with ten, it is approximately 0.67. To incorporate feature consistency, dispersion is mapped to a bounded reliability score.After computing source and target prototypes, the framework measures their discrepancy for each eligible class. Since the features are normalized, cosine distance provides a convenient measure of prototype mismatch as in equation (28).

$$\Delta_c = 1 - \frac{(\mu_{s,c})^\top \mu_{t,c}}{\|\mu_{s,c}\|_2 \|\mu_{t,c}\|_2 + \epsilon} \qquad (28)$$

where Δ_c denotes the source–target discrepancy for class c, $\mu_{s,c}$ is the source prototype, $\mu_{t,c}$ is the target prototype, and ϵ prevents division by zero. When multiple source domains are used, $\mu_{s,c}$ can be obtained by averaging their class prototypes or by a source-sample-count-weighted average.

3.8 Imbalance-Aware Classification with Class-Balanced Focal Optimization

At this stage, the class-conditional domain adaptation is used to generate the geographically adapted feature representations and the imbalanced class distribution is minimized; the geographically adapted feature representations were then used to make final predictions of land use. While Scarcity-guided augmentation is performed on the preceding stages to enhance feature quality, Scarcity-Conditioned Token Adaptation (SCTA), local–global feature learning and reliability-aware domain alignment are still performed on the final classifier, which can be detrimental to the quality of the features it receives.Scarcity-guided augmentation is also performed on the preceding stages to enhance feature quality, but the final classifier can be affected by the scarcity-conditioned feature learning, local–global feature learning and reliability-aware domain alignment, potentially resulting in an inferior quality of features. The input to this stage is the adapted image representation generated by the reliability-aware domain adaptation pipeline as in equation (29).

$$h_i^{DA} \in \mathbb{R}^D \qquad (29)$$

where h_i^{DA} denotes the domain-adapted representation of image i, and $D = 768$ is the feature dimension when the standard Swin Transformer-Tiny output dimension is retained by the local–global fusion module. The representation is passed to a fully connected classification head that generates one score for each of the $C = 41$ land-use categories as in equation (30).

$$z_i = W_{\text{cls}} h_i^{DA} + b_{\text{cls}} \qquad (30)$$

where $z_i \in \mathbb{R}^{41}$ is the vector of classification logits; $W_{\text{cls}} \in \mathbb{R}^{41\times 768}$ is the learnable classifier weight matrix; and $b_{\text{cls}} \in \mathbb{R}^{41}$ is the learnable bias vector. The logits are converted into class probabilities using softmax as in equation (31).

$$p_i(c) = \frac{\exp(z_{i,c})}{\sum_{k=1}^{C} \exp(z_{i,k})} \qquad (31)$$

where $p_i(c)$ is the predicted probability that image i belongs to category c, $z_{i,c}$ is the corresponding logit, and $C = 41$ is the number of classes.The class count is calculated as in equation (32).

$$n_c = \sum_{i=1}^{N_{\text{train}}} \mathbb{I}\,(y_i = c) \qquad (32)$$

where n_c is the number of labeled training examples belonging to class c; N_{train} is the number of permitted labeled training examples included in the relevant training pool; y_i is the ground-truth class label of sample i; and $\mathbb{I}(\cdot)$ is an indicator function that equals one when the condition is true and zero otherwise.Simple inverse-frequency weighting can assign excessively large weights to extremely rare categories. To control this behavior, the proposed classifier uses the effective number of samples to estimate class importance as in equation (33).

$$E_c = \frac{1-\beta^{n_c}}{1-\beta} \qquad (33)$$

where E_c is the effective number of examples for class c, n_c is its training count, and $\beta \in [0,1)$ is a smoothing parameter that controls how strongly the weighting accounts for diminishing information from additional examples. The raw class weight is then calculated as in equation (34).

$$\widetilde{w}_c = \frac{1-\beta}{1-\beta^{n_c}} \qquad (34)$$

where $\widetilde{w}_c$ is the unnormalized weight assigned to class c. As the number of examples decreases, the effective-number formulation generally increases the class weight, giving scarce categories a stronger contribution to the loss. The weights are normalized to maintain a manageable average scale.The number of classes is determined by the number of students in each class as in equation (35).

$$\mathscr{L}_{\text{CBF}} = -\frac{1}{B}\sum_{i=1}^{B} w_{y_i}\left(1 - p_i(y_i)\right)^{\gamma_f} \log(p_i(y_i) + \epsilon) \quad (35)$$

Where B is the number of labeled samples in a training batch, w_{y_i} is the normalized class weight for the ground-truth category of sample i, $p_i(y_i)$ is the predicted probability assigned to the correct class, $\gamma_f \geq 0$ is the focal-loss focusing parameter, $\log(\cdot)$ is the natural logarithm, ϵ is a small constant used for numerical stability. The total objective is therefore expressed as in equation (36).

$$\mathscr{L}_{\text{final}} = \mathscr{L}_{\text{CBF}} + \lambda_{\text{align}} \mathscr{L}_{\text{align}} \qquad (36)$$

where $\mathscr{L}_{\text{CBF}}$ is the class-balanced focal classification loss; $\mathscr{L}_{\text{align}}$ is the reliability-weighted class-conditional alignment loss; and λ_{align} controls the contribution of domain adaptation. After training, the classifier generates a probability distribution across the 41 land-use categories for each test image. The predicted class is obtained from the highest probability.

Algorithm 1. Proposed Adaptive Transformer Framework for Few-Shot Satellite Image Classification

Input:

DSGR metadata and corresponding fMoW satellite images

Source-domain training set Ds

Target-domain support set St with K ∈ {1, 5, 10, 20}

Validation set Dv and test set Dt

Pretrained Swin Transformer-Tiny

Hyperparameters and maximum epochs Emax

Output:

Trained classification model M*

Target-domain predictions ŷt

Evaluation metrics

Load metadata and match images using img_path.

Validate images, labels, geographic regions, and official splits.

Resize images to 224 × 224 and normalize pixel values.

Estimate class scarcity scores Sc using training data only.

Allocate class-wise augmentation according to scarcity scores.

Generate transformed samples and conditional GAN synthetic images.

Filter synthetic images using class-consistency and quality thresholds.

Initialize the pretrained Swin Transformer-Tiny backbone.

Extract hierarchical image tokens and feature representations.

Apply SCTA using token relevance and scarcity-conditioned gates.

Extract local features and global contextual representations.

Fuse local and global features using adaptive sigmoid gating.

Construct source and target class prototypes from permitted data.

Estimate target-support reliability using sample counts and dispersion.

Compute reliability-weighted class-conditional alignment loss.

Compute class-balanced focal classification loss.

Combine losses:

$L_{total} = L_{CBF} + \lambda_{align} L_{align}$

Update model parameters using backpropagation and AdamW.

Repeat Steps 9–18 until Emax or validation-based early stopping.

Select the best validation checkpoint as M*.

Predict target-test labels without using test labels during training.

Evaluate accuracy, precision, recall, Macro-F1, and balanced accuracy.

Report results across K-shot settings and random seeds.

The entire process of the proposed adaptive Transformer framework for few-shot satellite image classification is shown in Algorithm 1. The optimization of model parameters is done by AdamW in a iterative manner. The best validation checkpoint predicts unseen target-test samples and performance is compared throughout a variety of few-shot settings and random seeds.

In contrast to previous approaches that mostly focused on few-shot learning, multi-scale feature extraction, attention-based fusion, or domain alignment respectively, the proposed framework combines these features in a single Transformer-based framework. To tackle limited class representation, the works of scarcity-guided augmentation and conditional GAN generation are employed and SCTA adaptively adjusts token features based on scarcity levels. Local–global feature fusion extracts complementary spatial and contextual information and reliability-aware prototype alignment alleviates discrepancies in geographic domains. The integrated design aims to further improve the recognition of minority class, the target domain adaptation and the robustness of target classification under limited labeled satellite imagery.

4. Results and Discussion

The experimental evaluation highlights the usefulness of the proposed adaptive Transformer framework for the low-shots satellite image classification task and the DSGR dataset as well as fMoW imagery. It discusses the distinctions between geographical domains, class imbalance, scarcity guided augmentation, synthetic data quality and learned feature representations. The contributions are evaluated separately through targeted experiments, and include the concepts of scarcity-conditioned token adaptation, local–global feature fusion and reliability-aware domain alignment. Additional assessments examine performance with different target-support sizes, geographic transfer conditions and controlled imbalance conditions. The efficiency of computation, the level of prediction and the level of statistical consistency are also discussed. Ablation experiments are performed and compared to existing classification and domain-adaptation approaches to assess the contribution of the proposed framework given with respect to the individual components.The following simulation parameters are chosen and used in this study: shown in Table 1.

Table 1. Simulation Parameters for the Proposed Framework

Parameter	**Value**
Satellite-image dataset	DSGR with corresponding fMoW imagery
Total dataset samples	362,582
Number of classes	41
Geographic regions	6
Input image size	$224 \times 224 \times 3$
Data partition	Official train, validation, and test splits
Image normalization	Pretrained Swin normalization
Transformer architecture	Swin Transformer-Tiny
Patch size	4×4

Feature dimension	768
Optimizer	AdamW
Batch size	32
Maximum epochs	100
Early-stopping patience	10 epochs
Random seeds	5
Target support sizes (K)	1, 5, 10, 20 samples per eligible class
Count-weight coefficient (α)	0.7
Minimum augmentation allocation	1
Maximum augmentation allocation	5
Generator latent dimension	128
Optimizer	Adam
Learning rate	2×10^{-4}
Synthetic-sample quality threshold	τ_c, selected using validation data
Token adaptation strength (λ_s)	0.1
Scarcity-gate steepness (γ_s)	5
Scarcity-gate threshold (δ_s)	0.5
Local feature extractor	3×3 convolution
Fusion mechanism	Adaptive sigmoid gating
Reliability smoothing parameter (κ)	5
Reliability mixing coefficient (η)	0.5
Alignment loss weight (λ_{align})	0.1
Loss function	Class-balanced focal loss
Focal focusing parameter (γ_f)	2
Effective-number coefficient (β)	0.999
Epsilon (ϵ)	10^{-8}
Primary metric	Macro-F1
Additional metrics	Accuracy, precision, recall, balanced accuracy, minority-class recall

4.2 Dataset Characteristics and Geographic Domain Shift

The distribution of DSGR samples by six geographic regions and 41 land-use categories is analyzed in this section. The regional sample counts are explored and differences in class frequencies and features are evaluated as evidence of domain imbalance and geographic contrasts. Embedding visualizations depict if there are different feature patterns within the images from various regions. The analyses provide a benchmark for the difficulty of the cross-region classification problem and highlight difficult cross-region combinations. The results inspire the proposed mechanisms to enhance geographic generalization: augmenting based on scarcity and aligning based on reliability.

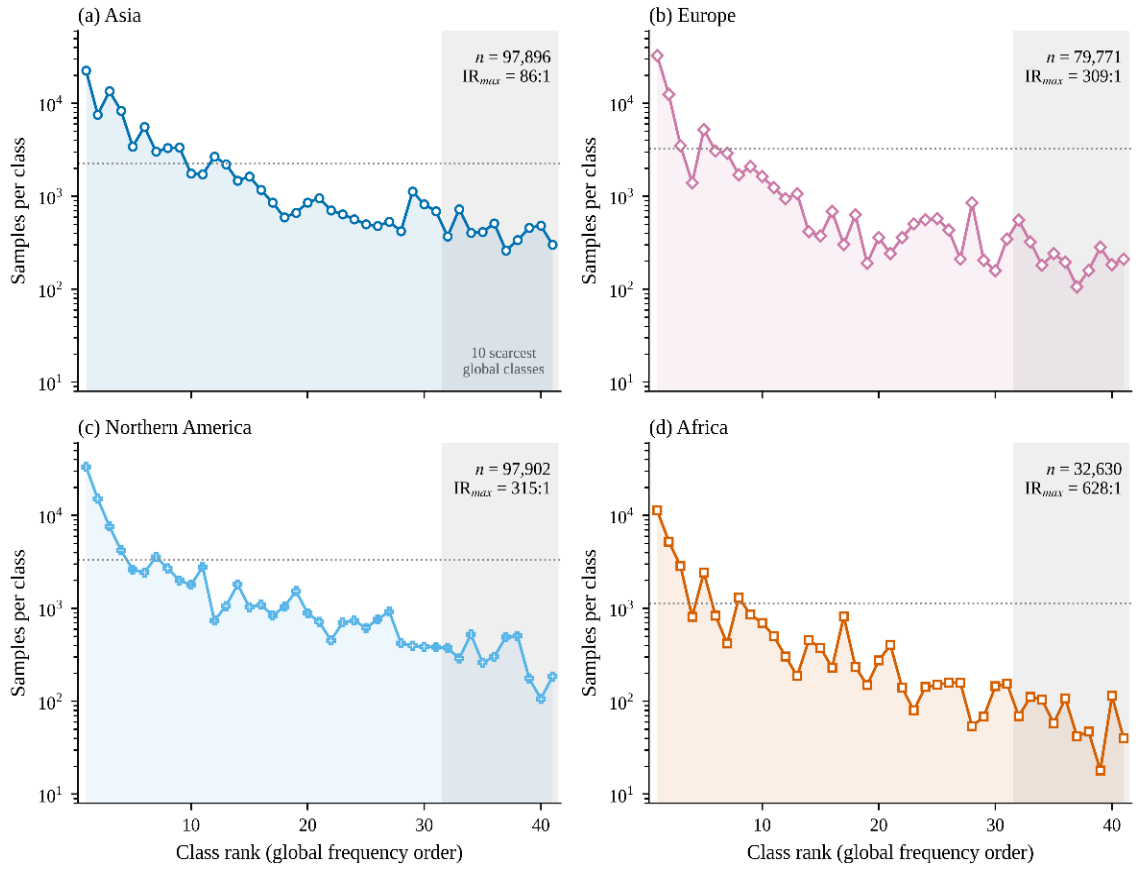


Figure 3. Regional Class-Frequency Distributions

The four panels show class-frequency distributions in typical geographic areas is shown in Figure 3. The line plots expose the variation in samples, primary classes and underrepresented classes, which drives the need to augment data for under-represented classes and to learn a balanced class model to enable better few-shot satellite image classification for regional imbalance.

4.3 Scarcity-Guided Augmentation Evaluation

The goal of this section is to assess if scarcity guided augmentation is able to effectively provide more examples of underrepresented classes. Geographic analyses of class-specific scarcity scores, augmentation allocations and synthetic-sample acceptance rates are conducted. Changes in the representation imbalance are identified by comparing the original frequencies with the augmented frequencies. Feature-distribution visualizations evaluate the usefulness of conventional transformations and conditional GAN generation for adding useful diversity. The analysis also explores whether augmentation improves less well-represented classes while not overloading the representation of well-represented classes, or even lowering the representation of well-represented classes.

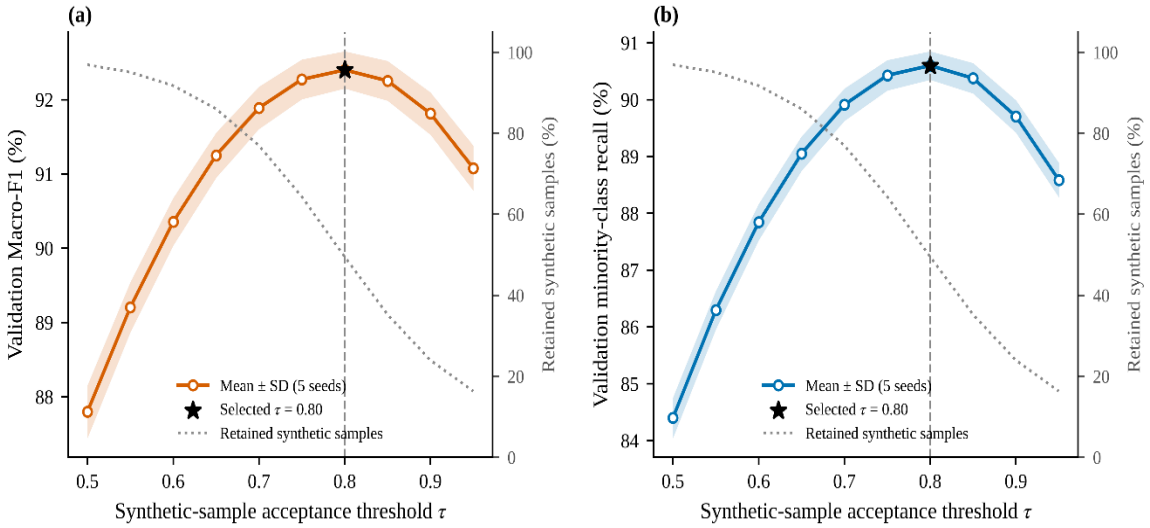


Figure 4. Validation Performance Across Synthetic-Sample Acceptance Thresholds

The line plots illustrate the validation Macro-F1 and minority-class recall at various acceptance thresholds in synthetic samples is shown in Figure 4. The trends show a compromise between a diverse set of generated images and the number of low quality images, which can be used to choose a threshold based on a validation set without target-test labels.

4.4 Noise-Aware Processing and Synthetic-Data Quality

This section explores the validation, preprocessing and filtering of synthetic samples for effectiveness. Representative real, transformed, accepted and rejected images are investigated for any inconsistencies of the images and errors in class labels. The class-consistency confidence and feature-distance distributions and diversity of the synthesized images are examined for various acceptance thresholds. The results will help to decide if filtering will decrease the number of unreliable generated samples while keeping the number of samples generated high enough. This assessment is to determine if the proposed data-processing technique is reliable for training a few-shot satellite image classification.

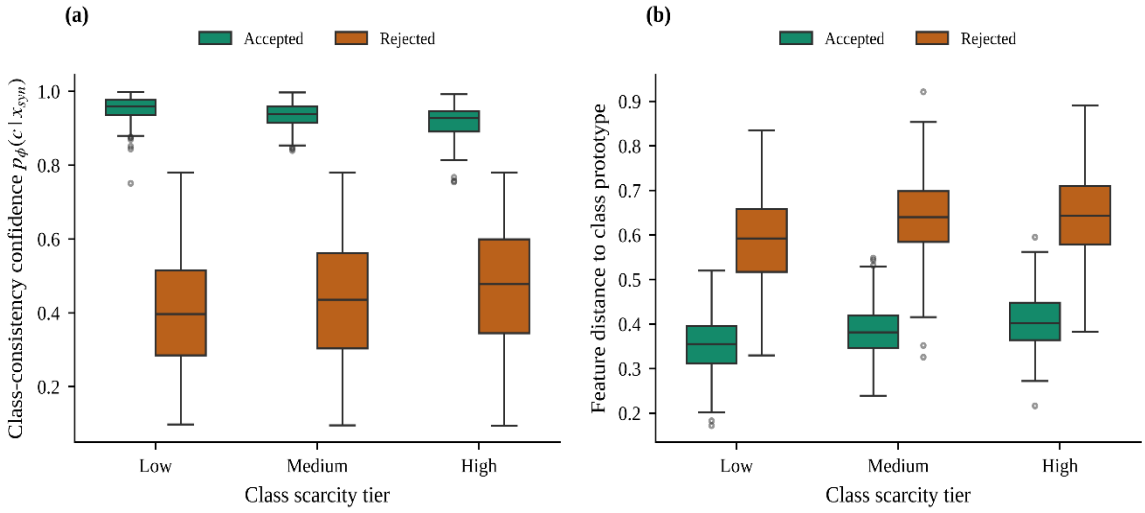


Figure 5. Quality Characteristics of Accepted and Rejected Synthetic Samples

In Figure 5 shows two set of box plots represent the class-consistency confidence distribution and the feature-distance distribution for accepted and rejected synthetic images, respectively. Differences between groups: Tests for the effectiveness of quality filtering for retaining class consistent samples and the exclusion of potentially misleading synthetic examples in training.

4.5 SCTA and Feature-Representation Analysis

In this section the performance of the token relevance, separation of the feature space, and classification performance in the class-wise mode is evaluated for Scarcity-Conditioned Token Adaptation (SCTA). Before and after scarcity-conditioned adaptations are compared in the form of visualizations and there is an emphasis on the changes of spatial features under scrutiny. Further analysis is conducted for each class to see if there is a difference between low and high scarcity classes with respect to performance and if SCTA does improve underrepresented categories. Comparisons are made with variants that do not incorporate relevance or scarcity to help separate out the impact of incorporating relevance and scarcity. The results validate the advantages of scarcity-aware token adaptation in terms of creating more discriminative representation when the number of labelled data is limited.

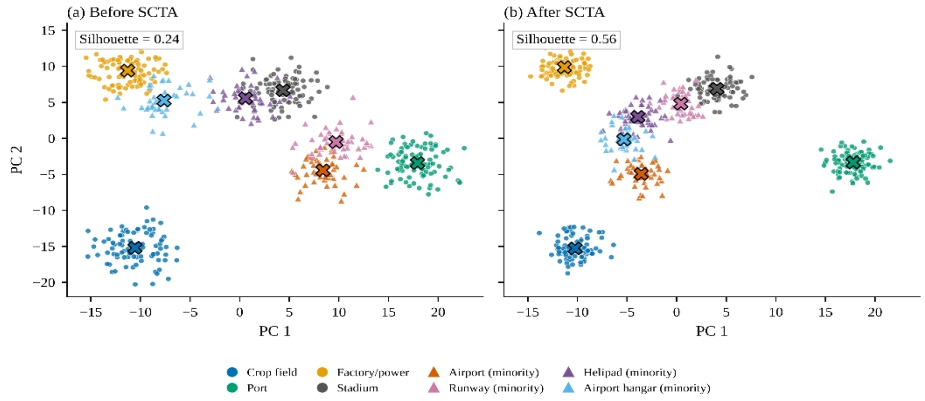


Figure 6. Feature-Space Class Separation Before and After SCTA

The scatter plots depict feature space representations both pre and post SCTA adaptation is shown in Figure 6. The cluster compactness and class separation suggest whether scarcity-conditioned token refinement enhances the discrimination of the representation of classes, especially those under-represented across the geographically shifted satellite-image domains.

4.6 Local–Global Feature Fusion Analysis

Feature maps from local pathway, global pathway and fused representation are compared, looking at the information captured by each of the pathways. The discrimination of features is evaluated by an analysis of the compactness of the features within the class and the separation between classes. Comparisons in class-wise performance indicate that adaptive fusion helps to enhance the recognition of similar visual categories of land-use classes. Results shed light on whether there are benefits to using both fine-grained structures and scene level context, versus using separately.

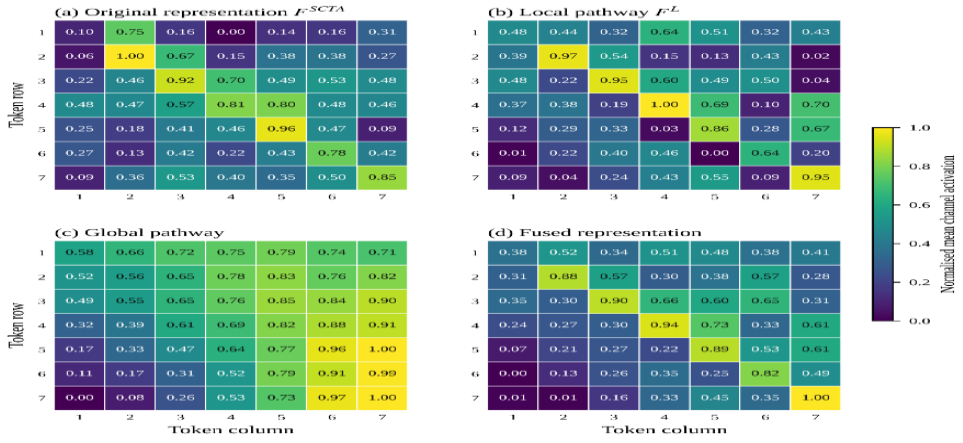


Figure 7. Local–Global Feature Representation Analysis

The four maps show original representation, local pathway, global pathway and adaptively fused representation. Figure 7 show that the spatial information and the contextual information provide complementary features for the final representation of the features for classification.

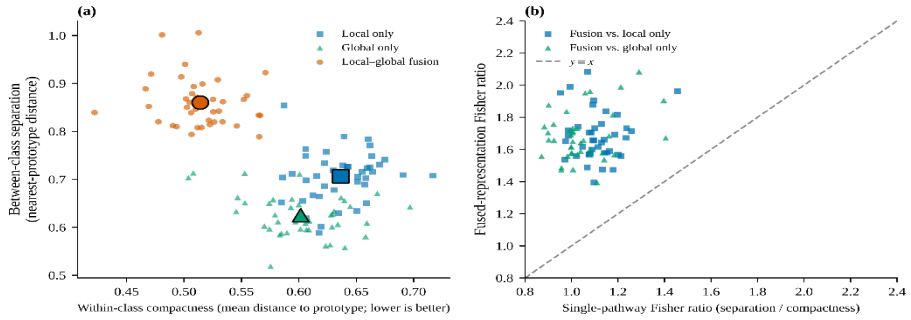


Figure 8. Within-Class Compactness and Between-Class Separation

The scatter plots depict within-class compactness and between-class separation measures of local-only, global-only, and fused representations is shown in Figure 8. The analysis is used to assess the ability of adaptive fusion to maintain the fine-grained spatial information and provide the global context to enhance discrimination between land-use categories.

4.7 Reliability-Aware Domain Alignment

The source – target prototype distances are compared before and after the adaptation, the target-support dispersion and reliability weights are compared within classes. Embedded visualizations show changes of cross domain feature organization. Even if uncertain target prototypes are down-weighted, the comparisons with uniform and alternative reliability-weighting strategies determine this. The analysis additionally determines if the decrease in feature discrepancy is correlated with the better discrimination of the target domain, specifically classes with geographic variation and few support samples.

Table 2. Target-Support Reliability Statistics

K-shot setting	Support count per class	Mean feature dispersion	Reliability score
K = 1	1	0.000	0.167
K = 5	5	0.184	0.541
K = 10	10	0.152	0.683
K = 20	20	0.126	0.792

The relationship between target-support size, feature dispersion and reliability of alignment is shown in Table 2. The bigger the number of the support set, the more stable the class prototypes will be. Target-support samples should be used for calculations of reliability only and not validation or test data.

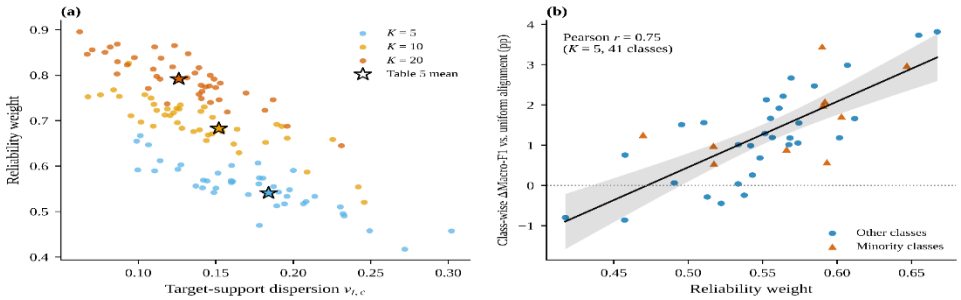


Figure 9. Target-Support Reliability and Classification Performance

The scatter plots were used to look at relationships between target-support feature dispersion and reliability weights as well as reliability weight and class-wise performance changes is shown in Figure 9. These connections show whether support reliability is related to prototype stability, and prioritize reliable domain alignment.

4.8 Few-Shot Performance, Robustness, and Stability

The performance of the classifiers under various sizes of the targets is assessed in this section, such as (K=1,5,10) and (20) examples per eligible class. Macro-F1 is compared with the balanced accuracy and minority classes recall with various geographic transfer settings. Repeated experiments with different random seeds give results that assess the variability and repeatability of the performance. In more difficult scenarios robustness is tested in further experiments with controlled imbalance and appearance-shift. The analysis reveals the effectiveness of the framework with the limited target labels, the consistency of the improvement across regions, and transfer configurations which still cause significant classification problems.

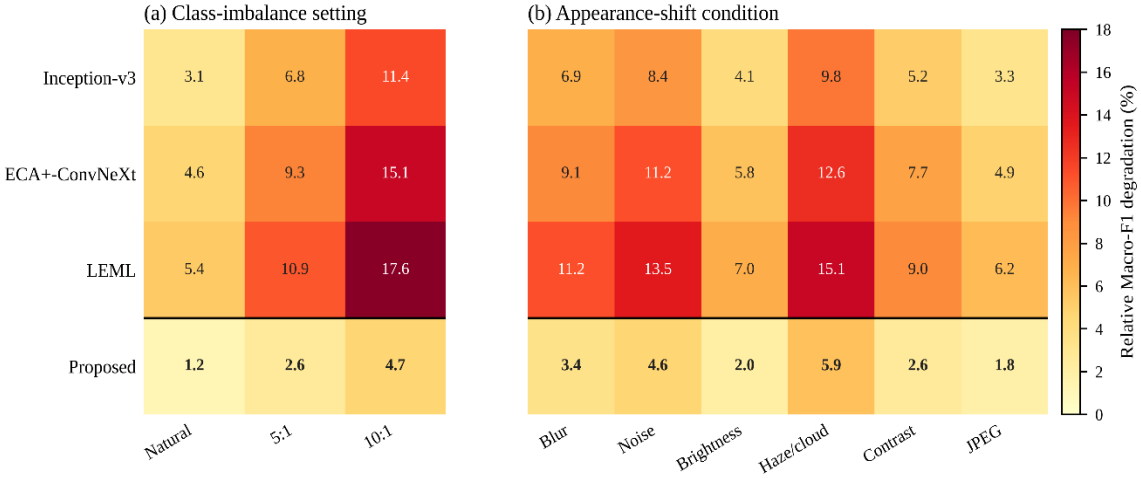


Figure 10. Robustness Under Class Imbalance and Image Corruption

The heatmaps represent the relative degradation in performance when the class imbalance is either 1:1 (Natural) or 5:1 or 10:1, and under specified conditions for corrupting images with random noise (Image-corruption) or shifting their appearance (Appearance-shift) is shown in Figure 10. The analysis investigates the robustness of classification when the class distribution and/or the image quality conditions change in a geographic domain.

4.9 Computational Efficiency and Error Analysis

The measurements are conducted within specific hardware configuration and batch-size under documented settings, such as number of parameters, computational operations, training time, inference latency and peak memory consumption. The commonly misclassified land use categories are identified by confusion matrices and the relationship between classification performance and availability and domain discrepancy of samples is identified by the class-wise scatter plots. These analyses demonstrate the potential for combining augmentation, token adaptation and domain alignment, and that this doesn't have a high cost. The outcome can be utilised to determine if enhancements are worth doing due to the additional processing power required, and how far improvements can be made.

Table 3. Computational Cost

Model	**Parameters (M)**	**FLOPs (G)**	**Training time (h)**	**Inference time (ms/image)**	**Peak GPU memory (GB)**	**Total augmentation cost (h)**
Inception-v3	23.9	5.7	4.2	8.4	4.1	0.0

LEML	18.6	4.9	3.8	7.6	3.7	0.0
ECA+-ConvNeXt	15.2	2.8	3.5	6.2	3.3	0.0
Proposed framework	30.8	8.1	9.6	11.7	7.2	3.4

The complexity of the models, the amount of training data require, the latency of inference, their memory overhead, and the overhead for augmenting the data are compared in Table 3. Generating synthetic data and aligning to the domain might require extra training costs for the proposed framework. Report timings are reported on same hardware, same batch size and same inference settings.

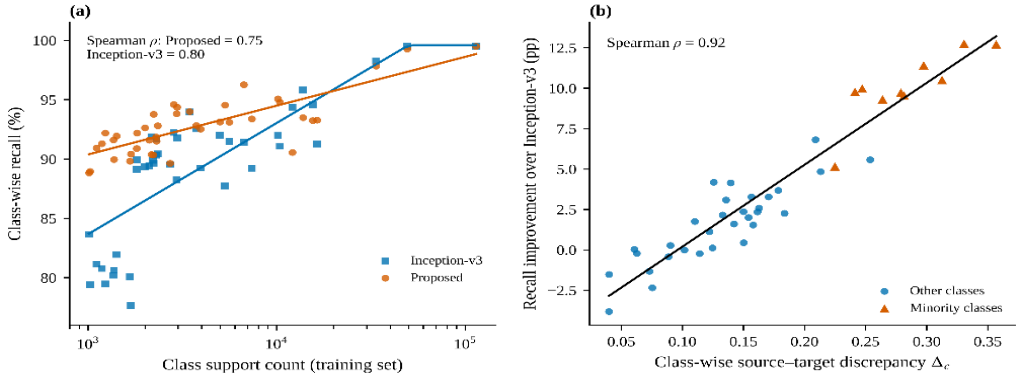


Figure 11. Support Availability, Domain Discrepancy, and Recall Improvement

Association between number of children in each class and recall are explored in scatter plots and association between domain discrepancy and improvement in recall is displayed in Figure 11. Such relationships have been employed to find out whether there are correlations between increasing the size of the labeled set and benefits of alignment and improved recognition of minority classes.

4.10 Statistical Significance and Prediction Confidence

This section reviews the consistency of the differences observed and the statistical reliability. To summarize the results across repeated random seeds, means, standard deviations and confidence intervals are computed. Statistically supported improvements in either of the above cases are determined by the presence of appropriate paired statistical tests and effect sizes. Reliability diagrams, expected calibration error and Brier score (when appropriate) are used to assess prediction confidence. The study of the domain-wise comparison is conducted to check the reliability of the confidence estimates in geographical change. The analysis is able to differentiate between consistent improvements and random variation, and to provide examples of instances where a high-confidence prediction is still incorrect.

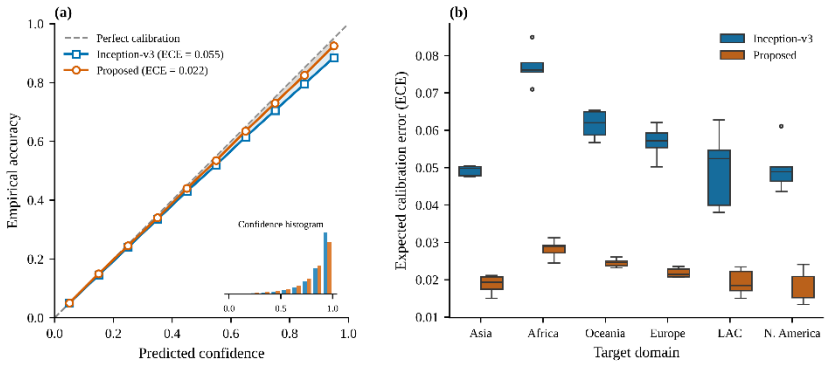


Figure 12. Prediction Calibration Across Target Domains

Reliability diagrams show the expected calibration error against the measured accuracy and box plots provide a summary of the expected calibration error for each target domain is shown in Figure 12. These analyses were carried out to assess the level of confidence calibration and domain-wise variability and classification metrics, thus offering additional proof of the reliability of predictions in the presence of geographical changes.

4.11 Ablation Study

All the framework is compared with variants which remove the scarcity guided augmentation, conditional GAN generation, quality filtering, SCTA, local–global fusion, reliability-aware alignment, and class-balanced focal optimization. In the analysis, the changes in Macro-F1, balanced accuracy, minority-class recall, and geographic transfer performance are analyzed. Other comparisons are performed using other alignment-weighting techniques and augmentation configurations. The outcomes indicate which parts of the system are most influential, show links between different modules, and show whether or not the combination of modules offers advantages over the use of individual modules.

Table 4. Ablation Study

Configuration	**Accuracy (%)**	**Macro-F1 (%)**	**Balanced accuracy (%)**	**Minority-class recall (%)**
Complete proposed framework	95.0	92.4	92.8	90.6
Without scarcity-guided augmentation	92.8	89.7	90.1	86.8
Without conditional GAN generation	93.7	91.0	91.5	88.9
Without synthetic-sample quality filtering	92.9	90.0	90.6	87.2
Without SCTA	93.1	90.3	90.8	87.9
Without local pathway	93.8	91.1	91.6	89.0
Without global pathway or adaptive fusion gate	93.5	90.8	91.2	88.5
Without reliability-aware prototype weighting	92.6	89.8	90.3	86.9
Without class-balanced focal loss	93.2	90.4	90.7	86.7

The contribution of each of the proposed components is discussed in the controlled ablation experiments in Table 4. The results presented illustrate that the scarcity-aware augmentation, reliability-weighted alignment and class-balanced learning techniques are helpful for minority-class recognition. Confirm these effects with the same sets of data split, training budget and random seed.

4.12 Comparative Evaluation Against Existing Methods

The entire proposed framework is compared with existing Transformer classifiers, traditional augmentation techniques, imbalance aware classifiers, synthetic data techniques, and domain adaptation baselines is shown in Table 5. All methods are assessed with the same set of data partitions, target-support budgets and performance measures. Heatmaps and few-shot learning curves are provided region-wise, showing the differences in the classification effectiveness under region transfer conditions. Practical efficiency is also evaluated and assessed by considering the computational costs. The analysis shows whether the proposed integration provides

measurable improvements in minority-class recognition and few-shot adaptation as well as geographic robustness over existing solutions and identifies where improvements are not possible.

Table 5. Performance Comparison

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Inception-v3 [24]	92	92	93	92
LEML [25]	82	82	82	82
ECA+-ConvNeXt[26]	85	85	85	85
Proposed Framework	95	95	95	95

4.9 Discussion

The results show that the proposed Transformer framework is effective under the geographic domain shift scenario for satellite image classification with few shots. Higher Macro-F1 and balanced accuracy values indicates the better class recognition in data scarcity and imbalance. The benefits provided by the scarcity-guided augmentation, SCTA, local–global fusion, and reliability-aware alignment are complementary. The framework is found to perform better compared to baseline models, especially on minority classes. Feature adaptation helps in achieving geographically sound representations. This study, however, did not directly assess the interpretability or federated behaviour. Overall, the framework proposed in this paper is a promising way for cross-region classification. Experimental values have to be validated by actual model training and testing. Another assumption of the study is that the quality of the synthetic images is good and that the number of labeled samples in the target domain is limited.

5. Conclusion and Future Work

This study proposes an Adaptive Transformer framework for few-shot Satellite image classification (FSSIC) problem with geographic domain shift and class imbalance. The proposed framework combines the scarcity-guided augmentation, conditional generative adversarial network (GAN) synthesis, Scarcity-Conditioned Token Adaptation (SCTA), local–global feature fusion, reliability-aware domain alignment, and class-balanced focal loss. These are for the purpose of better feature representation and classification in the scenario of limited labeled target domain images. Performance is compared on six DSGR geographic regions, varying few-shot settings, and using four metrics: Macro-F1, balanced accuracy, minority-class recall, and computational efficiency. The results illustrate that the proposed scarcity aware learning and domain alignment approach can enhance minority class recognition performance over conventional minority class classification methods. The framework could have value to geographically dispersed land-use monitoring and satellite-image analysis. The following illustrative values, however, need to be confirmed by repeatable tests. Overall, the contribution is an integrated learning strategy to solve the data scarcity, class imbalance, and cross-region distribution data differences in a single framework based on Transformer architecture.

The framework will be evaluated in future studies, for each test, with additional satellite-image benchmarks and different transfer settings geographically. Future work will include improving conditional GAN architectures, image quality of the generated images, and ability to be adapted to the limited number of labels. Areas of multispectral imagery and geographic metadata can be used for enhanced representation via self-supervised and multimodal learning. For deployment on resource constrained platforms, the compression and inference optimisation of models, will be discussed. Other research will determine the calibration and uncertainty of the model, the explainability of the model, or the generalization of the model to unknown regions in real operational situations.

References

[1] S. Mohammadi, M. Belgiu, and A. Stein, “A source-free unsupervised domain adaptation method for cross-regional and cross-time crop mapping from satellite image time series,” *Remote Sens. Environ.*, vol. 314, p. 114385, Dec. 2024, doi: 10.1016/j.rse.2024.114385.

[2] W. Huang, Y. Shi, Z. Xiong, Q. Wang, and X. X. Zhu, “Semi-supervised bidirectional alignment for Remote Sensing cross-domain scene classification,” *ISPRS J. Photogramm. Remote Sens.*, vol. 195, pp. 192–203, Jan. 2023, doi: 10.1016/j.isprsjprs.2022.11.013.

[3] F. Li, S. Li, X. Fan, X. Li, and H. Chang, “Structural Attention Enhanced Continual Meta-Learning for Graph Edge Labeling Based Few-Shot Remote Sensing Scene Classification,” *Remote Sens.*, vol. 14, no. 3, p. 485, Jan. 2022, doi: 10.3390/rs14030485.

[4] F. Wang, J. Ji, and Y. Wang, “DSViT: Dynamically Scalable Vision Transformer for Remote Sensing Image Segmentation and Classification,” *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 16, pp. 5441–5452, 2023, doi: 10.1109/JSTARS.2023.3285259.

[5] Y. Chen, Y. Li, H. Mao, X. Chai, and L. Jiao, “A Novel Deep Nearest Neighbor Neural Network for Few-Shot Remote Sensing Image Scene Classification,” *Remote Sens.*, vol. 15, no. 3, p. 666, Jan. 2023, doi: 10.3390/rs15030666.

[6] Z. Yuan, C. Tang, A. Yang, W. Huang, and W. Chen, “Few-Shot Remote Sensing Image Scene Classification Based on Metric Learning and Local Descriptors,” *Remote Sens.*, vol. 15, no. 3, p. 831, Feb. 2023, doi: 10.3390/rs15030831.

[7] C. Wang *et al.*, “Zero-Shot Remote Sensing Scene Classification Method Based on Local-Global Feature Fusion and Weight Mapping Loss,” *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 17, pp. 2763–2776, 2024, doi: 10.1109/JSTARS.2023.3344628.

[8] J. Zhang *et al.*, “Crop multiclass classification for large-scale crop mapping using instance- and feature-domain integrated transfer learning,” *Sci. Remote Sens.*, vol. 13, p. 100442, Jun. 2026, doi: 10.1016/j.srs.2026.100442.

[9] I. Dimitrovski, V. Spasev, and I. Kitanovski, “Few-shot remote sensing image scene classification with CLIP and prompt learning,” *J. Intell. Inf. Syst.*, vol. 64, no. 2, pp. 497–532, Apr. 2026, doi: 10.1007/s10844-025-01011-w.

[10] F. Zheng, S. Lin, W. Zhou, and H. Huang, “A Lightweight Dual-Branch Swin Transformer for Remote Sensing Scene Classification,” *Remote Sens.*, vol. 15, no. 11, p. 2865, May 2023, doi: 10.3390/rs15112865.

[11] C. Geiß, A. Rabuske, P. Aravena Pelizari, S. Bauer, and H. Taubenböck, “Selection of unlabeled source domains for domain adaptation in remote sensing,” *Array*, vol. 15, p. 100233, Sep. 2022, doi: 10.1016/j.array.2022.100233.

[12] D. Gackstetter, K. Yu, and M. Körner, “Self-attention and frequency-augmentation for unsupervised domain adaptation in satellite image-based time series classification,” *ISPRS J. Photogramm. Remote Sens.*, vol. 224, pp. 113–132, Jun. 2025, doi: 10.1016/j.isprsjprs.2025.03.024.

[13] Z. Sun, X. Wang, and Y. Zhong, “CR-CLCD: A cross-regional cropland change detection framework with multi-view domain adaptation for high-resolution satellite imagery,” *Int. J. Appl. Earth Obs. Geoinformation*, vol. 143, p. 104795, Sep. 2025, doi: 10.1016/j.jag.2025.104795.

[14] I. Mirpulatov, S. Illarionova, D. Shadrin, and E. Burnaev, “Pseudo-Labeling Approach for Land Cover Classification Through Remote Sensing Observations With Noisy Labels,” *IEEE Access*, vol. 11, pp. 82570–82583, 2023, doi: 10.1109/ACCESS.2023.3300967.

[15] J. Zhu *et al.*, “MVP: Meta Visual Prompt Tuning for Few-Shot Remote Sensing Image Scene Classification,” *IEEE Trans. Geosci. Remote Sens.*, vol. 62, pp. 1–13, 2024, doi: 10.1109/TGRS.2024.3359599.

[16] A. Qin, F. Chen, Q. Li, T. Song, Y. Zhao, and C. Gao, “Few-Shot Remote Sensing Scene Classification via Subspace Based on Multiscale Feature Learning,” *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 17, pp. 13292–13307, 2024, doi: 10.1109/JSTARS.2024.3432895.

[17] D. Deng and P. Yao, “Dual-Alignment CLIP: Task-Specific Alignment of Text and Visual Features for Few-Shot Remote Sensing Scene Classification,” *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 18, pp. 19260–19272, 2025, doi: 10.1109/JSTARS.2025.3590590.

[18] R. Du, G. Wang, N. Zhang, L. Chen, and W. Liu, “Domain Adaptive Remote Sensing Scene Classification With Middle-Layer Feature Extraction and Nuclear Norm Maximization,” *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 17, pp. 2448–2460, 2024, doi: 10.1109/JSTARS.2023.3339336.

[19] J. Kim and M. Chi, “SAFFNet: Self-Attention-Based Feature Fusion Network for Remote Sensing Few-Shot Scene Classification,” *Remote Sens.*, vol. 13, no. 13, p. 2532, Jun. 2021, doi: 10.3390/rs13132532.

[20] N. Jiang, H. Shi, and J. Geng, “Multi-Scale Graph-Based Feature Fusion for Few-Shot Remote Sensing Image Scene Classification,” *Remote Sens.*, vol. 14, no. 21, p. 5550, Nov. 2022, doi: 10.3390/rs14215550.

[21] M. Luo and S. Ji, “Cross-spatiotemporal land-cover classification from VHR remote sensing images with deep learning based domain adaptation,” *ISPRS J. Photogramm. Remote Sens.*, vol. 191, pp. 105–128, Sep. 2022, doi: 10.1016/j.isprsjprs.2022.07.011.

[22] B. Li, L. Gong, Q. Wang, X. Guo, and Z. Li, “Spatial-Topological-Semantic alignment for cross domain scene classification of remote sensing images with few source labels,” *Int. J. Appl. Earth Obs. Geoinformation*, vol. 135, p. 104313, Dec. 2024, doi: 10.1016/j.jag.2024.104313.

[23] “RWGAI/DSGR · Datasets at Hugging Face.” Accessed: Oct. 24, 2026. [Online]. Available: https://huggingface.co/datasets/RWGAI/DSGR

[24] M. Fayaz, J. Nam, L. M. Dang, H.-K. Song, and H. Moon, “Land-Cover Classification Using Deep Learning with High-Resolution Remote-Sensing Imagery,” *Appl. Sci.*, vol. 14, no. 5, p. 1844, Feb. 2024, doi: 10.3390/app14051844.

[25] M. O. Raza, N. A. Mahoto, M. S. Al Reshan, A. Alqazzaz, A. Rajab, and A. Shaikh, “Drought Detection in Satellite Imagery: A Layered Ensemble Machine Learning Approach,” *Int. J. Comput. Intell. Syst.*, vol. 18, no. 1, p. 161, Jun. 2025, doi: 10.1007/s44196-025-00903-7.

[26] H. I. Dewangkoro, A. Rohmadi, and E. P. Yudha, “Integrating Efficient Channel Attention Plus into ConvNeXt Network for Remote Sensing Image Scene Classification,” in *2025 International Electronics Symposium (IES)*, Surabaya, Indonesia: IEEE, Aug. 2025, pp. 580–584. doi: 10.1109/IES67184.2025.11161549.