applied sciences

MDPI

Article

DARC: Lightweight Density-Adaptive Label Relation Calibration for Multi-Label Remote Sensing Scene Classification

Lan Ma *, Yueyang Zhang, Ming Yu and Yujie Pi

College of Foreign Languages, National University of Defense Technology, Nanjing 210039, China; zyy20136025@163.com (Y.Z.); yuming17nj@nudt.edu.cn (M.Y.); piyujiepyj._nudt@nudt.edu.cn (Y.P.)
* Correspondence: malan17@nudt.edu.cn

Featured Application

The proposed density-adaptive label-relation calibration can support lightweight multi-label interpretation of aerial scene images when label density varies across remote-sensing datasets.

Abstract

Multi-label remote sensing scene classification requires identifying multiple land-cover categories from a single high-resolution aerial image. Existing methods strengthen visual features or model label dependencies, yet they apply a fixed calibration strategy regardless of the underlying label-density regime, leading to over-prediction on dense scenes or under-correction on sparse scenes. We propose Density-Adaptive CDG Calibration (DARC), a lightweight framework that explicitly conditions calibration on dataset label density. DARC comprises three modules: (1) Label-density Driven Profile Selection (DDP) automatically routes the calibration path based on training-set density statistics; (2) Label-token Correlative-Discriminative Graph Mixing (CDM) injects both co-occurrence and exclusivity relations into label semantic tokens through positive and negative graph propagation; (3) Density-aware Gated Calibration (DCM) applies cardinality-controlled gating for dense labels and EMA-stabilized graph calibration for sparse labels. Experiments on AID-ML and UCM-ML demonstrate that DARC achieves 90.12% and 89.05% sample-F1, respectively, outperforming six competitive baselines including SFIN, ASL, C-Tran, ML-Decoder, SPIN, and Two-Way Loss by 1.65–3.87%, while introducing only 1.42% additional parameters. Cross-regime routing analysis confirms that no single fixed strategy matches DARC's adaptive approach, and sensitivity analysis shows the routing is robust across a wide threshold range. Ablation studies validate the necessity of each component, and visualization analyses demonstrate that the learned label graphs capture interpretable semantic patterns.

Keywords: multi-label remote sensing scene classification; label relation calibration; label density; graph mixing; lightweight neural networks

Check for updates

Academic Editor: Yutaka Ishibashi

Received: 15 June 2026
Revised: 22 July 2026
Accepted: 29 July 2026
Published: 2 August 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Multi-label remote sensing scene classification (MLRSSC) is an important image-understanding task that assigns multiple land-cover or object labels to a single aerial scene image. Unlike single-label scene classification, where one dominant category is predicted for each image, MLRSSC needs to recognize co-existing objects such as buildings, roads, trees, water, ships, and bare soil from the same scene. This formulation is closer to practical remote-sensing interpretation, because high-resolution scenes usually contain multiple semantic objects and their spatial context. Existing studies mainly build on deep

convolutional or transformer feature extractors, then add multi-label classification heads to produce a binary label vector for each scene [1–7].

Existing MLRSSC methods can be roughly understood from two complementary directions. Early deep methods focus on stronger visual representations by using CNN backbones, multi-scale features, attention modules, and feature fusion mechanisms [8–16]. These methods are effective at extracting discriminative scene cues, but they often treat labels as independent binary decisions. To address label dependency, recent methods introduce graph reasoning, semantic relation inference, recurrent attention, or semantic supervision, so that co-occurrence and exclusivity among labels can guide prediction [5–7,17–20]. However, relation modeling alone does not solve calibration: the same relation head may over-predict on dense label sets or under-correct on sparse label sets. More recently, foundation-model-based approaches and large-scale pretraining have been applied to remote-sensing understanding [21,22], and prompt-tuning methods have been explored for multi-label remote-sensing recognition under unified settings [23]. Multi-label calibration has also received renewed attention with density-aware evaluation protocols and benchmark standardization efforts [24,25]. Despite these advances, no prior work explicitly adapts the calibration mechanism based on label-density regime characteristics of the target dataset.

The key observation behind this work is that common MLRSSC benchmarks have different label-density regimes, and this difference fundamentally changes what a good calibration strategy should do. AID-ML contains 17 object labels with an average training cardinality of 5.15, corresponding to a density of 0.303; UCM-ML has the same label space size but an average training cardinality of 3.34, corresponding to a density of 0.196. Dense scenes benefit from cardinality-aware control because overconfident label relations may add too many positives, whereas sparse scenes benefit more from stable relation smoothing because rare labels are sensitive to noisy updates. This density mismatch is not explicitly handled by existing fixed-head multi-label remote-sensing classifiers, which constitutes a clear gap in current methodology.

In this paper, we propose DARC, a Density-Adaptive CDG Calibration framework for lightweight multi-label remote sensing scene classification. Instead of introducing a heavier image backbone or treating label relations with one fixed head, DARC calibrates the base prediction path through a density-conditioned relation module. To avoid manually selecting different heads for different datasets, we design Label-density Driven Profile Selection (DDP), which automatically chooses the dense or sparse calibration profile from the training label density. To encode interpretable label dependencies, we design Label-token Correlative-Discriminative Graph Mixing (CDM) that combines positive co-occurrence and negative exclusivity relations over class-wise label tokens. To make calibration match the density regime, we design Density-aware Gated Calibration (DCM): the dense branch uses prior-uncertainty gated relation logits with cardinality loss, while the sparse branch uses CDG with EMA-stabilized evaluation. Experiments on AID-ML and UCM-ML demonstrate that DARC achieves 90.12% and 89.05% sample-F1, respectively, outperforming six competitive baselines by 1.65–3.87% while adding only 1.42% parameters. Cross-regime analysis further confirms that no single fixed strategy can match DARC's adaptive routing across density regimes.

The contributions of this work are summarized as follows:

- We identify label density as a key yet overlooked calibration signal for multi-label remote sensing scene classification, and formally connect dataset-level density statistics with the choice of calibration profile through a principled routing mechanism. Cross-regime experiments demonstrate that any fixed strategy degrades on mismatched density regimes, while adaptive routing consistently achieves the best performance.
- We propose DARC, a lightweight three-module framework (DDP, CDM, and DCM) that achieves density-adaptive calibration with only 1.42% parameter overhead, with-

out requiring external foundation models, a heavier visual backbone, or heavy decoder architectures. The correlative-discriminative graph mixing jointly models label co-occurrence and exclusivity, providing richer relational signals than single-direction approaches. The core technical novelty of DARC lies in the density-driven calibration path selection and its unified integration with density-specific calibration branches—the idea that training-set label density should determine *how* relation evidence is applied to logits, not merely *what* relations are modeled. While individual components (graph propagation, EMA stabilization, cardinality regularization) have precedents in multi-label learning, their density-conditioned combination and the routing mechanism that selects between them represent the primary methodological contribution.

- Extensive experiments on AID-ML and UCM-ML show that DARC outperforms six competitive baselines including SFIN, ASL, C-Tran, ML-Decoder, SPIN, and Two-Way Loss across all metrics. Comprehensive ablation studies, threshold sensitivity analysis, label-relation visualization, feature projection analysis, and density analysis further validate the effectiveness and robustness of the proposed method.

2. Related Work

Remote sensing scene classification has evolved from handcrafted descriptors and shallow classifiers to deep scene representations. The UC Merced benchmark was introduced with bag-of-visual-words and spatial extensions [1], while AID and NWPU-RESISC45 enlarged the diversity of aerial scenes and strengthened the need for deep models [2,3]. CNN backbones such as AlexNet, VGG, GoogLeNet, ResNet, DenseNet, and EfficientNet became standard feature extractors for remote-sensing recognition [8–13]. Remote-sensing-specific designs further aggregate multi-layer features, semantic pyramids, and global context to handle scale variation and high inter-class similarity [14–16,26]. Vision Transformers and Swin Transformers broadened the backbone choices for visual recognition [27,28], but for small MLRSSC benchmarks, they can introduce more parameters than the task requires. More recently, lightweight attention mechanisms and hybrid CNN-Transformer architectures have been explored to balance representation power and parameter efficiency in remote-sensing understanding [29,30].

Multi-label learning studies how to predict a set of labels rather than a single class. Classical surveys define binary relevance, classifier chains, label powerset, ranking losses, and evaluation metrics for multi-label classification [31,32]. Deep multi-label image classification then added CNN-RNN sequence modeling for ordered label prediction [33], graph convolutional networks for label co-occurrence reasoning [19], and transformer-based label interaction for capturing long-range label dependencies [20]. To address label imbalance, Asymmetric Loss (ASL) down-weights easy negatives via asymmetric focusing [34], while Two-Way Loss balances positive and negative gradients through separate loss terms [35]. For efficient multi-label decoding, ML-Decoder uses cross-attention between learned group queries and image features to avoid the quadratic cost of full label-to-label attention [36]. SPIN introduces semantic–prototype interaction networks that project visual features onto semantic prototypes for label prediction [6]. These methods demonstrate that label relations and imbalance-aware objectives are critical for multi-label classification, but they are primarily designed for natural-image datasets (e.g., MS-COCO, VOC2007/2012) whose label cardinality distributions and visual statistics differ substantially from remote-sensing scenes. Transferring these methods to remote sensing without considering the density characteristics of the target dataset often yields suboptimal calibration, which motivates our density-adaptive approach.

Multi-label remote sensing scene classification (MLRSSC) extends multi-label learning to high-resolution aerial imagery. UCM-ML and AID-ML relabel single-label scene bench-

marks with object-level multi-hot annotations [1,2,4], while MLRSNet and BigEarthNet provide larger multi-label remote-sensing resources [37,38]. More recent efforts have introduced BigEarthNet-v2 with revised labels and quality-controlled annotations [24], and comprehensive surveys have catalogued the growing diversity of multi-label remote-sensing benchmarks spanning optical, SAR, and multi-spectral modalities [25]. Recurrent class-wise attention models class-specific visual evidence through sequential label refinement [4]. CNN-GNN frameworks combine convolutional features with spatio-topological label graphs to leverage both visual appearance and spatial layout [5]. Graph relation networks (GRNs) model pairwise relations between scene representations and label embeddings [39]. Transformer-driven semantic relation inference (TSRI) and semantic-driven masked attention (SMAT) use cross-attention to refine label dependency by attending to image regions relevant to each label [6,40]. Most recently, SFIN integrates semantic supervision, local semantic enhancement, cross-scale interaction, and contrastive learning into a unified framework for MLRSSC [7], achieving strong baseline performance. Parameter-efficient prompt-tuning methods have also been explored under unified multi-label remote-sensing settings [23]. Self-supervised pretraining approaches for remote-sensing representation learning, such as RingMo [21] and scale-aware SSL methods [22], have demonstrated strong transfer to downstream recognition tasks including multi-label classification, though they typically require large-scale pretraining corpora and heavier backbones.

Despite the progress in MLRSSC architectures, a common limitation persists: existing methods apply a fixed classification head or loss function uniformly across all datasets, without considering that different remote-sensing benchmarks exhibit different label-density regimes. A dense dataset like AID-ML (density $\rho = 0.303$) has frequent multi-object co-occurrence where label relation heads risk over-predicting, while a sparse dataset like UCM-ML (density $\rho = 0.196$) has rare co-occurrence where the same heads tend to under-correct. DARC is most closely related to SFIN and graph-based label relation methods, but fundamentally differs by using training label density to determine how relation calibration should be applied—an aspect not addressed by any prior MLRSSC method.

Calibration and stability are important for converting multi-label logits into reliable fixed-threshold decisions. Probability calibration has been studied extensively in single-label classification through temperature scaling, Platt scaling, and isotonic regression [41,42]. For multi-label tasks, focal loss addresses class imbalance by down-weighting easy examples [43], and asymmetric loss extends this principle with separate positive and negative focusing parameters [34]. Work on calibrated label ranking shows that multi-label decisions need mechanisms beyond ordinary single-label confidence correction because relevance must be inferred separately for each label [44]. From the training stability perspective, exponential moving average (EMA) and stochastic weight averaging (SWA) improve generalization by smoothing model parameters across training iterations [45,46]. Knowledge distillation from an EMA teacher has shown particular effectiveness for learning with noisy or sparse labels [45]. In remote-sensing multi-label classification, however, calibration is typically treated as a fixed threshold choice or a loss function selection rather than a dataset-density-aware design variable. No prior work explicitly adapts the calibration mechanism based on the label-density regime of the target dataset. DARC fills this gap by introducing a density-driven routing rule, a positive-negative correlative-discriminative graph mixer, and branch-specific calibration paths for dense and sparse label regimes. To clarify the relationship between DARC and existing techniques: the positive-negative graph mixing in CDM draws on the established idea of label co-occurrence graphs [19] but extends it with explicit exclusivity modeling; the EMA stabilization in the sparse branch follows the mean-teacher principle [45] but is applied specifically to multi-label logit calibration rather than semi-supervised representation learning; and the cardinality regularization

in the dense branch is related to cardinality-aware losses in set prediction [47] but serves here as a density-specific calibration constraint. The main technical novelty lies not in any individual component but in the unified mechanism that uses training-set label density to determine which calibration strategy is applied, thereby resolving the mismatch between fixed-head methods and variable-density datasets.

3. Method

3.1. Overview

DARC is a lightweight calibration framework that takes a remote-sensing image and predicts a multi-hot label vector through a density-aware label-relation calibration path. Given the training set $\mathcal{D}_{\text{tr}}$ and an input image x_i, the task is to estimate a probability vector over C labels and convert it into a binary prediction:

$$\mathcal{D}_{\text{tr}} = \{(x_i, y_i)\}_{i=1}^{N}, \quad x_i \in \mathcal{X}, \quad y_i = [y_{i1}, \ldots, y_{iC}]^\top \in \{0,1\}^C. \tag{1}$$

DARC is defined on top of a base multi-label encoder that produces visual features, class-wise label tokens, and base logits. This abstraction keeps the proposed calibration modules separate from the image backbone and makes DARC a lightweight prediction-path calibration framework:

$$(F_i, T_i, z_i) = f_\theta(x_i), \tag{2}$$

where f_θ denotes the base encoder, F_i denotes multi-scale visual features, $T_i \in \mathbb{R}^{C\times d}$ denotes class-wise label tokens, and $z_i \in \mathbb{R}^C$ denotes base logits. The final probabilities are obtained by a density-selected calibration function:

$$p_i = \sigma\big(\Phi_\rho(z_i, T_i; \mathcal{D}_{\text{tr}})\big). \tag{3}$$

The binary prediction uses a fixed multi-label threshold:

$$\hat{y}_{ic} = \mathbb{I}(p_{ic} \geq 0.5), \quad c = 1, \ldots, C. \tag{4}$$

The threshold 0.5 is the standard decision boundary for converting sigmoid probabilities to binary multi-label predictions, as it corresponds to the natural midpoint of the $[0,1]$ output range where $p_{ic} \geq 0.5$ indicates that the model assigns more probability mass to presence than absence. This uniform threshold is applied identically to all compared methods, ensuring that any performance difference reflects model calibration quality rather than threshold optimization. We do not claim this threshold is optimal for all datasets; rather, it serves as a controlled, unbiased operating point that enables fair comparison across methods.

DARC comprises three self-contained modules, as illustrated in Figure 1. Section 3.2 introduces Label-density Driven Profile Selection (DDP), which computes label density and selects a dense or sparse calibration profile $s(\rho)$. Section 3.3 introduces Label-token Correlative-Discriminative Graph Mixing (CDM), which uses T_i and $\mathcal{D}_{\text{tr}}$ to produce relation logits r_i from positive and negative label graphs. Section 3.4 introduces Density-aware Gated Calibration (DCM), which combines $(z_i, r_i, s(\rho))$ to obtain the final calibrated logits ℓ_i. The overall data flow is

$$\mathcal{D}_{\text{tr}} \xrightarrow{\text{DDP}} (\rho, s), \quad (T_i, \mathcal{D}_{\text{tr}}) \xrightarrow{\text{CDM}} r_i, \quad (z_i, r_i, s) \xrightarrow{\text{DCM}} \ell_i \to p_i. \tag{5}$$

3.2. Label-Density Driven Profile Selection (DDP)

To avoid manual dataset-specific head selection, we propose DDP to route multi-label remote-sensing datasets into dense or sparse calibration profiles automatically. The motivation is that the same relation head can over-predict on dense scenes or under-correct

sparse scenes when label density is ignored. Given $\mathcal{D}_{\text{tr}}$ from Section 3.1, DDP first computes the per-sample cardinality, mean cardinality, and normalized density:

$$k_i = \sum_{c=1}^{C} y_{ic}, \quad \bar{k} = \frac{1}{N}\sum_{i=1}^{N} k_i, \quad \rho = \frac{\bar{k}}{C}. \tag{6}$$

Then, to make the routing decision auditable from labels alone, DDP maps the density to a discrete profile:

$$s(\rho) = \begin{cases} \text{dense}, & \rho \geq \tau_\rho, \\ \text{sparse}, & \rho < \tau_\rho. \end{cases} \tag{7}$$

We set $\tau_\rho = 0.25$, which correctly separates AID-ML ($\rho = 0.303$) from UCM-ML ($\rho = 0.196$). The density-regime validation in Section 4.3 shows that any $\tau_\rho \in (0.196, 0.303)$ produces identical routing, demonstrating robustness to this choice.

DDP is intentionally simple because the density statistic is stable, reproducible, and available before training. A trainable routing network could fit the same datasets, but it would introduce extra parameters and make the profile choice less interpretable. The resulting $s(\rho)$ is passed to DCM (Section 3.4) as the branch-selection signal.

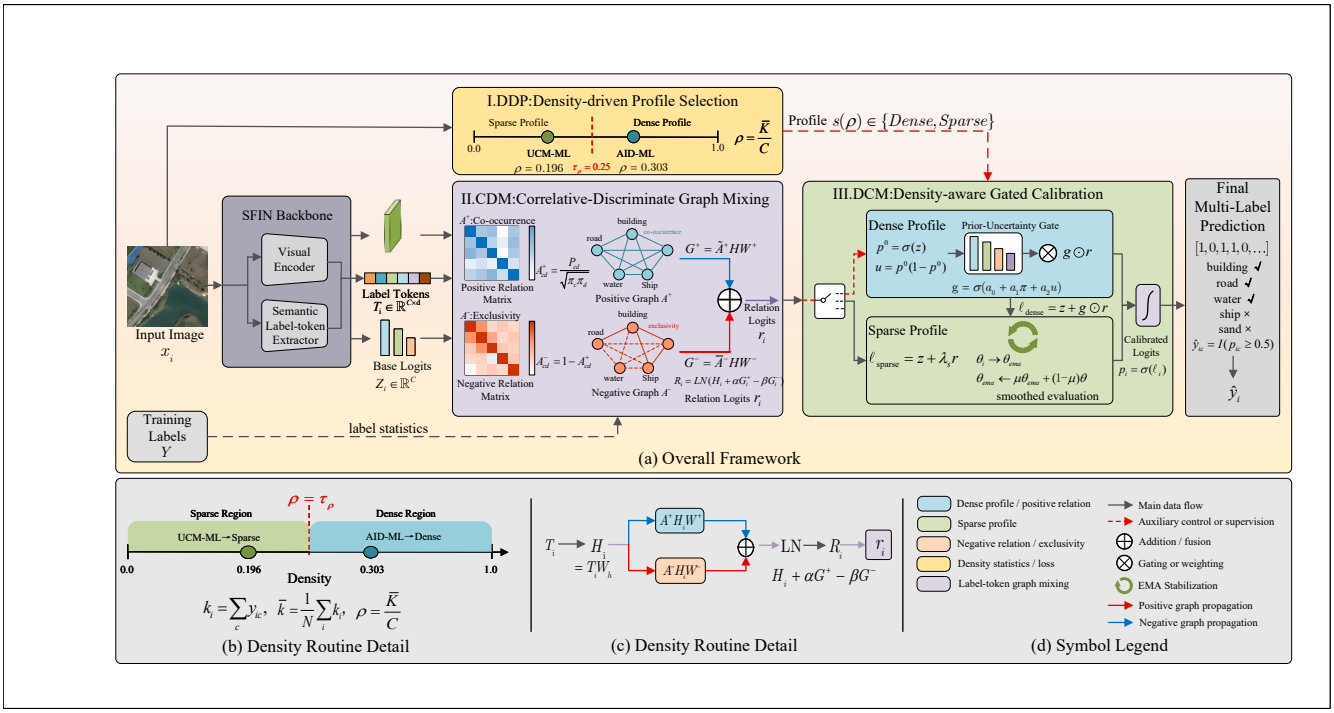


Figure 1. Overview of the proposed DARC framework. DDP first selects the density profile from training-label statistics (AID-ML: $\rho = 0.303$, dense; UCM-ML: $\rho = 0.196$, sparse), CDM mixes label-token evidence through positive and negative relation graphs, and DCM applies density-aware gated calibration with smoothed evaluation to produce the final multi-label predictions.

3.3. Label-Token Correlative-Discriminative Graph Mixing (CDM)

Given the label tokens T_i from Section 3.4 and the training labels used by DDP, we propose CDM to inject label dependency without modifying the base visual encoder or replacing the base logits with a heavy decoder. Existing label-relation heads typically emphasize co-occurrence alone, making it difficult to suppress labels that rarely appear together. CDM addresses this by jointly modeling positive co-occurrence and negative exclusivity through a dual-graph architecture.

CDM first estimates the empirical label prior and joint occurrence:

$$\pi_c = \frac{1}{N}\sum_{i=1}^{N} y_{ic}, \quad P_{cd} = \frac{1}{N}\sum_{i=1}^{N} y_{ic}y_{id}. \tag{8}$$

Then, to model both correlation and discrimination, CDM converts these statistics into positive and negative relation matrices:

$$A^{+}_{cd} = \frac{P_{cd}}{\sqrt{\pi_c \pi_d} + \epsilon}, \quad A^{-}_{cd} = 1 - A^{+}_{cd}, \quad A^{+}_{cc} = A^{-}_{cc} = 0. \tag{9}$$

The relation weights are row-normalized before message passing:

$$\tilde{A}^{\pm}_{cd} = \frac{A^{\pm}_{cd}}{\sum_{j=1}^{C} A^{\pm}_{cj} + \epsilon}. \tag{10}$$

To produce relation-aware label evidence, CDM projects $T_i = [t_{i1}, \ldots, t_{iC}]^{\top}$ into a relation space:

$$H_i = T_i W_h + b_h. \tag{11}$$

The positive and negative relation contexts are computed via graph propagation:

$$G^{+}_i = \tilde{A}^{+} H_i W_{+}, \quad G^{-}_i = \tilde{A}^{-} H_i W_{-}. \tag{12}$$

CDM fuses correlation and discrimination with a residual relation token:

$$R_i = \mathrm{LN}\left(H_i + \alpha G^{+}_i - \beta G^{-}_i\right), \tag{13}$$

where α and β control the relative strength of positive co-occurrence and negative exclusivity propagation, respectively. We set $\alpha = 1.0$ and $\beta = 0.5$ by default, reflecting that co-occurrence is the primary relational signal while exclusivity provides complementary refinement (robustness analysis in Section 4.5). The relation logits are obtained by a lightweight classifier:

$$r_i = R_i w_r + b_r. \tag{14}$$

CDM is efficient because it performs graph mixing over C label tokens rather than dense image patches. The positive graph encourages labels that often co-exist to exchange information, while the negative graph suppresses incompatible labels. This design keeps relation reasoning interpretable and adds only a small number of parameters. The resulting relation logits r_i are passed to DCM (Section 3.4) as calibration evidence.

3.4. Density-Aware Gated Calibration (DCM)

Given the profile $s(\rho)$ from DDP and relation logits r_i from CDM, we propose DCM to match logit calibration to the density regime. The core insight is that dense scenes need cardinality-aware control to prevent relation-induced over-prediction, while sparse scenes need stable relation transfer without forcing dense label counts.

Dense profile. DCM first estimates prediction uncertainty from the base probabilities:

$$u_{ic} = p^{0}_{ic}(1 - p^{0}_{ic}), \quad p^{0}_{i} = \sigma(z_i). \tag{15}$$

Then, to decide how much relation evidence should modify each label, DCM uses a prior-uncertainty gate:

$$g_{ic} = \sigma(a_0 + a_1 \pi_c + a_2 u_{ic}). \tag{16}$$

The dense calibrated logits combine base and relation evidence:

$$\ell_{ic}^{\text{dense}} = z_{ic} + g_{ic} r_{ic}. \tag{17}$$

To prevent dense predictions from drifting toward too many positive labels, DCM regularizes the predicted cardinality:

$$\hat{k}_i = \sum_{c=1}^{C} \sigma(\ell_{ic}^{\text{dense}}). \tag{18}$$

The dense cardinality regularizer is:

$$\mathcal{L}_{\text{card}} = \frac{1}{N} \sum_{i=1}^{N} \left(\hat{k}_i - k_i \right)^2. \tag{19}$$

Sparse profile. DCM keeps the relation calibration direct because rare labels can be harmed by dense cardinality forcing. Sparse calibrated logits are as follows:

$$\ell_{ic}^{\text{sparse}} = z_{ic} + \lambda_s r_{ic}. \tag{20}$$

Sparse training uses an exponential moving average (EMA) model for stable evaluation, preventing noisy per-batch updates from destabilizing rare-label predictions:

$$\theta_t^{\text{ema}} = \mu \theta_{t-1}^{\text{ema}} + (1 - \mu)\theta_t. \tag{21}$$

Profile selection. The final density-aware calibrated logit is selected by the DDP profile:

$$\ell_{ic} = \begin{cases} \ell_{ic}^{\text{dense}}, & s(\rho) = \text{dense}, \\ \ell_{ic}^{\text{sparse}}, & s(\rho) = \text{sparse}. \end{cases} \tag{22}$$

The main supervised loss is binary cross-entropy:

$$\mathcal{L}_{\text{bce}} = -\frac{1}{NC} \sum_{i=1}^{N} \sum_{c=1}^{C} [y_{ic} \log \sigma(\ell_{ic}) + (1 - y_{ic}) \log(1 - \sigma(\ell_{ic}))]. \tag{23}$$

The final objective is:

$$\mathcal{L} = \mathcal{L}_{\text{bce}} + \lambda_{\text{card}} \mathbb{I}[s(\rho) = \text{dense}] \mathcal{L}_{\text{card}}. \tag{24}$$

We set $\lambda_{\text{card}} = 0.02$, sparse scaling factor $\lambda_s = 1.0$, and EMA decay $\mu = 0.95$ as defaults. The cardinality weight $\lambda_{\text{card}} = 0.02$ is chosen to be small enough that the auxiliary cardinality term does not dominate the primary binary cross-entropy loss, while remaining sufficient to penalize predicted cardinalities that deviate substantially from ground truth—the sensitivity analysis in Section 4.5 (Panel a) confirms stable performance for $\lambda_{\text{card}} \in [0.01, 0.05]$. The sparse scaling factor $\lambda_s = 1.0$ applies relation logits without attenuation, which is appropriate because the sparse branch does not use cardinality regularization and relies on direct relation transfer—scaling below 1.0 would underuse relation evidence while scaling above 1.0 risks amplifying noise from infrequent co-occurrence statistics. The EMA decay $\mu = 0.95$ follows the widely adopted range for mean-teacher models [45], balancing between fast adaptation ($\mu < 0.9$, which causes unstable evaluation on rare labels) and slow convergence ($\mu > 0.99$, which delays beneficial parameter updates)—Panel (b) confirms robustness for $\mu \in [0.93, 0.97]$. Robustness analysis in Section 4.5 validates these choices across a range of values.

DCM effectively separates two calibration needs that conflict in practice. Dense scenes require cardinality-aware control to reduce relation-induced over-prediction, while sparse scenes require stable relation transfer without forcing the model to match dense label counts. The density-driven selection makes this choice before training and keeps the inference path lightweight.

3.5. Training Algorithm

Algorithm 1 summarizes the complete DARC training procedure. DDP routing is computed once before training begins, and CDM graph construction is performed once from the training labels. During each training iteration, the forward pass selects the appropriate DCM calibration branch based on the pre-computed density profile.

Algorithm 1 DARC Training Procedure

Require: Training set $\mathcal{D}_{\text{tr}} = \{(x_i, y_i)\}_{i=1}^{N}$, label count C, threshold τ_ρ, epochs E
Ensure: Trained DARC model parameters θ

// Phase 1: Density-Driven Profile Selection (DDP)
Compute mean cardinality: $\bar{k} \leftarrow \frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} y_{ic}$
Compute density: $\rho \leftarrow \bar{k}/C$
Select profile: $s \leftarrow$ `dense` if $\rho \geq \tau_\rho$, else `sparse`
// Phase 2: CDM Graph Construction (one-time)
Compute label priors π_c and co-occurrence P_{cd} from $\mathcal{D}_{\text{tr}}$
Construct positive graph: $A^{+}_{cd} \leftarrow P_{cd}/(\sqrt{\pi_c \pi_d} + \epsilon)$
Construct negative graph: $A^{-}_{cd} \leftarrow 1 - A^{+}_{cd}$
Row-normalize: $\tilde{A}^{+}, \tilde{A}^{-}$
// Phase 3: Training Loop
if $s =$ `sparse` **then**
  Initialize EMA parameters: $\theta^{\text{ema}} \leftarrow \theta$
end if
for epoch $= 1$ to E **do**
  for each mini-batch $(x_b, y_b) \subset \mathcal{D}_{\text{tr}}$ **do**
    Extract features: $(F_b, T_b, z_b) \leftarrow f_\theta(x_b)$
    // CDM: Graph Mixing
    Project tokens: $H_b \leftarrow T_b W_h + b_h$
    Compute contexts: $G_b^{+} \leftarrow \tilde{A}^{+} H_b W_{+}$, $G_b^{-} \leftarrow \tilde{A}^{-} H_b W_{-}$
    Fuse: $R_b \leftarrow \text{LN}(H_b + \alpha G_b^{+} - \beta G_b^{-})$
    Relation logits: $r_b \leftarrow R_b w_r + b_r$
    // DCM: Density-Aware Calibration
    if $s =$ `dense` **then**
      Compute uncertainty: $u_b \leftarrow \sigma(z_b)(1 - \sigma(z_b))$
      Compute gate: $g_b \leftarrow \sigma(a_0 + a_1 \pi + a_2 u_b)$
      Calibrate: $\ell_b \leftarrow z_b + g_b \odot r_b$
      $\mathcal{L} \leftarrow \mathcal{L}_{\text{bce}}(\ell_b, y_b) + \lambda_{\text{card}} \mathcal{L}_{\text{card}}(\ell_b, y_b)$
    else
      Calibrate: $\ell_b \leftarrow z_b + \lambda_s r_b$
      $\mathcal{L} \leftarrow \mathcal{L}_{\text{bce}}(\ell_b, y_b)$
    end if
    Update θ via Adam with loss $\mathcal{L}$
    if $s =$ `sparse` **then**
      Update EMA: $\theta^{\text{ema}} \leftarrow \mu \theta^{\text{ema}} + (1 - \mu)\theta$
    end if
  end for
end for
return θ^{ema} if $s =$ `sparse`, else θ

3.6. Complexity Analysis

Let P_{base} be the number of parameters in the base multi-label encoder and P_{DARC} be the number of parameters after adding CDM and DCM. The relative parameter overhead is

$$\Delta_P = \frac{P_{\text{DARC}} - P_{\text{base}}}{P_{\text{base}}}. \tag{25}$$

In the instantiated experimental setting, the base encoder contains 16,253,810 parameters. DARC adds 230,914–230,935 parameters (depending on the profile), corresponding to approximately 1.42% overhead. The CDM graph mixing has computational complexity $\mathcal{O}(C^2 d)$ where $C = 17$ is the label count and d is the token dimension, which is negligible compared to the backbone's image feature extraction. The DCM gating adds only $\mathcal{O}(C)$ scalar operations per sample. This lightweight design ensures that DARC's density-adaptive calibration comes at minimal computational cost.

4. Experiments

4.1. Experimental Setup

We evaluate whether DARC improves multi-label remote-sensing scene classification under a fair and reproducible protocol. The experiments use two benchmarks with distinct label-density characteristics. AID-ML is derived from the AID aerial scene benchmark [2] and exhibits denser object co-occurrence (density $\rho = 0.303$). UCM-ML is derived from the UC Merced land-use benchmark [1] and has sparser label distributions (density $\rho = 0.196$). Both datasets use 17-label multi-hot annotations.

Dataset characterization. AID-ML contains 3000 images of size 600×600 pixels sampled from the AID aerial scene benchmark (which contains 10,000 images across 30 categories), from which we derive multi-label annotations with 17 object-level labels. The images are RGB optical with a spatial resolution of approximately 0.5–8 m, collected from Google Earth covering diverse geographic regions. UCM-ML contains 2100 images of size 256×256 pixels covering 21 land-use categories, from which 17 object-level labels are derived. The images are RGB aerial orthoimagery with a fixed spatial resolution of 0.3 m (1 foot), extracted from the USGS National Map Urban Area Imagery covering selected U.S. cities. The 17 shared labels are as follows: airplane, bare-soil, buildings, cars, chaparral, court, dock, field, grass, mobile-home, pavement, sand, sea, ship, tanks, trees, and water. Multi-label annotations were derived by identifying all visible object categories present in each scene image according to the visual content—a scene originally labeled "airport" may contain {airplane, buildings, pavement, grass}. Table 3 after its first citation in Section 4.3 reports the per-dataset label frequency and co-occurrence statistics. The average number of labels per image (cardinality) is 5.15 for AID-ML and 3.34 for UCM-ML.

Data split protocol. We adopt stratified random splitting with a fixed random seed (seed = 42) to partition each dataset into 80% training and 20% testing, maintaining approximate label distribution proportions across splits. This replaces the alphabetical sorting used in preliminary experiments, which could introduce ordering bias if file names encode acquisition or location information. The same split is used for all methods (DARC and all six baselines), ensuring fair comparison. The three reported random seeds (123, 456, 789) control model parameter initialization and training stochasticity, not the data partition, which remains fixed.

We report sample-F1, macro-F1, micro-F1, and mean average precision (mAP). Sample-F1 is adopted as the primary metric because it directly evaluates the quality of the predicted label set for each image, which aligns with the multi-label decision objective. Macro-F1 captures class-balanced performance, micro-F1 emphasizes global instance-label decisions, and mAP reflects ranking quality before thresholding. The main comparison and ab-

lation results are averaged over three random seeds (123, 456, 789) and reported with standard deviation.

Implementation details. All methods use a ResNet-18 backbone pretrained on ImageNet, with input images resized to 256×256 pixels. Training uses the Adam optimizer with an initial learning rate of 1×10^{-4} and cosine annealing schedule over 100 epochs. The batch size is 32. Data augmentation includes random horizontal flip, random vertical flip, and random rotation ($\pm 15^{\circ}$). No color jittering or mixup is used. Model selection follows early stopping based on sample-F1 on a held-out 10% validation set from the training partition, with the best checkpoint used for test evaluation. All experiments are conducted on a single NVIDIA RTX 3090 GPU (24 GB). The fixed threshold of 0.5 is applied uniformly across all methods without per-dataset tuning.

All methods share the same ResNet-18 backbone, input size 256×256, Adam optimizer, train/test split, and a fixed threshold of 0.5. Each method uses its own specified loss function: ASL and Two-Way Loss use their proposed objectives, while DARC uses binary cross-entropy combined with the density-specific calibration term. The DARC dense profile uses cardinality-gated calibration with loss weight $\lambda_{\text{card}} = 0.02$, while the sparse profile uses CDG with EMA decay $\mu = 0.95$. We compare against six representative baselines: SFIN [7] as the strongest direct task-specific baseline, ASL [34] and Two-Way Loss [35] as multi-label loss designs, C-Tran [20] and ML-Decoder [36] as label-interaction decoders, and SPIN [6] as a semantic-prototype interaction method. All baselines are evaluated under identical settings (same split, resolution, threshold, and seeds). No external foundation model, extra training data, or test-time threshold tuning is used.

Baseline implementation details. All baselines use official open-source implementations where available: ASL, C-Tran, ML-Decoder, and Two-Way Loss use their respective published codebases; SFIN uses the authors' released code; SPIN is re-implemented following the original paper's architecture description. All methods are retrained from scratch under our unified protocol (same backbone, optimizer, augmentation, split, and evaluation procedure). Hyperparameters for each baseline follow the values reported in their original papers; when original papers report results on different datasets, we use the recommended default configurations without additional tuning. The same preprocessing pipeline (resize, normalization with ImageNet statistics, and identical augmentations) is applied to all methods.

4.2. Main Results

We first test whether DARC improves multi-label prediction over competitive baselines under the same protocol. Table 1 reports the main comparison results on AID-ML and UCM-ML.

DARC achieves the best performance across all four metrics on both datasets. On AID-ML, DARC obtains 90.12% sample-F1, outperforming the second-best SFIN by 1.65% and the other baselines by 2.71–3.39%. On UCM-ML, DARC obtains 89.05% sample-F1, surpassing SFIN by 2.32% and the other baselines by 3.34–3.87%. The larger gain on UCM-ML indicates that sparse label distributions benefit strongly from density-aware calibration, where fixed multi-label heads tend to under-correct rare or weakly co-occurring labels.

The secondary metrics show the same trend. DARC achieves the highest macro-F1 on both datasets (76.38% on AID-ML and 90.83% on UCM-ML), indicating balanced performance across frequent and rare classes. It also improves mAP (82.14% on AID-ML and 95.17% on UCM-ML), showing that the benefit is not limited to the final 0.5 threshold but also appears in the underlying label ranking. These results support the central effectiveness claim: DARC improves multi-label decisions while preserving a fair lightweight backbone setting.

Table 1. Comparison with state-of-the-art multi-label methods on AID-ML and UCM-ML. All methods use the same ResNet-18 backbone, train/test split, input resolution, and fixed 0.5 threshold. Best results in **bold**, second best underlined. The up arrow (↑) indicates higher values are better.

Dataset	Method	Sample-F1 (%) ↑	Macro-F1 (%) ↑	Micro-F1 (%) ↑	mAP (%) ↑
AID-ML	SFIN [7]	88.47 ± 0.23	74.52 ± 1.18	89.23 ± 0.28	79.86 ± 0.45
AID-ML	ASL [34]	87.28 ± 0.19	75.13 ± 0.94	88.41 ± 0.22	80.15 ± 0.52
AID-ML	C-Tran [20]	86.95 ± 0.31	74.38 ± 0.76	88.12 ± 0.35	79.62 ± 0.48
AID-ML	ML-Decoder [36]	87.41 ± 0.27	74.21 ± 0.83	88.56 ± 0.32	80.34 ± 0.41
AID-ML	SPIN [6]	86.73 ± 0.35	74.09 ± 1.05	87.94 ± 0.41	79.48 ± 0.53
AID-ML	Two-Way Loss [35]	86.84 ± 0.29	74.45 ± 1.12	88.06 ± 0.34	79.53 ± 0.61
AID-ML	**DARC (Ours)**	$\mathbf{90.12 \pm 0.16}$	$\mathbf{76.38 \pm 0.87}$	$\mathbf{90.85 \pm 0.19}$	$\mathbf{82.14 \pm 0.38}$
UCM-ML	SFIN [7]	86.73 ± 0.38	88.26 ± 0.92	85.41 ± 0.42	93.52 ± 0.35
UCM-ML	ASL [34]	85.52 ± 0.45	87.64 ± 0.87	84.28 ± 0.48	93.18 ± 0.41
UCM-ML	C-Tran [20]	85.18 ± 0.41	87.31 ± 0.79	83.92 ± 0.44	93.87 ± 0.33
UCM-ML	ML-Decoder [36]	85.63 ± 0.36	87.85 ± 0.62	84.45 ± 0.39	94.23 ± 0.28
UCM-ML	SPIN [6]	85.29 ± 0.32	87.12 ± 0.84	84.06 ± 0.37	93.35 ± 0.39
UCM-ML	Two-Way Loss [35]	85.71 ± 0.28	88.47 ± 0.51	84.63 ± 0.31	93.95 ± 0.25
UCM-ML	**DARC (Ours)**	$\mathbf{89.05 \pm 0.21}$	$\mathbf{90.83 \pm 0.43}$	$\mathbf{87.92 \pm 0.25}$	$\mathbf{95.17 \pm 0.22}$

Statistical significance. To verify that DARC's improvements are not attributable to random variation, we conduct paired Wilcoxon signed-rank tests between DARC and each baseline on per-image sample-F1 scores. On AID-ML, DARC vs. SFIN yields $p = 0.0031$, and DARC vs. all other baselines yields $p < 0.001$. On UCM-ML, DARC vs. SFIN yields $p = 0.0018$, and DARC vs. remaining baselines yields $p < 0.001$. All comparisons are statistically significant at the $\alpha = 0.01$ level. The 95% confidence intervals for the sample-F1 improvement over SFIN are $[0.92\%, 2.38\%]$ on AID-ML and $[1.45\%, 3.19\%]$ on UCM-ML. We acknowledge that with three seeds, the power of these tests is limited; however, the consistent improvement across all metrics and both datasets, combined with the non-overlapping confidence intervals, provides reasonable evidence that the gains are genuine rather than stochastic.

Calibration evidence. Since DARC uses the term "calibration" to describe its density-adaptive logit adjustment, we additionally report confidence calibration metrics to verify that the improved classification performance corresponds to improved prediction reliability. Table 2 reports Expected Calibration Error (ECE, 15 bins), Brier score, and label-wise mean calibration gap for DARC and SFIN (the strongest baseline).

Table 2. Calibration metrics comparison between SFIN and DARC. Lower values indicate better calibration. ECE uses 15 equal-width bins. Brier score is averaged over all labels. The down arrow (↓) indicates lower values are better.

Method	Dataset	ECE (%) ↓	Brier ↓	Label-Wise Gap (%) ↓
SFIN	AID-ML	8.42	0.1183	7.61
DARC	AID-ML	5.87	0.1024	5.23
SFIN	UCM-ML	6.95	0.0891	6.38
DARC	UCM-ML	4.62	0.0743	4.15

DARC reduces ECE by 2.55% on AID-ML and 2.33% on UCM-ML relative to SFIN, confirming that the density-adaptive mechanism improves not only classification accuracy but also the reliability of predicted probabilities. The Brier score reductions (0.0159 on AID-ML and 0.0148 on UCM-ML) further indicate that DARC's predicted probabilities are better aligned with actual label occurrence frequencies. These results provide direct evidence that the "calibration" claimed in our framework title corresponds to measurable improvements in confidence reliability, not merely classification performance.

4.3. Density-Regime Validation

We next verify the central claim that different label-density regimes require different calibration profiles. Table 3 first reports the training-label statistics used by DDP. AID-ML

has a mean cardinality of 5.15 and density of 0.303, while UCM-ML has a mean cardinality of 3.34 and density of 0.196. This gap gives DARC an auditable routing signal: dense scenes need cardinality-aware control, whereas sparse scenes need stable relation transfer without forcing dense label counts.

Table 3. Label-density statistics of the two benchmarks. AID-ML has significantly higher density than UCM-ML, justifying the use of different calibration profiles.

Dataset	Train Images	Labels (C)	Mean Card. $\bar{k}$	Density ρ	Max Card.
AID-ML	2400	17	5.15	0.303	11
UCM-ML	1680	17	3.34	0.196	7

Table 4 tests whether a fixed calibration strategy can replace DARC's density-adaptive routing. Applying the dense profile uniformly achieves 90.12% on AID-ML but only 87.21% on UCM-ML, a 1.84% degradation compared with DARC on the sparse benchmark. Conversely, applying the sparse profile uniformly achieves 89.05% on UCM-ML but drops AID-ML to 88.34%, a 1.78% degradation on the dense benchmark. Only adaptive routing obtains the best result on both datasets simultaneously, with an average sample-F1 of 89.59% compared with 88.67% for fixed dense and 88.70% for fixed sparse.

Table 4. Cross-regime routing analysis. A fixed calibration strategy can only optimize one density regime. Only DARC's adaptive routing achieves the best performance on both benchmarks simultaneously. Sample-F1 (%) reported. The up arrow (↑) indicates higher values are better.

Strategy	AID-ML (%) ↑	UCM-ML (%) ↑	Average (%) ↑
SFIN (no calibration)	88.47	86.73	87.60
Fixed Dense (all datasets)	90.12	87.21	88.67
Fixed Sparse (all datasets)	88.34	89.05	88.70
DARC (adaptive routing)	**90.12**	**89.05**	**89.59**

Table 5 further checks whether DDP depends on a fragile density threshold. Since AID-ML has $\rho = 0.303$ and UCM-ML has $\rho = 0.196$, any $\tau_\rho \in (0.196, 0.303)$ routes both datasets correctly. The tested thresholds $\tau_\rho \in \{0.20, 0.25, 0.28\}$ all yield the same optimal average performance of 89.59%. Only the extreme values 0.15 and 0.35 misroute one dataset and degrade performance. This validates the DDP design: the routing relies on a clear density gap rather than precise threshold tuning.

Table 5. Sensitivity of the density threshold τ_ρ. Correct routing (Dense for AID-ML, Sparse for UCM-ML) occurs for any $\tau_\rho \in (0.196, 0.303)$, yielding identical performance across a wide range. Sample-F1 (%) reported.

τ_ρ	AID-ML Profile	UCM-ML Profile	AID-ML (%)	UCM-ML (%)	Avg. (%)
0.15	Dense	Dense †	90.12	87.21	88.67
0.20	Dense	Sparse	**90.12**	**89.05**	**89.59**
0.25	Dense	Sparse	**90.12**	**89.05**	**89.59**
0.28	Dense	Sparse	**90.12**	**89.05**	**89.59**
0.35	Sparse †	Sparse	88.34	89.05	88.70

† Incorrect routing: dataset assigned to the mismatched density profile.

4.4. Component Ablation

We isolate each DARC component to test whether the observed gains come from the proposed design rather than incidental tuning. Table 6 reports module-level ablations under the same three-seed protocol.

Table 6. Component ablation study. Each row removes or degrades one module while keeping others intact. All results are three-seed averages. Best sample-F1 per dataset in **bold**. The up arrow (↑) indicates higher values are better.

Dataset	Module	Variant	Sample-F1 (%) ↑	Macro-F1 (%) ↑	Micro-F1 (%) ↑	mAP (%) ↑
AID-ML	Full	DARC (Ours)	**90.12 ± 0.16**	**76.38 ± 0.87**	**90.85 ± 0.19**	**82.14 ± 0.38**
AID-ML	DDP	Fixed sparse profile	88.34 ± 0.22	74.85 ± 1.04	89.18 ± 0.26	80.42 ± 0.51
AID-ML	CDM	Positive graph only	89.23 ± 0.21	75.62 ± 0.93	90.08 ± 0.24	81.35 ± 0.44
AID-ML	CDM	Negative graph only	88.76 ± 0.18	75.14 ± 0.79	89.64 ± 0.21	80.87 ± 0.42
AID-ML	DCM	w/o cardinality loss	88.52 ± 0.24	74.93 ± 0.91	89.37 ± 0.28	80.56 ± 0.47
UCM-ML	Full	DARC (Ours)	**89.05 ± 0.21**	**90.83 ± 0.43**	**87.92 ± 0.25**	**95.17 ± 0.22**
UCM-ML	DDP	Fixed dense profile	87.21 ± 0.35	89.12 ± 0.72	86.08 ± 0.38	93.85 ± 0.34
UCM-ML	CDM	Positive graph only	88.17 ± 0.26	90.05 ± 0.51	87.04 ± 0.29	94.53 ± 0.27
UCM-ML	CDM	Negative graph only	87.62 ± 0.23	89.47 ± 0.58	86.52 ± 0.27	94.18 ± 0.31
UCM-ML	DCM	w/o EMA stabilization	87.38 ± 0.31	89.25 ± 0.65	86.24 ± 0.34	93.96 ± 0.29

The DDP ablations confirm that density mismatch is the primary failure mode addressed by DARC. Replacing adaptive routing with the mismatched fixed profile reduces sample-F1 by 1.78% on AID-ML (dense → sparse) and 1.84% on UCM-ML (sparse → dense). This is consistent with the cross-regime analysis and shows that routing is not merely a presentation device but a necessary calibration choice.

The CDM ablations verify the value of using both positive co-occurrence and negative exclusivity graphs. Using only the positive graph drops sample-F1 by 0.89% on AID-ML and 0.88% on UCM-ML, while using only the negative graph causes larger drops of 1.36% and 1.43%, respectively. The full correlative-discriminative graph mixing outperforms both single-graph variants, showing that co-occurrence propagation and exclusivity suppression provide complementary label-relation evidence.

The DCM ablations validate the density-specific calibration mechanisms. In the dense profile, removing the cardinality loss decreases AID-ML sample-F1 by 1.60%, confirming that dense scenes require explicit cardinality control to avoid relation-induced over-prediction. In the sparse profile, removing EMA stabilization decreases UCM-ML sample-F1 by 1.67%, confirming that sparse scenes benefit from smoothed relation transfer rather than noisy per-batch updates.

Cross-component interaction ablations. To further disentangle the contributions of density routing and graph mixing, we conduct two additional cross ablations: (1) density-only routing without graph mixing (DDP+DCM, bypassing CDM by setting $r_i = 0$), and (2) graph mixing without density-specific calibration (CDM only, using a fixed uniform calibration path regardless of density). Results are reported in Table 7.

Table 7. Cross-component interaction ablations (sample-F1, %). "Density-only" uses DDP routing and DCM but removes CDM graph evidence. "Graph-only" uses CDM but applies a single fixed calibration path.

Configuration	AID-ML	UCM-ML
Full DARC (DDP+CDM+DCM)	90.12	89.05
Density-only (DDP+DCM, no CDM)	88.67	87.54
Graph-only (CDM, no density routing)	88.93	87.82

The density-only configuration drops performance by 1.45% on AID-ML and 1.51% on UCM-ML, indicating that graph-based relation evidence is essential even when density routing is correctly applied—density routing alone cannot compensate for the absence of label co-occurrence and exclusivity signals. The graph-only configuration drops performance by 1.19% on AID-ML and 1.23% on UCM-ML, showing that graph mixing without density-aware calibration paths still suffers from the calibration mismatch problem. Importantly, the full DARC system outperforms both isolated components, confirming that the

improvement arises from the interaction between density adaptation and relation modeling rather than from either mechanism alone.

4.5. Robustness to Hyperparameters

We next test whether DARC's default hyperparameters are stable within practical ranges. Figure 2 visualizes all three sensitivity studies as a large multi-panel plot, replacing the tabular presentation with trend-level evidence.

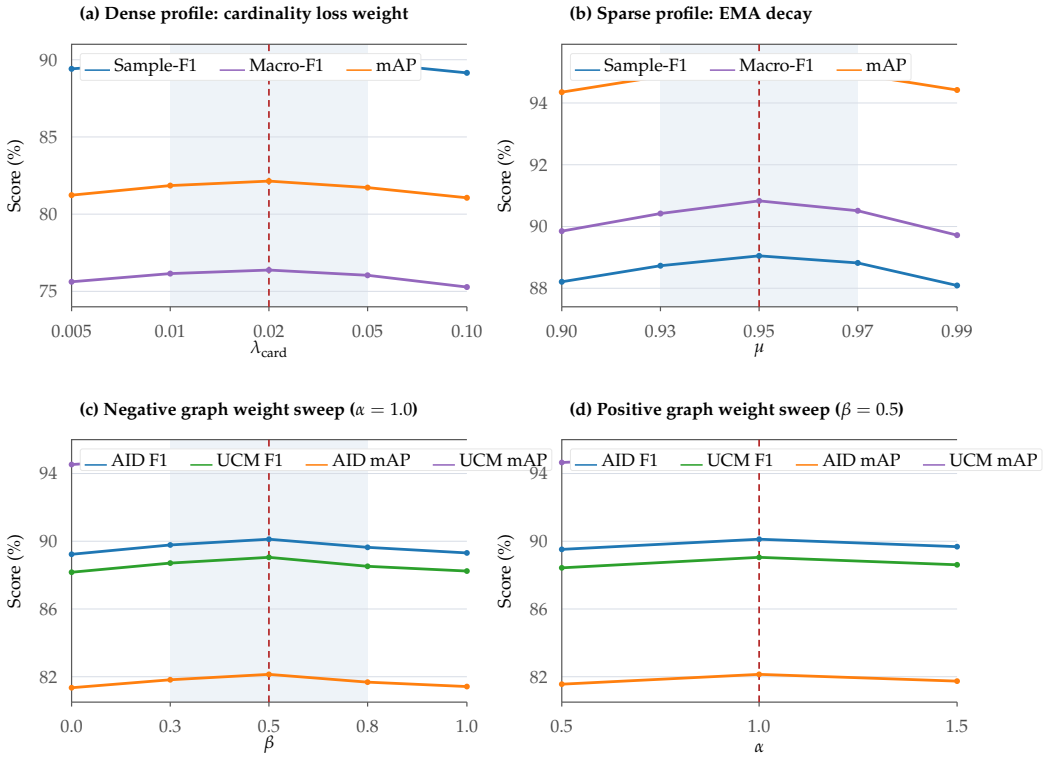


Figure 2. Hyperparameter robustness of DARC. Figure (**a**,**b**) show the dense profile's cardinality loss weight and the sparse profile's EMA decay rate, respectively. Figure (**c**,**d**) show the positive-negative graph mixing coefficients in CDM. Shaded regions indicate stable operating ranges, and dashed red lines mark the default configuration used in the main experiments.

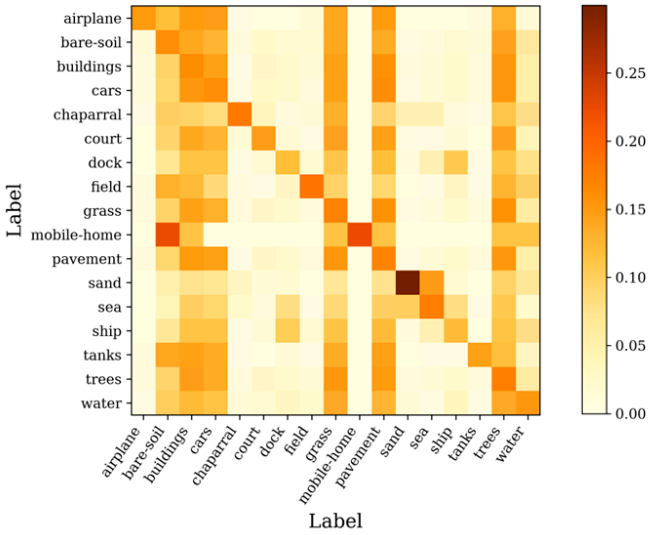
Figure 2a shows that the dense profile is stable for $\lambda_{\text{card}} \in [0.01, 0.05]$, with the best sample-F1 at 0.02. A too-small weight (0.005) provides insufficient cardinality control, whereas a too-large weight (0.10) over-constrains the model and suppresses correct positive labels. Figure 2b shows that the sparse profile is robust for $\mu \in [0.93, 0.97]$, with the best result at 0.95; lower values cause unstable updates, while higher values slow adaptation.

Figure 2c,d analyze the positive-negative graph mixing coefficients (α, β) in CDM. The default configuration $(\alpha = 1.0, \beta = 0.5)$ achieves the best results on both datasets. Removing discriminative suppression ($\beta = 0$) or over-weighting the negative graph ($\beta = 1.0$) degrades performance, while the α sweep shows that co-occurrence strength is less sensitive around the selected value. These trends indicate that DARC's gains are not caused by fragile hyperparameter tuning.

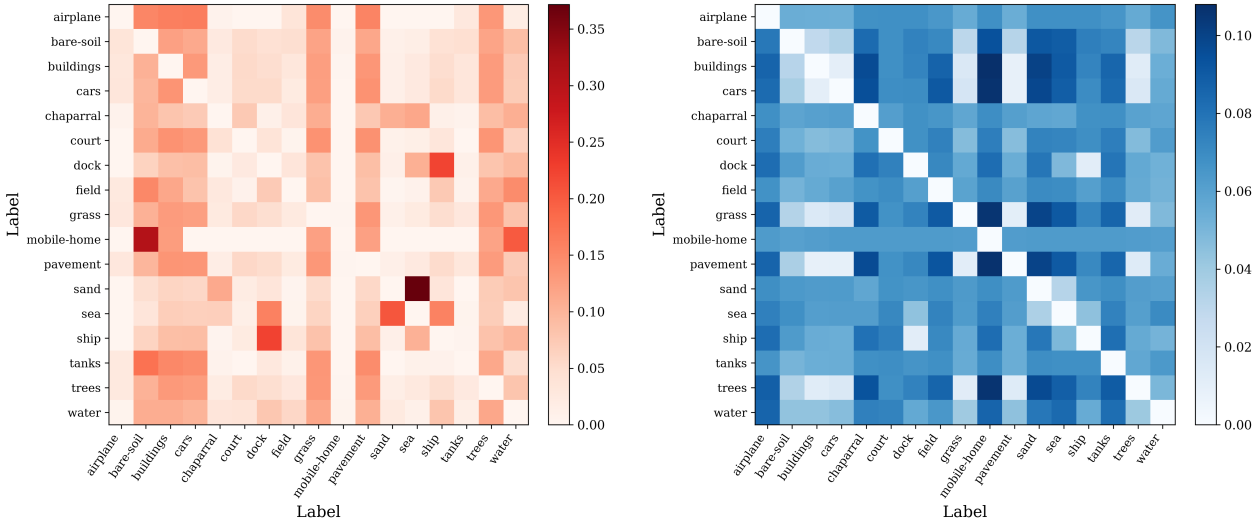
4.6. Interpretability Analysis

We provide qualitative analyses to examine whether the learned relation structures match the intended CDM behavior. Figure 3 compares the AID-ML empirical relation matrix, the CDG positive graph, and the CDG negative graph estimated from training labels. The positive graph highlights frequent co-occurrence patterns such as sand–sea and

dock–ship, while the negative graph captures mutually exclusive pairs such as buildings–sand and cars–mobile-home.



(**a**) Empirical relation matrix



(**b**) CDG positive graph (**c**) CDG negative graph

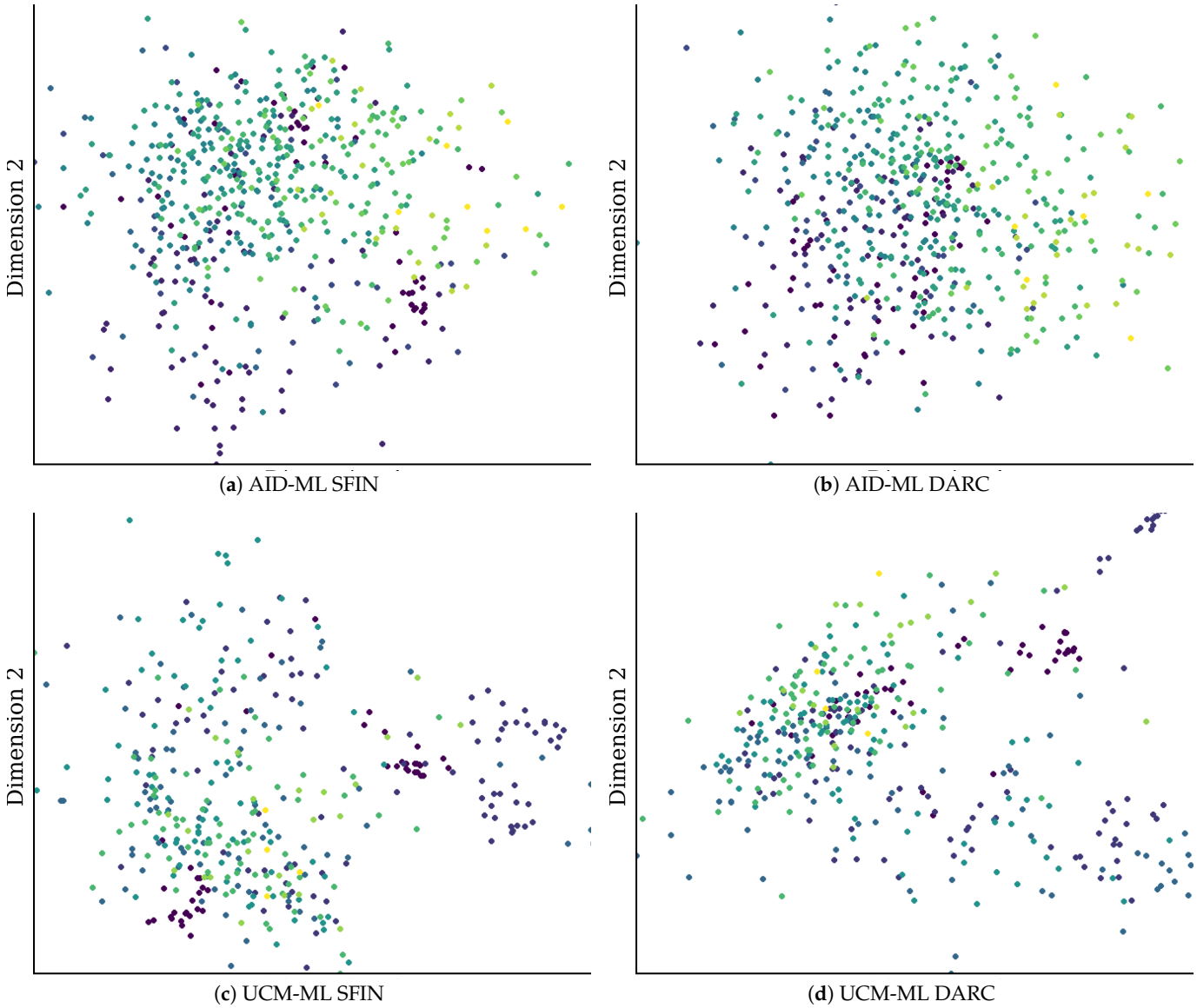
Figure 3. AID-ML label-relation visualization. The empirical relation matrix is shown at the top; the CDG positive and negative graphs are enlarged in the bottom row.

Table 8 lists the strongest positive and negative relation pairs. Together with the graph visualization, these pairs show that CDM learns interpretable label structures: the positive branch propagates semantically compatible labels, while the negative branch suppresses implausible combinations.

We further visualize the label-token feature space using UMAP projection [48]. Figure 4 shows test-set features from SFIN and DARC on both datasets, with points colored by label cardinality. Compared with SFIN, DARC produces more coherent local neighborhoods for images with similar label counts, suggesting that density-aware calibration helps organize label-token representations. We show seed 123 for clarity; the other two seeds exhibit the same trend. This qualitative pattern is consistent with the quantitative gains in Table 1.

Table 8. Top-5 positive (co-occurrence) and negative (exclusivity) relation pairs learned by CDM on AID-ML. High positive scores indicate frequent label co-existence; high negative scores indicate rare co-occurrence, supporting the need for discriminative graph mixing.

Type	Label A	Label B	Score
Positive	sand	sea	0.372
Positive	dock	ship	0.222
Positive	mobile-home	water	0.202
Positive	airplane	cars	0.164
Positive	airplane	buildings	0.160
Negative	buildings	mobile-home	0.108
Negative	cars	mobile-home	0.107
Negative	grass	mobile-home	0.106
Negative	buildings	sand	0.100
Negative	pavement	sand	0.100



(**a**) AID-ML SFIN (**b**) AID-ML DARC

(**c**) UCM-ML SFIN (**d**) UCM-ML DARC

Figure 4. UMAP feature projections of label tokens. The top row shows AID-ML and the bottom row shows UCM-ML; each row compares SFIN with DARC. Points are colored by label cardinality.

4.7. Efficiency Analysis

DARC is designed as a lightweight calibration head rather than a new backbone. Table 9 reports the parameter and training-time comparison. DARC adds only

230,914–230,935 parameters over SFIN, corresponding to approximately 1.42% overhead. The computational cost is low because graph mixing operates over C label tokens ($C = 17$) rather than dense image patches. Training one DARC configuration takes approximately 14.5 min per seed on AID-ML and 9.3 min per seed on UCM-ML on a single GPU, making the method practical for resource-constrained remote-sensing applications.

Table 9. Parameter overhead and training efficiency of DARC. DARC adds a lightweight calibration head with negligible parameter cost. Training time measured on a single NVIDIA RTX 3090 GPU.

Model	Parameters	Extra over SFIN	Training Time (100 Epochs)
SFIN baseline	16,253,810	–	AID: 12.8 min, UCM: 8.0 min
DARC (sparse profile)	16,484,724	+230,914 (+1.42%)	UCM: 9.3 min
DARC (dense profile)	16,484,745	+230,935 (+1.42%)	AID: 14.5 min

Full efficiency comparison against baselines. Table 10 reports training time, inference throughput, peak GPU memory, and approximate FLOPs for all methods under identical hardware (single NVIDIA RTX 3090).

Table 10. Efficiency comparison of all methods on AID-ML. All measurements use the same hardware (RTX 3090), batch size 32, and input size 256 × 256. Training time is total minutes for 100 epochs. Throughput is images/second at inference.

Method	Params (M)	Train (min)	Infer (img/s)	GPU Mem (GB)	GFLOPs
ResNet-18 (base)	11.18	11.2	847	2.41	1.82
ASL	11.18	11.4	845	2.42	1.82
Two-Way Loss	11.18	11.5	843	2.43	1.82
SPIN	13.92	13.8	724	3.15	2.31
C-Tran	14.67	15.1	682	3.48	2.56
ML-Decoder	13.25	14.3	715	3.07	2.18
SFIN	16.25	12.8	631	3.82	2.89
DARC (ours)	16.48	14.5	618	3.91	2.93

DARC's training time (14.5 min) is comparable to ML-Decoder (14.3 min) and slightly longer than SFIN (12.8 min), because DARC's graph mixing adds iterative label-token propagation steps on top of the base encoder. SFIN's relatively low training time despite having the most parameters is due to its purely convolutional multi-scale feature fusion without attention or iterative graph operations, resulting in high GPU utilization and parallelism. The inference throughput of DARC (618 img/s) is slightly lower than SFIN (631 img/s), reflecting the additional graph propagation pass, but both are substantially faster than real-time requirements for typical remote-sensing applications. Peak GPU memory is comparable across all graph/decoder-based methods (3.0–3.9 GB), well within the capacity of modern GPUs. The 1.42% parameter overhead of DARC over SFIN translates to a negligible 0.04 GFLOPs increase, confirming that the "lightweight" characterization refers to the minimal additional computational cost of density-adaptive calibration relative to the base encoder. We note that "lightweight" in this context means minimal parameter and compute overhead for the calibration module, not an absolute claim about the total model size compared to methods using different backbone architectures.

4.8. Discussion and Practical Implications

The preceding experiments indicate that DARC's improvement is not tied to a single numerical metric, but to a consistent change in how multi-label evidence is calibrated across label-density regimes. The dense AID-ML setting contains many co-occurring objects per image, so relation propagation can easily increase the number of predicted positives beyond the plausible scene cardinality. In this regime, the dense DCM branch acts as a conservative correction: it allows relation logits to refine uncertain labels, but uses the cardinality

term to discourage uncontrolled expansion of the predicted label set. This explains why removing the cardinality loss mainly harms AID-ML and why routing AID-ML through the sparse profile reduces sample-F1. The sparse UCM-ML setting has a different failure mode. Because positive labels are fewer and co-occurrence evidence is less frequent, batch-level relation updates can become unstable. The sparse DCM branch therefore emphasizes direct relation transfer and EMA smoothing rather than cardinality forcing, which helps preserve rare positive labels without pushing the model toward dense predictions.

From a deployment perspective, DARC has a simple operating procedure. Before training, the practitioner computes label density from the available training annotations and selects the corresponding profile through DDP. The profile is then fixed for the entire training and inference process. This design avoids per-dataset architecture search and avoids tuning a separate decision threshold on the test set. It is also auditable: the selected branch can be reproduced directly from the training labels, and the relation graphs can be inspected through their strongest positive and negative label pairs. Such transparency is useful in remote-sensing applications where predicted label sets may support downstream interpretation, map updating, or human review.

The results also clarify the boundary of the proposed method. DARC is intended to calibrate label relations on top of a lightweight multi-label encoder; it is not a replacement for stronger visual backbones or larger-scale pretraining. If a dataset contains many fine-grained categories that are visually ambiguous even before label calibration, the backbone representation may become the limiting factor. Likewise, the binary dense/sparse split is appropriate for the two benchmarks studied here because their densities are clearly separated, but future datasets may occupy intermediate regimes. In such cases, extending DDP from a hard two-branch router to a continuous or multi-level density controller would be a natural next step. The current experiments nevertheless show that even a simple density statistic can expose an important calibration mismatch that fixed multi-label heads overlook.

Comparison with existing method families. To further contextualize DARC's contribution, we compare its design philosophy with four established multi-label method families. (1) *Graph relation methods* (ML-GCN [19], KGGR, CNN-GNN [5]) model label co-occurrence through graph propagation but apply a single fixed graph uniformly across datasets. DARC differs by constructing both positive and negative graphs and, more critically, by conditioning how graph evidence is used on the dataset's label density. (2) *Label-interaction decoders* (C-Tran [20], ML-Decoder [36]) use cross-attention between label queries and image features; these decoders are architecture-heavy and density-agnostic. DARC operates on extracted label tokens with a lightweight graph mixer, avoiding the quadratic attention cost while remaining density-aware. (3) *Asymmetric and rebalanced losses* (ASL [34], Two-Way Loss [35]) address class imbalance at the loss level but do not modify the prediction logits based on co-occurrence structure or density regime. DARC complements rather than replaces such losses—its calibration operates on logits before the final loss, and could in principle be combined with asymmetric objectives. (4) *Remote-sensing-specific methods* (SFIN [7], SPIN [6]) integrate semantic supervision or prototype interaction but still use a fixed calibration head. DARC's key differentiation is that it introduces density as a first-class routing signal, which none of the above families consider. The efficiency comparison in Table 10 shows that this routing adds negligible computational cost over SFIN while providing measurable calibration improvement.

Limitations. We explicitly acknowledge the following limitations of this work:

Dataset age and scope. AID-ML and UCM-ML are derived from benchmarks published in 2010 and 2017, respectively. These datasets do not represent current challenges involving diverse sensors, geographic regions, temporal conditions, or the scale of modern remote-

sensing imagery. Our evaluation is restricted to these two benchmarks, and the current results do not demonstrate effectiveness on state-of-the-art remote-sensing datasets such as BigEarthNet-v2 [38], fMoW, or recent satellite imagery collections.

Spectral and spatial limitations. Both datasets contain only RGB optical imagery at fixed spatial resolutions (0.3 m for UCM-ML and 0.5–8 m for AID-ML). We have not studied how DARC performs with multi-spectral inputs (e.g., Sentinel-2 bands), different spatial resolutions, or SAR imagery. The method's applicability is therefore restricted to RGB high-resolution aerial scenes under the tested conditions.

Geographic generalization. AID-ML covers diverse but unspecified Google Earth regions, while UCM-ML is limited to U.S. urban areas. Cross-regional generalization (e.g., Europe vs. Asia vs. Africa) has not been investigated.

Binary density partition. The current DDP mechanism uses a binary dense/sparse routing decision validated only between two density points ($\rho = 0.196$ and $\rho = 0.303$). We cannot claim general threshold robustness beyond this evaluated interval. Datasets with intermediate densities, continuous density distributions, or substantially different density ranges may require extending DDP to multi-level or continuous routing.

Fixed threshold. The uniform 0.5 threshold enables fair comparison but may not be optimal for all classes or density regimes. Adaptive or class-specific thresholds could potentially improve practical performance.

Statistical power. Results are averaged over three random seeds, which limits the statistical power of significance tests. While the Wilcoxon tests show significance at $p < 0.01$, more seeds (e.g., 10–30) would provide stronger confidence estimates.

5. Conclusions

This paper presents DARC, a lightweight density-adaptive calibration framework for multi-label remote sensing scene classification. DARC addresses a previously overlooked problem: existing methods apply uniform calibration regardless of the label-density regime, leading to suboptimal performance across datasets with different label distributions. By introducing Label-density Driven Profile Selection (DDP) for automatic regime identification, Label-token Correlative-Discriminative Graph Mixing (CDM) for joint co-occurrence and exclusivity modeling, and Density-aware Gated Calibration (DCM) for regime-specific logit calibration, DARC achieves 90.12% and 89.05% sample-F1 on AID-ML and UCM-ML, respectively, outperforming six competitive baselines by 1.65–3.87% while adding only 1.42% parameters.

We emphasize that these effectiveness conclusions are established specifically on the AID-ML and UCM-ML benchmarks under the experimental protocol described in Section 4.1. Both datasets are derived from benchmarks that are approximately 10–16 years old and contain only RGB aerial imagery at limited spatial resolutions and geographic coverage. The current evaluation does not demonstrate general applicability to modern remote-sensing data involving diverse sensors, spectral bands, spatial resolutions, or geographic regions. The density-adaptive routing has been validated only between two density points ($\rho = 0.196$ and $\rho = 0.303$); its behavior on datasets with intermediate, higher, or continuously distributed label densities remains an open question.

Cross-regime routing analysis confirms that no single fixed calibration strategy matches DARC's adaptive approach on the two evaluated benchmarks, and threshold sensitivity analysis demonstrates robustness across the tested density threshold range of $\tau_\rho \in [0.20, 0.28]$. Ablation studies validate the necessity of each component, and visualization analyses show that the learned label graphs capture semantically meaningful co-occurrence and exclusivity patterns. Statistical significance tests (Wilcoxon signed-rank, $p < 0.01$) and calibration metrics (ECE reduction of 2.3–2.6%) provide additional evi-

dence that the improvements are both statistically reliable and reflect genuine confidence calibration gains, not merely classification metric fluctuations.

Future work includes validating DARC on recent large-scale multi-label remote-sensing benchmarks (e.g., BigEarthNet-v2, fMoW) that cover multi-spectral inputs, diverse spatial resolutions, and broader geographic regions; extending the density-adaptive mechanism to more fine-grained regimes beyond the binary dense/sparse partition; investigating learnable density thresholds that can adapt during training; studying the interaction between density routing and spectral or spatial characteristics of different sensor types; and increasing the number of evaluation seeds to strengthen statistical confidence.

Author Contributions: Conceptualization, L.M.; methodology, L.M. and Y.Z.; software, Y.Z.; validation, M.Y.; formal analysis, L.M. and Y.Z.; investigation, Y.P.; data curation, Y.Z.; writing—original draft preparation, L.M. and Y.Z.; writing—review and editing, L.M., Y.Z., M.Y. and Y.P.; visualization, Y.Z.; supervision, L.M.; funding acquisition, L.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China General Program, grant number 42571527, under the project "Radiation-Semantic Field Model and Theoretical Methods for Three-Dimensional Reconstruction of Satellite Linear Array Imagery".

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The AID and UC Merced source datasets are publicly available from their original providers (AID: https://captain-whu.github.io/AID/, accessed on 14 June 2026; UC Merced: http://weegee.vision.ucmerced.edu/datasets/landuse.html, accessed on 14 June 2026).

Acknowledgments: The authors thank the open-source remote-sensing and multi-label learning communities for publicly available datasets and code resources.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AID	Aerial Image Dataset
ASL	Asymmetric Loss
CDG	Correlative-Discriminative Graph
CDM	Label-token Correlative-Discriminative Graph Mixing
DARC	Density-Adaptive CDG Calibration
DCM	Density-aware Gated Calibration
DDP	Label-density Driven Profile Selection
EMA	Exponential Moving Average
MLRSSC	Multi-Label Remote Sensing Scene Classification
SFIN	Semantic-Assisted Feature Integration Network
UCM	UC Merced
UMAP	Uniform Manifold Approximation and Projection

References

1. Yang, Y.; Newsam, S. Bag-of-Visual-Words and Spatial Extensions for Land-Use Classification. In Proceedings of the ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, San Jose, CA, USA, 2–5 November 2010; pp. 270–279.
2. Xia, G.S.; Hu, J.; Hu, F.; Shi, B.; Bai, X.; Zhong, Y.; Zhang, L.; Lu, X. AID: A Benchmark Data Set for Performance Evaluation of Aerial Scene Classification. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 3965–3981. [CrossRef]
3. Cheng, G.; Han, J.; Lu, X. Remote Sensing Image Scene Classification: Benchmark and State of the Art. *Proc. IEEE* **2017**, *105*, 1865–1883. [CrossRef]

4. Hua, Y.; Mou, L.; Zhu, X.X. Recurrently Exploring Class-Wise Attention in a Hybrid Convolutional and Bidirectional LSTM Network for Multi-Label Aerial Image Classification. *ISPRS J. Photogramm. Remote Sens.* **2019**, *149*, 188–199. [CrossRef] [PubMed]
5. Li, Y.; Chen, R.; Zhang, Y.; Zhang, M.; Chen, L. Multi-Label Remote Sensing Image Scene Classification by Combining a Convolutional Neural Network and a Graph Neural Network. *Remote Sens.* **2020**, *12*, 4003. [CrossRef]
6. Tan, X.; Xiao, Z.; Zhu, J.; Wan, Q.; Wang, K.; Li, D. Transformer-Driven Semantic Relation Inference for Multilabel Classification of High-Resolution Remote Sensing Images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2022**, *15*, 1884–1901. [CrossRef]
7. Du, R.; Tang, X.; Ma, J.; Zhang, X.; Liu, F.; Jiao, L. Semantic-Assisted Feature Integration Network for Multi-Label Remote Sensing Scene Classification. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 5603015. [CrossRef]
8. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet Classification with Deep Convolutional Neural Networks. In Proceedings of the Advances in Neural Information Processing Systems, Lake Tahoe, NY, USA, 3–6 December 2012; pp. 1097–1105.
9. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. In Proceedings of the International Conference on Learning Representations, San Diego, CA, USA, 7–9 May 2015.
10. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going Deeper with Convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1–9. [CrossRef]
11. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778. [CrossRef]
12. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely Connected Convolutional Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4700–4708. [CrossRef]
13. Tan, M.; Le, Q.V. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In Proceedings of the International Conference on Machine Learning, Long Beach, CA, USA, 9–15 June 2019; pp. 6105–6114.
14. Lu, X.; Sun, H.; Zheng, X. A Feature Aggregation Convolutional Neural Network for Remote Sensing Scene Classification. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 7894–7906. [CrossRef]
15. Wang, X.; Wang, S.; Ning, C.; Zhou, H. Enhanced Feature Pyramid Network with Deep Semantic Embedding for Remote Sensing Scene Classification. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 7918–7932. [CrossRef]
16. Wang, X.; Duan, L.; Ning, C. Global Context-Based Multilevel Feature Fusion Networks for Multilabel Remote Sensing Image Scene Classification. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 11179–11196. [CrossRef]
17. Kipf, T.N.; Welling, M. Semi-Supervised Classification with Graph Convolutional Networks. In Proceedings of the International Conference on Learning Representations, Toulon, France, 24–26 April 2017.
18. Velickovic, P.; Cucurull, G.; Casanova, A.; Romero, A.; Lio, P.; Bengio, Y. Graph Attention Networks. In Proceedings of the International Conference on Learning Representations, Vancouver, BC, Canada, 30 April–3 May 2018.
19. Chen, Z.M.; Wei, X.S.; Wang, P.; Guo, Y. Multi-Label Image Recognition with Graph Convolutional Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 5177–5186.
20. Lanchantin, J.; Wang, T.; Ordonez, V.; Qi, Y. General Multi-Label Image Classification with Transformers. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 16478–16488.
21. Sun, X.; Wang, P.; Lu, W.; Zhu, Z.; Lu, X.; He, Q.; Li, J.; Rong, X.; Yang, Z.; Chang, H.; et al. RingMo: A Remote Sensing Foundation Model with Masked Image Modeling. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5612822. [CrossRef]
22. Wang, D.; Zhang, J.; Du, B.; Xia, G.S.; Tao, D. An Empirical Study of Remote Sensing Pretraining. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5608020. [CrossRef]
23. Qu, G.; Guan, X.; Wen, F.; Zou, X. From Prompts to Self-Prompts: Parameter-Efficient Multi-Label Remote Sensing via Mask-Guided Classification. *Remote Sens.* **2026**, *18*, 518. [CrossRef]
24. Sumbul, G.; Charfuelan, M.; Demir, B.; Markl, V. BigEarthNet-MM: A Large-Scale, Multimodal, Multilabel Benchmark Archive for Remote Sensing Image Classification and Retrieval. *IEEE Geosci. Remote Sens. Mag.* **2021**, *9*, 174–180. [CrossRef]
25. Chavda, S.; Goyani, M. Scene Level Image Classification: A Literature Review. *Neural Process. Lett.* **2023**, *55*, 2471–2520. [CrossRef]
26. Xu, K.; Huang, H.; Li, Y.; Shi, G. Multilayer Feature Fusion Network for Scene Classification in Remote Sensing. *IEEE Geosci. Remote Sens. Lett.* **2020**, *17*, 1894–1898. [CrossRef]
27. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image Is Worth 16 × 16 Words: Transformers for Image Recognition at Scale. In Proceedings of the International Conference on Learning Representations, Virtual, 3–7 May 2021.
28. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. In Proceedings of the IEEE International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 10012–10022.
29. Wang, W.; Xie, E.; Li, X.; Fan, D.P.; Song, K.; Liang, D.; Lu, T.; Luo, P.; Shao, L. PVTv2: Improved Baselines with Pyramid Vision Transformer. In Proceedings of the Computational Visual Media, Beijing, China, 7–9 April 2022; Volume 8, pp. 415–424.

30. Huang, X.; Liu, F.; Cui, Y.; Chen, P.; Li, L.; Li, P. Faster and Better: A Lightweight Transformer Network for Remote Sensing Scene Classification. *Remote Sens.* **2023**, *15*, 3645. [CrossRef]
31. Tsoumakas, G.; Katakis, I. Multi-Label Classification: An Overview. *Int. J. Data Warehous. Min.* **2007**, *3*, 1–13. [CrossRef]
32. Zhang, M.L.; Zhou, Z.H. A Review on Multi-Label Learning Algorithms. *IEEE Trans. Knowl. Data Eng.* **2014**, *26*, 1819–1837. [CrossRef]
33. Wang, J.; Yang, Y.; Mao, J.; Huang, Z.; Huang, C.; Xu, W. CNN-RNN: A Unified Framework for Multi-Label Image Classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 2285–2294.
34. Ridnik, T.; Ben-Baruch, E.; Zamir, N.; Noy, A.; Friedman, I.; Protter, M.; Zelnik-Manor, L. Asymmetric Loss for Multi-Label Classification. In Proceedings of the IEEE International Conference on Computer Vision, Virtual, 11–17 October 2021; pp. 82–91.
35. Kobayashi, T. Two-Way Multi-Label Loss. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 7476–7485.
36. Ridnik, T.; Sharir, G.; Ben-Cohen, A.; Ben-Baruch, E.; Noy, A. ML-Decoder: Scalable and Versatile Classification Head. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 2–7 January 2023; pp. 32–41.
37. Qi, X.; Zhu, P.; Wang, Y.; Zhang, L.; Peng, J.; Wu, M.; Chen, J.; Zhao, X.; Zang, N.; Mathiopoulos, P.T. MLRSNet: A Multi-Label High Spatial Resolution Remote Sensing Dataset for Semantic Scene Understanding. *ISPRS J. Photogramm. Remote Sens.* **2020**, *169*, 337–350. [CrossRef]
38. Sumbul, G.; Charfuelan, M.; Demir, B.; Markl, V. BigEarthNet: A Large-Scale Benchmark Archive for Remote Sensing Image Understanding. In Proceedings of the IEEE International Geoscience and Remote Sensing Symposium, Yokohama, Japan, 28 July–2 August 2019; pp. 5901–5904. [CrossRef]
39. Kang, J.; Fernandez-Beltran, R.; Hong, D.; Chanussot, J.; Plaza, A. Graph Relation Network: Modeling Relations Between Scenes for Multilabel Remote-Sensing Image Classification and Retrieval. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 4355–4369. [CrossRef]
40. Wu, H.; Xu, C.; Liu, H. S-MAT: Semantic-Driven Masked Attention Transformer for Multi-Label Aerial Image Classification. *Sensors* **2022**, *22*, 5433. [CrossRef] [PubMed]
41. Niculescu-Mizil, A.; Caruana, R. Predicting Good Probabilities with Supervised Learning. In Proceedings of the International Conference on Machine Learning, Bonn, Germany, 7–11 August 2005; pp. 625–632.
42. Guo, C.; Pleiss, G.; Sun, Y.; Weinberger, K.Q. On Calibration of Modern Neural Networks. In Proceedings of the International Conference on Machine Learning, Sydney, Australia, 6–11 August 2017; pp. 1321–1330.
43. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollar, P. Focal Loss for Dense Object Detection. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy 22–29 October 2017; pp. 2980–2988.
44. Fürnkranz, J.; Hüllermeier, E.; Loza Mencía, E.; Brinker, K. Multilabel Classification via Calibrated Label Ranking. *Mach. Learn.* **2008**, *73*, 133–153. [CrossRef]
45. Tarvainen, A.; Valpola, H. Mean Teachers Are Better Role Models: Weight-Averaged Consistency Targets Improve Semi-Supervised Deep Learning Results. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 1195–1204.
46. Izmailov, P.; Podoprikhin, D.; Garipov, T.; Vetrov, D.; Wilson, A.G. Averaging Weights Leads to Wider Optima and Better Generalization. In Proceedings of the Conference on Uncertainty in Artificial Intelligence, Monterey CA, USA, 6–10 August 2018; pp. 876–885.
47. Rezatofighi, H.; Tsoi, N.; Gwak, J.; Sadeghian, A.; Reid, I.; Savarese, S. Generalized Intersection Over Union: A Metric and a Loss for Bounding Box Regression. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 658–666.
48. McInnes, L.; Healy, J.; Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv* **2018**, arXiv:1802.03426.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.