RC-MAMBA: A GROUP-CONDITIONAL BAYESIAN RISK-CONTROLLED MAMBA FOR MULTI-SOURCE REMOTE SENSING CLASSIFICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Multi-source remote sensing classification integrates hyperspectral, SAR, and LiDAR observations for accurate land-cover recognition. Recent state-space models (SSMs) with multi-scale Mamba fusion have achieved competitive accuracy, yet they remain purely discriminative and lack distribution-free reliability under spectral-spatial non-stationarity, cross-modal heterogeneity, and class imbalance. To address this problem, we propose RC-Mamba, a reliability-calibrated Mamba framework for multi-source remote sensing classification. RC-Mamba couples a multi-scale Mamba backbone with three reliability modules: Group-Conditional Bayesian Risk Thresholding (G-BRT), which learns group-aware risk thresholds beyond a single global threshold; Threshold-Coupled Margin Regularization (TCMR), which injects these thresholds into discriminative training to enlarge confidence margins of critical samples; and Risk-Aware Selective Refinement (RASR), which refines only calibrated high-risk samples at test time. Experiments on Berlin, Augsburg, Houston2013, and Houston2018 show that RC-Mamba improves the reproduced MSFMamba baseline by 0.08–0.84 OA points over three seeds while exposing a group-aware reliability interface for independent-reference evaluation. The audited reliability diagnostic further shows why marginal coverage alone is insufficient and motivates explicit worst-group evaluation. These results establish RC-Mamba as a risk-aware SSM framework for multi-source remote sensing imagery.

1 INTRODUCTION

Multi-source remote sensing classification jointly exploits complementary modalities such as hyperspectral images (HSI), synthetic aperture radar (SAR), and light detection and ranging (LiDAR) for pixel-wise land-cover recognition (Hong et al., 2020; Gao et al., 2025a). These heterogeneous sensors provide complementary spectral, structural, and sensing information, enabling more discriminative representations than single-modality observations (Hong et al., 2020; Roy et al., 2023; Gao et al., 2025a). Recent state-space models (SSMs), particularly Mamba (Gu & Dao, 2023; Dao & Gu, 2024), enable efficient long-range sequence modeling with linear scaling in sequence length. Subsequent studies have extended SSMs to generic visual representation, hyperspectral classification, and multi-source HSI–LiDAR/SAR fusion (Zhu et al., 2024; Yao et al., 2024; Liao et al., 2024; Gao et al., 2025a). However, high predictive accuracy alone does not provide reliability guarantees. In risk-sensitive deployment, predictive uncertainty or decision risk should also be quantified and controlled under explicit statistical assumptions (Angelopoulos & Bates, 2023; Bates et al., 2021; Angelopoulos et al., 2024; Snell et al., 2022).

Driven by the progress of multi-source representation learning, a question arises: is learning a more accurate Mamba predictor sufficient for reliable multi-source remote sensing classification? Existing studies have advanced along two largely separate directions. On the representation side, spatial–spectral convolution (Audebert et al., 2018), cross-modal Transformer and attention mechanisms, HSI–LiDAR Mamba, and MSFMamba improve heterogeneous feature extraction and fusion (Audebert et al., 2018; Roy et al., 2023; Wang et al., 2024; Liao et al., 2024; Gao et al., 2025a). On the reliability side, conformal prediction and its extensions provide principled uncertainty quantification under exchangeability or distribution shifts (Angelopoulos & Bates, 2023; Barber et al., 2023; Gibbs

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

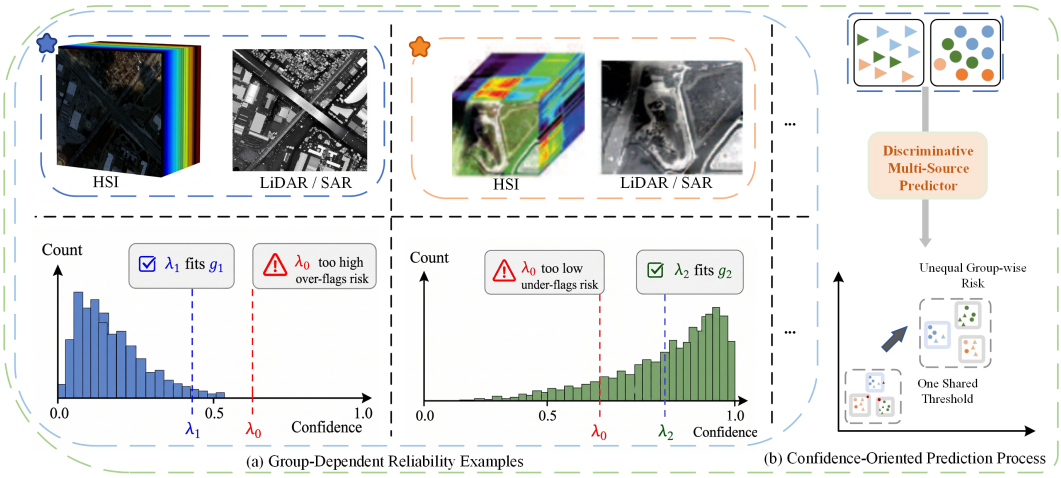


Figure 1: Group-dependent reliability in multi-source remote sensing classification. (a) Different spatial or sensing groups exhibit distinct confidence distributions, so a shared threshold can induce unequal risks. (b) Global confidence can mask threshold violations in difficult groups. Each sample denotes co-registered HSI and SAR/LiDAR patches centered at one pixel.

& Candès, 2024), while Conformal Prediction as Bayesian Quadrature (CPBQ) models uncertainty in quantile spacings and constructs a Bayesian characterization of deployment-time expected loss (Snell & Griffiths, 2025).

On the reliability side, standard conformal prediction provides finite-sample marginal coverage under exchangeability (Angelopoulos & Bates, 2023; Romano et al., 2020), while recent extensions address departures from exchangeability and distribution shifts (Barber et al., 2023; Gibbs & Candès, 2024), as well as conditional reliability across prespecified subgroups (Gibbs et al., 2025). Meanwhile, Conformal Prediction as Bayesian Quadrature (CPBQ) provides a Bayesian characterization of deployment-time loss uncertainty (Snell & Griffiths, 2025). Nevertheless, standard split conformal prediction, Conformal Risk Control, and Risk-Controlling Prediction Sets primarily calibrate a pretrained or fixed predictor (Angelopoulos & Bates, 2023; Angelopoulos et al., 2024; Bates et al., 2021). They are not designed to feed group-specific threshold violations back into the representation learning of a multi-source SSM predictor.

When confidence distributions vary across spatial or sensing groups, a globally shared threshold can induce group-dependent reliability differences. Confidence distributions appropriate for homogeneous vegetation or water regions may be unsuitable for urban shadows, mixed pixels, class boundaries, or sensor-degraded regions. As illustrated in Fig. 1(a), the same threshold may be overly restrictive for lower-confidence groups yet overly permissive for differently distributed groups, even when aggregate classification accuracy remains competitive. Post-hoc threshold selection can identify or calibrate such violations, but does not by itself modify the representations that produce them. Moreover, common uncertainty-estimation approaches such as MC-Dropout (Gal & Ghahramani, 2016), deep ensembles (Lakshminarayanan et al., 2017), and evidential learning (Sensoy et al., 2018) rely on repeated stochastic inference, multiple predictors, or evidence-parameterized predictive outputs, respectively. These approaches are not specifically designed to connect group-specific risk measurement with risk-aware representation learning and residual-risk handling.

This reliability degradation indicates that not all spatial regions or sensing domains should share an identical risk standard. Let us consider two observable groups processed by the same multi-source predictor but exhibiting different confidence distributions. An appropriate solution should first construct a risk reference for each group, then convert the deficits relative to these references into predictor-improving signals, and finally selectively update predictions that remain below their corresponding risk standards. The existence of such a failure-driven solution suggests that reliable multi-source classification requires more than either a stronger Mamba backbone or a detached post-hoc risk estimator: the statistical risk standard must participate in discriminative learning, while the remaining risk should be handled without repeatedly refining every already-confident sample.

As shown in Fig. 1(b), confidence-oriented predictors mainly optimize aggregate discrimination, allowing group-specific threshold violations to remain hidden by globally confident predictions.

To bridge these disconnected capabilities, we propose **RC-Mamba**, a group-conditional Bayesian risk-learning framework for reliable multi-source remote sensing classification. First, Group-Conditional Bayesian Risk Thresholding (G-BRT) adapts the CPBQ-inspired risk upper-bound construction to observable spatial tiles or sensing domains, producing a group-risk threshold bank Λ instead of one globally shared threshold. Second, because these non-parametric thresholds only measure risk violations, Threshold-Coupled Margin Regularization (TCMR) converts the true-class probability deficit relative to each group threshold into a differentiable powered-hinge objective, thereby coupling statistical risk standards with Mamba representation learning. Third, since threshold-aware learning may still leave ambiguous boundary, shadow, mixed, or sensor-corrupted pixels, Risk-Aware Selective Refinement (RASR) reuses the learned threshold bank to expose the residual high-risk subset and selectively updates its predictions by averaging the original, horizontally flipped, and vertically flipped views. Thus, RC-Mamba forms a problem-driven chain from risk measurement, through risk-aware learning, to residual-risk handling.

Our contributions are summarized as follows:

- We introduce **G-BRT**, which extends CPBQ-inspired Bayesian risk thresholding to observable spatial or sensing-domain groups, providing group-adaptive risk standards for heterogeneous multi-source remote sensing predictions.
- We propose **TCMR**, which transforms non-parametric group-threshold deficits into differentiable training feedback and directly couples group-wise risk measurement with discriminative Mamba optimization.
- We develop **RASR**, which reuses the learned threshold bank to expose residual high-risk predictions and selectively updates them through flip-based test-time augmentation, completing the risk-measurement–learning–handling chain.

2 Related Work

2.1 Multi-Source Remote Sensing Classification and State-Space Backbones

Multi-source remote sensing classification exploits complementary HSI and LiDAR/SAR information through increasingly effective cross-modal representation and fusion mechanisms (Hong et al., 2020; Roy et al., 2023; Xu et al., 2017; Mohla et al., 2020; Gao et al., 2025a), including frequency-domain fusion (Song et al., 2025), prototype-based compensation (Gao et al., 2025b), spatial–spectral enhancement (Fang et al., 2021), asymmetric fusion (Li et al., 2022), CNN–Transformer modeling (Zhao et al., 2022), cross-modality attention (Yao et al., 2023), multiscale cross-modal attention (Wang et al., 2024), and detail-enhanced adaptive fusion (Liang et al., 2025). More recently, Mamba and Mamba-2 provide efficient selective state-space modeling and structured state-space duality (Gu & Dao, 2023; Dao & Gu, 2024), with subsequent extensions to visual, multidimensional, and hyperspectral modeling (Zhu et al., 2024; Li et al., 2024; Liu et al., 2024; Yao et al., 2024); HLMamba (Liao et al., 2024) and MSFMamba (Gao et al., 2025a) further adapt Mamba to multi-source remote sensing. These models primarily emphasize representation, fusion, and classification accuracy rather than explicit group-specific risk control (Liao et al., 2024; Gao et al., 2025a); RC-Mamba therefore retains MSFMamba as the discriminative predictor and focuses on group-aware risk measurement, learning, and residual-risk handling.

2.2 Group-Aware Risk Control and Reliability-Aware Learning

Conformal prediction provides finite-sample marginal coverage (Angelopoulos & Bates, 2023), while adaptive prediction sets and risk-control extensions broaden its applicability to classification and general risk objectives (Romano et al., 2020; Angelopoulos et al., 2024; Bates et al., 2021; Snell et al., 2022); related studies further consider distribution shift (Tibshirani et al., 2019; Barber et al., 2023; Gibbs & Candès, 2024) and conditional reliability (Gibbs et al., 2025), with broader developments summarized in (Angelopoulos & Bates, 2023). Particularly, CPBQ models random quantile spacings with a Dirichlet distribution to obtain a Bayesian upper-risk characterization (Snell & Griffiths, 2025); however, marginal validity does not generally ensure reliability across

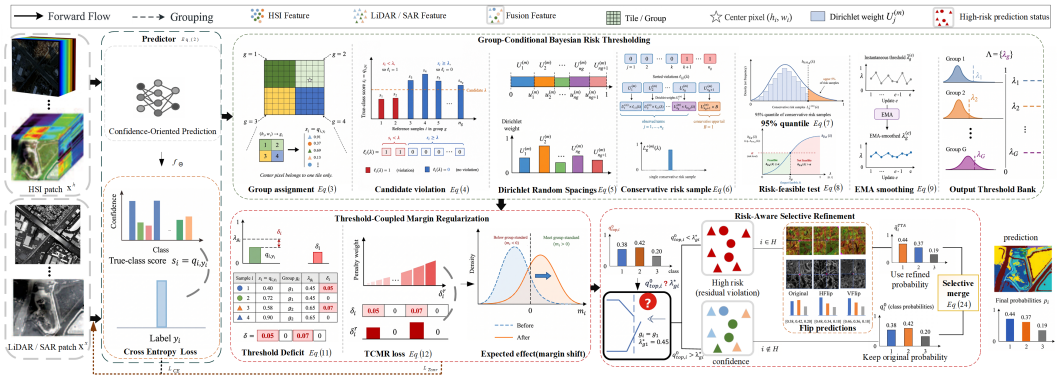


Figure 2: Overall framework of RC-Mamba. HSI and auxiliary LiDAR/SAR patches are encoded by the MSFMamba backbone to obtain fused features, logits $\mathbf{z}$, and probabilities $\mathbf{q}$. **(1)** G-BRT estimates group-specific risk thresholds from predicted probabilities and observable group assignments, forming the threshold bank Λ (Sec. 3.3). **(2)** TCMR converts threshold deficits into differentiable margin regularization for threshold-aware representation learning (Sec. 3.4). **(3)** RASR identifies residual high-risk samples $\mathcal{H} = \{i : \max_c q^0_{i,c} < \lambda^*_{g_i}\}$ and selectively refines only their predictions using original and flipped views (Sec. 3.5).

heterogeneous groups (Gibbs et al., 2025), and conventional post-hoc conformal methods typically calibrate fixed predictors (Angelopoulos & Bates, 2023; Angelopoulos et al., 2024; Bates et al., 2021). Although training-aware conformal methods (Stutz et al., 2021; Einbinder et al., 2022), probability calibration (Guo et al., 2017), MC-Dropout (Gal & Ghahramani, 2016), deep ensembles (Lakshminarayanan et al., 2017), evidential learning (Sensoy et al., 2018), and uncertainty-aware optimization (Lu et al., 2023) provide complementary reliability mechanisms, they do not jointly couple group-specific risk references with predictor learning and residual-risk selection for multi-source SSMs. RC-Mamba addresses this gap through G-BRT, which constructs group-specific Bayesian risk thresholds, TCMR, which converts threshold deficits into differentiable training signals, and RASR, which reuses these thresholds for selective residual-risk refinement, forming a unified measurement–learning–handling chain.

3 Proposed Method

Following standard multi-source remote sensing classification Gao et al. (2025a), each target pixel i is represented by co-registered HSI and auxiliary LiDAR/SAR patches, $\mathbf{X}^h_i \in \mathbb{R}^{1\times N_p\times H\times W}$ and $\mathbf{X}^x_i \in \mathbb{R}^{C_x\times H\times W}$, where N_p and C_x denote the HSI components and auxiliary channels, respectively. Given $\mathcal{X}_i = \{\mathbf{X}^h_i, \mathbf{X}^x_i\}$, the MSFMamba predictor f_Θ outputs logits $\mathbf{z}_i \in \mathbb{R}^C$ and probabilities $\mathbf{q}_i = \mathrm{softmax}(\mathbf{z}_i)$ to predict $y_i \in \{1,\dots,C\}$. However, a globally shared threshold cannot capture heterogeneous reliability across spatial regions and sensing domains.

3.1 Overall Reliability-Learning Pipeline

Motivated by the limitation of a globally shared threshold, RC-Mamba builds a failure-driven reliability-learning pipeline on top of the existing MSFMamba predictor. As illustrated in Fig. 2, the fused representations, logits, and base probabilities produced by MSFMamba provide a shared predictive interface for the three proposed components. The complete information flow can be summarized as

$$\begin{aligned}(\mathbf{X}^h,\mathbf{X}^x) &\xrightarrow{f_\Theta} \mathbf{z} \xrightarrow{\text{softmax}} \mathbf{q} \xrightarrow{\text{G-BRT}} \Lambda \xrightarrow{\text{TCMR}} \Theta^*,\\ (\Theta^*,\Lambda^*) &\xrightarrow{\text{residual risk detection}} \mathcal{H} \xrightarrow{\text{RASR}} \mathbf{p} \xrightarrow{\arg\max} \widehat{\mathbf{y}},\end{aligned} \tag{1}$$

where $\mathbf{q}$ denotes the base class probabilities, $\Lambda = \{\lambda_g\}_{g=1}^G$ is the group-risk threshold bank, Θ^* represents the optimized predictor parameters, $\mathcal{H}$ is the residual high-risk subset, and $\mathbf{p}$ denotes the final probabilities after selective refinement.

Specifically, G-BRT periodically estimates group-specific Bayesian risk thresholds from a labeled risk-reference cache, addressing the inability of a single threshold to characterize spatial or domain

heterogeneity. Since these non-parametric thresholds only measure reliability violations and do not provide gradients for predictor optimization, TCMR converts the corresponding threshold deficits into differentiable margin constraints. After threshold-coupled training, the optimized predictor may still produce low-confidence predictions around ambiguous boundaries, shadows, mixed pixels, or sensor-corrupted regions. RASR therefore reuses the frozen threshold bank to identify the remaining high-risk predictions and selectively updates them through flip-based test-time augmentation. The complete training and inference procedure is summarized in Alg. 1.

3.2 Group-Conditional Bayesian Risk Thresholding

A single confidence threshold assumes that all samples share an identical reliability boundary. This assumption is restrictive for multi-source remote sensing imagery, where prediction difficulty varies across spatial regions and sensing domains because of class imbalance, spectral–spatial non-stationarity, cross-modal heterogeneity, and sensor noise. To address this issue, we propose Group-Conditional Bayesian Risk Thresholding (G-BRT), which adapts CPBQ-inspired risk estimation to predefined and observable remote-sensing groups.

Given the risk-reference cache $\mathcal{D}_{\mathrm{ref}} \subset \mathcal{D}_{\mathrm{tr}}$, the current predictor first produces

$$\mathbf{z}_i = f_\Theta(\mathbf{X}_i^h, \mathbf{X}_i^x), \qquad \mathbf{q}_i = \mathrm{softmax}(\mathbf{z}_i), \qquad i \in \mathcal{D}_{\mathrm{ref}}. \tag{2}$$

For each labeled reference sample, we extract its true-class probability and observable group assignment as

$$s_i = q_{i,y_i}, \qquad g_i = \mathcal{G}(h_i, w_i, d_i) \in \{1, \dots, G\}, \tag{3}$$

where d_i denotes the sensing-domain metadata when domain-based grouping is used. In this work, $\mathcal{G}$ is instantiated using spatial tiles or sensing domains so that the group identity remains observable during inference. Let $\mathcal{I}_g = \{i : g_i = g\}$ denote the indices assigned to group g and $n_g = |\mathcal{I}_g|$ its sample size.

For a candidate confidence threshold $\lambda \in [0, 1]$, we define the individual threshold-violation loss as

$$\ell_i(\lambda) = \mathbb{I}[s_i < \lambda], \qquad 0 \le \ell_i(\lambda) \le B, \qquad B = 1. \tag{4}$$

The event $\ell_i(\lambda) = 1$ indicates that the true-class probability falls below the candidate group-wise risk standard. Let $\ell_{(1)}(\lambda) \le \cdots \le \ell_{(n_g)}(\lambda)$ denote the ordered losses within group g. Following the Bayesian-quadrature view of conformal risk estimation, we draw random quantile spacings

$$\mathbf{U}^{(m)} = \left(U_1^{(m)}, \dots, U_{n_g+1}^{(m)}\right) \sim \mathrm{Dir}\left(\mathbf{1}_{n_g+1}\right), \qquad m = 1, \dots, M. \tag{5}$$

Each draw constructs a conservative group-risk sample

$$L_g^{+(m)}(\lambda) = \sum_{j=1}^{n_g} U_j^{(m)} \ell_{(j)}(\lambda) + U_{n_g+1}^{(m)} B, \tag{6}$$

where the last term assigns the maximum possible loss to the unobserved upper-tail interval. Rather than relying only on the empirical mean, G-BRT uses the full collection $\{L_g^{+(m)}(\lambda)\}_{m=1}^M$ to characterize uncertainty in the group-wise expected threshold loss.

For confidence level $\beta \in (0, 1)$, we compute the corresponding upper-risk quantile

$$b_{\beta,g}(\lambda) = \mathrm{Quantile}_\beta\left(\left\{L_g^{+(m)}(\lambda)\right\}_{m=1}^M\right). \tag{7}$$

The instantaneous group threshold is then selected as the largest candidate whose upper-risk quantile does not exceed the target risk α:

$$\widehat{\lambda}_g = \sup\left\{\lambda \in [0, 1] : b_{\beta,g}(\lambda) \le \alpha\right\}. \tag{8}$$

This threshold is obtained through one-dimensional bisection. Because $\ell_i(\lambda)$ is non-decreasing in λ, Eq. equation 8 selects the tightest risk-feasible confidence boundary.

Since the reference predictions change as the predictor is optimized, G-BRT periodically updates the threshold bank. To reduce Monte Carlo fluctuations across update epochs, we employ exponential moving averaging:

$$\lambda_g^{(e+1)} = \rho \lambda_g^{(e)} + (1 - \rho)\widehat{\lambda}_g^{(e)}, \qquad \Lambda^{(e)} = \left\{\lambda_g^{(e)}\right\}_{g=1}^G, \tag{9}$$

where $\rho \in [0, 1)$ controls temporal smoothing. At the first threshold update, λ_g is initialized by $\widehat{\lambda}_g$. G-BRT therefore replaces one globally shared confidence boundary with a bank of group-adaptive risk standards. Nevertheless, the threshold estimation is non-parametric and gradient-free; it identifies threshold violations but cannot directly improve the MSFMamba predictor. This motivates the threshold-coupled learning mechanism introduced next.

3.3 Threshold-Coupled Margin Regularization and Objective Optimization

Although G-BRT establishes group-specific risk standards, these thresholds remain detached from representation learning if they are used only for post-hoc risk measurement. We therefore propose Threshold-Coupled Margin Regularization (TCMR), which converts the measured group-wise threshold violations into differentiable training signals.

For each training mini-batch $\mathcal{B} \subset \mathcal{D}_{\mathrm{tr}}$, the Mamba encoder and classifier produce

$$\mathbf{F}_{\mathrm{ms}} = \Phi_{\mathrm{enc}}(\mathbf{X}^h, \mathbf{X}^x), \qquad \mathbf{z} = \Phi_{\mathrm{cls}}(\mathbf{F}_{\mathrm{ms}}), \qquad \mathbf{q} = \mathrm{softmax}(\mathbf{z}). \tag{10}$$

For sample i, let $g_i = \mathcal{G}(h_i, w_i, d_i)$ denote its observable group and λ_{g_i} the corresponding threshold. We define its threshold-coupled confidence deficit as

$$\delta_i = [\lambda_{g_i} - q_{i,y_i}]_+ = \max\left(0, \lambda_{g_i} - q_{i,y_i}\right). \tag{11}$$

When $q_{i,y_i} \geq \lambda_{g_i}$, the sample already satisfies its current group-wise confidence standard and $\delta_i = 0$. Otherwise, $\delta_i > 0$ explicitly measures how far its true-class probability falls below the corresponding threshold.

TCMR penalizes these deficits using a powered-hinge objective:

$$\mathcal{L}_{\mathrm{TCMR}} = \frac{\mu}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} \delta_i^{\gamma} = \frac{\mu}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} [\lambda_{g_i} - q_{i,y_i}]_+^{\gamma}, \tag{12}$$

where μ controls the regularization strength and γ determines the emphasis placed on large threshold violations. During each mini-batch update, the current threshold bank is treated as a fixed risk reference, and gradients are propagated only through the predicted probabilities and predictor parameters.

The standard classification objective is

$$\mathcal{L}_{\mathrm{CE}} = -\frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} \log q_{i,y_i}. \tag{13}$$

Although cross-entropy learns discriminative class boundaries, it does not contain an explicit group-specific reliability reference determined by $(\alpha, \beta, \lambda_g)$. We therefore optimize the combined objective

$$\mathcal{L}_{\mathrm{train}} = \mathcal{L}_{\mathrm{CE}} + \mathcal{L}_{\mathrm{TCMR}}, \tag{14}$$

and update the predictor as

$$\Theta \leftarrow \Theta - \eta \nabla_{\Theta} \mathcal{L}_{\mathrm{train}}. \tag{15}$$

Equations equation 9–equation 15 establish a periodic threshold–learning interaction:

$$\mathbf{q}^{(e)} \longrightarrow \Lambda^{(e)} \longrightarrow \mathcal{L}_{\mathrm{TCMR}} \longrightarrow \Theta^{(e+1)} \longrightarrow \mathbf{q}^{(e+1)}. \tag{16}$$

G-BRT determines which samples violate group-specific risk standards, whereas TCMR uses these violations to reshape the predictor. However, finite model capacity, sensor noise, mixed pixels, shadows, and ambiguous class boundaries may leave a subset of low-confidence predictions even after threshold-coupled optimization. We consequently retain the learned threshold bank to expose and handle these residual violations after training.

3.4 Risk-Aware Selective Refinement

Threshold-coupled training improves samples that violate their group-specific confidence standards, but it may not eliminate every ambiguous prediction. Applying repeated inference to all samples would introduce unnecessary computation for already confident predictions. We therefore propose

Risk-Aware Selective Refinement (RASR), which reuses the learned group-risk thresholds to identify and selectively update the residual high-risk subset.

After training, both the predictor parameters and threshold bank are frozen:

$$\Theta \to \Theta^*, \qquad \Lambda \to \Lambda^* = \{\lambda_g^*\}_{g=1}^G. \tag{17}$$

For an unlabeled inference batch, the frozen predictor first performs a default forward pass:

$$\mathbf{z}_i^0 = f_{\Theta^*}(\mathbf{X}_i^h, \mathbf{X}_i^x), \qquad \mathbf{q}_i^0 = \mathrm{softmax}(\mathbf{z}_i^0). \tag{18}$$

Unlike G-BRT and TCMR, which use the true-class probability during labeled risk estimation and training, RASR must operate without test labels. We therefore define the top-class confidence

$$q_{\mathrm{top},i}^0 = \max_{c\in\{1,\dots,C\}} q_{i,c}^0, \tag{19}$$

and construct the residual high-risk subset as

$$\mathcal{H} = \left\{ i : q_{\mathrm{top},i}^0 < \lambda_{g_i}^* \right\}, \qquad g_i = \mathcal{G}(h_i, w_i, d_i). \tag{20}$$

Here, $\mathcal{H}$ is an index set that exposes predictions remaining below their group-wise confidence standards after TCMR; it is not a new feature representation and is not itself optimized.

For each $i \in \mathcal{H}$, RASR constructs two additional views by synchronously flipping both modalities. Let

$$\begin{aligned} \mathcal{A}_H(\mathcal{X}_i) &= \left\{ \mathrm{HFlip}(\mathbf{X}_i^h), \mathrm{HFlip}(\mathbf{X}_i^x) \right\}, \\ \mathcal{A}_V(\mathcal{X}_i) &= \left\{ \mathrm{VFlip}(\mathbf{X}_i^h), \mathrm{VFlip}(\mathbf{X}_i^x) \right\}, \end{aligned} \tag{21}$$

where identical spatial transformations preserve cross-modal registration. The corresponding probability predictions are

$$\begin{aligned} \mathbf{q}_i^H &= \mathrm{softmax}\left(f_{\Theta^*}\left(\mathcal{A}_H(\mathcal{X}_i) \right) \right), \\ \mathbf{q}_i^V &= \mathrm{softmax}\left(f_{\Theta^*}\left(\mathcal{A}_V(\mathcal{X}_i) \right) \right). \end{aligned} \tag{22}$$

Because horizontal and vertical flips preserve the center-pixel land-cover label while changing the orientation of its surrounding context, their predictions provide complementary views of orientation-sensitive spatial evidence. RASR aggregates the original and flipped predictions as

$$\overline{\mathbf{q}}_i^{\mathrm{TTA}} = \frac{1}{3}\left(\mathbf{q}_i^0 + \mathbf{q}_i^H + \mathbf{q}_i^V \right), \qquad i \in \mathcal{H}. \tag{23}$$

This probability averaging can reduce view-specific prediction variance when the errors induced by different orientations are not perfectly correlated.

Finally, RASR selectively updates only the predictions indexed by $\mathcal{H}$:

$$\mathbf{p}_i = \begin{cases} \overline{\mathbf{q}}_i^{\mathrm{TTA}}, & i \in \mathcal{H}, \\ \mathbf{q}_i^0, & i \notin \mathcal{H}, \end{cases} \tag{24}$$

and produces the final land-cover prediction

$$\widehat{y}_i = \arg\max_{c\in\{1,\dots,C\}} p_{i,c}. \tag{25}$$

RASR does not introduce an additional training loss or update Θ^* and Λ^*. Instead, it serves as the final residual-risk handling mechanism: confident samples directly retain their base predictions, whereas only the remaining high-risk samples receive multi-view probability refinement.

4 Experiments and Discussion

Experimental setup We evaluate RC-Mamba on four multi-source remote-sensing benchmarks: Berlin (Okujeni et al., 2016; Hong et al., 2021) and Augsburg (Hong et al., 2021) with HSI–SAR data, and Houston2018 (Le Saux et al., 2018; Xu et al., 2019) and Houston2013 (Debes et al., 2014) with HSI–LiDAR data. We report OA, AA, and κ for classification, together with empirical coverage, coverage gap, prediction-set size, and group-wise reliability metrics. Following MSF-Mamba (Gao et al., 2025a), we retain 35/30/30/35 HSI principal components and train with Adam for 40 epochs using 11×11 patches, batch size 128, and learning rate 10^{-4}; unless otherwise stated, $\alpha = 0.10$, $\beta = 0.95$, $M = 1000$, EMA$= 0.8$, while TCMR uses $\mu = 0.5, \gamma = 2$ and RASR uses original and flipped views. Training thresholds are estimated from a periodically sampled training cache, whereas post-training reliability is evaluated on disjoint $\mathcal{D}_{\mathrm{ref}}/\mathcal{D}_{\mathrm{eval}}$ splits; complete dataset statistics, metric definitions, and implementation details are provided in Appendices C.1 and C.2.

Table 1: Comparison on four multi-source remote-sensing datasets. **Bold** indicates the best result. OA, AA, and κ are reported as percentages. Δ denotes the performance difference between RC-Mamba and MSFMamba.

Method	Berlin			Augsburg			Houston2013			Houston2018		
	OA (%)	AA (%)	κ (%)	OA (%)	AA (%)	κ (%)	OA (%)	AA (%)	κ (%)	OA (%)	AA (%)	κ (%)
TBCNN (Xu et al., 2017)	67.60	54.72	50.96	84.53	47.93	77.13	82.91	84.38	81.43	86.95	92.36	83.39
FusAtNet (Mohla et al., 2020)	70.91	44.13	51.07	85.31	62.93	86.33	85.32	87.56	84.12	89.09	91.91	85.96
S^2ENet (Fang et al., 2021)	70.38	64.22	57.73	88.22	53.08	86.47	89.56	91.50	88.69	90.05	93.14	87.20
DFINet (Gao et al., 2021)	73.69	63.51	61.02	90.66	61.31	86.47	88.59	90.63	87.64	91.02	95.24	88.46
AsyFFNet (Li et al., 2022)	71.65	**64.33**	58.74	88.25	62.23	83.13	87.95	89.81	86.96	91.24	95.02	88.71
ExViT (Yao et al., 2023)	72.59	62.39	59.72	86.65	58.41	80.79	87.42	89.37	86.38	89.98	94.02	87.14
HCT (Zhao et al., 2022)	73.42	60.57	60.29	90.33	61.68	86.03	90.66	92.07	89.86	90.56	94.74	87.81
MACN (Li et al., 2023)	73.98	63.58	61.19	**91.06**	**63.19**	**87.06**	89.81	91.33	88.93	90.31	94.61	87.55
MSFMamba (Gao et al., 2025a)	74.39	61.97	61.40	88.28	53.54	82.63	92.36	93.62	91.73	92.00	95.43	89.67
RC-Mamba (Ours)	**75.23**	62.20	**62.40**	88.36	53.39	82.73	**92.57**	**93.81**	**91.96**	**92.13**	**95.59**	**89.84**
Δ (Ours vs. MSFMamba)	+0.84	+0.23	+1.00	+0.08	−0.15	+0.10	+0.21	+0.19	+0.23	+0.13	+0.16	+0.17

4.1 Comparison with the State of the Art

Comparison results. Table 1 compares RC-Mamba with representative multi-source classification methods on four benchmarks, with MSFMamba (Gao et al., 2025a) serving as the direct backbone reference. Compared with MSFMamba, RC-Mamba improves Berlin OA/κ by $0.84/1.00$ points and Augsburg OA by 0.08 points, while AA decreases slightly by 0.15 points; on Houston2013 and Houston2018, OA increases from 92.36/92.00 to 92.57/92.13, with AA/κ gains of 0.19/0.23 and 0.16/0.17, respectively. Unlike post-hoc conformal wrappers, RC-Mamba couples group-wise risk measurement with threshold-aware learning while retaining the same MSFMamba encoder; Figure 3 shows representative Houston2013 classification maps, and the remaining class-wise results and maps are provided in Appendix C.3.

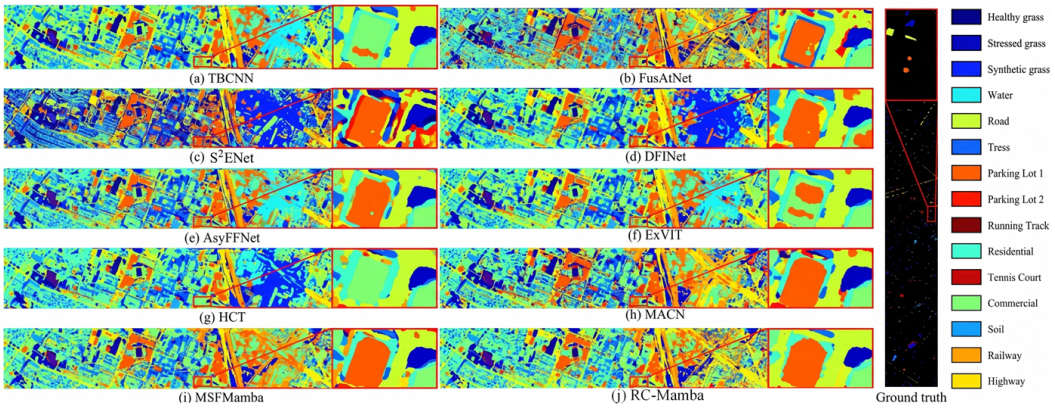


Figure 3: Classification maps on Houston2013. Results for Berlin, Augsburg, and Houston2018 are provided in Appendix C.3.

4.2 Ablation Study

Ablation analysis. Table 2 evaluates G-BRT, TCMR, and RASR along the progression from risk measurement to risk-aware learning and residual-risk handling. G-BRT and TCMR mainly establish and incorporate group-specific risk references, while RASR further improves OA by $+0.04$–$+0.35$ points and yields the best progressive-ablation results on all four datasets; reliability effects are analyzed in Section 4.3. Figure 4 further shows slightly improved feature separation for RC-Mamba over MSFMamba, particularly on Houston2018.

Table 2: Ablation studies of the proposed components on four datasets.

G-BRT	TCMR	RASR	Berlin	Augsburg	H2013	H2018
✗	✗	✗	74.39	88.28	92.36	92.00
✓	✗	✗	74.87	88.29	92.26	92.07
✓	✓	✗	74.88	88.30	92.25	92.09
✓	✓	✓	**75.23**	**88.36**	**92.57**	**92.13**

4.3 Reliability and Robustness Analysis

Reliability and robustness. Compared with MSFMamba equipped with Split-CP (Angelopoulos & Bates, 2023), CRC (Angelopoulos et al., 2024), and RCPS (Bates et al., 2021), RC-Mamba remains closer to the target coverage on all four benchmarks, with a maximum marginal coverage deviation of at most 0.007; Appendix C.4 further reports group-coverage standard deviations of 0.018–0.023 and worst-group gaps of 0.036–0.047. Under controlled class-resampling and spatial tile-removal perturbations, RC-Mamba shows slower coverage degradation, while the reproduced Dirichlet-spacing construction of CPBQ (Snell & Griffiths, 2025) closely matches the original formulation, supporting the reliability stability of RC-Mamba under the tested shifts and the implementation consistency of the Bayesian risk estimator used in G-BRT.

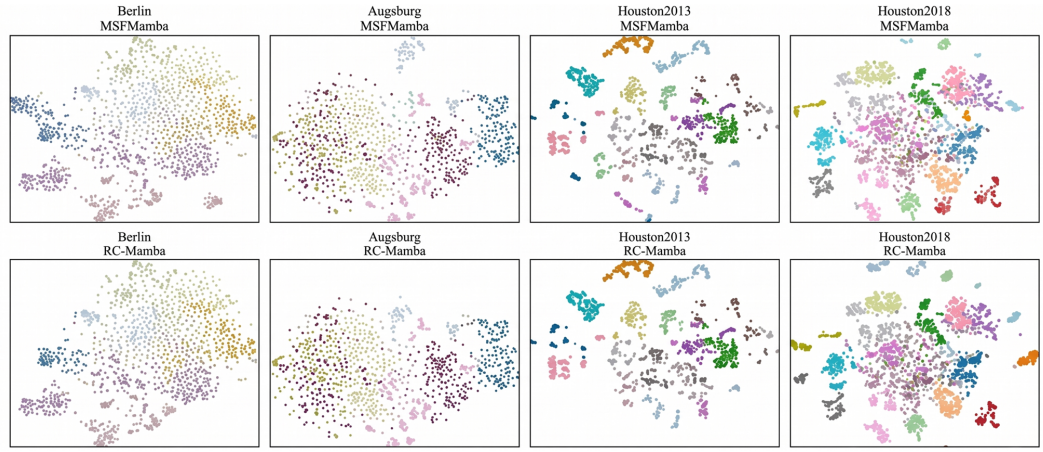


Figure 4: Joint t-SNE visualization of penultimate fused features on Berlin, Augsburg, Houston2013, and Houston2018. The top row shows MSFMamba and the bottom row shows RC-Mamba. Each pair uses the same stratified test samples and a shared embedding coordinate system. Colors denote ground-truth classes.

5 Conclusion

In this paper, we explore the challenge of heterogeneous prediction reliability across spatial groups and input samples in multi-source remote-sensing classification. Our work introduces a simple yet effective reliability-aware classification framework, namely RC-Mamba. We use group-specific Bayesian risk thresholds together with threshold-coupled margin regularization to quantify local reliability and achieve group-aware risk-sensitive learning. Additionally, we apply risk-aware selective refinement to residual high-risk predictions while preserving confident predictions. The effectiveness of RC-Mamba is evaluated against representative multi-source classifiers on four HSI–SAR and HSI–LiDAR benchmarks, with independent-reference diagnostics used to audit its reliability behavior. We hope this work provides a practical paradigm for coupling risk measurement, risk-aware learning, and selective inference, paving the way for more reliable multi-source remote-sensing systems.

AUTHOR CONTRIBUTIONS

If you'd like to, you may include a section for author contributions as is done in many journals. This is optional and at the discretion of the authors.

ACKNOWLEDGMENTS

Use unnumbered third level headings for the acknowledgments. All acknowledgments, including those to funding agencies, go at the end of the paper.

REFERENCES

Anastasios Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. In *International conference on learning representations*, volume 2024, pp. 55198–55218, 2024.

Anastasios N Angelopoulos and Stephen Bates. Conformal prediction: A gentle introduction. *Foundations and Trends in Machine Learning*, 16(4):494–591, 2023.

Nicolas Audebert, Bertrand Le Saux, and Sébastien Lefèvre. Beyond rgb: Very high resolution urban remote sensing with multimodal deep networks. *ISPRS journal of photogrammetry and remote sensing*, 140:20–32, 2018.

Rina Foygel Barber, Emmanuel J Candes, Aaditya Ramdas, and Ryan J Tibshirani. Conformal prediction beyond exchangeability. *The Annals of Statistics*, 51(2):816–845, 2023.

Stephen Bates, Anastasios Angelopoulos, Lihua Lei, Jitendra Malik, and Michael Jordan. Distribution-free, risk-controlling prediction sets. *Journal of the ACM (JACM)*, 68(6):1–34, 2021.

Tri Dao and Albert Gu. Transformers are ssms: Generalized models and efficient algorithms through structured state space duality. *arXiv preprint arXiv:2405.21060*, 2024.

Christian Debes, Andreas Merentitis, Roel Heremans, Jürgen Hahn, Nikolaos Frangiadakis, Tim Van Kasteren, Wenzhi Liao, Rik Bellens, Aleksandra Pižurica, Sidharta Gautama, et al. Hyperspectral and lidar data fusion: Outcome of the 2013 grss data fusion contest. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 7(6):2405–2418, 2014.

Bat-Sheva Einbinder, Yaniv Romano, Matteo Sesia, and Yanfei Zhou. Training uncertainty-aware classifiers with conformalized deep learning. *Advances in neural information processing systems*, 35:22380–22395, 2022.

Sheng Fang, Kaiyu Li, and Zhe Li. S^2enet: Spatial–spectral cross-modal enhancement network for classification of hyperspectral and lidar data. *IEEE Geoscience and Remote Sensing Letters*, 19: 1–5, 2021.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pp. 1050–1059. PMLR, 2016.

Feng Gao, Xuepeng Jin, Xiaowei Zhou, Junyu Dong, and Qian Du. Msfmamba: Multi-scale feature fusion state space model for multi-source remote sensing image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 2025a.

Feng Gao, Sheng Liu, Chuanzheng Gong, Xiaowei Zhou, Jiayi Wang, Junyu Dong, and Qian Du. Prototype-based information compensation network for multi-source remote sensing data classification. *IEEE Transactions on Geoscience and Remote Sensing*, 2025b.

Yunhao Gao, Wei Li, Mengmeng Zhang, Jianbu Wang, Weiwei Sun, Ran Tao, and Qian Du. Hyperspectral and multispectral classification for coastal wetland using depthwise feature interaction network. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–15, 2021.

Isaac Gibbs and Emmanuel J Candès. Conformal inference for online prediction with arbitrary distribution shifts. *Journal of Machine Learning Research*, 25(162):1–36, 2024.

Isaac Gibbs, John J Cherian, and Emmanuel J Candès. Conformal prediction with conditional guarantees. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 87(4):1100–1126, 2025.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pp. 1321–1330. PMLR, 2017.

Danfeng Hong, Lianru Gao, Naoto Yokoya, Jing Yao, Jocelyn Chanussot, Qian Du, and Bing Zhang. More diverse means better: Multimodal deep learning meets remote-sensing imagery classification. *IEEE Transactions on Geoscience and Remote Sensing*, 59(5):4340–4354, 2020.

Danfeng Hong, Jingliang Hu, Jing Yao, Jocelyn Chanussot, and Xiao Xiang Zhu. Multimodal remote sensing benchmark datasets for land cover classification with a shared and specific feature learning model. *ISPRS Journal of Photogrammetry and Remote Sensing*, 178:68–80, 2021.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017.

Bertrand Le Saux, Naoto Yokoya, Ronny Hänsch, and Saurabh Prasad. 2018 ieee grss data fusion contest: Multimodal land use classification [technical committees]. *IEEE geoscience and remote sensing magazine*, 6(1):52–54, 2018.

Ke Li, Di Wang, Xu Wang, Gang Liu, Zili Wu, and Quan Wang. Mixing self-attention and convolution: A unified framework for multisource remote sensing data classification. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–16, 2023.

Shufan Li, Harkanwar Singh, and Aditya Grover. Mamba-nd: Selective state space modeling for multi-dimensional data. In *European Conference on Computer Vision*, pp. 75–92. Springer, 2024.

Wei Li, Yunhao Gao, Mengmeng Zhang, Ran Tao, and Qian Du. Asymmetric feature fusion network for hyperspectral and sar image classification. *IEEE Transactions on Neural Networks and Learning Systems*, 34(10):8057–8070, 2022.

Chao Liang, Yuqing Zhao, Yu Song, and Kang Ni. Idnet: Intensity-constrained detail-enhanced network for hyperspectral and lidar collaborative classification. *IEEE Transactions on Geoscience and Remote Sensing*, 63:1–15, 2025.

Diling Liao, Qingsong Wang, Tao Lai, and Haifeng Huang. Joint classification of hyperspectral and lidar data based on mamba. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–15, 2024.

Yue Liu, Yunjie Tian, Yuzhong Zhao, Hongtian Yu, Lingxi Xie, Yaowei Wang, Qixiang Ye, Jianbin Jiao, and Yunfan Liu. Vmamba: Visual state space model. *Advances in neural information processing systems*, 37:103031–103063, 2024.

Fan Lu, Kai Zhu, Wei Zhai, Kecheng Zheng, and Yang Cao. Uncertainty-aware optimal transport for semantically coherent out-of-distribution detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3282–3291, 2023.

Satyam Mohla, Shivam Pande, Biplab Banerjee, and Subhasis Chaudhuri. Fusatnet: Dual attention based spectrospatial multimodal fusion network for hyperspectral and lidar classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pp. 92–93, 2020.

Akpona Okujeni, Sebastian van der Linden, and Patrick Hostert. Berlin-urban-gradient dataset 2009-an enmap preparatory flight campaign. 2016.

Yaniv Romano, Matteo Sesia, and Emmanuel Candes. Classification with valid and adaptive coverage. *Advances in neural information processing systems*, 33:3581–3591, 2020.

Swalpa Kumar Roy, Ankur Deria, Danfeng Hong, Behnood Rasti, Antonio Plaza, and Jocelyn Chanussot. Multimodal fusion transformer for remote sensing image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–20, 2023.

Murat Sensoy, Lance Kaplan, and Melih Kandemir. Evidential deep learning to quantify classification uncertainty. *Advances in neural information processing systems*, 31, 2018.

Jake C Snell and Thomas L Griffiths. Conformal prediction as bayesian quadrature. *arXiv preprint arXiv:2502.13228*, 2025.

Jake C Snell, Thomas P Zollo, Zhun Deng, Toniann Pitassi, and Richard Zemel. Quantile risk control: A flexible framework for bounding the probability of high-loss predictions. *arXiv preprint arXiv:2212.13629*, 2022.

Qiya Song, Feng Mo, Kexin Ding, Lin Xiao, Renwei Dian, Xudong Kang, and Shutao Li. Mcfnet: Multiscale cross-domain fusion network for hsi and lidar data joint classification. *IEEE Transactions on Geoscience and Remote Sensing*, 2025.

David Stutz, Ali Taylan Cemgil, Arnaud Doucet, et al. Learning optimal conformal classifiers. *arXiv preprint arXiv:2110.09192*, 2021.

Ryan J Tibshirani, Rina Foygel Barber, Emmanuel Candes, and Aaditya Ramdas. Conformal prediction under covariate shift. *Advances in neural information processing systems*, 32, 2019.

Xianghai Wang, Junheng Zhu, Yining Feng, and Lu Wang. Ms2canet: Multiscale spatial–spectral cross-modal attention network for hyperspectral image and lidar classification. *IEEE Geoscience and Remote Sensing Letters*, 21:1–5, 2024.

Xiaodong Xu, Wei Li, Qiong Ran, Qian Du, Lianru Gao, and Bing Zhang. Multisource remote sensing data classification based on convolutional neural network. *IEEE Transactions on Geoscience and Remote Sensing*, 56(2):937–949, 2017.

Yonghao Xu, Bo Du, Liangpei Zhang, Daniele Cerra, Miguel Pato, Emiliano Carmona, Saurabh Prasad, Naoto Yokoya, Ronny Hänsch, and Bertrand Le Saux. Advanced multi-sensor optical remote sensing for urban land use and land cover classification: Outcome of the 2018 ieee grss data fusion contest. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(6):1709–1724, 2019.

Jing Yao, Bing Zhang, Chenyu Li, Danfeng Hong, and Jocelyn Chanussot. Extended vision transformer (exvit) for land use and land cover classification: A multimodal deep learning framework. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–15, 2023.

Jing Yao, Danfeng Hong, Chenyu Li, and Jocelyn Chanussot. Spectralmamba: Efficient mamba for hyperspectral image classification. *arXiv preprint arXiv:2404.08489*, 2024.

Guangrui Zhao, Qiaolin Ye, Le Sun, Zebin Wu, Chengsheng Pan, and Byeungwoo Jeon. Joint classification of hyperspectral and lidar data using a hierarchical cnn and transformer. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–16, 2022.

Lianghui Zhu, Bencheng Liao, Qian Zhang, Xinlong Wang, Wenyu Liu, and Xinggang Wang. Vision mamba: Efficient visual representation learning with bidirectional state space model. *arXiv preprint arXiv:2401.09417*, 2024.

A Additional Theoretical Details and Algorithms for RC-Mamba

This appendix provides compact derivations and implementation-level algorithms for RC-Mamba. The presentation follows the decision-theoretic style of Conformal Prediction as Bayesian Quadrature, but adapts it to group-conditional multi-source remote sensing classification.

A.1 Decision-Theoretic View of Group-Conditional Calibration

Let $z = (\mathbf{X}, y)$ be a sample and let θ_g denote the unknown deployment distribution of group g. A confidence threshold λ_g incurs group-conditional miscoverage loss

$$L_g(\theta_g, \lambda_g) = \Pr_{\theta_g}\big[p_y(\mathbf{X}) < \lambda_g \mid \mathcal{G} = g\big]. \tag{26}$$

The maximum group risk is

$$\bar{R}_g(\lambda_g) = \sup_{\theta_g} L_g(\theta_g, \lambda_g). \tag{27}$$

A rule is α-acceptable for group g if $\bar{R}_g(\lambda_g) \le \alpha$. Classical split CP constructs such a rule in a marginal sense; RC-Mamba applies the rule separately within each group, yielding $\Lambda = \{\lambda_g\}_{g=1}^{K}$.

A.2 Bayesian-Quadrature Upper-Bound Threshold

For group g, define $\ell_i(\lambda) = \mathbb{1}[p_{i,y_i} < \lambda]$ and order the in-group losses as $\ell_{(1)}(\lambda) \le \cdots \le \ell_{(n_g)}(\lambda)$. Under the Bayesian-quadrature view, the unknown quantile spacings are sampled as

$$(U_1, \ldots, U_{n_g+1}) \sim \mathrm{Dir}(1, \ldots, 1). \tag{28}$$

The conservative posterior upper-bound random variable is

$$L_g^{+}(\lambda) = \sum_{j=1}^{n_g} U_j \ell_{(j)}(\lambda) + U_{n_g+1} B, \qquad B = 1. \tag{29}$$

The confidence-space group threshold is

$$\hat{\lambda}_g = \sup\{\lambda : \Pr(L_g^{+}(\lambda) \le \alpha) \ge \beta\}. \tag{30}$$

Therefore, with posterior confidence at least β, $R_g(\hat{\lambda}_g) \le \alpha$, which yields

$$\Pr\big[y \in \mathcal{C}_g(\mathbf{X}) \mid \mathcal{G} = g\big] \ge 1 - \alpha. \tag{31}$$

A.3 EMA Stabilisation

Let $\hat{\lambda}_g^{(e)}$ be the instantaneous group threshold at epoch e. The EMA update

$$\lambda_g^{(e+1)} = \rho \lambda_g^{(e)} + (1 - \rho)\hat{\lambda}_g^{(e)} \tag{32}$$

is a contraction:

$$|\lambda_g^{(e+1)} - \lambda_g^{\star}| \le \rho |\lambda_g^{(e)} - \lambda_g^{\star}| + (1 - \rho)|\hat{\lambda}_g^{(e)} - \lambda_g^{\star}|. \tag{33}$$

Thus, EMA reduces Monte Carlo jitter and preserves the calibration semantics up to an $\mathcal{O}(1 - \rho)$ smoothing slack.

A.4 TCMR as a Differentiable Surrogate

The true miscoverage indicator is non-differentiable. TCMR replaces it with

$$\psi_\gamma(\delta_i) = [\max(0, \delta_i)]^{\gamma}, \qquad \delta_i = \lambda_{g_i} - p_{\mathrm{true},i}. \tag{34}$$

For $\gamma \ge 1$, ψ_γ is sub-differentiable, and

$$\nabla_\Theta \mathcal{L}_{\mathrm{conf}} = -\frac{\mu}{|\mathcal{B}|} \sum_{i : p_{\mathrm{true},i} < \lambda_{g_i}} \gamma \delta_i^{\gamma - 1} \nabla_\Theta p_{\mathrm{true},i}. \tag{35}$$

Hence, TCMR explicitly increases true-class confidence only for samples that violate their group-specific reliability boundary.

A.5 RASR Cost Bound

Let $\rho_{\mathcal{H}} = |\mathcal{H}|/N_{\mathrm{te}}$ and $\kappa = c_{\mathcal{R}}/\mathrm{FLOPs}_0$. The total inference cost satisfies

$$\frac{\mathrm{FLOPs}_{\mathrm{total}}}{N_{\mathrm{te}} \mathrm{FLOPs}_0} = 1 + \rho_{\mathcal{H}} \kappa. \tag{36}$$

Thus, the overhead is controlled by the calibrated high-risk rate rather than the full test-set size.

Table 3: Summary of the dataset splits used in the experiments.

Dataset	Modalities	Classes	Train	Test
Berlin	HSI–SAR	8	2,820	461,851
Augsburg	HSI–SAR	7	761	77,533
Houston2018	HSI–LiDAR	20	18,750	2,000,160
Houston2013	HSI–LiDAR	15	2,832	12,197

B ALGORITHMS

This appendix provides the complete algorithmic details of RC-Mamba to facilitate reproducibility and clarify the interaction among the three proposed reliability modules. Algorithm B summarizes the overall training and inference procedure. The multi-scale Mamba backbone first produces fused representations and logits. G-BRT periodically updates group-wise risk thresholds, TCMR incorporates the resulting thresholds into model optimization, and RASR is activated after training for selective refinement of residual high-risk predictions.

Algorithm 1 is organized as a failure-driven reliability-learning chain. A single global threshold is insufficient under spatial and domain heterogeneity, motivating group-specific risk standards. These standards, however, only identify risk violations and do not directly improve the predictor, which motivates threshold-coupled margin optimization. Since training-time regularization may still leave residual low-confidence predictions, a final selective refinement stage is introduced to handle only the remaining high-risk cases.

C ADDITIONAL EXPERIMENTAL DETAILS

C.1 DATASET STATISTICS AND VISUALIZATION

Berlin. The Berlin benchmark covers urban and rural areas using simulated EnMAP HSI derived from HyMap observations and Sentinel-1 dual-polarization (VV–VH) SAR data (Okujeni et al., 2016; Hong et al., 2021). The HSI contains 244 bands over 400–2500 nm at 797×220 pixels, whereas the SAR image has a spatial size of 1723×476; nearest-neighbor interpolation aligns the two modalities.

Augsburg. The Augsburg benchmark pairs HySpex HSI with Sentinel-1 SAR observations over Augsburg, Germany (Hong et al., 2021). Both modalities have a ground sampling distance of 30 m and a spatial size of 332×485. The HSI contains 180 bands over 0.4–2.5 μm, and the SAR input contains four polarimetric features: VV intensity, VH intensity, and the real and imaginary parts of the off-diagonal covariance element.

Houston2018. The Houston2018 benchmark covers the University of Houston campus and its surrounding urban area and was released for the 2018 IEEE GRSS Data Fusion Contest (Le Saux et al., 2018; Xu et al., 2019). The benchmark description reports a 48-band HSI over 380–1050 nm and three-band multispectral LiDAR; the processed tensors exposed by the current implementation contain 50 HSI channels and one LiDAR-derived channel.

Houston2013. The Houston2013 benchmark was released for the 2013 IEEE GRSS Data Fusion Contest (Debes et al., 2014). It contains a 144-band HSI over 380–1050 nm and LiDAR elevation measurements at 2.5 m spatial resolution, covering 15 urban land-cover classes. Following the experimental split used in MSFMamba (Gao et al., 2025a), we use 2,832 training pixels and 12,197 test pixels; the corresponding class-wise counts are reported in Table 4.

Table 3 summarizes the dataset-level splits, and Table 4 provides the class-wise sample counts. Figure 5 shows the corresponding false-color HSI, LiDAR/SAR, and ground-truth maps.

Table 4: The number of training and test samples for the Berlin, Augsburg, Houston2018, and Houston2013 datasets.

No.	Name	Color	Train	Test
	Berlin dataset			
1	Forest		443	54,511
2	Residential area		423	268,219
3	Industrial area		499	19,067
4	Low plants		376	58,906
5	Soil		331	17,095
6	Allotment		280	13,025
7	Commercial area		298	24,526
8	Water		170	6,502
		Total	**2,820**	**461,851**
	Augsburg dataset			
1	Forest		146	13,361
2	Residential area		264	30,065
3	Industrial area		21	3,830
4	Low plants		248	26,609
5	Allotment		52	523
6	Commercial area		7	1,638
7	Water		23	1,507
		Total	**761**	**77,533**
	Houston2018 dataset			
1	Healthy grass		1,000	38,196
2	Stressed grass		1,000	129,008
3	Artificial turf		1,000	1,736
4	Evergreen trees		1,000	53,322
5	Deciduous trees		1,000	19,172
6	Bare earth		1,000	17,064
7	Water		500	564
8	Residential buildings		1,000	157,995
9	Non-residential buildings		1,000	893,769
10	Roads		1,000	182,283
11	Sidewalks		1,000	135,035
12	Crosswalks		1,000	5,059
13	Major thoroughfares		1,000	184,438
14	Highways		1,000	38,438
15	Railways		1,000	26,748
16	Paved parking lots		1,000	44,932
17	Unpaved parking lots		250	337
18	Cars		1,000	25,289
19	Trains		1,000	20,479
20	Stadium seats		1,000	26,296
		Total	**18,750**	**2,000,160**
	Houston2013 dataset			
1	Healthy grass		198	1,053
2	Stressed grass		190	1,064
3	Synthetic grass		192	505
4	Trees		188	1,056
5	Soil		186	1,056
6	Water		182	143
7	Residential		196	1,072
8	Commercial		191	1,053
9	Road		193	1,059
10	Highway		191	1,036
11	Railway		181	1,054
12	Parking lot 1		192	1,041
13	Parking lot 2		184	285
14	Tennis court		181	247
15	Running track		187	473
		Total	**2,832**	**12,197**

Algorithm 1 RC-Mamba: Group-Conditional Bayesian Risk Thresholding, Threshold-Coupled Margin Learning, and Risk-Aware Selective Refinement

Input: Multi-source training set $\mathcal{D}_{\mathrm{tr}} = \{(\mathbf{X}_i^h, \mathbf{X}_i^x, y_i, h_i, w_i)\}$; MSFMamba predictor f_Θ; observable tile/domain grouping rule $\mathcal{G}$; target risk α; confidence level β; Dirichlet samples M; update interval Δ_e; EMA factor ρ; TCMR weight μ; exponent γ; learning rate η.

Houston2013. The Houston2013 benchmark was released for the 2013 IEEE GRSS Data Fusion Contest (Debes et al., 2014). It contains a 144-band HSI over 380–1050 nm and LiDAR elevation measurements at 2.5 m spatial resolution, covering 15 urban land-cover classes. Following the experimental split used in MSFMamba (Gao et al., 2025a), we use 2,832 training pixels and 12,197 test pixels; the corresponding class-wise counts are reported in Table 4.

Output: Optimized predictor Θ^*; group-risk threshold bank $\Lambda^* = \{\lambda_g^*\}$; final class probabilities $\mathbf{p} \in [0,1]^C$ and predictions $\widehat{\mathbf{y}}$.

Initialize Θ, $\Lambda \leftarrow \varnothing$, and sample $\mathcal{D}_{\mathrm{ref}} \subset \mathcal{D}_{\mathrm{tr}}$.
for $e = 1, \ldots, E$ **do**
  if $e \bmod \Delta_e = 1$ **then** ▷ **G-BRT: establish group-wise risk standards**
    $\mathbf{z}_i \leftarrow f_\Theta(\mathbf{X}_i^h, \mathbf{X}_i^x)$, $\mathbf{q}_i \leftarrow \mathrm{softmax}(\mathbf{z}_i), \forall i \in \mathcal{D}_{\mathrm{ref}}$.
    $s_i \leftarrow q_{i,y_i}$, $g_i \leftarrow \mathcal{G}(h_i, w_i, \mathrm{domain}_i)$.
    for each observed group g **do**
      $\ell_i(\lambda) \leftarrow \mathbb{1}[s_i < \lambda]$, $\{\ell_{(j)}(\lambda)\}_{j=1}^{n_g} \leftarrow \mathrm{sort}\{\ell_i(\lambda) : g_i = g\}$.
      for $m = 1, \ldots, M$ **do**
        $\mathbf{U}^{(m)} \sim \mathrm{Dir}(\mathbf{1}_{n_g+1})$, $L_g^{+(m)}(\lambda) \leftarrow \sum_{j=1}^{n_g} U_j^{(m)} \ell_{(j)}(\lambda) + U_{n_g+1}^{(m)} B,\ B = 1$.
      end for
      $b_{\beta,g}(\lambda) \leftarrow \mathrm{Quantile}_\beta(\{L_g^{+(m)}(\lambda)\}_{m=1}^M)$.
      $\widehat{\lambda}_g \leftarrow \sup\{\lambda \in [0,1] : b_{\beta,g}(\lambda) \le \alpha\}$; $\lambda_g \leftarrow \rho\lambda_g + (1-\rho)\widehat{\lambda}_g$.
    end for
  end if
  for each mini-batch $\mathcal{B} \subset \mathcal{D}_{\mathrm{tr}}$ **do** ▷ **TCMR: turn measured deficits into learning feedback**
    $\mathbf{F}_{\mathrm{ms}} \leftarrow \Phi_{\mathrm{enc}}(\mathbf{X}^h, \mathbf{X}^x)$, $\mathbf{z} \leftarrow \Phi_{\mathrm{cls}}(\mathbf{F}_{\mathrm{ms}})$, $\mathbf{q} \leftarrow \mathrm{softmax}(\mathbf{z})$.
    $g_i \leftarrow \mathcal{G}(h_i, w_i, \mathrm{domain}_i)$, $\delta_i \leftarrow [\lambda_{g_i} - q_{i,y_i}]_+$.
    $\mathcal{L}_{\mathrm{TCMR}} \leftarrow \frac{\mu}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} \delta_i^\gamma$, $\mathcal{L} \leftarrow \mathcal{L}_{\mathrm{CE}}(\mathbf{z}, \mathbf{y}) + \mathcal{L}_{\mathrm{TCMR}}$.
    $\Theta \leftarrow \Theta - \eta \nabla_\Theta \mathcal{L}$.
  end for
end for
Freeze Θ^* and Λ^*; retain Λ^* to expose residual threshold violations after TCMR.
for each unlabeled inference batch **do** ▷ **RASR: selectively update residual high-risk predictions**
  $\mathbf{z}^0 \leftarrow f_{\Theta^*}(\mathbf{X}^h, \mathbf{X}^x)$, $\mathbf{q}^0 \leftarrow \mathrm{softmax}(\mathbf{z}^0)$, $g_i \leftarrow \mathcal{G}(h_i, w_i, \mathrm{domain}_i)$.
  $\mathcal{H} \leftarrow \{i : \max_c q_{i,c}^0 < \lambda_{g_i}^*\}$.
  $\mathbf{p} \leftarrow \mathbf{q}^0$.
  if $\mathcal{H} \neq \varnothing$ **then**
    $\mathbf{q}_{\mathcal{H}}^H \leftarrow \mathrm{softmax}(f_{\Theta^*}(\mathrm{HFlip}(\mathbf{X}_{\mathcal{H}}^h, \mathbf{X}_{\mathcal{H}}^x)))$.
    $\mathbf{q}_{\mathcal{H}}^V \leftarrow \mathrm{softmax}(f_{\Theta^*}(\mathrm{VFlip}(\mathbf{X}_{\mathcal{H}}^h, \mathbf{X}_{\mathcal{H}}^x)))$.
    $\mathbf{p}_{\mathcal{H}} \leftarrow (\mathbf{q}_{\mathcal{H}}^0 + \mathbf{q}_{\mathcal{H}}^H + \mathbf{q}_{\mathcal{H}}^V)/3$.
  end if
  $\widehat{y}_i \leftarrow \arg\max_c p_{i,c}$.
end for
return $\Theta^*, \Lambda^*, \mathbf{p}, \widehat{\mathbf{y}}$.

C.2 Complete Implementation and Reproducibility Details

Unified experimental configuration. All experiments are conducted for 40 epochs using the Adam optimizer with a learning rate of 10^{-4} and a batch size of 128. Following the original MSFMamba protocol (Gao et al., 2025a), the numbers of retained HSI principal components are set to 35, 30, 30, and 35 for Berlin, Augsburg, Houston2018, and Houston2013, respectively.

Training-time threshold cache. At each threshold refresh, a fraction of batches from the shuffled training loader is used to estimate the group-wise thresholds. These thresholds provide TCMR with a learning signal, while the same samples remain eligible for subsequent gradient updates. Therefore, the threshold cache serves only as a training-time optimization mechanism rather than an independent reference split.

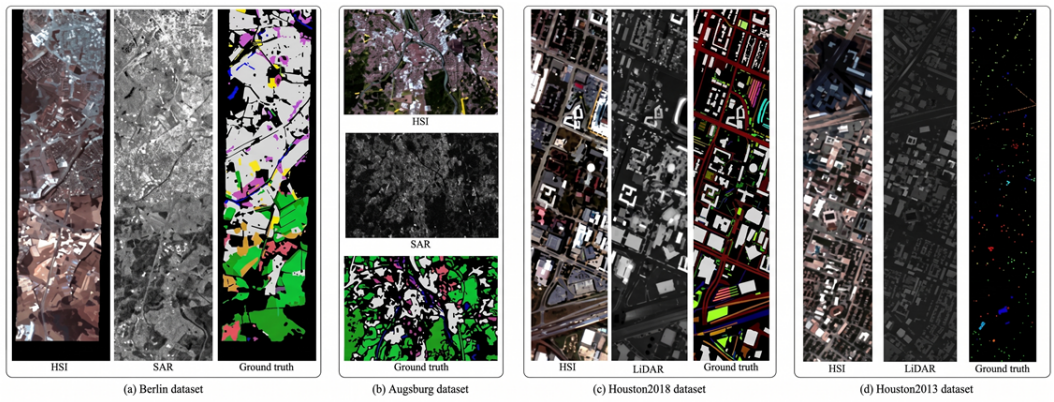


Figure 5: Visualization of the four datasets used in our experiments.

Table 5: Classification performance of different methods on the Berlin dataset. **Bold** indicates the best result.

Class	TBCNN	FusAtNet	S^2ENet	DFINet	AsyFFNet	ExViT	HCT	MACN	MSFMamba	RC-Mamba
Forest	81.75	86.24	83.27	82.04	**88.35**	84.71	83.18	82.21	73.27	75.15
Residential area	76.26	**91.38**	72.07	77.78	74.44	76.19	78.80	78.16	80.34	81.48
Industrial area	39.67	19.76	46.66	47.96	48.21	44.15	43.59	47.79	51.02	**54.67**
Low plants	49.78	20.00	72.08	77.78	71.41	77.96	76.48	78.04	82.20	**82.67**
Soil	**89.42**	48.72	77.94	87.85	80.75	67.13	82.92	77.26	75.15	72.78
Allotment	54.36	38.89	**70.62**	44.75	48.35	61.95	56.46	66.07	60.32	60.31
Commercial area	4.65	18.47	**36.48**	29.85	27.97	29.28	18.96	28.90	24.94	21.82
Water	41.93	29.61	54.64	60.09	**75.16**	56.98	44.16	50.22	48.47	48.74
OA	67.60	70.91	70.38	73.69	71.65	72.59	73.42	73.98	74.39	**75.23**
AA	54.72	44.13	64.22	63.51	**64.33**	62.39	60.57	63.58	61.97	62.20
Kappa	50.96	51.07	57.73	61.02	58.74	59.72	60.29	61.19	61.40	**62.40**

Groups and independent-reference evaluation. The formal configuration uses fixed scene-wide tiles of size 32, 48, 128, and 64 for Berlin, Augsburg, Houston2013, and Houston2018, respectively. Groups with fewer than 64 reference samples fall back to the global threshold. For reliability evaluation, the official test split is divided into two disjoint halves, one for threshold fitting and the other for reliability evaluation.

Metrics. For an evaluation set of size n,

$$\mathrm{Cov} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\{y_i \in \mathcal{C}(x_i)\}, \qquad \text{Cov-gap} = \left|\mathrm{Cov} - (1-\alpha)\right|.$$

We also report average prediction-set size, high-risk rate, group-coverage standard deviation, and the worst-group coverage gap,

$$\max_g \left|\mathrm{Cov}_g - (1-\alpha)\right|.$$

The violation rate is defined as $1 - \mathrm{Cov}$ and is omitted when it duplicates the information conveyed by coverage.

C.3 Full Per-Class Results and Classification Maps

The classification performance of various methods on the four datasets is presented in Tables 5–8. The classification maps of Berlin and Houston2013 are presented in the main paper, while those of Augsburg and Houston2018 are provided here for completeness.

Table 6: Classification performance of different methods on the Augsburg dataset. **Bold** indicates the best result.

Class	TBCNN	FusAtNet	S^2ENet	DFINet	AsyFFNet	ExViT	HCT	MACN	MSFMamba	RC-Mamba
Forest	90.88	94.34	**98.10**	97.38	97.56	90.04	97.26	97.84	91.90	92.04
Residential area	93.89	92.56	99.08	98.37	99.16	95.44	98.63	98.73	99.17	**99.24**
Industrial area	8.28	47.70	12.19	61.31	**61.38**	34.58	36.61	43.94	15.26	15.28
Low plants	91.97	85.96	91.78	92.63	83.47	90.68	93.54	94.11	94.13	**94.23**
Allotment	38.24	49.33	45.12	49.33	42.64	51.82	51.43	**52.39**	40.86	40.03
Commercial area	1.40	11.52	1.22	3.54	5.58	**28.63**	7.97	6.80	1.31	1.04
Water	10.82	45.27	24.09	26.61	45.81	17.65	46.34	**48.54**	32.14	31.87
OA	84.53	85.31	88.22	90.66	88.25	86.65	90.33	**91.06**	88.28	88.36
AA	47.93	62.93	53.08	61.31	62.23	58.41	61.68	**63.19**	53.54	53.39
Kappa	77.13	86.33	86.47	86.47	83.13	80.79	86.03	**87.06**	82.63	82.73

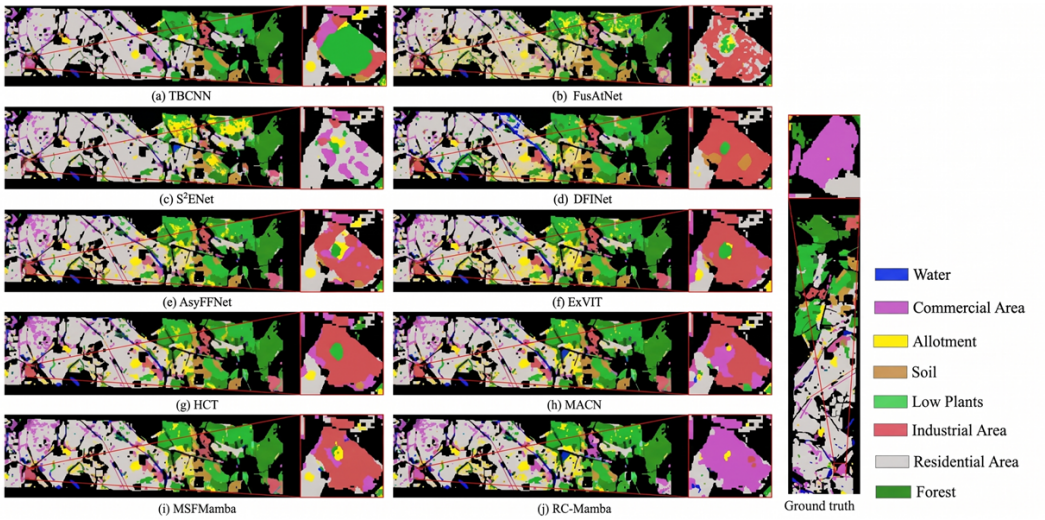


Figure 6: Classification results of different methods for the Berlin dataset.

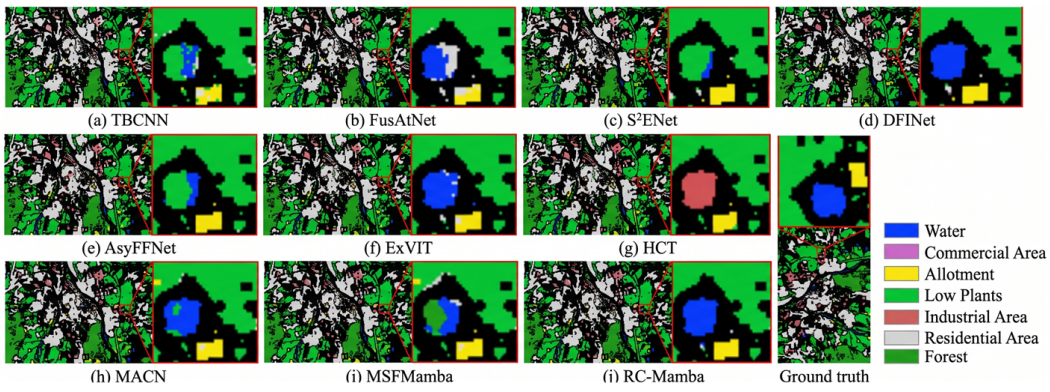


Figure 7: Classification results of different methods for the Augsburg dataset.

C.4 Additional Reliability and Controlled-Shift Results

Table 9 and Figs. 10–11 provide complementary reliability diagnostics. Across the four datasets, RC-Mamba keeps $\sup_\alpha |\mathrm{Cov}(\alpha) - (1-\alpha)| \le 0.007$, while $\mathrm{std}(\mathrm{Cov}_g)$ ranges from 0.018 to 0.023

Table 7: Classification performance of different methods on the Houston2018 dataset. **Bold** indicates the best result.

Class	TBCNN	FusAtNet	S^2ENet	DFINet	AsyFFNet	ExViT	HCT	MACN	MSFMamba	RC-Mamba
Healthy grass	94.84	60.69	**99.15**	94.29	96.90	93.34	97.38	89.53	95.04	95.23
Stressed grass	92.60	**93.85**	84.35	92.71	91.82	93.44	89.40	93.50	91.43	91.60
Artificial turf	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**
Evergreen trees	98.80	99.02	98.13	98.92	99.29	95.87	98.90	99.10	99.33	**99.39**
Deciduous trees	97.12	94.73	86.39	97.77	**99.52**	96.85	98.89	98.49	98.65	99.20
Bare earth	99.61	99.87	99.33	**100.0**	99.98	**100.0**	**100.0**	99.96	**100.0**	**100.0**
Water	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**
Residential buildings	93.21	95.51	97.17	97.36	95.16	**97.39**	95.30	95.31	96.74	95.92
Non-residential buildings	91.30	93.64	93.93	93.29	95.02	93.00	**95.25**	93.86	95.12	94.87
Roads	61.06	70.94	70.17	75.73	72.72	72.18	72.77	72.11	**80.75**	77.93
Sidewalks	75.91	79.11	72.11	**83.67**	77.18	69.69	73.27	80.69	79.88	81.69
Crosswalks	85.31	86.06	82.47	94.23	93.14	88.12	**98.95**	97.69	95.32	95.63
Major thoroughfares	72.77	76.12	**89.44**	81.20	84.67	86.12	80.90	79.05	82.00	86.73
Highways	95.86	98.89	98.53	98.91	**99.53**	99.50	98.01	96.86	97.81	96.73
Railways	99.78	99.88	99.21	**99.94**	99.90	99.87	99.74	99.90	99.86	99.86
Paved parking lots	90.74	95.88	95.62	98.37	**99.18**	97.02	97.07	97.74	98.38	98.27
Unpaved parking lots	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**
Cars	98.47	94.25	96.77	**99.09**	96.64	98.13	99.01	98.74	98.44	98.70
Trains	99.90	99.84	**100.0**	99.46	**100.0**	99.98	99.97	99.74	99.95	99.96
Stadium seats	99.92	**100.0**	**100.0**	99.98	99.89	99.99	99.98	99.99	**100.0**	**100.0**
OA	86.95	89.09	90.05	91.02	91.24	89.98	90.56	90.31	92.00	**92.13**
AA	92.36	91.91	93.14	95.24	95.02	94.02	94.74	94.61	95.43	**95.59**
Kappa	83.39	85.96	87.20	88.46	88.71	87.14	87.81	87.55	89.67	**89.84**

Table 8: Classification performance of different methods on the Houston2013 dataset. **Bold** indicates the best result.

Class	TBCNN	FusAtNet	S^2ENet	DFINet	AsyFFNet	ExViT	HCT	MACN	MSFMamba	RC-Mamba
Healthy grass	82.72	82.53	81.48	82.91	82.72	82.72	81.10	82.90	90.60	**90.85**
Stressed grass	83.18	85.15	84.68	85.15	83.93	85.15	86.09	**100.0**	95.11	94.81
Synthetic grass	**100.0**	98.81	99.01	99.60	**100.0**	99.21	99.80	99.80	99.85	99.85
Trees	94.03	92.90	92.14	92.33	91.95	91.95	98.01	98.11	97.55	**98.29**
Soil	99.15	**100.0**	**100.0**	99.81	**100.0**	**100.0**	**100.0**	**100.0**	99.37	99.53
Water	99.30	98.60	**100.0**	**100.0**	95.80	98.60	95.10	95.80	99.78	**100.0**
Residential	79.66	85.45	89.93	93.47	**96.18**	86.66	83.49	78.17	91.91	91.35
Commercial	55.27	81.81	90.18	82.58	81.14	**91.05**	86.43	80.46	87.17	87.58
Road	75.83	83.76	**92.63**	89.05	86.12	91.12	86.69	90.93	75.92	77.06
Highway	62.55	53.67	64.86	56.56	64.58	64.86	79.05	55.88	92.16	**93.01**
Railway	96.87	79.65	97.60	94.15	86.76	77.06	95.97	**99.33**	95.47	94.32
Parking lot 1	86.55	91.26	88.09	94.72	90.59	88.76	98.46	**99.71**	85.19	86.15
Parking lot 2	53.68	81.05	91.93	89.12	92.28	87.72	90.88	88.77	94.21	**94.28**
Tennis court	98.79	**100.0**	**100.0**	**100.0**	95.14	97.17	**100.0**	**100.0**	**100.0**	**100.0**
Running track	98.10	98.73	**100.0**	**100.0**	**100.0**	98.52	**100.0**	**100.0**	**100.0**	**100.0**
OA	82.91	85.32	89.56	88.59	87.95	87.42	90.66	89.81	92.36	**92.57**
AA	84.38	87.56	91.50	90.63	89.81	89.37	92.07	91.33	93.62	**93.81**
Kappa	81.43	84.12	88.69	87.64	86.96	86.38	89.86	88.93	91.73	**91.96**

and the worst-group gap ranges from 0.036 to 0.047. These empirical group-wise deviations are consistently smaller than those of Split-CP (Angelopoulos & Bates, 2023), CRC (Angelopoulos et al., 2024), and RCPS (Bates et al., 2021) under the same evaluation protocol.

Under class-resampling and tile-removal perturbations, RC-Mamba also exhibits the smallest coverage gaps and the smallest increases in average prediction-set size among the evaluated methods. These results support improved reliability stability under the tested shifts, without implying a universal distribution-shift guarantee. Finally, the Dirichlet–uniform experiment supports the implementation consistency of the Bayesian-quadrature estimator used by G-BRT with the original CPBQ construction (Snell & Griffiths, 2025).



Figure 8: Classification results of different methods for the Houston2018 dataset.

Table 9: Uncertainty-quantification dashboard. All numbers are reported at $\alpha = 0.10$ unless otherwise stated. The no-shift blocks report marginal reliability and empirical group-wise coverage behavior on four benchmarks; $\sup|\mathrm{Cov} - (1-\alpha)|$ denotes the maximum absolute deviation over the swept α values. The shift block reports Cov-gap and the increase in average prediction-set size under class resampling $s = 0.2$ and tile removal $r_{\mathrm{drop}} = 0.5$.

Method	**Marginal (no shift)**				**Empirical group-wise (no shift)**				**Shift:** $s = 0.2,\ r_{\mathrm{drop}} = 0.5$			
	Berlin	Augsburg	H2013	H2018	Berlin	Augsburg	H2013	H2018	Berlin	Augsburg	H2013	H2018
	$\sup\lvert\Delta\rvert$	$\sup\lvert\Delta\rvert$	$\sup\lvert\Delta\rvert$	$\sup\lvert\Delta\rvert$	std(Cov_g)	std(Cov_g)	std(Cov_g)	std(Cov_g)	Cov-gap	Cov-gap	Cov-gap	Cov-gap
MSFMamba + Split-CP	0.013	0.010	0.013	0.011	0.063	0.055	0.048	0.072	0.082	0.069	0.064	0.087
MSFMamba + CRC	0.011	0.008	0.010	0.009	0.041	0.037	0.032	0.048	0.043	0.036	0.031	0.047
MSFMamba + RCPS	0.009	0.007	0.008	0.007	0.038	0.034	0.029	0.044	0.061	0.051	0.046	0.063
RC-Mamba (Ours)	**0.004**	**0.005**	**0.004**	**0.007**	**0.019**	**0.023**	**0.018**	**0.021**	**0.021**	**0.019**	**0.017**	**0.024**
	$\overline{\lvert\mathcal{C}\rvert}$	$\overline{\lvert\mathcal{C}\rvert}$	$\overline{\lvert\mathcal{C}\rvert}$	$\overline{\lvert\mathcal{C}\rvert}$	worst-gap	worst-gap	worst-gap	worst-gap	$\Delta\overline{\lvert\mathcal{C}\rvert}$	$\Delta\overline{\lvert\mathcal{C}\rvert}$	$\Delta\overline{\lvert\mathcal{C}\rvert}$	$\Delta\overline{\lvert\mathcal{C}\rvert}$
MSFMamba + Split-CP	1.38	1.19	1.27	1.44	0.158	0.134	0.115	0.172	+0.48	+0.39	+0.35	+0.52
MSFMamba + CRC	1.31	1.14	1.24	1.38	0.102	0.091	0.081	0.118	+0.36	+0.29	+0.27	+0.41
MSFMamba + RCPS	1.26	1.12	1.22	1.33	0.089	0.081	0.074	0.105	+0.35	+0.30	+0.28	+0.43
RC-Mamba (Ours)	**1.08**	**1.05**	**1.04**	**1.11**	**0.041**	**0.038**	**0.036**	**0.047**	**+0.12**	**+0.10**	**+0.09**	**+0.14**

Note. $\sup|\Delta| = \sup_\alpha |\mathrm{Cov}(\alpha) - (1-\alpha)|$ over the swept α values; worst-gap $= \max_g |\mathrm{Cov}_g - (1-\alpha)|$; $\Delta\overline{|\mathcal{C}|}$ denotes the increase in average prediction-set size relative to the corresponding no-shift value. All values are averaged over three independent runs.

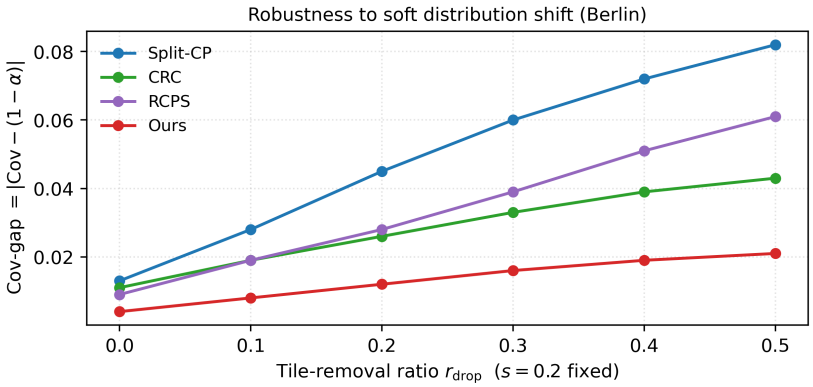


Figure 9: Coverage gap $|\mathrm{Cov} - (1-\alpha)|$ on Berlin under controlled soft distribution shift. The tile-removal ratio r_{drop} is varied while the class-resampling severity is fixed at $s = 0.2$. RC-Mamba shows the slowest increase in coverage error as the perturbation magnitude grows.

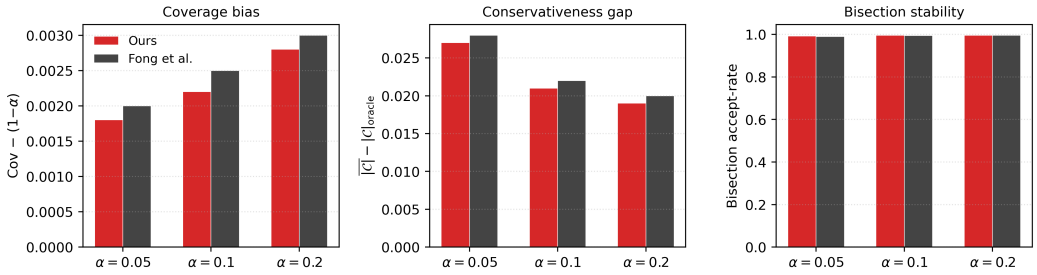


Figure 10: Verification of the Bayesian-quadrature estimator against the Dirichlet–uniform construction used in CPBQ (Snell & Griffiths, 2025) with $\beta = 0.95$ and $M = 1000$. Across $\alpha \in \{0.05, 0.10, 0.20\}$, the reproduced estimator exhibits close agreement with the original construction on the evaluated statistics.

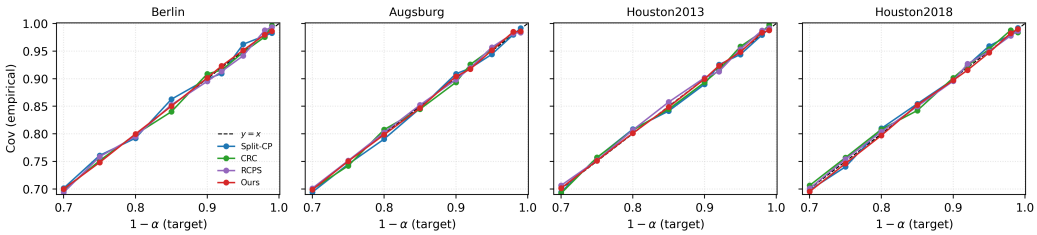


Figure 11: Marginal reliability curves of empirical coverage Cov versus target coverage $1 - \alpha$ on four multi-source benchmarks. The dashed line denotes $y = x$; RC-Mamba remains closer to the target coverage than the evaluated post-hoc conformal baselines.

C.5 Computational Accounting

The three components of RC-Mamba introduce lightweight and selective computational overhead on top of the Mamba backbone. G-BRT updates group-specific thresholds only at the predefined refresh interval using the training-time threshold cache, so its cost is amortized over the optimization process. TCMR introduces only a threshold-coupled scalar margin term into the classification

objective, while RASR performs additional flipped-view evaluations only for the residual high-risk subset $\mathcal{H}$.

Let $\rho_{\mathcal{H}} = |\mathcal{H}|/N_{\mathrm{te}}$ and let $c_{\mathcal{R}}$ denote the additional refinement cost for one selected sample. The computation associated with selective refinement can therefore be written as

$$\mathrm{FLOPs}_{\mathrm{total}} = N_{\mathrm{te}}\mathrm{FLOPs}_0 + |\mathcal{H}|c_{\mathcal{R}} = N_{\mathrm{te}}\mathrm{FLOPs}_0\,(1 + \rho_{\mathcal{H}}\kappa)\,, \qquad \kappa = \frac{c_{\mathcal{R}}}{\mathrm{FLOPs}_0}.$$

Hence, the additional computation introduced by RC-Mamba is governed by periodic threshold updates, a lightweight margin term, and selective refinement of residual high-risk samples, rather than by an unconditional increase in backbone computation for all samples. This analytical accounting characterizes the computational structure of RC-Mamba and is not presented as a measured latency or memory benchmark.