CRED: CLASS-SPECIFIC REASONING OVER TYPED EVIDENCE WITH RELIABILITY-HARDNESS PRIORITIZED KNOWLEDGE DISTILLATION FOR MULTISOURCE REMOTE SENSING IMAGE CLASSIFICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

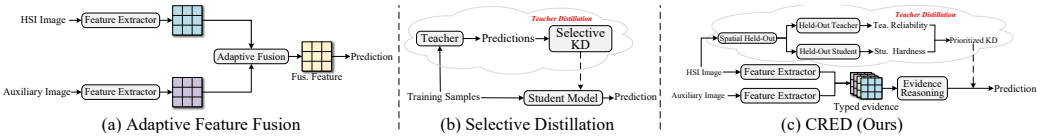
Joint classification of hyperspectral imagery (HSI) and LiDAR/SAR data demands the effective exploitation of complementary information across heterogeneous sensors. However, existing fusion methods primarily model modality-specific features or fused representations, while conventional distillation methods directly use teacher predictions as supervision, making it difficult to identify discriminative information and effective distillation signals for different classes and samples. To address these limitations, we propose CRED, a framework for Typed Reasoning over Multimodal Evidence with Reliability- and Hardness-Prioritized Knowledge Distillation. Specifically, Typed Multimodal Evidence Construction (TMEC) separately encodes spectral, spatial, and complementary modality information, organizes heterogeneous sensor representations into explicitly typed evidence, rather than directly collapsing them into a single fused feature representation. Class-Specific Typed Evidence Budgeting (CSTEB) models interactions among evidence types using Mamba, employs type-specific class queries to assess their support for each candidate class, and combines the resulting scores using class- and sample-specific weights. Finally, Reliability- and Hardness-Prioritized Distillation (RH-PD) constructs teacher reliability and student hardness indicators from spatial-block held-out prediction probabilities, and prioritizes distillation samples accordingly, transferring trustworthy teacher knowledge to difficult samples that benefit most from additional supervision. Experiments on three public HSI–LiDAR/SAR datasets demonstrate that CRED consistently outperforms existing methods across multiple evaluation metrics, achieving OA values of 94.06%, 80.60%, and 92.77% on Houston2013, Berlin, and Augsburg, respectively.

1 INTRODUCTION

Hyperspectral imagery (HSI) provides detailed spectral responses, whereas LiDAR and SAR offer complementary elevation, structural, and scattering information for land-cover classification (Gómez-Chova et al., 2015; Ghamisi et al., 2015). Multimodal learning seeks to exploit this complementarity through joint representation learning (Hong et al., 2021). However, information sources vary in their relevance to different classes and samples. Likewise, teacher predictions differ in reliability, and samples do not require equal correction during knowledge transfer. The central problem is therefore how to identify decision-relevant information for both classification and distillation.

Existing multimodal classifiers improve heterogeneous feature extraction and interaction through feature enhancement, asymmetric fusion, Transformer architectures, and state-space models (Fang et al., 2022; Li et al., 2023; Zhao et al., 2023; 2024; Zhang et al., 2025; Gao et al., 2025). Recent uncertainty-aware and adaptive approaches further regulate modality contributions according to observation quality (Han et al., 2023; Gao et al., 2024; Cho et al., 2025; Chen et al., 2025; Wang et al., 2025b; Chu et al., 2027). Nevertheless, these methods generally reason over modality fea-

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.



(a) Adaptive Feature Fusion (b) Selective Distillation (c) CRED (Ours)

Figure 1: Comparison of representative paradigms. (a) Adaptive multimodal fusion regulates modality contributions (Wang et al., 2025b; Cho et al., 2025; Chu et al., 2027). (b) Selective distillation adapts knowledge transfer to categories or samples (Shi et al., 2025; Zheng et al., 2026). (c) CRED performs typed evidence reasoning during prediction and reliability–hardness-prioritized knowledge distillation during training.

tures or their fused representations, as illustrated in Fig. 1(a). Consequently, the support provided by individual information types for each candidate class remains implicit.

Knowledge distillation transfers predictive knowledge from a teacher to a student (Hinton et al., 2015). In HSI classification, spatial dependence can place strongly correlated samples in both fitting and prediction sets under random partitioning (Liang et al., 2017). Cross-fitting constructs predictions from models that exclude the corresponding samples during fitting, mitigating teacher overfitting caused by data reuse (Dao et al., 2021); however, generic cross-fitting does not explicitly preserve spatial separation when constructing teacher targets. Recent methods improve knowledge transfer through student-oriented teacher training and adaptation (Dong et al., 2024; Qian et al., 2025), category-selective transfer and reliability-based sample screening (Chen et al., 2024; Shi et al., 2025; Kong et al., 2026), and student-compatible teacher selection or hardness-aware distillation (Panigrahi et al., 2026; Zheng et al., 2026). Despite these advances, existing methods still do not simultaneously account for spatial dependence in teacher-target construction and the joint effects of teacher reliability and student learning difficulty, as summarized in Fig. 1(b).

To address these limitations, we propose **CRED**, a framework for Typed Reasoning over Multimodal Evidence with Reliability- and Hardness-Prioritized Knowledge Distillation. As shown in Fig. 1(c), CRED connects prediction-time evidence reasoning with training-time supervision selection. Specifically, TMEC transforms heterogeneous observations into typed evidence representations, CSTEB enables class-aware reasoning over different evidence types, and RH-PD selectively transfers reliable teacher knowledge according to teacher reliability and student hardness. During training, RH-PD leverages spectral and spatial HyperSIGMA teachers (Wang et al., 2025a) to construct reliable distillation targets. Experiments on three public HSI–LiDAR/SAR datasets demonstrate consistent improvements over recent multimodal classification methods.

Our contributions are summarized as follows:

- **Typed Multimodal Evidence Construction.** TMEC constructs spectral, spatial, auxiliary, and cross-modal evidence. The spectral encoder models center responses, local context, and first- and second-order spectral variations, while the cross-modal relation encoder captures discrepancy, joint response, and similarity between aligned spatial and auxiliary features. By preserving the identity of heterogeneous information sources, TMEC exposes complementary decision cues for subsequent evidence reasoning.

- **Class-Specific Typed Evidence Budgeting.** We delay evidence aggregation until class-dependent relevance can be assessed. It first contextualizes the typed evidence through cross-type interaction and then uses type-specific class queries to estimate how strongly spectral, spatial, and auxiliary evidence support each candidate class. Cross-modal relational evidence further conditions sample- and class-specific contribution weights, enabling the final prediction to depend on both class-specific evidence support and sample-dependent evidence relevance.

- **Reliability- and Hardness-Prioritized Distillation.** We constructs teacher and student predictions through spatial-block held-out training. RH-PD constructs teacher reliability and student learning difficulty indicators based on prediction probabilities, and jointly determines the distillation priority accordingly, so that trustworthy teacher knowledge is preferentially transferred to difficult samples that the student has not yet sufficiently mastered.

2 Related Work

Multimodal Hyperspectral Classification and Adaptive Fusion. Multimodal HSI classification has evolved from modality-specific enhancement and asymmetric fusion toward global cross-modal modeling. S^2ENet enhances spatial and spectral information across modalities, while HCT and AsyFFNet introduce hierarchical Transformer interaction and asymmetric heterogeneous fusion (Fang et al., 2022; Zhao et al., 2023; Li et al., 2023). FrIT subsequently combines local feature extraction with global contextual modeling (Zhao et al., 2024). More recent methods employ mutual guidance, multiscale state-space modeling, and representative spectral selection to strengthen cross-source interaction (Zhang et al., 2025; Gao et al., 2025; Gong et al., 2026). TCMax further addresses modality competition by maximizing total correlation between multimodal features and labels (Yu et al., 2026). Together, these approaches improve how modality features are extracted, exchanged, and fused. A related line adapts multimodal integration according to observation quality. Trusted multi-view classification represents class-level uncertainty through evidential distributions, while EAU incorporates unimodal aleatoric uncertainty into robust fusion (Han et al., 2023; Gao et al., 2024). Cocoon and HyperDUM extend uncertainty estimation to object- and feature-level multimodal representations (Cho et al., 2025; Chen et al., 2025). UDFNet develops uncertainty-aware dynamic fusion, trusted robustness filters unreliable evidence, and UGMFT models feature and fusion uncertainty through fuzzy membership (Wang et al., 2025b; Liu et al., 2025; Chu et al., 2027).

Knowledge Distillation and Selective Teacher Supervision. Knowledge distillation transfers predictive distributions from a teacher to a student (Hinton et al., 2015). Cross-fitting constructs teacher predictions from models that exclude the corresponding samples during fitting, mitigating overfitting caused by data reuse (Dao et al., 2021). In remote sensing, cross-modal graph distillation transfers relational knowledge between heterogeneous sources (Wang et al., 2023). Subsequent methods improve teacher supervision through student-oriented training and reliability- or informativeness-based sample selection (Dong et al., 2024; Chen et al., 2024). SpectralGPT and HyperSIGMA additionally provide pretrained spectral and spatial–spectral representations that can serve as transferable knowledge sources (Hong et al., 2024; Wang et al., 2025a). Recent methods increasingly adapt transferred knowledge to particular modalities, categories, teachers, or samples. DPMamba addresses missing-modality remote-sensing classification, G^2D alleviates modality imbalance through gradient-guided distillation, and SelKD selects category-specific knowledge (Yang et al., 2025; Rakib & Bagavathi, 2025; Shi et al., 2025). Teacher adaptation and GRACES respectively improve the compatibility and selection of teachers for a given student (Qian et al., 2025; Panigrahi et al., 2026). For multimodal applications, MPTMKD selects reliable samples and refines teacher targets in HSI–LiDAR classification, while PMKD incorporates sample hardness into multispectral distillation (Kong et al., 2026; Zheng et al., 2026).

3 Method

3.1 Problem Formulation and Framework Overview

For the i-th input sample, given an HSI patch X_i^h, a spatially co-registered auxiliary patch X_i^a, and its center-pixel label $y_i \in \{1, \ldots, C\}$, CRED predicts the class logits $\mathbf{z}_i = f_\theta(X_i^h, X_i^a) \in \mathbb{R}^C$. The auxiliary modality is LiDAR or SAR. The original HSI is used for spectral modeling, while its PCA-reduced counterpart X_i^p represents spatial structure.

As shown in Fig. 2, the CRED student first uses Typed Multimodal Evidence Construction (TMEC) to construct typed multimodal evidence. Class-Specific Typed Evidence Budgeting (CSTEB) then models cross-type interactions and determines class- and sample-specific evidence contributions. During training, Reliability- and Hardness-Prioritized Distillation (RH-PD) constructs spatially held-out predictions from complementary spectral and spatial teachers and the CRED student, and prioritizes distillation according to teacher reliability and student-estimated sample hardness. RH-PD is removed after training, leaving only TMEC and CSTEB for inference.

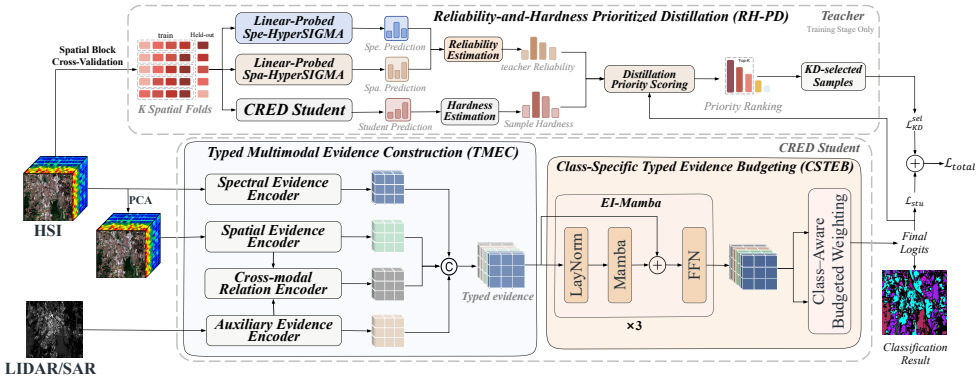


Figure 2: Overview of CRED. TMEC and CSTEB construct and reason over typed multimodal evidence, while training-only RH-PD builds spatially held-out predictions and prioritizes distillation according to teacher reliability and sample hardness.

3.2 Typed Multimodal Evidence Construction

TMEC constructs spectral, spatial, auxiliary, and cross-modal relation evidence while preserving their distinct information types. The first three provide complementary discriminative information, while relation evidence characterizes cross-modal agreement and discrepancy to inform subsequent reasoning, as illustrated in Fig. 3(a).

Spectral evidence (E_i^{spec}). Given the original HSI patch X_i^h, TMEC combines the center spectrum with its local mean and spectral differences:

$$E_i^{\text{spec}} = \text{MLP}_{\text{spec}}\big(\text{Concat}\left(\bar{\mathbf{s}}_i,\, \mathbf{s}_i,\, \Delta\mathbf{s}_i,\, \Delta^2\mathbf{s}_i\right)\big)\,. \tag{1}$$

Here, $\mathbf{s}_i$ and $\bar{\mathbf{s}}_i$ denote the center and local-mean spectra, respectively, while $\Delta\mathbf{s}_i$ and $\Delta^2\mathbf{s}_i$ denote the first- and second-order spectral differences.

Spatial evidence (E_i^{spa}). Given the PCA-HSI patch X_i^p, TMEC combines pooled spatial responses with neighborhood context obtained through masked pooling:

$$F_i^s = \text{Conv}_{3\times3}(X_i^p), \quad E_i^{\text{spa}} = \text{Concat}(\text{AvgPool}(F_i^s), \text{MaskedAvgPool}(F_i^s))\,. \tag{2}$$

Masked average pooling aggregates responses within selected neighborhood regions, preserving contextual information relative to the target pixel.

Auxiliary evidence (E_i^{aux}). For the aligned auxiliary patch X_i^a, let $\bar{X}_i^a = \text{AvgPool}(X_i^a)$ denote its local mean and $V_i^a = (X_i^a - \bar{X}_i^a)^2$ denote the local variation. The auxiliary evidence is constructed as

$$F_i^a = \text{Conv}_{3\times3}(\text{Concat}(X_i^a, V_i^a))\,, \quad E_i^{\text{aux}} = \text{AvgPool}(F_i^a). \tag{3}$$

The squared deviation V_i^a characterizes local variation, while the final pooling aggregates the encoded auxiliary responses.

Cross-modal relational evidence (E_i^{rel}). Using the aligned spatial and auxiliary feature maps F_i^s and F_i^a defined above, TMEC models their cross-modal relationships through agreement and discrepancy cues:

$$C_i = \text{Concat}(F_i^s, F_i^a, |F_i^s - F_i^a|\,, F_i^s \odot F_i^a, \cos(F_i^s, F_i^a))\,, \quad E_i^{\text{rel}} = \sigma(\text{Conv}_{1\times1}(C_i))\,. \tag{4}$$

Here, $\odot$ denotes element-wise multiplication, cos denotes channel-wise cosine similarity, and σ is the sigmoid function. The resulting relational representation captures cross-modal consistency and discrepancy patterns for subsequent cross-type interaction.

TMEC finally yields the typed evidence set

$$\mathcal{E}_i = \left\{E_i^{\text{spec}}, E_i^{\text{spa}}, E_i^{\text{aux}}, E_i^{\text{rel}}\right\}, \tag{5}$$

which is passed to CSTEB for cross-type interaction and class-specific contribution assessment.

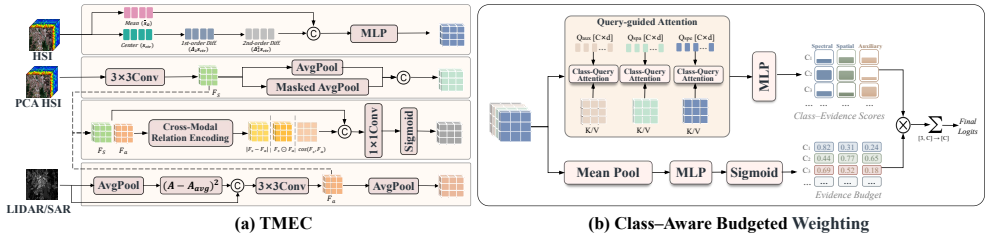


Figure 3: Detailed architectures of (a) TMEC and (b) the Class-Aware Budgeted Weighting block within CSTEB. TMEC constructs four typed evidence representations. CSTEB combines query-guided class–evidence scores with class- and sample-specific evidence budgets (CABW). Routine projection and tokenization operations are omitted for clarity.

3.3 Class-Specific Typed Evidence Budgeting

TMEC preserves heterogeneous information as distinct evidence types, but independent construction neither captures their dependencies nor determines their relevance to different classes. CSTEB first contextualizes the typed evidence with Evidence Integration Mamba (EI-Mamba) and then applies Class-Aware Budgeted Weighting to evaluate and combine it. The evidence-type dimension is retained until the final prediction.

Cross-type evidence contextualization. Let $\mathcal{E}_i = \{E_i^{\text{spec}}, E_i^{\text{spa}}, E_i^{\text{aux}}, E_i^{\text{rel}}\}$ denote the typed evidence constructed by TMEC. After adding type embeddings, the evidence tokens are processed by EI-Mamba:

$$\widetilde{\mathcal{E}}_i = \text{EI-Mamba}\left(\mathcal{E}_i\right). \tag{6}$$

As shown in Fig. 2, each EI-Mamba block comprises normalization, a Mamba layer, residual aggregation, and a feed-forward network. Stacking three blocks propagates information across evidence types while preserving their type assignments, producing contextualized spectral, spatial, auxiliary, and relation evidence.

Class-specific evidence scoring. The upper branch of Fig. 3(b) evaluates how strongly each discriminative evidence type supports every candidate class. For each $t \in \mathcal{T} = \{\text{spec}, \text{spa}, \text{aux}\}$, CSTEB learns a type-specific query matrix $\mathbf{Q}^t \in \mathbb{R}^{C\times d}$, whose rows represent the queries of the C classes. The interacted evidence tokens of type t serve as keys and values:

$$\mathbf{A}_i^t = \text{Attn}\left(\mathbf{Q}^t, \widetilde{\mathbf{E}}_i^t, \widetilde{\mathbf{E}}_i^t\right), \quad \mathbf{S}_i^t = \text{MLP}_t\left(\mathbf{A}_i^t\right), \quad \mathbf{S}_i = \text{Concat}_{t\in\mathcal{T}}\left(\mathbf{S}_i^t\right) \in \mathbb{R}^{C\times 3}. \tag{7}$$

Here, $\mathbf{A}_i^t \in \mathbb{R}^{C\times d}$ contains the class-conditioned evidence retrieved from type t, and $\mathbf{S}_i^t \in \mathbb{R}^C$ contains its support scores for the C candidate classes. Concatenating the three score vectors along the evidence-type dimension produces the class–evidence score matrix $\mathbf{S}_i$.

Class- and sample-specific evidence weighting. A high evidence score does not imply that the corresponding evidence type should contribute equally to every class and sample. The lower branch of Fig. 3(b) therefore integrates the contextualized typed evidence and predicts an evidence budget. The final logit for class c is obtained by applying the corresponding budget to the three evidence scores:

$$\mathbf{B}_i = \text{MLP}_B\left(\text{MeanPool}\left(\widetilde{\mathcal{E}}_i\right)\right) \in \mathbb{R}^{C\times 3}, \quad z_{i,c} = \sum_{t\in\mathcal{T}} S_{i,c}^t B_{i,c}^t, \quad \mathbf{z}_i \in \mathbb{R}^C. \tag{8}$$

Because $\mathbf{B}_i$ is inferred from sample i and contains one row for each candidate class, it provides class- and sample-specific contribution weights for spectral, spatial, and auxiliary evidence. The integrated representation also includes cross-modal relation evidence, allowing modality consistency and discrepancy patterns to inform evidence allocation. CSTEB therefore delays aggregation until each evidence type has been contextualized and evaluated with respect to candidate classes. This design enables the final prediction to depend on both class-specific evidence support and sample-specific evidence demand.

3.4 RELIABILITY- AND HARDNESS-PRIORITIZED DISTILLATION

Foundation-model teachers vary in reliability, while samples differ in their need for correction. RH-PD constructs reliable teacher targets and prioritizes distillation using teacher support and student difficulty. As shown in Fig. 2, it comprises spatially held-out prediction construction, teacher fusion, hardness estimation, and priority scoring. Here, j indexes labeled samples, HO denotes held-out predictions, and S denotes the CRED student.

Spatial-block held-out prediction construction. To limit spatial dependence between fitting and prediction samples, RH-PD partitions the labeled set $\mathcal{D}$ into K spatial folds $\{\mathcal{D}_k\}_{k=1}^{K}$ while preserving class coverage. For each fold, linear heads attached to the spectral and spatial HyperSIGMA encoders (Wang et al., 2025a), together with the CRED student, are fitted on $\mathcal{D} \setminus \mathcal{D}_k$ and evaluated on $\mathcal{D}_k$:

$$\mathbf{p}_j^{q,\mathrm{HO}} = \mathrm{softmax}\left(\mathbf{z}_j^{q,\mathrm{HO}}\right), \quad q \in \{\mathrm{spec}, \mathrm{spa}, S\}, \quad j \in \mathcal{D}_k, \tag{9}$$

where $\mathbf{z}_j^{q,\mathrm{HO}}$ denotes the logits produced by model q without fitting on $\mathcal{D}_k$. Repeating this process provides held-out teacher and student predictions for every labeled sample, avoiding evaluation by models fitted on the corresponding fold. These folds construct training signals rather than select hyperparameters.

Reliability Estimation. The spectral and spatial teachers provide complementary but sample-dependent supervision. The Prediction Fusion block weights their held-out predictions by reliability. For $m \in \{\mathrm{spec}, \mathrm{spa}\}$,

$$r_j^m = p_{j,y_j}^{m,\mathrm{HO}} \sqrt{\mu_j^{m,\mathrm{HO}} \left(1 - \bar{H}_j^{m,\mathrm{HO}}\right)}, \quad \mathbf{p}_j^T = \frac{\sum_m r_j^m \mathbf{p}_j^{m,\mathrm{HO}}}{\sum_m r_j^m}, \tag{10}$$

where $p_{j,y_j}^{m,\mathrm{HO}}$, $\mu_j^{m,\mathrm{HO}}$, and $\bar{H}_j^{m,\mathrm{HO}}$ denote ground-truth probability, top-two probability margin, and normalized entropy, respectively. Thus, the more reliable teacher contributes more to the fused target $\mathbf{p}_j^T$, whose overall reliability is $r_j^T = \frac{1}{2}\sum_m r_j^m$.

Hardness Estimation. Teacher reliability determines whether supervision is trustworthy, whereas student hardness determines whether it is needed. The Hardness Estimation block derives a sample-hardness prior from the held-out student prediction:

$$m_j^{S,\mathrm{HO}} = z_{j,y_j}^{S,\mathrm{HO}} - \max_{c \neq y_j} z_{j,c}^{S,\mathrm{HO}}, \quad h_j^{\mathrm{HO}} = \lambda_e \mathbb{I}\left[\hat{y}_j^{S,\mathrm{HO}} \neq y_j\right] + 1 - p_{j,y_j}^{S,\mathrm{HO}} + \max\left(0, 1 - \frac{m_j^{S,\mathrm{HO}}}{\tau_m}\right). \tag{11}$$

Here, $\hat{y}_j^{S,\mathrm{HO}} = \arg\max_c p_{j,c}^{S,\mathrm{HO}}$ is the held-out predicted class, λ_e weights prediction error, and τ_m controls the margin scale. The three terms measure prediction error, insufficient ground-truth confidence, and insufficient class separation. A larger h_j^{HO} indicates greater student difficulty.

Distillation priority scoring. RH-PD combines held-out hardness with the student's current learning state. It forms $\mathcal{C}_{\mathrm{KD}} = \{j : r_j^T \geq \tau_r,\, p_{j,y_j}^T \geq \tau_p\}$, where τ_r and τ_p are the minimum reliability and ground-truth support thresholds. Let $\mathbf{p}_j^S = \mathrm{softmax}(\mathbf{z}_j^S)$ denote the current student prediction. Eligible samples are ranked by

$$\pi_j = \left(1 - p_{j,y_j}^S\right) + \max\left(0, \frac{m_T - m_j^{S,\mathrm{cur}}}{\tau_m}\right) + p_{j,y_j}^T + \lambda_h h_j^{\mathrm{HO}}, \quad \mathcal{S}_{\mathrm{KD}} = \mathrm{TopK}_{j \in \mathcal{C}_{\mathrm{KD}}}\left(\pi_j\right), \tag{12}$$

where $m_j^{S,\mathrm{cur}} = z_{j,y_j}^S - \max_{c \neq y_j} z_{j,c}^S$ is the current student margin, m_T is the target margin, and λ_h controls the hardness contribution. The four terms capture current confidence deficiency, margin deficiency, teacher support, and held-out hardness, respectively. Thus, $\mathcal{S}_{\mathrm{KD}}$ selects samples with trustworthy targets and substantial correction needs.

3.5 TRAINING OBJECTIVE AND INFERENCE

The student is trained on all labeled samples with the supervised objective $\mathcal{L}_{\mathrm{stu}}$, while RH-PD applies distillation only to the selected set $\mathcal{S}_{\mathrm{KD}}$ using the fused teacher predictions $\mathbf{p}_j^T$ as targets.

The overall objective is

$$\mathcal{L} = \mathcal{L}_{\text{stu}} + \eta(e)\mathcal{L}_{\text{KD}}, \tag{13}$$

where $\mathcal{L}_{\text{KD}}$ is evaluated on $\mathcal{S}_{\text{KD}}$, and $\eta(e)$ controls its contribution at training epoch e.

At inference, RH-PD is removed together with the spatial folds, foundation teachers, held-out prediction banks, and sample-priority components, leaving only TMEC and CSTEB for prediction.

4 Experiments

4.1 Experimental Setup

Datasets and protocols. We evaluate CRED on three public multimodal remote-sensing benchmarks: Houston2013 with paired HSI–LiDAR observations, and Berlin and Augsburg with paired HSI–SAR observations. Following MSFMamba (Gao et al., 2025) and RSCNet (Gong et al., 2026), we use the established training–test partitions and keep the sample coordinates fixed across all comparison, ablation, and sensitivity experiments. Table 1 summarizes the dataset statistics.

Table 1: Statistics of the three multimodal remote-sensing datasets.

Dataset	Modalities	Spatial size	Bands	Classes	Train	Test
Houston2013	HSI+LiDAR	349×1905	144	15	2,832	12,197
Berlin	HSI+SAR	1723×476	244	8	2,820	461,851
Augsburg	HSI+SAR	332×485	180	7	761	77,533

Evaluation protocol. We report overall accuracy (OA), average accuracy (AA), macro-averaged F1 score (Macro-F1), and Cohen's kappa coefficient (κ). All methods use identical data partitions, evaluation criteria, and checkpoint-selection criteria. For publicly available implementations, we retain the original architectures and recommended optimization settings while adapting their data interfaces to the common evaluation pipeline. Because architectures converge at different rates, each method is trained with a sufficient training horizon. The same protocol is used for all ablations.

Training configuration. CRED is optimized using AdamW with a batch size of 96, an initial learning rate of 10^{-4}, and a weight decay of 5×10^{-4}. The learning rate is linearly warmed up for four epochs and then decayed using a cosine schedule. Training runs for at most 100 epochs on Houston2013 and Augsburg and 70 epochs on Berlin. RH-PD uses five spatial folds to construct held-out predictions.

4.2 Comparison with State-of-the-Art Methods

We compare CRED with nine representative multimodal remote-sensing classification methods, including S^2ENet (Fang et al., 2022), HCT (Zhao et al., 2023), AsyFFNet (Li et al., 2023), FrIT (Zhao et al., 2024), MGMNet (Zhang et al., 2025), MSFMamba (Gao et al., 2025), MSAF (Shi et al., 2026), RSCNet (Gong et al., 2026), and T-APT (Gao et al., 2026). These baselines cover cross-modal feature enhancement, attention- and Transformer-based fusion, state-space modeling, and adaptive multimodal learning. Table 2 reports OA, AA, Macro-F1, and κ on Houston2013, Berlin, and Augsburg. The best and second-best results are highlighted in bold and underlined, respectively.

The strongest baseline varies across datasets and metrics. On Houston2013, the leading baseline results are distributed among MGMNet, RSCNet, and T-APT. In comparison, CRED consistently achieves the best results across all four metrics, improving the strongest baseline by 0.27, 0.40, 0.57, and 0.30 percentage points in OA, AA, Macro-F1, and κ, respectively. Although several recent methods already approach 94% accuracy on this dataset, CRED retains consistent gains across both overall and class-averaged measures.

The improvement is most pronounced on Berlin. Compared with the strongest baseline for each metric, CRED improves OA, AA, Macro-F1, and κ by 2.42, 1.60, 3.77, and 4.40 percentage points, respectively. On Augsburg, where the baseline OA values are already high, CRED improves OA by 0.93 points while increasing Macro-F1 by 3.04 points. These results indicate that the gains

Table 2: Comparison with state-of-the-art methods on three datasets.

Dataset	Metric	S^2ENet GRSL'22	HCT TGRS'23	AsyFFNet TNNLS'23	FrIT TNNLS'24	MGMNet JSTARS'25	MSFMamba TGRS'25	MSAF TIP'26	RSCNet TGRS'26	T-APT AAAI'26	CRED Ours
Houston2013	OA	91.42	91.19	92.36	91.79	93.73	92.47	93.21	93.49	93.79	**94.06**
	AA	92.65	92.94	92.98	92.47	94.24	93.95	93.90	94.49	93.44	**94.89**
	Macro-F1	92.60	91.62	93.20	92.49	94.24	91.65	93.84	91.82	93.48	**94.81**
	κ	90.69	90.45	91.70	91.09	93.19	91.83	92.63	92.94	93.25	**93.55**
Berlin	OA	76.72	75.00	76.10	73.59	75.94	76.91	76.59	78.18	76.38	**80.60**
	AA	61.89	67.54	64.12	66.35	66.28	64.64	62.49	62.17	66.68	**69.14**
	Macro-F1	60.58	59.64	58.98	60.78	58.74	59.84	59.35	58.85	62.03	**65.80**
	κ	64.11	62.83	64.02	60.82	63.49	64.70	63.61	65.38	64.13	**69.78**
Augsburg	OA	90.76	91.84	90.05	89.06	89.19	90.82	89.90	91.53	91.58	**92.77**
	AA	66.20	65.08	69.15	69.18	66.20	62.82	66.14	65.77	66.83	**70.67**
	Macro-F1	65.99	63.18	63.70	62.70	62.74	63.94	65.70	64.75	67.40	**70.44**
	κ	86.62	88.20	85.82	84.48	84.52	86.56	85.51	87.81	87.94	**89.63**

extend beyond overall accuracy to more balanced class-wise performance. The contributions of the individual components are examined in the following ablation studies.

4.3 Component Ablation and Evidence Reasoning Analysis

Table 3 evaluates the main components of CRED. The baseline directly fuses multimodal features. TMEC constructs four typed evidence representations, while EI-Mamba models their cross-type interactions. Class-Aware Budgeted Weighting (CABW) uses type-specific class queries to estimate the support of each evidence type for every candidate class and adaptively weights these scores for each sample. When CABW is disabled, the interacted evidence is combined using a fixed aggregation rule. RH-PD is applied only during training.

Table 3: Ablation study of the main components of CRED.

TMEC	EI-Mamba	CABW	RH-PD	Houston2013				Berlin				Augsburg			
				OA	AA	Macro-F1	κ	OA	AA	Macro-F1	κ	OA	AA	Macro-F1	κ
×	×	×	×	91.75	92.83	91.88	91.05	75.11	67.40	61.58	63.06	90.97	68.96	63.05	87.09
✓	×	×	×	92.70	93.95	92.86	92.08	78.44	66.57	62.74	67.04	92.49	69.46	66.95	89.27
✓	✓	×	×	93.01	94.44	92.78	92.42	77.36	67.14	62.17	65.57	92.41	67.86	67.52	89.11
✓	✓	✓	×	92.67	93.95	92.84	92.05	80.28	68.66	65.55	69.34	92.69	69.95	70.22	89.51
✓	✓	✓	✓	**94.06**	**94.89**	**94.81**	**93.55**	**80.60**	**69.14**	**65.80**	**69.78**	**92.77**	**70.67**	**70.44**	**89.63**

TMEC consistently improves OA over direct feature fusion, with gains of 0.95, 3.33, and 1.52 percentage points on Houston2013, Berlin, and Augsburg, respectively. This supports preserving heterogeneous multimodal information as typed evidence before cross-type reasoning instead of directly collapsing it into a unified representation.

The CSTEB ablation further separates cross-type interaction from class-aware evidence weighting. EI-Mamba alone produces mixed results, showing that evidence interaction does not by itself determine the contribution of each evidence type. On Houston2013, adding CABW retains performance close to the EI-Mamba configuration before RH-PD is introduced. On Berlin and Augsburg, CABW increases OA from 77.36% to 80.28% and Macro-F1 from 67.52% to 70.22%, respectively. These results reflect the complementary roles of the two stages: EI-Mamba establishes dependencies across evidence types, whereas CABW evaluates their class-specific support and adjusts their contributions for each sample.

RH-PD further improves all four metrics on all three datasets without changing the inference architecture. The effect is most pronounced on Houston2013, where OA, Macro-F1, and κ increase by 1.39, 1.97, and 1.50 percentage points, respectively. Since the student architecture remains unchanged, these gains isolate the effect of teacher-target construction and reliability- and hardness-based distillation prioritization. The following subsection examines these mechanisms separately.

4.4 Analysis of RH-PD

Teacher targets and distillation priorities. Table 4 evaluates teacher-target construction and priority criteria on Houston2013. *w/o Spatial Folds* uses predictions from teachers fitted on the same samples they predict, whereas full RH-PD uses spatially held-out teacher predictions.

Table 4: Analysis of teacher-target construction and distillation prioritization on Houston2013. All metrics are reported in percentage.

Teacher-target construction					Distillation prioritization				
Setting	OA	AA	Macro-F1	κ	Setting	OA	AA	Macro-F1	κ
Spe-HyperSIGMA	92.83	94.05	93.32	92.22	All-sample distillation	93.69	94.61	94.48	93.15
Spa-HyperSIGMA	93.58	94.40	94.46	93.03	w/o Teacher Reliability	92.50	93.69	92.48	91.87
w/o Spatial Folds	93.38	94.22	93.82	92.81	w/o Sample Hardness	93.03	94.08	93.38	92.44
Full model	**94.06**	**94.89**	**94.81**	**93.55**	Full model	**94.06**	**94.89**	**94.81**	**93.55**

Among the teacher-target construction strategies, both individual teachers and the no-fold setting remain inferior to the full model. Using only the spectral or spatial teacher achieves lower performance than the complete dual-teacher construction with spatially held-out predictions. Removing spatial folds also degrades performance, reducing OA from 94.06% to 93.38% and Macro-F1 from 94.81% to 93.82%. These results indicate that reliable teacher supervision benefits from both complementary teacher information and spatially held-out target construction.

For distillation prioritization, applying teacher supervision to all samples is inferior to the complete RH-PD strategy, reducing OA from 94.06% to 93.69%. Removing teacher reliability causes the largest degradation, reducing OA to 92.50%, while removing sample hardness lowers OA to 93.03%. These results demonstrate that teacher reliability and sample hardness provide complementary criteria for selecting informative distillation signals: reliability determines whether teacher knowledge is trustworthy, while hardness identifies samples that require further correction.

Effect of the number of spatial folds. We further vary the number of spatial folds $K \in \{2, 3, 5, 10\}$ on Houston2013. As shown in Fig. 4, Performance remains stable across different fold configurations, with $K = 5$ achieving the best results on all metrics. Increasing K beyond 5 provides no additional improvement. We therefore use $K = 5$ in the remaining experiments.

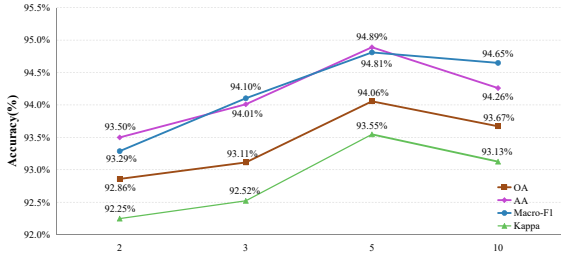


Figure 4: Effect of the number of spatial folds K on Houston2013 in terms of OA, AA, Macro-F1, and kappa.

5 Conclusion

We presented CRED for joint HSI–LiDAR/SAR classification, addressing how heterogeneous evidence should contribute to different decisions and which teacher signals should be transferred to the student. CRED constructs explicitly typed multimodal evidence with TMEC and uses CSTEB to determine its class- and sample-specific contributions. During training, RH-PD constructs spatially held-out teacher targets and prioritizes reliable supervision for samples that remain difficult for the student. Experiments on Houston2013, Berlin, and Augsburg demonstrate consistent improvements over existing methods, while the ablation studies support the respective roles of evidence construction, evidence budgeting, and prioritized distillation. RH-PD further improves the student without changing its inference architecture, parameter count, or peak inference memory. Future work will extend typed evidence modeling and selective knowledge transfer to broader modality combinations and larger remote-sensing scenes.

AI USE STATEMENT

Generative AI tools were used to assist with manuscript editing, translation, literature search, code and experiment-related assistance, and the preparation of figures and tables. All AI-assisted outputs were reviewed and verified by the authors. The experimental results, methodological claims, and conclusions reported in this paper were validated by the authors, who take full responsibility for the final content of the submission.

ETHICS STATEMENT

This work uses publicly available remote-sensing benchmark datasets and does not involve human subjects, personally identifiable information, or the collection of new sensitive data. The study focuses on methodological development and evaluation for multimodal remote-sensing image classification. We are not aware of additional ethical concerns specific to this work.

REPRODUCIBILITY STATEMENT

We provide the model architecture, training objectives, and complete algorithmic details in Sec. 3, together with the dataset protocols, evaluation metrics, and training configuration in Sec. 4.1. Additional implementation details, component analyses, OOF construction and sensitivity experiments, and supplementary quantitative and qualitative results are provided in the appendix. The submitted supplementary material also includes the source code, configuration files, and scripts required to reproduce the reported experiments.

REFERENCES

Luke Chen, Junyao Wang, Trier Mortlock, Pramod P. Khargonekar, and Mohammad Abdullah Al Faruque. Hyperdimensional uncertainty quantification for multimodal uncertainty fusion in autonomous vehicles perception. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22306–22316, 2025. doi: 10.1109/CVPR52734.2025.02078.

Yaofo Chen, Shuaicheng Niu, Yaowei Wang, Shoukai Xu, Hengjie Song, and Mingkui Tan. Towards robust and efficient cloud-edge elastic model adaptation via selective entropy distillation. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/b3b3e22dba07a71b0c0039af6af4d478-Abstract-Conference.html.

Minkyoung Cho, Yulong Cao, Jiachen Sun, Qingzhao Zhang, Marco Pavone, Jeong Joon Park, Heng Yang, and Zhuoqing Mao. Cocoon: Robust multi-modal perception with uncertainty-aware sensor fusion. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/6580845ce91a748a9bb09cec079aac1d-Abstract-Conference.html.

Yonghe Chu, Chi Wang, Jiangtao Peng, Weiping Ding, and Peng Li. UGMFT: Uncertainty guided multimodal fusion transformer for multi-source remote sensing classification. *Pattern Recognition*, 182:114812, 2027. doi: 10.1016/j.patcog.2026.114812.

Tri Dao, Govinda M. Kamath, Vasilis Syrgkanis, and Lester Mackey. Knowledge distillation as semiparametric inference. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=m4UCf24r0Y.

Chengyu Dong, Liyuan Liu, and Jingbo Shang. Toward student-oriented teacher network training for knowledge distillation. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/a1263ffa557506ea29c54481788d518f-Abstract-Conference.html.

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, 2022. doi: 10.1109/LGRS.2021.3121028.

Feng Gao, Xuepeng Jin, Xiaowei Zhou, Junyu Dong, and Qian Du. MSFMamba: Multiscale feature fusion state space model for multisource remote sensing image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 63:1–16, 2025. doi: 10.1109/TGRS.2025.3535622.

Qinghao Gao, Jiahui Qu, and Wenqian Dong. T-APT: Text-guided modality-aware prompt tuning for arbitrary multimodal remote sensing data joint classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 4158–4166, 2026. doi: 10.1609/aaai.v40i6.42411.

Zixian Gao, Xun Jiang, Xing Xu, Fumin Shen, Yujie Li, and Heng Tao Shen. Embracing unimodal aleatoric uncertainty for robust multimodal fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26866–26875, 2024. doi: 10.1109/CVPR52733.2024.02538.

Pedram Ghamisi, Jón Atli Benediktsson, and Stuart Phinn. Land-cover classification using both hyperspectral and LiDAR data. *International Journal of Image and Data Fusion*, 6(3):189–215, 2015. doi: 10.1080/19479832.2015.1055833.

Luis Gómez-Chova, Devis Tuia, Gabriele Moser, and Gustau Camps-Valls. Multimodal classification of remote sensing images: A review and future directions. *Proceedings of the IEEE*, 103(9):1560–1584, 2015. doi: 10.1109/JPROC.2015.2449668.

Chuanzheng Gong, Feng Gao, Junyan Lin, Junyu Dong, and Qian Du. Representative spectral correlation network for multisource remote sensing image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 64:1–14, 2026. doi: 10.1109/TGRS.2026.3684877.

Zongbo Han, Changqing Zhang, Huazhu Fu, and Joey Tianyi Zhou. Trusted multi-view classification with dynamic evidential fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2):2551–2566, 2023. doi: 10.1109/TPAMI.2022.3171983.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. URL `https://arxiv.org/abs/1503.02531`.

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, 2021. doi: 10.1109/TGRS.2020.3016820.

Danfeng Hong, Bing Zhang, Xuyang Li, Yuxuan Li, Chenyu Li, Jing Yao, Naoto Yokoya, Hao Li, Pedram Ghamisi, Xiuping Jia, Antonio Plaza, Paolo Gamba, Jón Atli Benediktsson, and Jocelyn Chanussot. SpectralGPT: Spectral remote sensing foundation model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(8):5227–5244, 2024. doi: 10.1109/TPAMI.2024.3362475.

Yi Kong, Shaocai Yu, Yuhu Cheng, C. L. Philip Chen, and Xuesong Wang. Multimodal prompt tuning with modified knowledge distillation for the joint classification of hyperspectral image and LiDAR data. *IEEE Transactions on Geoscience and Remote Sensing*, 64, 2026. doi: 10.1109/TGRS.2026.3699347. Art. no. 5515614.

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, 2023. doi: 10.1109/TNNLS.2022.3149394.

Jie Liang, Jun Zhou, Yuntao Qian, Lian Wen, Xiao Bai, and Yongsheng Gao. On the sampling strategy for evaluation of spectral-spatial methods in hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 55(2):862–880, 2017. doi: 10.1109/TGRS.2016.2616489.

Wei Liu, Yufei Chen, and Xiaodong Yue. Enhancing testing-time robustness for trusted multi-view classification in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15508–15517, 2025. doi: 10.1109/CVPR52734.2025.01445.

Abhishek Panigrahi, Bingbin Liu, Sadhika Malladi, Sham Kakade, and Surbhi Goel. In good GRACES: Principled teacher selection for knowledge distillation. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/59fa5d049b777d880fdcddba2b24738f-Abstract-Conference.html.

Chengyao Qian, Trung Le, and Mehrtash Harandi. A good teacher adapts their knowledge for distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1239–1248, 2025. doi: 10.1109/ICCV51701.2025.00123.

Mohammed Rakib and Arunkumar Bagavathi. G^2D: Boosting multimodal learning with gradient-guided distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4059–4068, 2025. doi: 10.1109/ICCV51701.2025.00387.

Liangliang Shi, Zhengyan Shi, and Junchi Yan. SelKD: Selective knowledge distillation via optimal transport perspective. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/cd687a58a13b673eea3fc1b2e4944cf7-Abstract-Conference.html.

Lulu Shi, Chunchao Li, Zhengchao Zeng, Puhong Duan, Behnood Rasti, and Antonio Plaza. Masked self-attention fusion network for joint classification of hyperspectral and LiDAR data. *IEEE Transactions on Image Processing*, 35:346–360, 2026. doi: 10.1109/TIP.2025.3648926.

Di Wang, Meiqi Hu, Yao Jin, Yuchun Miao, Jiaqi Yang, Yichu Xu, Xiaolei Qin, Jiaqi Ma, Lingyu Sun, Chenxing Li, Chuan Fu, Hongruixuan Chen, Chengxi Han, Naoto Yokoya, Jing Zhang, Minqiang Xu, Lin Liu, Lefei Zhang, Chen Wu, Bo Du, Dacheng Tao, and Liangpei Zhang. HyperSIGMA: Hyperspectral intelligence comprehension foundation model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(8):6427–6444, 2025a. doi: 10.1109/TPAMI.2025.3557581.

Hui Wang, Youxiang Huang, Hao Huang, Yu Wang, Jun Li, and Guan Gui. Uncertainty-aware dynamic fusion network with criss-cross attention for multimodal remote sensing land cover classification. *Information Fusion*, 123:103249, 2025b. doi: 10.1016/j.inffus.2025.103249.

Wenzhen Wang, Fang Liu, Wenzhi Liao, and Liang Xiao. Cross-modal graph knowledge representation and distillation learning for land cover classification. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–18, 2023. doi: 10.1109/TGRS.2023.3307604.

Yueguang Yang, Jiahui Qu, Ling Huang, and Wenqian Dong. DPMamba: Distillation prompt Mamba for multimodal remote sensing image classification with missing modalities. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pp. 2224–2232, 2025. doi: 10.24963/ijcai.2025/248.

Feng Yu, Xiangyu Wu, Yang Yang, and Jianfeng Lu. Multimodal classification via total correlation maximization. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/4ab56e2e7ea64ff654d28bc0fcaee108-Abstract-Conference.html.

Wei Zhang, Weiwei Song, Jie Wang, and Wen Gao. MGMNet: Mutual-guidance mechanism for joint classification of multisource remote sensing data. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 18:1085–1097, 2025. doi: 10.1109/JSTARS.2024.3502761.

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, 2023. doi: 10.1109/TGRS.2022.3232498. Art. no. 5500716.

Xudong Zhao, Mengmeng Zhang, Ran Tao, Wei Li, Wenzhi Liao, Lianfang Tian, and Wilfried Philips. Fractional fourier image transformer for multimodal remote sensing data classification. *IEEE Transactions on Neural Networks and Learning Systems*, 35(2):2314–2326, 2024. doi: 10.1109/TNNLS.2022.3189994.

Aihua Zheng, Pengyu Li, Zi Wang, and Jin Tang. Progressive multi-modal knowledge distillation for multi-spectral object re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 13351–13359, 2026. doi: 10.1609/aaai.v40i16.38338.

A Implementation and Reproducibility Details

A.1 Dataset Details and Preprocessing

We use the established training and test partitions of Houston2013, Berlin, and Augsburg. Table 5 reports their class-wise training and test distributions.

The original HSI is used for spectral evidence construction. For spatial evidence construction, PCA is applied to the HSI, and the first 30 whitened principal components are retained. The PCA features and auxiliary observations are standardized channel-wise. Horizontal and vertical random flipping is applied during training.

Table 5: Class-wise training and test sample distributions of the three multimodal remote-sensing datasets.

Houston2013

Class	Train	Test
Healthy grass	198	1,053
Stressed grass	190	1,064
Synthetic grass	192	505
Trees	188	1,056
Soil	186	1,056
Water	182	143
Residential	196	1,072
Commercial	191	1,053
Road	193	1,059
Highway	191	1,036
Railway	181	1,054
Parking lot 1	192	1,041
Parking lot 2	184	285
Tennis court	181	247
Running track	187	473
Total	**2,832**	**12,197**

Berlin

Class	Train	Test
Forest	443	54,511
Residential area	423	268,219
Industrial area	499	19,067
Low plants	376	58,906
Soil	331	17,095
Allotment	280	13,025
Commercial area	298	24,526
Water	170	6,502
Total	**2,820**	**461,851**

Augsburg

Class	Train	Test
Forest	146	13,361
Residential area	264	30,065
Industrial area	21	3,830
Low plants	248	26,609
Allotment	52	523
Commercial area	7	1,638
Water	23	1,507
Total	**761**	**77,533**

A.2 Model Configuration

CRED uses an evidence dimension of 192 on all three datasets. The spectral encoder produces 32 spectral tokens, while the spatial and auxiliary encoders organize their representations on a 5×5 token grid. All evidence types are projected to the same dimension before being processed by CSTEB.

CSTEB contains three EI-Mamba blocks, each consisting of normalization, a Mamba layer, residual aggregation, and a feed-forward network. The class queries and evidence tokens share the same representation dimension. Three type-specific query branches retrieve class-relevant spectral, spatial, and auxiliary evidence and produce a score matrix $\mathbf{S}_i \in \mathbb{R}^{C\times 3}$.

In parallel, the budget branch maps the mean-pooled evidence representation to a class- and sample-specific budget matrix $\mathbf{B}_i \in (0,1)^{C\times 3}$. Its output is passed through a sigmoid function, making every budget strictly positive and bounded above by one. Therefore, a budget can attenuate or retain an evidence score but cannot reverse its sign or amplify it beyond its original magnitude. The three budgets are constrained independently rather than normalized to sum to one, allowing multiple evidence types to support the same candidate class simultaneously. The final evidence logits are obtained by element-wise modulation and aggregation:

$$\mathbf{z}_i^{\mathrm{evi}} = \sum_{t\in\{\mathrm{spec,spa,aux}\}} \mathbf{B}_i^t \odot \mathbf{S}_i^t. \tag{14}$$

Thus, the score matrix represents the support provided by each evidence type for every candidate class, whereas the budget matrix determines how much that evidence contributes to the current

prediction. The same architecture is used on all datasets; only the input channels and number of output classes vary.

A.3 Training and Comparison Protocol

CRED is optimized using AdamW with a batch size of 96, an initial learning rate of 10^{-4}, and a weight decay of 5×10^{-4}. The learning rate is linearly warmed up for four epochs and subsequently decayed using a cosine schedule. The gradient norm is clipped to 0.5, and label smoothing is set to 0.04.

For comparison methods with public implementations, we retain their original architectures and recommended optimization settings while adapting their data interfaces to the common data partitions and evaluation pipeline. Since different architectures exhibit different convergence rates, the permitted training horizon and early-stopping allowance are relaxed when necessary to avoid premature termination. All methods use identical sample coordinates and evaluation criteria, and all ablation variants follow the protocol of the complete CRED model.

A.4 Additional Qualitative Results

Figure 5 compares the classification maps produced by representative recent methods on the three datasets. CRED produces more spatially coherent predictions and fewer isolated classification errors, particularly in heterogeneous regions and near class boundaries.

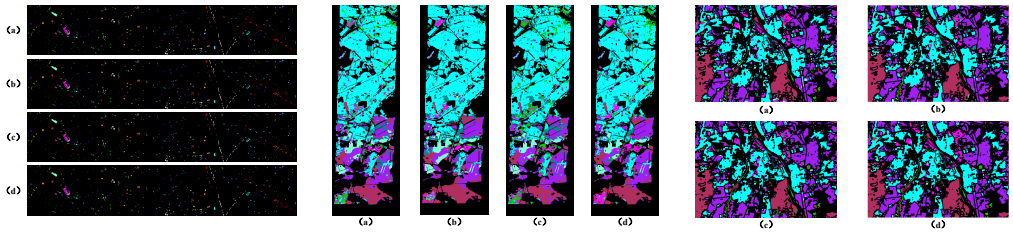


Figure 5: Classification maps on Houston2013, Berlin, and Augsburg, shown from left to right. The compared methods are (a) MSAF, (b) RSCNet, (c) T-APT, and (d) CRED.

A.5 Training Overhead and Inference Efficiency

RH-PD is a training-only supervision mechanism. The HyperSIGMA teachers are not executed during final student optimization; instead, CRED reads their fixed out-of-fold predictions from preconstructed banks. We therefore compare the complete model against CRED without RH-PD while keeping the student architecture, data split, batch size, optimization protocol, and hardware unchanged. Table 6 reports both the recurring training overhead and the resulting inference cost on Houston2013.

Table 6: Accuracy–efficiency trade-off of RH-PD on Houston2013. Relative training cost is normalized to CRED without RH-PD. Inference parameters and peak memory are measured under the same inference setting.

Model	Time/Epoch (s)	Relative Cost	Params (M)	Inference Mem. (MB)	OA (%)
CRED w/o RH-PD	16.10	1.00×	4.98	51.34	92.67
CRED	18.16	1.13×	4.98	51.34	**94.06**

RH-PD increases the average training time by 12.79% (2.06 seconds per epoch) while improving OA by 1.39 percentage points. This additional computation is confined to training: both configurations retain the same 4.98-M parameter student and exhibit identical inference memory. Hence, RH-PD provides a favorable accuracy–efficiency trade-off without increasing the deployed model size or inference cost. The construction of the out-of-fold supervision banks is a one-time offline operation and is excluded from the per-epoch measurements.

A.6 CONSTRUCTION OF RH-PD TRAINING SIGNALS

RH-PD constructs held-out predictions using five spatial folds. Labeled pixels are grouped into non-overlapping spatial blocks, and each block is assigned entirely to one fold. The assignment additionally considers class composition to maintain class coverage across folds.

For each fold, spectral and spatial HyperSIGMA representations are extracted for the labeled samples. The HyperSIGMA encoders remain frozen, and separate linear classification heads are trained on the remaining folds. Each head then predicts the excluded fold, producing one held-out spectral prediction and one held-out spatial prediction for every labeled sample.

Held-out student predictions are constructed using the same spatial partition. For each fold, an independent CRED student is trained on the remaining folds using the supervised student objective and then applied to the excluded fold. Repeating this procedure produces one held-out student prediction for every labeled sample. The resulting teacher and student prediction banks are fixed before the final student is trained.

During final training, the held-out teacher predictions are fused according to their ground-truth probability, top-two probability margin, and normalized entropy. The held-out student predictions are used to estimate whether each sample remains difficult for the student. RH-PD retains samples satisfying the teacher-quality requirements and combines teacher support with student difficulty to rank the eligible samples within each mini-batch.

The selected distillation objective contains three complementary terms. First, *prediction distillation* aligns the final student distribution with the fused teacher target, transferring the teachers' class-level predictions. Second, *evidence distillation* aligns the logits produced by Class-Aware Budgeted Weighting with the same teacher target, directly supervising evidence aggregation before it is combined with the remaining prediction branches. Third, *margin correction* encourages the ground-truth logit to exceed the logit of a competing class, focusing the transferred supervision on class distinctions that remain difficult for the student. These terms are evaluated only on the selected samples, while all labeled samples continue to contribute to the supervised classification objective.