

# Perceive, Act and Correct: Confidence Is Not Enough for Hyperspectral Classification

Muzhou Yang\*, Wuzhou Quan\*, Mingqiang Wei†

Nanjing University of Aeronautics and Astronautics

## Abstract

Confidence alone is often misleading in hyperspectral image classification, as models tend to mistake high predictive scores for correctness while lacking awareness of uncertainty. This leads to confirmation bias, especially under sparse annotations or class imbalance, where models overfit confident errors and fail to generalize. We propose CABIN (Cognitive-Aware Behavior-Informed learNing), a semi-supervised framework that addresses this limitation through a closed-loop learning process of perception, action, and correction. CABIN first develops perceptual awareness by estimating epistemic uncertainty, identifying ambiguous regions where errors are likely to occur. It then acts by adopting an Uncertainty-Guided Dual Sampling Strategy, selecting uncertain samples for exploration while anchoring confident ones as stable pseudo-labels to reduce bias. To correct noisy supervision, CABIN introduces a Fine-Grained Dynamic Assignment Strategy that categorizes pseudo-labeled data into reliable, ambiguous, and noisy subsets, applying tailored losses to enhance generalization. Experimental results show that a wide range of state-of-the-art methods benefit from the integration of CABIN, with improved labeling efficiency and performance.

**Code** — <https://github.com/Muzhou-Yang/CABIN>

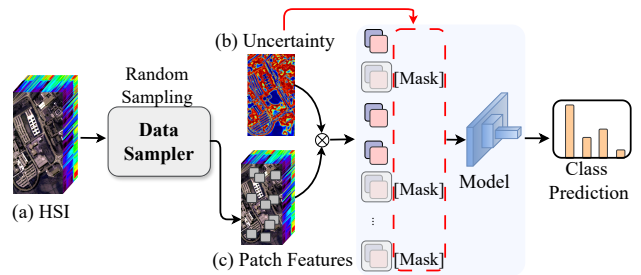
## Introduction

Hyperspectral image (HSI) classification enables fine-grained analysis of land cover and material composition by capturing rich spectral signatures across hundreds of contiguous bands. Its applications are vital in diverse fields such as urban planning (Weber et al. 2018; Yuan et al. 2022), military reconnaissance (Lampropoulos et al. 2008), and precision agriculture (Zhang et al. 2016; Sahadevan 2021). Recently, the paradigm of HSI classification has shifted due to the emergence of deep learning techniques. Advanced methods propose novel mechanisms such as spectral-spatial attention factorization (Sun et al. 2022; Hong et al. 2022; Guo and Liu 2024; Zhao et al. 2024), self-supervised pre-training (Scheibenreif, Mommert, and Borth 2023; Mohamed et al. 2024), and state-space sequence modeling (He

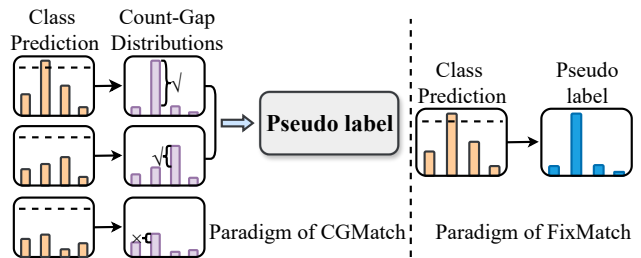
\*These authors contributed equally.

†Mingqiang Wei is the corresponding author.

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.



(a) Illustration of data sampling strategies and corresponding uncertainty usage in existing methods.



(b) Mainstream semi-supervised methods like FixMatch and CG-Match rely on proxy metrics such as static confidence or past consistency, ignoring the model's current cognitive state.

Figure 1: Existing methods depend on confidence for both sampling and pseudo-labeling, yet overlook cognitive gaps, supporting our core insight that confidence is not enough.

et al. 2024; Wang et al. 2025), achieving improved accuracy and cross-scene generalization.

Despite these remarkable advances, current methods often rely on the flawed assumption that confident predictions are inherently reliable, even when labels are noisy or ambiguous. However, due to the limited spatial resolution inherent in HSI data, especially in aerial and satellite imaging, this assumption often fails in practice (Bioucas-Dias et al. 2012). A single pixel may correspond to a mixture of multiple materials or lie on ambiguous boundaries, introducing semantic confusion that reduces the reliability of supervision. As a result, even carefully annotated ground truth cannot ensure precise or complete information. In such cases, naive sampling strategies may reinforce confident errors and lead

to confirmation bias, where models overfit to seemingly correct annotations that are in fact ambiguous. This undermines both label efficiency and generalization.

To address this issue, recent studies have turned to uncertainty-aware learning. Evidential deep learning (Sensoy, Kaplan, and Kandemir 2018; Shen et al. 2023) produces principled frameworks to estimate uncertainty for downstream tasks (Shi et al. 2024; Wentao Bao 2022; Su, Zhang, and Zhou 2023). Other methods integrate uncertainty as reference to adjust loss weights or guiding feature suppression (Yu, Lou, and Chen 2024; Ji et al. 2024), as illustrated in Fig. 1a. However, uncertainty should serve a more active role beyond guiding local optimization or passive learning. Specifically, it should drive learning behavior: identifying which samples to trust, which regions of the feature space to explore, and where to focus supervision. Then, we realize this behavior is naturally suitable for semi-supervised learning by providing a principled basis for sample selection. Existing semi-supervised methods such as FixMatch (Sohn et al. 2020) adopt fixed confidence thresholds to generate pseudo-labels, and CGMatch (Cheng et al. 2025) improves upon this by incorporating historical label consistency. Despite their strengths, these methods still depend on fixed rules or outdated feedback, making them slow to adapt to the model’s evolving uncertainty during training, as illustrated in Fig. 1b.

Inspired by these insights, we propose *Cognitive-Aware Behavior-Informed learning (CABIN)*, a semi-supervised framework that positions uncertainty as an active driver of both data selection and behavior correction. CABIN establishes a closed-loop learning process of perception, action, and correction, where the model learns to perceive sample uncertainty, take informed sampling actions, and iteratively correct its own supervision. This loop is anchored by the *Uncertainty-Guided Dual Sampling Strategy (UGDSS)*, which dynamically balances the exploration of uncertain feature regions and the reinforcement of confident predictions. It encourages the model to discover complex patterns while maintaining stability through reliable supervision. To further address the issue of unreliable supervision, CABIN introduces a novel criterion, the *Uncertainty-Gap metric ( $UG_\alpha$ )*, that quantifies the discrepancy between behavioral confidence and epistemic evidence. Based on  $UG_\alpha$ , we develop the *Fine-Grained Dynamic Assignment Strategy (FDAS)*, which categorizes pseudo-labeled data into reliable, ambiguous, and noisy subsets, applying targeted training objectives accordingly. This fine-grained supervision enhances robustness under class imbalance, sparse labels, and distribution shifts. Experimental results demonstrate that integrating CABIN into state-of-the-art methods enables them to achieve superior performance under semi-supervised settings (using only half of the original training labels) with virtually no additional computational costs.

Our contributions can be summarized as follows:

- We revisit low-supervision hyperspectral classification and find that confidence alone cannot reliably guide learning. To tackle this, we propose CABIN, an uncertainty-guided framework forming a closed loop of perception, action, and correction.

- We propose UGDSS, a dual sampling strategy that leverages epistemic uncertainty to balance the exploration of ambiguous regions and the exploitation of reliable pseudo-labels.
- We introduce FDAS, a dynamic assignment mechanism that uses the Uncertainty-Gap metric to correct pseudo-label noise through fine-grained supervision.
- Experiments confirm that CABIN, as a model-agnostic and plug-and-play strategy, can significantly improve the performance of various state-of-the-art methods under semi-supervised settings, even with as little as 75% of the original annotations.

## Method

### Framework Overview

In this section, we revisit the cognitive mechanism of EDL, and propose a novel framework **CABIN** which explicitly models a cognitive loop of perception, action, and correction to address the core challenge identified in this work: *confidence is not enough* for reliable decision making in HSI classification. The full pipeline is illustrated in Fig. 2.

Given a sample  $i$ , let  $\mathbf{e}_i = f(\mathbf{x}_i | \Theta)$  represent the non-negative evidence vector predicted by the model with parameters  $\Theta$  and features  $\mathbf{x}_i$ . The corresponding Dirichlet distribution is parameterized as  $\alpha_i = \mathbf{e}_i + 1$ , with total evidence  $S_i = \sum_{k=1}^K \alpha_{i,k}$ . The estimated class probabilities are computed as the mean of the Dirichlet distribution:  $\hat{\mathbf{p}}_i = \alpha_i / S_i$ . The epistemic uncertainty is denoted as  $u_i = K / S_i$ . This perceived uncertainty then guides the model to act through our UGDSS module. UGDSS partitions the candidate set  $\mathcal{D}_t$  based on an adaptive epistemic threshold:

$$\mathcal{D}_{hc}, \mathcal{D}_{qu} = \text{UGDSS}(\mathcal{D}_t, \{u_i\}_{i \in \mathcal{D}_t}), \quad (1)$$

where  $\mathcal{D}_{hc}$  comprises high uncertain samples for critical learning, and  $\mathcal{D}_{qu}$  includes reliable samples to maintain stability. To enhance the representation of  $\mathcal{D}_{qu}$ , UGDSS integrates two internal components: (i) Diverse-Representative Query Selection (DRQS) refines  $\mathcal{D}_{hc}$  via feature-space clustering:  $\mathcal{D}_{au} = \text{DRQS}(\mathcal{D}_{hc})$ . (ii) Gaussian Feature Perturbation (GFP) generates local variants of selected samples by injecting uncertainty-scaled noise in the feature space:  $\hat{\mathcal{D}}_{aug} = \text{GFP}(\mathcal{D}_{au})$ . Given the potential noise in the pseudo-labeled set  $\mathcal{D}_{qu}$ , we introduce Fine-Grained Dynamic Assignment Strategy (FDAS), which jointly considers softmax confidence and evidential separation to categorize:

$$\mathcal{D}_{re}, \mathcal{D}_{am}, \mathcal{D}_{no} = \text{FDAS}(\mathcal{D}_{qu}), \quad (2)$$

where  $\mathcal{D}_{re}$  contains reliable pseudo-labels,  $\mathcal{D}_{am}$  includes ambiguous ones, and  $\mathcal{D}_{no}$  is discarded.

The final training objective integrates all confident supervision sources. For labeled and highly reliable pseudo-labeled samples, we adopt the EDL loss:

$$\mathcal{L}_{\text{EDL}}^{(i)} = \sum_{j=1}^K y_{i,j} (\log(S_i) - \log(\alpha_{i,j})), \quad (3)$$

where  $\mathbf{y}_i \in \{0, 1\}^K$  is the one-hot supervision vector.

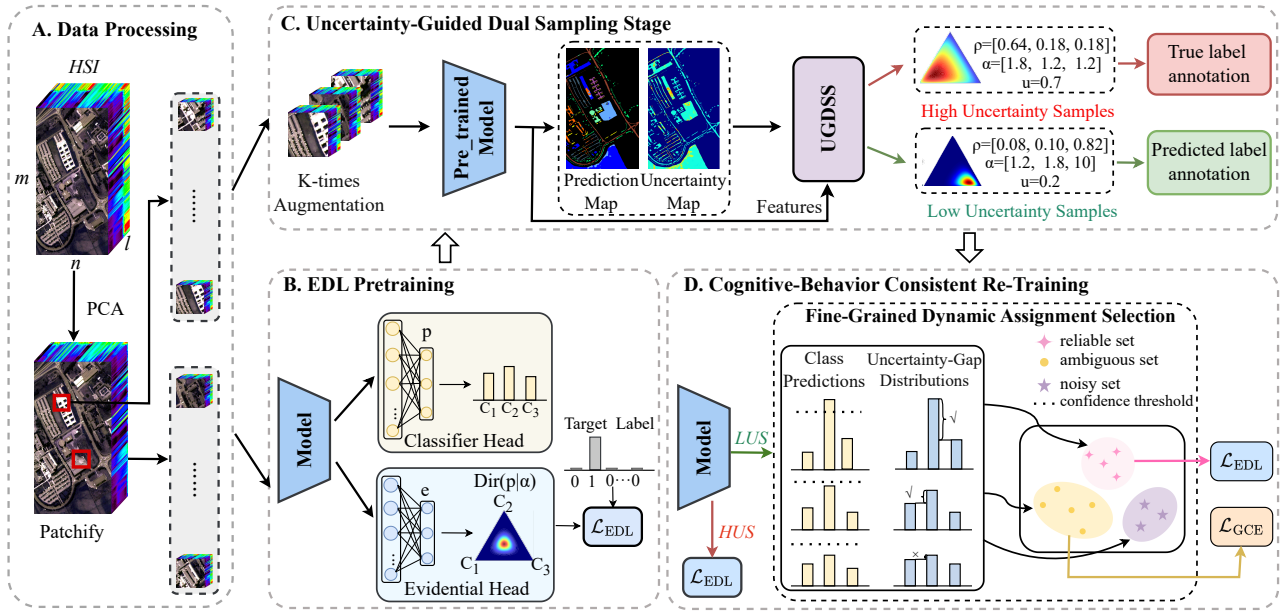


Figure 2: An overview of the proposed CABIN framework, designed to address the challenge that “confidence is not enough” for HSI classification. CABIN establishes a closed-loop process of perception, action and correction: the model first perceives sample uncertainty using EDL, then acts by selecting key samples via UGDSS, and finally corrects cognitive-behavioral gaps by dynamically assessing pseudo-label reliability with FDAS. This process enables more robust learning under limited supervision.

For ambiguous pseudo-labeled samples, we apply the noise-robust Generalized Cross Entropy (GCE) loss (Zhang and Sabuncu 2018):

$$\begin{aligned} \mathcal{L}_{\text{GCE}}^{(i)} &= \frac{1 - (\hat{p}_{i,y})^q}{q}, \\ \hat{p}_{i,y} &= \frac{\alpha_{i,y}}{S_i}, \end{aligned} \quad (4)$$

where  $y$  is the pseudo-label index for sample  $i$ , and the hyperparameter  $q \in (0, 1]$  is set to 0.7 following (Zhang and Sabuncu 2018).

The total training objective is:

$$\mathcal{L} = \mathcal{L}_{\text{EDL}}(\mathcal{D}_L \cup \hat{\mathcal{D}}_{\text{aug}}) + \lambda_r \mathcal{L}_{\text{EDL}}(\mathcal{D}_{re}) + \lambda_a \mathcal{L}_{\text{GCE}}(\mathcal{D}_{am}), \quad (5)$$

where  $\lambda_r$  and  $\lambda_a$  are balancing weights, both set to 0.3.

### Uncertainty-Guided Dual-Sampling Strategy

Building upon EDL, we propose UGDSS (in Fig. 3) which partitions the candidate set  $\mathcal{D}_t$  into two subsets for targeted exploration and exploitation. Since EDL-based uncertainty estimation is particularly sensitive to disturbances, which may lead to a misleading sampling process especially under sparse data conditions, we adopt a test-time augmentation method (Nalepa, Myller, and Kawulok 2020), generating  $K$  spectral-spatial variants of each  $\mathbf{x}_i$  through  $k$ -th random transforms  $\mathcal{T}_k$ , and averaging their EDL outputs:

$$\hat{u}_i = \frac{1}{K} \sum_{k=1}^K u(\mathcal{T}_k(\mathbf{x}_i)). \quad (6)$$

**Adaptive Thresholding.** The estimated uncertainty values  $\{\hat{u}_i\}_{i=1}^N$  exhibit a heterogeneous distribution, reflecting epistemic diversity across the feature space. However, using a static cutoff (e.g., fixed percentile) may introduce misaligned partitions due to the early-stage model’s prediction noise and instability.

To this end, we introduce a histogram-based adaptive thresholding strategy within the UGDSS module. Specifically, we sort  $\hat{u}_i$  in ascending order to form a histogram density distribution  $h[n]$  with  $N$  bins. The adaptive threshold  $T_u \in \mathbb{R}$  is set as the first local minimum satisfying:

$$T_u = \min \{h[n] \mid \Delta h[n] < \delta, \Delta^2 h[n] < 0\}, \quad (7)$$

where  $\Delta h[n] = h[n+1] - h[n]$  and  $\Delta^2 h[n] = \Delta h[n+1] - \Delta h[n]$  denote the first- and second-order discrete differences. The tolerance  $\delta$  avoids unstable extrema.

This adaptive threshold  $T_u$  is then used to divide the current candidate set  $\mathcal{D}_t$  into two subsets:  $\mathcal{D}_{hc} = \{i \in \mathcal{D}_t \mid \hat{u}_i \geq T_u\}$ ,  $\mathcal{D}_{qu} = \mathcal{D}_t \setminus \mathcal{D}_{hc}$ . This enables UGDSS to adaptively partition samples based on current uncertainty distribution, facilitating the robust selection of both uncertain and reliable samples.

**Diverse-Representative Query Selection.** To further refine the high-uncertainty candidate set  $\mathcal{D}_{hc}$  obtained via UGDSS, we propose DRQS strategy to eliminate sample redundancy and enhance selection diversity.

Specifically, for each candidate sample  $i \in \mathcal{D}_{hc}$ , we extract its semantic embedding  $f_i$  from the final feature layer of the model  $\Phi$ , which captures high-level discriminative information prior to classification. We then perform K-means++ (Arthur and Vassilvitskii 2007) clustering over the

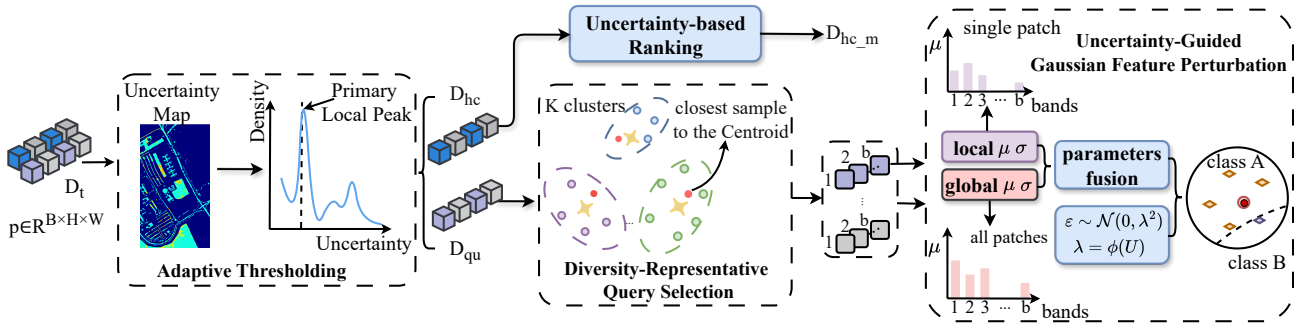


Figure 3: Schematic of the UGDSS module, which leverages epistemic uncertainty to orchestrate a dual-path selection strategy balancing exploration and exploitation. The exploration path combines DRQS and GFP to actively probe the model’s knowledge gaps, whereas the exploitation path selects confident samples for reliable pseudo-labeling.

embedding set  $\{\bar{f}_i\}$  and partition it into  $M$  clusters, where  $M$  is the predefined query budget. From each cluster, the sample closest to the centroid is selected to form the final query set  $\mathcal{D}_{au}$ :

$$\mathcal{D}_{au} = \bigcup_{k=1}^M \left\{ \arg \min_{i \in \mathcal{D}_{hc}} \|\bar{f}_i - C_k\|_2 \right\}, \quad (8)$$

where  $C_k$  denotes the centroid of the  $k$ -th cluster, and  $\|\cdot\|_2$  represents the Euclidean distance.

#### Uncertainty-Guided Gaussian Feature Perturbation.

To enhance the model’s robustness in highly uncertain regions, we introduce an uncertainty-guided Gaussian perturbation module that promotes feature-level exploration. The core idea is to make the perturbation strength proportional to the model’s uncertainty. For each sample  $i \in \mathcal{D}_{au}$ , we first compute its perturbation scale  $\lambda_i$  by linearly mapping the normalized epistemic uncertainty  $\hat{u}_i \in [0, 1]$  to a predefined range  $[\lambda_{\min}, \lambda_{\max}]$ :

$$\lambda_i = \lambda_{\min} + (\lambda_{\max} - \lambda_{\min}) \cdot \hat{u}_i. \quad (9)$$

This ensures that higher uncertainty yields stronger perturbations. We then generate a perturbed feature representation by injecting Gaussian noise scaled by  $\lambda_i$ . The noise is sampled from a weighted mixture of local and global statistics over  $\mathcal{D}_{au}$ . To maintain spectral fidelity and prevent unrealistic distortions, all perturbed values are clamped to their original dynamic range. This feature-space augmentation strategy, inspired by methods like Manifold Mixup (Verma et al. 2019) and CutMix (Yun et al. 2019), encourages the model to learn more invariant features. The resulting augmented set, denoted as  $\hat{\mathcal{D}}_{aug}$ , is then used for retraining.

#### Fine-Grained Dynamic Assignment Strategy

Existing methods like CGMatch (Cheng et al. 2025) identify ambiguous samples by tracking historical pseudo-label distributions using the Count-Gap metric. However, this method incurs delayed responses, high computational cost, and poorly reflects the model’s current confidence. To address these limitations, we propose a cognitive-driven pseudo-label selection framework named FDAS. By jointly

leveraging the behavioral confidence and internal evidential cues, FDAS enables a dynamic separation of the pseudo-labeled set  $\mathcal{D}_{pu}$ .

Specifically, behavioral confidence is denoted as the maximum classification probability  $c_i = \max(\hat{\mathbf{p}}_i)$ , where  $\hat{\mathbf{p}}_i$  denotes the predicted class distribution for sample  $i \in \mathcal{D}_{pu}$ . To capture the model’s discriminative certainty from a cognitive perspective, we introduce the *Uncertainty-Gap* ( $UG_\alpha$ ) based on EDL. Given the evidence vector  $\alpha_i = (\alpha_{i1}, \dots, \alpha_{iK})$  for a  $K$ -class sample  $x_i$ , we apply an Exponential Moving Average (EMA) to obtain a smoothed vector  $\bar{\alpha}_i$ , thereby mitigating fluctuations from single-batch estimations:

$$\bar{\alpha}_i \leftarrow m \cdot \bar{\alpha}_i + (1 - m) \cdot \alpha_i, \quad (10)$$

where  $m \in [0, 1)$  is the momentum coefficient. The UG is then computed using this smoothed evidence:  $UG_i^\alpha = \max_k(\bar{\alpha}_{ik}) - \text{second\_max}_k(\bar{\alpha}_{ik})$ , which quantifies the evidential margin between the most and second-most supported classes.

Based on this stable metric, we partition the pseudo-labeled set  $\mathcal{D}_{pu}$  into three disjoint subsets using dynamic thresholds for confidence ( $\tau_c$ ) and evidence ( $\tau_e$ ):

$$\begin{aligned} \mathcal{D}_{re} &= \{x_i \in \mathcal{D}_{pu} \mid c_i \geq \tau_c \ \& \ UG_i^\alpha \geq \tau_e\}, \\ \mathcal{D}_{no} &= \{x_i \in \mathcal{D}_{pu} \mid c_i < \tau_c \ \& \ UG_i^\alpha < \tau_e\}, \\ \mathcal{D}_{am} &= \mathcal{D}_{pu} \setminus (\mathcal{D}_{re} \cup \mathcal{D}_{no}), \end{aligned} \quad (11)$$

where  $\mathcal{D}_{re}$ ,  $\mathcal{D}_{no}$ , and  $\mathcal{D}_{am}$  correspond to the reliable, noisy, and ambiguous sets, respectively. This three-way partitioning allows for tailored learning strategies for each subset. Moreover, both thresholds  $\tau_c$  and  $\tau_e$  are also dynamically updated using EMA over their respective batch-wise statistics (i.e., the average confidence and the average uncertainty gap), allowing the model to adapt to its evolving confidence.

## Experiments

### Datasets and Evaluation Metrics

We conduct experiments on five widely used hyperspectral datasets, including Indian Pines, Salinas, Pavia University, WHU-Hi-HongHu, and WHU-Hi-LongKou (Zhong et al. 2020). These datasets vary in spatial resolution, spectral range, and scene complexity, covering agricultural, urban,

Class No.	CNN-based							Transformer-based			
	SVM	2DCNN	3DCNN	ReS <sup>2</sup>	CABIN	CLOLN	CABIN	SSFTT	CABIN	GSC-ViT	CABIN
1	19.76	65.63	75.12	100.00	<b>100.00</b>	100.00	<b>100.00</b>	100.00	<b>100.00</b>	100.00	<b>100.00</b>
2	71.60	63.44	70.40	48.63	<b>68.08</b>	64.91	<b>77.95</b>	<b>88.04</b>	85.23	59.08	<b>86.17</b>
3	41.90	60.25	68.43	71.14	<b>76.96</b>	61.01	<b>74.30</b>	80.13	<b>82.28</b>	<b>82.91</b>	76.20
4	79.81	41.11	69.73	99.49	<b>100.00</b>	91.37	<b>100.00</b>	95.94	<b>100.00</b>	<b>100.00</b>	96.45
5	25.35	87.05	77.16	88.26	<b>96.16</b>	90.97	<b>91.87</b>	87.58	<b>93.45</b>	84.88	<b>94.58</b>
6	86.30	97.21	85.70	95.0	<b>98.12</b>	95.65	<b>95.94</b>	<b>96.52</b>	94.93	91.01	<b>99.13</b>
7	34.00	89.47	70.63	100.00	<b>100.00</b>	100.00	<b>100.00</b>	100.00	<b>100.00</b>	100.00	<b>100.00</b>
8	95.12	96.66	90.21	100.00	<b>100.00</b>	100.00	<b>100.00</b>	100.00	<b>100.00</b>	<b>100.00</b>	96.12
9	44.44	32.26	68.20	100	<b>100.00</b>	100.00	<b>100.00</b>	100.00	<b>100.00</b>	100.00	<b>100.00</b>
10	52.75	73.84	75.48	62.98	<b>96.35</b>	<b>90.56</b>	83.05	75.86	<b>92.49</b>	<b>91.20</b>	89.06
11	87.19	84.36	78.57	81.16	<b>81.99</b>	<b>80.79</b>	79.38	85.05	<b>87.00</b>	<b>84.43</b>	78.88
12	42.78	42.89	65.23	83.54	<b>83.73</b>	<b>88.79</b>	66.00	74.14	<b>77.22</b>	80.83	<b>89.69</b>
13	92.39	98.58	90.17	100.00	<b>100.00</b>	<b>100.00</b>	99.39	100	<b>100.00</b>	99.39	<b>100.00</b>
14	96.84	94.02	94.65	98.04	<b>99.92</b>	<b>98.12</b>	95.84	94.37	<b>97.88</b>	<b>95.18</b>	91.10
15	63.40	42.65	64.89	95.38	<b>99.42</b>	96.23	<b>97.69</b>	<b>99.71</b>	97.40	92.49	<b>99.42</b>
16	90.36	92.19	93.07	100.00	<b>100.00</b>	98.11	<b>100.00</b>	100.00	<b>100.00</b>	<b>100.00</b>	96.23
OA(%)	72.51	76.39	78.28	79.70	<b>87.39</b> <i>+7.69</i>	83.77	<b>84.69</b> <i>+0.92</i>	87.35	<b>90.08</b> <i>+2.73</i>	84.64	<b>87.40</b> <i>+2.76</i>
AA(%)	65.50	72.18	76.94	88.98	<b>93.80</b> <i>+4.82</i>	90.34	<b>91.22</b> <i>+0.88</i>	93.65	<b>94.20</b> <i>+0.55</i>	91.33	<b>93.31</b> <i>+1.98</i>
$\kappa \times 100$	68.22	72.85	75.46	76.71	<b>85.68</b> <i>+8.97</i>	81.40	<b>82.60</b> <i>+1.20</i>	85.48	<b>88.66</b> <i>+3.18</i>	82.49	<b>85.60</b> <i>+3.11</i>

Table 1: Comparison of classification and overall performance on the *Indian Pines* dataset across different methods. For other methods, 20 samples per class (320 total) are used, while CABIN uses only 240 samples. CABIN results are highlighted in light gray. Red font indicates performance degradation, while Green font indicates improvement. The best results for both per-class accuracy and overall metrics, comparing CABIN and non-CABIN methods, are shown in **bold**.

Module		Indian Pines			Salinas		
UGDSS	FADS	OA (%)	AA (%)	$\kappa \times 100$	OA (%)	AA (%)	$\kappa \times 100$
×	×	87.35	93.65	85.48	96.00	98.27	95.55
✓	×	89.80	93.46	88.30	97.21	<b>98.97</b>	96.89
×	✓	88.87	93.66	87.27	96.46	98.58	96.06
✓	✓	<b>90.08</b>	<b>94.20</b>	<b>88.66</b>	<b>97.51</b>	98.90	<b>97.23</b>

Table 2: Module-wise ablation study of CABIN on Indian Pines and Salinas datasets under the SSFTT baseline.

and UAV-based remote sensing scenarios. Overall, they provide a comprehensive benchmark for evaluating robust hyperspectral classification methods.

Based on the above datasets, we evaluate the performance of our method using several standard classification metrics, including Overall Accuracy (OA), Average Accuracy (AA), and Cohen’s Kappa Coefficient (Cohen 1960) ( $\kappa$ ).

### Comparative Experiments

To validate the effectiveness, we integrate CABIN into four existing state-of-the-art methods, including CNN-based (ReS<sup>2</sup> (Meng et al. 2022), CLOLN (Li et al. 2024)) and Transformer-based (SSFTT (Sun et al. 2022), GSC-ViT (Zhao et al. 2024)). In addition, three representative baseline methods are included for comprehensive compar-

ison. Model architectures follow the original implementations to ensure fair comparison, while training hyperparameters are adjusted to fit our experimental setup.

Overall, CABIN delivers substantial performance gains across nearly all models and datasets. Especially on Indian Pines (Table 1), it improves ReS<sup>2</sup>’s OA by a remarkable **+7.69%** and SSFTT’s  $\kappa$  by **+3.18**. These superior performances are achieved with **25% fewer labeled samples**, highlighting CABIN’s exceptional sample efficiency. Class-wise accuracy further reveals that improvements mainly concentrate on classes where the original models struggle. Similar trends of robustness hold on the other four datasets (Table 3), e.g., a **+2.44%** OA boost for CLOLN on HongHu. Despite some minor trade-offs, e.g., SSFTT’s AA on PaviaU, the general results confirm CABIN’s ability to elevate performance and label-efficiency of existing supervised HSI classifiers. Results consistently demonstrate that CABIN is effective, model-agnostic and plug-and-play.

### Ablation Study

In this section, we conduct additional experiments to gain a deeper insights into how CABIN guides model learning more accurately and efficiently.

**Evaluation of Overall Effectiveness.** To evaluate the contribution of each module in CABIN, we perform an ablation study by selectively enabling UGDSS and FDAS across two datasets. As shown in Table 2, UGDSS alone boosts the

Datasets	Metrics	CNN-based							Transformer-based			
		SVM	2DCNN	3DCNN	ReS <sup>2</sup>	CABIN	CLOLN	CABIN	SSFTT	CABIN	GSC-viT	CABIN
Salinas	OA(%)	85.44	86.60	89.70	94.78	<b>95.84</b> +1.06	91.27	<b>93.19</b> +1.92	96.00	<b>97.51</b> +1.51	95.05	<b>96.52</b> +1.47
	AA(%)	90.31	92.89	94.88	98.07	<b>98.32</b> +0.25	96.20	<b>96.50</b> +0.3	98.27	<b>98.90</b> +0.63	98.03	<b>98.21</b> +0.18
	$\kappa \times 100$	83.85	85.10	84.41	94.21	<b>95.37</b> +1.16	90.33	<b>92.44</b> +2.11	95.55	<b>97.23</b> +1.68	94.50	<b>96.13</b> +1.63
PaviaU	OA(%)	76.60	77.63	87.52	90.71	<b>93.14</b> +2.43	95.55	<b>96.71</b> +1.16	91.88	<b>93.75</b> +1.87	93.03	<b>94.95</b> +1.92
	AA(%)	83.57	81.62	90.22	92.68	<b>94.08</b> +1.4	95.45	<b>96.53</b> +1.08	93.77	<b>92.85</b> -0.92	94.05	<b>95.77</b> +1.72
	$\kappa \times 100$	70.39	71.15	84.41	87.88	<b>90.99</b> +3.11	94.12	<b>95.64</b> +1.52	89.35	<b>91.70</b> +2.35	90.76	<b>93.35</b> +2.59
Longkou	OA(%)	87.81	89.27	91.51	95.99	<b>96.73</b> +0.74	96.87	<b>97.89</b> +1.02	97.22	<b>98.31</b> +1.09	98.09	<b>98.64</b> +0.55
	AA(%)	84.55	86.09	93.44	96.60	<b>95.55</b> -1.05	96.90	<b>97.71</b> +0.81	96.83	<b>97.23</b> +0.4	97.43	<b>97.74</b> +0.31
	$\kappa \times 100$	84.39	86.23	90.41	94.80	<b>95.74</b> +0.94	95.93	<b>97.32</b> +1.39	96.37	<b>97.78</b> +1.41	97.49	<b>98.22</b> +0.73
HongHu	OA(%)	77.49	79.23	81.66	87.82	<b>88.32</b> +0.5	88.47	<b>90.91</b> +2.44	87.68	<b>88.75</b> +1.07	91.52	<b>92.01</b> +0.49
	AA(%)	76.87	76.10	78.32	87.34	<b>88.72</b> +1.38	87.15	<b>89.98</b> +2.83	85.56	<b>89.36</b> +3.8	89.96	<b>89.14</b> -0.82
	$\kappa \times 100$	72.73	74.55	77.42	84.82	<b>86.76</b> +1.94	85.03	<b>86.45</b> +1.42	84.54	<b>85.95</b> +1.41	89.36	<b>89.89</b> +0.53

Table 3: Comparison of different types of methods on the Pavia University, WHU-Hi-Longkou, and WHU-Hi-HongHu datasets.

OA from 87.35% to **89.80%**, while FDAS alone raises it to **88.87%** on Indian Pines, which confirms that adding either module individually can lead to notable gains. Moreover, when we integrate UGDSS and FDAS, CABIN can achieve an OA of **90.08%** and a  $\kappa$  of **88.66** on Indian Pines, outperforming all other variants. The similar trend of robustness holds on Salinas. These results demonstrate that UGDSS and FDAS are not only effective individually, but also complementary components that can be combined to fully realize the potential of our framework.

**Diversification Capability of DRQS.** To compare the differences between the sampling strategies, we visualize the corresponding sampling results and overlay them on the uncertainty map generated by the model. The random sampling strategy (marked by red crosses) may over-select regions where the model has learned well. In contrast, the purely uncertainty-based method (marked by green circles) in Fig. 4a) shifts the focus to the critical high-uncertainty boundary regions. However, due to the strong spatial correlation of the HSI data, this strategy is prone to form dense and redundant sampling clusters in local regions (as shown in the enlarged illustration), resulting in diminishing learning returns. In Fig. 4b), the samples selected by DRQS (marked by green circles) are more spatially dispersed while still covering key uncertainty regions, thus capturing richer and more diverse information. This phenomenon shows that DRQS effectively reduces spatial redundancy on the basis of retaining the discriminant regions, thereby achieving a more

diverse sample selection capability.

**Efficiency Enhancement under Annotation Constraints.**

To evaluate the annotation efficiency of our selection strategy, we conduct an ablation study under different annotation ratios. For the Indian Pines and Salinas datasets, the baseline method randomly selects 20 samples from each class (320 in total) for training. Our method first uses 10 samples from each class (160 in total) for EDL pre-training, and the remaining 160 samples constitute the selection pool (i.e., the row with 100% ratio in the Table 4). According to the results, our method achieves peak performance with only 50% of the annotations. This demonstrates the high annotation efficiency of our strategy. By focusing on informative and diverse samples, the strategy can capture sufficient training signals without exhaustive annotations. When we increase the ratio to 75% or 100%, the performance degrades, suggesting that excessive annotation may introduce redundant or noisy supervision. These findings confirm the effectiveness of our strategy in real-world scenarios with limited annotation resources.

**Feature Enhancement Capability of GFP.**

To further explore the mechanism of GFP, we conduct more experiments for qualitative and quantitative analysis. In Figure 5, the qualitative visualization shows the features of enhanced samples generated around the original samples. The feature distribution of these enhanced samples does not exceed the category boundary, further strengthening the feature consis-

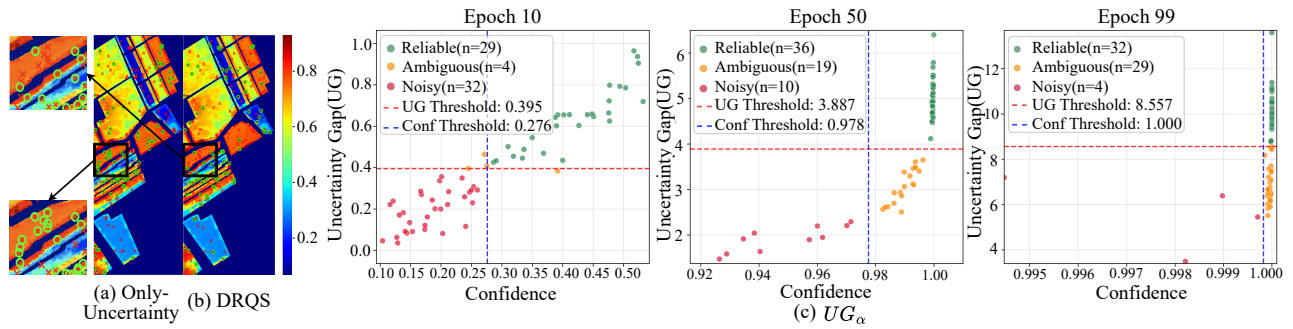


Figure 4: Left: Comparison of sampling strategies. (a) uses only dynamic thresholding, whereas (b) applies DRQS for diverse and non-redundant selection in feature space. Green circles: selected samples; red crosses: initial labeled samples. Right: Evolution of UG’s ability to discriminate sample difficulty throughout training.

	Indian Pines			Salinas		
Ratio	OA (%)	AA (%)	$\kappa \times 100$	OA (%)	AA (%)	$\kappa \times 100$
0%	78.92	89.08	75.74	94.76	98.00	95.28
25%	85.81	91.85	83.78	96.62	98.72	96.24
50%	<b>90.08</b>	<b>94.20</b>	<b>88.66</b>	<b>97.51</b>	<b>98.90</b>	<b>97.23</b>
75%	89.01	94.47	88.43	96.85	98.73	96.50
100%	89.15	93.58	87.58	97.11	98.89	96.79

Number	OA (%)	AA (%)	$\kappa \times 100$	OA (%)	AA (%)	$\kappa \times 100$
1	89.00	93.24	87.32	96.22	98.51	95.80
2	89.45	93.75	87.91	96.96	98.73	96.61
3	89.85	93.61	88.35	97.14	98.68	96.82
4	89.81	94.07	88.35	<b>97.51</b>	<b>98.90</b>	<b>97.23</b>
5	<b>90.22</b>	<b>93.76</b>	<b>88.75</b>	96.84	98.58	96.48
6	<b>90.08</b>	<b>94.20</b>	<b>88.66</b>	96.22	98.44	95.80
7	89.06	93.72	87.48	96.26	98.31	95.84

Table 4: Performance under different annotation ratios and augmentation numbers.

tency and intra-class compactness between samples of the same label. Quantitatively, Table 4 shows the impact of different numbers of enhanced samples on model performance. According to the results, although OA and  $\kappa$  perform better when the number is set to 5 on the Indian Pines dataset, the overall balance is better when the number is set to 6. On the Salinas dataset, the performance peaks when the number is set to 4. These findings confirm that moderate feature perturbation helps improve the robustness of the model, while too much may introduce noise, resulting in ambiguous decision boundary and decreased performance. By these experiments, the effectiveness of the GFP module and the importance of controlling the intensity of feature space perturbation on the enhancement effect can be verified.

#### Discriminative Power of $UG_\alpha$ across Training Epochs.

We track the evolution of  $UG_\alpha$ ’s discriminative capacity over training in Fig. 4c. Early on, both softmax confidence and  $UG_\alpha$  fail to separate samples by difficulty due to immature representations. As training proceeds, confidence val-

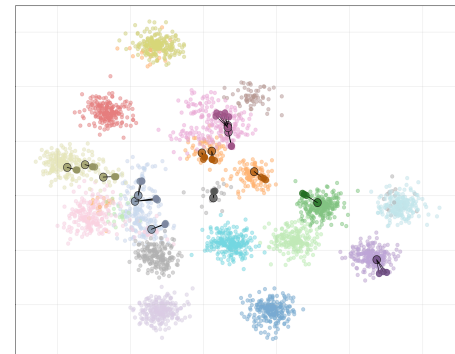


Figure 5: t-SNE projection of the original and augmented samples in feature space. Each color represents a semantic class. Light-colored points indicate original samples; dark points with outlines are selected samples; dark points without outlines are generated samples. Light and dark tones of the same color correspond to the same class.

ues saturate near 1.0, making them unreliable for identifying ambiguous or noisy cases. In contrast,  $UG_\alpha$  gradually sharpens its separation: by Epoch 50, easy samples start to emerge, and by Epoch 99, a clear boundary forms between high-UG (reliable) and low-UG (ambiguous/noisy) instances even under uniformly high confidence. This dynamic reveals UG’s growing ability to structure the sample space meaningfully, enabling reliable group-aware training.

## Conclusion

In this paper, we revisit the performance bottleneck in semi-supervised hyperspectral classification, tracing it to confirmation bias induced by over-reliance on confidence and neglect of uncertainty. From a cognitive-behavioral consistency perspective, we interpret this as a mismatch between model perception and decision-making. To address it, we introduce CABIN, a model-agnostic framework that closes the loop of perception, action, and correction through uncertainty-guided learning. Experiments show that CABIN consistently improves generalization and label efficiency across diverse baselines.

## Acknowledgements

This work was supported by the National Natural Science Foundation of China (No. T2322012, No. 62572240).

## References

- Arthur, D.; and Vassilvitskii, S. 2007. k-means++: the advantages of careful seeding. In *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, 1027–1035.
- Bioucas-Dias, J. M.; Plaza, A.; Dobigeon, N.; Parente, M.; Du, Q.; Gader, P.; and Chanussot, J. 2012. Hyperspectral Unmixing Overview: Geometrical, Statistical, and Sparse Regression-Based Approaches. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 5(2): 354–379.
- Cheng, B.; Lu, J.; Tian, Y.; Zhao, H.; Chang, Y.; and Du, L. 2025. CGMatch: A Different Perspective of Semi-supervised Learning. In *Computer Vision and Pattern Recognition Conference*, 15381–15391.
- Cohen, J. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1): 37–46.
- Guo, H.; and Liu, W. 2024. S3L: Spectrum Transformer for Self-Supervised Learning in Hyperspectral Image Classification. *Remote Sensing*, 16(6).
- He, Y.; Tu, B.; Jiang, P.; Liu, B.; Li, J.; and Plaza, A. 2024. IGroupSS-Mamba: Interval Group Spatial-Spectral Mamba for Hyperspectral Image Classification. *IEEE Transactions on Geoscience and Remote Sensing*, 62: 1–17.
- Hong, D.; Han, Z.; Yao, J.; Gao, L.; Zhang, B.; Plaza, A.; and Chanussot, J. 2022. SpectralFormer: Rethinking Hyperspectral Image Classification With Transformers. *IEEE Transactions on Geoscience and Remote Sensing*, 60: 1–15.
- Ji, F.; Zhao, W.; Wang, Q.; Emery, W. J.; Peng, R.; Man, Y.; Wang, G.; and Jia, K. 2024. Spectral-Spatial Evidential Learning Network for Open-Set Hyperspectral Image Classification. *IEEE Transactions on Geoscience and Remote Sensing*, 62: 1–17.
- Lampropoulos, G. A.; Liu, T.; Qian, S.-E.; and Fei, C. 2008. Hyperspectral Classification Fusion for Classifying Different Military Targets. In *IGARSS 2008 - 2008 IEEE International Geoscience and Remote Sensing Symposium*, volume 3, III – 262–III – 265.
- Li, C.; Rasti, B.; Tang, X.; Duan, P.; Li, J.; and Peng, Y. 2024. Channel-Layer-Oriented Lightweight Spectral-Spatial Network for Hyperspectral Image Classification. *IEEE Transactions on Geoscience and Remote Sensing*, 62: 1–14.
- Meng, Z.; Jiao, L.; Liang, M.; and Zhao, F. 2022. A Lightweight Spectral-Spatial Convolution Module for Hyperspectral Image Classification. *IEEE Geoscience and Remote Sensing Letters*, 19: 1–5.
- Mohamed, S.; Haghighat, M.; Fernando, T.; Sridharan, S.; Fookes, C.; and Moghadam, P. 2024. FactoFormer: Factorized Hyperspectral Transformers With Self-Supervised Pre-training. *IEEE Transactions on Geoscience and Remote Sensing*, 62: 1–14.
- Nalepa, J.; Myller, M.; and Kawulok, M. 2020. Training- and Test-Time Data Augmentation for Hyperspectral Image Segmentation. *IEEE Geoscience and Remote Sensing Letters*, 17(2): 292–296.
- Sahadevan, A. S. 2021. Extraction of spatial-spectral homogeneous patches and fractional abundances for field-scale agriculture monitoring using airborne hyperspectral images. *Computers and Electronics in Agriculture*, 188: 106325.
- Scheibenreif, L.; Mommert, M.; and Borth, D. 2023. Masked Vision Transformers for Hyperspectral Image Classification. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2166–2176.
- Sensoy, M.; Kaplan, L.; and Kandemir, M. 2018. Evidential deep learning to quantify classification uncertainty. In *International Conference on Neural Information Processing Systems*, 3183–3193.
- Shen, M.; Bu, Y.; Sattigeri, P.; Ghosh, S.; Das, S.; and Wornell, G. 2023. Post-hoc uncertainty learning using a dirichlet meta-model. In *AAAI Conference on Artificial Intelligence*, 9772–9781.
- Shi, R.; Duan, L.; Huang, T.; and Jiang, T. 2024. Evidential Uncertainty-Guided Mitochondria Segmentation for 3D EM Images. In *AAAI Conference on Artificial Intelligence*, 4847–4855.
- Sohn, K.; Berthelot, D.; Li, C.-L.; Zhang, Z.; Carlini, N.; Cubuk, E. D.; Kurakin, A.; Zhang, H.; and Raffel, C. 2020. FixMatch: simplifying semi-supervised learning with consistency and confidence. In *International Conference on Neural Information Processing Systems*.
- Su, B.; Zhang, H.; and Zhou, Z. 2023. HSIC-based Moving Weight Averaging for Few-Shot Open-Set Object Detection. In *ACM International Conference on Multimedia*, 5358–5369.
- Sun, L.; Zhao, G.; Zheng, Y.; and Wu, Z. 2022. Spectral-Spatial Feature Tokenization Transformer for Hyperspectral Image Classification. *IEEE Transactions on Geoscience and Remote Sensing*, 60: 1–14.
- Verma, V.; Lamb, A.; Beckham, C.; Najafi, A.; Mitliagkas, I.; Lopez-Paz, D.; and Bengio, Y. 2019. Manifold Mixup: Better Representations by Interpolating Hidden States. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, 6438–6447.
- Wang, G.; Zhang, X.; Peng, Z.; Zhang, T.; and Jiao, L. 2025. S2Mamba: A Spatial-Spectral State Space Model for Hyperspectral Image Classification. *IEEE Transactions on Geoscience and Remote Sensing*, 63: 1–13.
- Weber, C.; Aguejdad, R.; Briottet, X.; Avala, J.; Fabre, S.; Demuyneck, J.; Zenou, E.; Deville, Y.; Karoui, M.; Benhalouche, F.; Gadal, S.; Ourghemmi, W.; Mallet, C.; Bris, A. L.; and Chehata, N. 2018. Hyperspectral Imagery for Environmental Urban Planning. In *IEEE International Geoscience and Remote Sensing Symposium*, 1628–1631.
- Wentao Bao, Y. K., Qi Yu. 2022. OpenTAL: Towards Open Set Temporal Action Localization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2969–2979.

- Yu, L.; Lou, Y.; and Chen, F. 2024. Uncertainty-aware Graph-based Hyperspectral Image Classification. In *The Twelfth International Conference on Learning Representations*.
- Yuan, J.; Wang, S.; Wu, C.; and Xu, Y. 2022. Fine-Grained Classification of Urban Functional Zones and Landscape Pattern Analysis Using Hyperspectral Satellite Imagery: A Case Study of Wuhan. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 15: 3972–3991.
- Yun, S.; Han, D.; Oh, S. J.; Chun, S.; Choe, J.; and Yoo, Y. 2019. CutMix: Regularization Strategy to Train Strong Classifiers with Localizable Features. In *International Conference on Computer Vision*, 6022–6031.
- Zhang, X.; Sun, Y.; Shang, K.; Zhang, L.; and Wang, S. 2016. Crop Classification Based on Feature Band Set Construction and Object-Oriented Approach Using Hyperspectral Images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 9(9): 4117–4128.
- Zhang, Z.; and Sabuncu, M. R. 2018. Generalized cross entropy loss for training deep neural networks with noisy labels. In *International Conference on Neural Information Processing Systems*, 8792–8802.
- Zhao, Z.; Xu, X.; Li, S.; and Plaza, A. 2024. Hyperspectral Image Classification Using Groupwise Separable Convolutional Vision Transformer Network. *IEEE Transactions on Geoscience and Remote Sensing*, 62: 1–17.
- Zhong, Y.; Hu, X.; Luo, C.; Wang, X.; Zhao, J.; and Zhang, L. 2020. WHU-Hi: UAV-borne hyperspectral with high spatial resolution (H2) benchmark datasets and classifier for precise crop identification based on deep convolutional neural network with CRF. *Remote Sensing of Environment*, 250: 112012.