Context-Aware Masking and Learnable Diffusion-Guided Patch Refinement in Transformers via Sparse Supervision for Hyperspectral Image Classification

Supplementary Material

6. Appendix

This appendix presents supplementary analyses and extended experimental results that complement the findings reported in the main paper to offer a more comprehensive understanding of our proposed framework.

6.1. More Details on Datasets

Hyperspectral data consists of images captured across hundreds of narrow, contiguous spectral bands, allowing each pixel to contain a detailed reflectance spectrum. This rich spectral information enables fine-grained material identification and discrimination that is not possible with conventional RGB or multispectral imagery. HSI datasets [1] are essential in remote sensing, and provide detailed spectral information across hundreds of bands. Among the most frequently studied datasets are **Indian Pines**, **Salinas Scene**, & **Botswana**, each offering unique characteristics and applications. The **Indian Pines** dataset, collected by the AVIRIS sensor over Indiana, USA, contains **145×145** pixels and **220 spectral bands**, covering wavelengths from **0.4 µm to 2.5 µm**. It mainly consists of agricultural fields and forested areas, with **16 ground truth classes** and approximately **10,249 labeled samples**. Classification on this dataset is challenging due to class imbalance, high spectral similarity among crop types, and the presence of mixed pixels.

The **Salinas Scene** dataset, also captured by AVIRIS, represents agricultural land in California's Salinas Valley. It features higher spatial resolution with **512×217** pixels, **224 spectral bands** (excluding **20 bands** affected by water absorption), and **16 land-cover classes**, with a total of approximately **54,129 labeled samples**. Salinas Scene is the largest among the three in terms of both spatial resolution and labeled data, making it especially well-suited for detailed agricultural studies.

Botswana dataset, acquired by NASA's Hyperion sensor aboard the EO-1 satellite, covers the Okavango Delta—an ecologically rich wetland. After removing water absorption bands, it includes **145 spectral bands** and is commonly cropped to **145×145** pixels from its original **256×1476** dimensions. It has **14 land cover classes** and around **3,248 labeled samples**. Botswana exhibits high spectral variation due to the diverse natural vegetation and wetland features, making it particularly valuable for environmental monitoring. Table 7 provides a comparative summary of these datasets. Common challenges in working with HSI include high dimensionality, spectral redundancy, and difficulty in distinguishing between spectrally similar classes.

6.2. Additional Experimental Setup Details

In this work, we employ a warm restart learning rate scheduler to enhance model convergence. The scheduler begins training with a preset learning rate and gradually reduces it using exponential decay. To avoid the model getting trapped in local minima or plateaus, the learning rate is reset after a considerable period for all three datasets. This cyclical reset enables the optimizer to explore new regions of the loss landscape, striking an effective balance between exploration and exploitation. To enhance model generalization, we apply simple yet effective augmentation techniques to the HSI data. Each training patch is randomly subjected to horizontal and vertical flips, followed by rotations of $90°$, $180°$, or $270°$.

Observation 1: Flatter and Dataset-Invariant Loss Landscape of CPDGAL

CPDGAL forms a well-defined convex loss landscape with a broad, flatter global minimum compared to supervised ViT. This flatness correlates strongly with improved generalization, enabling the model to avoid saddle points and suboptimal traps across datasets. The loss topology remains consistent despite different initializations. This ensures reliable convergence to optimal solutions.

6.3. Details on Evaluation Metrics

In hyperspectral image (HSI) classification, we employ three widely used evaluation metrics: **Overall Accuracy (OA)**, **Average Accuracy (AA)**, and the **Kappa coefficient** (κ). These metrics provide complementary perspectives on classification performance, particularly in the presence of class imbalance. OA is defined as the proportion of correctly classified pixels across all classes. OA is intuitive and easy to interpret, but can be misleading for imbalanced datasets, as large classes dominate the metric.

AA computes mean of per-class accuracies. It balances classification performance across all classes. AA treats each class equally, regardless of its size, making it particularly important in HSI datasets where minority classes are often underrepresented. κ is the agreement between predicted and true labels, adjusted for chance. κ is more robust than OA in imbalanced settings, as it discounts the accuracy that could be achieved by random guessing. κ value close to 1 shows strong agreement, 0 indicates random performance, and negative values indicate disagreement.

Table 7. Details on Indian Pines, Salinas, and Botswana Hyperspectral Datasets

Feature	Indian Pines	Salinas	Botswana
Location	Indiana, USA	California, USA	Okavango Delta, Botswana
Sensor	AVIRIS (airborne)	AVIRIS (airborne)	Hyperion (satellite)
Spatial Size	145 × 145 px	512 × 217 px	256 × 1476 (cropped to 145×145) px
Spectral Bands	220	224 (20 removed)	145
Ground Truth Classes	16 (crops, forest)	16 (agriculture)	14 (natural land cover)
Labeled Samples	10,249	54,129	3,248
Primary Use	Agricultural land classification	Crop classification	Environmental analysis
Major Challenge	Spectral similarity between crops	High-resolution spectral variation	Complex vegetation-water interactions

6.4. Analysis of Results with Additional SOTA

To ensure a thorough evaluation, we compare our proposed CPDGAL approach with four additional state-of-the-art (SOTA) methods identified through an extensive literature review. **SSAN** [27] introduced the Spectral-Spatial Attention Network, which incorporates an attention module into a simple spectral-spatial framework to reduce the impact of interfering pixels near land-cover boundaries. **SST-FA** [10] proposed the Spatial-Spectral Transformer (SST), combining CNNs for spatial representation with a modified transformer to model spectral sequences. This hybrid architecture demonstrates how attention-based models can outperform traditional CNNs in hyperspectral image classification tasks. **3DSA-MFN** [20] presented a 3D Self-Attention Multiscale Feature Fusion Network that integrates multiscale convolutions with a 3D self-attention mechanism to effectively capture both local and long-range dependencies. **SSL** [13] developed a self-supervised learning framework that reconstructs the central pixel of a hyperspectral patch using global context. By embedding spatial priors into the transformer architecture, this method addresses the lack of inductive bias noted by [32]. Additionally, it combines pixel-wise reconstruction with metric space projections to learn both local and global features.

Table 8 presents a comparative evaluation of CPDGAL against these methods on the Indian Pines (IP) and Salinas datasets. CPDGAL achieves the highest performance across both datasets in terms of overall accuracy (OA) and Kappa coefficient (κ). On **IP**, CPDGAL achieves an OA of **97.34%** and κ of **97.21**, outperforming the second-best SSL method [13] by **+0.79%** in OA and **+1.11** in κ. It also surpasses 3DSA-MFN [20] (96.02% OA) and SSAN [27] (95.49% OA), while **SST-FA** [10] lags notably behind, emphasizing the performance gap between newer transformer-based/self-supervised models and earlier architectures. On the **Salinas** dataset, where most methods already yield near-saturation accuracy, CPDGAL still leads with **99.87%** OA and **99.81** κ, slightly outperforming SSL (99.85% OA, 99.75 κ) and 3DSA-MFN (99.72% OA, 99.13 κ). The improvement in the Δ row further highlights CPDGAL's edge, with gains of **+0.79% OA** and **+1.11** κ on IP, and smaller but meaningful margins on Salinas. The average per-class accuracy (AA) highlights the performance of the methods across all classes, particularly on minority classes. On the IP dataset, our method achieves an AA of **96.74%**, outperforming the best baseline (SSAN) by **2.57%**, and indicates a strong robustness to class imbalance. On Salinas, our technique attains an AA of 99.41%, slightly lower than SSL (99.73%), yet it achieves the highest **OA** and κ, showing that it maintains an overall classification balance while slightly trading off on uniform per-class performance. These show CPDGAL's strong classification performance and robust generalization across diverse data settings.

Observation 2: Optimal Model Complexity

Increasing embedding dimension D, number of heads h, and layers L improves accuracy until saturation at $(32, 8, 6)$. Further scaling leads to diminishing returns and possible overfitting.

6.5. Convergence Trends through Active Learning

Figure 8 shows the training and validation accuracy curves over different active learning cycles. For the *Indian Pines* dataset, using a 90% hold-out test set and a 10% training set, the configuration with 5% initial labeled data followed by 2% acquisitions per cycle demonstrates stable and efficient convergence. The model reaches approximately 80% accuracy by the 3^{rd} cycle, surpasses 90% at the 12^{th} cycle, and exceeds 97% accuracy after the 22^{nd} cycle. Beyond this point, the performance gradually stabilizes and indicates diminishing returns from further active learning cycles.

For *Botswana*, we use 90% hold-out test set & 10% training set similar to *IP*. With a 5% initial labeled fraction and iterative acquisition of 2% of the remaining unlabeled pool per cycle, the model exhibits a rapid and stable learning trajectory. Accuracy improves from approximately 50% in initial cycles to 90% by the 10^{th} cycle, surpasses 98% between the 19^{th} and 21^{st} cycles, and subsequently stabilizes. Beyond the 25^{th} cycle, performance plateaus, indicating that additional cycles yield negligible gains. This behavior underscores the efficiency of moderate initial supervision in driving convergence while avoiding early saturation effects.

For the *Salinas* dataset, we employ a more constrained

Table 8. Comparison with additional SOTA methods on Indian Pines and Salinas datasets.

Methods		Indian Pines			Salinas		
		OA (%)	AA (%)	κ	OA (%)	AA (%)	κ
SSAN [27]	*TGRS '20*	95.49	94.17	94.85	96.81	98.33	96.54
SST-FA [10]	*RS '21*	88.98	68.15	86.70	94.94	93.05	94.32
3DSA-MFN [20]	*RS '22*	96.02	93.89	94.78	99.72	99.32	99.13
SSL [13]	*ICLR '23*	96.55	93.12	96.10	99.85	**99.73**	99.75
OURS		**97.34**	**96.74**	**97.21**	**99.87**	99.41	**99.81**
Δ		**+0.79**	**+2.57**	**+1.11**	**+0.02**	-0.32	**+0.06**

training set of 2% and a 98% test set. Here, the initial labeled fraction is relatively high at 10%, and the same fraction is acquired in each subsequent cycle for optimal performance. Owing to the richer supervision from the outset, the model converges significantly faster, reaching near-perfect accuracy by the 6^{th} cycle. This shows the strong influence of initial sample richness on active learning efficiency in high-resolution hyperspectral datasets.

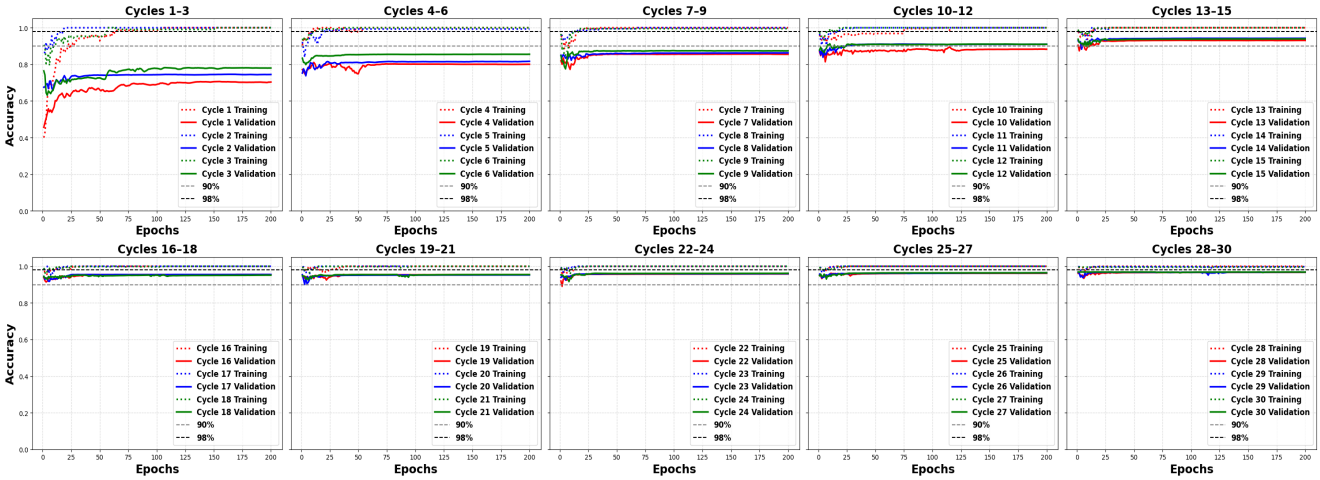
6.6. Impact of Pre-training with Limited Samples on Active Learning Convergence

Fig. 9 shows the impact of the pre-training phase with limited samples on active cycles for *Botswana* dataset. Fig. 9 (a) demonstrates that in the 1% initial pre-training setup, the model begins with an accuracy ranging from 35% to 50% in the first three cycles and shows a steady upward trend across subsequent active learning phases. By the 13^{th} cycle, accuracy surpasses the 90% threshold. Further improvements are observed until the 25^{th} to 27^{th} cycles, where the model reaches the 98% mark. Beyond this point, accuracy stabilizes slightly above 98%, indicating convergence with minimal fluctuations in the remaining cycles.

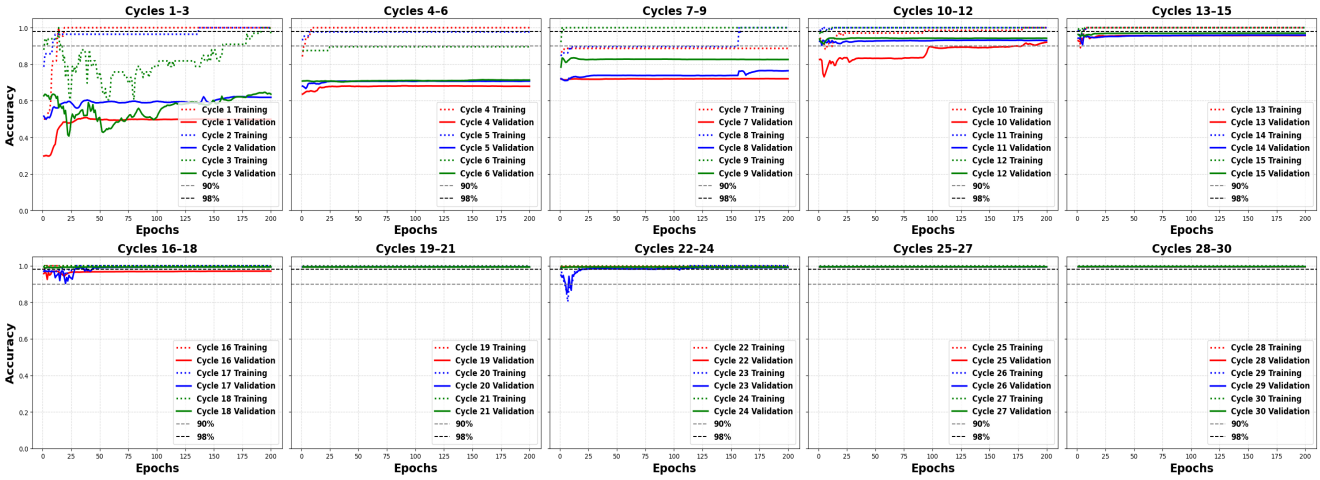
In the 5% initial pre-training configuration (Fig. 9 (b)), the model exhibits an early-phase accuracy between 50% and 60% across cycles 1–3, reflecting the beneficial effect of a richer pre-trained feature space in accelerating subsequent active learning gains. The performance trajectory demonstrates a rapid improvement, surpassing the 90% threshold by the 10^{th} cycle. High-confidence convergence is achieved between cycles 19–21, where accuracy exceeds 98%. Post-convergence, the performance profile remains remarkably stable, with minimal inter-cycle variance and substantial overlap of learning curves beyond the 25^{th} cycle, indicating diminishing marginal returns from further retraining. The empirical "sweet spot" for this setup is observed between cycles 20–25, the point where the cost of additional sampling no longer yields significant accuracy improvements.

In the 10% initial pre-training configuration (Fig. 9 (c)), the model benefits from a substantially more informative initialization, attaining close to 80% accuracy by cycle 3, which is markedly higher than the 5% setup during the same phase. The 90% benchmark is reached by cycles 10–12, followed by a progression to above 98% accuracy in the cycle 19–21 range. Stability is largely maintained between cycles 22–24, although a transient degradation in accuracy is observed around the 25^{th} cycle, which may be attributable to catastrophic forgetting effects. When the initial labeled set is large (e.g., 10% setup), the early-stage model exhibits higher accuracy due to richer supervision. However, this also means that the active learning acquisition in subsequent cycles has fewer truly informative or diverse samples to select from, and this can increase the chance of sample selection bias. Over multiple cycles, such bias can introduce catastrophic forgetting effects, particularly if newly acquired samples disproportionately represent a subset of the feature space, leading to temporary degradation in accuracy before re-stabilization.

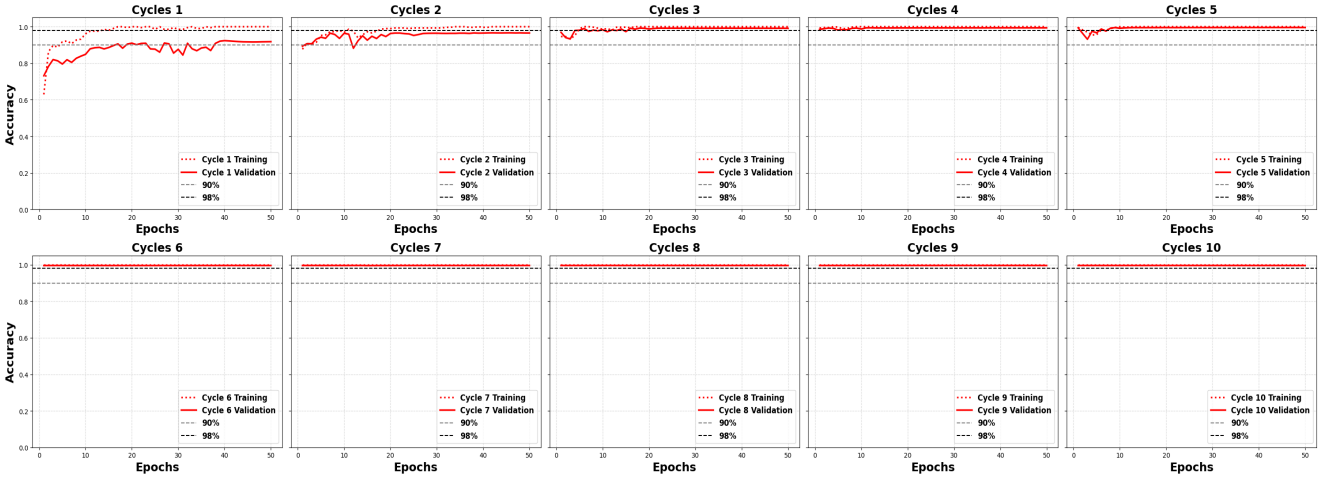
The system subsequently recovers, achieving late-phase stabilization between cycles 27–30. This final convergence occurs marginally later than in 5% setup. This implies that larger initial pre-training can introduce different long-term stability characteristics. Such delayed stabilization may reflect a more gradual resolution of residual high-uncertainty regions within the feature space, even after rapid early improvements. The experimental protocol is designed to evaluate the impact of varying initial labeled set sizes on the convergence dynamics of active learning. The full dataset is partitioned into 10% training and 90% holdout test subset to ensure unbiased evaluation. From the 10% training subset, an initial labeled pool is randomly sampled to initialize the model. In each active learning cycle, an additional 2% of samples is selected from the remaining training subset (labels were not used) based on our acquisition function and incorporated into the labeled set for retraining. This acquisition–retraining loop is repeated for a total of 30 cycles and visualized in Fig. 9. The only factor varied across the three experimental setups is the size of the initial labeled set: (a) 1% of the training subset (*low-resource initialization*), (b) 5% (*medium-resource initialization*), and (c) 10% (*high-resource initialization*). All other hyperparameters, model architectures, and training configurations are held constant across experiments.



(a)

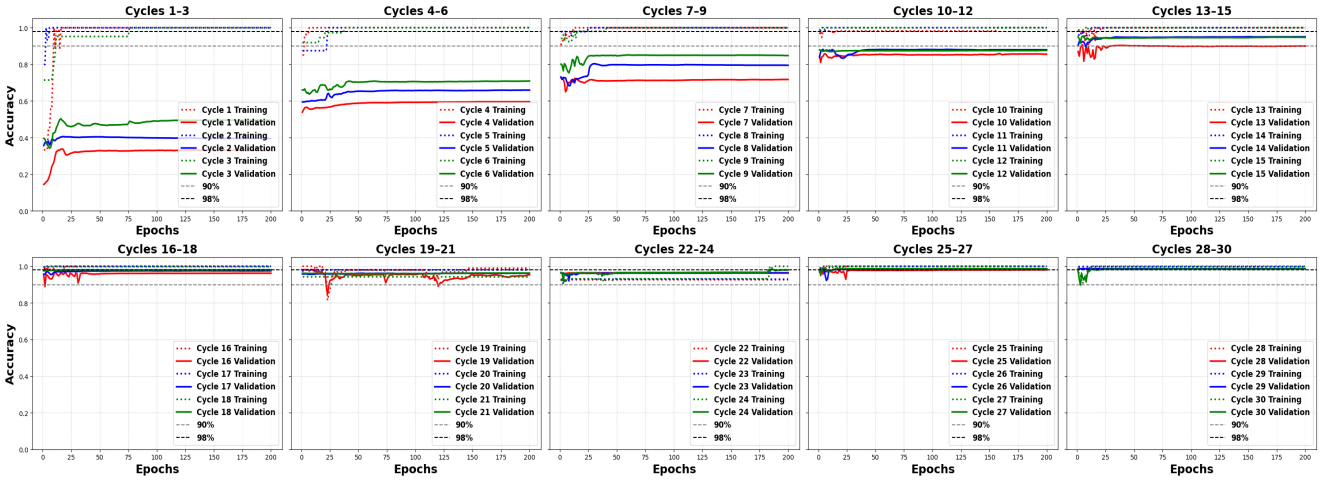


(b)

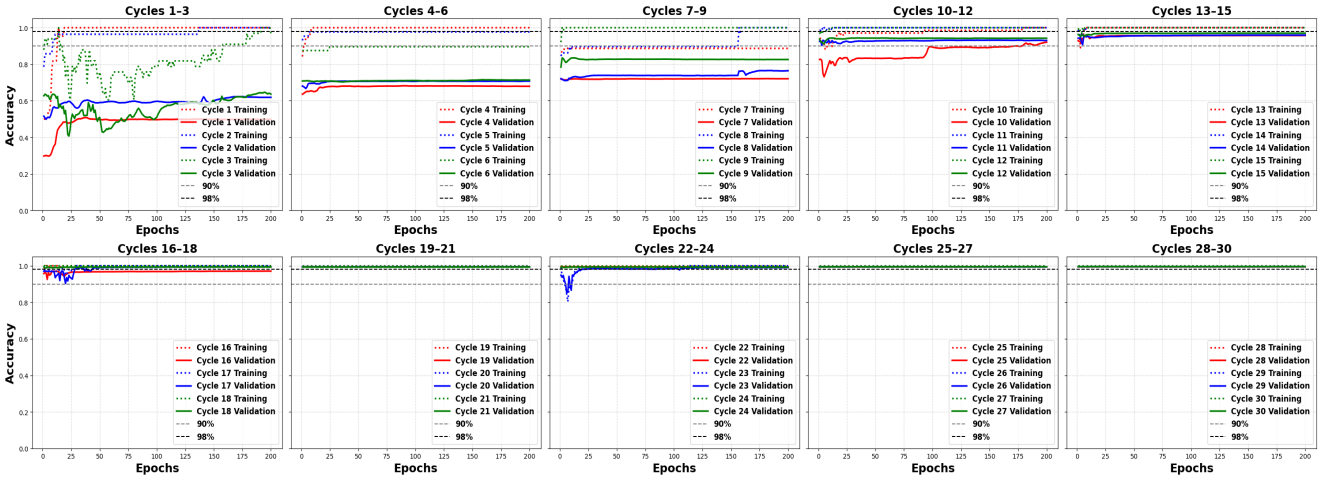


(c)

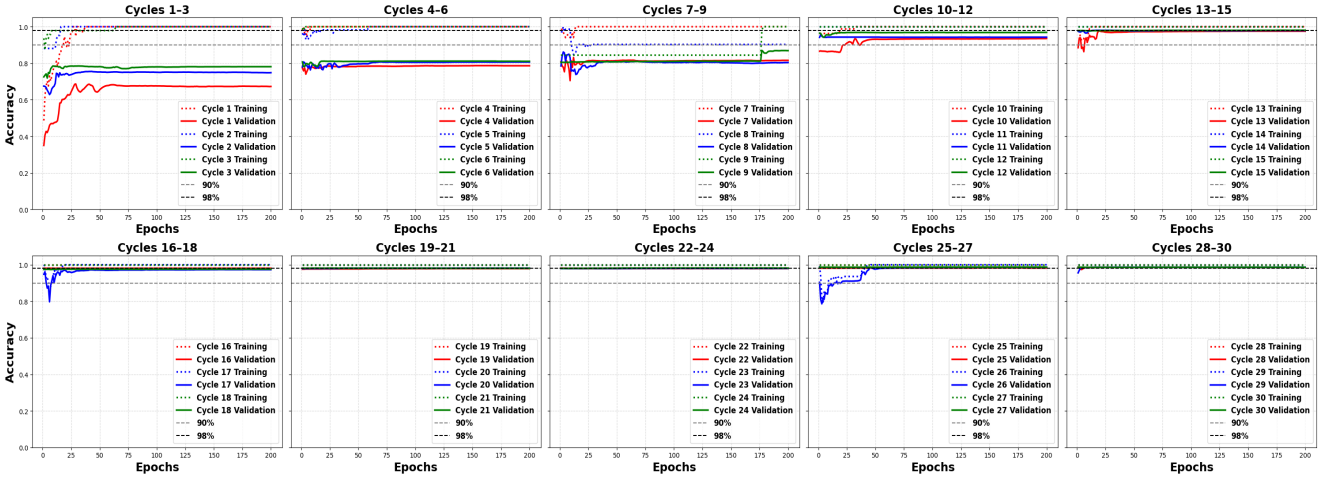
Figure 8. Evolution of validation and training accuracy across successive active learning cycles on three datasets. The top row corresponds to the **Indian Pines** dataset, the middle row to **Botswana**, and the bottom row to **Salinas**. For each dataset, the dotted lines depict the progression of training accuracy, while the bold lines present the corresponding validation accuracy on the held-out test set.



(a)



(b)



(c)

Figure 9. Impact of pre-training with limited samples on active learning convergence for Botswana: The dataset is split into 10% training and 90% hold-out test sets. From the 10% training set, a random initial labeled set is selected, followed by iterative retraining using 2% of samples from the remaining unlabeled set per cycle. The size of the initial labeled set is varied: (a) 1%, (b) 5%, and (c) 10%.

6.7. Expected Model Improvement (EMI) Analysis

To quantitatively assess the efficiency of label utilization in our active learning pipeline, we compute the *Expected Model Improvement* (EMI), formally defined as the ratio between the incremental change in validation accuracy and the corresponding increase in the number of labeled samples. This metric captures the *marginal utility per annotation*, and enables a fine-grained comparison of how effectively each newly acquired label contributes to model performance across acquisition cycles. It is defined as:

$$\text{EMI} = \frac{\Delta\text{Accuracy}}{\Delta N_{\text{labels}}}$$

For active learning, ΔAccuracy is measured between consecutive cycles, and ΔN_{labels} is the number of newly labeled samples acquired in that cycle. A high EMI value indicates that the model achieves a substantial performance gain per additional labeled sample, and this reflects high *label efficiency*. As active learning progresses, EMI typically decreases, revealing the onset of *diminishing returns*, where each new label provides progressively less improvement. Unlike absolute accuracy curves, EMI captures *marginal utility* of labeling decisions. This enables a direct comparison of the cost-effectiveness of different strategies. This perspective is crucial in scenarios where annotation costs are significant, and traditional accuracy plots fail to reveal the relative efficiency of label acquisition. EMI directly aligns with the core objective of active learning: maximizing performance gains while minimizing labeling effort. Trajectories of EMI for different acquisition budgets have been depicted in Fig. 6.

6.8. Evolution of Entropy Distribution During Active Learning

The entropy distribution heatmap captures the temporal evolution of model uncertainty within the *remaining unlabeled pool* over successive active learning cycles. Fig. 7 shows that, at the outset, the distribution is skewed towards higher-entropy bins, reflecting substantial predictive uncertainty across a large pool of unlabeled instances. As cycles progress, pool size decreases due to iterative acquisition and labeling, and the distribution's mass shifts towards lower-entropy bins. This indicates that the model becomes increasingly confident in the unlabeled remainder. Nevertheless, a non-negligible fraction of samples consistently occupy higher-entropy bins in later cycles, revealing the presence of intrinsically hard or ambiguous cases that resist confident classification. This persistence is a hallmark of uncertainty-driven querying: the model repeatedly targets the most informative (high-entropy) instances, thereby concentrating labeling resources on samples with the highest potential for marginal utility, even as the overall uncertainty landscape contracts.

As seen in the main paper, for the **2% acquisition budget**, the heatmap shows a globally lighter color distribution, which indicates fewer samples in the low-entropy region across all cycles. This pattern reflects that a larger acquisition rate rapidly removes low-uncertainty samples from the unlabeled pool (remaining training corpus), which accelerates model convergence and prevents the accumulation of redundant data points. The relative absence of deep yellow near the low-entropy bins towards the end implies that the model confidently generalizes across the dataset, with little residual ambiguity in predictions. The evolution of low-entropy mass can be formalized via the per-cycle entropy distribution $p_t(e)$, where $e \in [0, \log(14)] \approx [0, 2.64]$ (with 14 classes present in the Botswana dataset). After discretization into bins indexed by i, the entropy drop is

$$\Delta\mathbb{H}_t = \mathbb{H}(p_{t-1}) - \mathbb{H}(p_t),$$
$$\text{where, } \mathbb{H}(p_t) = -\sum_i p_t(e_i) \log p_t(e_i),$$

which quantifies the uncertainty removed from the unlabeled pool at cycle t. We infer that the reduction in entropy can be directly related to the expected model improvement (EMI), allowing us to quantify the marginal utility of labeling additional samples. This can be expressed as

$$\text{EMI}_t \approx -\frac{\Delta\mathbb{H}_t}{\Delta N_{\text{labels}}},$$

and reveals that lower acquisition budgets amplify marginal utility but with high variance, producing oscillations; whereas, larger budgets smoothen $\Delta\mathbb{H}_t$ and stabilize EMI, explaining consistent convergence observed for 2% setting.

6.9. Alignment with Sustainability

Beyond technical results, the proposed model, CPDGAL, supports key sustainable development goals. Its efficient, adaptable design can aid precision agriculture, environmental monitoring, and biodiversity assessment. By reducing energy use and resource demands, it also contributes to lowering barriers for adoption of this model in resource-limited regions for long-term access as a sustainable monitoring tool.