LUMEN: Language-guided Unified Memory for ENcoding Hyperspectral Images for Mars Mineral Identification

Anonymous CVPR submission

Paper ID *****

Abstract

Identifying Martian minerals from sparse hyperspectral observations is fundamentally constrained by extreme label scarcity and the absence of large-scale hyperspectral pretraining. To this end, we introduce LUMEN (Language-guided Unified Memory for ENcoding hyperspectral images), a memory-augmented cross-modal framework that formulates hyperspectral recognition tasks as structured relational alignment between spectral embeddings and language-derived semantic prototypes. LUMEN employs an ultralight 3D spectral encoder to distill spatial–spectral structure into a compact latent manifold, which is softly aligned to textual embeddings through a learnable bilinear projection. To reconcile linguistic priors with spectral evidence, semantic prototypes are refined via learnable offsets and a confidence-aware gating mechanism that adaptively regulates the contribution of language priors and visual memory. A unified momentum-updated memory bank further stabilizes learning by integrating exponentially smoothed visual representations with language-anchored prototypes, reducing cross-modal drift. Unlike prior language-guided or CLIP-style approaches, LUMEN explicitly enforces eigenspace-level cross-modal consistency through a spectral relational distillation objective that promotes global manifold alignment beyond pointwise embedding matching. Extensive evaluations of three Martian hyperspectral datasets show that LUMEN consistently exceeds strong transformer-based approaches while requiring fewer parameters and reduced training time.

1. Introduction

Hyperspectral imagery (HSI) [10, 14] captures dense spectral measurements across hundreds of contiguous wavelength bands, and provides detailed reflectance signatures that encode the physicochemical properties of surface materials. Unlike multispectral imaging, which samples sparsely across broad bands, hyperspectral sensing enables fine-grained discrimination of minerals through subtle absorption features associated with vibrational overtones, electronic transitions, hydration states, and crystal chemistry [6]. These spectral fingerprints make HSI a foundational tool for material identification in geology, environmental monitoring, and planetary science.

Hyperspectral remote sensing has significantly advanced our understanding of the geological and climatic history of **Mars**. Orbital instruments such as the **Compact Reconnaissance Imaging Spectrometer for Mars (CRISM)** [13] and the **Observatoire pour la Minéralogie, l'Eau, les Glaces et l'Activité (OMEGA)** [1] have acquired high-dimensional visible-to-near-infrared spectral cubes, enabling large-scale mineralogical mapping. These observations have revealed widespread phyllosilicates, sulfates, carbonates, hydrated silica, and mafic minerals, providing strong evidence for past aqueous alteration, volcanic activity, and surface weathering. Spectral analysis has therefore played a crucial role in reconstructing paleoenvironmental conditions and identifying astrobiologically relevant targets. Despite these advances, automated analysis of Martian hyperspectral imagery (HSI) remains highly challenging [2]. Planetary HSI is characterized by extremely high spectral dimensionality, complex noise induced by sensor artifacts and atmospheric effects, variable illumination geometry, and sub-pixel mixing, all of which increase uncertainty and complicate spectral interpretation.

More critically, labeled data are severely limited. Unlike terrestrial HSI applications—where large annotated datasets enable supervised deep learning [10, 14]—Martian labels are typically derived from expert spectral library matching or limited rover-based measurements, making large-scale pixel-level supervision infeasible. **In this work, we focus on accurate identification of scientifically significant minerals at prospective landing sites, where reliable compositional mapping is essential for mission planning and astrobiological exploration.**

Although convolutional neural networks and transformer-based models have achieved strong performance in terrestrial remote sensing, they rely on either

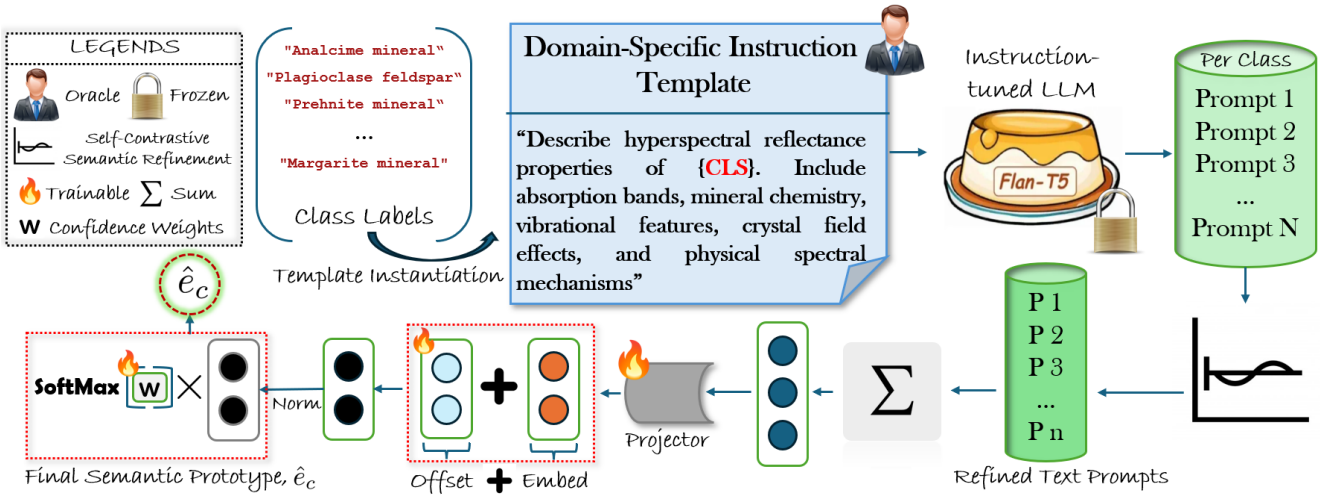

Figure 1. **Text Generation in LUMEN**: LLM-based Prompt Generation and Refinement Strategy.

substantial labeled data or large-scale pretraining. **In planetary hyperspectral settings, neither condition is satisfied**. The absence of hyperspectral foundational models and the domain gap between RGB imagery and high-dimensional spectral data hinder effective transfer learning. Consequently, the central challenge in Martian HSI analysis is to develop learning frameworks capable of inducing discriminative spectral structure under extreme label scarcity.

These constraints motivate approaches that move beyond purely data-driven supervision and instead integrate structured semantic knowledge into the learning process. Mineral classes possess well-documented spectral characteristics rooted in vibrational physics and crystal chemistry. Encoding such domain knowledge through language or semantic priors offers a promising avenue for guiding representation learning when pixel-level supervision is sparse. In this work, we, therefore, explore cross-modal alignment between hyperspectral embeddings and language-derived semantic prototypes, seeking to propagate semantic constraints onto spectral manifolds through structured memory and geometric consistency objectives. We next review related work in hyperspectral representation learning for planetary remote sensing.

Traditional Deep Learning Techniques. Early Hyperspectral image classification (HSIC) methods mainly employed CNNs. Yu et al. [26] introduced 2D CNNs for spectral–spatial feature extraction, later extended to 3D CNNs by Li et al. [11] for joint spectral–spatial modeling, while Zhong et al. proposed SSRN [27] to enhance discriminative learning.

Attention-based Methods. Attention mechanisms such as squeeze-and-excitation (SE) [8] and efficient channel attention (ECA) [21] enhance HSIC by emphasizing informative features [5, 20]. Notable methods include A2S^{2}K-ResNet [17]; however, CNN-based attention remains constrained by local receptive fields and limits global context.

Transformer-based Methods. Transformer architectures [3, 5, 20] have gained increasing attention due to their ability to capture long-range dependencies via multi-head self-attention. Representative approaches include SpectralFormer [7], which employs group-wise spectral embedding and adaptive fusion, the group-aware hierarchical transformer (GAHT) [12] for localized spectral–spatial attention, and SPRLT-Net [24], which models fine-grained spatial relationships by treating each pixel as a token.

Hybrid Techniques. CNN–Transformer hybrids combine local feature extraction with global context modeling. Representative methods include SSFTT [19], CDSFT [15], and MorphFormer [18]. More recently, MCTGCL [22] integrates spectral–spatial modeling, attention, morphological features, and graph contrastive learning for Martian HSIC.

Multimodal Approaches. Recent advances in vision–language adaptation highlight different strategies for improving generalization under limited supervision. CoOp [28] optimizes learnable prompt tokens for models like CLIP [16] to replace handcrafted text prompts, which improves few-shot performance but relies on pretrained models. Another work TCP [25] introduces class-aware prompt tuning through textual knowledge embedding, and enables dynamic generation of semantically enriched prompts that better generalize to unseen classes. Moving beyond prompt-level adaptation, RAda [4] focuses on recalibrating the final fused cross-modal representation (the rational matrix) using a lightweight attention-based mask, allowing targeted adjustment of decision-level interactions without modifying backbone encoders. While these methods significantly advance low-shot adaptation in natural images, hyperspec-

tral image (HSI) analysis—especially for planetary mineral identification—faces distinct challenges: strong inter-band correlation, spectral variability caused by illumination and atmospheric effects, domain shift across sensors and terrains, and the absence of large-scale hyperspectral pre-training. Additionally, HSI data lie on complex spectral manifolds where subtle absorption features determine class boundaries, making prior prompt-based strategies insufficient. These constraints require lightweight, domain-aware, and semantically guided frameworks that can robustly align the spectral structure with prior knowledge while maintaining stability under limited supervision.

To this end, we propose a language-guided hyperspectral few-shot learning framework that adaptively reconciles LLM-derived semantic priors with spectral evidence through a unified memory architecture. Instead of relying on fixed text prototypes, class semantics are dynamically refined via learnable embedding offsets and confidence-aware prompt gating, while a momentum-updated visual memory stabilizes prototype evolution. Additionally, a spectral relational distillation objective aligns the eigenspace geometry of semantic and visual manifolds, unlocking discriminative cross-modal learning under extreme data scarcity. **Our main contributions are summarized as follows:**

- We propose LUMEN, a language-guided hyperspectral classification framework that injects large language model semantic priors into few-shot spectral learning via a class-specific semantic refinement module. Learnable offsets and confidence-aware gating enable adaptive prototype evolution under low-shot scenarios.
- We introduce a momentum-updated visual memory that stabilizes cross-modal prototype dynamics and mitigates representation drift in label-scarce regimes.
- We formulate a relational distillation objective that aligns the eigenspace geometry of visual and semantic similarity matrices, through a bilinear projection, and enforces structural cross-modal consistency beyond pointwise embedding matching.
- Experiments on Martian hyperspectral benchmarks show improvements over strong transformer-based baselines with significantly fewer parameters.

2. Methodology

Let $\mathcal{X} \subset \mathbb{R}^{H\times W\times B}$ denote the hyperspectral input space, where H, W, and B represent height & width of the image and spectral bands, respectively. We consider a few-shot classification problem with C classes and a limited labeled set $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$, where N is small relative to the intrinsic dimensionality of $\mathcal{X}$. In such regimes, purely data-driven models struggle to learn discriminative spectral representations.

To address this drawback, *LUMEN* injects mineralogical prior knowledge through large language models (LLMs). For each class c, multiple descriptive prompts are generated and encoded into semantic embeddings. After redundancy reduction and adaptive projection into a shared semantic space of dimension D', the refined embeddings are aggregated into a per-class semantic prototype $\hat{e}_c \in \mathbb{R}^{D'}$.

In parallel, a hyperspectral encoder f_θ maps each input sample to a visual embedding. A learnable cross-modal projection aligns visual features with semantic prototypes via similarity matching in the shared space. To enhance stability under low supervision, LUMEN maintains a momentum-updated memory bank that integrates visual representations with language-derived prototypes, enabling structured cross-modal alignment beyond simple instance-level matching. *The detailed components (figs. 1 and 2) of the framework are presented in the following subsections.*

Semantic Prototypes via Language Prompts. For each class $c \in \{1, \dots, C\}$, we generate a set of P textual prompts $\mathcal{T}_c = \{t_{c,j}\}_{j=1}^{P}$ using an LLM to describe class-specific hyperspectral characteristics, including absorption bands, vibrational features, and crystal chemistry. Each prompt $t_{c,j}$ is embedded into a semantic vector using a frozen LLM encoder h_ψ: $e_{c,j} = h_\psi(t_{c,j}), \quad e_{c,j} \in \mathbb{R}^{D_t}$, where D_t denotes the raw textual embedding dimension.

Since directly using all generated prompts may introduce redundancy and propagate linguistic inaccuracies, we apply a self-contrastive refinement mechanism. While LLMs provide rich semantic priors, they may generate partially redundant or hallucinated descriptions, particularly in specialized domains such as hyperspectral mineralogy. The initial prompts thus serve as informative but imperfect semantic initializations and hence the refinement process quantifies semantic overlap among prompts and iteratively selects a diverse and informative subset. First, the pairwise cosine similarity between embeddings within class c is computed:

$$S_{ij}^{(c)} = e_{c,i}^{\top} e_{c,j}, \quad i, j \in \{1, \dots, P\}. \tag{1}$$

The first selected prompt minimizes mean similarity to all others:

$$i^* = \arg\min_i \frac{1}{P} \sum_{j=1}^{P} S_{ij}^{(c)}. \tag{2}$$

Let $\mathcal{S}_c$ denote the selected index set and define the remaining candidate set as $\mathcal{R}_c = \{1, \dots, P\} \setminus \mathcal{S}_c$. Additional prompts are selected iteratively to minimize similarity to the already chosen set:

$$r^* = \arg\min_{r\in\mathcal{R}_c} \frac{1}{|\mathcal{S}_c|} \sum_{s\in\mathcal{S}_c} e_{c,r}^{\top} e_{c,s}, \tag{3}$$

until a refined subset of P_{refined} prompts is obtained. This diversity-driven selection reduces redundancy and filters unreliable semantic descriptions prior to cross-modal alignment. Each refined embedding is then projected into a

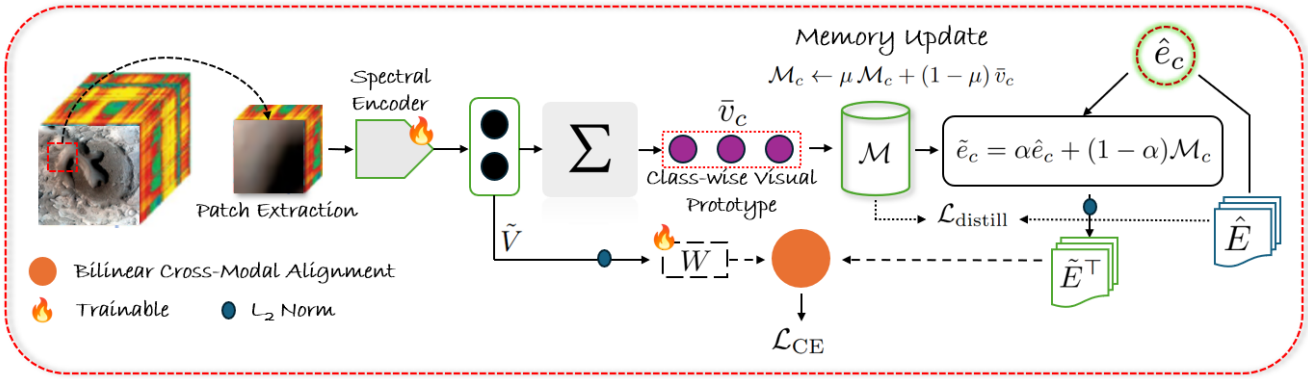

Figure 2. **Cross-Modal Alignment in LUMEN:** End-to-end alignment with semantic prototypes generated from LLM prompts.

lower-dimensional cross-modal space via a learnable linear projection:

$$\tilde{e}_{c,j} = \mathrm{Proj}(e_{c,j}), \quad \tilde{e}_{c,j} \in \mathbb{R}^{D'}, \quad D' \ll D_t. \tag{4}$$

To enable adaptive domain-specific refinement, a class-specific learnable offset $\Delta_{c,j} \in \mathbb{R}^{D'}$ is added, followed by L2 normalization:

$$z_{c,j} = \frac{\tilde{e}_{c,j} + \Delta_{c,j}}{\|\tilde{e}_{c,j} + \Delta_{c,j}\|_2}, \quad z_{c,j} \in \mathbb{R}^{D'}. \tag{5}$$

The refined embeddings are then aggregated using a gated attention mechanism. Learnable scalars $a_{c,j}$ are transformed via softmax to produce normalized weights:

$$w_{c,j} = \frac{\exp(a_{c,j})}{\sum_{k \in \mathcal{S}_c} \exp(a_{c,k})}. \tag{6}$$

The final semantic prototype for class c is computed as the weighted sum of normalized embeddings:

$$\hat{e}_c = \sum_{j \in \mathcal{S}_c} w_{c,j}\, z_{c,j}, \quad \hat{e}_c \in \mathbb{R}^{D'}. \tag{7}$$

This formulation explicitly enforces semantic diversity through self-contrastive selection, adapts LLM-derived embeddings via learnable offsets, and employs gated aggregation to emphasize informative prompts while down-weighting redundant or less reliable descriptions.

Hyperspectral Embedding and Cross-Modal Alignment. Let the hyperspectral encoder be denoted by $f_\theta : \mathcal{X} \to \mathbb{R}^L$, which maps an input hyperspectral sample $x \in \mathcal{X}$ to a latent spectral-spatial representation $z = f_\theta(x)$ of dimension L. The encoder is fully learnable and designed to extract discriminative spectral–spatial features under few-shot supervision.

The encoder output is projected into the shared cross-modal space via a learnable vision projection module $\phi_v : \mathbb{R}^L \to \mathbb{R}^{D'}$:

$$v = \phi_v(f_\theta(x)), \quad v \in \mathbb{R}^{D'}. \tag{8}$$

The projected embedding is then L2-normalized:

$$\tilde{v} = \frac{v}{\|v\|_2}, \quad \tilde{v} \in \mathbb{R}^{D'}. \tag{9}$$

Memory-Augmented Prototype Stabilization. To enhance stability under low-shot conditions, we maintain a memory bank $\mathcal{M} \in \mathbb{R}^{C \times D'}$ that stores class-wise visual prototypes. For each class c, the memory vector is updated using a momentum-based moving average:

$$\mathcal{M}_c \leftarrow \mu\,\mathcal{M}_c + (1-\mu)\,\bar{v}_c, \tag{10}$$

where $\bar{v}_c$ denotes the mean normalized embedding of class c in the current batch, and $\mu \in [0,1]$ controls temporal smoothing. This mechanism reduces batch-level variance and provides a robust estimate of the evolving visual class geometry.

Prototype Fusion via Memory Blending. While semantic prototypes encode prior knowledge from language prompts, visual memory captures evolving spectral geometry. To integrate both sources, we compute fused class prototypes:

$$\tilde{e}_c = \alpha\hat{e}_c + (1-\alpha)\mathcal{M}_c, \tag{11}$$

where $\alpha \in [0,1]$ controls the trade-off between semantic prior and visual memory. The fused prototypes are L2-normalized before similarity computation.

Cross-Modal Similarity Computation. Let $\tilde{E} = [\tilde{e}_1, \ldots, \tilde{e}_C]^\top \in \mathbb{R}^{C \times D'}$ denote the fused class prototype matrix. To align hyperspectral embeddings with fused prototypes, we introduce a learnable cross-modal transformation matrix $W \in \mathbb{R}^{D' \times D'}$.

For a batch of N samples, stack normalized embeddings into $\tilde{V} \in \mathbb{R}^{N \times D'}$. The similarity matrix is computed as:

$$S = (\tilde{V}W)\,\tilde{E}^\top, \quad S \in \mathbb{R}^{N \times C}, \tag{12}$$

where each entry $S_{i,c} = \tilde{v}_i^\top W \hat{e}_c$ represents the bilinear similarity between sample i and class c.

A learnable logit scaling parameter β is applied before softmax normalization:

$$P = \mathrm{softmax}(\beta S), \quad P \in \mathbb{R}^{N \times C}. \tag{13}$$

The cross-entropy loss over the batch is expressed in matrix form as:

$$\mathcal{L}_{\mathrm{CE}} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{c=1}^{C} Y_{i,c} \log P_{i,c}, \tag{14}$$

where $Y \in \{0,1\}^{N \times C}$ denotes the one-hot encoded ground-truth labels. *All similarity computations are thus fully vectorized, and enables efficient parallel optimization*.

The memory bank $\mathcal{M}$ captures the evolving visual class manifold through exponential temporal smoothing in $\mathbb{R}^{D'}$, which can be viewed as a low-rank filtering operator that preserves dominant spectral directions while attenuating high-frequency fluctuations from small-batch sampling. The learnable matrix W performs a linear transformation that rotates and scales the visual embedding space to better align with semantic prototypes. The similarity formulation $S = (\tilde{V}W)\hat{E}^\top$ therefore defines a bilinear alignment capturing both normalized angular relationships and adaptive cross-modal scaling.

Spectral Relational Distillation. While pointwise cross-modal similarity enforces instance-level alignment, it does not explicitly constrain the global relational structure between classes. To preserve manifold geometry across modalities, we introduce a spectral relational distillation loss that aligns the dominant eigenspaces of the hyperspectral memory and semantic prototype similarity matrices.

Let the visual memory matrix be $\mathcal{M} \in \mathbb{R}^{C \times D'}$ and the semantic prototype matrix be $\hat{E} = [\hat{e}_1, \ldots, \hat{e}_C]^\top \in \mathbb{R}^{C \times D'}$. We emphasize that relational distillation operates on semantic prototypes prior to memory blending. This ensures that global alignment is enforced directly between language-derived class geometry and the evolving visual memory manifold. We construct their corresponding class-level similarity (Gram) matrices:

$$S_v = \mathcal{M}\mathcal{M}^\top, \quad S_t = \hat{E}\hat{E}^\top, \tag{15}$$

where both $S_v, S_t \in \mathbb{R}^{C \times C}$ encode inter-class relationships. We compute the top-k (Here, $k = 3$) eigenvectors of each similarity matrix: $S_v = V_v \Lambda_v V_v^\top, \quad S_t = V_t \Lambda_t V_t^\top$, and retain the dominant subspaces $V_v \in \mathbb{R}^{C \times k}$ and $V_t \in \mathbb{R}^{C \times k}$.

Instead of directly matching eigenvectors (which are sign-ambiguous), we align the associated projection operators:

$$\mathcal{L}_{\mathrm{distill}} = \left\| V_v V_v^\top - V_t V_t^\top \right\|_F^2. \tag{16}$$

This formulation enforces consistency between the principal relational subspaces of visual and semantic representations, and promotes alignment of global class geometry beyond pairwise embedding similarity.

Overall Objective. The final training objective combines classification and relational distillation:

$$\mathcal{L} = \mathcal{L}_{\mathrm{CE}} + \lambda_{\mathrm{distill}} \mathcal{L}_{\mathrm{distill}}, \tag{17}$$

where $\lambda_{\mathrm{distill}}$ balances discriminative supervision and structural regularization.

3. Analysis of Results

Datasets. We evaluate the proposed framework on **THREE** Martian hyperspectral datasets obtained from **CRISM** [9, 13, 23], covering diverse geological regions: **Holden Crater** (HC, 418×595 pixels, 440 bands) with evidence of past water activity; **Nili Fossae** (NF, 478×593 pixels, 425 bands), rich in carbonates and hydrothermal alteration products; and **Utopia Planitia** (UP, 478×595 pixels, 432 bands), a northern plain with subtle spectral contrasts. These datasets provide complementary test cases of varying geological complexity.

Experimental Setup. We adopt a unified preprocessing and training protocol across all datasets. Each hyperspectral cube is first reduced using Principal Component Analysis (PCA) to retain the top 15 components. Spatial–spectral patches of size $13 \times 13 \times 15$ are then extracted and normalized. The model is trained for 80 epochs using the Adam optimizer with learning rate $\eta = 1 \times 10^{-4}$ and batch size of 4. We set hyperparameter values at ($\mu = 0.9$, $\alpha = 0.5$, $P = 4$, $D' = 64$, $\lambda_{\mathrm{distill}} = 50$). Performance is consistently evaluated using Overall Accuracy (OA), Average Accuracy (AA), and Cohen's Kappa (κ).

Comparison with SOTA Methods. Table 1 shows that our approach consistently achieves the best performance across all three Martian hyperspectral datasets. On Holden Crater, our method attains **98.22% OA**, **97.15% AA**, and a κ score of **97.39**, outperforming both conventional CNN-based models and recent transformer-based architectures. Similar superiority is observed on Nili Fossae and Utopia Planitia, where our framework achieves substantial margins over strong multimodal approaches such as TCP [25] and RAda [4]. Notably, while multimodal variants require over 140 seconds of training time, our method completes training in only **12.9 seconds**, despite having a comparable parameter count (241.8K). This highlights a key advantage of our prototype-level alignment strategy: it delivers state-of-the-art accuracy with reduced computational cost.

Computational Efficiency. The proposed framework achieves computational efficiency by replacing instance-level contrastive comparisons with class-level prototype alignment. Unlike contrastive learning, which requires $\mathcal{O}(N^2 D')$ pairwise similarities within a batch, our method performs $\mathcal{O}(NCD')$ operations by comparing each sample only to C class prototypes. This eliminates the need for negative sampling, large memory banks of individual instances,

Table 1. Comparison of our method with SOTA approaches on three Martian Hyperspectral datasets. Model complexity is also reported for the Holden Crater dataset. Competitive methods are highlighted.

Methods	Holden Crater			Nili Fossae			Utopia Planitia			Model Complexity	
	OA (%)	AA (%)	κ	OA (%)	AA (%)	κ	OA (%)	AA (%)	κ	Parameters (K)	Training Time (s)
3DCNN$_{RS'17}$ [11]	80.39	85.08	72.53	62.76	69.80	54.77	78.04	81.29	70.85	132.7	**1.550**
SSRN$_{TGRS'18}$ [27]	87.89	91.42	82.76	73.19	77.34	67.02	80.89	86.16	74.44	135.9	2.200
A^{2}S^{2}K$_{TGRS'21}$ [17]	79.97	86.91	72.11	71.20	76.49	64.72	82.06	87.27	75.97	**105.3**	2.720
Spectral-F$_{TGRS'22}$ [7]	82.33	86.91	75.14	72.21	74.92	65.79	79.47	84.50	72.56	130.1	21.72
GAHT$_{TGRS'22}$ [12]	87.28	89.63	81.81	74.25	76.25	68.42	82.18	85.51	76.01	946.8	3.380
SSFTT$_{TGRS'22}$ [19]	86.32	90.23	80.50	71.91	76.02	65.44	83.56	86.86	77.85	152.8	1.820
Morph-F$_{TGRS'23}$ [18]	81.47	87.14	73.73	63.11	70.31	55.41	81.16	85.25	74.63	141.7	4.190
MCTGCL$_{TGRS'25}$ [22]	91.99	93.30	88.39	81.20	82.68	76.57	86.93	88.68	82.19	247.0	64.53
TCP$_{CVPR'24}$ [25]	96.96	93.47	95.50	90.07	85.74	87.56	86.45	88.18	81.41	203.3	144.0
RAda$_{ICCV'25}$ [4]	97.61	96.32	96.48	88.25	88.03	85.36	89.79	91.70	85.97	203.3	145.6
OURS	**98.22**	**97.15**	**97.39**	**94.31**	**93.16**	**92.84**	**93.57**	**93.70**	**91.16**	241.8	12.9

and cross-batch similarity computation. Furthermore, unlike recent multimodal approaches, which construct interaction tensors of size $N \times C \times D'$ and apply gating or attention layers leading to $\mathcal{O}(NCD'^2)$ complexity, our framework avoids per-class interaction modeling and dense transformations over expanded tensors. Instead, alignment is performed through compact prototype-level projection, resulting in linear scaling with respect to the batch size N.

Efficiency in Practice: By operating at the prototype level rather than the instance level, LUMEN reduces computational and memory overhead while preserving global class structure via memory stabilization and spectral relational alignment. The model contains only **241.8K trainable parameters** and requires **12.9 seconds** of training on the Holden Crater dataset. Although slightly larger than lightweight CNNs such as 3DCNN (132.7K) and SSRN (135.9K), it remains substantially more efficient than transformer-based and multimodal approaches (e.g., MCTGCL: 247.0K parameters, 64.53 s; multimodal variants: >128 s), while keeping performance intact.

Few-Shot Robustness and Scalability. Table 2 reports 1-, 2-, and 5-shot results across three Martian hyperspectral datasets. In the extreme 1-shot regime, all methods degrade as expected; however, our approach maintains strong class balance and competitive performance. It achieves the highest AA on Holden Crater (64.26%) and clearly outperforms TCP and RAda on Utopia Planitia (64.68% OA, 52.32 κ), while remaining competitive on Nili Fossae.

Under 2-shot supervision, performance gaps widen. Our method attains the best OA and κ on Holden Crater (91.82% OA, 88.07 κ) and consistently surpasses SOTAs on Utopia Planitia and Nili Fossae. This shows improved stability with moderate supervision. In the 5-shot setting, our method shows consistent superiority across datasets, achieving 96.71% OA (95.18 κ) on Holden Crater and leading results on Nili Fossae (85.37% OA) and Utopia Planitia (88.82% OA). Collectively, these results highlight two key properties: *strong robustness under extreme label scarcity and consistent scalability as supervision increases, enabled by prototype-level alignment and memory stabilization.*

Ablation Study. Table 3 presents an ablation study evaluating the contribution of each core component of LUMEN on the Holden Crater. The results show the importance of proper memory stabilization, balanced prototype blending, appropriate prompt capacity, and spectral supervision.

Effect of Memory Momentum (μ): The memory momentum parameter plays a critical role in stabilizing prototype updates. When μ is small (0.3–0.6), performance collapses (OA $< 50\%$) due to unstable memory accumulation. As μ increases, performance improves significantly, and shows stronger temporal consistency in class representations. The best performance is achieved at $\mu = 0.9$ (OA 98.22%, AA 97.15%, $\kappa = 0.9739$), while $\mu = 1.0$ slightly degrades accuracy due to overly rigid memory.

Table 2. Few-shot comparisons of LUMEN with other methods. Best results are highlighted in bold.

Holden Crater									
Method	1-Shot			2-Shot			5-Shot		
	OA	AA	κ	OA	AA	κ	OA	AA	κ
TCP [25]	50.68	65.38	38.10	91.73	**88.04**	87.98	94.00	92.14	91.25
RAda [4]	43.47	63.50	29.06	90.20	85.23	85.71	91.56	92.03	87.75
OURS	45.56	**64.26**	29.67	**91.82**	86.42	**88.07**	**96.71**	**95.72**	**95.18**
Nili Fossae									
Method	1-Shot			2-Shot			5-Shot		
	OA	AA	κ	OA	AA	κ	OA	AA	κ
TCP [25]	**36.07**	**43.82**	**25.65**	58.09	60.27	49.59	83.24	81.30	79.20
RAda [4]	30.06	38.75	19.39	64.94	63.49	57.43	81.56	80.99	77.18
OURS	35.33	37.27	22.90	**62.19**	**62.29**	**53.36**	**85.37**	**81.12**	**81.80**
Utopia Planitia									
Method	1-Shot			2-Shot			5-Shot		
	OA	AA	κ	OA	AA	κ	OA	AA	κ
TCP [25]	59.97	49.93	48.18	69.71	64.93	59.48	82.57	84.37	76.44
RAda [4]	53.81	52.40	41.01	72.11	64.57	61.73	79.00	81.62	71.31
OURS	**64.68**	**54.20**	**52.32**	**75.01**	**65.88**	**65.78**	**88.82**	**91.65**	**84.71**

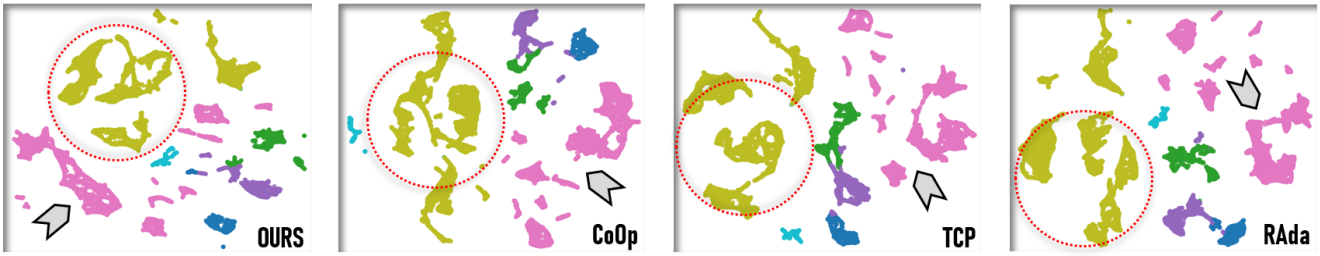

Figure 3. **UMAP Visualizations of the Embedding**: LUMEN Qualitatively Compared with Other Multimodal SOTA Approaches.

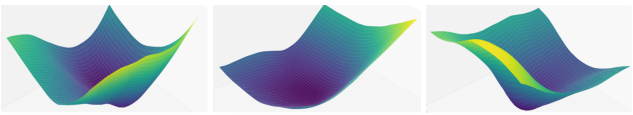

Figure 4. **Loss Landscape Visualization across Datasets.** From left to right: Holden Crater, Nili Fossae, and Utopia Planitia. The smooth and well-formed basins indicate stable optimization behavior and consistent convergence of the proposed framework across varying spectral distributions.

Effect of Memory Blending Coefficient (α): The blending coefficient controls the trade-off between current features and stored prototypes. Without blending ($\alpha = 0.0$), performance drops substantially (OA 89.80%), and shows the limitation of static prototypes. Performance peaks at $\alpha = 0.5$, and confirms that equal weighting between dynamic updates and memory retention yields the best res. Larger values ($\alpha > 0.8$) slightly reduce adaptability.

Effect of Number of Prompts (P): Increasing the number of prompts initially improves performance, with $P = 4$ achieving strong results (OA 98.22%). However, further scaling leads to diminishing returns, and excessive prompts (e.g., $P = 64$) cause performance degradation due to redundancy and over-parameterization.

Effect of Latent Dimension (D'): Latent dimensionality directly affects representational capacity. Very small dimensions ($D' = 4$ or 16) restrict feature expressiveness and significantly reduce accuracy. Increasing to $D' = 64$ yields the best performance (OA 98.22%), while larger dimensions ($D' = 256$) provide no additional benefit and slightly reduce stability.

Effect of Spectral Distillation Loss ($\mathcal{L}_{\text{distill}}$): The classification loss ($\mathcal{L}_{\text{CE}}$) serves as the primary supervision signal. Interestingly, even when both $\mathcal{L}_{\text{CE}}$ and $\mathcal{L}_{\text{distill}}$ are disabled, the model still achieves non-trivial performance (OA $\approx$ 48.61%). Since the image encoder is randomly initialized, this behavior does not stem from pretrained visual knowledge. Instead, it arises from the fixed semantic text prototypes and the structural bias of the architecture, which induces partial alignment between visual embeddings and class prototypes even without effective gradient-driven optimization. In this setting, predictions behave similarly to a weakly structured zero-shot classifier driven by randomly projected but normalized features. Enabling $\mathcal{L}_{\text{distill}}$ alone slightly improves performance by enforcing spectral consistency, but it cannot replace discriminative supervision. In contrast, activating $\mathcal{L}_{\text{CE}}$ restores strong class separation (OA 97.82%), and the joint optimization of $\mathcal{L}_{\text{CE}}$ and $\mathcal{L}_{\text{distill}}$ achieves the highest performance (OA 98.22%). This demonstrates that the two losses play orthogonal roles.

Table 3. **[All-in-one]** Ablation study to verify the efficacy of different components of LUMEN on Holden Crater dataset.

Setting	OA (%)	AA (%)	κ
Effect of Memory Momentum (μ)			
0.3	22.51	24.55	0.1335
0.4	42.46	39.73	0.2910
0.5	48.30	46.86	0.3623
0.6	50.32	47.69	0.3790
0.7	77.55	61.01	0.6736
0.8	96.58	92.72	0.9495
0.9	**98.22**	**97.15**	**0.9739**
1.0	97.36	95.42	0.9612
Effect of Memory Blending Coefficient (α)			
0.0	89.80	73.32	0.8485
0.2	89.27	74.29	0.8417
0.4	97.76	95.21	0.9669
0.5	**98.22**	**97.15**	**0.9739**
0.6	**98.07**	**96.25**	**0.9715**
0.8	97.75	96.10	0.9670
1.0	96.63	95.70	0.9507
Effect of Number of Prompts (P)			
2	97.84	96.53	0.9683
4	**98.22**	**97.15**	**0.9739**
8	**98.14**	**96.75**	**0.9726**
16	97.75	95.93	0.9669
32	**98.23**	**96.69**	**0.9739**
64	94.99	86.81	0.9259
Effect of Latent Dimension (D')			
4	78.64	53.63	0.6831
16	89.03	73.75	0.8392
64	**98.22**	**97.15**	**0.9739**
256	97.23	96.69	0.9594
Effect of Spectral Distillation Loss ($\mathcal{L}_{\text{distill}}$)			
$\mathcal{L}_{\text{cls}}$ (×), $\mathcal{L}_{\text{sd}}$ (×)	48.61	48.97	0.3623
$\mathcal{L}_{\text{cls}}$ (×), $\mathcal{L}_{\text{sd}}$ (✓)	54.12	51.61	0.4177
$\mathcal{L}_{\text{cls}}$ (✓), $\mathcal{L}_{\text{sd}}$ (×)	97.82	96.66	0.9680
$\mathcal{L}_{\text{cls}}$ (✓), $\mathcal{L}_{\text{sd}}$ (✓)	**98.22**	**97.15**	**0.9739**

Embedding Analysis. The UMAP visualizations in

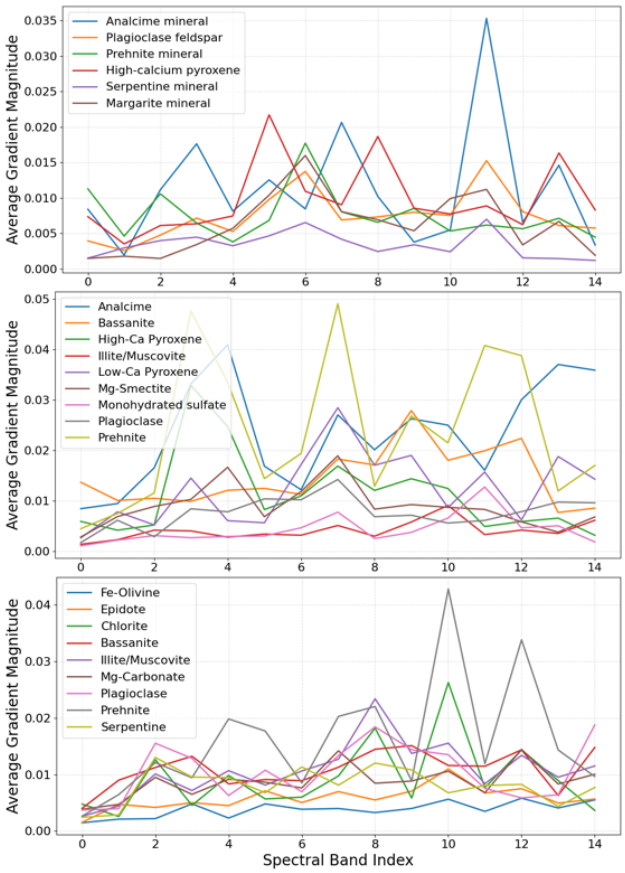

Figure 5. Class-wise average gradient magnitude across spectral bands (top to bottom: Holden Crater, Utopia Planitia, and Nili Fossae). The plots reveal dataset-dependent spectral sensitivity patterns, highlighting how the model adapts band importance according to mineral complexity and inter-class separability.

Fig. 3 qualitatively focus on the structural advantages of our framework. Compared with CoOp [28], TCP [25], and RAda [4], our method produces more compact and well-separated class clusters, with dense intra-class cohesion and clear inter-class boundaries. In contrast, CoOp exhibits fragmented clusters, TCP shows higher intra-class dispersion, and RAda presents residual class overlap, which indicates weaker cross-modal alignment. CoOp underperforms as it relies on large-scale CLIP pretraining, which is unavailable for hyperspectral imagery; without domain-aligned pretraining, prompt tuning alone cannot ensure strong spectral–semantic alignment. In contrast, our prototype alignment enforce structured cross-modal consistency, producing a more discriminative latent space.

Loss-Landscape Visualization. The loss landscape visualizations in Fig. 4 reveal stable optimization behavior across the three Martian hyperspectral datasets. From Holden Crater to Nili Fossae and Utopia Planitia, the surfaces exhibit relatively smooth curvature with well-defined low-loss basins. Holden Crater presents the broadest and

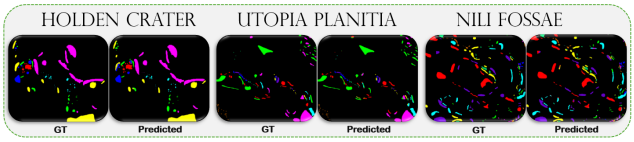

Figure 6. Mineralogical maps of Mars inferred by our model for three regions of interest.

smoothest basin, consistent with its strongest quantitative performance. Nili Fossae shows slightly increased curvature, reflecting greater spectral variability, yet retains a stable basin structure. Utopia Planitia exhibits mild asymmetry, but preserves a dominant low-loss region.

Spectral Band Sensitivity Analysis. Fig. 5 shows the class-wise average gradient magnitude across the spectral bands for all three datasets. This reflects how the model distributes discriminative importance over wavelengths. Holden Crater exhibits sharp, localized peaks, indicating reliance on distinct absorption bands due to strong spectral separability. Utopia Planitia presents smoother and more distributed responses, suggesting subtler inter-class differences requiring broader spectral integration. Nili Fossae shows higher and overlapping gradients across multiple bands, consistent with greater mineralogical diversity and spectral ambiguity. As gradient-based sensitivity is model-dependent, these patterns show learned spectral attention and the results indicate that the network adaptively reallocates spectral focus according to geological complexity, concentrating on distinct bands in simpler settings and distributing emphasis in more challenging environments.

Classification Maps. The qualitative results in Fig. 6 depict that *LUMEN* produces accurate mineral maps across Martian terrains, with most classes clearly identified and spatially well delineated. Minor errors occur in small or spatially sparse regions, likely due to limited training samples and spectral similarity with dominant mineral classes.

4. Conclusion

We presented *LUMEN*, a language-guided memory-augmented framework for few-shot Martian hyperspectral classification. By aligning spectral embeddings with LLM-derived semantic prototypes through momentum-stabilized memory and eigenspace-level relational distillation, the proposed method effectively induces discriminative structure under extreme label scarcity. Experiments across multiple CRISM datasets demonstrate consistent improvements over strong baselines with reduced model complexity. Nevertheless, LUMEN depends on the quality of language-generated priors and currently assumes closed-set mineral categories. Future work will explore physics-aware semantic grounding, open-set recognition, and large-scale self-supervised hyperspectral pretraining to further improve robustness and generalization in planetary spectral analysis.

References

[1] Jean-Pierre Bibring, Yves Langevin, John F. Mustard, François Poulet, Raymond Arvidson, Aline Gendrin, and Gerhard Neukum. Global mineralogical and aqueous mars history derived from OMEGA/Mars express data. *Science*, 312(5772):400–404, 2006. 1

[2] José M. Bioucas-Dias, Antonio Plaza, Gustavo Camps-Valls, Paul Scheunders, Nasser Nasrabadi, and Jocelyn Chanussot. Hyperspectral remote sensing data analysis and future challenges. *IEEE Geoscience and Remote Sensing Magazine*, 1 (2):6–36, 2013. 1

[3] Abhiroop Chatterjee, Susmita Ghosh, and Ashish Ghosh. Context-aware masking and learnable diffusion-guided patch refinement in transformers via sparse supervision for hyperspectral image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2906–2915, 2025. 2

[4] Liang Chen, Ghazi Shazan Ahmad, Tianjun Yao, Lingqiao Liu, and Zhiqiang Shen. One last attention for your vision-language model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025. 2, 5, 6, 8

[5] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2021. 2

[6] Glenn Healey and Dan Slater. Models and methods for automated material identification in hyperspectral imagery acquired under unknown illumination and atmospheric conditions. *IEEE Transactions on Geoscience and Remote Sensing*, 37(6):2706–2717, 1999. 1

[7] Danfeng Hong et al. SpectralFormer: Rethinking hyperspectral image classification with transformers. *IEEE Transactions on Geoscience and Remote Sensing*, 60, 2022. 2, 6

[8] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7132–7141, 2018. 2

[9] O. M. Kamps, R. D. Hewson, F. J. A. van Ruitenbeek, and F. D. Van Der Meer. Defining surface types of mars using global crism summary product maps. *Journal of Geophysical Research: Planets*, 125(8):e2019JE006337, 2020. 5

[10] Muhammad Javed Khan, Hammad Shahid Khan, Adeel Yousaf, Kashif Khurshid, and Ahsan Abbas. Modern trends in hyperspectral image analysis: A review. *IEEE Access*, 6: 14118–14129, 2018. 1

[11] Yushi Li, Haokui Zhang, and Qiang Shen. Spectral–spatial classification of hyperspectral imagery with 3D convolutional neural network. *Remote Sensing*, 9(1):67, 2017. 2, 6

[12] Shanshan Mei, Chao Song, Min Ma, and Feng Xu. Hyperspectral image classification using group-aware hierarchical transformer. *IEEE Transactions on Geoscience and Remote Sensing*, 60, 2022. 2, 6

[13] Scott Murchie, Raymond Arvidson, Philippe Bedini, Kirk Beisser, Jean-Pierre Bibring, Janice Bishop, and Michael Wolff. Compact reconnaissance imaging spectrometer for mars (CRISM) on mars reconnaissance orbiter (MRO). *Journal of Geophysical Research: Planets*, 112(E5), 2007. 1, 5

[14] Antonio Plaza, Jon Atli Benediktsson, Joseph W. Boardman, Justin Brazile, Lorenzo Bruzzone, Gustau Camps-Valls, and Giuseppe Trianni. Recent advances in techniques for hyperspectral image processing. *Remote Sensing of Environment*, 113:S110–S122, 2009. 1

[15] Zhen Qiu, Jie Xu, Jun Peng, and Wei Sun. Cross-channel dynamic spatial–spectral fusion transformer for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 61, 2023. 2

[16] Alec Radford, Jong Wook Kim, Chris Hallacy, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, 2021. 2

[17] Sankar K. Roy, Subhankar Manna, Tiejun Song, and Lorenzo Bruzzone. Attention-based adaptive spectral–spatial kernel ResNet for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 59(9):7831–7843, 2021. 2, 6

[18] Sankar K. Roy, Abhijit Deria, Chintan Shah, Juan M. Haut, Qian Du, and Antonio Plaza. Spectral–spatial morphological attention transformer for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 61: 5503615, 2023. 2, 6

[19] Lei Sun, Guang Zhao, Yuhui Zheng, and Zhiwei Wu. Spectral–spatial feature tokenization transformer for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 60, 2022. 2, 6

[20] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. 2

[21] Qilong Wang, Banggu Wu, Pengfei Zhu, Peihua Li, Wangmeng Zuo, and Qinghua Hu. ECA-Net: Efficient channel attention for deep convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11534–11542, 2020. 2

[22] Bobo Xi, Yun Zhang, Jiaojiao Li, Tie Zheng, Xunfeng Zhao, Haitao Xu, Changbin Xue, Yunsong Li, and Jocelyn Chanussot. MCTGCL: Mixed CNN–Transformer for mars hyperspectral image classification with graph contrastive learning. *IEEE Transactions on Geoscience and Remote Sensing*, 63: 1–14, 2025. 2, 6

[23] Bobo Xi, Yun Zhang, Jiaojiao Li, et al. HyMars: Mars hyperspectral image classification benchmark datasets, 2025. 5

[24] Zhaohui Xue, Qiang Xu, and Meng Zhang. Local transformer with spatial partition restore for hyperspectral image classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 15:4307–4325, 2022. 2

[25] Hantao Yao, Rui Zhang, and Changsheng Xu. Textual-based class-aware prompt tuning for vision-language models. In

Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 23438–23448, 2024. 2, 5, 6, 8

[26] Shuang Yu, Sen Jia, and Chunyan Xu. Convolutional neural networks for hyperspectral image classification. *Neurocomputing*, 219:88–98, 2017. 2

[27] Zilong Zhong, Jun Li, Zhiwei Luo, and Mark Chapman. Spectral–spatial residual network for hyperspectral image classification: A 3-D deep learning framework. *IEEE Transactions on Geoscience and Remote Sensing*, 56(2):847–858, 2018. 2, 6

[28] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13472–13481, 2022. 2, 8