



Article

Multimodal Prompt Tuning for Hyperspectral and LiDAR Classification

Zhengyu Liu ¹, Xia Yuan ¹, Shuting Yang ², Guanyiman Fu ¹, Chunxia Zhao ¹ and Fengchao Xiong ^{1,*}

¹ School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China; zhengyuliu@njust.edu.cn (Z.L.); yuanxia@njust.edu.cn (X.Y.); gym.fu@njust.edu.cn (G.F.); zhaochx@mail.njust.edu.cn (C.Z.)

² Institute of Agricultural Economics and Information Technology, Ningxia Academy of Agriculture and Forestry Sciences, Yinchuan 750002, China; nxnkyyst@163.com

* Correspondence: fcxiong@njust.edu.cn

Abstract

The joint classification of hyperspectral imaging (HSI) and Light Detection and Ranging (LiDAR) data holds significant importance for various practical uses, including urban mapping, mineral prospecting, and ecological observation. Achieving robust and transferable feature representations is essential to fully leverage the complementary properties of HSI and LiDAR modalities. However, existing methods are often constrained to scene-specific training and lack generalizability across datasets, limiting their discriminative power. To tackle this challenge, we introduce a new dual-phase approach for the combined classification of HSI and LiDAR data. Initially, a transformer-driven network is trained on various HSI-only datasets to extract universal spatial-spectral features. In the second stage, LiDAR data is incorporated as a task-specific prompt to adapt the model to HSI-LiDAR scenes and enable effective multimodal fusion. Through extensive testing on three benchmark datasets, our framework proves highly effective, outperforming all competing approaches.

Keywords: hyperspectral image; light detection and ranging (LiDAR); multimodal fusion; prompt tuning



Academic Editors: Xia Xu, Sen Lei, Yuanchao Su and Xuanwen Tao

Received: 23 June 2025

Revised: 28 July 2025

Accepted: 5 August 2025

Published: 14 August 2025

Citation: Liu, Z.; Yuan, X.; Yang, S.; Fu, G.; Zhao, C.; Xiong, F. Multimodal Prompt Tuning for Hyperspectral and LiDAR Classification. *Remote Sens.* **2025**, *17*, 2826. <https://doi.org/10.3390/rs17162826>

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Land cover classification is a fundamental task in Earth observation, essential for urban surveying, resource exploration, and environmental monitoring. Hyperspectral images (HSIs), which capture continuous spectra across the visible, near-infrared, and mid-infrared regions, significantly enhance land-cover classification [1]. However, HSIs face limitations in certain applications due to their relatively low spatial resolution and sensitivity to external environmental factors—such as atmospheric conditions and varying illumination—which can hinder classification accuracy [2]. To overcome these challenges, the integration of multisource remote sensing data has emerged as a promising approach, leveraging intermodal complementarity to improve classification performance.

Light Detection and Ranging (LiDAR) generates point clouds that represent elevation information. Unlike HSIs, LiDAR is an active sensing technology that is unaffected by ambient lighting or atmospheric conditions and can penetrate vegetation canopies to capture the underlying terrain. These advantages make LiDAR a natural complement to hyperspectral data. Consequently, recent years have witnessed significant progress in joint HSI and LiDAR classification [3,4], encompassing both traditional methods [5,6] and deep learning-based approaches [7–9].

Effective feature extraction is a critical step in enhancing multimodal HSI-LiDAR classification. Although substantial advancements have been achieved using CNNs [10,11], GNNs [8], and transformer-based methods [12], most existing models are designed for specific scenes and often require total retraining when applied to new ones. As a result, they do not take advantage of knowledge between datasets, leading to insufficiently discriminatory features. This limitation is largely due to the scarcity of co-registered HSI and LiDAR datasets.

In contrast, standalone HSI datasets are significantly more abundant, owing to the long-standing advancement of hyperspectral imaging technology. This imbalance raises a critical question: can standalone HSI datasets be leveraged to enhance joint HSI-LiDAR classification? Intuitively, the answer is affirmative—robust representations learned from diverse HSI datasets can provide a strong foundation for learning joint HSI-LiDAR representations, thereby improving classification performance in multimodal settings.

To this end, we propose a two-stage network framework for joint HSI-LiDAR classification. In the first stage, a transformer-based model is pre-trained collaboratively on multiple HSI-only datasets to learn shared and transferable spatial-spectral representations. In the second stage, LiDAR data is incorporated as a prompt to adapt the learned transformer network for HSI-LiDAR classification. This strategy facilitates knowledge transfer from auxiliary HSI data and enhances the model's representation capacity. Thorough evaluations on standard benchmark datasets confirm that our approach outperforms existing methods.

The remainder of this article is organized systematically: existing research is summarized in Section 2, the technical approach is developed in Section 3, comprehensive experiments are analyzed in Section 4, and final remarks are given in Section 5.

2. Related Work

This review concentrates on three essential components: HSI representation learning, multimodal fusion of HSI and LiDAR data, and recent advances in prompt tuning in computer vision.

2.1. HSI Representation

Early approaches in hyperspectral image (HSI) classification relied on raw spectral profiles as features, combined with traditional classifiers such as k-nearest neighbors (k-NN), Bayesian estimation, and support vector machines (SVMs) [13–15]. However, these methods often struggled with the high dimensionality of HSI data, leading to suboptimal performance. To mitigate these issues, dimensionality reduction techniques like principal component analysis (PCA) [16], linear discriminant analysis (LDA) [17], and non-negative matrix factorization [18] have been introduced for HSI representation. These methods help compactly represent HSIs while preserving discriminative information. Additionally, to exploit the spatial and spectral features of HSIs, Plaza et al. [19] proposed the extended morphological profile (EMP), which leverages mathematical morphology to automatically derive endmembers. When combined with PCA, EMP-based methods effectively reduced redundancy while capturing spatial-spectral correlations [20]. Further advancements include 3D morphological profiles (3DMPs) [21], 3D Wavelet [22], 3D Gabor filters [23], and 3D Local Binary Patterns [24]. Despite these innovations, traditional machine learning approaches remain limited in extracting high-level semantic features, ultimately constraining classification accuracy.

Recent advancements in deep learning have significantly enhanced HSI representation. For example, Chen et al. [25] demonstrated that stacked convolutional layers can effectively extract hierarchical features for HSI classification and target detection. To better capture the spatial-spectral structure of HSIs, 3D convolutional operations were introduced

in [26]. However, the high computational cost of 3D convolutions limits their scalability in deeper architectures. To address this issue, more efficient designs such as 3D Asymmetric Networks [27] and Pseudo-3D Networks [28] have been proposed, which decompose 3D convolutions to improve both efficiency and performance. While CNNs are effective in extracting deep spatial–spectral features, recurrent neural networks (RNNs) have also been employed to model spectral band dependencies by treating hyperspectral data as sequential inputs [29], thus capturing long-range spectral correlations. In addition, to mitigate the challenges of computational complexity and limited annotated data, deep unfolding-based methods [30,31] have been developed. These approaches unroll traditional iterative optimization algorithms into interpretable and trainable neural network layers, integrating domain-specific priors with the learning capabilities of deep models.

Graph Convolutional Networks (GCNs) [32] have emerged as a powerful alternative for HSI representation and analysis through graph-based structures. With the widespread success of transformers in natural language processing and computer vision [33,34], attention mechanisms for non-local modeling have also garnered significant interest in hyperspectral analysis. For instance, HSI-BERT [35] leverages transformers to provide a global receptive field and strong generalization. SpectralFormer [36] utilizes spectrally local sequence modeling with cross-layer skip connections to preserve essential features, surpassing traditional CNN-based methods. He et al. [37] adopt transformers for spectral sequence modeling, using multilayer perceptrons for classification.

The limitations of current HSI classification approaches can be summarized as follows. Traditional machine learning methods struggle with high-dimensional data and rely heavily on hand-crafted features, which limits their ability to capture high-level semantic representations. Deep learning models, while more powerful, depend on large labeled datasets and lack strong generative capabilities.

2.2. Fusion of HSI and LiDAR Data

Current methodologies for combining HSI and LiDAR information mainly operate at three distinct levels: pixel-level, feature-level, and decision-level fusion. Pixel-level fusion [15,38,39], the most fundamental form of image fusion, retains the original information but is highly susceptible to noise and typically demands more computational resources. Therefore, most research resorts to feature-level fusion, which involves extracting spatial features based on object shapes and neighborhood information, along with spectral and topographical features. For instance, Xu et al. [40] proposed a dual-branch CNN framework that leverages separate spatial and spectral networks to collaboratively classify HSI and LiDAR data. Su et al. [41] introduced a meta-heuristic optimization strategy to refine the graph structure for more effective feature representation. Wu et al. [42] developed a cross-modal reconstruction strategy to learn more compact and informative fusion representations. Hang et al. [10] presented a framework that integrates hyperspectral and LiDAR data using two interconnected CNNs: one designed to extract spectral–spatial features from HSI and the other to capture elevation information from LiDAR. Similarly, EndNet [43], an encoder–decoder fusion network, mitigates the limitations of single-modality remote sensing data through a reconstruction strategy that stimulates neuron activation across both data types. The FusAtNet model [44] incorporates self-attention and cross-attention mechanisms to facilitate effective interaction between modalities, emphasizing the harmonious fusion of HSI and LiDAR information.

Decision-level fusion, the highest level of data integration, combines classification outputs from multiple classifiers. The MFSuDF approach [6] applied KPCA to HSI data guided by superpixels, extracted multiple features, and fused them at the superpixel level, with final predictions obtained via decision fusion. A recent work [45] proposed

a multi-probability decision fusion strategy, where four types of features were fed into separate KELM classifiers, and their probability matrices were combined to achieve optimal classification accuracy.

The limitations of current HSI-LiDAR fusion can be summarized as follows: Pixel-level fusion, while theoretically capable of retaining the richest raw information, is highly sensitive to noise and computationally expensive, which constrains its practicality for large-scale applications. In contrast, decision-level fusion occurs at a much later stage, relying heavily on the outputs of individual classifiers. This reliance can propagate errors from weaker modalities and, because it integrates information only after independent decisions have been made, it often overlooks fine-grained cross-modal correlations embedded in the raw data or intermediate features. Positioned between these two extremes, feature-level fusion seeks to capture richer interactions by aligning heterogeneous features before the decision stage.

2.3. Prompt Tuning

Foundation models have shown impressive performance in computer vision and remote sensing tasks [46,47]. Traditionally, adapting these models to downstream tasks requires full fine-tuning, but prompt tuning has emerged as an efficient alternative that freezes backbone parameters and introduces minimal trainable prompts in the input space [48,49]. Initially developed for language models [50], prompt tuning has been extended to vision tasks, with Visual Prompt Tuning (VPT) [49] achieving performance comparable to full fine-tuning by embedding learnable tokens into transformer inputs. Recent advancements showcase diverse prompt engineering strategies in HSI representation: SAGFFNet [45] incorporates spectral prompts with a frozen SAM encoder; Tan et al. [51] introduce low-rank prompts to optimize transformer structures; and Kong et al. [52] fuse LiDAR-driven spatial prompts with spectral cues. Moreover, prompt tuning has demonstrated unique advantages in multimodal fusion. By training only a small number of prompt parameters, it avoids the overfitting risks associated with full-parameter fine-tuning. Zhou et al. [53] bridged modality gaps using learnable vision–language prompts, achieving higher accuracy than feature concatenation in few-shot scenarios. Similarly, Li et al. [31] introduced modality-specific prompts while keeping the CLIP backbone frozen, matching the performance of direct fusion with only 1/50 of the parameters. The development of hyperspectral foundation models [54–56] highlights the power of large-scale pre-training. However, integrating LiDAR with HSI-specific pre-trained backbones remains underexplored.

3. Methods

This section details our method, including the overall framework and two-stage training.

3.1. Overall Framework

The two-stage structure of our framework is depicted in Figure 1. In Stage I, a spatial–spectral transformer is pre-trained in a supervised manner using multiple HSI-only datasets to learn robust and transferable spatial–spectral representations. In Stage II, the pre-trained transformer is adapted to specific HSI-LiDAR datasets by introducing LiDAR data as a prompt, which is fused with hyperspectral features to enhance multimodal representation. This two-stage strategy—combining cross-dataset pre-training with prompt-based adaptation—significantly improves the model’s capacity for accurate and robust HSI-LiDAR classification.

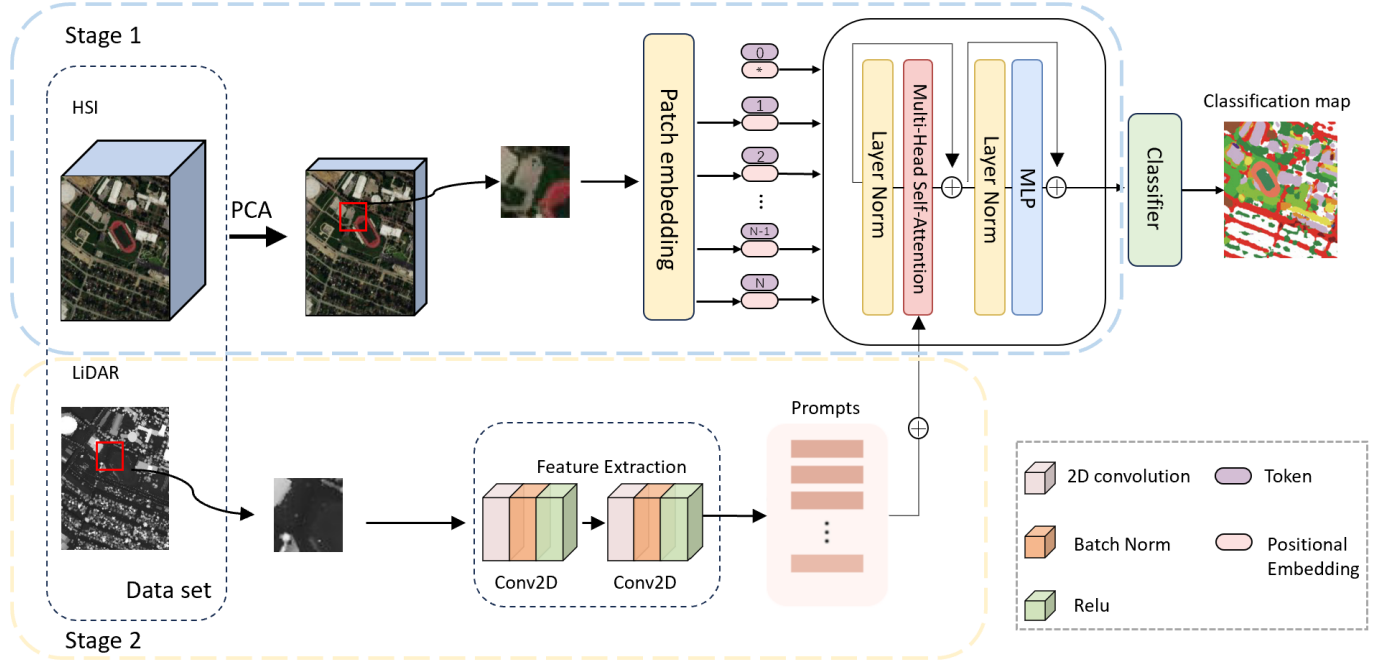


Figure 1. Methodological framework overview. Our framework consists of two stages. In the first stage, a spatial–spectral transformer is trained to extract robust features across multiple hyperspectral datasets. In the second stage, LiDAR data is introduced as a prompt to guide the network in learning a joint HSI–LiDAR representation.

3.2. First Stage: Learning the Spatial–Spectral Representation Across Multiple Datasets

Hyperspectral datasets are often captured using different sensors, resulting in varying numbers of spectral bands across datasets. To facilitate unified representation learning via a shared network, it is crucial to standardize the token length for subsequent processing. Our solution implements grouped PCA along the spectral axis, capitalizing on neighboring band correlations to transform the data into a coherent feature representation. The HSI data $I \in \mathbb{R}^{H \times W \times B}$ (with H , W , and B indicating height, width, and band quantity) is processed by dividing the spectral dimension into T uniform subgroups. PCA is then applied to each group to extract features of length L . Finally, the features from all groups are concatenated, resulting in the output $I_p \in \mathbb{R}^{H \times W \times T \cdot L}$. The process of groupwise PCA can be mathematically expressed as followed:

$$\begin{aligned} I_1, I_2, \dots, I_T &= \text{Split}(I), \\ P_i &= \text{PCA}(I_i), \forall i \in [1, T], \\ I_p &= \text{Concat}(P_1, P_2, \dots, P_T). \end{aligned} \quad (1)$$

In this study, T was set to 4 and L was set to 8, resulting in final features with a dimensionality of 32. For simplicity, we continue to refer to the resulting feature dimension after PCA as “bands”, although it no longer strictly represents spectral bands.

To achieve accurate material classification, capturing fine-grained variations or absorption patterns in spectral signatures is crucial. Inspired by the SpectralFormer [36], we propose a novel method that learns groupwise spectral embeddings instead of traditional bandwise representations in hyperspectral imagery. Given a spectral sequence $x = [x_1, x_2, \dots, x_c] \in \mathbb{R}^{1 \times c}$, where c is the number of spectral channels, we extract overlapping neighboring bands to embed local spectral profiles. The groupwise spectral embedding is defined as follows:

$$\mathbf{X} = g(\mathbf{x}) = [\mathbf{x}_1, \dots, \mathbf{x}_q, \dots, \mathbf{x}_c] \in \mathbb{R}^{j \times c} \quad (2)$$

Here, $x_q = [x_{q-\lfloor(i/2)\rfloor}, \dots, x_q, \dots, x_{q+\lfloor(i/2)\rfloor}]^t \in R^{i \times 1}$, $\lfloor \cdot \rfloor$ represents the rounding function, where i and j correspond to the count of adjacent pixels and total pixels in the HS image, respectively. Through experimental analysis, the parameter i was determined to be optimally set at 3, where $g(\cdot)$ represents the overlapping grouping function associated with $\mathbf{x}$.

To further capture the local spatial–spectral properties of HSIs, we extract local patches and formulate patch cubes of size $\mathbf{X}^+ \in R^{c \times h \times w}$, where h and w are the height and length of the patch, respectively. These patches are unfolded along the spatial dimension, resulting in the following spatial–spectral embedding:

$$\hat{\mathbf{X}} = [\vec{\mathbf{x}}_1, \dots, \vec{\mathbf{x}}_i, \dots, \vec{\mathbf{x}}_c] \quad (3)$$

where $\vec{\mathbf{x}}_i \in R^{d \times 1}$ ($d = h \times w$) denotes the unfolded patch vector for the i th band. Before feeding the feature embeddings into the transformer encoder, we prepend a learnable class embedding to the sequence, formulated as follows:

$$\mathbf{F}_1 = \text{concat}(\mathbf{C}, \hat{\mathbf{X}}) = [\mathbf{x}_{\text{hsel\_cls}}, \vec{\mathbf{x}}_1, \dots, \vec{\mathbf{x}}_i, \dots, \vec{\mathbf{x}}_c]. \quad (4)$$

Here, $\mathbf{x}_{\text{hsel\_cls}}$ corresponds to the adaptable classification token. For preserving the sequential structure of spectral tokens and encoding their positions, we integrate positional embeddings $\mathbf{P}_1$, which are summed with the spectral embeddings $\mathbf{F}_1$, resulting in

$$\mathbf{F}_2 = \mathbf{F}_1 + \mathbf{P}_1 \quad (5)$$

Specifically, the positional embeddings are defined as $\mathbf{P}_1 = [\mathbf{p}_1, \dots, \mathbf{p}_{c+1}]$, where each $\mathbf{p}_i$ encodes the positional information for the corresponding token. The resulting feature representation $\mathbf{F}_2$ is then passed through a transformer encoder, which comprises alternating multi-head attention (MHA) and multilayer perceptron (MLP) layers, each followed by residual connections and layer normalization. The computation at the l -th encoder block is given by

$$\begin{aligned} \mathbf{F}'_l &= \text{MHA}(\text{LN}(\mathbf{F}_{l-1})) + \mathbf{F}_{l-1} \\ \mathbf{F}_l &= \text{MLP}(\text{LN}(\mathbf{F}'_l)) + \mathbf{F}'_l \end{aligned} \quad (6)$$

Here, $\mathbf{F}_l$ denotes the output spectral feature embedding of the l -th transformer encoder block, and $l \in \{1, 2, \dots, k\}$, where k is the total number of layers. The residual connections help preserve gradient flow, while the LayerNorm LN ensures stable training.

A linear classifier is appended to the output of the transformer encoder to generate the classification map. To account for variations in class distributions across different datasets, multiple classifiers are employed. By learning from multiple datasets, the transformer encoder gains enhanced representational capacity, enabling it to more effectively model the characteristics of hyperspectral images.

3.3. Second Stage: LiDAR as Prompt for Tuning

As shown in Figure 2, we propose LiDAR Prompt Tuning (LPT) to adapt the transformer model, trained in the first stage, for joint HSI–LiDAR classification. Given a LiDAR data patch $\mathbf{X}_{\text{LiDAR}} \in R^{1 \times h \times w}$, we first extract its features using a lightweight CNN comprising two convolutional layers. This process generates a compact feature map of size $64 \times 1 \times 1$, and can be mathematically formulated as follows:

$$\mathbf{F}_{\text{LiDAR}} = \text{GAP}(\text{Conv2D}_2(\text{MaxPool}(\text{Conv2D}_1(\mathbf{X}_{\text{LiDAR}})))) \quad (7)$$

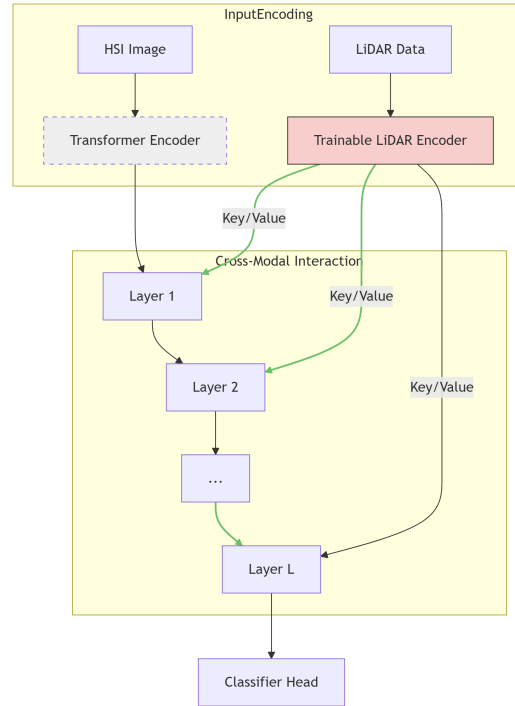


Figure 2. LiDAR Prompt Tuning.

Here, Conv2D_1 and Conv2D_2 contain 32 and 64 filters, respectively. The MaxPool operation reduces the spatial resolution by half, and GAP denotes global average pooling along the spatial dimensions. The resulting feature vector $\mathbf{F}_{\text{LiDAR}}$ is treated as a prompt token and is fused with hyperspectral features at each transformer layer via cross-attention:

$$\text{CrossAttn}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{D}}\right)\mathbf{V} \quad (8)$$

where the query is computed from the hyperspectral feature tokens as $\mathbf{Q} = \mathbf{F}_l \mathbf{W}_Q$, and the key/value projections are derived from the LiDAR prompt: $\mathbf{K} = \mathbf{F}_{\text{LiDAR}} \mathbf{W}_K$, $\mathbf{V} = \mathbf{F}_{\text{LiDAR}} \mathbf{W}_V$.

The output of the cross-attention module is added back to the input hyperspectral features via a residual connection:

$$\mathbf{F}_{l+1} = \mathbf{F}_l + \text{CrossAttn}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) \quad (9)$$

This mechanism enables the transformer to seamlessly incorporate structural and elevation cues from LiDAR data to achieve robust representation for joint classification.

4. Experiments

In this section, we evaluate our framework on three HSI-LiDAR datasets to demonstrate its advantages.

4.1. Dataset Description

Dataset in the first stage: We trained the HSI representation transformer on six datasets collected from various hyperspectral sensors: Indian Pines, Salinas Valley, Pavia University, Kennedy Space Center (KSC), Botswana, and Washington DC Mall. The patch size was set to 9×9 for training.

The detailed information of these datasets is given as follows:

1. Indian Pines: This dataset was acquired by the AVIRIS sensor over an agricultural area, with an image size of 145×145 pixels and 224 spectral bands. The dataset includes 16 classes and contains 21,025 labeled pixels.
2. Salinas Valley: This dataset was collected by the AVIRIS sensor and originally consisted of 224 spectral bands. After removing 20 water absorption bands, 204 bands were retained. The imagery size is 512×217 pixels. It contains 16 classes and 54,129 labeled pixels.
3. Pavia University: This dataset was captured by the ROSIS-3 sensor and includes 115 spectral bands (reduced to 103 bands after discarding 12 noise bands) covering the range of 430–860 nm. The imagery size is 610×340 pixels, with a spatial resolution of 1.3 m. It comprises 9 land-cover classes and 42,776 labeled pixels.
4. KSC: Acquired by the NASA AVIRIS instrument over Florida on 23 March 1996, this dataset originally consisted of 224 bands. After removing water absorption and low-SNR bands, 176 bands were retained. The imagery size is 512×614 pixels and it comprises 13 land-cover classes.
5. Botswana: This dataset was acquired by the Hyperion sensor on NASA's EO-1 satellite and consists of 242 spectral bands. The spatial resolution is 30 m, and the dataset captures a diverse range of environmental features, with 1476×256 pixels divided into 14 classes.
6. Washington DC Mall: Released by the Spectral Information Technology Application Center of Virginia in 2013, this dataset includes 191 spectral bands. It has an image size of 1280×307 pixels and includes 7 land-cover classes, representing urban and natural environments.

Dataset for Evaluation: Three HSI-LiDAR classification datasets were utilized to evaluate the performance of the proposed framework: Houston 2013, MUUFL Gulfport, and Trento. The details of each dataset are as follows:

1. Houston 2013: Provided by the Hyperspectral Image Analysis Group and the NSF-funded Airborne Laser Mapping Center (NCALM). Originally created for the 2013 GRSS Data Fusion Competition, this dataset was provided through collaboration between the Hyperspectral Image Analysis Group and NSF's NCALM. It features a 144-band hyperspectral image with 349×1905 pixel dimensions, representing 15 land-use classes. Table 1 enumerates the specific class labels and their respective training/test sample sizes in the Houston scene.
2. MUUFL Gulfport: The ROSIS imaging spectrometer captured this dataset, which contains 72-band spectral data with 325×220 pixel spatial coverage. Table 2 enumerates the land-use categories along with their respective training and validation sample sizes in the MUUFL study area.
3. Trento: The hyperspectral imagery was acquired by the AISA Eagle system, comprising 63 spectral channels. Synchronized LiDAR measurements collected with an Optech ALTM 3100EA sensor share identical spatial coverage (166×600 pixels at 1 m resolution). Table 3 provides the complete classification schema with training/testing sample allocations for the Trento study area.

Table 1. Sample size distribution for the Houston dataset.

Class	Land-Cover Type	Training	Testing	Total
1	Health grass	198	1053	1251
2	Stressed grass	190	1064	1254
3	Synthetic grass	192	505	697
4	Trees	188	1056	1244
5	Soil	186	1056	1242
6	Water	182	143	325
7	Residual	196	1072	1268
8	Commercial	191	1053	1244
9	Road	193	1059	1252
10	Highway	191	1036	1227
11	Railway	181	1054	1235
12	Parking lot 1	192	1041	1233
13	Parking lot 2	184	285	469
14	Tennis court	181	247	428
15	Running track	187	473	660

Table 2. Sample size distribution for the MUUFL dataset.

Class	Land-Cover Type	Training	Testing	Total
1	Trees	150	23,096	23,246
2	Mostly Grass	150	4120	4270
3	Mixed Ground Surface	150	6732	6882
4	Dirt and Sand	150	1676	1826
5	Road	150	6537	6687
6	Water	150	316	466
7	Building Shadow	150	2083	2233
8	Building	150	6090	6240
9	Sidewalk	150	1235	1385
10	Yellow Curb	150	33	183
11	Cloth Panels	150	119	269

Table 3. Sample size distribution for the Trento dataset.

Class	Land-Cover Type	Training	Testing	Total
1	Apple Trees	129	3905	4034
2	Buildings	125	2778	2903
3	Ground	105	374	479
4	Wood	154	8969	9123
5	Vineyard	184	10,317	10,501
6	Roads	122	3052	3174

4.2. Experimental Setup

4.2.1. Implementation Details

Our method was developed on PyTorch 1.13.1 and evaluated on a workstation equipped with an Intel Core i5-12400F processor, 64 GB of system memory, and an NVIDIA RTX 3080 graphics card (12 GB VRAM). The proposed architecture employed 12 transformer blocks in its initial phase. Hyperspectral images (HSIs) were first reduced to 32 dimensions via PCA and then reshaped into 32-length feature vectors and partitioned into $3 \times 3 \times 32$ patches. Model optimization was performed using Adam optimizer, with the first-stage pre-training configured as follows: batch size = 32, initial learning rate = 0.005, embedding dimension = 512, and 50 training epochs. The second stage, dedicated to LiDAR prompt feature extraction, utilized 100 epochs.

4.2.2. Compared Methods

Our experimental framework incorporates a comprehensive suite of comparative methods spanning three methodological paradigms: (1) established machine learning classifiers comprising support vector machines (SVMs), Random Forest (RF), Multinomial Logistic Regression (MLR), and K-nearest neighbors (KNN); (2) contemporary convolutional neural network architectures including S²ENet and FusAtNet; and (3) advanced vision transformer implementations featuring ViT, MTNet, and SpectralFormer. The subsequent section provides complete implementation specifications for each evaluated approach.

- **SVM:** The classification process was executed using the support vector machine (SVM) implementation from the sklearn library, which utilizes a radial basis function kernel configuration. Key model parameters consisted of the regularization coefficient ($C = 100$) and convergence tolerance threshold (1×10^{-7}), with both values being determined through empirical optimization studies.
- **RF:** The Random Forest algorithm was implemented through the sklearn library, with four key hyperparameters configured as follows: the ensemble comprised 200 decision trees with a maximum depth of 10 for each tree, while node splitting required a minimum of three samples per leaf and considered up to 10 features at each split. All parameter values were carefully selected based on empirical validation to ensure optimal model performance.
- **MLR:** The logistic regression model was implemented using scikit-learn's linear_model module, employing an L2 regularization penalty with the limited-memory Broyden–Fletcher–Goldfarb–Shanno (L-BFGS) optimization algorithm. To ensure convergence, the maximum iteration count was set to 5000. This classifier exclusively used hyperspectral data as input.
- **KNN:** The classification accuracy is highly sensitive to the neighborhood size (K). Using five-fold cross-validation on training data, we optimized K by testing candidate values $\{5, 10, 15, 20\}$ and retaining the top-performing configuration.
- **S²ENet [57]:** This architecture includes two feature enhancement modules, the spatial attention enhancement module and the spectral enhancement module, which refine the spatial and spectral features of HS and LiDAR data, respectively. All other hyperparameters were set consistently with those in the original paper.
- **FusAtNet [44]:** This architecture processes hyperspectral (HS) data through a self-attention (SA) module to produce spectral–spatial attention maps, while strategically integrating LiDAR-derived features via a cross-attention fusion mechanism to enhance spatial representation learning. All architectural configurations and hyperparameters were kept consistent with the baseline implementation described in the reference study.
- **ViT [58]:** For ViT, the LiDAR and HS data were concatenated along the channel dimension. The model architecture consisted solely of encoder layers to facilitate joint classification of HS and LiDAR data.
- **SpectralFormer [36]:** This architecture integrates LiDAR and hyperspectral (HS) data through channel-wise concatenation while maintaining the original vision transformer (ViT) backbone structure for comparative consistency. The 64D spectral embeddings are processed through five transformer blocks, each comprising (1) four-head self-attention, (2) an 8D hidden layer MLP, and (3) GeLU activations.
- **MTNet [12]:** The fundamental concept involves employing transformer architectures to effectively extract both modality-specific characteristics and cross-modal correlations from hyperspectral and LiDAR datasets. Our implementation of MTNet was developed using the PyTorch 1.13.1 framework, with model optimization performed through the Adam adaptive learning algorithm to achieve parameter convergence.

All architectural configurations and hyperparameters were kept consistent with the baseline implementation described in the reference study.

4.2.3. Evaluation Indicators

For comprehensive performance assessment and comparative analysis with existing methods, we employed four established quantitative metrics: (1) overall accuracy (OA), measuring total classification correctness, (2) average accuracy (AA), representing mean class-wise performance, (3) kappa coefficient (κ), evaluating agreement beyond chance, and (4) per-class accuracy for detailed category-specific evaluation. All metrics follow the convention where increased numerical values correspond to enhanced classification capability.

4.3. Classification Results

4.3.1. Quantitative Comparison

The systematic comparison presented in Tables 4–6 reveals that conventional classification approaches (SVM, RF, MLR, and KNN) generate substantially less accurate results throughout all dataset evaluations, a consequence of their fundamentally restricted ability to model sophisticated data representations. The comparative assessment of classification performance, as documented in Tables 4–6, reveals that traditional machine learning approaches—specifically SVM, RF, MLR, and KNN—exhibit consistently lower accuracy metrics when evaluated across all three experimental datasets. This performance limitation stems fundamentally from their restricted capacity to learn and represent complex feature relationships in high-dimensional data spaces. CNN-based networks, including S²ENet and FusAtNet, achieve more competitive results, benefiting from their strong feature extraction abilities learned from data. Transformer-based models, such as ViT, SpectralFormer, and MTNet, further improve performance by effectively capturing global spatial-spectral representations from hyperspectral data.

Table 4. Performance comparison of various methods on Houston dataset using OA (%), AA (%), and kappa $\times 100$ (%). Top performances highlighted in bold.

Class Name	Traditional Classifiers					Deep Learning-Based Methods				
	SVM	RF	MLR	KNN	S ² ENet	FusAtNet	ViT	SpectralFormer	MTNet	Ours
Health grass	86.32	85.75	95.44	84.43	83.38	82.43	82.62	79.20	82.05	82.10
Stressed grass	97.18	98.12	96.15	96.05	94.92	98.40	97.18	96.05	97.74	98.54
Synthetic grass	99.80	97.62	99.80	99.60	97.23	87.33	100.00	93.27	96.24	100.00
Trees	98.11	96.78	92.05	98.39	99.53	97.35	96.88	96.67	97.82	100.00
Soil	98.20	95.83	96.88	96.69	100.00	99.05	97.35	99.91	99.05	100.00
Water	97.90	96.50	99.30	97.20	100.00	97.90	95.11	81.12	93.71	91.86
Residual	89.09	85.35	82.00	80.03	83.49	93.75	82.00	86.01	89.65	90.13
Commercial	54.32	47.20	59.35	57.46	92.78	94.68	57.17	75.78	88.32	95.79
Road	81.30	71.11	70.16	72.43	84.80	88.01	69.50	70.54	82.81	85.71
Highway	69.02	56.27	62.74	61.97	90.73	62.74	61.10	49.32	77.51	80.43
Railway	88.43	81.78	77.42	85.01	93.93	82.16	77.99	81.03	91.37	90.36
Parking lot 1	64.65	44.19	69.74	51.97	80.60	87.61	63.31	75.41	76.08	81.13
Parking lot 2	69.83	60.70	77.90	40.00	77.90	81.05	74.04	82.16	76.84	83.53
Tennis court	99.20	98.38	99.19	97.57	99.19	98.79	99.19	100.00	99.60	100.00
Running track	98.31	96.19	97.04	98.52	98.10	93.02	97.89	93.23	100.00	100.00
OA (%)	85.32	82.14	84.91	82.56	89.12	88.34	82.97	83.45	87.89	91.70
AA (%)	86.45	83.67	85.23	83.89	90.01	89.56	84.12	84.78	88.95	92.11
Kappa $\times 100$ (%)	82.38	78.14	81.29	79.01	86.32	85.27	80.03	81.05	84.21	91.00

Table 5. Performance comparison of various methods on Trento dataset using OA (%), AA (%), and kappa $\times 100$ (%). Top performances highlighted in bold.

Class Name	Traditional Classifiers					Deep Learning-Based Methods				
	SVM	RF	MLR	KNN	S ² ENet	FusAtNet	ViT	SpectralFormer	MTNet	Ours
Apples	92.45	85.30	76.57	92.96	96.29	99.85	87.81	91.91	92.73	95.89
Buildings	84.59	85.57	67.28	83.12	99.93	99.32	82.83	89.02	96.36	99.33
Ground	98.93	86.90	94.39	96.52	93.32	66.58	95.99	93.85	95.19	97.12
Woods	96.91	95.56	80.28	93.79	99.99	98.46	96.68	93.56	97.00	98.87
Vineyard	77.67	80.94	62.46	66.68	99.74	96.40	77.47	94.60	87.63	99.77
Roads	70.22	64.65	72.18	68.91	83.55	90.47	68.58	60.98	86.11	92.31
OA (%)	84.47	83.00	72.52	82.67	97.80	95.01	84.89	87.32	92.36	98.13
AA (%)	86.79	83.15	75.53	83.66	95.47	91.84	84.89	87.32	92.50	97.22
Kappa $\times 100$ (%)	79.84	77.42	63.21	77.12	96.93	93.12	79.86	83.21	89.65	97.80

Table 6. Performance comparison of various methods on MUUFL dataset using OA (%), AA (%), and kappa $\times 100$ (%). Top performances highlighted in bold.

Class Name	Traditional Classifiers					Deep Learning-Based Methods				
	SVM	RF	MLR	KNN	S ² ENet	FusAtNet	ViT	SpectralFormer	MTNet	Ours
Trees	82.54	82.04	80.69	81.87	87.86	93.39	80.49	85.51	83.47	92.89
Mostly grass	81.04	79.10	81.55	80.68	86.14	84.64	81.29	75.46	79.61	87.71
Mixed ground surface	74.82	68.12	72.19	65.06	80.35	77.06	68.64	73.98	81.94	82.04
Dirt and sand	84.73	82.88	84.90	75.78	94.21	92.48	86.69	86.52	90.04	95.47
Road	87.50	84.86	73.58	88.11	89.17	86.57	86.97	88.73	90.52	90.79
Water	91.77	91.46	98.42	91.46	99.68	98.42	93.67	95.25	97.15	86.14
Building shadow	88.81	87.13	85.45	84.97	93.09	89.53	83.68	88.48	83.53	91.46
Building	78.77	72.56	72.17	71.61	90.59	94.98	83.89	77.19	81.13	92.98
Sidewalk	77.65	68.74	72.15	59.92	75.06	85.26	69.07	75.63	80.65	83.36
Yellow curb	97.27	94.54	95.63	86.34	93.94	81.82	95.42	93.94	91.77	100.00
Cloth panels	96.64	97.48	99.16	97.03	99.16	98.32	100.00	98.32	97.94	98.33
OA (%)	81.55	79.43	78.90	77.33	88.76	89.12	81.89	83.45	85.67	90.89
AA (%)	85.10	82.04	81.85	80.12	89.42	89.99	84.56	85.89	86.78	90.13
Kappa $\times 100$ (%)	78.42	75.89	75.21	73.24	86.79	87.32	78.54	81.23	83.21	89.35

Our proposed method, empowered by a two-stage training strategy, demonstrates significant advantages. The first stage enables learning robust and transferable representations from multiple HSI datasets, while the second stage adapts effectively to HSI-LiDAR joint scenarios. This design allows our model to capture comprehensive global spatial-spectral features and integrate them seamlessly with LiDAR data. As a result, the model exhibits stronger discriminative capability, leading to noticeable improvements in classification performance across all evaluation benchmarks.

4.3.2. Qualitative Comparison

To complement the quantitative analysis, we performed visual assessments by generating comparative classification maps across all experimental datasets. As illustrated in Figures 3–5 (representing the Houston, Trento, and MUUFL datasets, respectively), conventional machine learning approaches, including support vector machines (SVMs), Random Forest (RF), Multinomial Logistic Regression (MLR), and K-nearest neighbors (KNN), exhibited noise in their predictions, demonstrating their inherent limitations in precise material discrimination. In contrast, CNN-based models (e.g., S²ENet and FusAtNet) generated smoother classification maps due to their strong ability to model nonlinear patterns and leverage the additional elevation information provided by the LiDAR data. Transformer-based architectures, such as ViT and MTNet, also exhibited excellent feature representation capabilities, similarly to CNNs.

The experimental results demonstrate that our method generates classification maps with enhanced spatial consistency and finer-grained details while achieving a substantial reduction in classification noise compared to competing approaches. It effectively avoided over-smoothing at object boundaries and preserved small semantic structures, resulting in more accurate and visually consistent classification outputs. This improvement can be attributed to two key factors: (1) the ability to collaboratively learn discriminative spatial-spectral representations from multiple HSI datasets and (2) the seamless incorporation of LiDAR data as prompts to adapt the model to diverse HSI-LiDAR scenes.

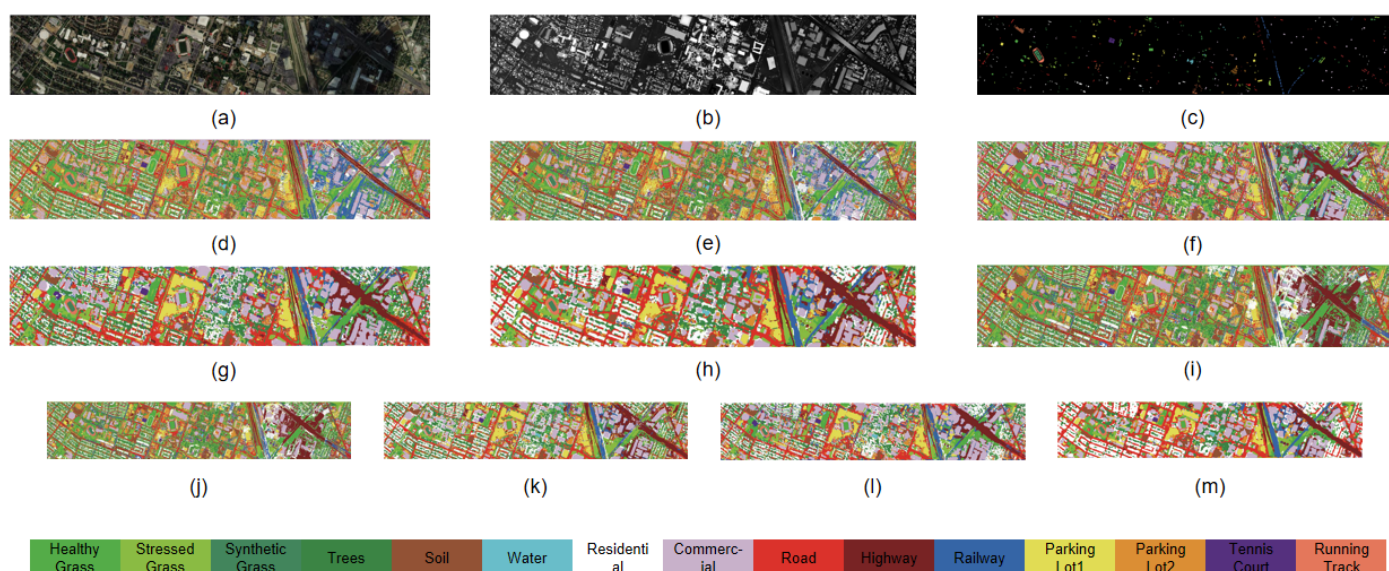


Figure 3. Comparative visualization of classification maps across models: Houston 2013 dataset. (a) Pseudo-color composite image based on bands 64, 43, and 20 for HSI. (b) Grayscale image for LiDAR-based DSM. (c) Ground-truth map. (d) SVM. (e) RF. (f) MLR. (g) KNN. (h) S²ENet. (i) FusAtNet. (j) ViT. (k) SpectralFormer. (l) MTNet. (m) Proposed method.

To comprehensively assess the computational efficiency of different architectures, we performed systematic comparisons between CNN-based and transformer-based models using two key metrics: parameter count and training time. All baseline models were implemented using their originally reported configurations to ensure fair comparisons, with quantitative results summarized in Table 7. Among the CNN-based models, except for FusAtNet, which exhibited significantly higher computational complexity due to its multiple convolutional layers and large number of convolutional filters, the other models showed minor differences in complexity. The ViT-based models, owing to their massive parameter counts and complex multi-head attention (MHA) computations, generally had higher computational complexity than the CNN-based models. Notably, on the three datasets, the training time of our model was mostly lower than that of FusAtNet and S²ENet. Overall, our model demonstrates strong competitiveness in achieving lightweight design and high classification accuracy.

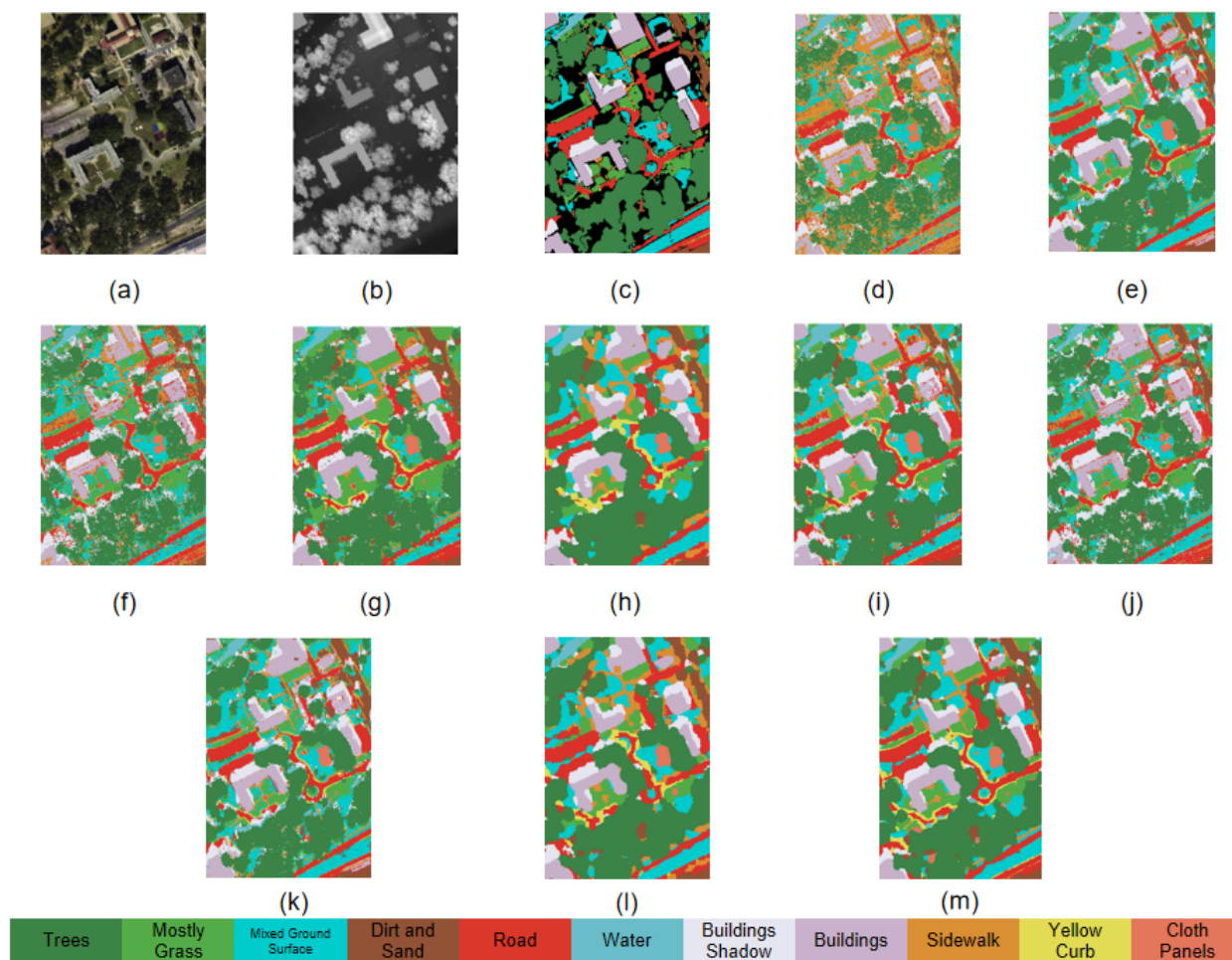


Figure 4. Comparative visualization of classification maps across models: MUUFL dataset. (a) Pseudo-color composite image based on bands 20, 15, and 5 for HSIs. (b) Grayscale image for LiDAR-based DSM. (c) Ground-truth map. (d) SVM. (e) RF. (f) MLR. (g) KNN. (h) S²ENet. (i) FusAtNet. (j) Vit. (k) SpectralFormer. (l) MTNet. (m) Proposed method.

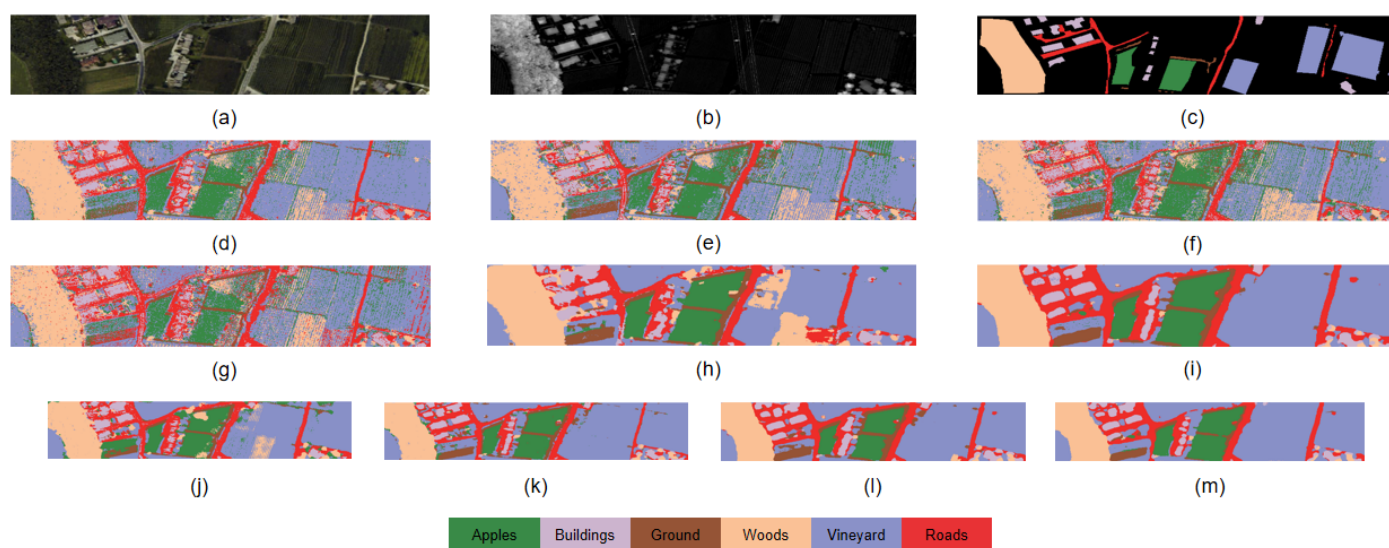


Figure 5. Comparative visualization of classification maps across models: Trendo dataset. (a) Pseudo-color composite image based on bands 20, 15, and 5 for HSIs. (b) Grayscale image for LiDAR-based DSM. (c) Ground-truth map. (d) SVM. (e) RF. (f) MLR. (g) KNN. (h) S²ENet. (i) FusAtNet. (j) Vit. (k) SpectralFormer. (l) MTNet. (m) Proposed method.

Table 7. Comparison of model complexity and efficiency.

Dataset	Complexity	S ² ENet	FusAtNet	ViT	SpectralFormer	MTNet	Ours
Houston	Parameters (M)	0.289	37.315	0.089	0.279	0.308	0.41
	Training Time (s)	261.31	1588.20	323.07	411.78	397.16	260.78
Trento	Parameters (M)	0.172	41.073	0.089	0.131	0.127	0.19
	Training Time (s)	77.31	399.74	73.15	79.48	81.74	76.61
MUUFL	Parameters (M)	0.165	38.54	0.089	0.176	0.173	0.142
	Training Time (s)	142.74	790.15	159.74	181.74	179.41	149.81

4.4. Ablation Study

4.4.1. Comparison of Different Components

To assess the contribution of each key component in our framework, we conducted an ablation study focusing on two stages: learning spatial–spectral representation from multiple HSI datasets and LiDAR-based prompt tuning. The results are summarized in Table 8.

Table 8. Ablation study on different components of our method.

Transformer	First Stage	Second Stage	Houston (OA)	Trento (OA)	MUUFL (OA)
✓			83.01	89.95	82.75
✓	✓		88.34	92.34	88.26
✓		✓	86.97	95.67	81.86
✓	✓	✓	92.21	98.03	91.39

Effect of First Stage Training: Models pre-trained on multiple HSI datasets consistently outperformed the baseline across all benchmarks, highlighting the benefit of learning generalizable spatial–spectral features from diverse scenes. This is especially effective under limited data conditions like those faced in HSI–LiDAR joint classification.

Impact of LiDAR as a Prompt: Integrating LiDAR as a modality-specific prompt further improved performance by enhancing scene understanding through complementary elevation information. However, a slight performance drop was observed on the MUUFL dataset, likely due to the already discriminative spectral content of HSI, where LiDAR may introduce redundant or noisy features.

Combining Two-Stage Training: The joint application of two-stage training yielded the best performance across all datasets, demonstrating the effectiveness of the proposed framework in exploiting both cross-domain generalization and modality-specific adaptation.

4.4.2. Feature Distribution Analysis

To further analyze the representativeness of the training data, we employed t-SNE visualization, as shown in Figure 6, which reveals the following key observations: First, the visualization confirms that our training data spans a broad range of scenarios, demonstrating strong representativeness across diverse environments. However, when relying solely on hyperspectral imagery (HSI), certain challenging classes, such as Healthy Grass, exhibit substantial feature overlap, leading to reduced class separability and a lower overall accuracy (OA) of 88.34%. By integrating LiDAR data, this limitation is substantially mitigated; the fused representation enhances feature discriminability, particularly for critical and visually similar classes, ultimately improving OA to 91.70%. These findings highlight not only the diversity of the dataset but also the clear advantage of incorporating LiDAR data to strengthen class separation and overall model performance.

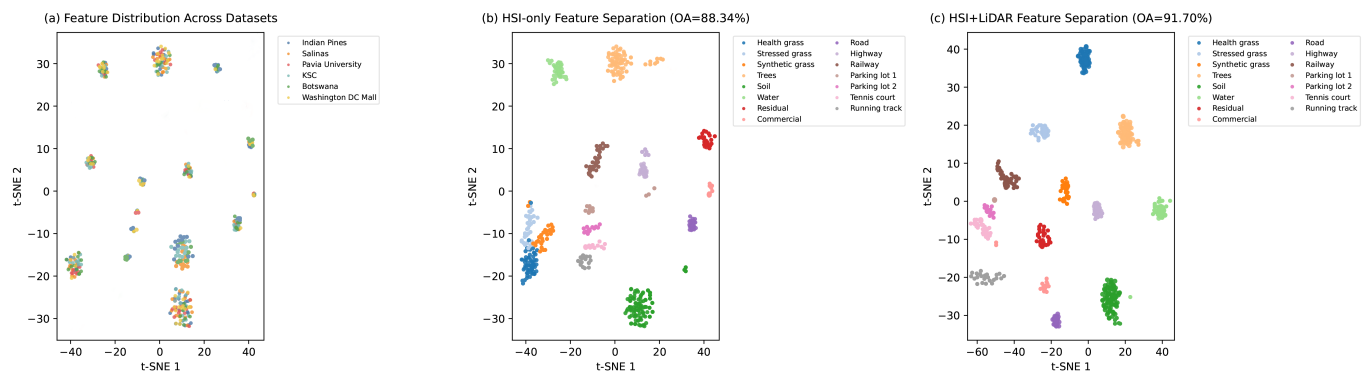


Figure 6. t-SNE visualization of (a) cross-dataset feature distribution (colors = data sources), (b) Houston test set classification with HSI only (colors = land-cover classes), and (c) multimodal feature space comparison (colors = land-cover classes). Axes represent relative similarity in t-SNE space.

4.4.3. Prompt Method Comparison

Prompt-based methods adapt pre-trained models by inserting learnable guidance signals, avoiding the need for full model fine-tuning. In the context of HSI–LiDAR fusion, two dominant paradigms can be distinguished. Input-level prompt tuning injects LiDAR-derived tokens at the model’s input stage, while cross-modal attention prompt tuning dynamically modulates intermediate representations through attention mechanisms. In the input-level approach, LiDAR data (e.g., elevation maps) are processed via a lightweight MLP to generate prompt tokens, which are concatenated with HSI patch embeddings at the frozen ViT’s input layer. By contrast, the cross-modal approach adopted in this work encodes LiDAR information into high-level features and introduces them as dynamic prompts through cross-attention layers. Here, LiDAR features serve as key/value pairs to iteratively refine HSI queries, enabling richer semantic interaction and deeper feature fusion. Empirical results shown in Table 9 confirm that the cross-attention-based prompt tuning outperforms the input-level strategy, as it allows more effective interaction between HSI and LiDAR features, leading to enhanced fusion and stronger performance.

Table 9. Comparison of different prompt methods.

Methods	Houston (OA)	Trento (OA)	MUUFLL (OA)
Input-level prompt strategy	87.34	95.91	90.17
Cross-attention-based prompt	92.21	98.03	91.39

5. Conclusions

This paper presents a novel framework for HSI–LiDAR joint classification by leveraging extra hyperspectral data for robust spatial–spectral representation learning and introducing LiDAR as a prompt for efficient domain adaptation. Extensive evaluations on three benchmark datasets demonstrate that learning from multiple HSI sources enhances feature generalization, which in turn improves LiDAR-guided classification performance. Future work will focus on reducing the reliance on labeled data in the pre-training stage to enable fully unsupervised spatial-spectral representation learning.

Author Contributions: All authors contributed to this manuscript: Conceptualization, Z.L.; Methodology, Z.L.; Resources, S.Y. and F.X.; Supervision, C.Z.; Validation, Z.L. and G.F.; Writing—Original Draft, Z.L.; Writing—Review and Editing, X.Y. and C.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by NingXia Academy of Agriculture and Forestry Sciences Science and Technology Innovation Guidance Technology Research Project, “Research and Demonstration of Key Technologies for Smart Planting of Wine Grapes in Ningxia”, under grant NKYG-23-02.

Data Availability Statement: Data are contained within the article.

Conflicts of Interest: The authors declare no conflicts of interest

References

- Landgrebe, D. Hyperspectral image data analysis. *IEEE Signal Process. Mag.* **2002**, *19*, 17–28. [\[CrossRef\]](#)
- Gu, Y.; Liu, T.; Gao, G.; Ren, G.; Ma, Y.; Chanussot, J.; Jia, X. Multimodal hyperspectral remote sensing: An overview and perspective. *Sci. China Inf. Sci.* **2021**, *64*, 121301. [\[CrossRef\]](#)
- Ghamisi, P.; Rasti, B.; Yokoya, N.; Wang, Q.; Hofle, B.; Bruzzone, L.; Bovolo, F.; Chi, M.; Anders, K.; Gloaguen, R.; et al. Multisource and multitemporal data fusion in remote sensing: A comprehensive review of the state of the art. *IEEE Geosci. Remote Sens. Mag.* **2019**, *7*, 6–39. [\[CrossRef\]](#)
- Pedergrana, M.; Marpu, P.R.; Dalla Mura, M.; Benediktsson, J.A.; Bruzzone, L. Classification of remote sensing optical and LiDAR data using extended attribute profiles. *IEEE J. Sel. Top. Signal Process.* **2012**, *6*, 856–865. [\[CrossRef\]](#)
- Liao, W.; Pižurica, A.; Bellens, R.; Gautama, S.; Philips, W. Generalized graph-based fusion of hyperspectral and LiDAR data using morphological features. *IEEE Geosci. Remote Sens. Lett.* **2014**, *12*, 552–556. [\[CrossRef\]](#)
- Jia, S.; Zhan, Z.; Zhang, M.; Xu, M.; Huang, Q.; Zhou, J.; Jia, X. Multiple feature-based superpixel-level decision fusion for hyperspectral and LiDAR data classification. *IEEE Trans. Geosci. Remote Sens.* **2020**, *59*, 1437–1452. [\[CrossRef\]](#)
- Roy, S.K.; Deria, A.; Shah, C.; Haut, J.M.; Du, Q.; Plaza, A. Spectral–spatial morphological attention transformer for hyperspectral image classification. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5503615. [\[CrossRef\]](#)
- Xiu, D.; Pan, Z.; Wu, Y.; Hu, Y. MAGE: Multisource attention network with discriminative graph and informative entities for classification of hyperspectral and LiDAR data. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5539714. [\[CrossRef\]](#)
- Xue, Z.; Tan, X.; Yu, X.; Liu, B.; Yu, A.; Zhang, P. Deep hierarchical vision transformer for hyperspectral and LiDAR data classification. *IEEE Trans. Image Process.* **2022**, *31*, 3095–3110. [\[CrossRef\]](#)
- Hang, R.; Li, Z.; Ghamisi, P.; Hong, D.; Xia, G.; Liu, Q. Classification of hyperspectral and LiDAR data using coupled CNNs. *IEEE Trans. Geosci. Remote Sens.* **2020**, *58*, 4939–4950. [\[CrossRef\]](#)
- Roy, S.K.; Krishna, G.; Dubey, S.R.; Chaudhuri, B.B. HybridSN: Exploring 3-D–2-D CNN feature hierarchy for hyperspectral image classification. *IEEE Geosci. Remote Sens. Lett.* **2019**, *17*, 277–281. [\[CrossRef\]](#)
- Zhang, S.; Meng, X.; Liu, Q.; Yang, G.; Sun, W. Feature-decision level collaborative fusion network for hyperspectral and LiDAR classification. *Remote Sens.* **2023**, *15*, 4148. [\[CrossRef\]](#)
- Ma, L.; Crawford, M.M.; Tian, J. Local Manifold Learning-Based k -Nearest-Neighbor for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2010**, *48*, 4099–4109. [\[CrossRef\]](#)
- Li, J.; Bioucas-Dias, J.M.; Plaza, A. Hyperspectral Image Segmentation Using a New Bayesian Approach With Active Learning. *IEEE Trans. Geosci. Remote Sens.* **2011**, *49*, 3947–3960. [\[CrossRef\]](#)
- Melgani, F.; Bruzzone, L. Classification of hyperspectral remote sensing images with support vector machines. *IEEE Trans. Geosci. Remote Sens.* **2004**, *42*, 1778–1790. [\[CrossRef\]](#)
- Farrell, M.; Mersereau, R. On the impact of PCA dimension reduction for hyperspectral detection of difficult targets. *IEEE Geosci. Remote Sens. Lett.* **2005**, *2*, 192–195. [\[CrossRef\]](#)
- Bandos, T.V.; Bruzzone, L.; Camps-Valls, G. Classification of hyperspectral images with regularized linear discriminant analysis. *IEEE Trans. Geosci. Remote Sens.* **2009**, *47*, 862–873. [\[CrossRef\]](#)
- Ye, Q.; Zhao, H.; Li, Z.; Yang, X.; Gao, S.; Yin, T.; Ye, N. L1-Norm distance minimization-based fast robust twin support vector k -plane clustering. *IEEE Trans. Neural Netw. Learn. Syst.* **2017**, *29*, 4494–4503. [\[CrossRef\]](#) [\[PubMed\]](#)
- Plaza, A.; Martinez, P.; Perez, R.; Plaza, J. A new method for target detection in hyperspectral imagery based on extended morphological profiles. In Proceedings of the 2003 IEEE International Geoscience and Remote Sensing Symposium, Toulouse, France, 21–25 July 2003 ; Volume 6, pp. 3772–3774. [\[CrossRef\]](#)
- Licciardi, G.; Marpu, P.R.; Chanussot, J.; Benediktsson, J.A. Linear versus nonlinear PCA for the classification of hyperspectral data based on the extended morphological profiles. *IEEE Geosci. Remote Sens. Lett.* **2011**, *9*, 447–451. [\[CrossRef\]](#)
- Hou, B.; Huang, T.; Jiao, L. Spectral–Spatial Classification of Hyperspectral Data Using 3-D Morphological Profile. *IEEE Geosci. Remote Sens. Lett.* **2015**, *12*, 2364–2368. [\[CrossRef\]](#)
- Qian, Y.; Ye, M.; Zhou, J. Hyperspectral image classification based on structured sparse logistic regression and three-dimensional wavelet texture features. *IEEE Trans. Geosci. Remote Sens.* **2012**, *51*, 2276–2291. [\[CrossRef\]](#)
- Jia, S.; Liao, J.; Xu, M.; Li, Y.; Zhu, J.; Sun, W.; Jia, X.; Li, Q. 3-D Gabor convolutional neural network for hyperspectral image classification. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 5509216. [\[CrossRef\]](#)

24. Jia, S.; Hu, J.; Zhu, J.; Jia, X.; Li, Q. Three-dimensional local binary patterns for hyperspectral imagery classification. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 2399–2413. [\[CrossRef\]](#)
25. Chen, Y.; Jiang, H.; Li, C.; Jia, X.; Ghamisi, P. Deep feature extraction and classification of hyperspectral images based on convolutional neural networks. *IEEE Trans. Geosci. Remote Sens.* **2016**, *54*, 6232–6251. [\[CrossRef\]](#)
26. Li, Y.; Zhang, H.; Shen, Q. Spectral–Spatial Classification of Hyperspectral Imagery with 3D Convolutional Neural Network. *Remote Sens.* **2017**, *9*, 67. [\[CrossRef\]](#)
27. Zhang, H.; Gong, C.; Bai, Y.; Bai, Z.; Li, Y. 3-D-ANAS: 3-D Asymmetric Neural Architecture Search for Fast Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5508519. [\[CrossRef\]](#)
28. Li, S.; Song, W.; Fang, L.; Chen, Y.; Ghamisi, P.; Benediktsson, J.A. Deep Learning for Hyperspectral Image Classification: An Overview. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 6690–6709. [\[CrossRef\]](#)
29. Wu, H.; Prasad, S. Convolutional Recurrent Neural Networks for Hyperspectral Data Classification. *Remote Sens.* **2017**, *9*, 298. [\[CrossRef\]](#)
30. Ma, Q.; Jiang, J.; Liu, X.; Ma, J. Deep Unfolding Network for Spatsiospectral Image Super-Resolution. *IEEE Trans. Comput. Imaging* **2022**, *8*, 28–40. [\[CrossRef\]](#)
31. Li, C.; Zhang, B.; Hong, D.; Yao, J.; Chanussot, J. LRR-Net: An Interpretable Deep Unfolding Network for Hyperspectral Anomaly Detection. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5513412. [\[CrossRef\]](#)
32. Chen, R.; Vivone, G.; Li, G.; Dai, C.; Chanussot, J. An Offset Graph U-Net for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 1–15. [\[CrossRef\]](#)
33. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is all you need. In *Proceedings of the Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2017; Volume 30, pp. 5998–6008.
34. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image is Worth 16 × 16 Words: Transformers for Image Recognition at Scale. In *Proceedings of the International Conference on Learning Representations*, Virtual, 3–7 May 2021.
35. He, J.; Zhao, L.; Yang, H.; Zhang, M.; Li, W. HSI-BERT: Hyperspectral Image Classification Using the Bidirectional Encoder Representation From Transformers. *IEEE Trans. Geosci. Remote Sens.* **2020**, *58*, 165–178. [\[CrossRef\]](#)
36. Hong, D.; Han, Z.; Yao, J.; Gao, L.; Zhang, B.; Plaza, A.; Chanussot, J. SpectralFormer: Rethinking hyperspectral image classification with transformers. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 5518615. [\[CrossRef\]](#)
37. He, C.; Sun, L.; Huang, W.; Zhang, J.; Zheng, Y.; Jeon, B. TSLRLN: Tensor subspace low-rank learning with non-local prior for hyperspectral image mixed denoising. *Signal Process.* **2021**, *184*, 108060. [\[CrossRef\]](#)
38. Wang, Z.; Ziou, D.; Armenakis, C.; Li, D.; Li, Q. A comparative analysis of image fusion methods. *IEEE Trans. Geosci. Remote Sens.* **2005**, *43*, 1391–1402. [\[CrossRef\]](#)
39. Morchhale, S.; Pauca, V.P.; Plemmons, R.J.; Torgersen, T.C. Classification of pixel-level fused hyperspectral and lidar data using deep convolutional neural networks. In *Proceedings of the 2016 8th Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS)*, Los Angeles, CA, USA, 21–24 August 2016; pp. 1–5. [\[CrossRef\]](#)
40. Xu, X.; Li, W.; Ran, Q.; Du, Q.; Gao, L.; Zhang, B. Multisource remote sensing data classification based on convolutional neural network. *IEEE Trans. Geosci. Remote Sens.* **2017**, *56*, 937–949. [\[CrossRef\]](#)
41. Su, Y.; Chen, J.; Gao, L.; Plaza, A.; Jiang, M.; Xu, X.; Sun, X.; Li, P. ACGT-Net: Adaptive cuckoo refinement-based graph transfer network for hyperspectral image classification. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5521314. [\[CrossRef\]](#)
42. Wu, X.; Hong, D.; Chanussot, J. Convolutional neural networks for multimodal remote sensing data classification. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 5517010. [\[CrossRef\]](#)
43. Hong, D.; Gao, L.; Hang, R.; Zhang, B.; Chanussot, J. Deep encoder–decoder networks for classification of hyperspectral and LiDAR data. *IEEE Geosci. Remote Sens. Lett.* **2020**, *19*, 5500205. [\[CrossRef\]](#)
44. Mohla, S.; Pande, S.; Banerjee, B.; Chaudhuri, S. Fusatnet: Dual attention based spectrospatial multimodal fusion network for hyperspectral and lidar classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, Seattle, WA, USA, 14–19 June 2020; pp. 92–93. [\[CrossRef\]](#)
45. Chen, T.; Chen, S.; Chen, L.; Chen, H.; Zheng, B.; Deng, W. Joint Classification of Hyperspectral and LiDAR Data via Multiprobability Decision Fusion Method. *Remote Sens.* **2024**, *16*, 4317. [\[CrossRef\]](#)
46. Wang, W.; Dai, J.; Chen, Z.; Huang, Z.; Li, Z.; Zhu, X.; Hu, X.; Lu, T.; Lu, L.; Li, H.; et al. InternImage: Exploring Large-Scale Vision Foundation Models With Deformable Convolutions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Vancouver, BC, Canada, 17–24 June 2023; pp. 14408–14419. [\[CrossRef\]](#)
47. Sun, X.; Wang, P.; Lu, W.; Zhu, Z.; Lu, X.; He, Q.; Li, J.; Rong, X.; Yang, Z.; Chang, H.; et al. RingMo: A Remote Sensing Foundation Model with Masked Image Modeling. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5612822. [\[CrossRef\]](#)

48. Lester, B.; Al-Rfou, R.; Constant, N. The Power of Scale for Parameter-Efficient Prompt Tuning. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Punta Cana, Dominican Republic, 7–11 November 2021; pp. 3045–3059. [\[CrossRef\]](#)
49. Jia, M.; Tang, L.; Chen, B.C.; Cardie, C.; Belongie, S.; Hariharan, B.; Lim, S.N. Visual Prompt Tuning. In Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23–27 October 2022; Springer: Cham, Switzerland, 2022; pp. 709–727. [\[CrossRef\]](#)
50. Liu, P.; Yuan, W.; Fu, J.; Jiang, Z.; Hayashi, H.; Neubig, G. Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing. *ACM Comput. Surv.* **2023**, *55*, 195. [\[CrossRef\]](#)
51. Tan, X.; Shao, M.; Qiao, Y.; Liu, T.; Cao, X. Low-Rank Prompt-Guided Transformer for Hyperspectral Image Denoising. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5520815. [\[CrossRef\]](#)
52. Kong, Y.; Cheng, Y.; Chen, Y.; Wang, X. Joint Classification of Hyperspectral Image and LiDAR Data Based on Spectral Prompt Tuning. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5521312. [\[CrossRef\]](#)
53. Zhou, L.; Geng, J.; Jiang, W. Joint classification of hyperspectral and LiDAR data based on position-channel cooperative attention network. *Remote Sens.* **2022**, *14*, 3247. [\[CrossRef\]](#)
54. Wang, D.; Hu, M.; Jin, Y.; Miao, Y.; Yang, J.; Xu, Y.; Qin, X.; Ma, J.; Sun, L.; Li, C.; et al. HyperSIGMA: Hyperspectral Intelligence Comprehension Foundation Model. *arXiv* **2024**, arXiv:2406.11519. [\[CrossRef\]](#) [\[PubMed\]](#)
55. Braham, N.A.A.; Albrecht, C.M.; Mairal, J.; Chanussot, J.; Wang, Y.; Zhu, X.X. SpectralEarth: Training Hyperspectral Foundation Models at Scale. *arXiv* **2024**, arXiv:2408.08447. [\[CrossRef\]](#)
56. Hong, D.; Zhang, B.; Li, X.; Li, Y.; Li, C.; Yao, J.; Yokoya, N.; Li, H.; Ghamisi, P.; Jia, X.; et al. SpectralGPT: Spectral Remote Sensing Foundation Model. *IEEE Trans. Pattern Anal. Mach. Intell.* **2024**, *46*, 5227–5244. [\[CrossRef\]](#)
57. Fang, S.; Li, K.; Li, Z. S²ENet: Spatial–spectral cross-modal enhancement network for classification of hyperspectral and LiDAR data. *IEEE Geosci. Remote Sens. Lett.* **2021**, *19*, 6504205. [\[CrossRef\]](#)
58. Zhao, Z.; Xu, X.; Li, S.; Plaza, A. Hyperspectral Image Classification Using Groupwise Separable Convolutional Vision Transformer Network. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5511817. [\[CrossRef\]](#)

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.