DPMamba: Distillation Prompt Mamba for Multimodal Remote Sensing Image Classification with Missing Modalities

Yueguang Yang, **Jiahui Qu***, **Ling Huang**, **Wenqian Dong**
State Key Laboratory of Integrated Service Network, Xidian University, Xi'an 710071, China
{yyg, hling}@stu.xidian.edu.cn, {jhqu, wqdong}@xidian.edu.cn

Abstract

Multimodal remote sensing image classification (RSIC) has emerged as a key focus in Earth observation, driven by its capacity to extract complementary information from diverse sources. Existing methods struggle with modality absence caused by weather or equipment failures, leading to performance degradation. As a solution, knowledge distillation-based methods train student networks (SN) using a full-modality teacher, but they usually require training separate SN for each modality absence scenario, increasing complexity. To this end, we propose a unified Distillation Prompt Mamba (DPMamba) framework for multimodal RSIC with missing modalities. DPMamba leverages knowledge distillation in a shared text semantic space to optimize learnable prompts, transforming them from "placeholder" to "adaptation" states by enriching missing modality information with full-modality knowledge. To achieve this, we focus on two main aspects: first, we propose a new modality-aware Mamba for dynamically and hierarchically extracting cross-modality interactive features, providing richer, contextually relevant representations for backpropagation-based optimization of prompts; and second, we introduce a novel text-bridging distillation method to efficiently transfer full-modality knowledge, guiding the inclusion of missing modality information into prompts. Extensive evaluations demonstrate the effectiveness and robustness of the proposed DPMamba.

1 Introduction

Remote sensing image classification (RSIC) plays a vital role in earth observation, including applications such as urban planning [Quan *et al.*, 2018], environmental monitoring [Nguyen and Liou, 2019] and more. With the advancement of image sensor technology, remote sensing images tend to be diversified. Numerous methods for multimodal RSIC [Yang *et al.*, 2024; Roy *et al.*, 2024] have emerged within the community, all of which have demonstrated promising results.

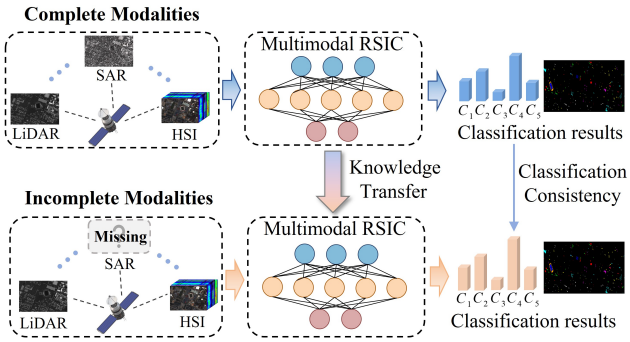

Figure 1: Illustration of multimodal RSIC with missing modalities.

Note that these methods are based on an assumption that the input multimodal data is consistent in both the training and inference. However, in practical applications, due to limitations such as harsh shooting environment or low sampling frequency of certain sensors [Huang *et al.*, 2024], the issue of missing modalities may arise during the inference. Existing multimodal classification methods are not designed and optimized for scenarios of missing modalities. The performance of these methods deteriorates significantly when certain modalities of the input data are missing during the inference.

Existing methods for addressing the problem of missing modalities can be categorized into three main approaches. The first approach utilizes generative models to synthesize missing modalities from the available modalities [Woo *et al.*, 2023; Liu *et al.*, 2023]. This approach requires initially training a generative model, followed by fine-tuning for scenarios with missing modalities. The classification performance largely depends on the quality of the synthesized modality. The second approach aims to learn a shared space that captures invariant information across the accessible modalities [Zhou *et al.*, 2021; Zhang *et al.*, 2022], which have demonstrated strong performance. Nevertheless, their effectiveness drastically decreases when only single modality data is available [Wang *et al.*, 2023b]. To address this issue, the community has developed various knowledge distillation-based methods [Li *et al.*, 2022; Wei *et al.*, 2023a], which transfer the full-modality information from a

*Corresponding Author

teacher network (TN) (pre-trained with full-modality data) to a student network (SN) (input incomplete or even single-modality data), as shown in Fig. 1. Due to the significant differences in inputs between the TN and the SN, existing knowledge distillation-based methods may encounter inefficiency and even overfitting issues during knowledge transfer. Furthermore, these methods often require training different models through knowledge distillation for various missing cases, which greatly increases computational costs. Recently, prompt learning has demonstrated potential in addressing the challenge of modality absence by designing task-aware prompts, which can significantly enhance the adaptability of multimodal tasks[Lee *et al.*, 2023]. However, the optimization of prompts is typically confined to task-specific loss functions, failing to specifically optimize for missing modality information and thereby hindering the full potential of prompts in complex scenarios with missing modalities.

Based on the above observations, we propose a Distillation Prompt Mamba (DPMamba) framework for multimodal RSIC with missing modalities, as shown in Fig. 2. The core idea of DPMamba is to utilize a knowledge distillation technique based on shared text semantic space to guide the optimization of missing-modality aware prompts (MMAPs), effectively leveraging full-modality knowledge to supplement and enrich the missing modality information in MMAPs, thereby transforming them from "placeholder" to "adaptation" states. Specifically, to effectively enable the transition of MMAPs, we focus on two critical components. First, we introduce a novel modality-aware Mamba backbone for both the TN and the SN, which is designed to capture dynamic cross-modal interactions at both global and local levels, providing essential feature representations for optimizing MMAPs via backpropagation. Second, we propose a new teacher-student paradigm called text bridging distillation (TBD) to enable efficient transfer of full-modality knowledge. By projecting the features of TN and SN into a shared semantic space, TBD facilitates knowledge transfer through feature alignment and inter-class relational constraints, offering robust guidance for optimization of MMAPs. In summary, our main contributions are threefold:

- We propose a unified DPMamba framework for multimodal RSIC with missing modalities, leveraging TBD for optimizing MMAPs from "placeholder" to "adaptation" states by enriching missing modality information within MMAPs with full-modality knowledge.
- We design a modality-aware Mamba as the backbone for extracting dynamic and multi-level feature, providing critical representations that facilitate the backpropagation-based optimization of MMAPs.
- We design a text bridging distillation (TBD) to facilitate the efficient transfer of full-modality knowledge, providing robust guidance for the optimization of MMAPs.

2 Related Work

2.1 Multimodal RSIC with Missing Modalities

Multimodal RSIC enables high-precision classification of complex scenes by integrating complementary information from multiple modalities [Qu *et al.*, 2024; Wang *et al.*, 2024]. However, most existing methods are not optimized for missing modality scenarios. Recent approaches address this by using autoencoders to impute missing modalities [Hafner and Ban, 2023; Chen *et al.*, 2024b] or transferring knowledge from teacher models trained on complete modalities to student models with partial data [Wei *et al.*, 2023c; Liu *et al.*, 2024a]. However, these methods encounter notable limitations when addressing complex scenarios, such as ensuring high-quality generation by generative models on large-scale datasets and the dependency of knowledge distillation-based approaches on training separate models for each missing modality case. This reliance significantly complicates the overall model design and training process.

2.2 Knowledge Distillation

Knowledge distillation (KD) is a deep learning technique for compressing large models into smaller, more efficient ones. In incomplete multimodal learning, KD enables the transfer of full-modality knowledge from a TN to a SN. KD methods fall into three categories: response-based, feature-based, and relationship-based. Response-based distillation [Wang *et al.*, 2023c] uses the soft target output of TN as the training target for SN but lacks internal knowledge transfer. Feature-based distillation [Li *et al.*, 2022] minimizes discrepancies between TN and SN hidden features, though mismatched features in incomplete multimodal data can cause overfitting [Garcia *et al.*, 2019]. Relationship-based distillation [Xin *et al.*, 2024] addresses this by leveraging inter-layer or inter-sample relationships for more effective knowledge transfer. Inspired by CLIP [Radford *et al.*, 2021] and DIST [Huang *et al.*, 2022], this paper introduces a new TBD that maps the representations of TN and SN into a shared text semantic space and utilizes feature constraints and inter-class relationship constraints to collaboratively guide the efficient transfer of full-modality information.

2.3 State Space Models

State Space Models (SSMs) [Gu *et al.*, 2021] are effective in capturing global context in deep learning due to their scalability. Gu et al.[Gu and Dao, 2023] introduced Mamba, an SSM-based model that even outperforms the Transformer [Vaswani, 2017] across multiple modalities. Recently, VMamba [Liu *et al.*, 2024b] extends the Mamba architecture to 2D images, achieving strong performance in computer vision. The Mamba architecture is now being applied to remote sensing tasks like change detection [Chen *et al.*, 2024a] and image classification [Yao *et al.*, 2024]. However, most of these works only utilize single-modality remote sensing image data, and are not optimized for scenarios of missing modalities. The potential of the Mamba architecture in the multimodal RSIC with missing modalities remains to be explored.

3 Method

3.1 Problem Definition

In this paper, we focus on the task of multimodal RSIC with missing modalities that generates a classification map categorizing each pixel into one of the C classes. For simplicity and

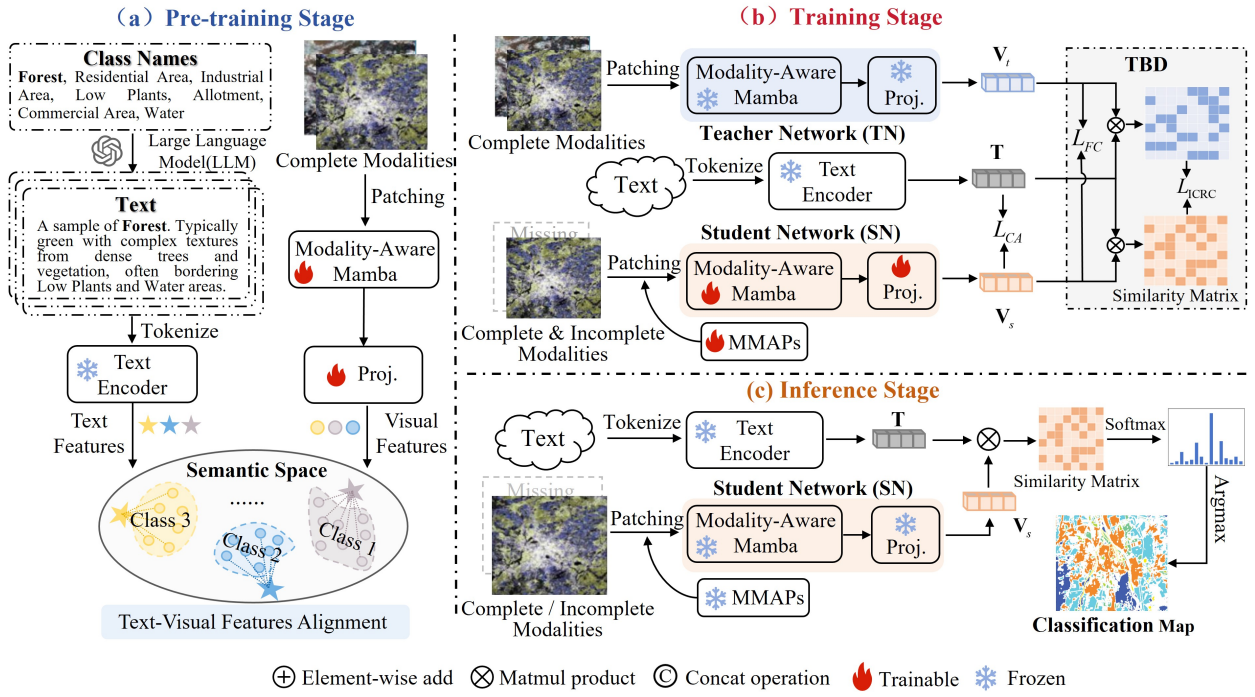

Figure 2: Overview of our proposed DPMamba.

without loss of generality, we consider a multimodal dataset $\mathcal{D} = \{\mathbf{X}_i^{m_1}, \mathbf{X}_i^{m_2}, y_i\}_{i=1}^{N_c}$, where $\mathbf{X}_i^{m_1}$ and $\mathbf{X}_i^{m_2}$ denote the i-th pair of multimodal patches from m_1 and m_2, respectively, y_i represents the label, N_c is the number of patches. There is a nonlinear mapping $M : \{\mathbf{X}_i^{m_1}, \mathbf{X}_i^{m_2}\} \rightarrow y_i$ from image patches to labels. Since certain modalities may be absent during inference, the mapping could be rephrased as $M_1 : \{\tilde{\mathbf{X}}_i^{m_1}, \mathbf{X}_i^{m_2}\} \rightarrow y_i$ or $M_2 : \{\mathbf{X}_i^{m_1}, \tilde{\mathbf{X}}_i^{m_2}\} \rightarrow y_i$, where $\tilde{\mathbf{X}}_i^{m_1}$ and $\tilde{\mathbf{X}}_i^{m_2}$ denote dummy inputs that keep the input structure consistent. Our goal is to optimize a unified and robust M that performs well regardless of whether the input data is complete or has missing certain modalities by fully exploiting the privileged modality information available only during training. The subscript i will be omitted in the following for brevity.

3.2 Overall Design

To achieve the above goal, we propose a unified framework DPMamba (Fig. 2), which leverages a knowledge distillation algorithm based on a shared semantic space to guide the optimization of missing-modality aware prompts (MMAPs) from "placeholder" to "adaptation" states, thereby enabling robust multimodal RSIC with missing modalities. To effectively facilitate the transition of MMAPs, we focus on two key aspects: first, we propose a novel modality-aware Mamba as the backbone of TN and SN to capture cross-modal dynamic interaction information at the global and local levels, which is important for backpropagation optimization of MMAPs. Second, we propose a novel teacher-student paradigm, termed text bridging distillation (TBD), for the efficient transfer of full-modality knowledge. TBD projects the features of the TN and the SN into a shared text semantic space and achieves knowledge transfer through feature alignment and inter-class relational constraints.

3.3 Modality-Aware Mamba

Multimodal remote sensing images typically contain rich information from different perspectives, and effectively extracting their complementary features remains a core research challenge. Inspired by the Mamba structure, which achieves global modeling with linear complexity, we propose a modality-aware Mamba that enables dynamic and multi-level visual feature extraction, based on the DWConv, a novel attention fusion (AF) and modality information scanning mechanism (MISM), as shown in Fig. 3. Different modality data are processed in separate branches, with injected fusion information enhancing cross-modal collaboration. The AF enables dynamic fusion, and the MISM collaborates with DWConv to capture multi-level features. Subsequently, we establish semantic associations through text-visual features alignment to achieve classification.

Attention Fusion

Due to the dynamic changes in the redundancy and complementarity of information between modalities in the case of missing data, a fixed fusion strategy is insufficient to adapt to these variations. To address this, we introduce a simple attention fusion (AF) mechanism that enables dynamic integration of information, thereby enhancing the network's robustness to modality missing scenarios. Formally, let $\{x_1, x_2\}$ represent the collection of input feature tensors. The attention mechanism computes attention scores for each input feature

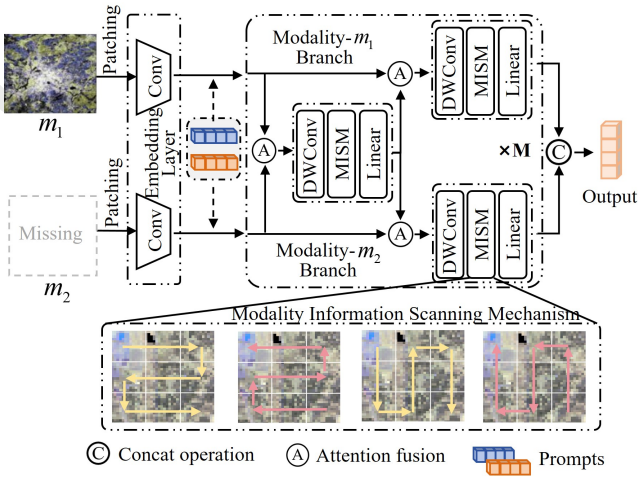

Figure 3: Ilustration of our proposed modality-aware Mamba.

tensor. Specifically, the attention score $score_i, i \in \{1, 2\}$ for each input tensor x_i is derived through a linear transformation, followed by a scalar output:

$$score_i = attention\_layer(x_i) = Wx_i + b \tag{1}$$

where W and b are the learnable parameters of the attention layer. Then, to transform the scores into attention weights, the attention scores are normalized using the softmax function, thereby ensuring that the weights are non-negative and sum to one:

$$\alpha_i = \frac{\exp(score_i - \max(\{x_1, x_2\}))}{\sum_{j=1}^{2} \exp(score_j - \max(\{x_1, x_2\}))} \tag{2}$$

where α_i is the attention weight for x_i, and subtracting the maximum score ensures numerical stability during the softmax operation. The final fused output h_a is obtained by computing the weighted sum of the input tensors, with the weights determined by the attention scores:

$$h_a = \alpha_1 x_1 + \alpha_2 x_2 \tag{3}$$

Modality Information Scanning Mechanism

SSMs process input sequences in a causal manner, but simply flattening the non-causal image patch into sequential input fails to fully capture modal information. To this end, we introduce a modality information scanning mechanism (MISM), which models the context by interacting each pixel with its neighboring pixels through SSMs. The combination of MISM and DWConv enables efficient multi-level feature extraction while maintaining relatively low computational complexity. Specifically, as shown in Fig. 3, we follow four zigzag paths to scan the patch pixel by pixel, i.e., top left to bottom right (width-priority scan and height-priority scan) , and their respective reversals. Given the output h_a of AF, we flatten the $p \times p$ feature patches into a sequence according to the scan path, and then calculate each element of the sequence $\mathcal{H}_{\text{seq}} = \{[h_{i,1}, h_{i,2}, \cdots, h_{i,p\times p}] \,|\, i \in \{1, 2, 3, 4\}\}$ by iteratively using the equation:

$$\begin{cases} \mathbf{h}_j^i = \bar{\mathbf{A}}\mathbf{h}_{j-1}^i + \bar{\mathbf{B}}h_{i,j} \\ \mathbf{y}_{i,j}^{m_1} = \mathbf{C}\mathbf{h}_j^i \end{cases} \tag{4}$$

where $\bar{\mathbf{A}}$, $\bar{\mathbf{B}}$, and $\mathbf{C}$ are trainable parameters of MISM. After rearranging to a consistent order, the outputs of each path are fused for comprehensive contextual information.

Text-Visual Features Alignment

For remote sensing image classification tasks, annotation information typically includes class labels and corresponding class names. Class text information has also been demonstrated as an effective form of supervisory signal for training, e.g., CLIP[Radford *et al.*, 2021]. To be exact, class text descriptions provide additional semantic constraints, enabling the model to better capture inter-class relationships and improve discriminative capability in complex scenarios. Technically speaking, we utilize a large language model to generate text based on the *class intrinsic attributes* and *inter-class relationships*. The text is subsequently tokenized and encoded into text features using a frozen text encoder, with its parameters initialized from CLIP's text encoder. Then, the projected visual features are aligned with the text features in the semantic space, thereby establishing deep semantic associations and enhancing classification robustness.

3.4 Text Bridging Distillation

To take full advantage of the privileged modal information available only during training, we design a text bridging distillation (TBD). Specifically, we employ contrast constraints to align the visual features of both the TN and SN with shared text features, mapping the visual features to a shared text semantic space. Subsequently, feature constraints and inter-class relationship constraints are utilized to collectively optimize the SN and MMAPs. The shared semantic space provides a unified context, ensuring semantic consistency between the TN and SN, thereby reducing information loss during the distillation process. Unlike directly imitating the prediction output of TN, TBD allows the SN to acquire a more generalized full-modality representations from TN. TBD guides the optimization of prompts, while the prompts alleviate distillation overfitting caused by significant input disparities due to modality absence, with both elements supporting and enhancing each other. Given the visual output features of the TN and SN as $\mathbf{V}_t \in \mathbb{R}^{C\times D_1}$ and $\mathbf{V}_s \in \mathbb{R}^{C\times D_1}$, and the text features as $\mathbf{T} \in \mathbb{R}^{C\times D_1}$, where C and D_1 denotes the number of class and channel, respectively. In more detail, we employ the feature constraint loss L_{FC} to distill fine-grained full-modality information as follows:

$$L_{\text{FC}} = \frac{1}{C}\sum_{i=1}^{C} \left(\mathbf{V}_t^{(i)} - \mathbf{V}_s^{(i)}\right)^2 \tag{5}$$

Furthermore, to capture the inherent inter-class relationship from the TN and enhance the classification performance of the SN, we initially derive the similarity matrices $\mathbf{S}^t = \cos(\mathbf{V}_t, \mathbf{T})$ and $\mathbf{S}^s = \cos(\mathbf{V}_s, \mathbf{T})$, where $\cos(\cdot, \cdot)$ represents the cosine similarity between two features. Following [Huang *et al.*, 2022], we then utilize the following inter-class relationship constraint loss L_{ICRC} to facilitate the transfer of inter-class information:

$$L_{\text{ICRC}} = \frac{1}{C}\left\{\sum_{i=1}^{C} d_{\text{P}}(\mathbf{S}_{i,:}^t, \mathbf{S}_{i,:}^s) + \sum_{j=1}^{C} d_{\text{P}}(\mathbf{S}_{:,j}^t, \mathbf{S}_{:,j}^s)\right\} \tag{6}$$

where $d_{\text{P}}(\cdot, \cdot)$ is a function for calculating Pearson's distance between column vectors or row vectors of $\mathbf{S}^t$ and $\mathbf{S}^s$.

In conclusion, the text bridging distillation loss L_{TBD} can be expressed as follows:

$$L_{\text{TBD}} = L_{\text{FC}} + L_{\text{ICRC}} \tag{7}$$

3.5 Prompt Injection

To handle classification tasks involving missing modalities, we integrate trainable MMAPs into modality-aware Mamba seamlessly, enabling the learning of modality-specific information and mitigating the issue of distillation overfitting caused by input inconsistencies by dynamically optimizing. Specifically, we define two modality-specific prompts, p_{m_1} and p_{m_2}, which are collaboratively optimized by task-specified criterion and distillation constraints. Given the complete input modalities $\{x_1, x_2\}$, if modality m_1 is missing during inference, the prompt p_{m_1} is introduced to replace the missing modality, resulting in the input $\{p_{m_1}, x_2\}$. This modified input is then fed into the backbone network, where p_{m_1} serves as a surrogate to guide feature extraction for the missing modality. By dynamically updating modality-specific prompts, the model can effectively leverage prior knowledge to mitigate the performance degradation caused by missing modalities.

3.6 Training Objective

The total loss L during the training stage is defined as follows:

$$L = L_{\text{TBD}} + L_{\text{CA}} \tag{8}$$

where L_{TBD} has been described above, and L_{CA} is the contrastive alignment loss. In both the pre-training stage of the TN and the training of the SN, we utilize the following L_{CA} as a task-specific criterion to establish the association between visual and text features for classification and map the visual features into a shared semantic space for distillation:

$$L_{\text{CA}} = \frac{(L_{\text{T}\to\text{V}} + L_{\text{V}\to\text{T}})}{2} \tag{9}$$

which is composed of a text-to-vison contrastive loss defined as:

$$L_{T\to V} = -\frac{1}{\text{C}} \sum_{n=1}^{\text{C}} \log \left(\frac{\exp\left(\tau \cdot \cos(\mathbf{T}^n, \mathbf{V}_{t/s}^n)\right)}{\sum_{k=1}^{\text{C}} \exp\left(\tau \cdot \cos(\mathbf{T}^n, \mathbf{V}_{t/s}^k)\right)} \right) \tag{10}$$

and a vison-to-text contrastive loss defined as:

$$L_{T\to V} = -\frac{1}{\text{C}} \sum_{n=1}^{\text{C}} \log \left(\frac{\exp\left(\tau \cdot \cos(\mathbf{V}_{t/s}^n, \mathbf{T}^n)\right)}{\sum_{k=1}^{\text{C}} \exp\left(\tau \cdot \cos(\mathbf{V}_{t/s}^k, \mathbf{T}^n)\right)} \right) \tag{11}$$

4 Experiments

4.1 Datasets

We conducted experiments on three public multimodal remote sensing datasets: 1) Houston dataset [Debes *et al.*, 2014], obtained in 2012, includes hyperspectral images (HSI) and LiDAR-derived digital surface model (DSM) data of the University of Houston campus and surrounding urban areas. It contains 15 object categories and 15,029 labeled samples. 2) Trento dataset [Rasti *et al.*, 2017], captured in a rural area near Trento, Italy, includes HSI and LiDAR-derived DSM data with 6 object categories and 30,214 samples. 3) Augsburg dataset [Baumgartner *et al.*, 2012], captured over Augsburg, Germany, contains HSI, LiDAR-derived DSM data and a dual-polarized SAR image with 4 feature channels. This dataset includes 7 object categories and 78,294 samples.

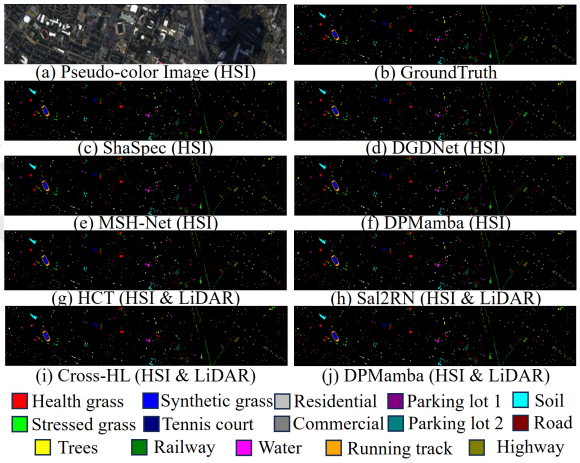

Figure 4: Classification maps of the Houston dataset.

4.2 Implementation Details

All experiments are conducted on Ubuntu 18.04 with an NVIDIA GeForce RTX 3090 GPU, using the PyTorch framework for model building, training, and evaluation. The training process of DPMamba consists of pre-training and training stages, with the AdamW optimizer used in both stages. The batch sizes for the Houston, Trento, and Augsburg datasets are 150, 120, and 140, respectively, and the optimal patch sizes are 11, 15, and 9. Pre-trained text encoder parameters from CLIP are loaded and fixed. During pre-training, the initial learning rate is set to 3e-4 for 500 epochs. In the training stage, the model parameters from pre-training are used for both TN and SN, with the learning rate maintained at 3e-4 for training the SN over 500 epochs. Code is available at https://github.com/Jiahuiqu/DPMamba.

4.3 Performance Comparison

In this section, we conduct a comprehensive comparison with concurrent baselines. We set up three different experimental configurations: 1) Training and testing are conducted within a single modality in the proposed method. "HSI-Net" refers to the model trained and tested using only the HSI modality, with other modalities following the same naming convention. 2) Advanced methods for joint classification in scenarios with modality absence, including ShaSpec [Wang *et al.*, 2023a], DGDNet [Wei *et al.*, 2023b] and MSH-Net [Wei *et al.*, 2023d]. 3) Advanced methods for joint classification with complete modalities, including HCT [Zhao *et al.*, 2023], Sal2RN [Li *et al.*, 2023] and Cross-HL [Roy *et al.*, 2024]. All methods are trained using 40 samples per class, with all samples used for testing. As for evaluation metrics, we report overall accuracy (OA), average accuracy (AA), and the kappa coefficient (κ), with κ values scaled by a factor of 100.

Method	Houston					Trento					Augsburg				
	Training Modalities	Testing Modalities	OA(%)	AA(%)	κ	Training Modalities	Testing Modalities	OA(%)	AA(%)	κ	Training Modalities	Testing Modalities	OA(%)	AA(%)	κ
	single modality					single modality					single modality				
HSI-Net	HSI	HSI	96.80	97.41	96.54	HSI	HSI	97.97	97.67	97.30	HSI	HSI	87.40	82.97	82.72
LiDAR/SAR-Net	LiDAR	LiDAR	60.24	63.51	57.17	LiDAR	LiDAR	92.87	93.24	90.63	SAR	SAR	75.37	62.34	67.25
	W/o HSI modality					W/o HSI modality					W/o HSI modality				
ShaSpec	HSI, LiDAR	LiDAR	49.44	50.43	45.62	HSI, LiDAR	LiDAR	87.85	87.85	84.04	HSI, SAR	SAR	48.01	32.62	33.12
DGDNet	HSI, LiDAR	LiDAR	59.07	60.39	55.92	HSI, LiDAR	LiDAR	93.46	91.71	91.37	HSI, SAR	SAR	51.92	30.66	31.71
MSH-Net	HSI, LiDAR	LiDAR	60.50	62.19	57.50	HSI, LiDAR	LiDAR	92.24	88.93	89.74	HSI, SAR	SAR	66.59	52.77	56.42
DPMamba (Ours)	HSI, LiDAR	LiDAR	**61.45**	**64.74**	**58.55**	HSI, LiDAR	LiDAR	**94.40**	**93.30**	**92.60**	HSI, SAR	SAR	**79.65**	**62.47**	**72.48**
	W/o LiDAR modality					W/o LiDAR modality					W/o SAR modality				
ShaSpec	HSI, LiDAR	HSI	94.24	95.18	93.78	HSI, LiDAR	HSI	96.08	96.03	94.79	HSI, SAR	HSI	86.42	73.97	80.93
DGDNet	HSI, LiDAR	HSI	93.51	94.53	92.98	HSI, LiDAR	HSI	97.60	96.88	96.80	HSI, SAR	HSI	78.40	74.69	71.06
MSH-Net	HSI, LiDAR	HSI	96.50	97.00	96.22	HSI, LiDAR	HSI	**98.41**	**97.59**	**97.88**	HSI, SAR	HSI	88.04	78.82	83.42
DPMamba (Ours)	HSI, LiDAR	HSI	**97.04**	**97.62**	**96.80**	HSI, LiDAR	HSI	98.35	97.13	97.79	HSI, SAR	HSI	**88.35**	**84.44**	**83.92**
	Complete modalties					Complete modalities					Complete modalities				
HCT	HSI, LiDAR	HSI, LiDAR	97.32	97.57	97.10	HSI, LiDAR	HSI, LiDAR	99.07	98.44	98.76	HSI, SAR	HSI, SAR	89.11	81.92	84.82
Sal2RN	HSI, LiDAR	HSI, LiDAR	97.68	98.01	97.49	HSI, LiDAR	HSI, LiDAR	99.14	98.64	98.86	HSI, SAR	HSI, SAR	91.18	82.05	87.63
Cross-HL	HSI, LiDAR	HSI, LiDAR	97.01	97.59	96.77	HSI, LiDAR	HSI, LiDAR	98.44	98.14	97.93	HSI, SAR	HSI, SAR	89.14	81.03	84.82
DPMamba (Ours)	HSI, LiDAR	HSI, LiDAR	**98.27**	**98.60**	**98.13**	HSI, LiDAR	HSI, LiDAR	**99.40**	**98.90**	**99.21**	HSI, SAR	HSI, SAR	**92.29**	**85.06**	**89.17**

Table 1: Classification accuracy of different methods on Houston, Trento and Augsburg HSI-SAR datasets. "W/o" denotes the missing modality in inference. "HSI/LiDAR/SAR-Net" indicates the proposed method trained and tested with only HSI/LiDAR/SAR data.

Method	Training Modalities	Testing Modalities	OA(%)	AA(%)	κ
Baseline	HSI,SAR,LiDAR	HSI	58.75	42.52	39.67
DPMamba (Ours)	HSI,SAR,LiDAR	HSI	89.47	83.98	85.26
Baseline	HSI,SAR,LiDAR	LiDAR	47.82	18.59	16.96
DPMamba (Ours)	HSI,SAR,LiDAR	LiDAR	45.35	45.70	32.76
Baseline	HSI,SAR,LiDAR	SAR	34.30	14.29	0.00
DPMamba (Ours)	HSI,SAR,LiDAR	SAR	68.39	61.24	59.07
Baseline	HSI,SAR,LiDAR	HSI,LiDAR	84.32	61.01	77.00
DPMamba (Ours)	HSI,SAR,LiDAR	HSI,LiDAR	90.18	85.15	86.26
Baseline	HSI,SAR,LiDAR	HSI,SAR	57.15	45.76	38.41
DPMamba (Ours)	HSI,SAR,LiDAR	HSI,SAR	91.42	85.10	87.96
Baseline	HSI,SAR,LiDAR	SAR,LiDAR	59.57	23.46	36.22
DPMamba (Ours)	HSI,SAR,LiDAR	SAR,LiDAR	69.59	67.21	60.61

Table 2: Classification accuracy of different methods.

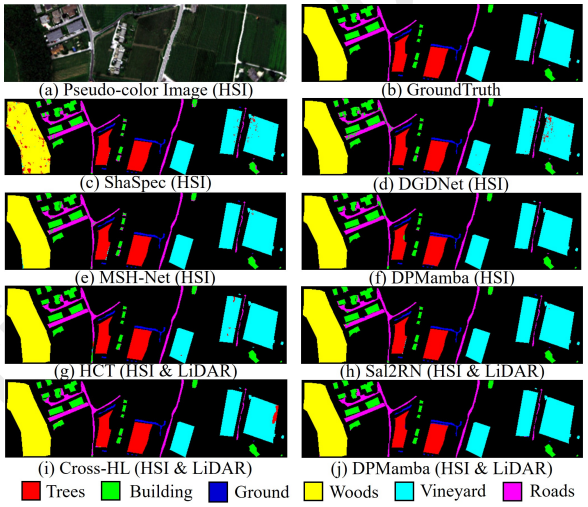

Figure 5: Classification maps of the Trento dataset.

Performance Analysis of HSI-LiDAR Classification

We conducted experiments on the Houston and Trento HSI-LiDAR datasets. As shown in Table 1, the proposed method achieves an OA of 98.27% on the Houston dataset with the complete modality, compared to 97.68% from baseline methods. In the W/o HSI scenario on the Trento dataset, the OA is 94.40%, and in the W/o LiDAR scenario, the OA is 98.35%, demonstrating robustness in scenarios with missing modalities. Due to its limited capacity, the unified DPMamba may perform slightly worse in certain scenarios compared to scenario-specific baseline methods. Additionally, visualizations in Fig. 4 and Fig. 5 show that the method accurately captures complex boundaries and heterogeneous regions, particularly in areas with high spectral variability, such as urban and vegetation zones. This highlights the effective use of the modality-aware Mamba to explore the consistency of spectral features, spatial structures and elevation information. The enriched text contextual semantics improve the recognition of subtle data distributions, thereby enhancing classification accuracy and reliability.

Performance Analysis of HSI-SAR Classification

Comparative experiments on the Augsburg HSI-SAR dataset are summarized in Table 1. The heterogeneity between HSI and SAR data, stemming from their distinct imaging mechanisms, challenges information fusion. The proposed method achieves an OA of 92.29% with complete modality, exceeding Sal2RN by 1.11%, and demonstrates notable improvements over MSH-Net in the W/o HSI (+13.06%) and W/o SAR (+0.31%) scenarios. This is attributed to the proposed backbone's effective fusion of heterogeneous information and highlights DPMamba's ability to efficiently transfer knowledge from the complete modality and accurately optimize SN and MMAPs. Visualization in Fig. 6 further confirms the robustness of DPMamba, even in complex urban distributions.

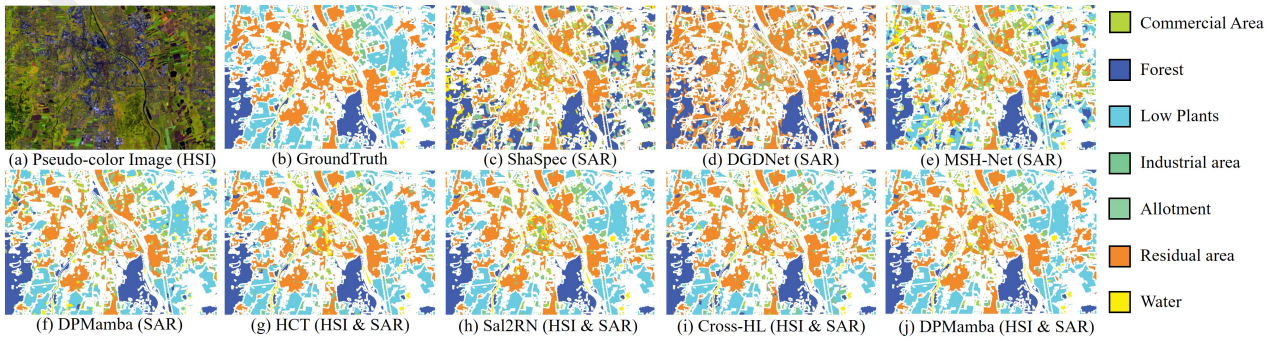

Figure 6: Classification maps of the Augsburg HSI-SAR dataset.

Datasets	Methods	W/o LiDAR/SAR			W/o HSI		
		OA (%)	AA (%)	κ	OA (%)	AA (%)	κ
Houston	Variant-1 (TBD replaced)	96.74	97.30	96.48	57.75	62.04	54.61
	Variant-2a (Backbone replaced)	94.70	95.32	94.27	50.60	52.68	46.93
	Variant-2b (AF replaced)	96.83	97.43	96.57	52.20	53.07	48.52
	Variant-2c (MISM replaced)	96.52	97.21	96.24	59.47	64.55	56.48
	Variant-3 (W/o MMAPs)	95.25	95.20	94.86	53.46	54.93	49.79
	DPMamba (ours)	**97.04**	**97.62**	**96.80**	**61.45**	**64.74**	**58.55**
Augsburg	Variant-1 (TBD replaced)	88.29	84.32	83.85	71.22	59.88	62.79
	Variant-2a (Backbone replaced)	87.08	80.77	82.24	71.01	52.12	61.84
	Variant-2b (AF replaced)	88.19	84.43	83.73	75.16	59.02	66.67
	Variant-2c (MISM replaced)	87.25	84.13	82.48	76.65	55.25	68.44
	Variant-3 (W/o MMAPs)	86.01	81.33	80.80	74.66	53.40	65.31
	DPMamba (ours)	**88.35**	**84.44**	**83.92**	**79.65**	**62.47**	**72.48**

Table 3: Results of ablation experiments.

Performance Analysis of HSI-LiDAR-SAR Classification
This section presents experiments on the Augsburg HSI-LiDAR-SAR dataset to assess the scalability of DPMamba in scenarios with multiple missing modalities. The results are shown in Table 2, where the baseline models are trained on the complete modality and evaluated under missing modality conditions, without optimizing for scenarios of missing modalities. We observe that DPMamba consistently outperforms the baseline models, with the OA showing a maximum increase of 34.27% (HSI & SAR). These results demonstrate that when multiple modalities are missing, the optimized MMAPs effectively alleviate the performance degradation caused by modality absence.

4.4 Ablation Study

Effectiveness of TBD
To assess the effectiveness of TBD, we conduct an ablation experiment by replacing L_{TBD} with KL Divergence in Variant-1. TBD enables full-modality information transfer through feature alignment and inter-class relation constraints based on a shared text semantic space. As shown in Table 3, in the W/o HSI scenario, DPMamba outperforms Variant-1 by 3.70% and 8.43% in OA on the Houston and Augsburg HSI-SAR datasets, respectively, demonstrating the advantages of TBD over KL Divergence replacement.

Effectiveness of Modality-Aware Mamba
To evaluate the modality-aware Mamba backbone and its components, we design three variants: Variant-2a (modality-aware Mamba replaced with ViT [Dosovitskiy, 2020]), Variant-2b (AF replaced with Add Fusion), and Variant-2c (MISM replaced with DWConv). Table 3 shows that DPMamba outperforms all variants, achieving average OA improvements of 6.60%, 4.73%, and 1.25% on the Houston dataset, and 4.96%, 2.33%, and 2.05% on the Augsburg HSI-SAR dataset across all scenarios of missing modalities. The performance drops in the Variants highlight the importance of dynamic, multi-level feature extraction, enabled by the AF and MISM, for robust multimodal fusion and handling missing modalities.

Effectiveness of Prompt Injection
The MMAPs are seamlessly integrated into the modality-aware Mamba, enabling the acquisition of missing modality information and mitigating performance degradation caused by modality absence. To further validate the effectiveness of prompt injection, we design Variant-3, which omits MMAPs and focuses solely on optimizing the SN for different scenarios of missing modalities. As shown in Table 3, DPMamba achieves average OA improvements of 4.89% and 3.67% over Variant-3 across all scenarios of missing modalities on the Houston and Augsburg HSI-SAR datasets, respectively.

5 Conclusion

We propose DPMamba, a unified framework for RSIC with missing modalities. By leveraging knowledge distillation in a shared text semantic space, DPMamba optimizes MMAPs with full-modality knowledge for enriching missing modality information, transforming them from "placeholder" to "adaptation" states. To enable the transition of MMAPs, we focus on two key elements. First, we introduce a novel modality-aware Mamba backbone for both TN and SN, designed to capture dynamic cross-modal interactions at global and local levels. Second, we propose a new TBD to efficiently transfer full-modality knowledge for optimizing MMAPs. Experimental results show that DPMamba outperforms baseline methods on the Houston, Trento, and Augsburg datasets, underscoring its effectiveness in multimodal fusion and handling missing modalities. Ablation studies further validate the contributions of TBD, modality-aware Mamba and prompt injection.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant 62201423 and 62471359, the Key Research and Development Program of Shaanxi under Grant 2025SF-YBXM-513, the Young Talent Fund of Association for Science and Technology in Shaanxi under Grant 20230117 and 20250133, the Young Talent Fund of Xi'an Association for Science and Technology under Grant 959202313052, in part by the Fundamental Research Funds for the Central Universities under Grant QTZX25084.

References

[Baumgartner *et al.*, 2012] Andreas Baumgartner, Peter Gege, Claas Henning Köhler, Karim Lenhard, and Thomas Schwarzmaier. Characterisation methods for the hyperspectral sensor hyspex at dlr's calibration home base. In *Remote Sensing*, 2012.

[Chen *et al.*, 2024a] Hongruixuan Chen, Jian Song, Chengxi Han, Junshi Xia, and Naoto Yokoya. Changemamba: Remote sensing change detection with spatio-temporal state space model. *arXiv preprint arXiv:2404.03425*, 2024.

[Chen *et al.*, 2024b] Yuxing Chen, Maofan Zhao, and Lorenzo Bruzzone. A novel approach to incomplete multimodal learning for remote sensing data fusion. *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

[Debes *et al.*, 2014] Christian Debes, Andreas Merentitis, Roel Heremans, Jürgen Hahn, Nikolaos Frangiadakis, Tim van Kasteren, Wenzhi Liao, Rik Bellens, Aleksandra Pižurica, Sidharta Gautama, Wilfried Philips, Saurabh Prasad, Qian Du, and Fabio Pacifici. Hyperspectral and lidar data fusion: Outcome of the 2013 grss data fusion contest. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 7(6):2405–2418, 2014.

[Dosovitskiy, 2020] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[Garcia *et al.*, 2019] Nuno C Garcia, Pietro Morerio, and Vittorio Murino. Learning with privileged information via adversarial discriminative modality distillation. *IEEE transactions on pattern analysis and machine intelligence*, 42(10):2581–2593, 2019.

[Gu and Dao, 2023] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

[Gu *et al.*, 2021] Albert Gu, Isys Johnson, Karan Goel, Khaled Saab, Tri Dao, Atri Rudra, and Christopher Ré. Combining recurrent, convolutional, and continuous-time models with linear state space layers. *Advances in neural information processing systems*, 34:572–585, 2021.

[Hafner and Ban, 2023] Sebastian Hafner and Yifang Ban. Multi-modal deep learning for multi-temporal urban mapping with a partly missing optical modality. In *IGARSS 2023-2023 IEEE International Geoscience and Remote Sensing Symposium*, pages 6843–6846. IEEE, 2023.

[Huang *et al.*, 2022] Tao Huang, Shan You, Fei Wang, Chen Qian, and Chang Xu. Knowledge distillation from a stronger teacher. *Advances in Neural Information Processing Systems*, 35:33716–33727, 2022.

[Huang *et al.*, 2024] Ling Huang, Wenqian Dong, Song Xiao, Jiahui Qu, Yuanbo Yang, and Yunsong Li. Language-guided visual prompt compensation for multimodal remote sensing image classification with modality absence. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 5161–5170, 2024.

[Lee *et al.*, 2023] Yi-Lun Lee, Yi-Hsuan Tsai, Wei-Chen Chiu, and Chen-Yu Lee. Multimodal prompting with missing modalities for visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14943–14952, 2023.

[Li *et al.*, 2022] Xiao Li, Lin Lei, Caiguang Zhang, and Gangyao Kuang. Dense adaptive grouping distillation network for multimodal land cover classification with privileged modality. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–14, 2022.

[Li *et al.*, 2023] Jiaojiao Li, Yuzhe Liu, Rui Song, Yunsong Li, Kailiang Han, and Qian Du. Sal²rn: A spatial–spectral salient reinforcement network for hyperspectral and lidar data fusion classification. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–14, 2023.

[Liu *et al.*, 2023] Hong Liu, Dong Wei, Donghuan Lu, Jinghan Sun, Liansheng Wang, and Yefeng Zheng. M3ae: Multimodal representation learning for brain tumor segmentation with missing modalities. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 1657–1665, 2023.

[Liu *et al.*, 2024a] Xiao Liu, Fei Jin, Shuxiang Wang, Jie Rui, Xibing Zuo, Xiaobing Yang, and Chuanxiang Cheng. Multimodal online knowledge distillation framework for land use/cover classification using full or missing modalities. *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

[Liu *et al.*, 2024b] Yue Liu, Yunjie Tian, Yuzhong Zhao, Hongtian Yu, Lingxi Xie, Yaowei Wang, Qixiang Ye, and Yunfan Liu. Vmamba: Visual state space model. *arXiv preprint arXiv:2401.10166*, 2024.

[Nguyen and Liou, 2019] Kim-Anh Nguyen and Yuei-An Liou. Global mapping of eco-environmental vulnerability from human and nature disturbances. *Science of the total environment*, 664:995–1004, 2019.

[Qu *et al.*, 2024] Jiahui Qu, Lijian Zhang, Wenqian Dong, Nan Li, and Yunsong Li. Shared-private decoupling-based multilevel feature alignment semi-supervised learning for hsi and lidar classification. *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

[Quan *et al.*, 2018] Sinong Quan, Boli Xiong, Deliang Xiang, and Gangyao Kuang. Derivation of the orientation parameters in built-up areas: With application to model-based decomposition. *IEEE Transactions on Geoscience and Remote Sensing*, 56(8):4714–4730, 2018.

[Radford *et al.*, 2021] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[Rasti *et al.*, 2017] Behnood Rasti, Pedram Ghamisi, and Richard Gloaguen. Hyperspectral and lidar fusion using extinction profiles and total variation component analysis. *IEEE Transactions on Geoscience and Remote Sensing*, 55(7):3997–4007, 2017.

[Roy *et al.*, 2024] Swalpa Kumar Roy, Atri Sukul, Ali Jamali, Juan M. Haut, and Pedram Ghamisi. Cross hyperspectral and lidar attention transformer: An extended self-attention for land use and land cover classification. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–15, 2024.

[Vaswani, 2017] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

[Wang *et al.*, 2023a] Hu Wang, Yuanhong Chen, Congbo Ma, Jodie Avery, Louise Hull, and Gustavo Carneiro. Multi-modal learning with missing modality via shared-specific feature modelling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15878–15887, 2023.

[Wang *et al.*, 2023b] Shuai Wang, Zipei Yan, Daoan Zhang, Haining Wei, Zhongsen Li, and Rui Li. Prototype knowledge distillation for medical segmentation with missing modality. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, 2023.

[Wang *et al.*, 2023c] Shuai Wang, Zipei Yan, Daoan Zhang, Haining Wei, Zhongsen Li, and Rui Li. Prototype knowledge distillation for medical segmentation with missing modality. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[Wang *et al.*, 2024] Haoyu Wang, Xiaomin Liu, Zhenzhuang Qiao, Guoqing Wang, and Haotian Chen. Multimodal remote sensing data classification based on gaussian mixture variational dynamic fusion network. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–14, 2024.

[Wei *et al.*, 2023a] Shicai Wei, Chunbo Luo, and Yang Luo. Mmanet: Margin-aware distillation and modality-aware regularization for incomplete multimodal learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20039–20049, 2023.

[Wei *et al.*, 2023b] Shicai Wei, Yang Luo, and Chunbo Luo. Diversity-guided distillation with modality-center regularization for robust multimodal remote sensing image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–14, 2023.

[Wei *et al.*, 2023c] Shicai Wei, Yang Luo, Xiaoguang Ma, Peng Ren, and Chunbo Luo. Msh-net: Modality-shared hallucination with joint adaptation distillation for remote sensing image classification using missing modalities. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–15, 2023.

[Wei *et al.*, 2023d] Shicai Wei, Yang Luo, Xiaoguang Ma, Peng Ren, and Chunbo Luo. Msh-net: Modality-shared hallucination with joint adaptation distillation for remote sensing image classification using missing modalities. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–15, 2023.

[Woo *et al.*, 2023] Sangmin Woo, Sumin Lee, Yeonju Park, Muhammad Adi Nugroho, and Changick Kim. Towards good practices for missing modality robust action recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 2776–2784, 2023.

[Xin *et al.*, 2024] Xiaomeng Xin, Heping Song, and Jianping Gou. A new similarity-based relational knowledge distillation method. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3535–3539. IEEE, 2024.

[Yang *et al.*, 2024] Yueguang Yang, Jiahui Qu, Wenqian Dong, Tongzhen Zhang, Song Xiao, and Yunsong Li. Tmcfn: Text-supervised multidimensional contrastive fusion network for hyperspectral and lidar classification. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–15, 2024.

[Yao *et al.*, 2024] Jing Yao, Danfeng Hong, Chenyu Li, and Jocelyn Chanussot. Spectralmamba: Efficient mamba for hyperspectral image classification. *arXiv preprint arXiv:2404.08489*, 2024.

[Zhang *et al.*, 2022] Yao Zhang, Nanjun He, Jiawei Yang, Yuexiang Li, Dong Wei, Yawen Huang, Yang Zhang, Zhiqiang He, and Yefeng Zheng. mmformer: Multimodal medical transformer for incomplete multimodal learning of brain tumor segmentation. *Springer, Cham*, pages 107–117, 2022.

[Zhao *et al.*, 2023] Guangrui Zhao, Qiaolin Ye, Le Sun, Zebin Wu, Chengsheng Pan, and Byeungwoo Jeon. Joint classification of hyperspectral and lidar data using a hierarchical cnn and transformer. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–16, 2023.

[Zhou *et al.*, 2021] Tongxue Zhou, Stéphane Canu, Pierre Vera, and Su Ruan. Latent correlation representation learning for brain tumor segmentation with missing mri modalities. *IEEE Transactions on Image Processing*, 30:4263–4274, 2021.