T-APT: Text-Guided Modality-Aware Prompt Tuning for Arbitrary Multimodal Remote Sensing Image Joint Classification

Qinghao Gao, Jiahui Qu, Wenqian Dong*

State Key Laboratory of Integrated Service Network, Xidian University, Xi'an 710071, China
qhgao@stu.xidian.edu.cn, jhqu@xidian.edu.cn, wqdong@xidian.edu.cn

Abstract

Multimodal remote sensing image joint classification has achieved significant progress. However, existing methods primarily focus on designing modality-specific networks, lacking adaptive generalization capabilities in diverse and dynamic modality combinations encountered in real-world scenarios. Leveraging vision foundation models pretrained on large-scale natural image datasets with proven effectiveness in heterogeneous downstream tasks, we propose a unified Text-guided Arbitrary Modality Prompt Tuning (T-APT) framework, which leverages complementary fused features to drive the foundation model and employs text-guided modality-specific prior knowledge as cross-modal prompts to fine-tune a pretrained Vision Transformer (ViT) model. Specifically, a Mamba-Based Arbitrary Modal-Focused Feature Capture (MAMF-FC) module is designed to extract complementary joint features and modality-specific prior knowledge from arbitrary modalities through a shared-specific scanning encoder-decoder architecture. Subsequently, a Text-Guided Modality-Aware Prompt Tuning (TMPT) module is proposed to support the adaptation of fused features to the foundation model, enabling our arbitrary remote sensing image classification task. Extensive experiments on public datasets spanning multispectral (MS), hyperspectral (HS), light detection and ranging (LiDAR), and synthetic aperture radar (SAR) modalities demonstrate that our T-APT achieves classification performance comparable to specialized networks across arbitrary modal combinations. The code is available at https://github.com/Jiahuiqu/TAPT.

Introduction

Joint classification of multi-source remote sensing images is an important research direction. Thanks to the development of sensing technology, multi-source remote sensing image classification technology has made great progress, and its application in the fields of environmental monitoring(Hu et al. 2023), disaster monitoring(Zhao et al. 2024) and so on has become more and more important. In previous work, researchers have achieved success in designing modality-specific dedicated networks. However, in real-world scenarios, the combination of modalities is flexible, as shown in Figure 1, in remote sensing data, models for specific modality combinations may not be able to adapt to specific feature structures when redeployed to new combinations, which can lead to performance issues or even prevent training. Consequently, the solution of the Arbitrary Modal Joint Classification (AMJC) task is crucial.

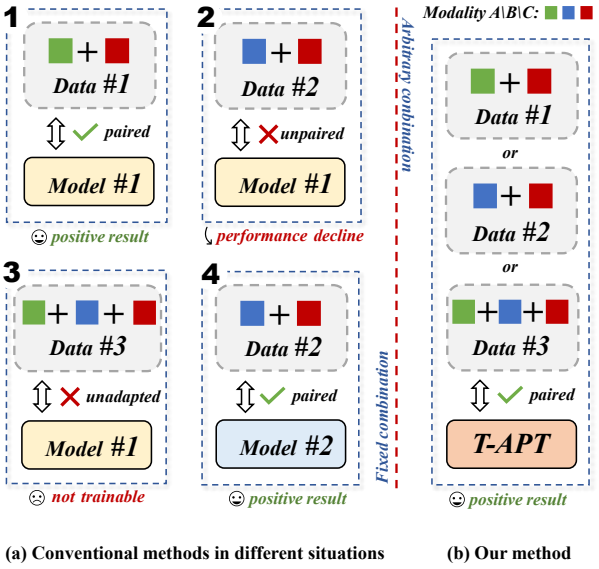

Figure 1: Comparison between conventional methods and our method, # represents the correspondence relationship between model and dataset. (a) Specific design methods in different situations. (b) Our proposed T-APT qdapts to arbitrarily datasets with different modality combinations.

The core of AMJC lies in effectively extracting complementary information from different modalities and constructing a unified feature interpretation method. This requires the model to not only thoroughly understand the specific and complementary information from different modalities but also to maintain robust generalization capabilities. Recently, thanks to substantial advances in visual foundation models, networks trained on massive datasets have demonstrated powerful feature analysis and generalization capabilities(Han et al. 2024), providing reliable support for downstream tasks in remote sensing classification. In our view, for

*Corresponding Authors

uncertain or dynamically changing classification scenarios, designing a new large foundation model holds far less value than fine-tuning existing foundation models with minimal parameter updates under resource-constrained conditions to adapt to specific tasks. The limitations of previous networks lie in their focus on prompt fine-tuning for single modalities or specific modal combinations rather than generalized multimodal tasks. Furthermore, for remote sensing image processing with rich feature properties and significant structural differences across modalities, relying solely on simple mapping-based embedding makes it difficult to effectively extract key features before the frozen net. Moreover, due to the significant differences between modalities, a unified network may lose information from "less important" modalities during training, leading to the modal inertia problem.

To address these challenges, we propose a unified framework text-guided arbitrary-modal prompt tuning (T-APT) specifically designed for arbitrary-modality remote sensing image classification. The core of this method lies in fusing complementary multi-modal information and decoupling modality-specific features from arbitrary modalities to learn text-guided visual prompts, thereby enabling the effective adaptation of multi-modal remote sensing imagery to foundation visual models. Specifically, we first design a unified mamba-based arbitrary modal-focused feature capture (MAMF-FC) module, where a specific-shared integrated Mamba fusion block performs multi-directional scanning over the common spatial domain and modality-specific channels with linear computational complexity. During this process, MAMF-FC dynamically interacts with modality-specific information, enabling unified modeling of both modality-specific and complementary information distributions. In addition, we propose an aligned Mamba module that performs cross-modal scanning to fuse multi-source data, enhancing fine-grained modality correspondence. The fused multi-modal information is subsequently fed into a frozen foundation model, and a text-guided modality-aware prompt tuning (TMPT) mechanism is proposed to perform prompt-based fine-tuning, guiding cross-modal information interaction and enhancing the model's generalization ability. Meanwhile, to address the modal inertia challenge arising from disparate modality contributions in the AMJC task, we propose a strategy that regulates and guides image prompt information through dual text feature.

To summarize, the contributions of this work are as follows:

- We propose a unified fine-tuning model T-APT for arbitrary multi-source remote sensing classification, which learns complementary and modality-specific features of multimodal data and fine-tuning the foundation model for downstream tasks by prompting.
- We design a unified MAMF-FC module, which obtains fused information and modality-specific information through an specific-shared integrated Mamba fusion block structure.
- We develop TMPT, which converts modality-specific information into text-guided prompt content and introduces cross-modal prompt information through a modality-aware control (MAC) module to guide multimodal interaction and model generalization.
- We propose a dual text-guided modality-balanced prompt generation method, which balances the strength of different modality-specific data through text guidance, preserving image modality features while constructing a unified text modality prompt space.

Related Work

Multi-modal Based Classification in RS

In recent years, with the development of deep learning techniques, deep learning-based methods for multisource remote sensing (RS) image classification have been extensively studied. Wang(Wang et al. 2022a) proposed an adaptive mutual learning-based multimodal data fusion network (AM3Net), enhancing fusion and classification performance through adaptive mutual learning. Jha(Jha, Bose, and Banerjee 2023) modeled complex intra- and inter-modal relationships by integrating memory mechanisms and quaternion operations, while enforcing global feature consistency through cross-modal contrastive learning (CMCL). Yang(Yang et al. 2024) utilized textual information to guide models in learning cross-modal correlations, introducing a text-supervised mechanism to enhance contrastive fusion networks. Although these methods have successfully achieved multi-source remote sensing image classification, the networks are modality-specific and cannot cope with changing realities.

Prompt Learning

In recent years, prompt learning has made significant progress in the application of large-scale pre-trained foundation models. These methods optimize the model's adaptability to specific tasks by introducing learnable prompts, thereby improving performance on downstream tasks(Dong, Gu, and Liu 2024). Prompt learning has also been applied in downstream domains such as medicine(Wu and Xu 2024; Ahmed et al. 2024), remote sensing(Fang et al. 2023; Liu et al. 2023), and 3D scenes(Zha et al. 2023; Wang et al. 2022b; Tao et al. 2025). These studies demonstrate that prompt learning, by introducing more flexible and efficient prompt generation methods, significantly enhances the adaptability of pre-trained models for specific tasks. It enables models to flexibly adapt to a variety of tasks in different input scenarios, providing a promising approach to integrating prompt learning into multimodal learning for solving the problem of arbitrary modality inputs.

Method

Overview

We denote the dataset of arbitrarily modal for remote sensing images classification by $\mathcal{D} = \{(X_{\mathcal{M}_1}, ..., X_{\mathcal{M}_n}, y)\}$, where the sample in $X_{\mathcal{M}_j} \in \mathbb{R}^{H \times W \times C_j}$ is a three-dimensional tensor input, representing the j-th modality data of the sample, where H is the height, W is the width, and C_j is the number of channels, and y is the category labels. For different $\mathcal{D}$, the length of $\mathcal{X} = \{X_{\mathcal{M}_1}, ..., X_{\mathcal{M}_n}\} \in \mathcal{D}$

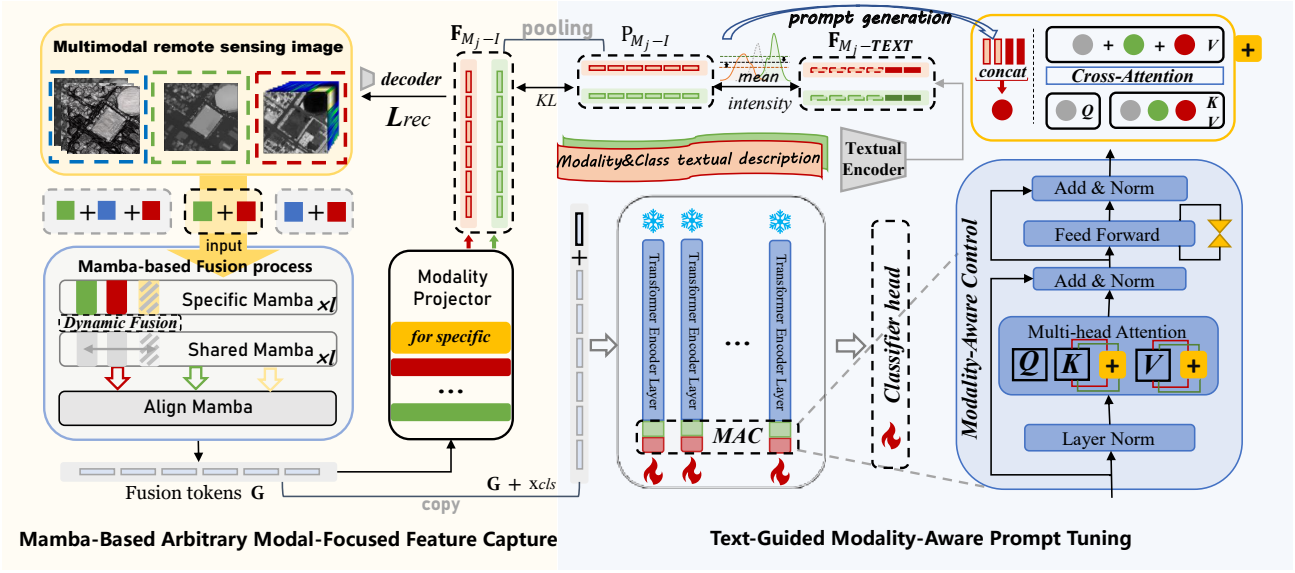

Figure 2: Overall architecture of the proposed T-APT. The method consists of two stage:1) Mamba-Based Arbitrary Modal-Focused Feature Capture. 2) Text-Guided Modality-Aware Prompt Tuning.

(i.e., the number of modalities) may vary, and each $X_{\mathcal{M}_j}$ may also differ. Our goal is to learn an optimal mapping $f_\phi : \mathcal{X} \rightarrow y$, which adapts to the diverse number and compositions of multimodal combinations across different datasets.

Considering the different data characteristics of uncertain numbers, the proposed method jointly optimizes f_ϕ by leveraging both complementary and specific information. During the optimization process, complementary information from different modalities is integrated to enhance the overall representation of the foundation model backbone input data. Additionally, the features of each individual modality are fully exploited and utilized in forming modality-aware prompts to fine-tune and compensate for representation bias due to data imbalance.The ultimate goal is to minimize the error between predicted and true labels in the f_ϕ process when applying the foundation model to any modal combination $\mathcal{X}$.

Our overall network is illustrated in Figure 2. We propose a T-APT model for arbitrary-modality remote sensing image classification, which learns to unifiedly embed complementary and modality-specific features of multimodal data and generates modality-aware visual prompt information. By reintroducing multimodal feature prompts, it achieves fine-tuning adaptation of the foundation model for downstream tasks. Specifically, we design a Mamba-Based Arbitrary Modal-Focused Feature Capture (MAMF-FC) module, which extracts and fuses multimodal features through an encoder-decoder structure. Additionally, we develop a Text-Guided Modality-Aware Prompt Tuning (TMPT) module, which constructs a unified prompt space using text guidance and introduces prompt information through a cross-modal approach.

Mamba-Based Arbitrary Modal-Focused Feature Capture

Traditional methods are usually designed for specific remote sensing modalities, and directly applying them to other modality combinations may lead to the loss of complementary information and degraded performance. Leveraging the multi-view relational dependencies of the Mamba architecture, we propose a unified encoder-decoder module, MAMF-FC, which incorporates a specific-shared scanning structure. This design maximizes the preservation of modality-specific features while enabling effective cross-modal interaction. Specifically, given a set of entities $\mathcal{X}$, for each $X_{\mathcal{M}_j}$, the corresponding unified channel feature $\mathbf{H}^0_{\mathcal{M}_j} \in \mathbb{R}^{H\times W\times C}$ is first obtained:

$$\mathbf{H}^0_{\mathcal{M}_j} = \Theta(C_j, C)\left(X_{\mathcal{M}_j}\right) \tag{1}$$

where $\Theta(\cdot)$ represents the convolution operation, and the subsequent specific-shared interleaved Mamba feature forward extraction is a multi-layer process, which can be expressed as:

$$\mathbf{H}^{l_{spe}}_{\mathcal{M}_j} = \Phi_{Spe}\left(\mathbf{H}^{l-1}_{\mathcal{M}_j}\right) \tag{2}$$

$$\mathbf{H}^{l}_{\mathcal{M}_j} = \Phi_{Sha}\left(\mathbf{H}^{l_{spe}}_{\mathcal{M}_j} \oplus \ell\left(\sum_{j=1}^{M}\sigma\left(\mathbf{H}^{l_{spe}}_{\mathcal{M}_j}\right)\right)\right) \tag{3}$$

where $\mathbf{H}^{l}_{\mathcal{M}_j}$ represents the features output by the l-th layer ($l \in \{1, 2\}$), $\mathbf{H}^{l_{spe}}_{\mathcal{M}_j}$ represents the specific coding features of the j-th modality, ℓ denotes the linear mapping, σ represents the activation function (PReLU is used here), $\oplus$ denotes the addition operation, $\Phi_{Spe}(\cdot)$ and $\Phi_{Sha}(\cdot)$ represent the specific and shared Mamba(Liu et al. 2024) blocks. $\Phi_{Sha}(\cdot)$ captures modality-common information through

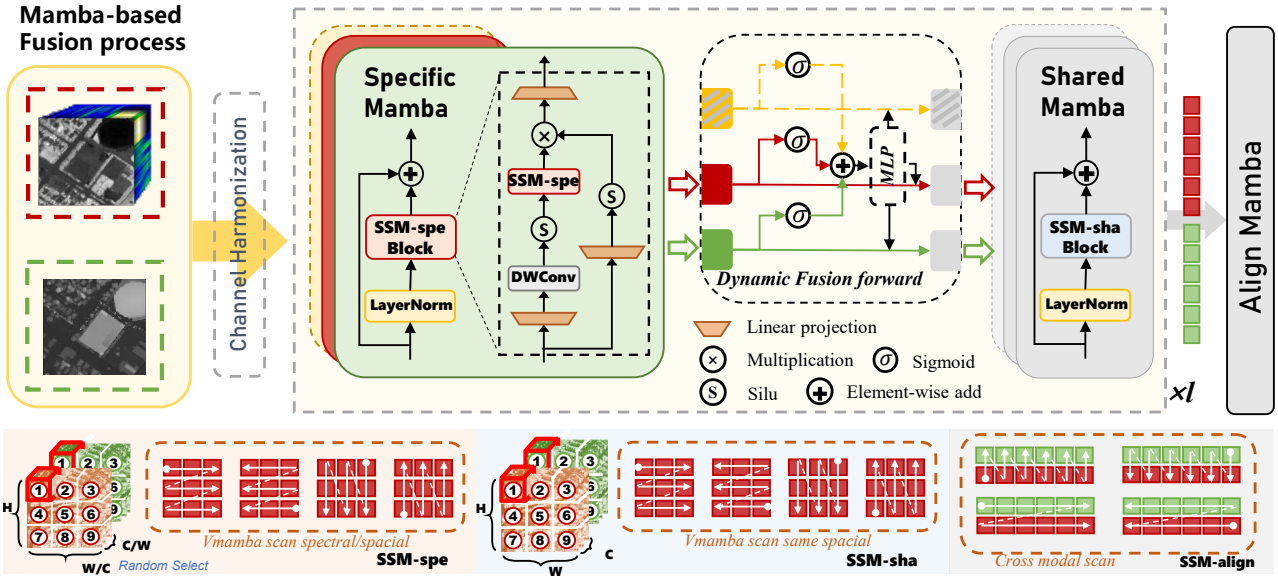

Figure 3: The structure of proposed Mamba-based Fusion process.

spatial multi-directional scanning, while $\Phi_{Spe}(\cdot)$ introduces channel scanning to extract unique characteristics.

As shown in Figure 3, the Mamba-based fusion process involves a shared-specific intertwined scanning mechanism and dynamic forward injection between modalities, followed by a cross-modal alignment and integration through the aligned Mamba module.

The results of the multi-layer forward propagation are denoted as $\mathbf{G}_{\mathcal{M}_j} \in \mathbb{R}^{\frac{H}{l} \times \frac{W}{l} \times C}$. In the final stage of the fusion process, we propose a cross-modal scanning alignment method, which unifies the modality relationships among multiple features while aligning the data distribution with the base model. The fusion result $\mathbf{G}$ is represented as:

$$\mathbf{G} = \ell_{adapt}\left(\Phi_{Align}\left(\mathbf{G}_{\mathcal{M}_1}, \ldots, \mathbf{G}_{\mathcal{M}_n}\right)\right), \mathbf{G} \in \mathbb{R}^{N \times D} \tag{4}$$

where N denotes the number of input tokens, and D represents the length of the tokens, $\Phi_{Align}(\cdot)$ represents the scanning Mamba for modality alignment, which fuses multi-modal information through cross-modal scanning, and $\ell_{adapt}(\cdot)$ denotes the adaptive linear layer. The decoupling process is expressed as follows:

$$\mathbf{F}_{\mathcal{M}_j - I} = \ell_{dec}(\mathbf{G}) \tag{5}$$

$$\mathcal{L}_{rec} = \frac{1}{M} \sum_{j=1}^{M} \| doc(\mathbf{F}_{\mathcal{M}_j - I}) - X_{\mathcal{M}_j} \|_2^2 \tag{6}$$

where $\mathbf{F}_{\mathcal{M}_j - I} \in \mathbb{R}^{N^1 \times D}$ represents the image feature of the j-th modality, obtained by applying a linear mapping $\ell_{dec}(\cdot)$ to $\mathbf{G}$, $\mathcal{L}_{rec}$ is the reconstruction error, $doc(\cdot)$ represents the decoding operation, and $\|\cdot\|$ denotes the mean squared error loss calculation.

Text-Guided Modality-Aware Prompt Tuning

The fine-tuning process proposed in this paper employs the pre-trained ViT-B/14(Dosovitskiy et al. 2021) version as the base model framework. To adapt the original backbone network to the downstream task and avoid potential modality laziness issues in multi-modal training, the text-guided modality-aware prompt fine-tuning process can be divided into two parts:

Dual Text-Guided Modality Balance Prompt Generation In this module, we map modality-specific features $\mathbf{F}_{\mathcal{M}_j\text{-}I}$ into $\mathbf{P}_{\mathcal{M}_j\text{-}I}$ through pooling. Our goal is to retain modality-specific characteristics in the prompt content while eliminating modality bias. Specifically, we design two types of text templates: class-level and modality-level descriptions (as shown in Figure 4), and encode these descriptions using CLIP's text encoder:

$$\begin{aligned} \mathbf{F}_{\mathcal{M}_{class}} &= \mathcal{E}_T(T_{class}), \\ \mathbf{F}_{\mathcal{M}_j\text{-}modality} &= \mathcal{E}_T(T_{j-modality}) \end{aligned} \tag{7}$$

where T_{class} denotes class descriptions and $T_{j-modality}$ represents modality-specific text descriptions for the j-th modality. $\mathbf{F}_{\mathcal{M}_{class}}$ and $\mathbf{F}_{\mathcal{M}_j\text{-}modality}$ are the outputs of the text encoder $\mathcal{E}_T$. Subsequently, we employ a Kullback–Leibler divergence loss to align the image feature distribution with modality-specific characteristics, and an L2-norm mean constraint to eliminate modality magnitude discrepancies:

$$\begin{aligned} \mathcal{L}_{align} = &\frac{1}{k} \sum_{j=1}^{k} \mathrm{KL}\left(\mathbf{F}_{\mathcal{M}_j\text{-}I} \parallel \mathbf{P}_{\mathcal{M}_j\text{-}I}\right) \\ &+ \left| \|\mathbf{P}_{\mathcal{M}_j\text{-}I}\|_2^{mean} - \|\mathbf{F}_{\mathcal{M}_{class}}\|_2^{mean} \right|. \end{aligned} \tag{8}$$

We define $\mathbf{P}_{\mathcal{M}_j} = \left\{\mathbf{P}_{\mathcal{M}_j\text{-}I}, \mathbf{F}_{\mathcal{M}_j\text{-}modality}\right\}$ as the generated prompt, where the invariant $\mathbf{F}_{\mathcal{M}_j\text{-}modality}$ acts as a task-specific prompt.

Modality-Aware Control In the feature fusion stage, we obtain the multi-modal complementary joint information $\mathbf{G} \in \mathbb{R}^N \times D$. In the frozen network, we take $\mathbf{Z}_0 =$

$\{\mathbf{G}; \mathbf{x}_{cls}\} \in \mathbb{R}^{(N+1)\times D}$ as the input to the L-layer Transformer encoder, where $\mathbf{x}_{cls}$ is an extra learnable classification token to form an extension feature. We propose a modality-aware control (MAC) prompt tuning mechanism, which injects cross-modal information into the attention blocks of the Transformer encoder layers and incorporates a similar low-rank adapter structure in the feed-forward layers. Thus, the training process in the frozen network can be expressed as:

$$\mathbf{Z}_{l+1} = \mathcal{E}_l(\mathbf{Z}_l), \quad l = \{0, 1, \ldots, L\} \tag{9}$$

$$\mathbf{A}_l = att_l(\mathbf{Q}_l, \mathbf{K}_l, \mathbf{V}_l), \begin{cases} \mathbf{Q}_l = W_q(\mathbf{Z}_l), \\ \mathbf{K}_l = W_k(\mathbf{Z}_l; \mathbf{P}_{\mathcal{M}_1}; \ldots; \mathbf{P}_{\mathcal{M}_n}), \\ \mathbf{V}_l = W_v(\mathbf{Z}_l; \mathbf{P}_{\mathcal{M}_1}; \ldots; \mathbf{P}_{\mathcal{M}_n}). \end{cases} \tag{10}$$

where $\mathcal{E}_l$ denote the l-th frozen Transformer encoder layer, $\mathbf{Z}_l$ be its input, and att_l represent the attention operation. The output of the attention layer is denoted as $\mathbf{A}_l$. According to the MAC mechanism, the key $\mathbf{K}_l$ and value $\mathbf{V}_l$ of the l-th layer are augmented by injecting cross-modal information through the query $\mathbf{Q}_l$, we define $(:)$ as the concatenated representation of $\mathbf{Z}_l$ and $\mathbf{P}_{\mathcal{M}_j}$, which serves as the input to the Key and Value projections at the l-th layer, and W_k, W_v, W_q are the pre-trained attention key, value, and query weight matrices, respectively. Let $\mathbf{z}_l$ and $\mathbf{x}_l$ be the input and output of the feed-forward network (FFN) in the Transformer encoder layer. Three projection matrices $W \in \mathbb{R}^{d\times k}$, $W_{\text{up}} \in \mathbb{R}^{d\times r}$, and $W_{\text{down}} \in \mathbb{R}^{r\times k}$ satisfy $r \ll \min(d, k)$, where σ is the activation function. During the forward propagation in the frozen network, we integrate information across arbitrary modalities using the MAC prompt tuning technique:

$$\mathbf{z}_l = W\mathbf{x}_l + W_{up}(\sigma(W_{down}\mathbf{x}_l)) \tag{11}$$

Finally, we append a classification head and perform classification on the output $\mathbf{x}_{cls}^l$ of the last layer using a MLP to obtain the predicted label $\hat{y}$:

$$\hat{y} = MLP(\mathbf{x}_{cls}^l) \tag{12}$$

Training Objective

The overall optimization loss is composed of classification loss, reconstruction loss and alignment loss, given by the following equation:

$$\mathcal{L} = \mathcal{L}_{cls} + \alpha_1\mathcal{L}_{rec} + \alpha_2\mathcal{L}_{align} \tag{13}$$

where the hyperparameter α_1 and α_2 control the balance of multiple losses, and L_{cls} is the classification loss, and here we use cross-entropy loss, which is calculated using the following formula:

$$\mathcal{L}_{cls} = -\frac{1}{N}\sum_{i=1}^{N} y_i \log(\hat{y}_i) \tag{14}$$

where N is the number of sample classes, y_i is the true label represented by one-hot encoding, and $\hat{y}_i$ is the predicted result.

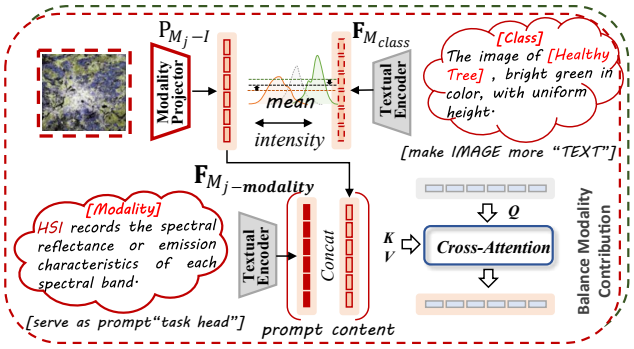

Figure 4: Eliminate modality scale differences through category text, use modality text as task tokens, thereby balancing modality contributions in visual tasks.

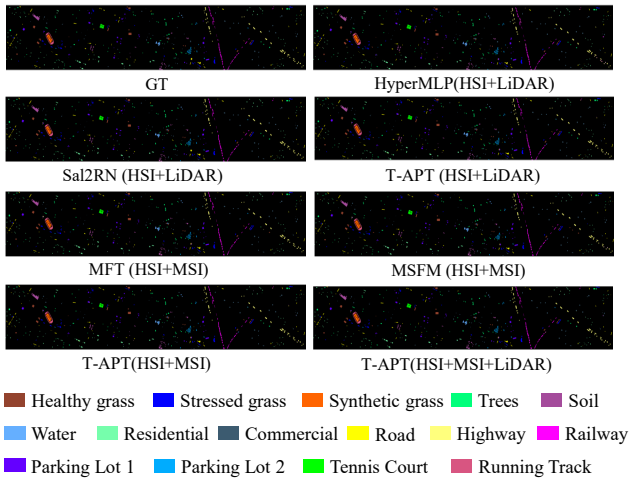

Figure 5: Classification maps of Houston2013 dataset.

Experiments

Datasets Description

We conduct experiments on three publicly multi-modal datasets for performance evaluation: The Houston2013 (Contest 2013) dataset includes LiDAR, hyperspectral imagery (HSI), and multispectral imagery (MSI) data, the dataset consists of 349 × 1905 pixels. The MUUFL(Gader et al. 2013) dataset contains registered HSI and LiDAR-based digital surface model (DSM) data, the dataset consists of 325 × 220 pixels. The Augsburg (Baumgartner et al. 2012) dataset contains HS, synthetic aperture radar (SAR), and LiDAR, the scene consists of 332 × 485 pixels.

Experimental Setup

Evaluation Metrics The classification accuracies are rigorously quantified through: Overall accuracy (OA), average accuracy (AA), and kappa coefficient.

Implementation Details The proposed method is implemented on the PyTorch platform and trained on one NVIDIA GeForce 3090 GPU using the Adam optimizer. We use the off-the-shelf Vit from HuggingFace as our pre-trained

Dataset	Method	Modality	OA(%)	AA(%)	$\kappa\times100$
Houston	HyperMLP(24)	H-L	97.85	98.19	97.68
	Sal2RN(23)	H-L	96.99	97.39	96.74
	T-APT(ours)	H-L	**98.89**	**99.09**	**98.80**
	MFT(23)	H-M	97.73	98.15	97.55
	MSFM(25)	H-M	98.51	98.68	98.39
	T-APT(ours)	H-M	**98.60**	**98.88**	**98.49**
	T-APT(ours)	H-M-L	98.94	99.14	98.85
MUUFL	MSFE-IFN(24)	H-L	89.37	92.23	86.30
	AM3Net(22)	H-L	85.10	86.15	80.93
	AMSSE(23)	H-L	90.19	92.51	87.30
	M2FNet(24)	H-L	87.89	91.01	84.42
	T-APT(ours)	H-L	**90.90**	**93.60**	**88.20**
Augsburg	S2ENet(22)	H-L	89.61	83.46	85.66
	FDNet(24)	H-L	91.81	87.07	88.57
	T-APT(ours)	H-L	**92.59**	**88.31**	**89.58**
	DSTD(23)	H-S	91.78	88.20	88.36
	MACN(23)	H-S	89.70	89.83	85.81
	T-APT(ours)	H-S	**91.80**	**90.63**	**88.59**
	T-APT(ours)	H-S-L	93.17	90.99	90.41

Table 1: Methods Comparison on Multi-modal Datasets.

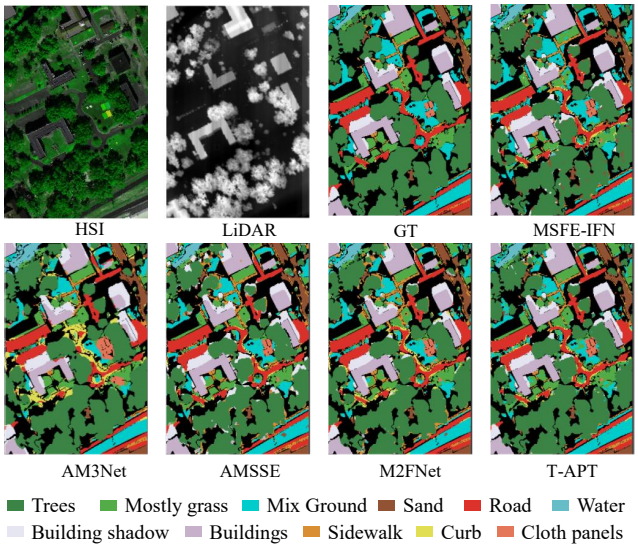

Figure 6: Classification maps of MUUFL dataset.

model. The model is trained for 250 epochs and the learning rate is set to 1e-4. α_1 and α_2 are set to 0.1 and 0.5.

Competing Methods To thoroughly assess the effectiveness of the proposed network in multi-source remote sensing data fusion, we benchmark it against several state-of-the-art deep learning models tailored for joint classification tasks across three representative datasets: 1) **Houston2013**: HyperMLP (Li et al. 2024) and Sal2RN (Li et al. 2023a) for LiDAR–HS fusion; MFT (Roy et al. 2023) and MSFM (Gao et al. 2025) for MS–HS fusion; 2) **MUUFL**: MSFE-IFN (Guo et al. 2024), AM3Net (Wang et al. 2022a), AMSSE (Gao et al. 2023), and M2FNet (Sun et al. 2024) for LiDAR–HS fusion; 3) **Augsburg**: S2ENet (Fang, Li, and Li 2022) and FDNet (Ni et al. 2024) for LiDAR–HS fusion; DSTD (Xu et al. 2023) and MACN (Li et al. 2023b) for HS–SAR fusion.

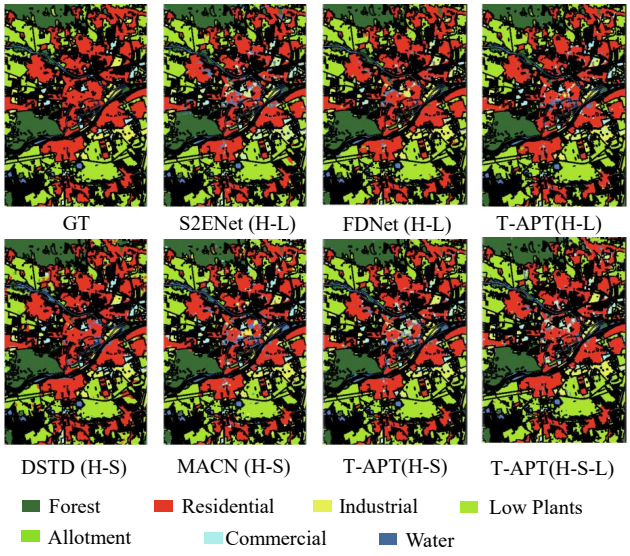

Figure 7: Classification maps of Augsburg dataset.

Method	Houston2013 H-M			Augsburg H-S		
	OA(%)	AA(%)	$\kappa\times100$	OA(%)	AA(%)	$\kappa\times100$
HyperMLP	96.69	97.11	96.42	81.53	77.93	75.08
AMSSE	97.25	97.72	97.03	56.10	47.54	45.41
M2FNet	95.83	96.56	95.50	87.32	86.67	82.61
S2ENet	90.33	91.66	89.54	87.70	81.83	82.98
T-APT	**98.60**	**98.88**	**98.85**	**91.80**	**90.63**	**88.59**

Table 2: Classification accuracy of different data combinations.

Performance Comparison

We conducted comprehensive quantitative and qualitative experiments with the competing methods on two datasets that combine two modalities. Furthermore, we performed classification experiments using three-modality combination. Table 1 presents the quantitative comparison between our proposed method and several state-of-the-art approaches across three benchmark datasets. The proposed T-APT consistently achieves superior performance in terms of OA, AA, and the Kappa coefficient under bimodal modality combinations on all three datasets.The classification outcomes illustrated in Figs. 5-7 highlight the method's superiority over alternatives. Qualitative comparisons reveal that our results align more closely with ground truth references. In Table 2, we analyzed the classification accuracy of different data combinations and compared the performance of specialized models on different datasets. The performance of certain proprietary methods has declined or even failed.

Ablation Method	Houston2013(MS+HS)			MUUFL(LiDAR+HS)			Augsburg(HS+SAR)		
	OA(%)	AA(%)	$\kappa\times100$	OA(%)	AA(%)	$\kappa\times100$	OA(%)	AA(%)	$\kappa\times100$
$\mathcal{A}_1$ Full tune baseline	98.25	98.59	98.12	90.74	92.05	87.91	**92.38**	87.84	**89.31**
$\mathcal{A}_{3-1}$ w/o $\mathcal{L}_{rec}$	98.19	98.51	98.04	89.21	91.87	86.08	90.07	87.25	87.54
$\mathcal{A}_{3-2}$ w/o $\mathcal{L}_{align}$	97.95	98.35	97.78	89.84	91.73	86.83	90.28	88.90	86.43
$\mathcal{A}_{3-3}$ w/o $\mathcal{L}_{rec}$ +$\mathcal{L}_{align}$	97.48	97.97	97.24	88.07	90.68	84.59	90.21	86.84	86.34
proposed method	**98.60**	**98.88**	**98.80**	**90.90**	**93.60**	**88.20**	91.80	**90.63**	88.59

Table 3: Effect of different fine-tuning components.

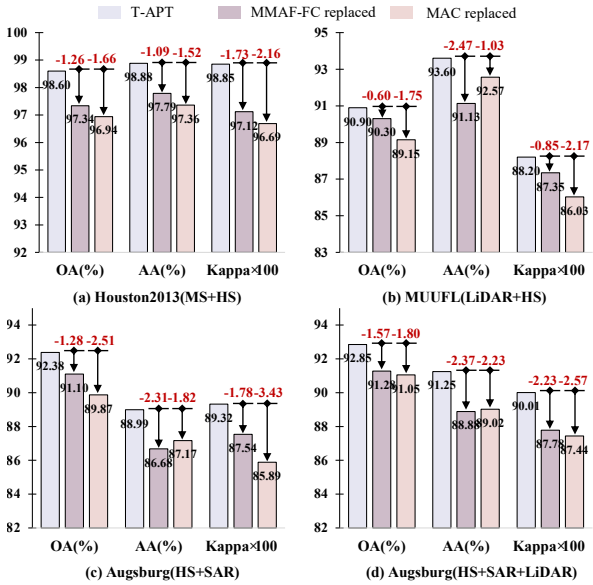

Figure 8: Effects of replacing function module components.

Dataset	Modality	Parms	OA(%)	AA(%)	$\kappa\times100$
Houston	HS-only	5.0M	96.21	96.81	95.90
	MS-only	4.9M	92.41	93.75	91.80
	LiDAR-only	4.9M	70.95	73.96	68.70
	H-L ‡	8.2M	98.36	98.57	98.23
	H-L	9.0M	98.89	99.09	98.80
	H-M-L ‡	11.8M	98.56	98.83	98.44
	H-M-L	13.0M	98.94	99.14	98.85
Augsburg	HS-only	5.0M	86.15	82.68	80.97
	SAR-only	4.9M	68.15	51.91	58.58
	LiDAR-only	4.9M	53.35	57.11	40.06
	H-S ‡	8.2M	91.99	87.87	88.74
	H-S	9.0M	91.80	90.63	88.59
	H-S-L ‡	11.8M	91.84	90.10	88.62
	H-S-L	12.9M	93.17	90.99	90.41

Table 4: Comparison with imbalanced multimodal learning function. The ‡ represents T-APT without text-guided prompt.

Ablation Study

Effectiveness of Mamba Structure in MAMF-FC To validate the effectiveness of the proposed specific-shared interleaved Mamba structure in cross-modal feature extraction, we investigated a variant of visual feature encoding-decoding that learns mapping relationships under the same optimization constraints. As shown in Figure 8 of comparative experiments, the OA decreased after replacing the Mamba structure in MAMF-FC. This demonstrates that the incorporation of proposed structure significantly enhances the ability to extract more discriminative supplementary information for cross-modal representation learning.

Effectiveness of Constraints on Prompt To validate the effectiveness of the proposed loss function, we experimented with two losses $\mathcal{L}_{rec}$, $\mathcal{L}_{align}$ for constrained cue information. After removing the modal reconstruction supervision and the text alignment loss, the results of the experiments as shown in Table 3 all decrease.

Effectiveness of TMPT 1)**MAC** To validate the effectiveness of the proposed prompt-based fine-tuning approach, we explored a variant of the fine-tuning strategy. This variant employs adapter-only to efficiently finetune the feed-forward layers of the foundation model. Experimental results shown in Figure 8, showed performance degradations of 1.32%, 1.43%, and 2.76%. Our proposed architecture demonstrates superior advantages in guiding the foundation model to adapt to arbitrary modality classification downstream tasks. 2)**Text-Guided Prompt** We evaluated the classification accuracy under a single-modal baseline, as shown in Table 4. It can be observed that each dataset contains a dominant mode. A performance deviation appeared after eliminating the text-guided mechanism. Further experiments will be conducted in subsequent studies to verify this issue.

Conclusion

In this work, we solve the problem of arbitrary-modality remote sensing image classification. We propose a unified model T-APT for arbitrary-modality classification, with Mamba-Based Arbitrary Modal-Focused Feature Capture (MAMF-FC) to fuse and disentangle complementary and modality-specific features and Text-Guided Modality-Aware Prompt Tuning (TMPT) to fine-tune pre-trained foundation model by introducing prompt information through a cross-modal approach. T-APT enables effective adaptation of the backbone model to downstream tasks and demonstrates a strong capability in mitigating modality inertia across arbitrary modality combinations. Experimental validation on three datasets demonstrates the effectiveness of T-APT.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant 62201423, 62471359, and 62476013, 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 20250133 and 20230117, in part by the Fundamental Research Funds for the Central Universities under Grant QTZX25084.

References

Ahmed, A.; Zeng, X.; Xi, R.; Hou, M.; and Shah, S. A. 2024. MED-Prompt: A novel prompt engineering framework for medicine prediction on free-text clinical notes. *J. King Saud Univ. Comput. Inf. Sci.*, 36(2).

Baumgartner, A.; Gege, P.; Köhler, C.; Lenhard, K.; and Schwarzmaier, T. 2012. Characterisation methods for the hyperspectral sensor HySpex at DLR's calibration home base. In *Sensors, Systems, and Next-Generation Satellites XVI*, volume 8533, 371–378. SPIE.

Contest, D. F. 2013. IEEE GRSS Data Fusion Contest Fusion of Hyperspectral and LiDAR Data.

Dong, Z.; Gu, Y.; and Liu, T. 2024. UPetu: A Unified Parameter-Efficient Fine-Tuning Framework for Remote Sensing Foundation Model. *IEEE Transactions on Geoscience and Remote Sensing*, 62: 1–13.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *ICLR*.

Fang, L.; Kuang, Y.; Liu, Q.; Yang, Y.; and Yue, J. 2023. Rethinking Remote Sensing Pretrained Model: Instance-Aware Visual Prompting for Remote Sensing Scene Classification. *IEEE Transactions on Geoscience and Remote Sensing*, 61: 1–13.

Fang, S.; Li, K.; and Li, Z. 2022. S²ENet: Spatial–Spectral Cross-Modal Enhancement Network for Classification of Hyperspectral and LiDAR Data. *IEEE Geoscience and Remote Sensing Letters*, 19: 1–5.

Gader, P.; Zare, A.; Close, R.; Aitken, J.; and Tuell, G. 2013. MUUFL Gulfport Hyperspectral and LiDAR Airborne Data Set. Technical Report Rep. REP-2013-570, University of Florida, Gainesville, FL.

Gao, F.; Jin, X.; Zhou, X.; Dong, J.; and Du, Q. 2025. MSFMamba: Multiscale Feature Fusion State Space Model for Multisource Remote Sensing Image Classification. *IEEE Transactions on Geoscience and Remote Sensing*, 63: 1–16.

Gao, H.; Feng, H.; Zhang, Y.; Xu, S.; and Zhang, B. 2023. AMSSE-Net: Adaptive Multiscale Spatial–Spectral Enhancement Network for Classification of Hyperspectral and LiDAR Data. *IEEE Transactions on Geoscience and Remote Sensing*, 61: 1–17.

Guo, F.; Meng, Q.; Li, Z.; Ren, G.; Wang, L.; Zhang, J.; Xin, R.; and Hu, Y. 2024. Multisource Feature Embedding and Interaction Fusion Network for Coastal Wetland Classification With Hyperspectral and LiDAR Data. *IEEE Transactions on Geoscience and Remote Sensing*, 62: 1–16.

Han, B.; Zhang, S.; Shi, X.; and Reichstein, M. 2024. Bridging Remote Sensors with Multisensor Geospatial Foundation Models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 27852–27862.

Hu, M.; Wu, C.; Du, B.; and Zhang, L. 2023. Binary Change Guided Hyperspectral Multiclass Change Detection. *IEEE Transactions on Image Processing*, 32: 791–806.

Jha, A.; Bose, S.; and Banerjee, B. 2023. GAF-Net: Improving the Performance of Remote Sensing Image Fusion using Novel Global Self and Cross Attention Learning. In *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 6343–6352.

Li, J.; Liu, Y.; Song, R.; Li, Y.; Han, K.; and Du, Q. 2023a. Sal²RN: A Spatial–Spectral Salient Reinforcement Network for Hyperspectral and LiDAR Data Fusion Classification. *IEEE Transactions on Geoscience and Remote Sensing*, 61: 1–14.

Li, J.; Liu, Y.; Song, R.; Liu, W.; Li, Y.; and Du, Q. 2024. HyperMLP: Superpixel Prior and Feature Aggregated Perceptron Networks for Hyperspectral and LiDAR Hybrid Classification. *IEEE Transactions on Geoscience and Remote Sensing*, 62: 1–14.

Li, K.; Wang, D.; Wang, X.; Liu, G.; Wu, Z.; and Wang, Q. 2023b. Mixing Self-Attention and Convolution: A Unified Framework for Multisource Remote Sensing Data Classification. *IEEE Transactions on Geoscience and Remote Sensing*, 61: 1–16.

Liu, C.; Zhao, R.; Chen, J.; Qi, Z.; Zou, Z.; and Shi, Z. 2023. A Decoupling Paradigm With Prompt Learning for Remote Sensing Image Change Captioning. *IEEE Transactions on Geoscience and Remote Sensing*, 61: 1–18.

Liu, Y.; Tian, Y.; Zhao, Y.; Yu, H.; Xie, L.; Wang, Y.; Ye, Q.; and Liu, Y. 2024. VMamba: Visual State Space Model. *arXiv preprint arXiv:2401.10166*.

Ni, K.; Wang, D.; Zhao, G.; Zheng, Z.; and Wang, P. 2024. Hyperspectral and LiDAR Classification via Frequency Domain-Based Network. *IEEE Transactions on Geoscience and Remote Sensing*, 62: 1–17.

Roy, S. K.; Deria, A.; Hong, D.; Rasti, B.; Plaza, A.; and Chanussot, J. 2023. Multimodal Fusion Transformer for Remote Sensing Image Classification. *IEEE Transactions on Geoscience and Remote Sensing*, 61: 1–20.

Sun, L.; Wang, X.; Zheng, Y.; Wu, Z.; and Fu, L. 2024. Multiscale 3-D–2-D Mixed CNN and Lightweight Attention-Free Transformer for Hyperspectral and LiDAR Classification. *IEEE Transactions on Geoscience and Remote Sensing*, 62: 1–16.

Tao, W.; Lei, B.; Liu, K.; Lu, S.; Cui, M.; and Xie, X. 2025. DivAvatar: Diverse 3D Avatar Generation with a Single Prompt. In *Proceedings of the Winter Conference on Applications of Computer Vision (WACV)*, 2568–2577.

Wang, J.; Li, J.; Shi, Y.; Lai, J.; and Tan, X. 2022a. AM3Net: Adaptive Mutual-learning-based Multimodal Data Fusion

Network. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(8): 5411–5426.

Wang, Z.; Yu, X.; Rao, Y.; Zhou, J.; and Lu, J. 2022b. P2P: Tuning Pre-trained Image Models for Point Cloud Analysis with Point-to-Pixel Prompting. *arXiv preprint arXiv:2208.02812*.

Wu, J.; and Xu, M. 2024. One-Prompt to Segment All Medical Images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 11302–11312.

Xu, L.; Zhu, H.; Jiao, L.; Zhao, W.; Li, X.; Hou, B.; Ren, Z.; and Ma, W. 2023. A Dual-Stream Transformer With Diff-Attention for Multispectral and Panchromatic Classification. *IEEE Transactions on Geoscience and Remote Sensing*, 61: 1–14.

Yang, Y.; Qu, J.; Dong, W.; Zhang, T.; Xiao, S.; and Li, Y. 2024. TMCFN: Text-Supervised Multidimensional Contrastive Fusion Network for Hyperspectral and LiDAR Classification. *IEEE Transactions on Geoscience and Remote Sensing*, 62: 1–15.

Zha, Y.; Wang, J.; Dai, T.; Chen, B.; Wang, Z.; and Xia, S.-T. 2023. Instance-aware Dynamic Prompt Tuning for Pre-trained Point Cloud Models. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 14115–14124.

Zhao, X.; Zhang, M.; Tao, R.; Li, W.; Liao, W.; Tian, L.; and Philips, W. 2024. Fractional Fourier Image Transformer for Multimodal Remote Sensing Data Classification. *IEEE Transactions on Neural Networks and Learning Systems*, 35(2): 2314–2326.