Frequency-Aware Vision-Language Multimodality Generalization Network for Remote Sensing Image Classification

Junjie Zhang[1], **Feng Zhao**[1*], **Hanqiang Liu**[2], **Jun Yu**[3]

[1]School of Communications and Information Engineering, Xi'an University of Posts and Telecommunications
[2]School of Artificial Intelligence and Computer Science, Shaanxi Normal University
[3]Department of Automation, University of Science and Technology of China
junjiezhang@stu.xupt.edu.cn, zhaofeng201@xupt.edu.cn, liuhq@snnu.edu.cn, harryjun@ustc.edu.cn

Abstract

The booming remote sensing (RS) technology is giving rise to a novel multimodality generalization task, which requires the model to overcome data heterogeneity while possessing powerful cross-scene generalization ability. Moreover, most vision-language models usually describe surface materials using universal texts, lacking proprietary linguistic prior knowledge specific to different RS modalities. In this work, we formalize RS multimodality generalization (RSMG) as a learning paradigm, and propose a frequency-aware vision-language multimodality generalization network (FVMGN) for RS image classification. Specifically, a diffusion-based training-test-time augmentation (DTAug) strategy is designed to reconstruct multimodal land-cover distributions, enriching input information for FVMGN. Following that, to overcome multimodal heterogeneity, a multimodal wavelet disentanglement (MWDis) module is developed to learn cross-domain invariant features by resampling low and high frequency components in the frequency domain. Considering the characteristics of RS vision modalities, shared and proprietary class texts is designed as linguistic inputs for the transformer-based text encoder to extract diverse text features. For multimodal vision inputs, a spatial-frequency-aware image encoder (SFIE) is constructed to realize local-global feature reconstruction and representation. Finally, a multiscale spatial-frequency feature alignment (MSFFA) module is suggested to construct a unified semantic space, ensuring refined multiscale alignment of different text and vision features in spatial and frequency domains. Extensive experiments show that FVMGN has the excellent multimodality generalization ability compared with state-of-the-art methods.

Code — https://github.com/ZJier/FVMGN

Introduction

Deep learning, as a catalyst for the rapid development of remote sensing image classification (RSIC), has elevated the efficiency and accuracy to a new degree for RSIC task, promoting innovative research and practical applications within the RS community (Qu et al. 2024; Kuckreja et al. 2024; Guo et al. 2024b). Recently, with the development of sensor and satellite technologies, types of RS images are becoming increasingly diverse, such as hyperspectral (HS) image, light detection and ranging (LiDAR) image, and synthetic aperture radar (SAR) image, which have distinct data characteristics (Li et al. 2022). For different RS data, numerous advanced **single-modal RSIC** methods have emerged (Wang et al. 2018; Imani 2025; Zhang et al. 2025).

Real-world tasks usually require models to adapt to unseen RS scenes, while single-modal RSIC methods face limitations in domain generalization (DG) (Zhang et al. 2023a; Wang et al. 2024a). Based on this, **single-modal RSDG** methods have sprouted up everywhere. It refers to training a model using RS data from a scene, hoping that the model achieves good generalization performance on new and unseen scenes. This task requires the model to learn robust and domain-invariant features that can adapt to data distribution characteristics from the target domain (Chu et al. 2024; Zhao et al. 2023b; Qin et al. 2024). In addition, modality types are becoming increasingly diverse, which reflects geomorphic structures comprehensively while posing challenges to single-modal RSIC methods. Given this, **multimodal RSIC** methods have developed rapidly. It refers to performing classification tasks using RS data obtained from different sensors, namely various modality data, such as HS and LiDAR data (Guo et al. 2024a; Zhang et al. 2024e). This task requires that the model learn the complementarity and similarity between different modalities, realizing multimodal interaction while overcoming the data heterogeneity (Tang et al. 2024; Jia et al. 2020).

However, multimodal RSIC-based methods are usually required to have the capability to effectively adapt to unseen new scenes, where multimodal heterogeneity and distribution differences are urgent problems that need to be addressed in practical applications. Therefore, a novel and interesting task, **RS multimodality generalization (RSMG)** for image classification, comes into being, which requires the algorithm to consider multimodality heterogeneity and generalization simultaneously. RSMG requires that a well-trained model on the source domain can show excellent generalization ability on the target domain under the premise where multimodal training data and test data are mutually unseen. RSMG aims to more effectively reflect the geographical environment and urbanization process in different regions, which can provide crucial support for multi-perspective and cross-regional disaster monitoring, resource management and land planning.

*Corresponding author.

In this paper, we propose a frequency-aware vision-language multimodality generalization network (FVMGN) for RSIC, which can achieve the joint representation for land-cover distribution in spatial and frequency domains. In the pre-processing phase, we design a diffusion-based training-test-time augmentation (DTAug) strategy to achieve multimodal land-cover distribution reconstructions, enhancing the input diversity. In the feature interaction phase, we develop a multimodal wavelet disentanglement (MWDis) module, which can learn cross-domain invariant features while achieving multimodal feature interaction. In the feature extraction phase, we design a spatial-frequency-aware image encoder (SFIE) to achieve spatial and multi-frequency analysis by integrating wavelet transform into CNN and ViT. In the feature alignment phase, we develop a multiscale spatial-frequency feature alignment (MSFFA) module to realize multi-level supervision between text and vision features, thereby further improving multimodality generalization ability. Contributions of this work are as follows.

- We formalize RSMG as a learning paradigm, and propose a FVMGN for RSIC.
- Shared and proprietary texts are designed to reflect general and modality-specific geospatial attributes.
- MWDis module is designed to realize the multimodal information interaction, helping the network learn domain-invariant generalization features.
- SFIE is developed by integrating wavelet transform, convolution, and self-attention to realize local-global spatial-frequency feature extraction.
- MSFFA module is developed to achieve fine multiscale vision-vision and vision-text feature alignments in spatial and frequency domains.

Related Works

Vision-Language Model (VLM)

VLM aims to learn the complex relationships between image and text, enabling more accurate decision-making and reasoning (Radford et al. 2021; Zhang et al. 2024b). Currently, diverse VLM-based methods have emerged in various fields, such as visual question answering (Guo et al. 2023a), image captioning (Rotstein et al. 2024), text-to-image retrieval (Zhang et al. 2024a), and more. In view of this, researchers have explored how VLMs perform on the RSIC task (Wang et al. 2024b). Zhang et al. (2023b) described coarse-grained and fine-grained texts as linguistic prior knowledge based on class names for HS image classification, serving as effective text-based soft connections for different targets. Yang et al. (2024) refined the class name descriptions based on intrinsic attributes, and utilized the attention mechanism to fuse text and vision features, effectively achieving joint HS and LiDAR data classification.

Denoising Diffusion Probabilistic Model (DDPM)

DDPM is a generative model that reconstructs data distribution by adding noise in the forward process and removing noise in the reverse process (Ho, Jain, and Abbeel 2020). DDPM has been widely applied in various fields, such as image generation (Zhao et al. 2024), image super-resolution (Yuan and Yuan 2024), and image restoration (Zhu et al. 2023). In recent years, DDPM with powerful data generation capability has shone in RSIC field (Chen et al. 2023; Zhang et al. 2024d). Chen et al. (2023) reconstructed the spectral-spatial land-cover distribution through the noise addition and denoising processes in DDPM, confirming its effectiveness in HS image classification tasks. Zhang et al. (2024d) considered the land-cover distribution reconstructed by DDPM as an unsupervised generative knowledge, which is combined with HS data to achieve competitive classification results.

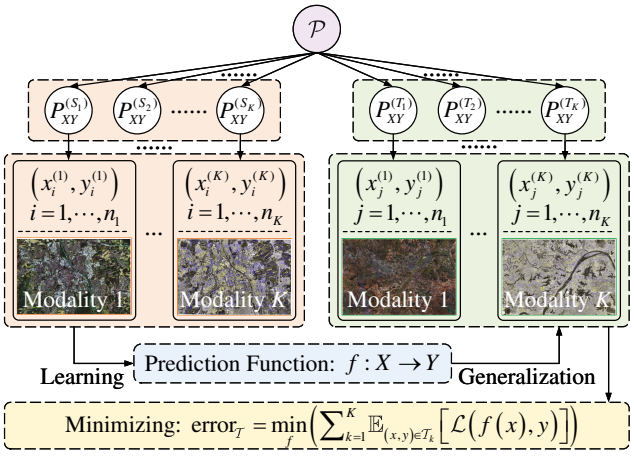

Figure 1: Problem definition for RSMG.

Wavelet Transform (WT)

Recently, the synthesis of WT and neural network has made a remarkable success in the computer vision field (Korkmaz, Tekalp, and Dogan 2024). WT can decompose image data into multiple frequency signals, realizing multi-frequency analysis and feature reconstruction for image information (Seydi, Bozorgasl, and Chen 2024; Gao et al. 2024). Finder et al. (2025) proposed wavelet convolution by combining WT and convolution operations, which proved to have a larger receptive field than vanilla convolution. Inspired by the above works, researchers have begun to explore the application of WT on RSIC tasks. Ahmad et al. (2024) applied WT for reversible down-sampling to achieve lossless compression for feature maps while enhancing structure and shape information in spatial and channel dimensions. Seydi, Bozorgasl, and Chen (2024) treated the wavelet function as a learnable activation function to achieve nonlinear mapping, which allows the Kolmogorov-Arnold network to mine multiscale spatial-spectral patterns.

Methodology

Problem Definition

As shown in Fig. 1. Let $\mathcal{X}$ be the input space, $\mathcal{Y}$ be the output (label) space, and domain is represented by the joint distribution P_{XY} over the input samples X and labels Y on $\mathcal{X} \times \mathcal{Y}$. Given K RS modalities $\mathcal{S} = \{S_k = \{(x_{n_k}^{(k)}, y_{n_k}^{(k)})\}\}_{k=1}^{K} \sim$

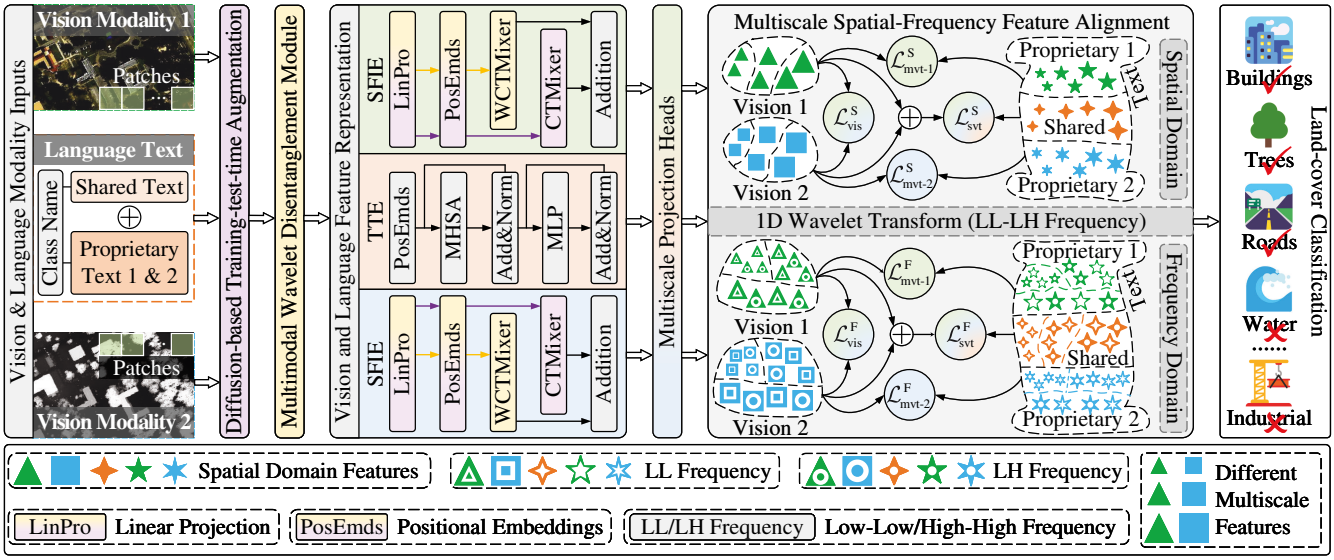

Figure 2: Overview of FVMGN. It consists of five components: DTAug, MWDis, SFIE, TTE, and MSFFA.

$P_{XY}^{(S_k)}$, where n_k denotes the number of samples in the k-th modality S_k, and $P_{XY}^{(S_k)} \neq P_{XY}^{(S_{k'})}, k \neq k'; k, k' \in \{1, \ldots, K\}$. RSMG aims to learn a prediction model $f : \mathcal{X} \rightarrow \mathcal{Y}$ using only the multimodal data from the source domain, such that the model f minimizes the prediction error $\text{error}_{\mathcal{T}}$ on K unseen target modalities $\mathcal{T} = \{T_k = \{(x_{m_k}^{(k)}, y_{m_k}^{(k)})\}\}_{k=1}^{K} \sim P_{XY}^{(T_k)}, \forall k \in \{1, \ldots, K\}$. Here, m_k denotes the number of samples in the k-th target modality T_k, and $P_{XY}^{(T_k)} \neq P_{XY}^{(T_{k'})} \neq P_{XY}^{(S_k)} \neq P_{XY}^{(S_{k'})}, k \neq k'; k, k' \in \{1, \ldots, K\}$. This means that prediction model f trained on the source domain can generalize well even without access to the target domain during the training phase. The $\text{error}_{\mathcal{T}}$ is defined as:

$$\text{error}_{\mathcal{T}} = \min_{f} \left(\sum_{k=1}^{K} \mathbb{E}_{(x,y)\in T_k} \left[\mathcal{L}\left(f\left(x \right), y \right) \right] \right), \quad (1)$$

where $\mathbb{E}$ denotes the expectation, and $\mathcal{L}$ is the loss function.

Overview

As shown in Fig. 2, FVMGN is composed of the DTAug, MWDis, SFIE, transformer-based text encoder (TTE), and MSFFA. These sub-parts form a powerful whole, each with different tasks. 1) DTAug utilizes the DDPM to reconstruct multimodal land-cover distributions, enriching model inputs. 2) MWDis performs feature disentanglement in the frequency domain, learning cross-domain-invariant feature representation. 3) SFIE organically integrates wavelet transform (WT), convolution operations, and attention mechanism, which can realize multi-frequency analysis in the frequency domain while extracting local-global features in the spatial domain. 4) MSFFA performs multiscale feature alignments in wavelet and spatial domains, thereby achieving more refined positive sample pair matching. Finally, the classification result can be obtained by the maximum score strategy with two linear classifiers.

DTAug

Motivated by DDPM (Ho, Jain, and Abbeel 2020), we explore its effectiveness in data augmentation, and design a novel multimodal RS data augmentation strategy, called DTAug, which enriches the input of FVMGN by reconstructing multimodal land-cover distribution, enhancing the cross-scene generalization ability (refer to *Supp. Fig.* 1). Firstly, we feed the RS data from the two modalities into the DDPM based on 3D UNet (Zhang et al. 2024d) to generate the reconstructed and unsupervised land-cover distributions. Secondly, we perform principal component analysis (PCA: 30) on the original and diffusion-enhanced RS data, and concatenate the reduced data on the channel dimension, thereby obtaining obtain new multimodal input data. In addition, the obtained RS data is cropped into patches with the same size, and geometry augmentation based on patches is performed to further increase the diversity, such as flip augmentation and radiation augmentation. Notably, the entire land-cover distribution generation process is unsupervised in DTAug, and the multimodal RS data from source and target domains are mutually unseen in training and test phase.

MWDis

For various types of RS data, the interaction and fusion process between modalities are key to achieve accurate land-cover classification. As shown in Fig. 3, we utilize wavelet decomposition to achieve image disentanglement for multimodal data, which allows for learning cross-domain invariant features while overcoming multimodality heterogeneity. Taking HS and LiDAR modalities as examples, they are decomposed into a low-frequency component $\mathbf{F}^{ll}$ (low-low (LL) frequency) and a set of high-frequency components $\mathbf{F}^{h}$ (high-low/low-high/high-high (HL/LH/HH) frequencies). LL frequency component usually contains higher energy, and resampling it can change the source sample style (Guo et al. 2023b). Therefore, to enhance sample diver-

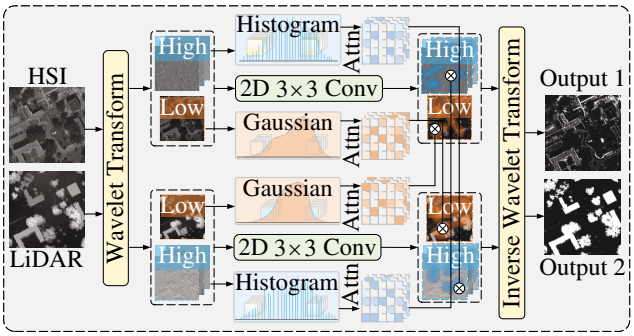

Figure 3: Structure of MWDis module.

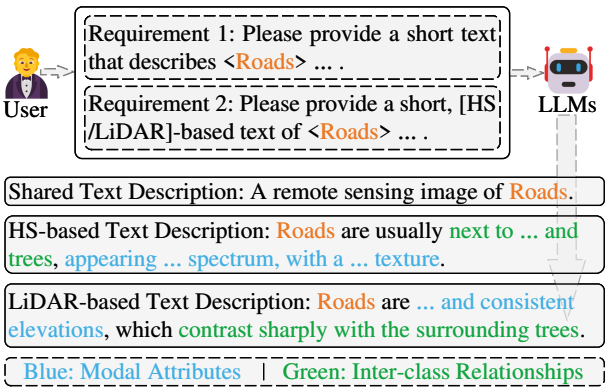

Figure 4: Different types of class text descriptions (Class: [*Roads*] as an example).

sity, we model LL frequency component as the multivariate Gaussian distribution, as follows:

$$\begin{aligned} \hat{\mathbf{F}}^{ll} &= \mathbf{F}^{ll} + \rho\sigma\left[\mathbf{F}^{ll}\right], \\ \sigma^2\left[\mathbf{F}^{ll}\right] &= \frac{1}{N}\sum_{n=1}^{N}\left[\mathbf{F}_n^{ll} - \mu\left[\mathbf{F}_n^{ll}\right]\right]^2, \end{aligned} \tag{2}$$

where $\{\rho \backsim \mathcal{N}(0,\alpha), \alpha \in [0,1]\}$ denotes the strength of domain shifts, $\hat{\mathbf{F}}^{ll}$ is the resampled LL frequency feature, and N is batch size. $\mu[\cdot]$ and $\sigma[\cdot]$ represent the mean and standard deviation, respectively. In addition, the resampled LL frequency features $\hat{\mathbf{F}}^{ll}_{\mathrm{v1}}$ and $\hat{\mathbf{F}}^{ll}_{\mathrm{v2}}$ are calculated as spatial attention (SpatAttn) (Woo et al. 2018) matrix, which is acted on the convolution (Conv) features of another modality. The subscripts v1 and v2 represent two vision modalities, respectively. This process can be expressed as follows:

$$\mathbf{F}^{ll}_{\mathrm{I1}} = \mathrm{SpatAttn}(\hat{\mathbf{F}}^{ll}_{\mathrm{v2}}) \otimes \mathrm{Conv}\left(\mathbf{F}^{ll}_{\mathrm{v1}}\right), \tag{3}$$

where $\mathbf{F}^{ll}_{\mathrm{I1}}$ is the weighted low-frequency feature of vision modality 1, and $\otimes$ denotes element-wise multiplication operation. Similarly, we can obtain the weighted LL frequency feature $\mathbf{F}^{ll}_{\mathrm{I2}}$ of vision modality 2.

HL/LH/HH frequencies are usually sharp semantics, with the most noticeable intensity changes occurring at boundaries between adjacent ground objects (Li, Gao, and He 2023). Given this, we perform gradient-map-based resampling on high frequencies, which can capture relative intensity variations as the modality generalization representation. Gradient-map resampling $G(\cdot)$ is formulated by:

$$G\left(\mathbf{F}^{h}\right) = HE\left(\sqrt{\left(\mathbf{F}^{hl}\right)^2 + \left(\mathbf{F}^{lh}\right)^2 + \left(\mathbf{F}^{hh}\right)^2}\right), \tag{4}$$

where $HE(\cdot)$ represents the histogram equalization. Similarly, we can obtain the weighted features of HL/LH/HH frequencies through Eq.(3). After the frequency interaction mentioned above, we perform an inverse wavelet transform on the resampling frequency components to achieve the effective feature reconstruction.

Vision & Language Feature Representation

The vision and language feature representation phase contains a TTE and two SFIEs, as shown in Fig. 2.

TTE As shown in Fig. 4, considering the uniqueness and commonality of multimodal RS data, we design the shared and modality-specific (proprietary) text descriptions for class names by the large language model and manual fine-tuning. The shared text provides general linguistic information. Proprietary texts provide linguistic prior knowledge for vision feature representation in modal attributes and inter-class relationships, such as Class [*Roads*]: 1) HS modal attributes [*Roads appear gray and brown in the visible spectrum, with a uniform and smooth texture*]. 2) Inter-class relationships [*Roads are usually next to sidewalks, buildings, and trees*]. As shown in Fig. 2, TTE is a language model-based transformer without pre-training (Vaswani et al. 2017; Radford et al. 2019). Similar to CLIP, TTE utilizes a lower-cased byte pair encoding (BPE) for feature representation of the different text descriptions, where the vocabulary size and maximum sequence length are 49,152 and 76 (Sennrich, Haddow, and Birch 2015), respectively.

SFIE To achieve full-domain spatial-frequency feature extraction, we design the SFIE with a fully residual grouping convolution module (FRGCM) and a spatial-frequency dual-branch backbone. For FRGCM, we apply 3×3 Conv layers between two 1×1 Conv (channel transformation) layers to facilitate inter-group feature interaction and fusion, enhancing feature diversity (refer to *Supp. Fig.* 2).

Spatial-frequency dual-branch backbone mainly consists of a wavelet convolution transformer mixer (WCTMixer) and a convolution transformer mixer (CTMixer), where CTMixer is applied to extract local-global spatial feature according to (Zhao et al. 2023a), as shown in Fig. 2. For WCTMixer, we design a multi-head wavelet self-attention (MHWSA) mechanism using WT. Specifically, the input feature is decomposed into different frequency components using WT, which are considered as different parallel heads according to their properties for self-attention. Here, we take LL frequency as an example to illustrate the above process:

$$\begin{aligned} \mathbf{Z}^{ll}_{\mathrm{out}} &= \mathrm{CPool}\left(\mathrm{WSA}\left(\mathbf{Q}^{ll}, \mathbf{K}^{ll}, \mathbf{V}^{ll}, \mathbf{C}^{ll}\right)\right), \\ \mathbf{Q}^{ll}, \mathbf{K}^{ll}, \mathbf{V}^{ll}, \mathbf{C}^{ll} &= \mathrm{Re}\left(\mathrm{Chunk}\left(\mathrm{Conv}\left(\mathbf{Z}^{ll}\right)\right)\right), \\ \mathbf{Z}^{ll}, \mathbf{Z}^{hl}, \mathbf{Z}^{lh}, \mathbf{Z}^{hh} &= \mathrm{WT}\left(\mathbf{Z}_{\mathrm{in}}\right), \end{aligned} \tag{5}$$

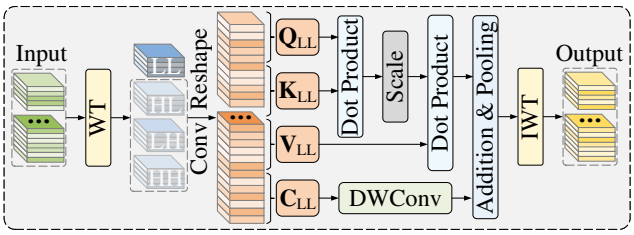

Figure 5: Structure of MHWSA mechanism.

where $\mathbf{Z}_{\mathrm{in}}$ represents the input of the MHWSA mechanism. $\mathbf{Z}^{ll}$, $\mathbf{Z}^{hl}$, $\mathbf{Z}^{lh}$ and $\mathbf{Z}^{hh}$ are different frequency components. $\mathrm{Re}(\cdot)$, $\mathrm{Chunk}(\cdot)$, $\mathrm{CPool}(\cdot)$ and $\mathrm{WSA}(\cdot)$ represent reshape, chunking, channel pooling, and wavelet self-attention. $\mathbf{Q}^{ll}$, $\mathbf{K}^{ll}$, $\mathbf{V}^{ll}$, and $\mathbf{C}^{ll}$ are intermediate features. $\mathbf{Z}^{ll}_{\mathrm{out}}$ is the output feature of $\mathrm{CPool}(\cdot)$. Similarly, we apply Eq.(5) to obtain the output features ($\mathbf{Z}^{hl}_{\mathrm{out}}$, $\mathbf{Z}^{lh}_{\mathrm{out}}$, $\mathbf{Z}^{hh}_{\mathrm{out}}$) for the other frequencies. Finally, output feature $\mathbf{Z}_{\mathrm{out}}$ of MHWSA is acquired through the inverse WT (IWT), as follows:

$$\mathbf{Z}_{\mathrm{out}} = \mathrm{IWT}\left(\mathbf{Z}^{ll}_{\mathrm{out}}, \mathbf{Z}^{hl}_{\mathrm{out}}, \mathbf{Z}^{lh}_{\mathrm{out}}, \mathbf{Z}^{hh}_{\mathrm{out}}\right). \tag{6}$$

In summary, SFIE utilizes the FRGCM to mine grouping-interactive feature information, and then applies WCTMixer and CTMixer to capture discriminative local-global vision features from the spatial and frequency domains, enhancing the model representation ability.

MSFFA

In VLMs, the contrastive learning-based vision-text feature alignment allows matching image-text pairs to be closer in the same embedding space, thereby understanding the relationships between image and text (Radford et al. 2021).

As shown in Fig. 6, compared with traditional RS vision-text feature alignment methods (Zhang et al. 2023b; Wang et al. 2024b; Yang et al. 2024), we consider the following three aspects in MSFFA. **1) Multiscale characteristics**: MSFFA aligns the multiscale vision features and text features obtained from the multiscale projection heads, which enables a more refined understanding for vision-text pairs, enhancing overall generalization on RSMG tasks. **2) Joint spatial-frequency domain**: MSFFA performs the multiscale feature alignment in the spatial domain while utilizing WT to decouple spatial features into the LL and LH frequency components, achieving multiscale and multi-frequency feature alignment in the frequency domain. **3) Multimodal vision-vision feature alignment**: Most VLMs focus on the feature alignment for vision-text pairs, while neglecting the complementarity between vision features, especially for RSMG tasks. In view of this, MSFFA applies cosine similarity to design vision-vision feature alignment losses, enhancing multimodal representation capability. In addition, vision-text feature alignment losses refer to the contrastive learning (Radford et al. 2021) in implementation, and the final loss contains multiscale spatial-frequency vision-vision and vision-text alignment losses, such as $\mathcal{L}^{\mathrm{S}}_{\mathrm{vis}}$, $\mathcal{L}^{\mathrm{S}}_{\mathrm{svt}}$, $\mathcal{L}^{\mathrm{F}}_{\mathrm{mvt\text{-}1}}$, and $\mathcal{L}^{\mathrm{F}}_{\mathrm{mvt\text{-}1}}$ in Fig. 2.

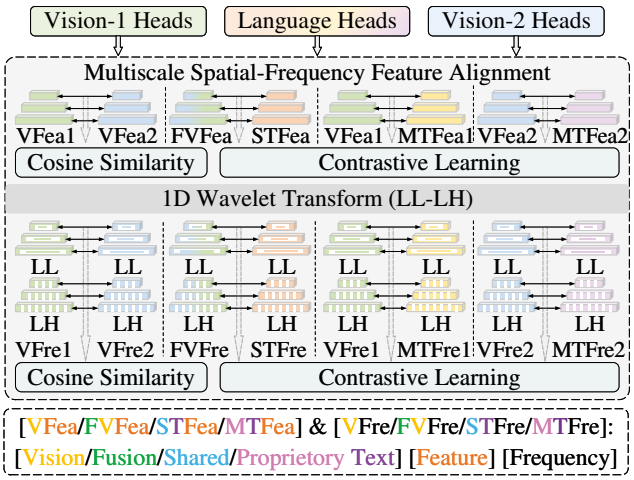

Figure 6: Structure of MSFFA module.

Experiments

Experimental Setup

Datasets We select three public multimodal datasets for the experiments, namely the HS-LiDAR MUUFL (Gader et al. 2013), Trento (Hong et al. 2020), and Houston 2013 (HU2013) (Debes et al. 2014) datasets. For the above three datasets, we select three common classes for generalization analysis, namely, [*Trees*], [*Roads*], and [*Buildings*], which show significant differences in land-cover distribution across different datasets (details are provided in *Supp. Table* 1 & *Fig*. 3). In addition, to accommodate various types of real-world tasks, we perform mutual RSMG among three datasets, meaning each dataset can serve as the source domain (SD) and the target domain (TD), namely SD↔TD: MUUFL↔Trento↔HU2013 (MU↔TR↔HU).

Implementation Details We select several types of SOTA methods for comparison, including 1) multimodality-based methods: MFT (Roy et al. 2023), MsFE-IFN (MsFE) (Guo et al. 2024a), and CMFAEN (CFEN) (Zhang et al. 2024e); 2) DDPM-based method: DKDMN (DKDN) (Zhang et al. 2024d); 3) DG-based methods: SDENet (SDEN) (Zhang et al. 2023a), LLURNet (LLURN) (Zhao et al. 2023b), FDGNet (FDGN) (Qin et al. 2024), TFTNet (TFTN) (Zhang et al. 2024c), ADNet (Zhao et al. 2025), and ISDGS (Gao et al. 2025); 4) VLM & DG-based methods: EHSNet (EHSN) (Wang et al. 2024b), and LDGNet (LDGN) (Zhang et al. 2023b). Secondly, the learning rate, batch size, patch size, and epoch of FVMGN are set to 0.001, 128, 11, and 20. Adam optimizer and cosine annealing scheduler are used for parameter optimization and learning rate adjustment. In addition, we select 10% samples in the source domain for training and generalize the well-trained model to all test samples in the target domain, and experiments are conducted on the PyTorch framework. *For the reproducibility, we will release the adjusted datasets and source code soon.*

Evaluation Metrics We apply overall accuracy (OA), average accuracy (AA), and kappa coefficient (Kappa) to as-

SD→TD	Metric	MFT	MsFE	CFEN	DKDN	SDEN	LLURN	FDGN	TFTN	ADNet	ISDGS	EHSN	LDGN	FVMGN
MU→TR	OA	77.59	83.98	82.82	78.95	78.14	78.23	80.11	76.32	80.36	80.27	77.58	83.17	**92.44**
	AA	61.46	73.60	70.75	64.48	63.75	63.17	65.62	60.44	66.03	66.51	63.09	72.36	**88.58**
	Kappa	53.54	69.25	66.81	58.08	60.22	60.49	64.14	56.15	64.67	64.59	58.49	69.65	**86.59**
MU→HU	OA	27.64	16.93	23.17	13.38	61.60	60.11	56.20	67.96	70.39	65.24	41.06	49.68	**93.19**
	AA	27.63	15.28	19.51	13.43	65.58	64.93	61.41	70.98	75.13	71.12	41.85	51.09	**93.28**
	Kappa	9.930	26.73	21.04	30.22	43.83	42.16	36.81	52.73	56.93	49.86	14.16	24.76	**89.64**
HU→TR	OA	23.32	20.66	26.75	29.05	69.11	66.68	57.80	66.66	73.90	51.12	71.01	74.88	**82.87**
	AA	38.00	35.13	45.34	48.15	60.96	59.19	55.28	60.43	64.02	51.45	60.24	63.01	**78.29**
	Kappa	1.530	4.120	6.300	10.10	49.01	45.42	36.06	45.81	55.49	28.18	47.78	55.55	**69.93**
HU→MU	OA	17.82	48.38	18.60	46.31	62.25	48.47	22.26	40.65	51.41	41.39	64.73	80.41	**82.33**
	AA	31.95	28.21	33.29	47.07	44.36	42.66	35.58	43.11	42.03	41.90	37.67	74.38	**76.85**
	Kappa	4.130	11.79	0.070	10.37	24.74	17.80	3.610	14.48	16.26	14.53	5.610	64.14	**66.44**
TR→HU	OA	28.94	11.11	23.36	25.06	53.84	58.51	56.17	77.89	69.53	72.75	43.76	67.74	**89.46**
	AA	27.76	10.51	19.41	27.12	56.80	61.11	59.27	80.21	71.84	75.53	44.25	68.30	**90.47**
	Kappa	6.450	32.19	20.27	9.390	31.64	38.26	35.32	67.10	54.61	59.62	16.59	51.32	**84.12**
TR→MU	OA	66.50	68.66	62.09	49.60	67.72	70.98	66.64	63.20	68.77	67.11	73.99	74.42	**90.01**
	AA	45.65	58.56	45.77	49.06	44.19	52.54	39.48	46.64	44.67	45.32	60.67	62.45	**86.94**
	Kappa	22.86	43.64	24.63	24.98	23.66	37.12	14.62	23.65	22.55	24.08	44.99	53.05	**81.27**

Table 1: Classification results (%) of different methods on several multimodality generalization combinations.

sess model performance. To ensure fairness, we calculated the mean values of OA, AA, and Kappa from 10 independent experiments for comparative analysis.

Comparison with SOTA Methods

Quantitative Analysis As shown in Table 1, comparative experiments are conducted on six dataset combinations (SD↔TD: MU↔TR↔HU). Overall, FVMGN achieves 92.44%, 93.19%, 82.87%, 82.33%, 89.46%, and 90.01% in OAs, respectively, demonstrating competitive classification performance compared with SOTA methods. It mainly benefits from two aspects: 1) FVMGN can focuses on multimodal heterogeneity and cross-scene generalization compared with DDPM-based (DKDN), multimodality-based (SDEN, LLURN, FDGN, TFTN, ADNet, and ISDGS) and DG-based (EHSN and LDGN) methods, achieving the effective multimodal cross-domain invariant feature extraction. 2) FVMGN designs proprietary class texts to characterize modal attributes while constructing a multiscale unified semantic space for spatial and frequency feature alignments compared with VLM & DG-based methods.

Qualitative Analysis Taking the TR→MU combination as an example, classification maps of different methods are shown in Fig. 7. According to distribution difference descriptions for different datasets in *Appl.Datasets*, it can be known that the source domain (Trento dataset) distribution is sparse, while the target domain (MUUFL dataset) distribution is dense, which poses certain challenges for RSMG. For example, CFEN fails to recognize roads as a separate class, and DKDN recognizes a few building regions with a relatively low accuracy. Nevertheless, FVMGN still has a high classification accuracy on different class regions, and its classification map contains less noise and contaminated regions with a better visual effect.

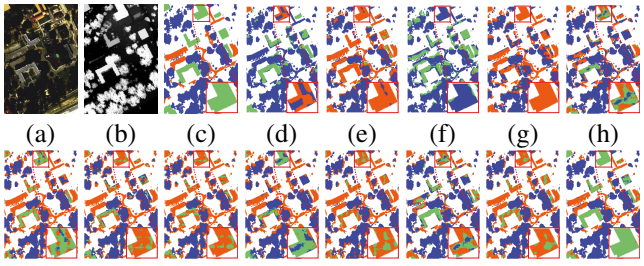

Figure 7: Classification maps (Blue/orange/green regions: Trees/Roads/Buildings) on the TR→MU dataset combination. (a) HSI. (b) LiDAR image. (c) Ground truth. (d) MFT. (e) MsFE. (f) CFEN. (g) DKDN. (h) SDEN. (i) LLURN. (j) FDGN. (k) TFTN. (l) ADNet. (m) ISDGS. (n) EHSN. (o) LDGN. (p) FVMGN.

Ablation Study

Contribution of Different Components We conduct performance evaluation on different key components, as shown in Table 2. Taking the MU→TR as an example, OA of **NET**-2 is 6.65% higher than that of **NET**-1, which indicates that DTAug is conducive to enhancing the input diversity. Secondly, **NET**-3 achieves a 2.85% OA improvement over **NET**-2. This is mainly attributed to the fact that MWDis enhances the ability to extract cross-domain invariant features. Furthermore, we apply the SFIE for **NET**-4, achieving a 2.14% OA improvement. This is mainly because SFIE can capture local-global features and realize the wavelet reconstruction effectively. Finally, **NET**-5 (FVMGN) has a 1.75% OA improvement by the MSFFA, which mainly benefits from that MSFFA can achieve more accurate matching between positive sample pairs.

SD→TD	**NET**-1	**NET**-2	**NET**-3	**NET**-4	**NET**-5
MU→TR	79.05	85.70	88.55	90.69	**92.44**
MU→HU	65.00	82.90	91.54	92.40	**93.19**
HU→TR	47.82	73.92	79.20	80.33	**82.87**
HU→MU	71.88	73.95	79.20	79.78	**82.33**
TR→HU	52.04	76.60	85.42	88.67	**89.46**
TR→MU	73.47	86.91	87.67	89.12	**90.01**

Table 2: OAs (%) of FVMGN with different parts. **NET**-1 is a Baseline with the dual-branch structure of a ViT and three residual blocks, **NET**-2 is **NET**-1 with DTAug, **NET**-3 is **NET**-2 with MWDis, **NET**-4 is **NET**-3 with SFIE, and **NET**-5 is **NET**-4 with MSFFA.

SD→TD	Spat-FA	Freq-FA	MSFFA
MU→TR	89.85	91.93	**92.44**
MU→HU	89.68	86.36	**93.19**
HU→TR	77.16	81.43	**82.33**
HU→MU	79.64	80.93	**82.87**
TR→HU	85.73	87.11	**89.46**
TR→MU	86.15	88.19	**90.01**

Table 3: OAs (%) of FVMGN with the feature alignments (FA). Spat-FA is the spatial FA, Freq-FA is the frequency FA, and MSFFA is the multiscale spatial-frequency FA.

Study on Different Feature Alignments As shown in Table 3, OA of FVMGN with MSFFA outperforms that of FVMGN with spatial FA (Spat-FA) and that of FVMGN with frequency FA (Freq-FA). The performance improvement brought by MSFFA mainly benefits from multiscale and multi-frequency FAs in spatial and frequency domains, which facilitates fine-grained matching of positive vision-vision and vision-text features.

Study on Different Text Descriptions As shown in Fig. 8, FVMGN with both shared and modality-specific texts achieves superior classification performance. This indicates that shared and modality-specific texts can provide general and proprietary linguistic prior knowledge for vision feature representation, respectively, thereby enhancing multimodality generalization ability.

Effect of Training Sample Proportions As shown in Fig. 9, FVMGN has relatively stable classification performance on most training sample ratios. Nevertheless, OA of FVMGN is poor when training sample ratio is 3% on the HU→MU combination, mainly because there is a huge gap in training sample ratios between HU and MU datasets, making it difficult for FVMGN to learn cross-domain invariant features. In addition, FVMGN exhibits minor fluctuations on some datasets, which may be attributed to the following two aspects: 1) Larger training sample size may overly focus on source domain-specific features, such as MU→TR; 2) sampling randomness, such as TR→HU.

Supplementary Notes Due to space limitations, more details are available in supplementary materials. If necessary, you can find them in **Code** link soon. Supplementary materials include but are not limited to: method details, data characteristics, computational complexity, modality extensibility, and more visualizations.

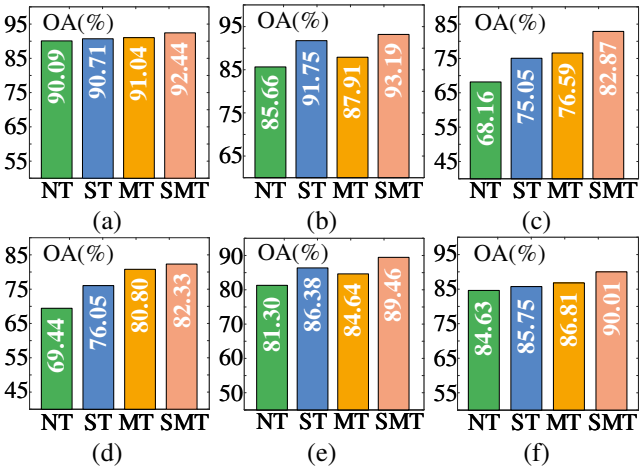

Figure 8: OAs of FVMGN with different texts. (a) MU→TR. (b) MU→HU. (c) HU→TR. (d) HU→MU. (e) TR→HU. (f) TR→MU. NT: No text, ST: Shared text, MT: Modality-specific (proprietary) texts, and SMT: Shared and modality-specific texts.

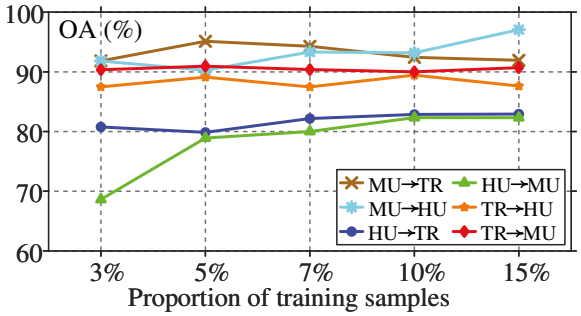

Figure 9: OAs of FVMGN on different datasets.

Conclusion

In this paper, we propose a FVMGN for RSIC that facilitates cross-scene multimodal information representation. Specifically, FVMGN leverages land-cover distributions generated by DTAug strategy to enrich input diversity, exploring the potential and effectiveness of DDPM in data augmentation. On the other hand, MWDis performs Gaussian modeling and histogram equalization on different frequency components in the frequency domain to learn cross-domain-invariant representations. Moreover, WCTMixer elegantly integrates the wavelet transform and attention mechanism to achieve fine multi-frequency analysis and feature reconstruction. Finally, MSFFA introduces multiscale properties and modality attributes based on the vanilla vision-text feature alignment, thereby enhancing the model ability to match and understand positive feature pairs. Extensive experiment results confirm the generalization of FVMGN.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (Grant No. 62071379), the National Science and Technology Major Project (Grant No. 2022ZD0119001), the Postgraduate Innovation Foundation of Xi'an University of Posts and Telecommunications (CXJJBDL2023005), and the Youth Innovation Team of Shaanxi Universities.

References

Ahmad, M.; Ghous, U.; Usama, M.; and Mazzara, M. 2024. WaveFormer: Spectral–spatial wavelet transformer for hyperspectral image classification. *IEEE Geoscience and Remote Sensing Letters*, 21: 5502405.

Chen, N.; Yue, J.; Fang, L.; and Xia, S. 2023. SpectralDiff: A generative framework for hyperspectral image classification with diffusion models. *IEEE Transactions on Geoscience and Remote Sensing*, 61: 5522416.

Chu, M.; Yu, X.; Dong, H.; and Zang, S. 2024. Domain-adversarial generative and dual feature representation discriminative network for hyperspectral image domain generalization. *IEEE Transactions on Geoscience and Remote Sensing*, 62: 5533213.

Debes, C.; Merentitis, A.; Heremans, R.; Hahn, J.; Frangiadakis, N.; Van Kasteren, T.; Liao, W.; Bellens, R.; Pižurica, A.; Gautama, S.; et al. 2014. 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.

Finder, S. E.; Amoyal, R.; Treister, E.; and Freifeld, O. 2025. Wavelet convolutions for large receptive fields. In *Proceedings of the European Conference on Computer Vision*, 363–380.

Gader, P.; Zare, A.; Close, R.; Aitken, J.; and Tuell, G. 2013. MUUFL Gulfport hyperspectral and LiDAR airborne data set. *University of Florida, Gainesville, Florida, United States, Technical Report REP-2013-570*.

Gao, J.; Ji, X.; Ye, F.; and Chen, G. 2025. Invariant semantic domain generalization shuffle network for cross-scene hyperspectral image classification. *Expert Systems with Applications*, 273: 126818.

Gao, X.; Qiu, T.; Zhang, X.; Bai, H.; Liu, K.; Huang, X.; Wei, H.; Zhang, G.; and Liu, H. 2024. Efficient multi-scale network with learnable discrete wavelet transform for blind motion deblurring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2733–2742.

Guo, F.; Meng, Q.; Li, Z.; Ren, G.; Wang, L.; Zhang, J.; Xin, R.; and Hu, Y. 2024a. Multisource feature embedding and interaction fusion network for coastal wetland classification with hyperspectral and LiDAR data. *IEEE Transactions on Geoscience and Remote Sensing*, 62: 5509516.

Guo, J.; Li, J.; Li, D.; Tiong, A. M. H.; Li, B.; Tao, D.; and Hoi, S. 2023a. From images to textual prompts: Zero-shot visual question answering with frozen large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10867–10877.

Guo, J.; Wang, N.; Qi, L.; and Shi, Y. 2023b. Aloft: A lightweight mlp-like architecture with dynamic low-frequency transform for domain generalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 24132–24141.

Guo, X.; Lao, J.; Dang, B.; Zhang, Y.; Yu, L.; Ru, L.; Zhong, L.; Huang, Z.; Wu, K.; Hu, D.; et al. 2024b. Skysense: A multi-modal remote sensing foundation model towards universal interpretation for earth observation imagery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 27672–27683.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.

Hong, D.; Gao, L.; Hang, R.; Zhang, B.; and Chanussot, J. 2020. Deep encoder–decoder networks for classification of hyperspectral and LiDAR data. *IEEE Geoscience and Remote Sensing Letters*, 19: 5500205.

Imani, M. 2025. Attention based network for fusion of polarimetric and contextual features for polarimetric synthetic aperture radar image classification. *Engineering Applications of Artificial Intelligence*, 139: 109665.

Jia, S.; Zhan, Z.; Zhang, M.; Xu, M.; Huang, Q.; Zhou, J.; and Jia, X. 2020. Multiple feature-based superpixel-level decision fusion for hyperspectral and LiDAR data classification. *IEEE Transactions on Geoscience and Remote Sensing*, 59(2): 1437–1452.

Korkmaz, C.; Tekalp, A. M.; and Dogan, Z. 2024. Training generative image super-resolution models by wavelet-domain losses enables better control of artifacts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5926–5936.

Kuckreja, K.; Danish, M. S.; Naseer, M.; Das, A.; Khan, S.; and Khan, F. S. 2024. Geochat: Grounded large vision-language model for remote sensing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 27831–27840.

Li, B.; Gao, Z.; and He, X. 2023. Gradient-map-guided adaptive domain generalization for cross modality MRI segmentation. In *Proceedings of Machine Learning Research*, 292–306.

Li, J.; Hong, D.; Gao, L.; Yao, J.; Zheng, K.; Zhang, B.; and Chanussot, J. 2022. Deep learning in multimodal remote sensing data fusion: A comprehensive review. *International Journal of Applied Earth Observation and Geoinformation*, 112: 102926.

Qin, B.; Feng, S.; Zhao, C.; Xi, B.; Li, W.; and Tao, R. 2024. FDGNet: Frequency disentanglement and data geometry for domain generalization in cross-scene hyperspectral image classification. *IEEE Transactions on Neural Networks and Learning Systems*, 36(6): 10297–10310.

Qu, J.; Yang, Y.; Dong, W.; and Yang, Y. 2024. LDS2AE: Local diffusion shared-specific autoencoder for multimodal remote sensing image classification with arbitrary missing modalities. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 14731–14739.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I.; et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1: 9.

Rotstein, N.; Bensaïd, D.; Brody, S.; Ganz, R.; and Kimmel, R. 2024. Fusecap: Leveraging large language models for enriched fused image captions. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 5689–5700.

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: 5515620.

Sennrich, R.; Haddow, B.; and Birch, A. 2015. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*, 1715–1725.

Seydi, S. T.; Bozorgasl, Z.; and Chen, H. 2024. Unveiling the power of wavelets: A wavelet-based kolmogorov-arnold network for hyperspectral image classification. *arXiv preprint arXiv:2406.07869*.

Tang, X.; Zou, Y.; Ma, J.; Zhang, X.; Liu, F.; and Jiao, L. 2024. Multiple information collaborative fusion network for joint classification of hyperspectral and LiDAR data. *IEEE Transactions on Geoscience and Remote Sensing*, 62: 5525416.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*, 30: 1–11.

Wang, A.; He, X.; Ghamisi, P.; and Chen, Y. 2018. LiDAR data classification using morphological profiles and convolutional neural networks. *IEEE Geoscience and Remote Sensing Letters*, 15(5): 774–778.

Wang, H.; Cheng, Y.; Liu, X.; and Wang, X. 2024a. Reinforcement learning based Markov edge decoupled fusion network for fusion classification of hyperspectral and LiDAR. *IEEE Transactions on Multimedia*, 26: 7174–7187.

Wang, X.; Dong, S.; Zheng, X.; Lu, R.; and Jia, J. 2024b. Explicit high-level semantic network for domain generalization in hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 62: 5538314.

Woo, S.; Park, J.; Lee, J.-Y.; and Kweon, I. S. 2018. CBAM: Convolutional block attention module. In *Proceedings of the European Conference on Computer Vision*, 3–19.

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: 5511015.

Yuan, Y.; and Yuan, C. 2024. Efficient conditional diffusion model with probability flow sampling for image super-resolution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 6862–6870.

Zhang, B.; Chen, Y.; Xiong, S.; and Lu, X. 2025. Hyperspectral image classification via cascaded spatial cross-attention network. *IEEE Transactions on Image Processing*, 34: 899–913.

Zhang, F.; Qu, S.; Shi, F.; and Xu, C. 2024a. Overcoming the pitfalls of vision-language model for image-text retrieval. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 2350–2359.

Zhang, J.; Huang, J.; Jin, S.; and Lu, S. 2024b. Vision-language models for vision tasks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(8): 5625–5644.

Zhang, J.; Zhang, C.; Liu, S.; Shi, Z.; and Pan, B. 2024c. Three-dimensional frequency domain transform network for cross-scene hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 62: 5409813.

Zhang, J.; Zhao, F.; Liu, H.; and Yu, J. 2024d. Data and knowledge-driven deep multiview fusion network based on diffusion model for hyperspectral image classification. *Expert Systems with Applications*, 249: 123796.

Zhang, Y.; Gao, H.; Zhou, J.; Zhang, C.; Ghamisi, P.; Xu, S.; Li, C.; and Zhang, B. 2024e. A cross-modal feature aggregation and enhancement network for hyperspectral and LiDAR joint classification. *Expert Systems with Applications*, 258: 125145.

Zhang, Y.; Li, W.; Sun, W.; Tao, R.; and Du, Q. 2023a. Single-source domain expansion network for cross-scene hyperspectral image classification. *IEEE Transactions on Image Processing*, 32: 1498–1512.

Zhang, Y.; Zhang, M.; Li, W.; Wang, S.; and Tao, R. 2023b. Language-aware domain generalization network for cross-scene hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 61: 5501312.

Zhao, F.; Zhang, J.; Meng, Z.; Liu, H.; Chang, Z.; and Fan, J. 2023a. Multiple vision architectures-based hybrid network for hyperspectral image classification. *Expert Systems with Applications*, 234: 121032.

Zhao, H.; Lin, L.; Wang, J.; Gao, S.; and Zhang, Z. 2025. Adversarial decoupling domain generalization network for cross-scene hyperspectral image classification. *Knowledge-Based Systems*, 318: 113432.

Zhao, H.; Zhang, J.; Lin, L.; Wang, J.; Gao, S.; and Zhang, Z. 2023b. Locally linear unbiased randomization network for cross-scene hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 61: 5526512.

Zhao, M.; Liu, M.; Ren, B.; Dai, S.; and Sebe, N. 2024. Denoising diffusion probabilistic models for action-conditioned 3D motion generation. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 4225–4229.

Zhu, Y.; Zhang, K.; Liang, J.; Cao, J.; Wen, B.; Timofte, R.; and Van Gool, L. 2023. Denoising diffusion models for plug-and-play image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1219–1229.