LDS^{2}AE: Local Diffusion Shared-Specific Autoencoder for Multimodal Remote Sensing Image Classification with Arbitrary Missing Modalities

Jiahui Qu*, Yuanbo Yang*, Wenqian Dong†, Yufei Yang

State Key Laboratory of Integrated Service Network, Xidian University, Xi'an 710071, China
jhqu@xidian.edu.cn, yuanboyang@stu.xidian.edu.cn, wqdong@xidian.edu.cn, yfyang_2021@stu.xidian.edu.cn

Abstract

Recent research on the joint classification of multimodal remote sensing data has achieved great success. However, due to the limitations imposed by imaging conditions, the case of missing modalities often occurs in practice. Most previous researchers regard the classification in case of different missing modalities as independent tasks. They train a specific classification model for each fixed missing modality by extracting multimodal joint representation, which cannot handle the classification of arbitrary (including multiple and random) missing modalities. In this work, we propose a local diffusion shared-specific autoencoder (LDS^{2}AE), which solves the classification of arbitrary missing modalities with a single model. The LDS^{2}AE captures the data distribution of different modalities to learn multimodal shared feature for classification by designing a novel local diffusion autoencoder which consists of a modality-shared encoder and several modality-specific decoders. The modality-shared encoder is designed to extract multimodal shared feature by employing the same parameters to map multimodal data into a shared subspace. The modality-specific decoders put the multimodal shared feature to reconstruct the image of each modality, which facilitates the shared feature to learn unique information of different modalities. In addition, we incorporate masked training to the diffusion autoencoder to achieve local diffusion, which significantly reduces the training cost of model. The approach is tested on widely-used multimodal remote sensing datasets, demonstrating the effectiveness of the proposed LDS^{2}AE in addressing the classification of arbitrary missing modalities. The code is available at https://github.com/Jiahuiqu/LDS2AE.

Introduction

Remote sensing images of the same geographic area captured from different sensors can provide complementary ground feature (Rasti, Ghamisi, and Gloaguen 2017a; Su et al. 2021; Ghamisi et al. 2018). The joint classification of multimodal remote sensing data is an effective technique to integrate the complementary information of different modalities to improve the classification accuracy, and has been widely used in urban planning (Zhang et al. 2020a; Dong et al. 2023), natural resources management (Chen et al. 2019), environmental monitoring (Li et al. 2020; Qu et al. 2023) and water quality monitoring (Mei et al. 2021). However, in practice, the case of missing modalities often occurs, due to sensor malfunctions, weather conditions, or other factors (Zhang et al. 2018). The most existing methods treat the joint classification with different modalities as independent tasks, which makes it hard to put into pratical use (Wang et al. 2020; Park et al. 2019; Zhang et al. 2020b). So far, multimodal remote sensing image classification with arbitrary missing modalities remains unfully explored.

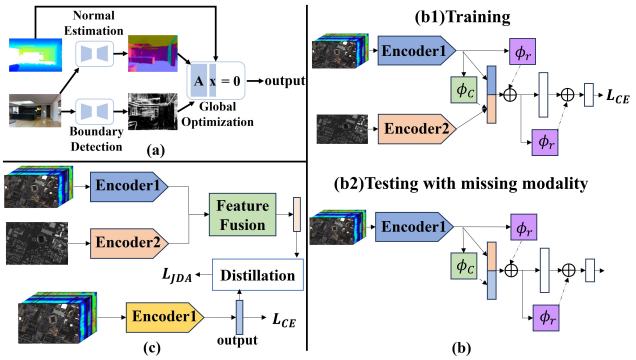

Figure 1: The methods for dealing with missing modalities. (a) Reconstruction of images with missing modalities (Xue, Zhang, and Cai 2016); (b) Reconstruction of feature with missing modalities (Ma et al. 2021); (c) Multimodal joint representation learning (Wei et al. 2023).

The mainstream approaches to address the challenge of multimodal image classification with missing modalities can be summarized in two ways: 1) generative methods (Holloway et al. 2019; Ma et al. 2021), and 2) multimodal joint representation learning (Zhang et al. 2018).

The generative-based methods can be divided into two categories: the reconstruction of images with missing modalities and the reconstruction of feature with missing modalities. For example, Xue et al. (Xue, Zhang, and Cai 2016) generates the deep channel of RGB-D images by using a low-rank matrix improved by low gradient regularization. Li et al. (Li et al. 2022) introduces the Dynamic Hierarchical Attention distillation module (DHAD) to gen-

*These authors contributed equally.
†Corresponding Authors

erate Synthetic aperture radar (SAR) image features from RGB images, aiming to train available modality to reconstruct representations of missing modality by directly matching their intermediate feature mappings.

The key of multimodal joint representation learning is to integrate and learn information from different modalities to better understand and represent cross-modal feature. Although each modality possesses unique characteristics, they often share common information within the semantic space. For example, Hazarika et al. (Hazarika, Zimmermann, and Poria 2020) introduces a shared subspace to discover potential commonalities among different modalities, aiming to diminish the effect of modal gaps. Dutt et al. (Dutt, Zare, and Gader 2022) develops a common shared manifolds model that learns shared feature representations from hyperspectral (HS) and Light Detection and Ranging (LiDAR) image.

While these methods have shown promising results for multimodal image classification with missing modalities, they still face some challenges and limitations: 1) They typically train one feature completion model for each certain missing modality and overlook the more prevalent scenario of missing multiple modalities. 2) These approaches predominantly concentrate on acquiring joint representation within a shared subspace, potentially leading to the omission of certain modality-specific features. These disadvantages limit the use of multimodal remote sensing image classification to real-world scenarios.

To put multimodal remote sensing image classification into more practical use, we propose a local diffusion shared-specific autoencoder (LDS^{2}AE) network to address multimodal remote sensing image classification with arbitrary missing modalities. For the first challenge, we present a modality-shared encoder to extract multimodal joint representation, which allows concurrent encoding of data from different modalities using the same set of parameters. After the pre-training phase, the modality-shared encoder can directly handle missing singular and multiple modalities. For the second challenge, the shared feature reconstructs the entire image of all modality through all modality-specific decoders, which allows the shared feature to learn specific feature of other modalities. We incorporate the whole encoding and decoding process into a denoising diffusion model with strong implicit learning capabilities to execute the self-reconstruction and cross-modal reconstruction tasks, which helps to deal with the difficulty of cross-modal reconstruction caused by the huge modal gap. Moreover, we introduce masked training into diffusion model to speed up training and reduce memory consumption, which is mainly based on the intuition that the image is highly redundant in space. The LDS^{2}AE denoises and reconstructs unmasked pixels based on the reconstruction of masked pixels. Therefore, we propose a new training objective to predict the denoising reconstruction score of unmasked patches while simultaneously reconstructing the masked patches.

To summarize, we make the following contributions:

1) We propose a novel framework LDS^{2}AE to deal with multimodal remote sensing image classification with arbitrary missing modalities, which designs a modality-shared encoder to learn joint representation and several modality-specific decoders to learn unique characteristics of different modalities.
2) We exploit the denoising diffusion model to achieve cross-modal reconstruction of remote sensing image with the large gap of data distribution, which helps the model to learn independent and complementary features in multi-modal data.
3) We quickly train the diffusion model by randomly masking a high proportion (e.g., 70%) of input patches, and add the masked reconstruction task to the denoising loss of the diffusion model.

Related Work

Denoising Diffusion Model Denoising diffusion model (DDPM) (Nichol and Dhariwal 2021) is a class of generative models that captures the potential probability distribution of data by gradually adding a standard gaussian noise to the sample and learning a model to reverse the process (Dhariwal and Nichol 2021). The DDPM holds significant advantages in terms of its high-level semantic feature capture ability, potential spatial continuity, exploration of feature space, denoising capability, and recovery potential (Sohl-Dickstein et al. 2015; Zhou et al. 2023; Yang et al. 2023). In this paper, we introduce diffusion model to capture the underlying patterns and relationships between different modalities to assist in reconstruction and cross-modal reconstruction tasks. Furthermore, we incorporate the idea of masked training into diffusion model to achieve local diffusion, which can speed up training and reduce memory overhead.

Masked Training Masked training begin as a task in the field of natural language processing to fill in or predict masked parts of text, thereby inferring missing words or phrases (Devlin et al. 2019; Liu et al. 2019). This kind of task is usually done with a pre-trained language model, such as BERT (Devlin et al. 2019) or GPTs (Radford and Narasimhan 2018). With the introduction of ViT (Dosovitskiy et al. 2020), a large number of self-supervised works are proposed to learn useful representations by predicting the content of masked or obscured areas in images, such as MAE (He et al. 2021), Simmim (Bao, Dong, and Wei 2021) and Beit (Xie et al. 2021), which all work well in a variety of downstream tasks. In particular, MAE employs an asymmetric architecture to accelerate pre-training, consisting of an encoder that operates only on the visible part, and a lightweight decoder that reconstructs the masked patches with the latent representation of the visible part and masked tokens. We also use a lightweight architecture similar to MAE to achieve local diffusion.

Methods

Overview

The proposed LDS^{2}AE aims to learn complementary representations from different modalities via the local diffusion shared-specific autoencoder, which can deal with multimodal remote sensing image classification with arbitrary missing modalities. The framework is shown in Figure 2. The method consists of two key stages: pre-training and

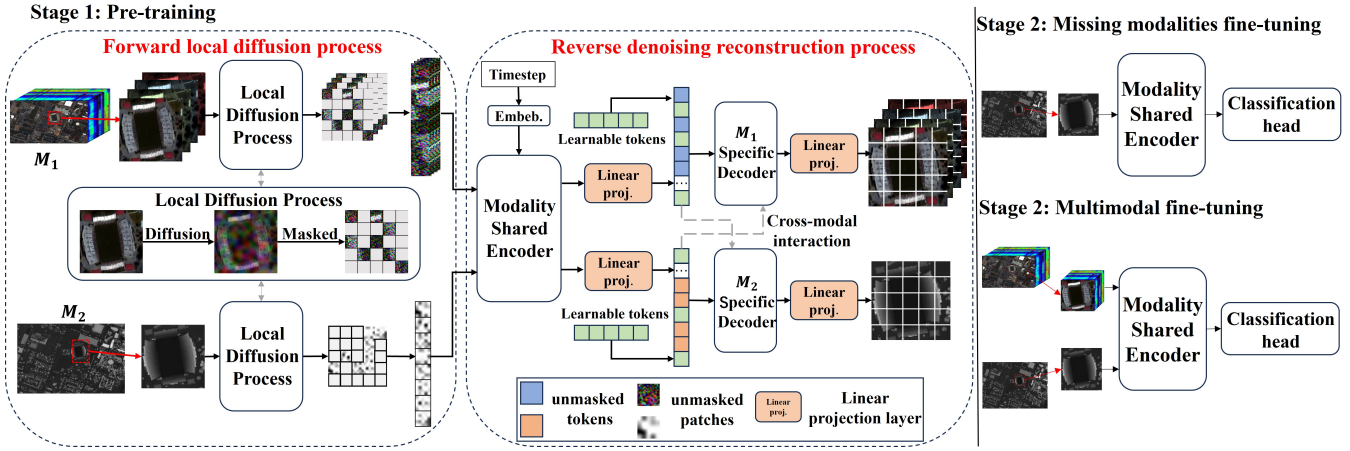

Figure 2: Overall architecture of the proposed LDS^{2}AE. The method consists of two stages: 1) In the multimodal pre-training stage, the encoding and decoding operations are incorporated into the local diffusion model to guide the network to learn the shared and specific knowledge among all input images; 2) Fine-tuning stage, arbitrary missing modality inference and multimodal inference tasks only need to fine-tune the modality-sharing encoder without modifying the model architecture.

fine-tuning. The pre-training stage includes two processes of the forward local diffusion and the reverse denoising reconstruction, which are parameterized Markov chains. The forward local diffusion is performed only on part of the input data to reduce the training cost of diffusion model. We design a modality-shared encoder in the reverse denoising reconstruction to learn the multimodal shared feature and several modality-specific decoders to facilitate the shared feature to learn specific properties of different modalities. The fine-tuning stage is a supervised learning stage where only the fully connected layer are fine-tuned for multimodal remote sensing image classification with arbitrary missing modalities, which helps the feature of available modality extracted by the encoder to learn the multimodal features of pre-training phase.

Forward Local Diffusion Process

In the forward local diffusion, the real data $x_0 \sim p_{data}$ is divided into non-overlapping patches. We randomly masked some patches as x_0^m according to a fixed masking ratio and treat the rest as visible patches x_0^v. Only visible patches x_0^v is performed diffusion and added gaussian noise n at time t with variance $\beta_t \in (0,1)$ to produce x_1^v through x_T^v as follows the Markov process below:

$$q(x_1^v, \ldots, x_T^v | x_0^v) = \prod_{t=1}^{T} L\_q(x_t^v | x_{t-1}^v) \tag{1}$$

$$L\_q(x_t^v | x_{t-1}^v) = N(x_t^v; \sqrt{1-\beta_t} x_{t-1}^v, \beta_t I), \tag{2}$$

where β_t for different t is pre-defined and undergoes a gradual linear decay, satisfied $\beta_1 < \beta_2 < \ldots < \beta_T$.

For the masked operation, LDS^{2}AE employs an asymmetric masking strategy for each modality, which facilitates the model to more efficiently capture the shared information of different modalities. The local diffusion minimizes visual redundancy and brings in highly sparse inputs, which reduces the computational cost of diffusion model.

Reverse Denoising Reconstruction Process

In the reverse process, the proposed model predicts input data x_0 based on the current sampling time t. This modification is based on Bayesian theory, a posteriori distribution $q(x_{t-1} | x_t, x_0)$ can be calculated in terms of $\tilde{\mu}_t(x_t, x_0)$ and $\tilde{\beta}_t$.

The LDS^{2}AE predicts input data x_0 from the noisy sample x_T^v by estimating the denoising reconstruction score of unmasked patches x_0^v and simultaneously reconstructing the masked patches x_0^m by designing a shared, learnable vector for each masked token. Under large T and small β_t, the x_T is approximated as a gaussian distribution and predicted by a learned neural network as follows:

$$p_\theta(x_{0:T}^v) = p(x_T^v) \prod_{t=1}^{T} p_\theta(x_{t-1}^v | x_t^v) \tag{3}$$

$$p_\theta(x_{t-1}^v | x_t^v) = \mathcal{N}(x_{t-1}^v; \mu_\theta(x_t^v, t), \sigma_\theta(x_t^v, t)), \tag{4}$$

where $\mu_\theta(x_t^v, t)$ is the expectation of x_t^v, $\sigma_\theta(x_t, t)$ is the variance of x_t^v.

At the end, we update the model by minimizing the denoising reconstruction loss:

$$E_{x_0 \sim p_{data}} E_{n \sim N(0, t^2 I)} ||x_0 - model(x_0 + n, t)||^2. \tag{5}$$

Specifically, the reverse denoising reconstruction process is executed through a modality-shared encoder and several modality-specific decoders, whose particular implementation is introduced in details as follows.

Modality-Shared Encoder The encoder backbone is based on ViT with some modifications. Specifically, the sine-cosine positional embeddings, diffusion timesteps and class tokens are added to the visible patches after diffusion and projected by the encoder $E_\phi(\cdot)$ into the shared latent space for subsequent reconstruction tasks.

$$\hat{x}^v = E_\phi(\{t_{cls} + t_{ts} + t_p; x^v\}), \tag{6}$$

Algorithm 1: Pseudocode of Pretraining for LDS^{2}AE

Input: the multi-model encoder E_ϕ, HSI-decoder D_{HSI}, HS data x, LiDAR data x_L, LiDAR-decoder D_{LiDAR}, masking ratio r, gaussian noise n, the diffusion timestep t, q_sample means forward process

for x, x_L, y **in loader:**
  $x^t, x_L^t = q\_sample(x, x_L, t, n)$
  $x^v, x^m, ids, m = \text{random\_mask}(x^t, \text{r})$
  $x_L^v, x_L^m, idx_L, m_L = \text{random\_mask}(x_L^t, \text{r})$
  $\hat{x}^v, \hat{x}_L^v = E_\phi\left(x^v, x_L^v\right)$
  $\hat{x} = D_{HSI}\left(\hat{x}^m, \hat{x}^v, idx\right)$
  $\hat{x}_L = D_{LiDAR}\left(\hat{x}_L^m, \hat{x}_L^v, idx_L\right)$
  $\mathcal{L}_{rec} = rec\_l(\hat{x}, x, m) + rec\_l(\hat{x}_L, x_L, m_L)$
  $\mathcal{L}_{v-rec} = mse\_l(\hat{x}, x, m) + mse\_l(\hat{x}_L, x_L, m_L)$
  $\hat{x}_c = D_{HSI}\left(\hat{x}_L^m, \hat{x}_L^v, idx_L\right)$
  $\hat{x}_{c-L} = D_{LiDAR}\left(\hat{x}^m, \hat{x}^v, idx\right)$
  $L_{c-rec} = rec\_l(\hat{x}_c, x, m) + rec\_l(\hat{x}_{c-L}, x_L, m_L) + mse\_l(\hat{x}_c, x, m) + mse\_l(\hat{x}_{c-L}, x_L, m_L)$
  $\mathcal{L}_{loss} = \lambda_{rec}\mathcal{L}_{rec} + \lambda_{c-rec}\mathcal{L}_{c-rec} + \lambda_{v-rec}\mathcal{L}_{v-rec}$
  loss.backward()
  update()

where $\hat{x}^v$ represents the encoded visible feature, t_{cls}, t_{ts}, t_p stand for class tokens, diffusion timestep and positional embeddings, respectively.

We employ a modality-shared encoder to map all input data into a shared subspace to effectively capture complementary features among diverse modalities, which facilitates the encoder to address the classification of arbitrary missing modalities in fine-tuning stage. The encoder codes multimodal visible tokens locally diffused by using a uniform set of parameters, resulting in a substantial reduction in the computational cost of acquiring shared representations. Moreover, each modality passes through its own linear projection layer after passing through the same encoder.

Modality-Specific Decoders Each modality requires its dedicated decoder due to different reconstruction tasks. The decoders are made up of a series of Transformer blocks that are narrower and shallower than the modality-shared encoder. The inputs of the decoder are the visible tokens $\hat{x}^v$ encoded by the encoder and a set of masked tokens $\hat{x}^m$. Each of masked tokens is a learnable vector initialized to zeros. Before passing all tokens to the decoder $D_\phi(\cdot)$, we add the same sine-cosine positional embeddings to each of them to indicate their positions within the image.

$$\hat{x} = D_\phi\left(\{t_p + \hat{x}^m; t_p + \hat{x}^v\}\right), \tag{7}$$

where $\hat{x}$ is the reconstructed data converted back to the original input space.

We introduce diffusion model to help the decoders to perform self-reconstruction tasks while achieve cross-modal reconstruction by exchanging the decoders' inputs, which facilitates shared features to capture the distinct representations of each modality.

The proposed LDS^{2}AE allows model to benefit from modality-shared of knowledge and modality-specific information. The modality-shared encoder ensures that the model learns the multimodal shared representation. At the same time, the modality-specific decoders assist the shared feature to accurately capture the unique feature of each modality through self-reconstruction and cross-modal reconstruction processes. By combining modality-shared encoder and modality-specific decoders, LDS^{2}AE can efficiently process multimodal data, leveraging shared knowledge while preserving the notable traits of each modality.

Training Objective

Stage1: Pre-training The pre-training process defines a hybrid optimization objective consisting of denoising self-reconstruction loss of visible tokens, self-reconstruction loss of masked tokens and cross-modal reconstruction loss:

$$\mathcal{L} = \lambda_{m-rec}\mathcal{L}_{m-rec} + \lambda_{c-rec}\mathcal{L}_{c-rec} + \lambda_{v-rec}\mathcal{L}_{v-rec}, \tag{8}$$

where the hyperparameter λ_{m-rec}, λ_{c-rec}, λ_{v-rec} control the balance of multiple losses. We assign different hyperparameters to drive the model together, due to the difficulty of different reconstruction tasks. We further demonstrate more detailed method like pytorch in Algorithm 1.

The denoising self-reconstruction loss of visible tokens $\mathcal{L}_{v-rec}$ is actually the denoising loss of the local diffusion model. The traditional diffusion models calculate the score of the general map, but it is difficult to reconstruct the whole image with only visible tokens after noise sampling. Therefore, the model utilizes the mean square error (MSE) loss to only compute the visible portions, aiming to prevent the model from overfitting to unmasked tokens. We consider this loss as the main loss of model, and the formula is as follows:

$$\mathcal{L}_{v-rec} = E_m \parallel \left(D_\phi\left(\hat{x}^m, \hat{x}^v\right) - x_0\right) \odot (1 - m)\|^2, \tag{9}$$

where $\odot$ denotes the element-wise multiplication along the token length dimension of the "patchify".

The self-reconstruction loss of masked tokens $\mathcal{L}_{m-rec}$ is a reconstruction loss similar to MAE. We calculate the reconstruction loss $\mathcal{L}_{m-rec}$ on the masked tokens and assign a lower training weight:

$$\mathcal{L}_{m-rec} = E_m \parallel \left(D_\phi\left(\hat{x}^m, \hat{x}^v\right) - x_0\right) \odot m\|^2. \tag{10}$$

The cross-modal reconstruction loss $\mathcal{L}_{c-rec}$ is composed of the denoising cross-reconstruction loss of visible tokens and the cross-reconstruction loss of masked tokens, which is the same as the self-reconstruction loss. We execute the cross-modal reconstruction obtained by exchanging the decoders' inputs to learn modality-specific knowledge.

$$\mathcal{L}_{c-rec} = \mathcal{L}_{v-rec}^{cross} + \mathcal{L}_{m-rec}^{cross}, \tag{11}$$

where $\mathcal{L}_{c-rec}$ only takes the cross-modal reconstruction of one modality to another as an example, it is actually a bidirectional process that allows multimodal joint representation to learn unique representations of other modality.

Stage2: Fine-tuning The fine-tuning process is optimized with classification loss, which uses the cross entropy loss to calculate the difference between the prediction $\hat{y}$ of the class tokens after passing through the classification head and the true label of the corresponding sample.

$$\mathcal{L}_{cls} = -\sum_{n=1}^{N} y_n \log\left(\hat{y}_n\right), \tag{12}$$

where N is the number of classes.

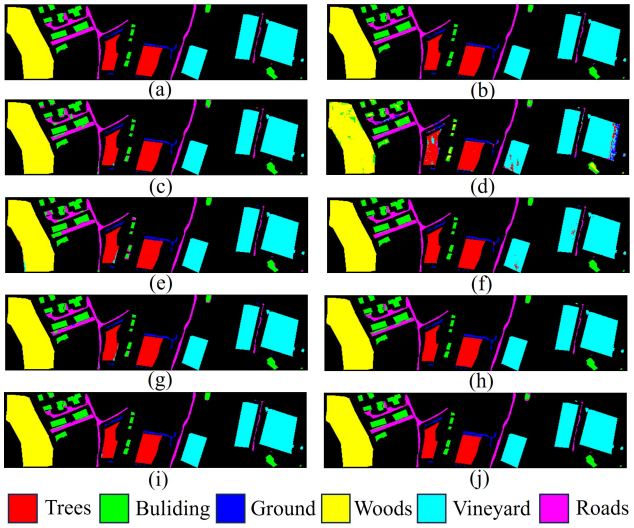

Figure 3: Classification maps of the Trento dataset. (a) Ground-truth map. (b) LDS^{2}AE (HS and LiDAR). (c) LDS^{2}AE (HS). (d) LDS^{2}AE (LiDAR). (e) DMAE (HS). (f) TBCNN. (g) Sal2RN. (h) HRWN. (i) MFT. (j) HRWN.

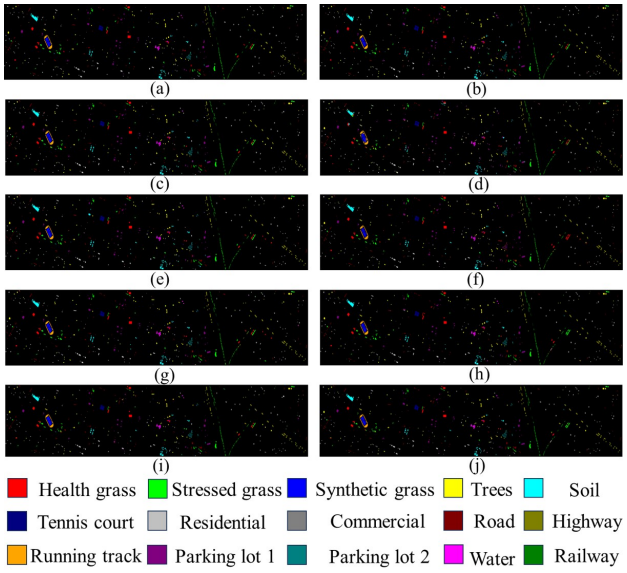

Figure 4: Classification maps of the Houston dataset. (a) Ground-truth map. (b) LDS^{2}AE (HS and LiDAR). (c) LDS^{2}AE (HS). (d) LDS^{2}AE (LiDAR). (e) DMAE (HS). (f) TBCNN. (g) Sal2RN. (h) HRWN. (i) MFT. (j) HRWN.

Experiments

Datasets Description

1) Houston: The Houston dataset is introduced in the IEEE GRSS data fusion contest held in 2013 and contains HS, multispectral (MS) and LiDAR images. It specifically focuses on a university campus and includes fifteen categories, encompassing both natural and man-made objects.

2) Trento: The Trento dataset (Rasti, Ghamisi, and Gloaguen 2017b) is collected in the rural area of southern Trento, Italy. It consists of HS and LiDAR data for six vegetated land cover categories.

3) Berlin: The Berlin dataset (Hong et al. 2021b) describes the geomorphological composition of the urban area of Berlin and its surrounding rural areas. It includes HS and SAR data, encompassing eight categories in total.

Experiments Setup

1) Evaluation Criteria: Three classification evaluation indexes are considered, which are overall accuracy (OA), average accuracy (AA) and kappa coefficient (Kappa).

2) Implementation Details: The proposed method is implemented on the PyTorch platform. The entire framework is trained using the Adam optimizer with a pre-training epoch of 400, a fine-tuning epoch of 150, a batch size of 1024, and learning rate of 1e-3. Additionally, we employ the CosineAnnealingLR strategy during the pre-training phase and the Multi StepLR policy during fine-tuning phase to update the learning rate.

3) Comparison Methods: To evaluate the effectiveness of LDS^{2}AE in dealing with multimodal remote sensing image classification with arbitrary missing modalities, we compare LDS^{2}AE with methods falling under three distinct categories: a) Multimodal training and inference: Sal2RN (Li et al. 2023), HRWN (Zhao et al. 2020), MFT (Roy et al. 2023), GLT-Net (Ding et al. 2022). b) Multimodal training and missing modality in inference: MDL-RS (Hong et al. 2021a). c) Single-modal training and inference: TBCNN (Xu, Du, and Zhang 2018), the variant of LDS^{2}AE that trains and tests with single-modal data (DMAE). Figure. 3-5 display classification results of Trento, Houston, and Berlin.

Joint Classification Analysis of HS and LiDAR

To evaluate the performance of the proposed network, comparative experiments are conducted using 40 samples for each class on the highly heterogeneous HS and LiDAR datasets Trento and Houston. The upper part and the middle part of Table 1 respectively depict the performance comparison of OA, AA and Kappa of Trento and Houston datasets under three different types of methods. Firstly, it is evident that the approaches employing multimodal training and testing yield superior results compared to the most advanced method used single-modal training and testing, which indicates the synergistic effect of multimodal data fusion in land cover classification. Under the absence of LiDAR or HS image, the OA of TBCNN decreases by 4.60% and 19.21% on the Trento datasets, and exhibits a more substantial decline of 8.47% and 39.23% on the Houston dataset with more complex object types. In contrast, LDS^{2}AE considerably improves this problem and maintains a performance improvement over the single-modal variant DMAE. Specifically, the OA of missing LiDAR image increases by 0.86% on the Trento dataset and by 1.09% on the Houston dataset. On the other hand, the OA for missing HS image shows a remarkable improvement, with an increase of 10.62% and 11.85%, respectively. When the LiDAR image is missing, LDS^{2}AE achieves comparable performance with the

Datasets	Method	Single-modal training and test				Missing multimodal in test				Multimodal training and multimodal test					
		TBCNN		DMAE		MDL-RS		our		TBCNN	Sal2RN	HRWN	MFT	GLT-Net	our
Trento	Training	HS	LiDAR	HS	LiDAR	HS, LiDAR		HS, LiDAR		HS, LiDAR		HS, LiDAR		HS, LiDAR	
	Testing	HS	LiDAR	HS	LiDAR	HS	LiDAR	HS	LiDAR	HS, LiDAR		HS, LiDAR		HS, LiDAR	
	OA(%)	94.60	79.99	97.47	81.21	89.23	67.93	98.33	91.83	99.20	99.03	99.00	99.21	99.07	99.53
	AA(%)	93.81	60.10	96.21	76.11	89.71	72.18	97.07	89.04	98.70	98.47	98.01	98.69	98.23	99.00
	Kappa	92.85	72.09	96.63	75.06	85.78	60.19	97.78	89.19	98.94	98.71	98.66	98.95	98.76	99.39
Houston	Training	HS	LiDAR	HS	LiDAR	HS, LiDAR		HS, LiDAR		HS, LiDAR		HS, LiDAR		HS, LiDAR	
	Testing	HS	LiDAR	HS	LiDAR	HS	LiDAR	HS	LiDAR	HS, LiDAR		HS, LiDAR		HS, LiDAR	
	OA(%)	79.14	48.38	94.69	67.94	86.4	69.93	95.78	79.79	87.61	94.78	95.28	94.69	95.12	**96.19**
	AA(%)	77.11	48.37	95.35	73.01	87.32	73.53	95.98	81.55	84.00	95.48	95.82	95.49	95.87	**96.83**
	Kappa	81.9	47.02	94.27	65.44	85.31	67.63	95.44	78.19	91.56	94.36	95.28	94.26	94.73	**95.89**
Berlin	Training	HS	SAR	HS	SAR	HS, SAR		HS, SAR		HS, SAR		HS, SAR		HS, SAR	
	Testing	HS	SAR	HS	SAR	HS	SAR	HS	SAR	HS, SAR		HS, SAR		HS, SAR	
	OA(%)	63.10	40.93	66.59	38.78	64.88	38.11	73.92	53.21	70.65	73.65	65.78	73.64	66.26	**76.83**
	AA(%)	66.16	26.25	67.88	46.46	64.36	43.20	74.10	47.10	67.73	68.63	63.81	61.20	68.25	**75.86**
	Kappa	50.75	16.54	54.31	25.50	51.99	25.40	62.84	37.63	58.48	51.57	52.21	59.83	53.99	**66.48**

Table 1: Classification accuracy of different methods on Trento, Houston and Berlin Datasets. The best one is shown in bold, and the best one under the case of missing modalities is underlined. Training represents the available modalities for training, and Testing represents the available modalities for testing.

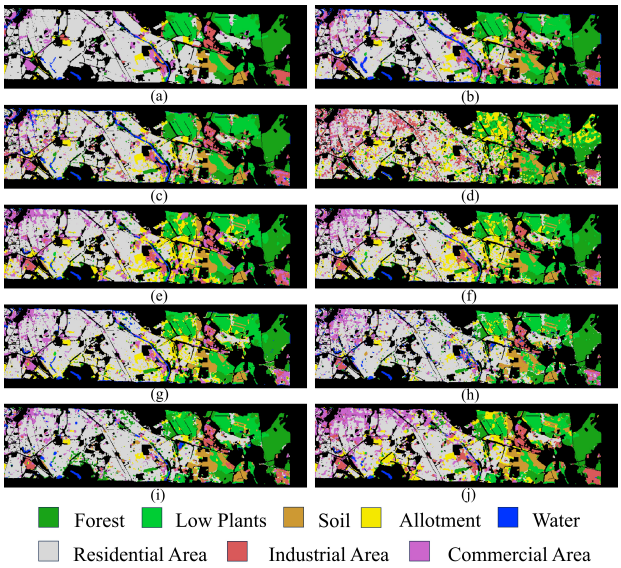

Figure 5: Classification maps of the Berlin dataset. (a) Ground-truth map. (b) LDS^{2}AE (HS and SAR). (c) LDS^{2}AE (HS). (d) LDS^{2}AE (SAR). (e) DMAE (HS). (f) TBCNN. (g) Sal2RN. (h) HRWN. (i) MFT. (j) HRWN.

methods that utilize multimodal training and testing, which demonstrates LDS^{2}AE can effectively learn and leverage both modality-shared information and modality-specific information to address the classification of missing modalities.

Joint Classification Analysis of HS and SAR

We also carry out comparative experiments on the joint classification of HS and SAR image using the standard training set of Berlin dataset. As presented in Table 1, the proposed

Training Modalities	Testing Modalities	OA(%)	AA(%)	Kappa
HS, LiDAR, MS	HS, LiDAR, MS	**96.99**	**97.47**	**96.75**
	HS, LiDAR	96.45	97.04	96.17
	HS, MS	96.74	97.20	96.48
	MS, LiDAR	96.14	96.29	95.84
	HS	96.20	96.43	95.79
	LiDAR	80.81	83.22	79.29
	MS	95.34	96.13	94.97

Table 2: Joint classification accuracy of combinations involving HS, LiDAR, and MS images on the Houston dataset.

LDS^{2}AE demonstrates superior performance compared to other multimodal models, achieving an OA that is 3.18% higher than the second-best performing model, Sal2RN. Furthermore, even in case of missing SAR image, it outperforms Sal2RN by an additional 0.27%. This demonstrates the potential of our method in the classification of missing modalities, as it surpasses the performance of multimodal models tested with all modalities while in case of missing modalities. Furthermore, in the absence of the HS modality, LDS^{2}AE exhibits a remarkable 14.43% improvement in OA compared to DMAE that trains and tests with SAR.

Joint Classification Analysis of HS, LiDAR and SAR

This section conducts experiments on the joint classification of HS, LiDAR, and MS images to evaluate the performance of LDS^{2}AE in case of missing multiple modalities. As shown in Table 2, The proposed method only reduces the OA by 0.85%, 0.54% and 0.25% in the absence of HS, MS or LiDAR images, respectively. The OA for single-modal testing also far exceeds the baseline of the Houston dataset in Table 1, which demonstrates that LDS^{2}AE still has good

Datasets		Cross-modal Reconstruction			Diffusion Model			our		
	Training Modalities	HS, LiDAR			HS, LiDAR			HS, LiDAR		
	Testing Modalities	HS, LiDAR	HS	LiDAR	HS, LiDAR	HS	LiDAR	HS, LiDAR	HS	LiDAR
Trento	OA(%)	98.60	97.05	89.84	98.65	97.01	90.61	99.53	98.33	91.83
	AA(%)	97.88	95.43	87.45	97.98	95.51	88.56	99.00	97.07	89.04
	Kappa	98.13	96.10	86.60	98.21	96.01	87.60	99.39	97.78	89.19
	Training Modalities	HS, SAR			HS, SAR			HS, SAR		
	Testing Modalities	HS, SAR	HS	SAR	HS, SAR	HS	SAR	HS, SAR	HS	SAR
Berlin	OA(%)	72.95	72.42	49.32	73.96	72.51	50.17	76.83	73.92	53.21
	AA(%)	73.32	72.89	45.68	74.00	73.93	47.09	75.86	74.10	47.10
	Kappa	61.35	61.03	33.56	62.84	61.19	34.78	66.49	62.84	37.63

Table 3: The initial two columns show classification results on Trento and Berlin datasets respectively with the removal of the cross-modal reconstruction branch or the diffusion model. The last column represents classification results of the full model.

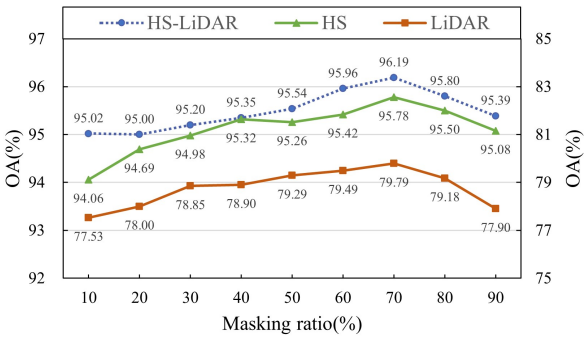

Figure 6: Classification accuracy of Houston dataset with different masking ratios (The testing OA of HS-LiDAR and HS correspond to the left vertical axis, and testing OA of LiDAR correspond to the right vertical axis).

robustness in case of missing multiple modalities.

Ablation Experiments

In this section, we conduct ablation experiments to verify the impact of different configurations on model performance.

Effect of Masking Ratio To investigate the effect of masking ratio, we compare the performance of LDS^{2}AE with different masking ratios. A higher masking ratio means fewer visible patches are subject to diffusion and encoding, reducing computational overhead while potentially losing some useful contextual information. As shown in Figure 6, the value of 70% performs best for the Houston dataset across three ways of inference, and a large range (40% - 80%) works well. Based on these findings, the masking ratio of 70% is chosen for the experiments.

Effect of Cross-Modal Reconstruction In this paper, the cross-modal reconstruction is constructed to help the shared feature learn unique features of other modalities. To confirm the efficacy of cross-modal reconstruction, we examine an altered version, only using encoder to learn the multimodal shared feature. Based on the results presented in Table 3, the OA of Trento dataset has improved by 0.93%, 1.28%, and 1.99% across the three experimental settings, respectively. Similarly, for the Berlin dataset, the OA increases by 3.88%, 1.50%, and 3.74%, in turn. These results show the superiority of introducing cross-modal reconstruction to learn modal-specific knowledge compared to model that only learns shared knowledge.

Effect of Diffusion Model Here the diffusion model is constructed to facilitate the model to learn modality-shared feature and modality-specific knowledge. To verify the effectiveness of diffusion model, we test the variant version of LDS^{2}AE by only using masked processing instead of local diffusion. As indicated by the OA, AA and Kappa of the proposed method and its variant in Table 3, the proposed method achieves the highest classification accuracy across all three forms of testing for the Trento and Berlin datasets. The results demonstrate that diffusion modal can enhance the robustness of the multimodal shared feature learned by the model, which can improve the performance of proposed method across various inference tasks.

Conclusion

In this article, we propose a local diffusion shared-specific autoencoder called LDS^{2}AE to address multimodal remote sensing image classification with arbitrary missing modalities. LDS^{2}AE designs a modality-shared encoder to learn multimodal shared feature by mapping multimodal data into a shared subspace, and several modality-specific decoders to facilitate the shared feature to learn unique information of different modalities by putting the multimodal shared feature to reconstruct the image of each modality. Moreover, LDS^{2}AE only performs diffusion on unmasked patches, which reduces the training overhead of the diffusion model. LDS^{2}AE facilitates the feature of available modalities in the fine-tuning stage to preserve the decision boundaries learned from multimodal features, effectively addressing feature heterogeneity resulting from input variations and structural distinctions between the missing modalities and multimodal model. In conclusion, we perform a detailed evaluation on three widely-used multimodal remote sensing datasets to demonstrate the effectiveness of our approach.

Acknowledgments

This work was supported in part by the the National Natural Science Foundation of China under Grant 62101414 and Grant 62201423, Young Talent Fund of Xi'an Association for Science and Technology under Grant 095920221320 and Grant 959202313052, the China Postdoctoral Science Special Foundation under Grant 2022T150508 and 2023T160502, the Youth Innovation Team of Shaanxi Universities, the Young Talent Fund of Association for Science and Technology in Shaanxi under Grant 20230117, and the China Postdoctoral Science Foundation under Grant 2021M702546 and 2021M702548.

References

Bao, H.; Dong, L.; and Wei, F. 2021. BEiT: BERT Pre-Training of Image Transformers. *ArXiv*, abs/2106.08254.

Chen, C.; Zhao, X.; Li, W.; Tao, R.; and Du, Q. 2019. Collaborative Classification of Hyperspectral and Lidar Data With Information Fusion and Deep Nets. In *IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium*, 2475–2478.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *ArXiv*, abs/1810.04805.

Dhariwal, P.; and Nichol, A. 2021. Diffusion Models Beat GANs on Image Synthesis. *ArXiv*, abs/2105.05233.

Ding, K.; Lu, T.; Fu, W.; Li, S.; and Ma, F. 2022. Global–Local Transformer Network for HSI and LiDAR Data Joint Classification. *IEEE Transactions on Geoscience and Remote Sensing*, 60: 1–13.

Dong, W.; Zhao, J.; Qu, J.; Xiao, S.; Li, N.; Hou, S.; and Li, Y. 2023. Abundance Matrix Correlation Analysis Network Based on Hierarchical Multihead Self-Cross-Hybrid Attention for Hyperspectral Change Detection. *IEEE Transactions on Geoscience and Remote Sensing*, 61: 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. 2020. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *ArXiv*, abs/2010.11929.

Dutt, A.; Zare, A.; and Gader, P. D. 2022. Shared Manifold Learning Using a Triplet Network for Multiple Sensor Translation and Fusion With Missing Data. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 15: 9439–9456.

Ghamisi, P.; Rasti, B.; Yokoya, N.; Wang, Q.; Hofle, B.; Bruzzone, L.; Bovolo, F.; Chi, M.; Anders, K.; Gloaguen, R.; Atkinson, P. M.; and Benediktsson, J. A. 2018. Multisource and Multitemporal Data Fusion in Remote Sensing: A Comprehensive Review of the State of the Art. *IEEE Geoscience and Remote Sensing Magazine*, 7: 6–39.

Hazarika, D.; Zimmermann, R.; and Poria, S. 2020. MISA: Modality-Invariant and -Specific Representations for Multimodal Sentiment Analysis. *Proceedings of the 28th ACM International Conference on Multimedia*.

He, K.; Chen, X.; Xie, S.; Li, Y.; Doll'ar, P.; and Girshick, R. B. 2021. Masked Autoencoders Are Scalable Vision Learners. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 15979–15988.

Holloway, J.; Helmstedt, K. J.; Mengersen, K. L.; and Schmidt, M. 2019. A Decision Tree Approach for Spatially Interpolating Missing Land Cover Data and Classifying Satellite Images. *Remote. Sens.*, 11: 1796.

Hong, D.; Gao, L.; Yokoya, N.; Yao, J.; Chanussot, J.; Du, Q.; and Zhang, B. 2021a. More Diverse Means Better: Multimodal Deep Learning Meets Remote-Sensing Imagery Classification. *IEEE Transactions on Geoscience and Remote Sensing*, 59(5): 4340–4354.

Hong, D.; Hu, J.; Yao, J.; Chanussot, J.; and Zhu, X. X. 2021b. Multimodal remote sensing benchmark datasets for land cover classification with a shared and specific feature learning model. *ISPRS Journal of Photogrammetry and Remote Sensing*, 178: 68–80.

Li, J.; Liu, Y.; Song, R.; Li, Y.; Han, K.; and Du, Q. 2023. 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.; Pei, Y.; Zhao, S.; Xiao, R.; Sang, X.; and Zhang, C. 2020. A Review of Remote Sensing for Environmental Monitoring in China. *Remote. Sens.*, 12: 1130.

Li, X.; Lei, L.; Sun, Y.; and Kuang, G. 2022. Dynamic-Hierarchical Attention Distillation With Synergetic Instance Selection for Land Cover Classification Using Missing Heterogeneity Images. *IEEE Transactions on Geoscience and Remote Sensing*, 60: 1–16.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *ArXiv*, abs/1907.11692.

Ma, M.; Ren, J.; Zhao, L.; Tulyakov, S.; Wu, C.; and Peng, X. 2021. SMIL: Multimodal Learning with Severely Missing Modality. In *AAAI Conference on Artificial Intelligence*.

Mei, S.; Yan, K.; Ma, M.; Chen, X.; Zhang, S.; and Du, Q. 2021. Remote Sensing Scene Classification Using Sparse Representation-Based Framework With Deep Feature Fusion. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14: 5867–5878.

Nichol, A.; and Dhariwal, P. 2021. Improved Denoising Diffusion Probabilistic Models. *ArXiv*, abs/2102.09672.

Park, W.; Kim, D.; Lu, Y.; and Cho, M. 2019. Relational Knowledge Distillation. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 3962–3971.

Qu, J.; Zhao, J.; Dong, W.; Xiao, S.; Li, Y.; and Du, Q. 2023. Feature Mutual Representation Based Graph Domain Adaptive Network for Unsupervised Hyperspectral Change Detection. *IEEE Transactions on Geoscience and Remote Sensing*, 1–1.

Radford, A.; and Narasimhan, K. 2018. Improving Language Understanding by Generative Pre-Training.

Rasti, B.; Ghamisi, P.; and Gloaguen, R. 2017a. Hyperspectral and LiDAR Fusion Using Extinction Profiles and Total Variation Component Analysis. *IEEE Transactions on Geoscience and Remote Sensing*, 55: 3997–4007.

Rasti, B.; Ghamisi, P.; and Gloaguen, R. 2017b. Hyperspectral and LiDAR Fusion Using Extinction Profiles and Total Variation Component Analysis. *IEEE Transactions on Geoscience and Remote Sensing*, 55(7): 3997–4007.

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.

Sohl-Dickstein, J. N.; Weiss, E. A.; Maheswaranathan, N.; and Ganguli, S. 2015. Deep Unsupervised Learning using Nonequilibrium Thermodynamics. *ArXiv*, abs/1503.03585.

Su, H.; Yu, Y.; Wu, Z.; and Du, Q. 2021. Random Subspace-Based k-Nearest Class Collaborative Representation for Hyperspectral Image Classification. *IEEE Transactions on Geoscience and Remote Sensing*, 59: 6840–6853.

Wang, Q.; Zhan, L.; Thompson, P. M.; and Zhou, J. 2020. Multimodal Learning with Incomplete Modalities by Knowledge Distillation. *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*.

Wei, S.; Luo, Y.; Ma, X.; Ren, P.; and Luo, C. 2023. 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.

Xie, Z.; Zhang, Z.; Cao, Y.; Lin, Y.; Bao, J.; Yao, Z.; Dai, Q.; and Hu, H. 2021. SimMIM: a Simple Framework for Masked Image Modeling. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 9643–9653.

Xu, Y.; Du, B.; and Zhang, L. 2018. Multi-Source Remote Sensing Data Classification via Fully Convolutional Networks and Post-Classification Processing. In *IGARSS 2018 - 2018 IEEE International Geoscience and Remote Sensing Symposium*, 3852–3855.

Xue, H.; Zhang, S.; and Cai, D. 2016. Depth Image Inpainting: Improving Low Rank Matrix Completion With Low Gradient Regularization. *IEEE Transactions on Image Processing*, 26: 4311–4320.

Yang, Y.; Fu, H.; Avilés-Rivero, A. I.; Schonlieb, C.-B.; and Zhu, L. 2023. DiffMIC: Dual-Guidance Diffusion Network for Medical Image Classification. *ArXiv*, abs/2303.10610.

Yao, J.; Hong, D.; Gao, L.; and Chanussot, J. 2022. Multimodal Remote Sensing Benchmark Datasets for Land Cover Classification. In *IGARSS 2022 - 2022 IEEE International Geoscience and Remote Sensing Symposium*, 4807–4810.

Zhang, C.; Cui, Y.; Han, Z.; Zhou, J. T.; Fu, H.; and Hu, Q. 2020a. Deep Partial Multi-View Learning. arXiv:2011.06170.

Zhang, C.; Fu, H.; Hu, Q.; Cao, X.; Xie, Y.; Tao, D.; and Xu, D. 2020b. Generalized Latent Multi-View Subspace Clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(1): 86–99.

Zhang, Q.; Yuan, Q.; Zeng, C.; Li, X.; and Wei, Y. 2018. Missing Data Reconstruction in Remote Sensing Image With a Unified Spatial–Temporal–Spectral Deep Convolutional Neural Network. *IEEE Transactions on Geoscience and Remote Sensing*, 56: 4274–4288.

Zhao, X.; Tao, R.; Li, W.; Li, H.-C.; Du, Q.; Liao, W.; and Philips, W. 2020. Joint Classification of Hyperspectral and LiDAR Data Using Hierarchical Random Walk and Deep CNN Architecture. *IEEE Transactions on Geoscience and Remote Sensing*, 58(10): 7355–7370.

Zhou, J.; Sheng, J.; Fan, J.; Ye, P.; He, T.; Wang, B.; and Chen, T. 2023. When Hyperspectral Image Classification Meets Diffusion Models: An Unsupervised Feature Learning Framework. *ArXiv*, abs/2306.08964.