Fusion from a Distributional Perspective: A Unified Symbiotic Diffusion Framework for Any Multisource Remote Sensing Data Classification

Teng Yang[1], **Song Xiao**[1,2,*], **Wenqian Dong**[1], **Jiahui Qu**[1] and **Yueguang Yang**[1]
[1]The State Key Laboratory of Integrated Service Network, Xidian University
[2]Beijing Electronic Science and Technology Institute
yangteng@stu.xidian.edu.cn, xiaosong@mail.xidian.edu.cn, wqdong@xidian.edu.cn, jhqu@xidian.edu.cn, yueguangyang001@gmail.com

Abstract

The joint classification of multisource remote sensing data is a prominent research field. However, most of the existing works are tailored for two specific data sources, which fail to effectively address the diverse combinations of data sources in practical applications. The importance of designing a unified network with applicability has been disregarded. In this paper, we propose a unified and self-supervised Symbiotic Diffusion framework (named SymDiffuser), which achieves the joint classification of any pair of different remote sensing data sources in a single model. The SymDiffuser captures the inter-modal relationship through establishing reciprocal conditional distributions across diverse sources step by step. The fusion process of multisource data is consistently represented within the framework from a data distribution perspective. Subsequently, features under the current conditional distribution at each time step is integrated during the downstream phase to accomplish the classification task. Such joint classification methodology transcends source-specific considerations, rendering it applicable to remote sensing data from any diverse sources. The experimental results showcase the framework's potential in achieving state-of-the-art performance in multimodal fusion classification task.

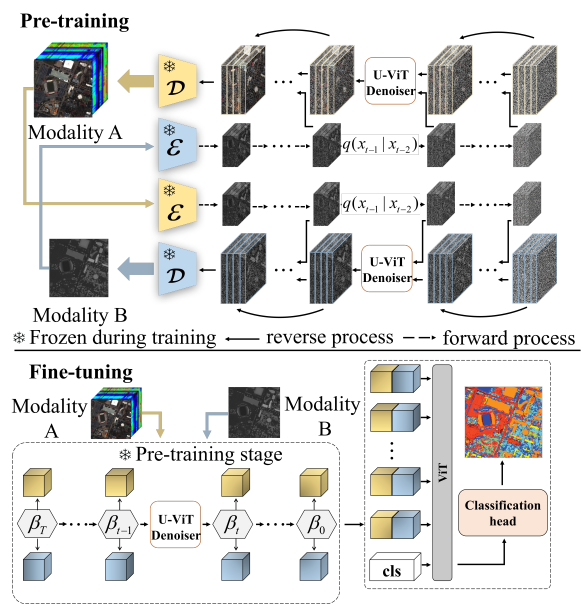

Figure 1: Overview of our proposed SymDiffuser. **(Up) Pre-training**: Different modalities of RS data are projected into the latent space through frozen encoder, and their mutual conditional distributions are modeled by a Coupled Symbiotic Diffusion Model (CSDiff) to capture the relationship between modalities. **(Down) Fine-tuning**: Downstream task fuses the features of each time step of CSDiff for fine-tuning to complete the joint classification.

1 Introduction

The classification of remote sensing (RS) images has become increasingly significant in various domains such as urban planning [Dong *et al.*, 2022], military applications [Ding *et al.*, 2022a], environmental monitoring [Hu *et al.*, 2023], and agricultural production [Meng *et al.*, 2022]. Due to the limitations of sensor technology, a single type of RS image cannot meet the requirements of classification for diverse tasks. Consequently, the integration of multisource modalities has attracted increasing research attention. Commonly utilized types of RS images include hyperspectral image (HSI) [Dong *et al.*, 2023], which provide rich spectral information, Synthetic Aperture Radar (SAR) images [Wang *et al.*, 2022a], which contain amplitude and phase information, and Light Detection and Ranging (LiDAR) data [Xue *et al.*, 2022], which reflects the precise height information. Relevant work has provided evidence of the advantages of utilizing multisource data in classification task.

The key point of joint classification of multisource data lies in effectively fusing information from diverse sources and harnessing their complementary attributes to enhance the expressive capacity of the model. Nevertheless, existing multisource fusion methods are typically tailored to two specific RS data sources, with the network architecture designed to accommodate their unique characteristics. The lack of generalizability in these data-specific networks become apparent

*Corresponding author

when facing with diverse types of data. In addition, the acquisition of labels for RS images is a labor-intensive task due to their coverage of vast geographical areas. Therefore, there exists a significant quantitative disparity between the labeled and unlabeled data. The exploration of unlabeled multisource data correlations in supervised joint classification methods remain largely unexplored. Self-supervised based methods in computer vision have been demonstrated the effectiveness of acquiring feature representations from unlabeled data. However, most of self-supervised based works solely focusing on visual features of a single modality. In contrast, the effective integration of information from multiple modalities should be primarily emphasized in the joint classification of multisource data.

To overcome this dilemma, as depicted in Fig. 1, we propose a unified and self-supervised symbiotic diffusion framework, which is capable of fusing any two RS modality data for classification. We point out that the process of multimodal fusion can be conceptualized as the process of capturing the reciprocal conditional distribution between two modalities. In this way, the relationships and dependencies among modalities can be deeply explored and modeled, which is helpful for an efficient and accurate fusion. Specifically, the reciprocal conditional distributions between different modalities are gradually captured by a coupled symbiotic diffusion model. The feature at each step in the two Markov chains defined by the diffusion steps is used as input for downstream tasks. At each time step of the reverse process of diffusion, a Modality Perception Block (MPB) is employed to enhance the correspondence between different modalities under the same degradation. Specially, textual information is leveraged as weak supervision to guide the fusion feature towards the classification task. Our proposed framework is specifically designed to extract fusion feature from multimodal RS data, offering a promising direction for the development of integrated multimodal remote sensing data fusion classification approach. Our major contributions are as follows:

1) We propose a unified symbiotic diffusion framework for any multisource RS data classification, which learns the conditional distribution between the two modalities by a coupled symbiotic diffusion model, guiding the fusion process and facilitating a more effective combination of information from each modality.

2) To effectively enhance the correlation between modalities, we propose a modality perception block (MPB). MPB helps the exploration of interactions between different modalities, thereby enhancing the accuracy of modeling the conditional distribution.

3) We introduced a task-oriented condition injection (TCI) module to inject task-related prior knowledge into the model. It ensures that the model gains a more informed understanding of the data, leading to improved adaptability in downstream tasks.

2 Related Work

2.1 Multisource Fusion Classification in RS

Deep-learning based multisource RS data classification methods have been widely studied in recent years. In an early attempt [Chen *et al.*, 2017], two independent Convolutional Neural Networks (CNNs) were proposed to extract the features of HSI and LiDAR respectively. To capture more specific information from different sources, some studies [Xu *et al.*, 2018; Zhao *et al.*, 2020] utilize two network branches to extract both spectral and spatial features from HSI, while focusing only on spatial feature for LiDAR. Such schemes simply cascade the features of different sources and fail to fully exploit the interrelationship between different modalities. To address this issue, Zhang et al. [Zhang *et al.*, 2020] attempted to extract fusion feature in hidden layers through encoding and decoding processes. Hong et al. [Hong *et al.*, 2021] investigated fusion strategies at various stages, which can be applicable to pixel-based and spatial–spectral classification respectively. Subsequently, attention mechanisms have been introduced to guide the complementary integration of multisource features [Mohla *et al.*, 2020; Xiu *et al.*, 2022; Li *et al.*, 2022a]. Li et al. [Li *et al.*, 2022a] proposed A^3CLNN, which incorporates a composite attention mechanism to fuse enhanced feature representations from both spatial and spectral domains in HSI and LiDAR data. Another group of works [Mohla *et al.*, 2020] explored semantic correlations between different source features by designing cross-attention mechanisms. Xue et al. [Xue *et al.*, 2022] developed a DHViT structure, employing cross-attention feature fusion pattern to fuse heterogeneous features from multi-modality data adaptively. Decision-level fusion was then employed to integrate these heterogeneous features. Methods such as NNC [Wang *et al.*, 2023] and CCL [Jia *et al.*, 2023] utilized contrastive learning to reduce the heterogeneity between different modalities. More recently, Huo et al. [Huo *et al.*, 2023] proposed a fusion method that captures the relationship between different perspectives within each modality at the patch level, which achieves superior results. However, the above methods utilize two networks tailored to two specific RS data sources to extract features , yet these methods do not fully explore the inter-modal relationships, thereby failing to capture the intricate structures and correlations within multisource RS data. Such deficiency leads to the modal with suboptimal utilization of multisource RS data.

2.2 Denoising Diffusion Probabilistic Model

The denoising diffusion probability model (DDPM) has emerged as a mainstream generation model, garnering notable achievements in various tasks such as image restoration [Zhu *et al.*, 2023; Wang *et al.*, 2022b; Luo *et al.*, 2023], image super-resolution [Metzger *et al.*, 2023; Rombach *et al.*, 2022], style transfer [Li *et al.*, 2023a; Zhang *et al.*, 2023], and text-to-image generation [Xu *et al.*, 2023a; Kumari *et al.*, 2023]. Compared to the Generative Adversarial Network (GAN), which aims to achieve feature space transformation through a direct forward process, DDPM decomposes the generation process into multiple denoising steps with well-defined data distribution. This decomposition approach provides enhanced stability in generation while offering effective controllability and flexibility. The controlled sampling process of DDPM has facilitated the design of numerous conditional diffusion models. Gao et al. [Gao *et al.*, 2023] developed a low-resolution (LR) conditioning network

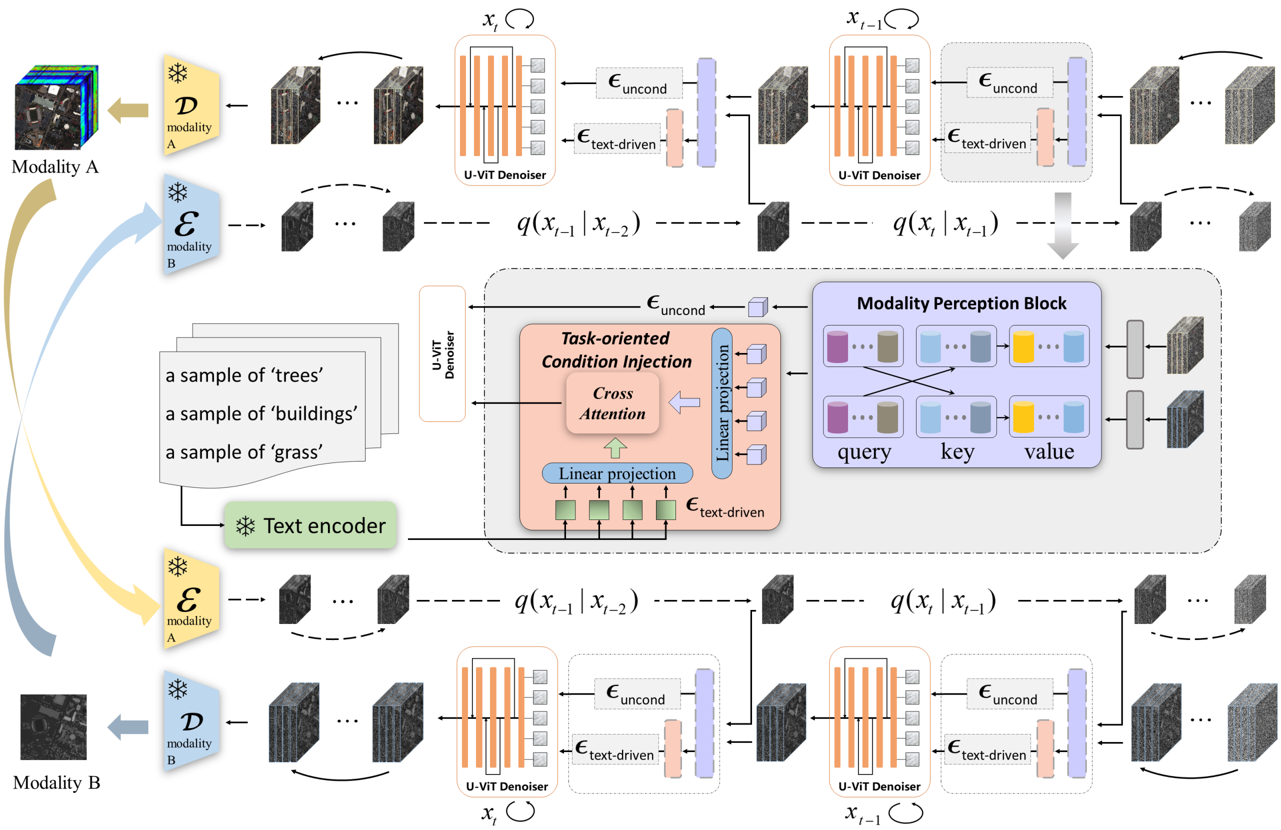

Figure 2: Overview of the Pre-training stage (CSDiff): The CSDiff framework incorporates MPB and TCI (represented by purple and pink blocks respectively) at each denoising step. In the MPB module, the forward degradation features of one modality are fused with the prediction features of another modality and fed into the U-Vit-based denoising network for subsequent time-step noise prediction. TCI is utilized to inject textual information into the model, enabling feature attention towards downstream classification tasks.

to encode the image without priors, and then combined it with a scaling factor in DDPM to facilitate continuous image super-resolution. Kawar et al. [Kawar *et al.*, 2023] introduced imagic for non-rigid edits on a single image, which employs natural language text as a prompt to pass the semantically meaningful mixture of image embedding and target text embedding to DDPM. Recent studies have started to employ the diffusion model to multimodal generation. MM-diffusion [Ruan *et al.*, 2023] used two coupled diffusion models to learn the joint distribution of audio and video for the generation of aligned audio-video pairs. The proposed SymDiffuser distinguishes itself from numerous existing diffusion models by placing emphasis on the modeling of the conditional distribution between two modes, thereby acquiring more effective modality fusion information for classification task.

3 Method

3.1 Motivation and Overview

In the joint classification of multisource RS data, the purpose of multimodal fusion is to combine information from different modalities and leverage the complementary characteristics between them to enhance the model's comprehension of classification task. During multimodal fusion, the exchange of pertinent information between modalities needs to be facilitated, which requires the model to discern the relationship between each modality and how they can mutually reinforce or complement one another. This relationship can be efficiently represented by a conditional distribution. Conditional distributions illustrate how the probability of one variable changes based on the observation of another variable, aiding in the comprehension of the influence of information between different modalities. Based on this fact, we adopt DDPM to model the conditional distribution between different modalities of RS data, with the aim of improving the mutual understanding between the two modalities and facilitating a more effective and balanced combination of modality during the fusion process.

Our proposed SymDiffuser is a unified framework for joint classification of multisource RS data, as shown in Fig. 1. it consists of two stages: pre-training and downstream finetuning. The high-level framework of pre-training incorporates a coupled symbiotic diffusion model, where we independently input the latent features of two modalities and perform a two-way modality conditional distribution modeling. The degraded feature of one modality and the corresponding diffusion feature from the other modality are correlated using a Modality Perception Block (MPB) during each reverse diffusion step. Subsequently, U-ViT based denoising networks are used for single-step modal transformation completion. We utilize the text captions produced by the supervised labeling process to train a Task-oriented Condition Injection module (TIC). It leverages the diffusion features from each step and the text encoder embedding of category names to provide guidance for the fusion process in classification task. Once pre-trained, we perform multimodal RS images classification with features from each time step in the diffusion models.

3.2 Preliminaries of Vanilla Diffusion

The diffusion model is a type of generative model which utilizes multi-step denoising to predict specific data distributions from random noise [Ho *et al.*, 2020]. It consists of two fundamental components: the forward process and the reverse process. In the forward process, multi-step Gaussian noise is incrementally added to data that needs to be predicted. This iterative process persists until the data distribution degenerates into a state of approximately pure Gaussian noise. Assume X as the data distribution to be predicted, x_0 as a sample in X, and x_t as the noise sample at time step t. The forward process of diffusion model can be expressed as:

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t\mathbf{I})$$
$$q(x_{1:T}|x_0) = \prod_{t=1}^{T} q(x_t|x_{t-1}) \tag{1}$$

where $t \in [1, T]$, and $\beta_0, \beta_1, ..., \beta_T$ represent a pre-defined variance schedule.

The reverse process of the diffusion model is characterized as a Markov chain that aims to recover a specific data distribution through multi-step denoising networks. At each time step t, the reverse process is defined as $p_\theta(x_{t-1}|x_t)$, where θ is typically implemented as a denoising network. By following this Markov process, the model progressively denoises the input x_t to ultimately obtain the final result. The reverse process of diffusion model can be expressed as:

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t))$$
$$p_\theta(x_{0:T}) = p(x_T)\prod_{t=1}^{T} p_\theta(x_{t-1}|x_t) \tag{2}$$

where μ_θ represents the predicted mean value by the network θ. The vanilla diffusion model can be expanded to encompass the modeling of the conditional distribution $p_\theta(x_0|c)$ by introducing conditions into the reverse process. Here, c represents a data distribution distinct from x_0, such as images or textual information with varying styles.

In our method, a coupled symbiotic conditional diffusion model is designed to promote the fusion of different RS modalities by delving into their mutual conditional distributions.

3.3 Coupled Symbiotic Diffusion Model

As mentioned in the motivation, we suggest to model the conditional distribution between different modalities to facilitate multimodal fusion. It is essential to consider the bidirectional relationship of two modalities from one to another when modeling the conditional distribution. Assume that there are two different types of RS data: modality $\mathcal{A}$ and modality $\mathcal{B}$. We design a coupled symbiotic diffusion model (CSDiff) to capture the conditional distributions $p(\mathcal{A}|\mathcal{B})$ and $p(\mathcal{B}|\mathcal{A})$ simultaneously. To enable the model to effectively handle complex data distributions, similar as Stable Diffusion [Rombach *et al.*, 2022] and UniDiffuser [Bao *et al.*, 2023b], the data of both modalities are encoded into latent space by two frozen ViT-based encoders $\varepsilon_{\mathcal{A}}$ and $\varepsilon_{\mathcal{B}}$, respectively. The latent features x_{A_0} and x_{B_0} obtained separately by $\varepsilon_{\mathcal{A}}$ and $\varepsilon_{\mathcal{B}}$ of the two modalities are subsequently degraded using two independent forward processes. Taking modality $\mathcal{A}$ as an example, its forward process at timestep t can be expressed as follows:

$$q(x_{A_t}|x_{A_{t-1}}) = \mathcal{N}(x_{A_t}; \sqrt{1-\beta_t}x_{A_{t-1}}, \beta_t\mathbf{I}) \tag{3}$$

where the schedule sequence β is shared of two modalities. The forward process for modality $\mathcal{B}$ is same as modality $\mathcal{A}$. Different from the forward process of the two modalities, the reverse process of the diffusion model in the generation task cannot be directly applied to CSDiffer, as it presents two key challenges. i) The general conditional diffusion model utilizes fixed conditional information and degenerate features from the current step for reconstruction at each step. However, this approach overlooks the correspondence between data from different modalities under the same degradation level, which is essential for multimodal fusion task. ii) In the context of complex multimodal data, the conventional cascade of conditions typically employed in conditional diffusion models faces challenges in capturing the intricate modal relationships. To tackle these problems, we redefine the reverse process in CSDiffer, which can be expressed as follows:

$$p_{\mathscr{F}_{\mathcal{A}}}(x^r_{\mathcal{A}_{t-1}}|(x^r_{\mathcal{A}_t}, x^f_{\mathcal{B}_t})) = \mathcal{N}(x^r_{\mathcal{A}_t}; \mathscr{F}(x^r_{\mathcal{A}_t}, x^f_{\mathcal{B}_t}, t))$$
$$p_{\mathscr{F}_{\mathcal{B}}}(x^r_{\mathcal{B}_{t-1}}|(x^f_{\mathcal{A}_t}, x^r_{\mathcal{B}_t})) = \mathcal{N}(x^r_{\mathcal{B}_t}; \mathscr{F}(x^f_{\mathcal{A}_t}, x^r_{\mathcal{B}_t}, t)) \tag{4}$$

where $x^f_{\mathcal{A}_t}$ denotes the feature of modality $\mathcal{A}$ at time step t in the forward process, $x^r_{\mathcal{A}_t}$ is the feature in the reverse process, which is same for modality $\mathcal{B}$. As depicted in Eq. (4), the output of the t-th step for each mode is computed based on its own degradation feature, the time step t, and the forward degradation feature of the corresponding t-th step in the other modality. This design is centered around the concept that the features of both modes at each time step are interdependent under the same level of degradation conditions. Consequently, this coupling strengthens the interconnectedness between the different levels of multimodal features. To train the denoising process, the CSDiff predict the added noise using the standard MSE loss:

$$L_{\text{mse}_{\mathcal{A}}} = \mathbb{E}_{t\sim[1,T]}\left\|\epsilon - \mathscr{F}_{\mathcal{A}}(x^r_{\mathcal{A}_t}, x^f_{\mathcal{B}_t}, t)\right\|^2$$
$$L_{\text{mse}_{\mathcal{B}}} = \mathbb{E}_{t\sim[1,T]}\left\|\epsilon - \mathscr{F}_{\mathcal{B}}(x^f_{\mathcal{A}_t}, x^r_{\mathcal{B}_t}, t)\right\|^2 \tag{5}$$

3.4 Modality Perception Block

To facilitate the alignment between two modalities in noise prediction and learn their correlation, a cross-attention based MPB is employed. MPB is embedded before each step of the denoising network, which aims to capture and strengthen the correlations between two modalities within the fused feature representation. Taking modality $\mathcal{A}$ in CSDiffer as an example, the cross-attention mechanism can be formulated as follows:

$$\begin{gathered}\boldsymbol{Q}=\phi_q(x^f_{\mathcal{B}_t}),\quad \boldsymbol{K}=\phi_k(x^r_{\mathcal{A}_t}),\quad \boldsymbol{V}=\phi_v(x^r_{\mathcal{A}_t})\\ \boldsymbol{F}_{att}=\frac{\boldsymbol{Q}\boldsymbol{K}^T}{\sqrt{C}},\quad \boldsymbol{F}_{fuse}=\boldsymbol{W}\text{softmax}(\boldsymbol{F}_{att})\boldsymbol{V}+x^r_{\mathcal{A}_t}\end{gathered} \tag{6}$$

where C is the dimension of feature, ϕ_q, ϕ_k, ϕ_v are Multilayer Perceptron (MLP) and $\boldsymbol{W}$ refers to learnable weights to generate final fused features $\boldsymbol{F}_{fuse}$. Once $\boldsymbol{F}_{fuse}$ is obtained, it is treated as the input to the denoising network. The U-ViT based network [Bao *et al.*, 2023a] is employed as the denoising network in each step, leveraging the significant advantages of ViT for multimodal data processing. In U-Vit, time steps t and fused feature $\boldsymbol{F}_{fused}$ are treated as tokens, and long-range skip connections are employed between shallow and deep layers. The reverse process of CSDiffer, represented by Eq. (4), can be further refactored as follows:

$$\begin{gathered}p_{\mathscr{F}_\mathcal{A}}(x^r_{\mathcal{A}_{t-1}}|(x^r_{\mathcal{A}_t},x^f_{\mathcal{B}_t}))=\mathcal{N}(\mathscr{F}_\mathcal{A}(\boldsymbol{F}^\mathcal{A}_{fuse},t))\\ p_{\mathscr{F}_\mathcal{B}}(x^r_{\mathcal{B}_{t-1}}|(x^f_{\mathcal{A}_t},x^r_{\mathcal{B}_t}))=\mathcal{N}(\mathscr{F}_\mathcal{B}(\boldsymbol{F}^\mathcal{B}_{fuse},t))\end{gathered} \tag{7}$$

where $\mathscr{F}_\mathcal{A}$ and $\mathscr{F}_\mathcal{B}$ are U-ViT based denoising networks in tth step for modality $\mathcal{A}$ and modality $\mathcal{B}$, respectively. After the last step of the reverse process, the features of two modalities are inputted into their corresponding frozen ViT-based decoder $\mathcal{D}_\mathcal{A}$ and $\mathcal{D}_\mathcal{B}$ to accomplish reconstruction.

After the conditional distributions $p(\mathcal{A}|\mathcal{B})$ and $p(\mathcal{B}|\mathcal{A})$ modeled by CSDiffer, in the subsequent downstream task, the features at each time step t are utilized to accomplish the fusion classification task involving multiple modalities. The comprehensive details of this process will be elaborated upon in the last subsection.

3.5 Task-oriented Condition Injection

In order to align the focus of the integrated features in pretrained models with downstream classification task, a Task-oriented Conditional Injection (TCI) module is introduced. Textual information about categories typically contains semantic details regarding the respective categories. By injecting this information into the reverse process, the model can benefit from richer semantic guidance, thus enhancing its ability to comprehend and leverage semantic differences among different categories. This facilitates the inclusion of more discriminative information in features during the pretraining phase, enabling the model to adapt to classification task. Hence, we advocate for the integration of visual features with text that carries rich semantic meaning. We generate corresponding text prompts for each category and serve these prompts as conditions for injecting the U-ViT denoising network in the reverse process of the diffusion model.

Assuming that the input multimodal data belongs to a known category Cls, where Cls represents the category name. We set "a sample of [Cls]" as the text prompt. The frozen CLIP text encoder $\mathcal{T}$ is utilized to extract the embedding of the text prompt. This text embedding, along with the fusion feature obtained from Eq. (6), is then passed through a cross-attention module to enhance the correlation between the diffusion feature and high-level semantics. The process of task-oriented condition injection can be formulated as follows:

$$\boldsymbol{\epsilon}_{\text{text-driven}}=\boldsymbol{\epsilon}_\theta(\boldsymbol{F}_{fuse},\mathcal{T}(Cls)) \tag{8}$$

where $\boldsymbol{\epsilon}_\theta$ is a cross-attention block. Significantly, the network performs semantic injection solely on labeled data, while unlabeled data undergoes the denoising process directly in the reverse process.

3.6 Multi-step Feature Fusion

Once the pre-training phase is completed, we input two different modalities of remote sensing data into the proposed framework during the downstream classification task. This process allows us to obtain fusion features at different levels. The continuous generation steps of the diffusion model enable a gradual conversion between different modalities. As a result, each time step contains varying degrees of modal fusion information. Based on this observation, we extract the fusion features from each time step within the co-occurrence diffusion model. These features capture the diverse levels of fusion information. We then combine these different levels of features and feed them into the classification head. The network architecture employed for processing the fusion features uses ViT. Each time step is encoded and embedded to provide specific time information for each step within the network. Finally, the fused feature representation is passed through a simple MLP to output the classification results.

4 Experimental

4.1 Datasets

We validate SymDiffuser on three real multi-source remote sensing datasets. To explore its potential in more types of remote sensing data fusion classification, we construct simulated multi-source remote sensing data on two additional datasets for further validation.

HSI-MSI *2012Houston data*: The 2012Houston dataset comprises HSI and MSI data over the University of Houston campus and the surrounding urban regions. This image consists of 349 × 1905 pixels and encompasses 15 different categories. HSI was collected by CASI-1500, encompasses 144 spectral bands ranging from 380 nm to 1050 nm. MSI is composed of the same size with HSI but 8 spectral bands.

HSI-LiDAR *MUUFL data*: The MUUFL dataset contains registered HSI and LiDAR-based DSM over the University of Southern Mississippi Gulf Park Campus. The spatial size of this data is 325 × 220 pixels, with the spatial resolution in 1 m and 11 categories. HSI consists 64 bands ranging from 375 nm to 1050 nm at a spectral sampling of 10 nm. The LiDAR data has the same spatial size and resolution.

HSI-SAR *Augsburg data*: The Augsburg dataset consists of HSI and PolSAR image, over the city of Augsburg, Germany. The scene comprises a total of 332 × 485 pixels and encompasses a spectral range spanning from 0.4μm to 2.5μm, consisting of 180 spectral bands for HSI and 7 categories. Additionally, the dual-Pol (VV-VH) SAR image contributes with four distinct features, namely VV intensity, VH

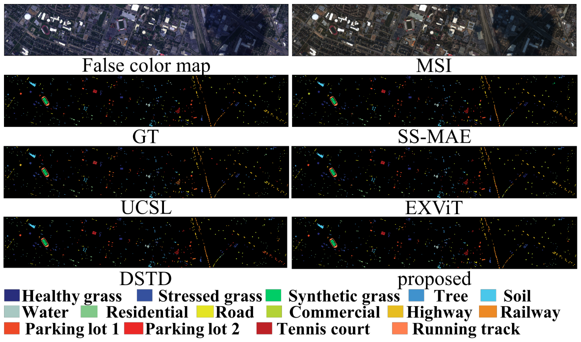

Figure 3: Classification maps of the 2012Houston dataset.

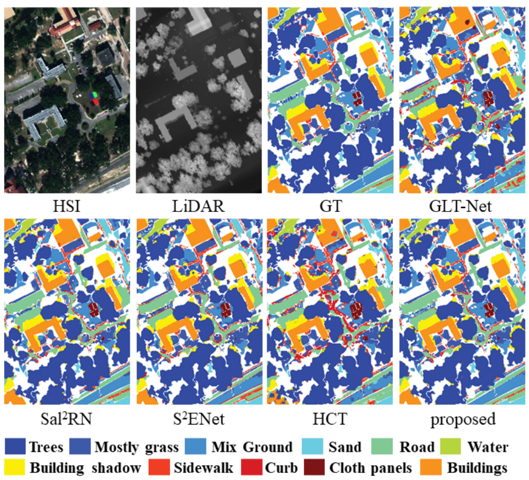

Figure 4: Classification maps of the MUUFL dataset.

Method \ Index	OA(%)	AA(%)	$\kappa \times 100$
2012houston HSI + MSI			
SS-MAE [Lin *et al.*, 2023]	85.25	87.85	84.04
UCSL [Yao *et al.*, 2023a]	79.73	79.95	78.10
ExVit [Yao *et al.*, 2023b]	88.18	90.15	87.23
DSTD [Xu *et al.*, 2023b]	87.61	89.86	86.62
Proposed	**91.63**	**92.89**	**90.95**
MUUFL HSI + LiDAR			
GLT-Net [Ding *et al.*, 2022b]	81.50	82.99	76.39
Sal^{2}RN [Li *et al.*, 2022b]	88.50	90.02	85.14
S^{2}ENet [Fang *et al.*, 2021]	86.68	86.53	82.73
HCT [Zhao *et al.*, 2022]	80.60	80.72	75.10
Proposed	**90.42**	**93.03**	**87.63**
Augsburg HSI + SAR			
CCRNET [Wu *et al.*, 2021]	73.07	65.26	64.16
MFT [Roy *et al.*, 2023]	86.10	78.95	81.57
MACN [Li *et al.*, 2023b]	89.02	81.59	84.73
SepDGConv [Yang *et al.*, 2022]	85.96	82.95	80.74
Proposed	**91.20**	**86.16**	**87.74**

Table 1: Classification accuracy of different methods for three datasets

	OA(%)	AA(%)	$\kappa \times 100$	OA(%)	AA(%)	$\kappa \times 100$
	Trento PAN + LiDAR			Trento RGB + LiDAR		
GLT-Net	98.34	97.89	97.79	98.23	97.35	97.64
Sal2RN	97.91	95.92	97.22	98.83	98.11	98.45
S2ENet	97.93	97.36	97.25	98.30	98.12	97.74
HCT	96.73	95.99	95.64	97.42	97.44	96.57
Proposed	**98.89**	**97.97**	**98.41**	**99.01**	**98.52**	**98.81**
	Berlin PAN + SAR			Berlin RGB + SAR		
CCRNet	61.84	55.71	47.12	64.08	57.45	49.72
MFT	63.62	64.05	50.30	64.93	66.05	52.00
MACN	59.59	54.76	44.56	59.46	62.84	46.66
SepDGConv	60.79	58.81	46.73	61.72	62.67	48.70
Proposed	**64.30**	**68.66**	**52.54**	**67.24**	**71.56**	**55.69**

Table 2: Classification accuracy of different methods for other data combinations

intensity, the real part, and the imaginary part of the off-diagonal element within the PolSAR covariance matrix.

More data combinations *Trento and Berlin data*: The Trento dataset comprises HSI and LiDAR data, while the Berlin dataset consists of HSI and SAR image. Utilizing the HSI data from these datasets, we generated PAN, RGB, and MSI. Subsequently, we conducted various combinations of these derived data with the original LiDAR and SAR data to validate the applicability of our approach across diverse types of data combinations.

4.2 Comparison with State-of-the-Art Methods

In order to verify the effectiveness of the proposed network in the joint classification of multisource RS data, it is comprehensively compared with the state-of-the-art DL-based methods. Specifically, competing methods designed specifically for different modalities were selected on three datasets. For HSI-MSI dataset, SS-MAE [Lin *et al.*, 2023], UCSL [Yao *et al.*, 2023a], ExVit [Yao *et al.*, 2023b], and DSTD [Xu *et al.*, 2023b] are employed. For HSI-LiDAR dataset, GLT-Net [Ding *et al.*, 2022b], Sal^{2}RN [Li *et al.*, 2022b], S^{2}ENet [Fang *et al.*, 2021], and HCT [Zhao *et al.*, 2022] are employed. For HSI-SAR dataset, CCRNET [Wu *et al.*, 2021], MFT [Roy *et al.*, 2023], MACN [Li *et al.*, 2023b], and SepDGConv [Yang *et al.*, 2022] are employed. Given the lack of joint classification research on PAN+LiDAR, RGB+LiDAR, PAN+SAR, and RGB+SAR data combinations, the methods for HSI+LiDAR is applied to combinations with LiDAR data, and the methods for HSI+SAR is applied to combinations with SAR data. For fair comparisons, we use the original code provided by the author.

The quantitative results of the proposed method and competing methods on three different multisource datasets are presented in Table 1. By leveraging the data correlation across different modalities, it is evident that our proposed method consistently achieves the highest accuracy on all datasets in terms of overall accuracy (OA), average accuracy (AA), and Kappa coefficient. For example, our overall accuracy achieved on the HSI-MSI dataset is 91.63%, representing a significant improvement of 6.38% over SS-MAE, 3.45% over ExVit, and 4.02% over DSTD. The classification results of the proposed method on the three datasets are illustrated in Fig. 3, Fig. 4 and Fig. 5, respectively, highlighting its superiority compared to other methods. In comparison to other methods, the proposed method yields qualitative results that are more closely aligned with ground truth. For instance, it is a challenge to distinguish "healthy grass" and "stressed grass" in 2012Houston dataset. As shown in Fig. 3, compared with other methods, the classification results of the proposed method on 2012Houston dataset show that different

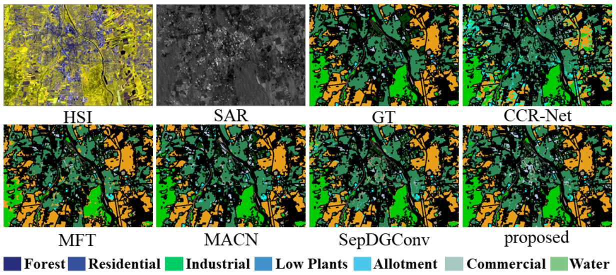

Figure 5: Classification maps of the Augsburg dataset.

Dataset(2012Houston)	OA(%)	AA(%)	$\kappa \times 100$
HSI+MSI+LiDAR	93.69	94.51	93.18

Table 3: Classification accuracy of dataset with three modals

grassland types are successfully distinguished. Furthermore, as depicted in Fig. 4, our method demonstrates fewer misclassification within the "road" category. Both quantitative and qualitative evaluations unequivocally establish the superiority of our approach.

Specially, the proposed SymDiffuser capture the intermodal relationship through establishing conditional distribution for fusion. SymDiffuser is applicable to any modal data and can naturally extend to accommodate any number of modal inputs by establishing a chained conditional distribution. The classification result of the extended scheme, which fuses three different modal data, are presented in Table 3.

4.3 Ablation Study

To verify the effectiveness of different metrics in our method, we conduct ablation studies with four variants of SymDiffuser. The details are as follows.

Variant-1

In Variant-1, CSDiff is replaced by VAE. CSDiff is used to explore the interdependence between different modal data, so that the model can enhance the understanding of modal relationships to effectively fuse information from different sources. The data from different sources are fed into two frozen mode-specific encoders $\varepsilon_{\mathcal{A}}$ and $\varepsilon_{\mathcal{B}}$, and the extracted latent space features are cascaded for fusion before classification. As shown in Table 4, the use of CSDiff increases by 5.91% on OA in 2012Houston dataset compared to using the cascaded way to fuse features for classification.

Variant-2

To further validate the effectiveness of the proposed CSDiff, the MPB and TIC modules are removed in Variant-2, and the experimental results are shown in the second row of Table 4. Compared to the result by using VAE shown in the first row, employing CSDiff without including MPB and TIC has improvements of 4.31%, 6.01%, and 4.35% in the OA on the three datasets.

Variant-3

The mode-aware block uses the fusion features of the two modalities to predict the noise, which enhances the correlation of the two modalities in the process of establishing the

	2012houston	MUUFL	Augsburg
Variant - 1 (CSDiff replaced)	85.72	84.41	86.85
Variant - 2 (w/CSDiff only)	90.03	87.96	88.25
Variant - 3a (w/o MPB)	89.67	88.25	89.14
Variant - 3b (MPB replaced)	89.51	88.93	88.82
Variant - 4 (w/o TIC)	91.18	88.84	89.39
Proposed	**91.63**	**90.42**	**91.20**

Table 4: Effect of different components

conditional distribution of each other. We remove the MPB in the model to verify its effectiveness in Variant-3a. Given that the primary objective of MPB is to enhance modal correlation, we adopted a direct strategy in Variant-3b to replace MPB: aligning and concatenating features from modality $\mathcal{A}$ and modality $\mathcal{B}$, subsequently inputting the merged features into the denoising network. From Table 4, we can find that compared with not using MPB, model with MPB has an average improvement of 2.06% in the three datasets.

Variant-4

The Task-oriented Conditional Injection module (TCI) seeks to associate the visual features of each category with their corresponding textual descriptions by integrating textual information into CSDiff. TCI facilitates the enrichment of features with more discriminative information during the pre-training stage, enhancing adaptation to multi-source classification tasks. In Varinat-4, we remove TCI to verify its impact on classification accuracy. As can be seen from Table 4, the model adopting the task guidance module has 0.45%, 1.58% and 1.81% improvements in the OA of the three datasets.

5 Conclusion

We proposed SymDiffuser, a unified framework for any multimodal RS data classification. By modeling the conditional distribution, SymDiffuser effectively captures the interdependence and interaction between different modalities. This profound understanding of the modality relationship enables the efficient integration and fusion of information from diverse sources. To further enhance the correlation between modalities and improve the model's ability to comprehend downstream classification task, we introduce the modality perception block and task-oriented conditional injection module. By offering insights into the modal relationship and fusion process, CSDiff Pre-training opens up new avenues for advancing the field of multimodal data fusion and classification in RS applications.

Acknowledgements

This work was supported in part by the 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 Young Talent Fund of Association for Science and Technology in Shaanxi under Grant 20230117, the National Natural Science Foundation of China under Grant 62101414 and Grant 62201423, and the China Postdoctoral Science Foundation under Grant 2021M702546 and 2021M702548.

References

[Bao *et al.*, 2023a] Fan Bao, Shen Nie, Kaiwen Xue, Yue Cao, Chongxuan Li, Hang Su, and Jun Zhu. All are worth words: A vit backbone for diffusion models. In *IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 22669–22679, 2023.

[Bao *et al.*, 2023b] Fan Bao, Shen Nie, Kaiwen Xue, Chongxuan Li, Shi Pu, Yaole Wang, Gang Yue, Yue Cao, Hang Su, and Jun Zhu. One transformer fits all distributions in multi-modal diffusion at scale. *arXiv preprint arXiv:2303.06555*, 2023.

[Chen *et al.*, 2017] Yushi Chen, Chunyang Li, Pedram Ghamisi, Xiuping Jia, and Yanfeng Gu. Deep fusion of remote sensing data for accurate classification. *IEEE Geosci. Remote Sens. Lett.*, 14(8):1253–1257, 2017.

[Ding *et al.*, 2022a] Jian Ding, Nan Xue, Gui-Song Xia, Xiang Bai, Wen Yang, Michael Ying Yang, Serge Belongie, Jiebo Luo, Mihai Datcu, Marcello Pelillo, and Liangpei Zhang. Object detection in aerial images: A large-scale benchmark and challenges. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(11):7778–7796, 2022.

[Ding *et al.*, 2022b] Kexing Ding, Ting Lu, Wei Fu, Shutao Li, and Fuyan Ma. Global–local transformer network for hsi and lidar data joint classification. *IEEE Trans. Geosci. Remote Sens.*, 60:1–13, 2022.

[Dong *et al.*, 2022] Yanni Dong, Quanwei Liu, Bo Du, and Liangpei Zhang. Weighted feature fusion of convolutional neural network and graph attention network for hyperspectral image classification. *IEEE Trans. Image Process.*, 31:1559–1572, 2022.

[Dong *et al.*, 2023] Wenqian Dong, Yufei Yang, Jiahui Qu, Song Xiao, and Yunsong Li. Local information-enhanced graph-transformer for hyperspectral image change detection with limited training samples. *IEEE Trans. Geosci. Remote Sens.*, 61:1–14, 2023.

[Fang *et al.*, 2021] Sheng Fang, Kaiyu Li, and Zhe Li. S^2enet: Spatial–spectral cross-modal enhancement network for classification of hyperspectral and lidar data. *IEEE Geoscience and Remote Sensing Letters*, 19:1–5, 2021.

[Gao *et al.*, 2023] Sicheng Gao, Xuhui Liu, Bohan Zeng, Sheng Xu, Yanjing Li, Xiaoyan Luo, Jianzhuang Liu, Xiantong Zhen, and Baochang Zhang. Implicit diffusion models for continuous super-resolution. In *IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 10021–10030, 2023.

[Ho *et al.*, 2020] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[Hong *et al.*, 2021] Danfeng Hong, Lianru Gao, Naoto Yokoya, Jing Yao, Jocelyn Chanussot, Qian Du, and Bing Zhang. More diverse means better: Multimodal deep learning meets remote-sensing imagery classification. *IEEE Trans. Geosci. Remote Sens.*, 59(5):4340–4354, 2021.

[Hu *et al.*, 2023] Meiqi Hu, Chen Wu, Bo Du, and Liangpei Zhang. Binary change guided hyperspectral multiclass change detection. *IEEE Trans. Image Process.*, 32:791–806, 2023.

[Huo *et al.*, 2023] Lu Huo, Jiahao Xia, Leijie Zhang, Haimin Zhang, and Min Xu. Multimodal hyperspectral image classification via interconnected fusion. *arXiv preprint arXiv:2304.00495*, 2023.

[Jia *et al.*, 2023] Sen Jia, Xi Zhou, Shuguo Jiang, and Ruyan He. Collaborative contrastive learning for hyperspectral and lidar classification. *IEEE Trans. Geosci. Remote Sens.*, 61:1–14, 2023.

[Kawar *et al.*, 2023] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. In *IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 6007–6017, 2023.

[Kumari *et al.*, 2023] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 1931–1941, 2023.

[Li *et al.*, 2022a] Heng-Chao Li, Wen-Shuai Hu, Wei Li, Jun Li, Qian Du, and Antonio Plaza. A3 clnn: Spatial, spectral and multiscale attention convlstm neural network for multisource remote sensing data classification. *IEEE Trans. Neural Netw. Learn. Syst.*, 33(2):747–761, 2022.

[Li *et al.*, 2022b] Jiaojiao Li, Yuzhe Liu, Rui Song, Yunsong Li, Kailiang Han, and Qian Du. Sal^2rn: A spatial–spectral salient reinforcement network for hyperspectral and lidar data fusion classification. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–14, 2022.

[Li *et al.*, 2023a] Bo Li, Kaitao Xue, Bin Liu, and Yu-Kun Lai. Bbdm: Image-to-image translation with brownian bridge diffusion models. In *IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 1952–1961, 2023.

[Li *et al.*, 2023b] Ke Li, Di Wang, Xu Wang, Gang Liu, Zili Wu, and Quan Wang. Mixing self-attention and convolution: A unified framework for multi-source remote sensing data classification. *IEEE Trans. Geosci. Remote Sens.*, 2023.

[Lin *et al.*, 2023] Junyan Lin, Feng Gao, Xiaochen Shi, Junyu Dong, and Qian Du. Ss-mae: Spatial–spectral masked autoencoder for multisource remote sensing image classification. *IEEE Trans. Geosci. Remote Sens.*, 61:1–14, 2023.

[Luo *et al.*, 2023] Ziwei Luo, Fredrik K Gustafsson, Zheng Zhao, Jens Sjölund, and Thomas B Schön. Refusion: Enabling large-size realistic image restoration with latent-space diffusion models. In *IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 1680–1691, 2023.

[Meng *et al.*, 2022] Qingfa Meng, Guoqing Ma, Taihan Wang, and Tingyi Wang. High-resolution density joint inversion method of airborne and ground gravity data with cross-constraint technique. *IEEE Trans. Geosci. Remote Sens.*, 60:1–9, 2022.

[Metzger *et al.*, 2023] Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Guided depth super-resolution by deep anisotropic diffusion. In *IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 18237–18246, 2023.

[Mohla *et al.*, 2020] Satyam Mohla, Shivam Pande, Biplab Banerjee, and Subhasis Chaudhuri. Fusatnet: Dual attention based spectrospatial multimodal fusion network for hyperspectral and lidar classification. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 416–425, 2020.

[Rombach *et al.*, 2022] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 10684–10695, 2022.

[Roy *et al.*, 2023] Swalpa Kumar Roy, Ankur Deria, Danfeng Hong, Behnood Rasti, Antonio Plaza, and Jocelyn Chanussot. Multimodal fusion transformer for remote sensing image classification. *IEEE Trans. Geosci. Remote Sens.*, 61:1–20, 2023.

[Ruan *et al.*, 2023] Ludan Ruan, Yiyang Ma, Huan Yang, Huiguo He, Bei Liu, Jianlong Fu, Nicholas Jing Yuan, Qin Jin, and Baining Guo. Mm-diffusion: Learning multimodal diffusion models for joint audio and video generation. In *IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 10219–10228, 2023.

[Wang *et al.*, 2022a] Junjie Wang, Wei Li, Yunhao Gao, Mengmeng Zhang, Ran Tao, and Qian Du. Hyperspectral and sar image classification via multiscale interactive fusion network. *IEEE Trans. Neural Netw. Learn. Syst.*, pages 1–15, 2022.

[Wang *et al.*, 2022b] Yinhuai Wang, Jiwen Yu, and Jian Zhang. Zero-shot image restoration using denoising diffusion null-space model. *arXiv preprint arXiv:2212.00490*, 2022.

[Wang *et al.*, 2023] Meng Wang, Feng Gao, Junyu Dong, Heng-Chao Li, and Qian Du. Nearest neighbor-based contrastive learning for hyperspectral and lidar data classification. *IEEE Trans. Geosci. Remote Sens.*, 61:1–16, 2023.

[Wu *et al.*, 2021] Xin Wu, Danfeng Hong, and Jocelyn Chanussot. Convolutional neural networks for multimodal remote sensing data classification. *IEEE Trans. Geosci. Remote Sens.*, 60:1–10, 2021.

[Xiu *et al.*, 2022] Di Xiu, Zongxu Pan, Yirong Wu, and Yuxin Hu. Mage: Multisource attention network with discriminative graph and informative entities for classification of hyperspectral and lidar data. *IEEE Trans. Geosci. Remote Sens.*, 60:1–14, 2022.

[Xu *et al.*, 2018] Xiaodong Xu, Wei Li, Qiong Ran, Qian Du, Lianru Gao, and Bing Zhang. Multisource remote sensing data classification based on convolutional neural network. *IEEE Trans. Geosci. Remote Sens.*, 56(2):937–949, 2018.

[Xu *et al.*, 2023a] Jiale Xu, Xintao Wang, Weihao Cheng, Yan-Pei Cao, Ying Shan, Xiaohu Qie, and Shenghua Gao. Dream3d: Zero-shot text-to-3d synthesis using 3d shape prior and text-to-image diffusion models. In *IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 20908–20918, 2023.

[Xu *et al.*, 2023b] Lin Xu, Hao Zhu, Licheng Jiao, Wenhao Zhao, Xiaotong Li, Biao Hou, Zhongle Ren, and Wenping Ma. A dual-stream transformer with diff-attention for multispectral and panchromatic classification. *IEEE Trans. Geosci. Remote Sens.*, 61:1–14, 2023.

[Xue *et al.*, 2022] Zhixiang Xue, Xiong Tan, Xuchu Yu, Bing Liu, Anzhu Yu, and Pengqiang Zhang. Deep hierarchical vision transformer for hyperspectral and lidar data classification. *IEEE Trans. Image Process.*, 31:3095–3110, 2022.

[Yang *et al.*, 2022] Yi Yang, Daoye Zhu, Tengteng Qu, Qiangyu Wang, Fuhu Ren, and Chengqi Cheng. Single-stream cnn with learnable architecture for multisource remote sensing data. *IEEE Trans. Geosci. Remote Sens.*, 60:1–18, 2022.

[Yao *et al.*, 2023a] Jing Yao, Danfeng Hong, Haipeng Wang, Hao Liu, and Jocelyn Chanussot. Ucsl: Towards unsupervised common subspace learning for cross-modal image classification. *IEEE Trans. Geosci. Remote Sens.*, 2023.

[Yao *et al.*, 2023b] Jing Yao, Bing Zhang, Chenyu Li, Danfeng Hong, and Jocelyn Chanussot. Extended vision transformer (exvit) for land use and land cover classification: A multimodal deep learning framework. *IEEE Trans. Geosci. Remote Sens.*, 2023.

[Zhang *et al.*, 2020] Mengmeng Zhang, Wei Li, Qian Du, Lianru Gao, and Bing Zhang. Feature extraction for classification of hyperspectral and lidar data using patch-to-patch cnn. *IEEE Trans. Cybern.*, 50(1):100–111, 2020.

[Zhang *et al.*, 2023] Yuxin Zhang, Nisha Huang, Fan Tang, Haibin Huang, Chongyang Ma, Weiming Dong, and Changsheng Xu. Inversion-based style transfer with diffusion models. In *IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 10146–10156, 2023.

[Zhao *et al.*, 2020] Xudong Zhao, Ran Tao, Wei Li, Heng-Chao Li, Qian Du, Wenzhi Liao, and Wilfried Philips. Joint classification of hyperspectral and lidar data using hierarchical random walk and deep cnn architecture. *IEEE Trans. Geosci. Remote Sens.*, 58(10):7355–7370, 2020.

[Zhao *et al.*, 2022] Guangrui Zhao, Qiaolin Ye, Le Sun, Zebin Wu, Chengsheng Pan, and Byeungwoo Jeon. Joint classification of hyperspectral and lidar data using a hierarchical cnn and transformer. *IEEE Trans. Geosci. Remote Sens.*, 61:1–16, 2022.

[Zhu *et al.*, 2023] Yuanzhi Zhu, Kai Zhang, Jingyun Liang, Jiezhang Cao, Bihan Wen, Radu Timofte, and Luc Van Gool. Denoising diffusion models for plug-and-play image restoration. In *IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 1219–1229, 2023.