Multimodal Informative ViT: Information Aggregation and Distribution for Hyperspectral and LiDAR Classification

Jiaqing Zhang, Jie Lei, *Member, IEEE*, Weiying Xie, *Senior Member, IEEE*, Geng Yang, Daixun Li, Yunsong Li, *Member, IEEE*

Abstract—In multimodal land cover classification (MLCC), a common challenge is the redundancy in data distribution, where task-irrelevant information from multiple modalities can hinder the effective integration of their unique features. To tackle this, we introduce the Multimodal Informative Vit (MIVit), a system with an innovative information aggregate-distributing mechanism. This approach redefines redundancy levels and integrates performance-aware elements into the fused representation, facilitating the learning of semantics in both forward and backward directions. MIVit stands out by significantly reducing redundancy in the empirical distribution of each modality's separate and fused features. It employs oriented attention fusion (OAF) for extracting shallow local shape features across modalities in horizontal and vertical dimensions, and a Transformer feature extractor for extracting deep global features through long-range attention. We also propose an information aggregation constraint (IAC) based on mutual information, designed to remove redundant information and preserve complementary information within embedded features. Additionally, the information distribution flow (IDF) in MIVit enhances performance-awareness by distributing global classification information across different modalities' feature maps. This architecture also addresses missing modality challenges with lightweight independent modality classifiers, reducing the computational load typically associated with Transformers. Our results show that MIVit's bidirectional aggregate-distributing mechanism between modalities is highly effective, achieving an average overall accuracy of 95.56% across three multimodal datasets. This performance surpasses current state-of-the-art methods in MLCC. The code for MIVit is accessible at https://github.com/icey-zhang/MIViT.

Index Terms—Hyperspectral image, multimodal fusion, land cover classification, mutual information, self-distillation.

I. INTRODUCTION

LAND Cover Classification (LCC) is a key area in remote sensing image processing, focusing on identifying and categorizing land types or cover classes on Earth's surface areas [1]–[3]. Advances in satellite technology have enabled the capture of multi-source imagery from various satellite
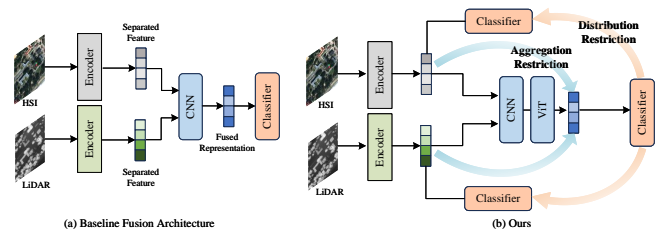
Fig. 1. Existing MLCC method *vs.* ours MIViT.

sensors, leading to the growth of multimodal fusion in remote sensing image analysis [4]–[6]. Each data source, such as Hyperspectral Image (HSI) and Light Detection and Ranging (LiDAR) Image, provides unique insights into the same geographical area [7], [8], enriching LCC by offering more comprehensive representations of scenes and objects in remote sensing images [9]. Despite these advancements, effectively utilizing information from various multimodal sources to improve the accuracy and reliability of LCC remains a significant challenge.

In recent years, many methods have been developed to address the challenges of multimodal land cover classification (MLCC) [10]–[12]. A common pipeline that demonstrated promising results utilizes CNN-based feature extraction in an end-to-end manner [13]–[18]. For instance, Chen *et al.* [13] initially employed deep neural networks to extract features from HSI and LiDAR for precise classification. Xu *et al.* designed a dual-branch CNN framework, where one branch was dedicated to extracting spectral-spatial features from HSI, and another was designed for LiDAR data feature extraction. Similarly, Dong *et al.* [19] developed the MB2FscgaNet, a multi-branch fusion network emphasizing self- and cross-guided attention to refine HSI and LiDAR classification, thereby enhancing the accuracy of spectral, spatial, and elevation feature estimations. Additionally, Wu *et al.* [17]introduced the CCR-Net, incorporating a novel cross-channel reconstruction module, and Zhang *et al.* [20] devised a spectral-spatial fractal residual convolutional neural network with data balance augmentation to address dataset-specific challenges such as imbalanced categories and limited labeled data. Li *et al.* [21] first considers on-orbit fusion with limited computing resources and provides a strategy manifold structure for multi-modal fusion transmission.

These methodologies generally involve two distinct phases: the fusion of multimodal data through feature extractors to create latent mapping representations, followed by the classification of these representations, as illustrated in Fig. 1a. Despite their effectiveness, these methods exhibit certain limitations. Firstly, conventional CNNs, being context-independent, primarily extract local information within a limited receptive field and struggle with global information acquisition [22]. Secondly, the opaque internal mechanisms of CNNs often lead to the insufficient and redundant extraction of cross-modality features. Thirdly, in single-modal feature extraction, the distance from the initial feature extraction to the final decision classifier often results in a significant loss of task-relevant knowledge [23]. To address these shortcomings, Transformer-based algorithms [10], [24]–[26] have been introduced, offering advantages in capturing global long-range sequence information. However, they also face challenges, particularly in terms of the high computational resources required [27], [28]. In light of these issues, our paper focuses on reducing information redundancy to enhance feature fusion and foster performance-aware learning and classification discrimination.

Aiming to improve the controllability and interpretability of multimodal fusion, we propose the Multimodal Informative ViT (MIViT) algorithm, as shown in Fig. 1b. This approach initially transforms multimodal data into low-dimensional separated features using CNNs for local context and efficiency, and Transformers for global attention and long-range dependency, particularly in addressing the LCC task. We utilize an Oriented Attention Fusion (OAF) based on CNN to extract cross-modality shallow local context features in both horizontal and vertical dimensions. Furthermore, the Transformer feature extractor ViT [29] is designed to leverage global attention and long-range dependency for handling deep global features, enhancing the overall effectiveness and accuracy of the LCC process.

Inspired by information theory, MIViT integrates an iterative refinement mechanism for information aggregation and distribution. It comprises two branches: the forward aggregation branch and the backward distribution branch. In the forward aggregation branch, mutual information (MI) [30] is utilized to mathematically quantify and define the redundancy present in feature information. Building upon this concept, we have developed an Information Aggregation Constraint (IAC), which is grounded in MI. The primary aim of IAC is to selectively filter out redundant and irrelevant information from tasks, thereby facilitating the formation of compact, yet highly reliable representations.

The backward distribution branch plays a crucial role in enhancing the granularity of classification results. It takes the globally fused features and effectively distributes them across separated feature maps. This is achieved through the Information Distribution Flow (IDF), which is based on the principles of self-distillation (SD) [23]. The IDF is instrumental in reducing discrepancies between different classification maps, thereby significantly improving the perceptual capabilities of individual modalities. This unique approach allows for a more holistic and comprehensive construction of multimodal fusion representations. Notably, this distillation method permits the

output of classification results through shallow classifiers for each modality, independent of the presence of multiple modalities. This feature is especially significant as it leverages a lightweight CNN architecture, enabling long-range knowledge transfer similar to that of deep Transformer networks, but without the associated computational burden.

Our contributions to this work are multifaceted and can be summarized as follows:

- We introduce the MIViT network, featuring an innovative information aggregate-distributing mechanism. This mechanism is adept at extracting and fusing both global and local features, thereby facilitating a dynamic and interactive flow of information throughout the multimodal fusion process. The result is the generation of compact representations that are acutely aware of performance dynamics.
- Aiming to remove less task-related redundant information, We formulate an information aggregation constraint based on mutual information. This constraint is a pioneering step in redefining the redundancy that exists between multimodal data and fusion representations, and it is integrated as a regularization term within the loss function.
- We design an information distribution flow, grounded in self-distillation constraints, allowing for the effective dissemination of global fusion information to feature maps of various modalities. This not only enhances the performance awareness of individual modalities but also leads to the creation of more comprehensive and enriched fusion representations.
- Our research marks the first exploration into the realm of missing modality learning via knowledge distillation. We delve into the capabilities of lightweight CNN models in handling missing modalities, showcasing their potential to mimic the learning efficiency of Transformers while evading the extensive computational demands typically associated with Transformer architectures.

II. RELATED WORK

This section provides an overview of significant advancements in deep learning-based fusion approaches and multimodal learning with mutual information, as employed in MIViT.

A. Deep learning-based Fusion methods

Nowadays, with the success of deep learning (DL) methods in various fields, the DL methods have also begun to be utilized in multi-modal learning and have made great progress. The integration of deep learning (DL) methods into multimodal learning has marked a significant shift in this field, leading to substantial progress. The fundamental premise of multimodal learning is the existence of both shared and distinct information across different modalities [6]. In the context of Land Cover Classification (LCC), datasets comprising Hyperspectral Images (HSIs) [31], LiDAR images [9], and SAR images [32] exhibit shared semantic information. While HSI predominantly encodes spectral information, LiDAR and
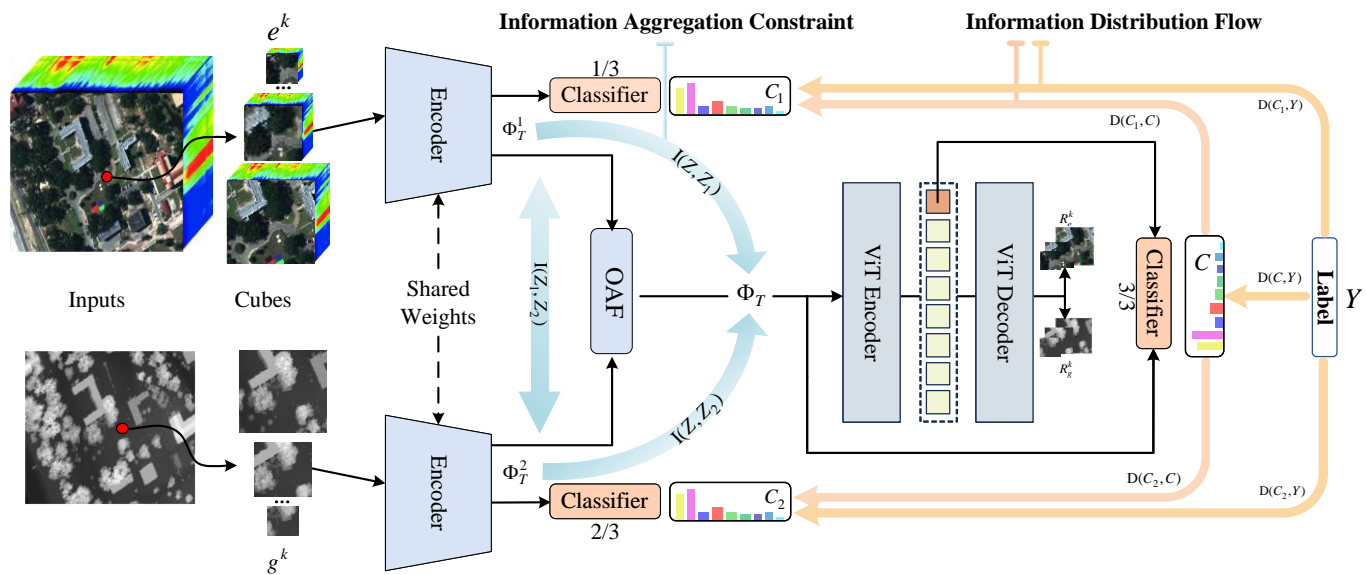
Fig. 2. The detailed frame of our proposed method. The different modalities are firstly fed into an Alignment Encoder (AE) to capture the aligned separated features (Φ_T^1 and Φ_T^2) from shallow to deep levels. Subsequently, we fuse the separated features to effectively model the complementary information of each mode by Oriented Attention fusion (OAF). To explicitly encourage complementary learning, eliminate information redundancy, and enhance the performance perception capability of multi-classifiers, we impose both the Information Aggregation Constraint (IAC) and Information Distribution Flow (IDF) over the separated and fused representations.

SAR are more focused on geometric details, like the relative elevation or localization of objects. This distinction between spectral and geometric information highlights the uniqueness of each modality. Popular approaches to achieve multimodal learning include early fusion, late fusion [33], and cross fusion [26], [34], [35]. Early fusion involves direct concatenation of modalities, whereas late fusion processes each modality separately, combining them only at the output layer. Cross fusion, as demonstrated in studies [7], [26], effectively strengthens the interaction and semantic correlation between different modalities.

Existing methods, while adept at fusing for multimodal learning, often do not explicitly address how the CNN network facilitates effective multimodal learning of fused representation. Our approach, with its focus on aggregation and distribution constraints, aims to minimize the redundancy of appearance and geometric features, thereby promoting a more performance-aware and efficient multimodal learning process.

B. Multi-modal learning with mutual information

In machine learning, mutual information is frequently used as a criterion to either encourage or limit dependencies between variables in a range of tasks [36]–[38]. In the realm of multimodal learning, deep mutual information maximization methods have been proposed for applications like cross-clustering [39]. These methods focus on preserving shared information across modalities while discarding extraneous individual modal information in an end-to-end manner. Another approach maximizes mutual information between 3D objects and their geometrically transformed versions to refine representations [40]. Contrarily, mutual information minimization is employed to explicitly foster multimodal information learning, as seen in the interaction between RGB images and depth

data [41]. Similarly, mutual information has been incorporated as an effective representation regularizer in tasks like Pan-sharpening [42].

The information redundancy in the feature extraction of the various modalities naturally exists in MLCC task [32] which leads to the limit of the effective feature fusion. To address this, we propose the minimization of mutual information between two modalities to reduce specific information redundancy. Additionally, we aim to maximize mutual information between the fusion features and other isolated modality features. This approach not only captures sufficient complementary information but also enhances the overall performance of fusion features. Importantly, our focus extends beyond just learning representations of different modalities; we also concentrate on the fused representation itself. This aspect of our work sets it apart from previous studies, emphasizing our unique approach to multimodal learning with mutual information.

III. METHOD

In this section, we give the mathematical definition of task-relevant and task-irrelevant information for multimodal learning and fusion in the LCC task. We then detail the process of achieving compact representation and describe the implementation of our information aggregate-distributing mechanism.

A. Problem Definition

The essence of LCC lies in the precise categorization of each pixel in an image into a specific land cover category. In the context of multimodal LCC, this involves the complex processes of multimodal learning and fusion. Given n heterogeneous remote sensing modalities, denoted as $X_n \in \mathbb{R}^{h \times w \times b}$,

the challenge is to extract a compact and reliable representation Z from the input image X. This representation is then used to map into a new probability distribution $C \in \mathbb{R}^{h \times w \times c}$, indicating the likelihood of each pixel belonging to different classes. In this scenario, h, w, b, and c represent height, width, number of channels, and classification categories, respectively. The success of this task hinges on the efficient coupling of information without redundancy and the perception of performance within the latent fusion representation, which significantly influences the accuracy of the final classification results. In our approach, multimodal LCC involves two key processes: eliminating task-irrelevant information r^* to reduce redundancy between fusion modes and enhancing task-relevant information s^*. This dual approach ensures that the classification results of fusion representations and individual modal features mutually constrain each other during the classification process. The problem can be defined as:

$$
\begin{aligned}
\min r^* &= \sum_{i=1}^{n} \sum_{j>i}^{n} \mathrm{I}\left(Z_i, Z_j\right) - \sum_{i=1}^{n} \mathrm{I}\left(Z, Z_i\right) \\
\max s^* &= -\left(\sum_{i=1}^{n} \mathrm{D}\left(C_i, C\right) + \sum_{i=1}^{n} \mathrm{D}\left(C_i, Y\right) + \mathrm{D}\left(C, Y\right)\right)
\end{aligned}
\tag{1}
$$

where Z_i are probability distributions of compressed embedding of modalities X_i, so minimizing mutual information $\mathrm{I}(Z_i, Z_j)$ can eliminate the redundant information between X_i and X_j. Z represents the fused representation and the complementary information of multiple modalities by maximizing mutual information between Z and Z_i. The similarity distribution distance [43] between the two classification results after a softmax operation is defined as $\mathrm{D}(\cdot)$. The maximum value of a negative distance reflects a greater degree of similarity between distributions. The notation Y refers to the training labels, which represent the labeled GroundTruth corresponding to the input data. To simplify the expression, we set $n = 2$ in the subsequent paper.

B. Multimodal Informative ViT

1) Alignment Encoder: We employ a multi-scale shared encoder to align latent feature distribution. This encoder utilizes a convolutional block to project multi-modal images to modality-aware maps, extracting multi-scale shallow features for each modality efficiently. The encoder firstly splits N types of patches from the original images, denoted as $\{e^k\}_{k=1}^N$ and $\{g^k\}_{k=1}^N$, for their respective modalities. For example, A 100×100 image will be cropped into 8×8 (or another) size patches [44], where the GroundTruth category corresponding to the center position is used as the patch label. Two branches, $\mathcal{P}(\cdot)$ and $\mathcal{Q}(\cdot)$, help obtain multi-scale shared shallow features of different modalities:

$$
\Phi_P^k = \mathcal{P}(e^k), \Phi_Q^k = \mathcal{Q}(g^k).
\tag{2}
$$

The features, Φ_P^k and Φ_Q^k, are then mapped to the same spatial dimension using a convolutional block and max-pooling operation. Assuming that the embedding functions are

$\{\mathcal{F}(\cdot)\}_{k=1}^N$ and $\{\mathcal{G}(\cdot)\}_{k=1}^N$. The base features extracted from the shared features can be represented as:

$$
\Phi_F^k = \mathcal{F}^k(\Phi_P^k), \Phi_G^k = \mathcal{G}^k(\Phi_Q^k).
\tag{3}
$$

The base features are adaptively linearly weighted to achieve multi-scale feature fusion:

$$
\Phi_T^1 = \alpha_k \Phi_F^k, \Phi_T^2 = \alpha_k \Phi_G^k,
\tag{4}
$$

where α_k represents the adaptive weighting coefficient, learnable and updatable by the network.
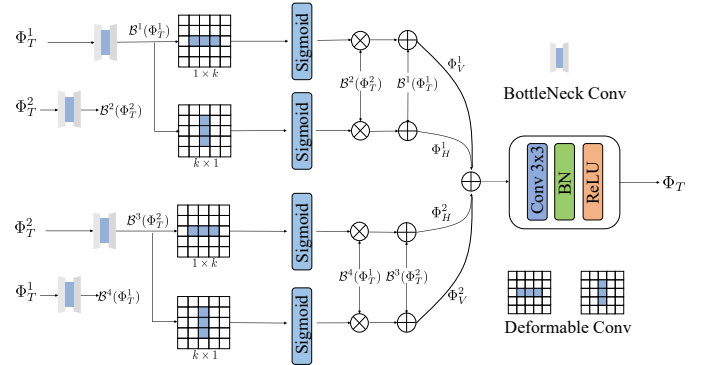

Fig. 3. The structure of oriented attention fusion (OAF) module.

2) Oriented Attention Fusion (OAF): Based on the premise of intermodal information interaction inspired by cross-attention [45], we intend to facilitate the effective exchange of information between two modalities. As mentioned in section III-B1, land cover classification tasks were entered according to patch-level data, but labels were only marked for patches' central coordinates, so the commonly used local spatial information extraction method often resulted in the redundancy of local information and failed to extract useful shape information. The OAF module fuses multi-modal cross-information through two-orientation attention aggregation. Each attention module in OAF generates an attention map along two directions (horizontal and vertical) to extract shape information and modulates the other modality features to accentuate the consistency features of the fused data. The process can be written as:

$$
\begin{aligned}
\Phi_H^1 &= \mathcal{S}(\mathcal{E}_{1 \times k}^1(\mathcal{B}^1(\Phi_T^1))) \times \mathcal{B}^2(\Phi_T^2) + \mathcal{B}^1(\Phi_T^1), \\
\Phi_V^1 &= \mathcal{S}(\mathcal{E}_{k \times 1}^1(\mathcal{B}^1(\Phi_T^1))) \times \mathcal{B}^2(\Phi_T^2) + \mathcal{B}^1(\Phi_T^1), \\
\Phi_H^2 &= \mathcal{S}(\mathcal{E}_{1 \times k}^2(\mathcal{B}^3(\Phi_T^2))) \times \mathcal{B}^4(\Phi_T^1) + \mathcal{B}^3(\Phi_T^2), \\
\Phi_V^2 &= \mathcal{S}(\mathcal{E}_{k \times 1}^2(\mathcal{B}^3(\Phi_T^2))) \times \mathcal{B}^4(\Phi_T^1) + \mathcal{B}^3(\Phi_T^2),
\end{aligned}
\tag{5}
$$

where $\mathcal{B}(\cdot)$ denotes the bottleneck structure which is similar to the role of Nonnegative Matrix Factorization (NMF) [46], which can retain useful features while filtering out redundant high-frequency noise to enhance the feature quality. $\mathcal{E}(\cdot)$ represents the $1 \times k$ deformable convolution in the horizontal direction and $k \times 1$ deformable convolution in the vertical direction. $\mathcal{S}$ is a sigmoid function utilized to obtain attention maps. The oriented attention fusion mechanism promotes extracting shape information from one modality-enhanced feature in two directions and accordingly guides the refinement of the other modality feature by element-wise multiplication

operation which is represented as $\times$. And then this complementary information from the other modality is added to the original modality to enhance the information extraction and reduce information loss.

The attentive features are then summed together:

$$\Phi_T = \mathcal{T}(\Phi_H^1 + \Phi_V^1 + \Phi_H^2 + \Phi_V^2), \qquad (6)$$

where $\mathcal{T}(\cdot)$ represents a base convolutional block that balances the fusion of different features. This two-orientation attention mechanism in OAF is pivotal in extracting cross-level features and refining modality-aware features. OAF is utilized to perform cross-level feature fusion. Simultaneously, the different multimodal classification results Y_1 and Y_2 are obtain by classifiers $\mathcal{C}$:

$$C_1 = \mathcal{C}(\Phi_T^1), C_2 = \mathcal{C}(\Phi_T^1). \qquad (7)$$

3) Transformer Reconstruction Decoder: To extract long-distance dependency features, we use a transformer encoder $\mathcal{V}(\cdot)$ based on ViT [29] to explore global spectral dependencies from CNN-based encoding features Φ_T containing rich spatial information. This is formulated as:

$$\Phi_V = \mathcal{V}(\Phi_T). \qquad (8)$$

To obtain the joint distinguish capacity from CNN and transformer structures, we complete the final classification by:

$$C = \mathcal{C}(\Phi_T, \Phi_V). \qquad (9)$$

The decoder $\mathcal{D}(\cdot)$ reconstructs input modalities:

$$R_e^k, R_g^k = \mathcal{D}(\Phi_V), \qquad (10)$$

where the $\mathcal{D}$ enhances the fused representation using ViT as a structure.
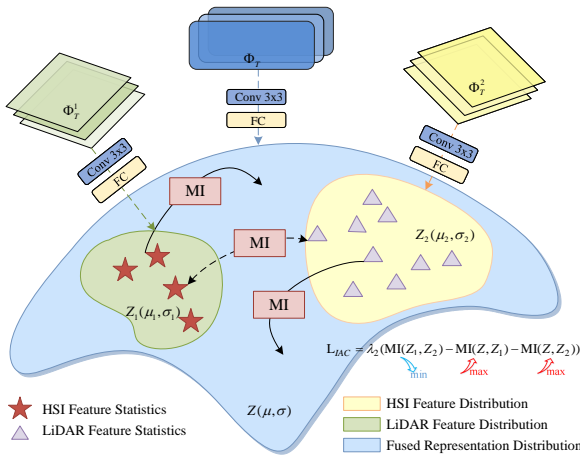

Fig. 4. The illustration of complementary information preservation and redundancy information elimination.

C. Information Aggregate-distributing Mechanism

1) Information Aggregation Constraint: To tackle the challenges of information redundancy and insufficient information in fusion maps, we employ mutual information maximization and minimization. We aim to enhance learning and fusion of complementary information between two modalities and reduce redundancy. The process of preserving complementary information and eliminating redundant information is illustrated in Fig. 4. Given unimodal features Φ_T^1 and Φ_T^2 and multi-modal fusion representation Φ_T, we map them into low-dimensional vectors for mutual information analysis. Particularly, these features Φ_T^1, Φ_T^2 and Φ_T are fed into an additional convolution layer with a kernel size of 3×3, and then followed by a fully connected layer that receives above-reshaped features to obtain the low-dimensional feature vectors Z_1, Z_2 and Z:

$$Z_1 = \mathcal{M}(\Phi_T^1), Z_2 = \mathcal{M}(\Phi_T^2), Z = \mathcal{M}(\Phi_T), \qquad (11)$$

where $\mathcal{M}_1$, $\mathcal{M}_2$ and $\mathcal{M}$ represent the convolution layers of kernel size 3×3 and fully-connected layers. In information theory, mutual information aims to measure the amount of information about a random variable by observing some other random variable or vice versa. The mutual information can be used to measure the difference between the entropy terms:

$$\mathrm{MI}(Z, Z_1) = \mathrm{H}(Z) - \mathrm{H}(Z \mid Z_1), \qquad (12)$$

$$\mathrm{MI}(Z, Z_2) = \mathrm{H}(Z) - \mathrm{H}(Z \mid Z_2). \qquad (13)$$

Take the calculation of Eq. 12 as an example:

$$\mathrm{H}(Z \mid Z_1) = \mathrm{H}(Z, Z_1) - \mathrm{H}(Z_1), \qquad (14)$$

where $\mathrm{H}(\cdot)$ is the entropy, $\mathrm{H}(Z)$ and $\mathrm{H}(Z_1)$ represents marginal entropies, and $\mathrm{H}(Z, Z_1)$ indicate joint entropy of Z and Z_1, $\mathrm{H}(Z, Z_1)$ is the conditional entropy. Combing above two equations, the mutual information $\mathrm{MI}(Z, Z_1)$ can be rewritten between the entropy terms:

$$\mathrm{MI}(Z, Z_1) = \mathrm{H}(Z) + \mathrm{H}(Z_1) - \mathrm{H}(Z, Z_1), \qquad (15)$$

where the marginal entropies $\mathrm{H}(Z)$ and $\mathrm{H}(Z_1)$ can be calculated with Kullback-Leibler divergence (KL) as:

$$\mathrm{H}(Z) = \mathrm{H}_{Z_1}(Z) - KL(Z\|Z_1), \qquad (16)$$

$$\mathrm{H}(Z_1) = \mathrm{H}_z(Z_1) - KL(Z_1\|Z), \qquad (17)$$

where $\mathrm{H}_{Z_1}(Z)$ is the cross-entropy. We then obtain:

$$\begin{aligned} \mathrm{MI}(Z, Z_1) = {}& \mathrm{H}_{Z_1}(Z) + \mathrm{H}_Z(Z_1) - \mathrm{H}(Z, Z_1) \\ & - \mathrm{KL}(Z\|Z_1) - \mathrm{KL}(Z_1\|Z). \end{aligned} \qquad (18)$$

Similarly, we can calculate the mutual information between Z and Z_2, as well as between Z_1 and Z_2:

$$\begin{aligned} \mathrm{MI}(Z, Z_2) = {}& \mathrm{H}_{Z_2}(Z) + \mathrm{H}_Z(Z_2) - \mathrm{H}(Z, Z_2) \\ & - \mathrm{KL}(Z\|Z_2) - \mathrm{KL}(Z_2\|Z), \end{aligned} \qquad (19)$$

$$\begin{aligned} \mathrm{MI}(Z_1, Z_2) = {}& \mathrm{H}_{Z_2}(Z_1) + \mathrm{H}_{Z_1}(Z_2) - \mathrm{H}(Z_1, Z_2) \\ & - \mathrm{KL}(Z_1\|Z_2) - \mathrm{KL}(Z_2\|Z_1). \end{aligned} \qquad (20)$$

Intuitively, $MI(Z_1 Z_2)$ measures the reduction of uncertainty in Z_1 when Z_2 is observed, or vice versa while $\mathrm{MI}(Z, Z_1)$
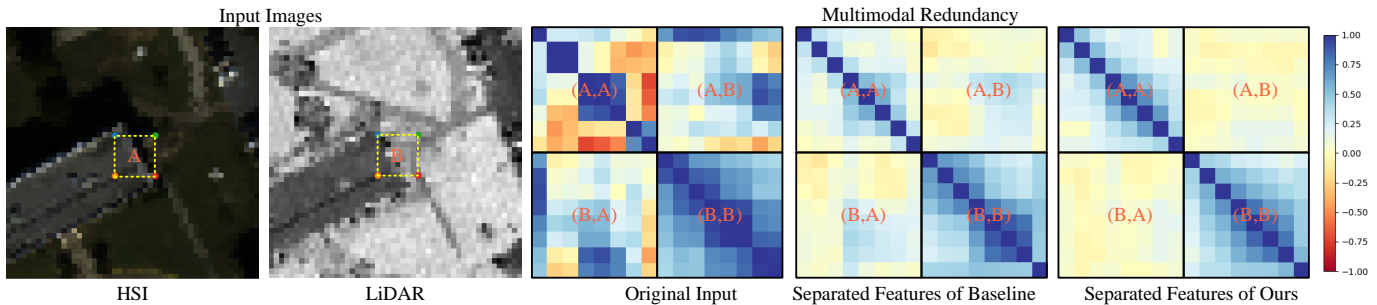
Fig. 5. Visualization of the multimodal redundancy in the existing MLCC method and ours MIViT.

and $\mathrm{MI}(Z, Z_2)$ measure the information transition from Z_1 and Z_2 to fusion features Z.

We use the correlation coefficient matrix to measure the redundancy between two modalities, as shown in Fig. 5. The original input images exhibit high correlation, but the redundancy of the features obtained through the neural network is reduced. This indicates that removing redundancy in multimodal learning is important for improving the network's ability to extract semantic information and recognize categories. Our algorithm has been validated to perform better in removing redundant information.

2) Information Distribution Flow (IDF): IDF is designed to transfer task-relevant knowledge from deep to shallow classifiers, integrating performance-aware elements into fused representations and enhancing the classification ability of isolated modality classifiers. This design provides a solution to address missing multimodal learning scenarios. In a classification problem, a widely-used loss function between two discrete distributions is cross-entropy loss, whose gradient is equivalent to the gradient of KL divergence between two distributions. We assume C is the deepest classifier and C_n is the shallow classier output. Cross entropy loss from label to not only the deepest classifier but also all the shallow classier. It is computed with the labels from the training dataset and the outputs of each classifier's softmax layer. In this way, the knowledge hidden in the dataset is introduced directly from labels to all the classifiers:

$$\mathrm{D}(C_1, Y) + \mathrm{D}(C_2, Y) = \mathrm{H}_{C_1}(Y) + \mathrm{H}_{C_2}(Y),$$
$$\mathrm{D}(C, Y) = \mathrm{H}_C(Y). \tag{21}$$

KL (Kullback-Leibler) divergence loss under the deepest classifier's guidance. The KL divergence is computed using softmax outputs between the shallow classifiers and the deepest classifier and introduced to the softmax layer of each shallow classifier. By introducing KL divergence, the self-distillation framework affects the deepest classifier, the deepest one, to each shallow classifier:

$$\mathrm{D}(C_1, C) + \mathrm{D}(C_2, C) = \mathrm{KL}(C_1 \| C) + \mathrm{KL}(C_2 \| C). \tag{22}$$

The popular solutions to the problem of missing multimodal learning [47]–[49] are to design a cross-modality reconstruction network or separate the tasks of two modalities. Generally, it is difficult to reconstruct heterogeneous modalities from each other, while modal separation may lead to impaired information interaction. In this paper, the IDF is considered,
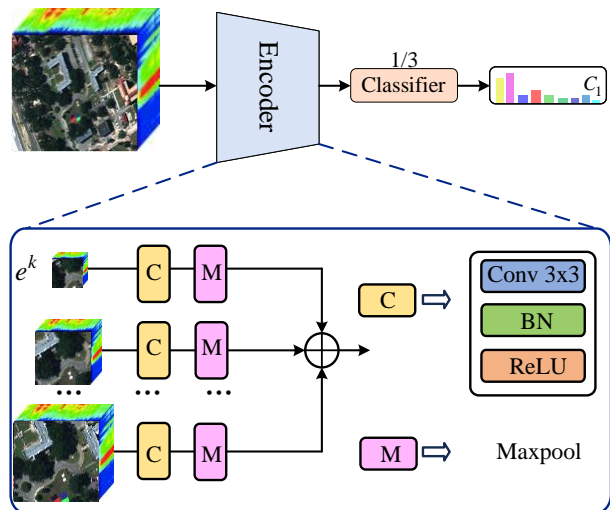

Fig. 6. The illustration of single-modal shallow classifier for the missing multimodal learning in the inference stage.

which distributes the refined classification results of globally fused information from deep classifiers to shallow classifiers in different modal branches, enhancing their performance perception ability. This enables independent modal classification tasks to be achieved even in the absence of a certain modality, using a single-modal shallow classifier trained on a single-modal dataset as shown in Fig. 6. This is a lightweight CNN network, comprising only k (k set as 3 in paper) convolution layers, k maxpool layers, and a fully connected layer.

D. Optimization

In this part, a variational optimization method is proposed to optimize the objective function of the MIViT method. As shown in Fig. 2, our final objective function is:

$$\mathcal{L} = \mathcal{L}_{IDF} + \mathcal{L}_{Re} + \mathcal{L}_{IAC}. \tag{23}$$

And the Algorithm 1 clarifies the process of our MIViT method. The reconstruction losses ensure that the information contained in the images is not lost during the encoding and decoding process, *i.e.*:

$$\mathcal{L}_{\mathrm{Re}} = \lambda_1 \frac{1}{N} \sum_{k=1}^{N} \left(e^k - R_e^k \right)^2 + \left(g^k - R_g^k \right)^2, \tag{24}$$

Algorithm 1 Training MIViT model for fusion classification of multimodal remote sensing data

---

Input: Multimodal images X_n (n=1,2), training label Y, training epochs *epochs*, types of patches N;

Output: Classification result C.

1: Divide X_n into contextual patches $\{e_k\}_{k=1}^N$ and $\{g_k\}_{k=1}^N$.

2: Initialize all weights.

3: **while** epoch $<$ *epochs* do **do**

4:    Extract the separated features Φ_T^1 and Φ_T^2 from the alignment encoder by Eqs. 2, 3 and 4 .

5:    Obtain the fused representation Φ from oriented attention fusion by Eq. 5 and Eq. 6.

6:    Generate the long-range information Φ_V from the ViT encoder by Eq. 8.

7:    Acquire the classification maps C_n and C in the shallow and latent classifiers according to Eq. 9 and Eq. 7, respectively.

8:    Measure the information redundancy among distributions of the separated features and fusion representations using mutual information entropy by Eqs. 18, 19 and 20.

9:    Distillate the performance-aware classification probability of fused representation to the separated features according to Eq. 21 and Eq. 22.

10:    Reconstruct patches utilizing the ViT decoder in the way of Eq. 10.

11:    Update the MIViT model according to Eq. 23.

12: **end while**

---

where λ_1 is the tuning parameter. By maximizing $\mathrm{MI}(Z, Z_1)$ and $\mathrm{MI}(Z, Z_2)$ can explore the complementary attributes of both modalities and enhance the feature representation of the multi-modal fusion. By minimizing $\mathrm{I}(Z_1, Z_2)$, we can effectively reduce the information redundancy between the two modalities. Hence the final IAC objective loss function is defined as:

$$\mathcal{L}_{IAC} = \lambda_2(\mathrm{MI}(Z_1, Z_2) - \mathrm{MI}(Z, Z_1) - \mathrm{MI}(Z, Z_2)), \quad (25)$$

where λ_2 is the tuning parameter. The IDF objective loss function is defined as:

$$\mathcal{L}_{IDF} = \lambda_3(\mathrm{D}(C_1, Y) + \mathrm{D}(C_2, Y)) + \mathrm{D}(C, Y) \\ + \lambda_4(\mathrm{D}(C_1, C) + \mathrm{D}(C_2, C)). \quad (26)$$

where λ_3 as well as λ_4 are the tuning parameters.

IV. EXPERIMENTS

A. Datasets and metrics

In our study, we conducted experiments on three widely recognized benchmarks to assess the performance of our proposed fusion method: **Houston2013**, **MUUFL**, and **Trento**. Detailed characteristics of these datasets are presented in Table I. To evaluate the classification outcomes quantitatively, we employed three key metrics: overall accuracy (OA), average accuracy (AA), and Kappa coefficient.

TABLE I
DATASET DESCRIPTION.

Dataset	Houston	MUUFL	Trento
Location	Houston,Texas,USA	Long Beach, Mississippi,USA	Trento,Italy
Sensor Type	HSI,DSM	HSI,DSM,DEM	HSI,DSM
Channel Number	144,1	64,2	63,1
Image Size (H)	349	325	600
Image Size (W)	1905	220	166
Train Samples	2832	2683	819
Test Samples	12197	51004	29595
Classes	15	11	6

B. Implement details

The experimental setup was executed on a system equipped with an NVIDIA GeForce RTX 3090 GPU. During the preprocessing phase, training samples were randomly cropped into patches of sizes 8×8, 16×16, and 24×24. For optimization, we utilized the Adam optimizer with an initial learning rate of 1e-4 and a weight decay set at 1e-3. We also employed a step scheduler with a step size of 50 and a gamma value of 0.9. The training process spanned 500 epochs. To align with most comparative experiments and ensure optimal performance, we set the batch size to 64. Regarding the network's hyperparameters, the decoder's configuration was mirrored to match that of the encoder.

C. Comparison with state-of-the-art methods

In assessing the performance of our proposed MIViT method, we selected several leading deep learning-based MLCC methods for comparison. These include **Endnet** [15], **CoupleCNN** [35], **Cross** [50], **CALC** [12], **CCR** [17], **HCT** [26], **ViT** [29], **MFT** [10], **ExViT** [24], **GLT** [11].

1) Quantitative Comparison: The comparative results over five datasets are presented in Table II and Table III, with the highest values emphasized in bold. Our method outperforms other advanced classification techniques across all remote sensing datasets. Specifically, compared to the second-best methods, our method shows average gains of 1.87%, 0.46%, and 0.14% in OA on the Houston2013, MUUFL, and Trento datasets, respectively. In addition to OA, we observed consistent improvements in other metrics, highlighting the distinctiveness of the fused features. The MIViT method distinguishes itself as the only approach capable of handling various dataset types, showcasing remarkable generalization capabilities and potential for MLCC tasks. It demonstrates SOTA classification performance on all six benchmark multi-modal datasets, surpassing all existing deep fusion strategies. This validates the effectiveness of the mutual information theory in feature-level fusion for multi-modal land cover classification.

2) Qualitative Comparison: We conducted a qualitative evaluation by visualizing the classification maps generated by different methods. Figs. 7, 8, and 9 display classification maps for hyperspectral datasets (Houston, MUUFL, Trento) alongside their corresponding LiDAR modalities. CNN-based

TABLE II
OA, AA, AND Kappa ON THE HOUSTON2013 DATASET BY CONSIDERING HSI AND LiDAR DATA. THE BEST RESULT IS **HIGHLIGHTED** AND THE SECOND RESULT IS <u>UNDERLINED</u>

| No.	Class(Train/Test)	Endnet	CoupleCNN	Cross	CALC	CCR	HCT	ViT	MFT	ExViT	GLT	MIViT
1	Healthy grass(198/1053)	83.86	83.10	83.10	78.63	<u>88.60</u>	82.91	82.59	82.34	81.2	83.10	**91.07**
2	Stressed grass(190/1064)	<u>98.40</u>	85.15	95.49	83.83	**99.34**	91.35	82.33	88.78	85.15	85.06	85.15
3	Synthetic grass(192/505)	**100.00**	92.28	96.63	93.86	**100.00**	100.00	97.43	98.15	<u>99.80</u>	99.21	99.00
4	Tree(188/1056)	98.39	99.24	98.77	86.55	<u>98.96</u>	91.10	92.93	94.35	91.38	91.19	**100.00**
5	Soil(186/1056)	99.15	**100.00**	100.00	99.72	<u>99.91</u>	100.00	99.84	99.12	99.62	100.00	100.00
6	Water(182/143)	93.71	91.61	97.20	<u>97.90</u>	95.10	95.80	84.15	**99.30**	93.01	81.12	95.80
7	Residential(196/1072)	83.86	85.63	**93.84**	91.42	<u>92.35</u>	81.06	87.84	88.56	91.51	90.21	91.14
8	Commercial(191/1053)	91.17	72.74	91.93	92.88	90.98	94.97	79.93	86.89	**97.44**	<u>96.87</u>	**97.44**
9	Road(193/1059)	85.08	90.75	**95.28**	87.54	87.54	88.29	82.94	87.91	88.48	<u>94.15</u>	91.60
10	Highway(191/1036)	**85.62**	62.55	72.97	68.53	78.86	76.45	52.93	64.70	<u>81.56</u>	75.58	78.47
11	Railway(181/1054)	82.35	<u>97.82</u>	91.65	93.36	91.18	97.25	80.99	**98.64**	94.31	94.97	97.53
12	Park lot 1(192/1041)	83.57	94.43	87.61	95.10	82.04	91.55	91.07	94.24	93.76	<u>97.89</u>	**98.94**
13	Park lot 2(184/285)	76.14	91.23	**92.98**	**92.98**	81.05	88.42	87.84	90.29	90.53	<u>91.58</u>	90.18
14	Tennis court(181/247)	100.00	87.85	93.12	100.00	93.93	100.00	100.00	<u>99.73</u>	97.57	99.19	100.00
15	Running track(187/473)	98.31	96.41	100.00	99.37	98.73	95.56	<u>99.65</u>	99.58	97.04	98.94	98.94
	OA(%)	89.93	87.91	<u>91.84</u>	88.97	91.56	90.43	85.05	89.80	91.27	91.64	**93.71**
	AA(%)	90.64	88.72	<u>92.70</u>	90.78	91.90	91.65	86.83	91.54	92.16	91.94	**94.35**
	Kappa(%)	89.07	86.88	<u>91.16</u>	88.06	90.84	89.62	83.84	88.93	90.53	90.94	**93.17**
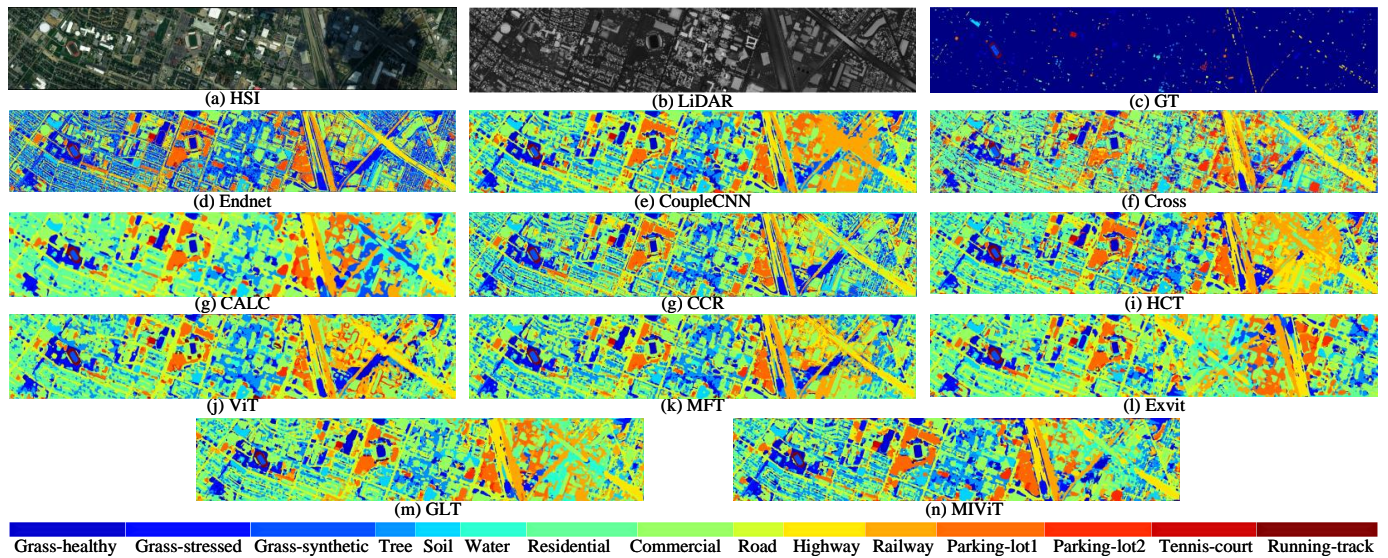
Fig. 7. Classification maps for the Trento data obtained with different methods. (a) Pseudocolor image for HSI. (b) LiDAR-based DSM. (c) Ground-truth map. (d) Endnet (89.93%). (e) CoupleCNN (87.91%). (f) Cross (91.84%). (g) CALC (88.97%). (h) CCR (91.56%). (i) HCT (90.43%). (j) ViT (85.05%). (k) MFT (89.80%). (l) ExViT (91.27%).GLT (91.64%). (m) MIViT (93.71%).

models, due to their learnable feature extraction capabilities, capture nonlinear relationships between input and output feature maps, producing relatively smooth classification maps. Transformer-based models, like ViT, MFT, and ExViT, excel in hyperspectral classification by extracting high-level abstract representations, resulting in superior visual quality in classification maps compared to CNNs. The MIViT, by more efficiently reducing redundancy between modalities and enhancing the perception of independently extracted modality features, produces highly accurate classification maps, particularly when compared to ViT. This indicates that MIViT facilitates more effective information interaction, allowing for

better fusion of spectral data from HSI with elevation data from LiDAR.

On the MUUFL dataset, as shown in Fig. 8, our MIViT method achieves the best visual performance, closely resembling the ground truth map. It effectively utilizes high-level semantic information, particularly in the classification of Trees and Yellow curbs, which appear more continuous compared to other methods. Notably, our method avoids issues like the omission of the Cloth panels category, a problem observed in the CALC method.

Fig. 9 offers a similar perspective, displaying classification maps for different methods on the Trento dataset. The classification results from MIViT exhibit smoother class transitions

TABLE III
OA, AA, AND KAPPA ON THE MUUFL DATASET BY CONSIDERING HSI AND LiDAR DATA. THE BEST RESULT IS **HIGHLIGHTED** AND THE SECOND RESULT IS <u>UNDERLINED</u>

No.	Class(Train/Test)	Endnet	CoupleCNN	Cross	CALC	CCR	HCT	ViT	MFT	ExViT	GLT	MIViT
1	Trees(1162/22084)	95.27	97.56	**98.55**	97.31	97.44	97.40	97.85	97.90	<u>98.47</u>	98.12	98.35
2	Mostly grass(214/4056)	75.52	86.19	65.85	**93.00**	77.93	90.31	76.06	92.11	90.34	90.41	<u>92.46</u>
3	Mixed ground surface(344/6538)	78.53	90.47	79.24	91.57	82.90	85.62	87.58	91.80	90.81	**91.94**	<u>91.92</u>
4	Dirt and sand(91/1735)	87.32	**96.25**	69.22	<u>95.10</u>	88.13	91.59	92.05	91.59	92.39	94.52	94.81
5	Road(334/6353)	94.16	**97.37**	<u>97.12</u>	95.91	96.49	94.66	94.71	95.60	93.69	95.73	94.13
6	Water(23/443)	90.07	95.03	60.95	<u>99.32</u>	94.36	97.52	82.02	88.19	90.97	95.49	**100.00**
7	Building shadow(112/2121)	69.31	88.21	64.55	92.69	83.36	86.28	87.11	90.27	91.37	<u>93.16</u>	**93.82**
8	Building(312/5928)	93.35	98.25	90.30	**98.45**	97.93	96.66	97.60	97.26	98.11	98.21	<u>98.41</u>
9	Sidewalk(69/1316)	64.82	42.17	32.75	51.60	52.89	55.62	57.83	61.35	66.57	<u>74.39</u>	**81.31**
10	Yellow curb(9/174)	<u>51.72</u>	3.45	0.00	0.00	2.30	21.26	31.99	17.43	32.76	28.74	**56.32**
11	Cloth panels(13/256)	92.19	76.17	43.36	0.00	85.16	69.14	58.72	72.79	85.94	<u>92.97</u>	**98.05**
	OA(%)	88.85	93.50	87.29	93.94	91.50	92.76	92.15	94.34	94.53	<u>95.20</u>	**95.66**
	AA(%)	81.11	79.24	63.81	74.09	78.08	80.55	78.50	81.48	84.67	<u>86.70</u>	**90.87**
	Kappa(%)	85.25	91.40	82.75	92.00	88.74	90.42	89.56	92.51	92.76	<u>93.67</u>	**94.27**

.

(a) HSI   (b) LiDAR   (c) GT   (d) Endnet   (e) CoupleCNN   (f) Cross   (g) CALC

(h) CCR   (i) HCT   (j) ViT   (k) MFT   (l) Exvit   (m) GLT   (n) MIViT

Trees   Mostly grass   Mixed ground surface   Dirt and sand   Road   Water   Building shadow   Building   Sidewalk   Yellow curb   Cloth panels
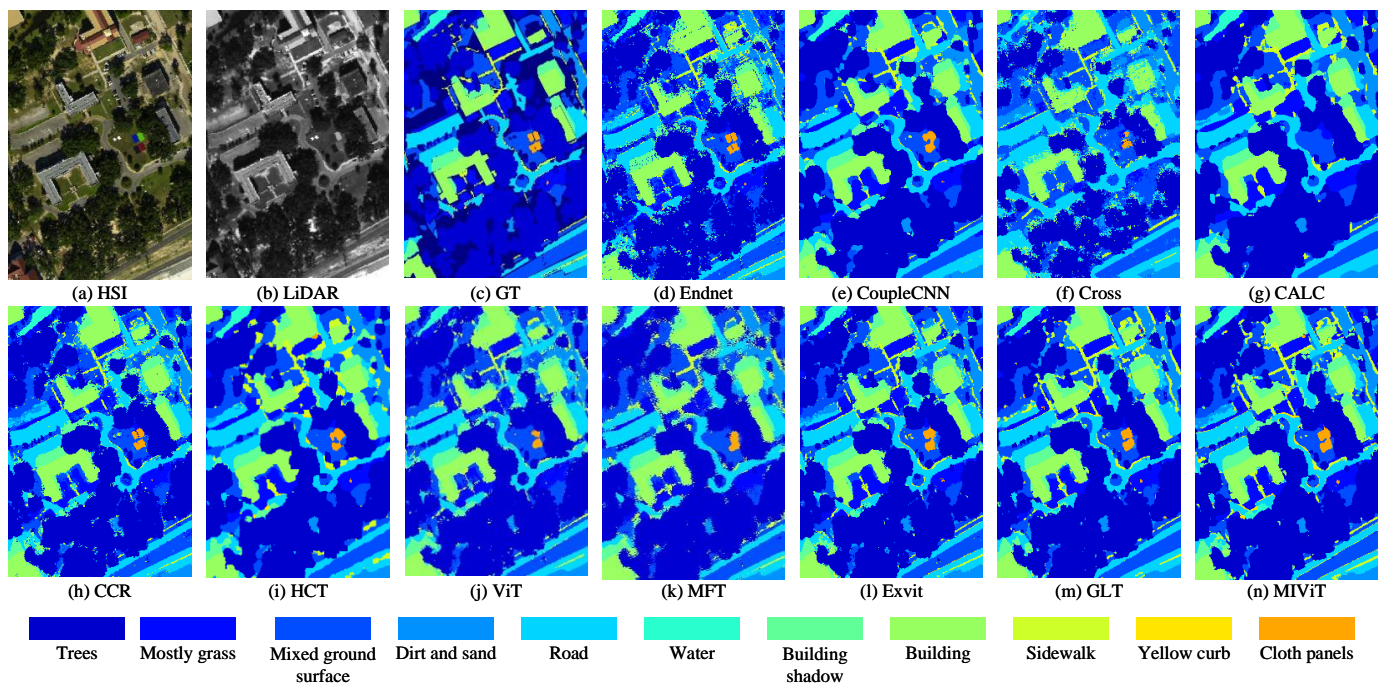
Fig. 8. Classification maps for the MUUFL data obtained with different methods. (a) Pseudocolor image for HSI. (b) LiDAR-based DSM. (c) Ground-truth map. (d) Endnet (88.85%). (e) CoupleCNN (93.50%). (f) Cross (87.29%). (g) CALC (93.94%). (h) CCR (91.50%). (i) HCT (92.76%). (j) ViT (92.15%). (k) MFT (94.34%). (l) ExViT (94.53%).GLT (95.20%). (m) MIViT (95.66%).

and less noise. Misclassification, a common challenge in such tasks, is markedly reduced in our method. Categories like Apple trees and Vineyards, often prone to confusion, are more accurately classified by our method. While other methods show misclassifications in categories like Roads, MIViT achieves completely accurate results. Additionally, in the Ground category, our method's classification performance is significantly superior to other state-of-the-art classifiers.

D. Ablation experiments

In this section, we conducted ablation studies to further validate the effectiveness of our model's components. Specifically,

we designed four scenarios to analyze the impact of key modules: (1) Exclusion of the Oriented Attention Fusion (OAF) and its replacement with a simple addition operation; (2) Removal of the Transformer Reconstruction Decoder (TRD) in the reconstruction stage; (3) Exclusion of the Information Aggregation Constraint (IAC) from the aggregation training of separated features and fused representations; (4) Omission of distillation from the information distribution flow (IDF) in the distribution training.

Effect of transformer reconstruction decoder module. We integrated a feature reconstruction transformer encoder (TRD) into our fusion process to minimize information loss

TABLE IV

OA, AA, AND KAPPA ON THE TRENTO DATASET BY CONSIDERING HSI AND LiDAR DATA. THE BEST RESULT IS **HIGHLIGHTED** AND THE SECOND RESULT IS <u>UNDERLINED</u>

| No.	Class(Train/Test)	Endnet	CoupleCNN	Cross	CALC	CCR	HCT	ViT	MFT	ExViT	GLT	MIViT
1	Apple trees(129/3905)	92.88	98.98	<u>99.87</u>	98.62	**100.00**	99.26	89.07	98.20	99.13	99.23	99.59
2	Buildings(125/2778)	97.05	96.83	98.63	**99.96**	98.67	**99.96**	84.74	91.65	98.56	98.74	<u>99.53</u>
3	Ground(105/374)	96.52	95.45	98.93	72.99	90.37	**100.00**	92.25	94.92	77.81	<u>99.20</u>	**100.00**
4	Woods(154/8969)	99.30	99.87	99.22	**100.00**	<u>99.91</u>	**100.00**	99.63	99.45	**100.00**	**100.00**	**100.00**
5	Vineyard(184/10317)	80.11	**100.00**	98.05	99.44	94.76	67.47	98.23	99.72	<u>99.92</u>	**100.00**	99.85
6	Roads(122/3052)	89.02	93.77	89.42	88.76	85.39	83.06	84.04	91.52	91.78	<u>95.35</u>	**95.97**
	OA(%)	90.40	98.82	97.82	98.11	96.37	86.72	94.62	97.76	98.58	<u>99.29</u>	**99.43**
	AA(%)	92.48	97.48	97.35	93.30	94.85	91.63	91.33	95.91	94.53	<u>98.75</u>	**99.16**
	Kappa(%)	87.35	98.42	97.09	97.46	95.16	82.95	92.81	97.00	98.10	<u>99.04</u>	**99.24**
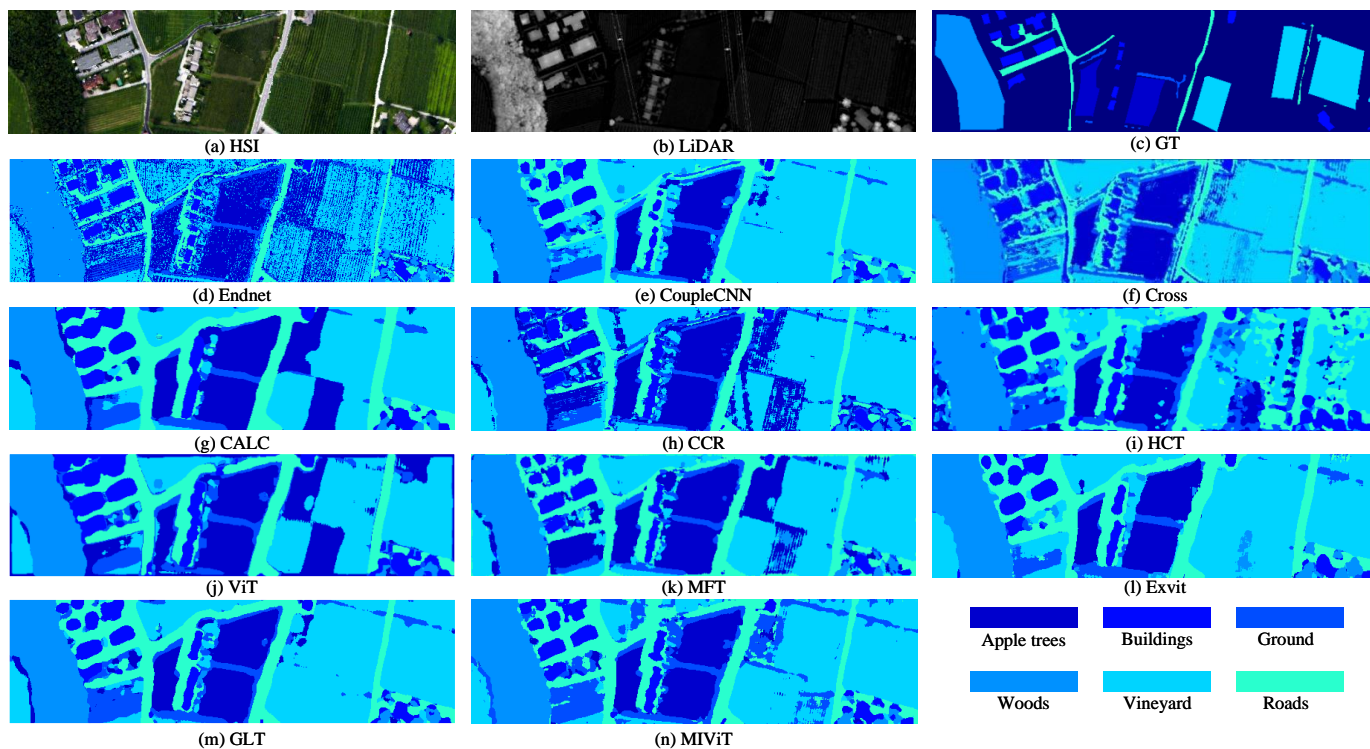
Fig. 9. Classification maps for the Trento data obtained with different methods. (a) Pseudocolor image for HSI. (b) LiDAR-based DSM. (c) Ground-truth map. (d) Endnet (90.40%). (e) CoupleCNN (98.82%). (f) Cross (97.82%). (g) CALC (98.11%). (h) CCR (96.37%). (i) HCT (86.72%). (j) ViT (94.62%). (k) MFT (97.76%). (l) ExViT (98.58%).GLT (99.29%). (m) MIViT (99.43%).

from different modalities during fusion, as illustrated in Fig. 2. To assess TRD's effectiveness in feature recognition tasks, we conducted ablation experiments, using final classification accuracy as the metric. Fig. 10 evaluates the impact of TRD on OA, AA, and Kappa statistics across three datasets. The inclusion of TRD led to improvements in all performance metrics. This suggests that focusing more on discriminative spatial features is beneficial for improving classification accuracy, particularly for HSI and LiDAR data. Notably, the most significant improvement was observed for the Houston dataset, with AA improved by nearly 1.6%. The Houston dataset contains a wide variety of complex ground object categories, resulting in significant variations in the probabilities for each classification. In such cases, TRD can have a more pronounced effect. This highlights TRD's significant role in enhancing

classification accuracy, especially for complex ground object categories.

Effect of oriented attention fusion module. As shown in Fig. 2, we incorporated oriented attention fusion (OAF) to facilitate the effective extraction and fusion of multi-layer features from different modalities, allowing for information interaction in various directions. To demonstrate OAF's effectiveness, we conducted ablation experiments, replacing OAF with a simple addition operation. Fig. 11 shows that the inclusion of OAF led to improvements in overall performance metrics across all datasets, confirming its advantage in enhancing model performance.

Effect of information aggregation constraint module. Fig. 2 depicts our introduction of IAC to reduce redundancy between modalities and ensure consistent information in fusion
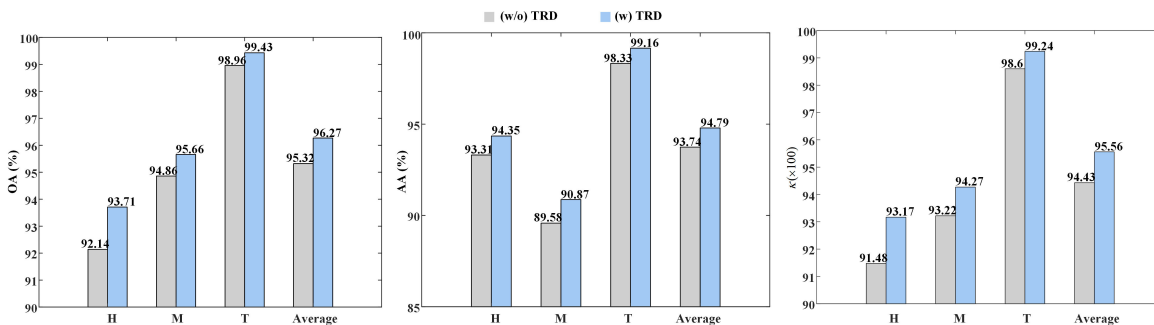
Fig. 10. Effects of transformer reconstruction decoder (TRD) on the Houston2013(H), MUUFL(M), Trento(T) datasets.
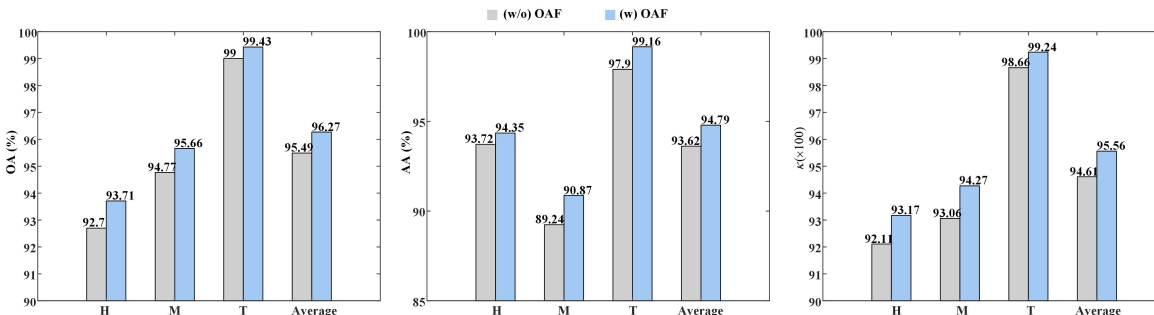

Fig. 11. Effects of oriented attention fusion (OAF) on the Houston2013(H), MUUFL(M), Trento(T) datasets.

representations. Ablation experiments assessing the impact of IAC on OA, AA, and Kappa are presented in Fig. 13. The addition of IAC led to enhancements in all metrics. This indicates that paying more attention to discriminative spatial features contributes to improved classification accuracy. The improvement in AA is relatively pronounced, especially for the Trento dataset, as IAC effectively alleviates the phenomenon of non-uniform classification difficulty. In such cases, IAC can play a more significant role. These results suggest that IAC indeed makes a positive contribution to the classification task. To visualize the efficacy of the IAC module in eliminating redundancy, we generated scatter plots for features extracted from different modalities. A more scattered distribution indicates a lower correlation between two features, implying less intermingled redundant information. Upon inspecting Fig 12, it is evident that the model incorporating the IAC module exhibits a more dispersed distribution among different features. This observation suggests that the IAC module plays a crucial role in effectively reducing redundancy in multimodal information.

Effect of information distribution flow module. As illustrated in Eq. 9, the final classification result is achieved through the decision fusion of CNN and ViT encoders, aiming for a globally optimal classification outcome. Individual modality-specific classification results are computed using Eq. 7. The globally optimal classification result exhibits strong performance awareness and, through the IDF loss based on self-distillation, can enhance the performance awareness of individual modalities by distributing information across different modalities. The IDF module's effectiveness was validated through ablation experiments, with the impact evaluated in terms of OA, AA, and Kappa (Fig. 14). With the inclusion
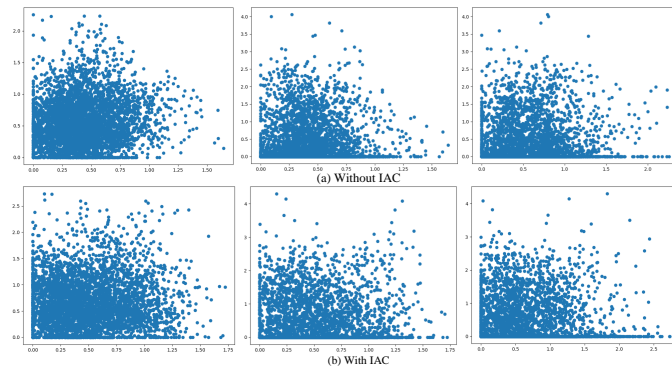

Fig. 12. The distribution visualization of correlation relationship between the separated features from different modalities.
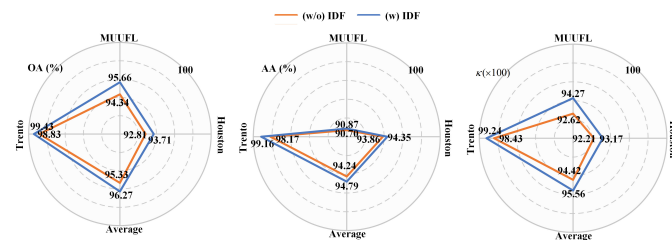

Fig. 13. Effects of information aggregation constraint (IAC) on the Houston2013, MUUFL, Trento datasets.

of the IDF module, shallow, individual modality features are paired with their respective classifiers. When the IDF module is omitted, these classifiers are removed. The inclusion of IDF improved all metrics, demonstrating its role in enhancing the perceptual capabilities of separated modality features and thereby strengthening fusion representation classification results. This confirms IDF's positive impact on the classification
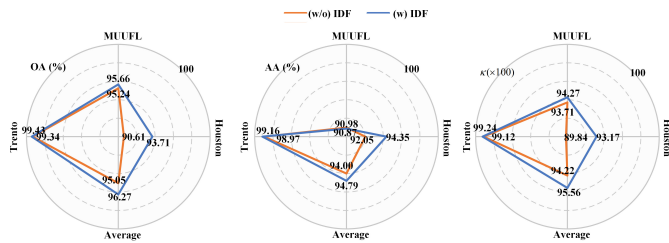
Fig. 14. Effects of information distribution flow (IDF) on the Houston2013, MUUFL, Trento datasets.

task.

E. Missing Multimodal Learning

Fig. 15 presents an analysis of the features across different classifiers. Classifiers 1/3 and 2/3 represent features of shallow separated modalities, while Classifier 3/3 denotes the latent fused representation. A notable observation is the enhanced discriminative performance of classifiers incorporating the Information Distribution Flow (IDF) compared to those without it. Specifically, Class 2 shows a clear distinction from other classes in the two-dimensional projection space. This underscores the significant accuracy enhancement achieved through IDF, emphasizing its crucial role in land cover classification tasks.
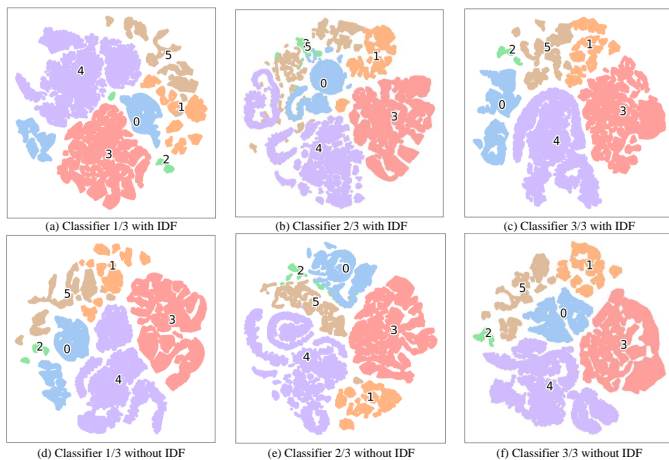

Fig. 15. tSNE visualization of feature distribution in different classifiers.

TABLE V

ACCURACY MEASUREMENT OF THE DIFFERENT CLASSIFIERS (CLASSIFIER 1/3, 2/3 AND 3/3) IN MIVIT. THE BEST RESULT OF EACH MODALITY IS **HIGHLIGHTED**.

Modality	Method	Houston			MUUFL			Trento		
		OA	AA	Kappa	OA	AA	Kappa	OA	AA	Kappa
HSI	Cross	83.73	86.62	82.41	86.22	64.72	81.69	87.59	87.32	83.68
	1/3	**88.47**	**89.88**	**87.49**	**92.43**	**82.92**	**89.99**	**96.87**	**94.94**	**95.80**
LiDAR	Cross	62.01	63.45	58.90	**70.19**	39.65	**59.84**	84.74	82.95	80.24
	2/3	**64.49**	**65.84**	**61.47**	69.41	**43.15**	57.56	**95.60**	**87.54**	**94.15**
Fusion	Cross	91.84	92.70	91.16	87.29	63.81	82.75	97.82	97.35	97.09
	MFT	89.80	91.51	88.93	94.34	81.48	92.51	97.76	95.91	97.00
	3/3	**93.71**	**94.35**	**93.17**	**95.66**	**90.87**	**94.27**	**99.43**	**99.16**	**99.24**

Table V summarizes the accuracy of each classifier, revealing that the global classifier (3/3) achieves higher classification accuracy (96.15% OA, 91.01% AA, and 95.04% Kappa) compared to the single-modality classifiers (1/3 and 2/3). This indicates that multimodal feature-level fusion markedly improves accuracy, highlighting its significance in remote sensing tasks. However, the classifier for HSI (1/3) records suboptimal accuracy, likely due to its wide spectral coverage and diverse spectral information. When compared with Cross [50], which specifically tackles missing multimodal learning, our method shows substantial improvements across three datasets and various modalities. For instance, on the Trento dataset, our approach achieved a 9.28% increase in OA, 7.35% in AA, and 12.12% in Kappa for the HSI modality. In the LiDAR modality, the improvements are even more pronounced, with increases of 10.86% in OA, 4.59% in AA, and 5.08% in Kappa. This demonstrates that our algorithm effectively channels global fusion information into modalities, thereby enhancing the performance of single-modal detectors. Moreover, we evaluated the complexity and computational load of different modal detectors within our algorithm, as shown in Table VI. Parameters (Paras) indicate the model size, while FLOPs measure computational complexity. The high-dimensional characteristics of HSI require more parameters and computational resources than LiDAR data. Our study introduces an innovative approach to addressing missing modalities. By using a global deep detector to distill information from shallow classifiers, we enhance the extraction of detailed information during the feature extraction process of shallow single modalities and augment semantic information. In cases of missing modalities, the classification task can still be effectively accomplished using the feature extractor and classifier from another modality.

TABLE VI

COMPARISON OF MODEL COMPLEXITY OF DIFFERENT MODALITIES ON ALL THREE DATASETS.

Modality	Method	Params/K	FLOPs/M
HSI	Classifier 1/3	84.48	80.45
LiDAR	Classifier 2/3	58.85	34.51
Fusion	Classifier 3/3	1867.00	900.12

V. CONCLUSIONS

In summary, our proposed Multimodal Informative ViT (MIViT) offers a highly effective solution for addressing empirical redundancy in multimodal land cover classification (MLCC). MIViT integrates an innovative information aggregate-distributing mechanism, redefining the redundancy level and enhancing performance-awareness in fused representations. This approach promotes a more efficient integration of complementary features from each modality. Central to MIViT's effectiveness is its bidirectional learning process, encompassing both forward and backward directions, which enables the fused representation to comprehensively capture semantic information. The incorporation of Oriented Attention Fusion (OAF) and a Transformer feature extractor is pivotal in

extracting both shallow local and deep global features, respectively. Furthermore, the implementation of an Information Aggregation Constraint (IAC), grounded in mutual information, adeptly reduces redundant information while preserving the correlation of complementary information. The Information Distribution Flow (IDF) augments the performance awareness of individual modalities, distributing global classification information across various modality feature maps. This sophisticated architecture is also adept at handling missing modality challenges, utilizing lightweight independent modality classifiers. Our extensive experimental analyses validate the superiority of MIViT, demonstrating its remarkable improvements over existing state-of-the-art methods across five multimodal datasets. Achieving an overall accuracy of 95.56%, MIViT establishes itself as a robust and innovative approach for multimodal land cover classification, effectively bridging the gap between multiple data sources and accurate, efficient classification.

REFERENCES

[1] J. Xie, N. He, L. Fang, and P. Ghamisi, "Multiscale densely-connected fusion networks for hyperspectral images classification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 1, pp. 246–259, 2020.

[2] H. Liu, Y. Jia, J. Hou, and Q. Zhang, "Global-local balanced low-rank approximation of hyperspectral images for classification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 4, pp. 2013–2024, 2021.

[3] J. Wang, J. Li, Y. Shi, J. Lai, and X. Tan, "Am^{3}net: Adaptive mutual-learning-based multimodal data fusion network," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 8, pp. 5411–5426, 2022.

[4] W. G. C. Bandara and V. M. Patel, "Hypertransformer: A textural and spectral feature fusion transformer for pansharpening," in *Proc. 922 IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2022, pp. 1767–1777.

[5] S. Xu, J. Zhang, Z. Zhao, K. Sun, J. Liu, and C. Zhang, "Deep gradient projection networks for pan-sharpening," in *Proc. 922 IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2021, pp. 1366–1375.

[6] Z. Zhao, H. Bai, J. Zhang, Y. Zhang, S. Xu, Z. Lin, R. Timofte, and L. Van Gool, "Cddfuse: Correlation-driven dual-branch feature decomposition for multi-modality image fusion," in *Proc. 922 IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2023, pp. 5906–5916.

[7] S. Mohla, S. Pande, B. Banerjee, and S. Chaudhuri, "Fusatnet: Dual attention based spectrospatial multimodal fusion network for hyperspectral and lidar classification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020, pp. 92–93.

[8] S. K. Roy, A. Deria, C. Shah, J. M. Haut, Q. Du, and A. Plaza, "Spectral–spatial morphological attention transformer for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–15, 2023.

[9] J. Wang and X. Tan, "Mutually beneficial transformer for multimodal data fusion," *IEEE Trans. Circuits Syst. Video Technol.*, 2023.

[10] S. K. Roy, A. Deria, D. Hong, B. Rasti, A. Plaza, and J. Chanussot, "Multimodal fusion transformer for remote sensing image classification," *IEEE Trans. Geosci. Remote Sens.*, 2023.

[11] K. Ding, T. Lu, W. Fu, S. Li, and F. Ma, "Global–local transformer network for hsi and lidar data joint classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–13, 2022.

[12] T. Lu, K. Ding, W. Fu, S. Li, and A. Guo, "Coupled adversarial learning for fusion classification of hyperspectral and lidar data," *Inf. Fusion*, vol. 93, pp. 118–131, 2023.

[13] Y. Chen, C. Li, P. Ghamisi, X. Jia, and Y. Gu, "Deep fusion of remote sensing data for accurate classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 8, pp. 1253–1257, 2017.

[14] X. Xu, W. Li, Q. Ran, Q. Du, L. Gao, and B. Zhang, "Multisource remote sensing data classification based on convolutional neural network," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 2, pp. 937–949, 2017.

[15] D. Hong, L. Gao, R. Hang, B. Zhang, and J. Chanussot, "Deep encoder–decoder networks for classification of hyperspectral and lidar data," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2020.

[16] X. Zhao, R. Tao, W. Li, H.-C. Li, Q. Du, W. Liao, and W. Philips, "Joint classification of hyperspectral and lidar data using hierarchical random walk and deep cnn architecture," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 10, pp. 7355–7370, 2020.

[17] X. Wu, D. Hong, and J. Chanussot, "Convolutional neural networks for multimodal remote sensing data classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–10, 2021.

[18] W. Dong, T. Yang, J. Qu, T. Zhang, S. Xiao, and Y. Li, "Joint contextual representation model-informed interpretable network with dictionary aligning for hyperspectral and lidar classification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 11, pp. 6804–6818, 2023.

[19] W. Dong, T. Zhang, J. Qu, S. Xiao, T. Zhang, and Y. Li, "Multibranch feature fusion network with self-and cross-guided attention for hyperspectral and lidar classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–12, 2022.

[20] X. Zhang, Y. Wang, N. Zhang, D. Xu, H. Luo, B. Chen, and G. Ben, "Spectral–spatial fractal residual convolutional neural network with data balance augmentation for hyperspectral classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 12, pp. 10 473–10 487, 2021.

[21] D. Li, W. Xie, Y. Li, and L. Fang, "Fedfusion: Manifold driven federated learning for multi-satellite and multi-modality fusion," *arXiv preprint arXiv:2311.09540*, 2023.

[22] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, and R. Timofte, "Swinir: Image restoration using swin transformer," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2021, pp. 1833–1844.

[23] L. Zhang, J. Song, A. Gao, J. Chen, C. Bao, and K. Ma, "Be your own teacher: Improve the performance of convolutional neural networks via self distillation," in *Proc. 922 IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2019, pp. 3713–3722.

[24] J. Yao, B. Zhang, C. Li, D. Hong, and J. Chanussot, "Extended vision transformer (exvit) for land use and land cover classification: A multimodal deep learning framework," *IEEE Trans. Geosci. Remote Sens.*, 2023.

[25] X. Zhao, M. Zhang, R. Tao, W. Li, W. Liao, L. Tian, and W. Philips, "Fractional fourier image transformer for multimodal remote sensing data classification," *IEEE Trans. Neural Netw. Learn. Syst.*, 2022.

[26] G. Zhao, Q. Ye, L. Sun, Z. Wu, C. Pan, and B. Jeon, "Joint classification of hyperspectral and lidar data using a hierarchical cnn and transformer," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–16, 2022.

[27] M. Zhang, C. Zhang, Q. Zhang, J. Guo, X. Gao, and J. Zhang, "Essaformer: Efficient transformer for hyperspectral image super-resolution," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2023, pp. 23 073–23 084.

[28] D. Li, W. Xie, J. Zhang, and Y. Li, "MDFL: Multi-domain Diffusion-driven Feature Learning," *arXiv e-prints*, p. arXiv:2311.09520, Nov. 2023.

[29] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2020.

[30] R. Bramon, I. Boada, A. Bardera, J. Rodriguez, M. Feixas, J. Puig, and M. Sbert, "Multimodal data fusion based on mutual information," *IEEE Trans. Vis. Comput. Graphics*, vol. 18, no. 9, pp. 1574–1587, 2011.

[31] B. Xi, J. Li, Y. Diao, Y. Li, Z. Li, Y. Huang, and J. Chanussot, "Dgssc: A deep generative spectral-spatial classifier for imbalanced hyperspectral imagery," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 4, pp. 1535–1548, 2022.

[32] W. Li, Y. Gao, M. Zhang, R. Tao, and Q. Du, "Asymmetric feature fusion network for hyperspectral and sar image classification," *IEEE Trans. Neural Netw. Learn. Syst.*, 2022.

[33] Z. Han, C. Zhang, H. Fu, and J. T. Zhou, "Trusted multi-view classification with dynamic evidential fusion," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 2, pp. 2551–2566, 2022.

[34] M. Wang, F. Gao, J. Dong, H.-C. Li, and Q. Du, "Nearest neighbor-based contrastive learning for hyperspectral and lidar data classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–16, 2023.

[35] R. Hang, Z. Li, P. Ghamisi, D. Hong, G. Xia, and Q. Liu, "Classification of hyperspectral and lidar data using coupled cnns," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 7, pp. 4939–4950, 2020.

[36] C. Zhao, L. Wu, P. Shao, K. Zhang, R. Hong, and M. Wang, "Fair representation learning for recommendation: A mutual information perspective," *Proc. AAAI Conf. Artif. Intell. (AAAI)*, vol. 37, no. 4, pp. 4911–4919, Jun. 2023.

[37] X. Wu, X. Dong, T. Nguyen, C. Liu, L.-M. Pan, and A. T. Luu, "Infoctm: A mutual information maximization perspective of cross-lingual topic modeling," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, vol. 37, no. 11, 2023, pp. 13 763–13 771.

[38] W. Su, X. Zhu, C. Tao, L. Lu, B. Li, G. Huang, Y. Qiao, X. Wang, J. Zhou, and J. Dai, "Towards all-in-one pre-training via maximizing multi-modal mutual information," in *Proc. 922 IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2023, pp. 15 888–15 899.

[39] Y. Mao, X. Yan, Q. Guo, and Y. Ye, "Deep mutual information maximin for cross-modal clustering," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, vol. 35, no. 10, 2021, pp. 8893–8901.

[40] A. Sanghi, "Info3d: Representation learning on 3d objects using mutual information maximization and contrastive learning," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Springer, 2020, pp. 626–642.

[41] J. Zhang, D.-P. Fan, Y. Dai, X. Yu, Y. Zhong, N. Barnes, and L. Shao, "Rgb-d saliency detection via cascaded mutual information minimization," in *Proc. 922 IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2021, pp. 4338–4347.

[42] M. Zhou, K. Yan, J. Huang, Z. Yang, X. Fu, and F. Zhao, "Mutual information-driven pan-sharpening," in *Proc. 922 IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2022, pp. 1798–1808.

[43] F.-Z. Ou, X. Chen, R. Zhang, Y. Huang, S. Li, J. Li, Y. Li, L. Cao, and Y.-G. Wang, "Sdd-fiqa: unsupervised face image quality assessment with similarity distribution distance," in *Proc. 922 IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2021, pp. 7670–7679.

[44] L. Zhang, L. Song, B. Du, and Y. Zhang, "Nonlocal low-rank tensor completion for visual data," *IEEE Trans. on Cybern.*, vol. 51, no. 2, pp. 673–685, 2019.

[45] C.-F. R. Chen, Q. Fan, and R. Panda, "Crossvit: Cross-attention multi-scale vision transformer for image classification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2021, pp. 357–366.

[46] M. Zhang, R. Zhang, Y. Yang, H. Bai, J. Zhang, and J. Guo, "Isnet: Shape matters for infrared small target detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 877–886.

[47] S. Woo, S. Lee, Y. Park, M. A. Nugroho, and C. Kim, "Towards good practices for missing modality robust action recognition," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, vol. 37, no. 3, 2023, pp. 2776–2784.

[48] M. Ma, J. Ren, L. Zhao, S. Tulyakov, C. Wu, and X. Peng, "Smil: Multimodal learning with severely missing modality," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, vol. 35, no. 3, 2021, pp. 2302–2310.

[49] Y.-L. Lee, Y.-H. Tsai, W.-C. Chiu, and C.-Y. Lee, "Multimodal prompting with missing modalities for visual recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2023, pp. 14 943–14 952.

[50] D. Hong, L. Gao, N. Yokoya, J. Yao, J. Chanussot, Q. Du, and B. Zhang, "More diverse means better: Multimodal deep learning meets remote-sensing imagery classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 5, pp. 4340–4354, 2020.

Weiying Xie (M'18) received a B.S. degree in electronic information science and technology from the University of Jinan in 2011. She received an M.S. degree in communication and information systems, Lanzhou University in 2014, and the Ph.D. degree in communication and information systems of Xidian University in 2017. Currently, she is an Associate Professor with the State Key Laboratory of Integrated Services Networks, Xidian University.

Her research interests include neural networks, machine learning, hyperspectral image processing, and high-performance computing.
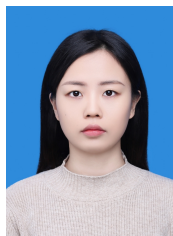
Geng Yang received the B.E. degree in Telecommunications Engineering from Xidian University, Xi'an, China in 2019. She is currently pursuing the Ph.D. degree in the Image Coding and Processing Center at State Key Laboratory of Integrated Services Networks, Xidian University, Xi'an, China. Her research interests include remote sensing image processing, computer vision, efficient deep learning.

Daixun Li received the B.E. degree in Telecommunications Engineering from Xidian University, Xi'an, China in 2022. He is currently pursuing the M.S. degree with the Image Coding and Processing Center at State Key Laboratory of Integrated Service Network, Xidian University, Xi'an, China. His research interests include multimodal image processing, remote sensing classification and federated learning.

Jiaqing Zhang received the B.E. degree in Telecommunications Engineering from Ningbo University, Zhejiang, China in 2019. She is currently pursuing the Ph.D. degree with the Image Coding and Processing Center at State Key Laboratory of Integrated Service Network, Xidian University, Xi'an, China.

Her research interests include multimodal image processing, remote sensing object detection and network compression.

Jie Lei (M'19) received his M.S. degree in Telecommunications and Information Systems and his Ph.D. degree in Signal and Information Processing from Xidian University, China, in 2006 and 2010, respectively. He was a Visiting Scholar at the Department of Computer Science at the University of California, Los Angeles, USA, from 2014 to 2015. He served as a Professor at the School of Telecommunications Engineering, Xidian University, until 2023. Currently, he is a Research Fellow at the School of Electrical and Data Engineering at the University of Technology Sydney.

His research interests include wireless communication, remote sensing image processing, machine learning, and customized computing for big data applications.

Yunsong Li received the M.S. degree in telecommunication and information systems and the Ph.D. degree in signal and information processing from Xidian University, China, in 1999 and 2002, respectively. He joins the school of Telecommunications Engineering, Xidian University in 1999 where he is currently a Professor. Prof. Li is the director of the image coding and processing center at the State Key Laboratory of Integrated Service Networks.

His research interests focus on image and video processing and high-performance computing.