Language-Aware Domain Generalization Network for Cross-Scene Hyperspectral Image Classification

Yuxiang Zhang, Student Member, IEEE, Mengmeng Zhang®, Wei Li®, Senior Member, IEEE, Shuai Wang, and Ran Tao®, Senior Member, IEEE

Abstract—Text information including extensive prior knowledge about land cover classes has been ignored in hyperspectral image (HSI) classification tasks. It is necessary to explore the effectiveness of linguistic mode in assisting HSI classification. In addition, the large-scale pretraining image-text foundation models have demonstrated great performance in a variety of downstream applications, including zero-shot transfer. However, most domain generalization methods have never addressed mining linguistic modal knowledge to improve the generalization performance of model. To compensate for the inadequacies listed above, a language-aware domain generalization network (LDGnet) is proposed to learn cross-domain-invariant representation from cross-domain shared prior knowledge. The proposed method only trains on the source domain (SD) and then transfers the model to the target domain (TD). The dual-stream architecture including the image encoder and text encoder is used to extract visual and linguistic features, in which coarse-grained and fine-grained text representations are designed to extract two levels of linguistic features. Furthermore, linguistic features are used as cross-domain shared semantic space, and visual–linguistic alignment is completed by supervised contrastive learning in semantic space. Extensive experiments on three datasets demonstrate the superiority of the proposed method when compared with the state-of-the-art techniques. The codes will be available from the website: https://github.com/YuxiangZhang-BIT/IEEE_TGRS_LDGnet.

Index Terms—Contrastive learning, cross-scene, domain generalization, hyperspectral image (HSI) classification, multimodality, natural language supervision.

I. INTRODUCTION

INSPIRED by the success of deep learning (DL) technology, remote sensing image classification based on advanced convolutional neural networks (CNNs) has received extensive attention and achieved excellent performance, particularly in hyperspectral image (HSI) classification [1], [2], [3], [4], [5], [6], [7], [8], [9]. However, most of the CNN-based classification methods are only suitable for fixed scenes, that is, training samples and testing samples are independently and identically distributed. The acquisition process of HSI is inevitably affected by various factors, such as sensor nonlinearities, seasonal, and weather conditions, which lead to variations in spectral reflectance between the source domain (SD) and the target domain (TD) of the same land cover classes. As a result, classification based on CNN has high generalization error and poor interpretation effect in the cross-scene classification task.

It is encouraging that transfer learning helps solve this problem, where domain adaptation (DA)-based approaches are widely used for cross-scene classification [10], [11], [12], [13]. Many DA methods have been developed for cross-scene classification from the perspective of DA, most notably related methods based on statistics, subspace learning, active learning, or DL. Among them, the maximum mean discrepancy (MMD) criterion [14] was the earliest statistical technique used in the cross-scene interpretation. To address the issue of dynamic distribution adaptation in an adversarial network, Yu et al. [15] developed the dynamic adversarial adaptation network (DAAN). Classwise distribution adaptation was created for HSI cross-scene classification [16], and the MMD method based on probability prediction was used in an adversarial adaptation network to obtain more accurate feature alignment.

In recent years, a more difficult task setting, domain generalization (DG), has been developed in computer vision. The training samples of DA are labeled SD and unlabeled TD, that is, TD is accessed by the model during training, while the training samples of DG are only labeled SD, and TD is not allowed to be accessed. The objective of DG is to learn a model from one or several different but related domains (i.e., diverse training datasets) that generalize well on TD [17]. Zhou et al. [18] adversarially trained a transformation network for data augmentation instead of directly updating the inputs by gradient ascent. To replicate diverse photometric and geometric alterations in TD, Li et al. [19] created the progressive domain expansion network (PDEN), a learning architecture that gradually produces multiple domains.

Remote sensing images combined with language to complete tasks such as image captioning, image classification, and image retrieval is one of the main research directions. Extracting visual features from remote sensing images and generating descriptions has become a challenging task. Zia et al. [20] proposed a technique that uses topic-sensitive word embedding to handle multiscale data using adaptive attention-based decoding
to create innovative captions. Lobry et al. [21] constructed two datasets of image/question/answer triplets, which can be used for training and evaluation models to solve remote sensing data visualization question answering (RSVQA) tasks. Great efforts have been made in the remote sensing community as a result of the rising need for captioning-related applications. To make remote sensing image captioning easier, a novel recurrent attention and semantic gate (RASG) framework is proposed [22], which combines competing visual elements with a recurrent attention mechanism.

At present, most DG works consider how to learn domain-invariant representation at the visual level from multiple SDs or from a single SD. They have never tried to use language knowledge to assist visual representation learning and realize generalization by visual–linguistic alignment. In addition, it has been proven that language is helpful for visual representation learning in multimodal learning [23], [24]. However, in HSI classification tasks, there is a lack of text information that is most able to reflect the prior knowledge of land cover classes. Therefore, how to construct a multimodal learning framework suitable for HSI is worthy of in-depth study.

To address the aforementioned problems, a straightforward multimodal DG framework for HSI, called language-aware domain generalization network (LDGnet), is proposed. Language knowledge is viewed as shared knowledge between SD and TD. The semantic space unifies visual modality and linguistic modality containing prior knowledge. A succinct visual–linguistic alignment strategy is designed to decrease the domain shift and increase the generalizability of the model. Specifically, LDGnet is split into three sections: image encoder, text encoder, and visual–linguistic alignment. In the visual modality, the deep residual 3-D CNN network is used by the image encoder to extract the visual features from the image patch. Second, to create semantic space in the linguistic modality, coarse- and fine-grained text descriptions are intended to supplement supervised signals with semantic information. The text encoder makes use of the universal language-model transformer. The core design of LDGnet is a visual–linguistic alignment strategy in which the semantic space is treated as a cross-domain shared space. To learn domain-invariant representations, supervised contrastive learning is used to project visual features into the semantic space.

The main contributions of this work are summarized as follows.

1) As far as we are aware, this is the first work to propose a language-guided framework for HSI classification, and the introduction of a linguistic mode with prior knowledge as a supervised signal improved visual representation learning.

2) The semantic space composed of linguistic features is regarded as cross-domain shared space, and the visual features are projected to the semantic space by visual–linguistic alignment to minimize the disparity across domains.

3) In the semantic space, coarse-grained and fine-grained text representations are designed, which enriches semantic supervised signals and promotes domain-invariant representation learning.

The rest of this article is organized as follows. Section II introduces relevant concepts of DG and image–text foundation models. Section III elaborates on the proposed LDGnet. The extensive experiments and analyses are presented in Section IV. Finally, conclusions are drawn in Section V.

II. RELATED WORK

A. Domain Generalization (DG)

DG is more challenging than DA, because DG aims to learn the model through SD data and does not need to access TD in the training phase. The model can be extended to TD in the inference stage. A comparison between DA and DG is listed in Table I, where \( S \) represents domain. The existing DG methods are divided into two categories: learning the domain-invariant representation and data manipulation.

The key idea of the first category is to reduce the domain shift between multiple SD domain representations, which is mainly applied to multisource DG. The most typical strategy is the explicit feature alignment. Some methods explicitly minimize the feature distribution divergence by minimizing the MMD [25], second-order correlation [26], and Wasserstein distance [27] of domains. Data manipulation is mainly applied to single-source DG. Such methods generally augment or generate out-of-domain samples related to SD, and then use these samples to train the model with the SD and transfer to TD. Data generation creates diversified and abundant data to help generalization. For example, variational auto-encoder (VAE) [28] and generative adversarial networks (GANs) [29] are often used for these purposes.

B. Image–Text Foundation Models

Recent researches have proposed image–text foundation models that learn strong joint representations between the two modalities through pretraining on large-scale image–text pairs and exhibit excellent transferability on some downstream visual and language tasks. The mainstream image–text foundation models are classified into two types based on their model architecture design: single stream (early fusion of the two input modalities) and dual stream (late fusion) [30]. The former one concatenates image and text input embeddings from VisualBERT [31] and unified-modal pre-training architecture (UNIMO) [32] to model both image and text representations in a single transformer in a single semantic space. The latter, however, uses a decoupled image encoder and a text encoder for high-level representations, such as contrastive language-image pre-training (CLIP) [23] and Coca [24], to encode

| Table I | COMPARISON BETWEEN DA AND DOMAIN GENERALIZATION |
|---|---|---|---|---|
| **Learning paradigm** | **Training data** | **Test data** | **Test access** |
| Domain adaptation | \( S^{scr}, S^{tar} \) | \( S^{tar} \) | ✓ |
| Single-source | \( S^{scr} \) | \( S^{tar} \) | × |
| Domain generalization | \( S^1, S^2, \ldots, S^n \) | \( S^{n+1} \) | × |
images and texts independently. Contrastive learning as training objectives is used by CLIP, DeCLIP, and Coca. The main principle of contrastive learning is to automatically build similar positive sample pairs and dissimilar negative sample pairs, with the goal of making positive sample pairs closer to one another in the projection space and negative sample pairs farther apart. In this article, the dual-stream architecture and contrastive learning are used for LDGnet for flexible design and relatively cheap computation.

III. PROPOSED LDGNET

Assume that \( X = \{ x_i \}_{i=1}^{N} \in \mathbb{R}^d \) is the data from SD, and \( Y = \{ y_i \}_{i=1}^{N} \) is the corresponding class labels. Here, \( d \) and \( N \) denote the dimension of data and the number of source samples, respectively. The proposed LDGnet is broken into three parts: an image encoder, a text encoder, and a visual–linguistic alignment. The flowchart for each part is illustrated in Fig. 1. The sample of \( 13 \times 13 \times d \) spatial patch in HSI is selected from SD, and each one is assigned a coarse-grained text and two fine-grained texts based on the class of the patch. In the visual mode, the image is delivered to the image encoder to extract features, and then the classification head Cls is used to calculate the cross-entropy loss, and the projection head Proj outputs visual features for visual–linguistic alignment. To extract the linguistic features and create the cross-domain shared semantic space, the text is passed to the text encoder. Finally, supervised contrastive learning is used for visual–linguistic alignment to align image and text by class.

A. Image Encoder

Given that HSI is a 3-D data cube made up of hundreds of spectral channels with a wealth of spatial and spectral information, the deep residual 3-D CNN network is used to extract spatial–spectral features, as shown in Fig. 2. The image encoder is made up of two 3-D residual Block-MaxPool3d modules and one Conv3d module, where the 3-D residual Block is the core module for extracting robust visual features. The 3-D residual Block is composed of two Conv3d-BN3d-rectified linear units (ReLUs) and one Conv3d in series. The output of the first Conv3d-BN3d-ReLU and the output of Conv3d are used for residual connection. The extracted spatial–spectral features are sent to the classification head and the cross-entropy is calculated with the ground truth. At the same time, it is fed into the projection head to obtain the visual features for feature alignment of visual modes and linguistic modes. The cross-entropy loss is defined as

\[
\mathcal{L}_{ce}(p_i, y_i) = - \sum_{c} y_i^c \log p_i^c
\]

where \( y_i \) is the one-hot encoding of the label information of \( x_i \), \( c \) is the index of class, and \( p_i \) is the predicted probability
output obtained by CLs. Therefore, the classification loss for SD is defined as:

\[
L_{SD}(\mathbf{X}, \mathbf{Y}) = \frac{1}{N} \sum_i L_{ce}(C(x_i), y_i).
\]

(2)

B. Text Encoder

The text encoder is a language-model transformer [33] modified in accordance with Radford et al. [34]. The text encoder starts with a base size of a 33M-parameter, three-layer, 512-wide model with eight attention heads. Similar to CLIP, the text is represented by the transformer using a lower-cased byte pair encoding (BPE) with a vocabulary size of 49,152 [35]. The maximum sequence length is limited to 76 for computational efficiency. This linguistic feature is then normalized by layer and linearly projected into the semantic space.

According to the prior knowledge of land cover classes in SD scene, coarse-grained and fine-grained text descriptions are assigned to each land cover classes. As shown in Fig. 3, A HSI of `<class name>` is used as a template to construct coarse-grained text descriptions for each class in a clone way, and the solid pentagon is shown in Fig. 3. As for fine-grained text, combined with prior knowledge, the color, shape, distribution, and adjacency relationship are described manually, such as The grass stress is pale green, The trees beside road, and The road appears as elongated strip shape. Each land cover class sets two fine-grained text, and the hollow pentagon is shown in Fig. 3.

C. Visual–Linguistic Alignment

The linguistic features obtained by the text encoder construct semantic space as a cross-domain shared space, as shown in Fig. 1. To reduce the distance between visual features and linguistic features by class, supervised contrastive learning [36] is performed, and the domain-invariant representation driven by language is developed to generalize the model to TD. First, a supervised contrastive learning is defined as:

\[
L_{supcon} = -\sum_{i=0}^N \frac{1}{|P(i)|} \sum_{p \in P(i)} \log \frac{\exp(x_i^T \mathbf{v}_p^+ / \tau)}{\sum_{a \in A(i)} \exp(x_i^T \mathbf{v}_a^- / \tau)}
\]

(3)

where for each embedding feature \(x_i\) in minibatch, \(P(i)\) and \(A(i)\) are the positive and negative sample sets, respectively,

\[
\left| P(i) \right| \text{ is the number of positive samples, and } \mathbf{x}_p^+ \text{ and } \mathbf{x}_a^- \text{ are one of the positive and negative samples, respectively.}
\]

Given all the image–text pairs in a minibatch, where image and text are one-to-one and have the same class label. The alignment losses \(L_{supcon}\) of image to text and text to image are calculated

\[
L_{coarse} = -\sum_{i=0}^N \frac{1}{|P(i)|} \left( \sum_{p \in P(i)} \log \frac{\exp(\mathbf{v}_p^T \mathbf{I}_p^+ / \tau)}{\sum_{a \in A(i)} \exp(\mathbf{v}_a^- T \mathbf{I}_p^+ / \tau)} + \sum_{p \in P(i)} \log \frac{\exp(\mathbf{I}_p^T \mathbf{v}_a^+ / \tau)}{\sum_{a \in A(i)} \exp(\mathbf{I}_p^T \mathbf{v}_a^- / \tau)} \right)
\]

(4)

where for each visual feature \(\mathbf{v}_i\) and linguistic feature \(\mathbf{I}_i\) in minibatch, \(|P(i)|\) and \(|A(i)|\) are the positive and negative sample sets of visual feature, respectively. \(P(i)\) and \(A(i)\) are the positive and negative sample sets of linguistic feature, respectively, and \(|P(i)| = |P(i)| = |P(i)|\). In addition, the temperature parameter \(\tau\), which governs the range of the logits in the softmax, is explicitly optimized as a log-parameterized multiplicative scalar during training. \(L_{coarse}\) represents the alignment loss of image to coarse-grained text. The alignment loss of image to fine-grained text \(L_{fine}\) is similar to \(L_{coarse}\),
with the difference being that linguistic features require two fine-grained linguistic features. In the process of optimizing, the visual features and linguistic features belonging to the same class are put into $P_v(i)$ and $P_l(i)$. The image encoder and text encoder are optimized by $L_{\text{coarse}}$ and $L_{\text{fine}}$ to bring features from the same class closer together and samples from different classes farther apart. This enables language knowledge to be integrated into the learning process of visual representation, which helps the image encoder obtain class-specific domain-invariant representations.

In LDGnet, the image encoder and text encoder are jointly trained. Integrating the above loss functions, the total loss of LDGnet is defined as

$$L_{\text{total}} = L_{\text{SD}} + \lambda((1-\alpha)L_{\text{coarse}} + \alpha L_{\text{fine}})$$

where $\lambda$ is a hyperparameter for balancing the alignment loss, and $\alpha$ is used to control the contribution of both the coarse- and fine-grained linguistic features. In the test phase, only image encoder and classification header are used to predict TD.

**IV. EXPERIMENTAL RESULTS AND DISCUSSION**

The experiments using three cross-scene HSI datasets, i.e., the Houston dataset, Pavia dataset, and Gaofen image dataset (GID), are conducted to validate the proposed LDGnet. For comparison algorithms, several state-of-the-art transfer learning algorithms are used, including DA methods, DAAN [15], deep subdomain adaption network (DSAN) [37], multirepresentation adaptation network (MRAN) [38], and heterogeneous transfer CNN (HTCNN) [39], DG methods, PDEN [19], learning to diversify for single domain generalization (LDSDG) [40], and style-agnostic network (SagNet) [41]. The class-specific accuracy (CA), the overall accuracy (OA) and the Kappa coefficient (KC) are used to evaluate the classification performance.

**A. Experimental Data**

1) **Houston Dataset**: The dataset includes Houston 2013 [42] and Houston 2018 [43] scenes, which were obtained by different sensors on the University of Houston campus and its vicinity in different years. The Houston 2013 dataset is composed of 349 × 1905 pixels, including 144 spectral bands, the wavelength range is 380–1050 nm, and the image spatial resolution is 2.5 m. The Houston 2018 dataset has the same wavelength range but contains 48 spectral bands, and the image has a spatial resolution of 1 m. There are seven consistent classes in their scene. We extract 48 spectral bands (wavelength range 0.38–1.05 um) from the Houston 2013 scene corresponding to the Houston 2018 scene and select the overlapping area of 209 × 955. The classes and the number of samples are listed in Table II. In addition, their false-color and ground-truth maps are shown in Fig. 4.

2) **Pavia Dataset**: The Pavia dataset includes University of Pavia (UP) and Pavia Center (PC). The PC has 1096 × 715 pixels and 102 bands. The UP has 103 spectral bands, 610 × 340 pixels, and 1.3-m spatial resolution, where the last band was removed to ensure the same number of spectral bands as PC. They all have the same seven classes, and the name of land cover classes and the number of samples are listed in Table III.

3) **GID Dataset**: GID dataset is constructed by Wuhan University [44], which contains multispectral images (MSIs) taken at different times in many regions of China. We selected GID-nc shot in Nanchang, Jiangxi Province, on January 3, 2015, as the SD and GID-wh shot in Wuhan, Hubei Province, on April 11, 2016, as the TD. GID-nc consists of 900 × 4400 pixels, including blue (0.45–0.52 um), green (0.52–0.59 um), red (0.63–0.69 um), and near-infrared (0.77–0.89 um) bands, and the spatial resolution is 4 m. GID-wh also has the same spatial and spectral resolution, but it is composed of 1600 × 1900 pixels. They have the same five classes, as listed in Table IV. The false-color and ground-truth images are shown in Fig. 5.

![Fig. 4. Pseudocolor image and ground-truth map of the Houston dataset.](image-url)
TABLE IV
NUMBER OF SOURCE AND TARGET SAMPLES FOR THE GID DATASET

| Class   | Number of Samples | GID-nc (Source) | GID-wh (Target) |
|---------|-------------------|----------------|-----------------|
| No.     | Name              | No.            | No.             |
| 1       | Rural residential | 5495           | 4729            |
| 2       | Irrigate land     | 3643           | 3643            |
| 3       | Garden Land       | 6171           | 6116            |
| 4       | River             | 2858           | 11558           |
| 5       | Lake              | 5172           | 2666            |
| Total   |                   | 23339          | 30812           |

Fig. 5. Pseudocolor image and ground-truth map of the GID dataset. (a) Pseudocolor image of GID-nc. (b) Ground-truth map of GID-nc. (c) Pseudocolor image of GID-wh. (d) Ground-truth map of GID-wh.

TABLE V
FINE- GRAINED TEXT DESCRIPTIONS FOR THE HOUSTON DATASET

| Class name       | Fine-grained text                                                                 |
|------------------|----------------------------------------------------------------------------------|
| Grass healthy    | The grass healthy is next to the road, The grass healthy is dark green           |
| Grass stressed   | The grass stressed is next to the road, The grass stressed is pale green         |
| Trees            | The trees are next to residential buildings or non-residential buildings, The trees appear as small circles |
| Water            | The water has a smooth surface, The water appears dark blue or black             |
| Residential buildings | Residential buildings are densely packed, Residential buildings appear as small blocks |
| Non-residential buildings | The shapes of the non-residential buildings are inconsistent, Non-residential buildings appear as large blocks |
| Road             | Trees grew along the road, The road appear as elongated strip shape               |

B. Experimental Setting

LDGnet is implemented on the PyTorch platform. The input is configured with a patch size of $13 \times 13$. The image encoder and the text encoder are optimized via Adaptive Moment Estimation (Adam). The default value for $\ell_2$-norm regularization of all the modules is set to $1e^{-4}$ for weight decay. As the initialization of the text encoder, load the ViT-B-32.pt that was completed by CLIP pretraining. A HSI of `<class name>` is used as a template to generate coarse-grained text descriptions for each class. Furthermore, we artificially set the fine-grained text descriptions of three datasets, as seen in Tables V–VII.

C. Parameter Tuning

A parameter sensitivity analysis is conducted to evaluate the sensitivity of LDGnet on the three TDs. The base learning rate $\eta$, regularization parameter $\lambda$, and weight $\alpha$, regarded as adjustable hyperparameters are selected from $\{1e^{-5}, 1e^{-4}, 1e^{-3}, 1e^-2, 1e^-1\}, \{1e^-3, 1e^-2, 1e^-1, 1e+0, 1e+1\}$, and $\{0.1, 0.3, 0.5, 0.7, 0.9\}$, respectively.

After being modified by the learning rate, the gradient of loss function is used in the gradient descent updates to estimate the model weight hyperparameters. Table VIII provides classification results corresponding to different base learning rates in three datasets. With respect to three datasets, $1e^-2$ is the ideal base learning rate. Tables IX and X include the OA of all the experimental datasets for regularization parameter $\lambda$ and weight $\alpha$, respectively. The optimal $\lambda$ is $1e+0$ for the Houston dataset and GID dataset and $1e^-2$ for the Pavia dataset, and $\alpha$ is $0.3$ for the Houston dataset and the Pavia dataset, and $0.1$ for the GID dataset.

D. Ablation Study

The image encoder and text encoder are the key components of LDGnet, and visual-linguistic alignment is the main
### TABLE IX

**Overall Classification Accuracy (%) with Different Regularization Parameters $\lambda$ for the Proposed LDGnet Using the Three Experimental Data**

| Target scene | Regularization parameter $\lambda$ |
|--------------|-----------------------------------|
|              | 1e-3 | 1e-2 | 1e-1 | 1e+0 | 1e+1 |
| Houston      | 75.64 | 73.82 | 77.49 | 80.25 | 78.44 |
| Pavia        | 81.58 | 82.75 | 82.09 | 79.06 | 81.87 |
| GID          | 73.21 | 70.44 | 73.13 | 78.69 | 73.02 |

### TABLE X

**Overall Classification Accuracy (%) with Different Weights $\alpha$ for the Proposed LDGnet Using the Three Experimental Data**

| Target scene | Weight $\alpha$ |
|--------------|-----------------|
|              | 0.1 | 0.3 | 0.5 | 0.7 | 0.9 |
| Houston      | 79.81 | 80.25 | 78.98 | 78.6 | 77.16 |
| Pavia        | 80.29 | 82.75 | 82.53 | 81.9 | 80.93 |
| GID          | 78.69 | 72.91 | 77.13 | 70.6 | 77.62 |

Fig. 6. Data visualization and classification maps for the target scene Houston 2018 data obtained with different methods including: (a) DAAN (71.13%), (b) MRAN (72.48%), (c) DSAN (78.52%), (d) HTCNN (77.42%), (e) PDEN (75.98%), (f) LSDDG (73.55%), (g) SagNet (73.64%), and (h) LDGnet (80.08%).

strategy for learning domain-invariant representation. Ablation analyses are carried out by eliminating each component from the total framework to evaluate the contribution of important LDGnet components.

There are three variants in the ablation analyses: 1) “LDGnet (cls)”: the text encoder and projection head are deleted from LDGnet, which is to evaluate the performance of the image encoder; 2) “LDGnet (coarse)”: the fine-grained text is deleted; and 3) “LDGnet (fine)”: the coarse-grained text is deleted. LDGnet (coarse) and LDGnet (fine) are the performance analysis of the text encoder and visual–linguistic alignment from different language levels. The text encoder and visual–linguistic alignment cannot be omitted and need to be used for LDGnet at the same time. It is clear from Table XI that the proposed LDGnet performs better than the existing variations and makes significant advancements. On the basis of LDGnet (cls), add transformer as the text encoder architecture, paired with coarse-grained text descriptions LDGnet (coarse) or fine-grained text descriptions LDGnet (fine), and the performance improves by around 6%. This suggests that linguistic modality aids in the learning of visual representations, and both coarse- and fine-grained visual–linguistic alignment can help improve the generalization ability of the model on TD. Furthermore, LDGnet performs 1%–4% better than LDGnet (coarse) and LDGnet (fine). The use of coarse-grained or fine-grained linguistic features alone is inferior to the fusion of them, so as to enrich language supervised signals and complete visual–linguistic alignment at two levels.

### E. Performance on Cross-Scene HSI Classification

1) **Comparison Method Setting:** Relevant algorithms such as DAAN, MRAN, DSAN, HTCNN, PDEN, LSDDG, and SagNet are used for comparison to assess the performance of LDGnet with only SD used for training. The training sample is set as follows. All the SD data with labels (80% for training and 20% for validation) and all the TD data without labels are used for training with DAAN, MRAN, DSAN, and HTCNN, which are considered DA techniques. Only SD with labels (80% for training and 20% for validation) are used as training examples for DG techniques, PDEN, LSDDG, and SagNet, and the patch size of LSDDG and SagNet is adjusted to $32 \times 32$ to accommodate the input size of Resnet18. In contrast to the other two datasets, the SD in the Houston dataset is also increased by four times through random flip and random radiation noise (illumination). The optimal base learning rate and regularization parameters of all the comparison algorithms are selected from $\{10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}\}$ and $\{10^{-3}, 10^{-2}, 10^{-1}, 0, 10^{0}, 10^{1}, 10^{2}\}$, respectively, and cross-validation is used to find the corresponding optimal parameters.

2) **Classification Performance Analysis:** The following analyses are obtained from Tables XII–XIV.

1) On all TDs, DSAN exhibits the best performance for the DA technique. In the comparison of DG approaches, PDEN performs well on the data from Houston 2018 and the PC, while LSDDG works well on the data from GID-wh. In particular, DSAN provides 2% improvement in OA over PDEN on the Houston 2018 data, while PDEN and LSDDG are 2% higher than DSAN on PC and GID-wh, respectively. This demonstrates that TD is not always used in the training process to produce the
greatest classification performance and that both the DA method and the DG approach have their own advantages in various scenes.

LDGnet outperforms DSN on all TDs by 2%–6%. DSN expressly uses the domain alignment technique during the training process to minimize the domain shift and directly accesses TD. In contrast, self-supervised contrastive learning is used in LDGnet to reduce the gap between visual and linguistic features by class while learning domain-invariant representation in cross-domain shared semantic space. The enhanced LDGnet classification performance demonstrates the superiority of this explicit shared space learning technique over the domain alignment strategy used in DA.

2) LDGnet outperforms the DG algorithms, PDEN and LDSDG, by 2%–6% on OA. Different from the data generation strategies in PDEN and LDSDG, coarse-grained and fine-grained text representations are introduced in LDGnet to learn more robust domain-invariant representations from multimodal features through visual–linguistic alignment.

3) Visual Classification Map: The classification maps are illustrated in Figs. 6 and 7. In Fig. 6, labeled pixels are displayed as ground truth and unlabeled pixels as backgrounds, and all the pixels are predicted for comparison in Fig. 7. In contrast, the proposed LDGnet obtains less noisy and more accurate results in some areas of the classification maps, such as third class (Brick) and seventh bare soil in the PC data are greatly improved compared with all the comparison methods. It is obvious from Fig. 7 that the third (Garden land) and fifth (Lake) in GID-wh data are better predicted.

4) Time Complexity Analysis: The one epoch training time on all the experimental data is presented in Table XV to illustrate the computational difficulty of various approaches. All the experiments are carried out using PyTorch on a advanced micro devices (AMD) EPYC 7542 32-Core Processor [48-GB random access memory (RAM)] powered with Nvidia GTX 3090 graphics processing unit (GPU) with 24-GB memory. The computational cost of LDGnet is not low compared with other methods, and the three-layer transformer used as the text encoder is the primary source of calculation expense. Its training parameters 33.43 M accounted for 97.7%
Fig. 7. Data visualization and classification maps for the target scene GID-wh data obtained with different methods including: (a) DAAN (68.06%), (b) MRAN (67.49%), (c) DSAN (73.69%), (d) HTCNN (57.20%), (e) PDEN (67.71%), (f) LDSDG (76.48%), (g) SagNet (61.64%), and (h) LDGnet (77.08%).

Fig. 8. Class separability of the proposed LDGnet using three datasets, where (a) and (b) are the Houston 2018 data, and (c) and (d) are the PC data. Best looking in a→b and c→d.

TABLE XV
EXECUTION TIME (IN SECONDS) OF ONE EPOCH TRAINING IN DIFFERENT METHODS

| Methods     | DAAN [18] | MRAN [38] | DSAN [37] | HTCNN [39] | PDEN [19] | LDSDG [40] | SagNet [41] | LDGnet |
|-------------|-----------|-----------|-----------|-------------|-----------|------------|-------------|--------|
| Houston 2018| 15.68     | 13.52     | 16.64     | 31.21       | 4.88      | 43.56      | 7.03        | 17.84  |
| Pavia Center| 28.32     | 29.01     | 31.76     | 46.24       | 14.21     | 63.38      | 20.44       | 68.21  |
| GID-wh      | 20.35     | 22.32     | 24.43     | 39.65       | 3.15      | 20.78      | 16.78       | 19.53  |

of the total number of parameters. The overall computational complexity of LDGnet is lower than that of LDSDG, where the Style-Complement module is designed to consider multiple potential style changes.

F. Generalization Performance of LDGnet

Language is used in LDGnet as an additional supervised signal. Visual and linguistic features are aligned in a cross-domain shared semantic space, and classwise domain-invariant representations are learned from multimodal features. Two TDs (Houston 2018 data and PC data) are inferred using fully trained LDGnet, and the visualization of class separability in the original space and semantic space is shown in Fig. 8. The interclass distribution is messed up in Fig. 8(a) and (c), and after features are embedded in the semantic space, the separability is greatly increased. For instance, in the first class of (b) and third class of (d), the aggregate of features belonging to the same class is greatly improved.

G. Fine-Grained Text Analysis

The fine-grained text descriptions provide prior knowledge in the proposed LDGnet. Tables V–VII show the fine-grained text artificially defined from the color, shape, distribution, and adjacency relationship. The following experiments are carried
TABLE XVI
IRRELEVANT DESCRIPTIONS AND RELEVANT DESCRIPTIONS FOR TREES

| Class name       | Fine-grained text                                                                 |
|------------------|----------------------------------------------------------------------------------|
| Trees (Irrelevant description) | The rigid tell can not crack the garden                                        |
|                  | It was then the formal bus met the imaginary difficulty                         |
|                  | The heartbreaking profession excites into the abash guarantee                    |
|                  | The cultural commission sneezes into the federal solution                        |
| Trees (Relevant description)   | In botany, a tree is a perennial plant with an elongated stem, or trunk, usually supporting branches and leaves |
|                  | Tree is a woody plant that regularly renews its growth                           |
|                  | Tree has many secondary branches supported clear of the ground on a single main stem or trunk with clear apical dominance |
|                  | Trees appear mostly at the surface and often in clusters with similar height      |

TABLE XVII
OVERALL CLASSIFICATION ACCURACY (%) FOR IRRELEVANT DESCRIPTIONS AND INTERNET DESCRIPTIONS USING THE HOUSTON 2018 DATA

| Irrelevant description | 1 time | 2 time | 3 time | 4 time | 5 time | 6 time | 7 time | 8 time | 9 time | 10 time | mean |
|------------------------|--------|--------|--------|--------|--------|--------|--------|--------|--------|---------|------|
| OA (%)                 | 77.96  | 77.20  | 76.91  | 78.08  | 77.04  | 76.82  | 76.02  | 77.86  | 75.99  | 76.46   | 77.03 |
| KC (x)                 | 59.15  | 57.85  | 57.22  | 59.64  | 57.50  | 56.73  | 55.91  | 58.87  | 55.7   | 56.12   | 57.47 |
| Relevant description   |        |        |        |        |        |        |        |        |        |         |      |
| OA (%)                 | 79.67  | 78.37  | 78.21  | 78.47  | 78.74  | 78.18  | 81.84  | 79.55  | 80.10  | 79.73   | 79.25 |
| KC (x)                 | 65.33  | 60.05  | 60.08  | 61.31  | 63.87  | 59.94  | 68.84  | 64.11  | 65.32  | 64.84   | 65.35 |

out with the Houston dataset as an example to analyze the sensitivity of LDGnet to how fine-grained text is defined.

1) Irrelevant description: four illogical sentences are randomly generated for each land cover class, and these sentences have nothing to do with the attributes of land cover class.

2) Relevant description: the name of each land cover class is taken as the entry, and four sentences corresponding to the description are selected from Google. Table XVI shows the irrelevant and Internet descriptions for the Trees used in the experiment. These sentences are closely related to the land cover class. In the experiment, two of the four fine-grained text descriptions of each class are randomly selected (two identical descriptions may be selected for a class).

Both the experiments are run ten times, and the classification accuracy is shown in Table XVII. When irrelevant descriptions are used as fine-grained text, it is decreased by 3% lower compared with 80.25% using artificially defined descriptions, and by 1% when relevant Internet descriptions are used. It is clear that relevant fine-grained text descriptions, whether derived from human definitions or Internet searches, perform better than unrelated ones, indicating that a correct description of the land cover class in the linguistic modality is necessary. In addition, LDGnet has considerable performance when using relevant fine-grained text from Internet, which indicates that the proposed method is compatible with multiple descriptions of a land cover class, not one defined in Tables V–VII.

V. CONCLUSION

The LDGnet, which combines the visual and language modes, is proposed. The visual and linguistic features are extracted concurrently using the dual-stream architecture. The text representation with prior knowledge is regarded as cross-domain shared knowledge, which is designed to guide the learning of cross-domain-invariant representation. Specifically, the coarse-grained and fine-grained text representations are designed, using the CNN-based image encoder and transformer-based text encoder to extract visual features, coarse-grained and fine-grained linguistic features, respectively. The semantic space constituted of linguistic features is treated as a cross-domain shared space, and supervised contrastive learning is used to gradually decrease the gap between visual and linguistic features to achieve visual–linguistic alignment. Finally, for TD scene generalization, the image encoder embedded with common language knowledge is applied. Comprehensive experiments on three datasets validate the effectiveness of the proposed LDGnet in domain generalization.

REFERENCES
[1] H. Yuan and Y. Y. Tang, “Spectral–spatial shared linear regression for hyperspectral image classification,” IEEE Trans. Cybern., vol. 47, no. 4, pp. 934–945, Apr. 2017.
[2] M. Zhang, W. Li, Q. Du, L. Gao, and B. Zhang, “Feature extraction for classification of hyperspectral and LiDAR data using patch-to-patch CNN,” IEEE Trans. Cybern., vol. 50, no. 1, pp. 100–111, Jan. 2020.
[3] M. Zhang, W. Li, Y. Zhang, R. Tao, and Q. Du, “Hyperspectral and LiDAR data classification based on structural optimization transmission,” IEEE Trans. Cybern., early access, May 13, 2022, doi: 10.1109/TCYB.2022.3169773.
[4] Z. Xue, M. Zhang, Y. Liu, and P. Du, “Attention-based second-order pooling network for hyperspectral image classification,” IEEE Trans. Geosci. Remote Sens., vol. 59, no. 11, pp. 9600–9615, Nov. 2021.
[5] Z. Xue, Q. Xu, and M. Zhang, “Local transformer with spatial partition restore for hyperspectral image classification,” IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens., vol. 15, pp. 4307–4325, 2022.
[6] Z. Xue, Y. Zhou, and P. Du, “S3Net: Spectral–spatial Siamese network for few-shot hyperspectral image classification,” IEEE Trans. Geosci. Remote Sens., vol. 60, pp. 1–19, 2022, Art. no. 5531219.
[7] S. Zhou, H. Wu, and Z. Xue, “Grouped subspace linear semantic alignment for hyperspectral image transfer learning,” IEEE Trans. Geosci. Remote Sens., vol. 60, no. 2, pp. 4412116.
[8] Y. Zhang, W. Li, M. Zhang, Y. Qu, R. Tao, and H. Qi, “Topological structure and semantic information transfer network for cross-scene hyperspectral image classification,” IEEE Trans. Neural Netw. Learn. Syst., early access, Sep. 16, 2021, doi: 10.1109/TNNLS.2021.3109872.
[11] Y. Zhang, W. Li, R. Tao, J. Peng, Q. Du, and Z. Cai, “Cross-scene hyperspectral image classification with discriminative cooperative aggregation,” IEEE Trans. Geosci. Remote Sens., vol. 59, no. 11, pp. 9646–9660, Nov. 2021.

[12] Y. Zhang, W. Li, M. Zhang, S. Wang, R. Tao, and Q. Du, “Graph information aggregation cross-domain few-shot learning for hyperspectral image classification,” IEEE Trans. Neural Netw. Learn. Syst., early access Jun. 2022. doi: 10.1109/TNNLS.2022.3185795.

[13] Y. Zhang, W. Li, W. Sun, R. Tao, and Q. Du, “Single-source domain expansion network for cross-scene hyperspectral image classification,” 2022, arXiv:2209.01634.

[14] B. Schölkopf, J. Platt, and T. Hofmann, “A kernel method for the two-sample-problem,” in Proc. Adv. Neural Inf. Process. Syst., vol. 19, 2007.

[15] C. Yu, J. Wang, Y. Chen, and M. Huang, “Transfer learning with dynamic adversarial adaptation network,” in Proc. IEEE Int. Conf. Data Mining (ICDM), Nov. 2019, pp. 778–786.

[16] Z. Liu, L. Ma, and Q. Du, “Class-wise distribution adaptation for unsupervised classification of hyperspectral remote sensing images,” IEEE Trans. Geosci. Remote Sens., vol. 59, no. 1, pp. 508–521, Jan. 2021.

[17] J. Wang et al., “Generalizing to unseen domains: A survey on domain generalization,” 2021, arXiv:2103.03097.

[18] K. Zhou, Y. Yang, T. Hospedales, and T. Xiang, “Deep domain-adversarial image generation for domain generalisation,” in Proc. AAAI Conf. Artif. Intell., 2020, vol. 34, no. 7, pp. 13025–13032.

[19] L. Li et al., “Progressive domain expansion network for single domain generalization,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2021, pp. 224–233.

[20] U. Zia, M. M. Riaz, and A. Ghafoor, “Transforming remote sensing images to textual descriptions,” Int. J. Appl. Earth Obs. Geoinf., vol. 108, Apr. 2022, Art. no. 102741.

[21] S. Lobry, D. Marcos, J. Murray, and D. Tuia, “RSVQA: Visual question answering for remote sensing data,” IEEE Trans. Geosci. Remote Sens., vol. 58, no. 12, pp. 8555–8566, Dec. 2020.

[22] Y. Li et al., “Recurrent attention and semantic gate for remote sensing image captioning,” IEEE Trans. Geosci. Remote Sens., vol. 60, pp. 1–16, 2022, Art. no. 5608816.

[23] A. Radford et al., “Learning transferable visual models from natural language supervision,” in Proc. Int. Conf. Mach. Learn., 2021, pp. 8748–8763.

[24] J. Yu, Z. Wang, V. Vasudevan, L. Yeung, M. Seyedhosseini, and Y. Wu, “CoCa: Contrastive captors are image-text foundation models,” 2022, arXiv:2205.01917.

[25] J. Wang, Y. Chen, W. Feng, H. Yu, M. Huang, and Q. Yang, “Transfer learning with dynamic distribution adaptation,” ACM Trans. Intell. Syst. Technol., vol. 11, no. 1, pp. 1–25, Feb. 2020.

[26] B. Sun and K. Saenko, “Deep coral: Correlation alignment for deep domain adaptation,” in Proc. Eur. Conf. Comput. Vis., Amsterdam, The Netherlands: Springer, 2016, pp. 443–450.

[27] F. Zhou, Z. Jiang, C. Shui, B. Wang, and B. Chaib-draa, “Domain generalization via optimal transport with metric similarity learning,” 2020, arXiv:2007.10573.

[28] D. P. Kingma and M. Welling, “Auto-encoding variational Bayes,” 2013, arXiv:1312.6114.

[29] I. Goodfellow et al., “Generative adversarial nets,” in Proc. Adv. Neural Inf. Process. Syst., vol. 27, 2014, pp. 1–9.

[30] Y. Gao et al., “PyramidCLIP: Hierarchical feature alignment for vision-language model pretraining,” 2022, arXiv:2204.14095.

[31] L. H. Li, M. Yatskar, D. Yin, C.-J. Hsieh, and K.-W. Chang, “VisualBERT: A simple and performant baseline for vision and language,” 2019, arXiv:1908.03557.

[32] W. Li et al., “UNIMO: Towards unified-modal understanding and generation via cross-modal contrastive learning,” in Proc. 59th Annu. Meeting Assoc. Comput. Linguistics 11th Int. Joint Conf. Natural Lang. Process., vol. 1, 2021, pp. 2592–2607.

[33] A. Vaswani et al., “Attention is all you need,” in Proc. Adv. Neural Inf. Process. Syst., vol. 30, 2017, pp. 1–11.

[34] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, “Language models are unsupervised multitask learners,” OpenAI Blog, vol. 1, no. 8, p. 9, 2019.

[35] R. Sennrich, B. Haddow, and A. Birch, “Neural machine translation of rare words with subword units,” in Proc. 54th Annu. Meeting Assoc. Comput. Linguistics, Stroudsburg, PA, USA: Association for Computational Linguistics, 2016, pp. 1715–1725.

[36] P. Khosla et al., “Supervised contrastive learning,” in Proc. 34th Int. Conf. Neural Inf. Process. Syst., 2020, pp. 18661–18673.

[37] Y. Zhu et al., “Deep subdomain adaptation network for image classification,” IEEE Trans. Neural Netw. Learn. Syst., vol. 32, no. 4, pp. 1713–1722, Apr. 2021.

[38] Y. Zhu et al., “Multi-representation adaptation network for cross-domain image classification,” Newsl., vol. 119, pp. 214–221, Nov. 2019.

[39] X. He, Y. Chen, and P. Ghamisi, “Heterogeneous transfer learning for hyperspectral image classification based on convolutional neural network,” IEEE Trans. Geosci. Remote Sens., vol. 58, no. 5, pp. 3246–3263, May 2019.

[40] Z. Wang, Y. Luo, R. Qiu, Z. Huang, and M. Baktashmotlagh, “Learning to diversify for single domain generalization,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2021, pp. 834–843.

[41] H. Nam, H. Lee, J. Park, W. Yoon, and D. Yoo, “Reducing domain gap by reducing style bias,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2021, pp. 8690–8699.

[42] C. Debes et al., “Hyperspectral and LiDAR data fusion: Outcome of the 2013 GRSS data fusion contest,” IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens., vol. 7, no. 6, pp. 2405–2418, Jun. 2014.

[43] B. L. Saux, N. Yokoya, R. Hansch, and S. Prasad, “2018 IEEE GRSS data fusion contest: Multimodal land use classification [technical committee],” IEEE Geosci. Remote Sens. Mag., vol. 6, no. 1, pp. 52–54, Mar. 2018.

[44] X.-Y. Tong et al., “Land-cover classification with high-resolution remote sensing images using transferable deep models,” Remote Sens. Environ., vol. 237, Feb. 2020, Art. no. 111322.

Yuxiang Zhang (Student Member, IEEE) received the B.S. degree in electronics and information engineering and the M.S. degree in control science and engineering from the Beijing University of Chemical Technology, Beijing, China, in 2017 and 2020, respectively. He is currently pursuing the Ph.D. degree with the Beijing Institute of Technology, Beijing.

His research interests include hyperspectral imagery processing and pattern recognition. His supervisor is Dr. Wei Li.

Mengmeng Zhang received the B.S. degree in computer science and technology from the Qingdao University of Science and Technology, Qingdao, China, in 2014, and the Ph.D. degree in control science and engineering from the Beijing University of Chemical Technology, Beijing, China, in 2019.

She is currently an Assistant Professor with the School of Information and Electronics, Beijing Institute of Technology, Beijing. Her research interests include remote sensing image processing and pattern recognition.

Wei Li (Senior Member, IEEE) received the B.E. degree in telecommunications engineering from Xidian University, Xi’an, China, in 2007, and the M.S. degree in information science and technology from Sun Yat-sen University, Guangzhou, China, in 2009, and the Ph.D. degree in electrical and computer engineering from Mississippi State University, Starkville, MS, USA, in 2012.

Subsequently, he spent one year as a Post-Doctoral Researcher with the University of California, Davis, CA, USA. He is currently with the School of Information and Electronics, Beijing Institute of Technology, Beijing. His research interests include hyperspectral image analysis, pattern recognition, and data compression.

Dr. Li is an Active Reviewer for the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING (TGRS), the IEEE GEOFISCE REMOTE SENSING LETTERS (GRSL), and the IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING (JSTARS). He is currently serving as an Associate Editor for both the IEEE SIGNAL PROCESSING LETTERS and the IEEE JSTARS. He has served as the Guest Editor for special issue of IEEE CEMUGE PROCESSING, REMOTE SENSING, and IEEE JSTARS. He received the 2015 Best Reviewer Award from the IEEE Geoscience and Remote Sensing Society (GRSS) for his service for IEEE JSTARS.
Shuai Wang received the bachelor’s and master’s degrees in chemical engineering and technology from the Beijing University of Chemical Technology, Beijing, China, in 2013 and 2020, respectively. He is currently pursuing the Ph.D. degree with the University of Hong Kong, Hong Kong.

His research interest is about machine learning (ML) applications into different fields, such as computational chemistry, materials science, and images.

Ran Tao (Senior Member, IEEE) received the B.S. degree from the Electronic Engineering Institute of PLA, Hefei, China, in 1985, and the M.S. and Ph.D. degrees from the Harbin Institute of Technology, Harbin, China, in 1990 and 1993, respectively.

He has been a Senior Visiting Scholar with the University of Michigan, Ann Arbor, MI, USA, and the University of Delaware, Newark, DE, USA, in 2001 and 2016, respectively. He is currently a Professor with the School of Information and Electronics, Beijing Institute of Technology, Beijing, China. His research interests include fractional Fourier transform and its applications, theory, and technology for radar and communication systems.

Dr. Tao is a Fellow of the Institute of Engineering and Technology and the Chinese Institute of Electronics. He was a recipient of the National Science Foundation of China for Distinguished Young Scholars in 2006, and a Distinguished Professor of Changjiang Scholars Program in 2009. He has been a Chief-Professor of the Creative Research Groups with the National Natural Science Foundation of China since 2014, and he was a Chief-Professor of the Program for Changjiang Scholars and Innovative Research Team in University from 2010 to 2012. He is currently the Vice-Chair of the IEEE China Council. He is also the Vice-Chair of the International Union of Radio Science (URSI) China Council and a member of Wireless Communication and Signal Processing Commission, URSI. He was a recipient of the First Prize of Science and Technology Progress in 2006 and 2007, and the First Prize of Natural Science in 2013, both awarded by the Ministry of Education.