Article

LIVEnet: Linguistic-Interact-with-Visual Engager Domain Generalization for Cross-Scene Hyperspectral Imagery Classification

Yuanyuan Dang , Xianhe Zhang * , Bing Liu

Posted Date: 8 February 2024

doi: 10.20944/preprints202402.0491.v1

Keywords: Hyperspectral image (HSI) classification; Contrastive learning; CLIP; Domain generalization; Linguistic-Visual alignment

Article

LIVEnet: Linguistic-Interact-with-Visual Engager Domain Generalization for Cross-Scene Hyperspectral Imagery Classification

Yuanyuan Dang [1], Xianhe Zhang [2,*] and Bing Liu [3]

[1] College of Computer Science and Engineering, Changchun University of Technology, Changchun, 130000, China

[*] Correspondence: liubing@ccut.edu.cn

Abstract: Domain generalization has led to remarkable achievements in Hyperspectral Image (HSI) classification. Inspired by contrastive language-image pre-training (CLIP), the language-aware domain generalization method has been explored to learn cross-domain-invariant representation. However, existing methods face some challenges: 1) The weak capacity to extract long-range contextual information and inter-class correlation. 2) Due to the inadequacies of the large-scale pre-training for HSI data, the spatial-spectral features of HSI and linguistic features can not be straightforwardly alignment. To address the above problems, a novel network has been proposed with a CLIP framework, which consists of an image encoder, based on an encoder-only transformer to obtain the global contextual information and inter-class correlation, a frozen text encoder, and a cross-attention mechanism, named Linguistic-Interact-with-Visual Engager (LIVE), enhances the interaction between two modalities. Extensive experiments demonstrating superior performance over state-of-the-art methods in HSI Domain Generalization with a CLIP framework.

Keywords: hyperspectral image (HSI) classification; contrastive learning; CLIP; domain generalization; Linguistic-Visual alignment

1. Introduction

With the development of Deep Learning (DL), many supervised methods for HSI classification have achieved extraordinary performance[1][2][3][4][5]. However, the traditional supervised methods need manual annotation, which is time-consuming and professional-knowledge-supported, and have a weak capacity to attain cross-domain-invariant representation. It causes a terrible generalization gap between the Source Domain (SD) to the Target Domain (TD). In other words, when encountering cross-scene classification tasks, such as data re-collection by advanced devices, HSIs are affected by sensor nonlinearities, seasonal variations, and weather conditions. These factors result in spectral reflectance variations between SD and TD of the same land cover classes. Despite the same land space and region, the well-pretrained model also fails to achieve exact performance. The naive supervised architecture is presented in Figure 1.

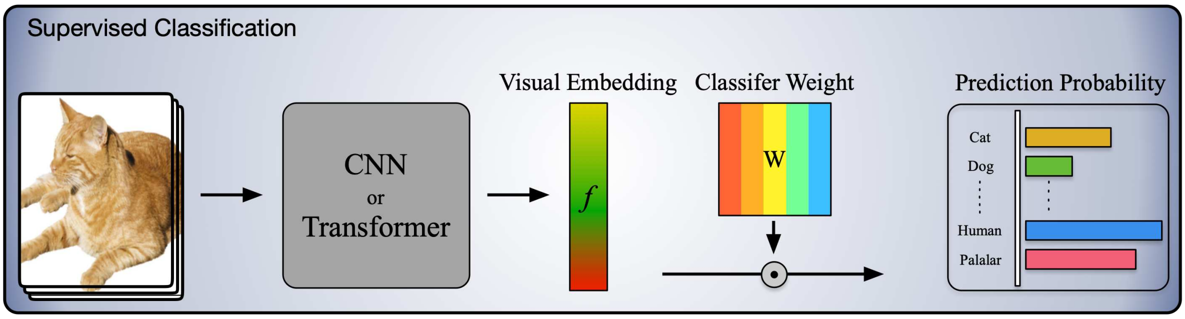

Figure 1. The pipeline of the traditional supervised method for image classification.

Domain Generalization (DG) has recently attracted significant attention as a more challenging task setting to tackle this problem. Distincting from Domain Adaption (DA), DG requests to produce generalization models and without any TD samples engaging in training. Currently, most DG works focus on learning visual-level domain-invariant representation from multiple SDs or a single SD [6][7][8].

With the rise of the Contrastive Language-Image Pre-training (CLIP) [9] model, which is trained on vast (image, text) pairs and achieves a remarkable generalization and zore-shot learning capability. It has become increasingly appealing to adapt a CLIP framework with linguistic information to DG tasks for natural RGB image classification. A question arises: Can we adapt the CLIP domain generalization for cross-scene hyperspectral imagery classification, directly? A markable work [10] proposed a straightforward multimodal DG framework for HSI, called LDGnet, utilizing the language knowledge to learn visual representation and achieve visual-linguistic alignment. However, due to the substantial disparity between the natural RGB image and the HSI data, the SOTA work, LDGnet, still achieves an inferior performance compared with some traditional DG classification models. This seems a depressing assessment, and the question is why we need to employ the CLIP. There are three potential advantages: Firstly, traditional DG models tackle the domain alignment with domain extension and data augment methods based on variational auto-encoder (VAE) [11] and generative adversarial networks (GANs) [12], which special design for single or multiple modalities. As opposed to the traditional methods, fine-tuning is a key step in CLIP, which is demonstrated by [13][14], for the diversity of downstream tasks. Once the CLIP's fine-tuning provides an effective for one modality, it can be easily extended to other modalities, such as multispectral and visible light data. This is the exciting generalization capacity distinguished from the traditional method. Secondly, the language knowledge auxils have been demonstrated to be helpful for multi-modal learning by [15][16][17]. Thirdly, The training time of the CLIP framework has an advantage over traditional discriminative generative (DG) classification tasks due to the preservation of a large portion of pre-trained weights. Based on the above analysis, we have reconsidered the LDGnet. The LDGnet's architecture is specially designed and can be split into three sections: An Image encoder, a frozen text encoder, and a visual–linguistic alignment strategy. Image encoder extracts the visual features from the image patch and splits it into two branches: one is fed into the classification head for the design of the auxiliary loss and another is fed into the projection layer for the visual–linguistic alignment, where the image encoder is designed by the deep residual 3-D CNN network. the text encoder extracts the linguistic features. Finally, the visual–linguistic alignment strategy can make the semantic space treated as a cross-domain shared space. Fine-grained feature recognition, however, benefits from the contextual information in the background, where this information encapsulates the correlation between each pixel and its neighboring pixels. Yet, the image encoder proposed by LDGnet can not effectively extract long-range contextual information. In addition, this image encoder has a weak ability to extract inter-class correlation, where the inter-class correlation plays a significant role in enhancing across-domain category transferability and classification performance, especially under a sample imbalance situation. Distinct from traditional RGB images, which represent the category information by visual characters, the prior knowledge of land cover classes is mainly reflected by the abundant spectrum of the hyperspectral image so that spatial-spectral features extracted by the image encoder are necessary for classification performance. However, the CLIP has already achieved well-aligned knowledge between image features and linguistic features. In contrast, the joint spatial-spectral features provide additional resistance for feature alignment, due to the large-scale pre-trained CLIP specifically designed for HIS characteristics not yet perfect. In summary, the deficiency exists in the following two stages: 1) In the image feature extraction stage, the image encoder can not attain the inter-class correlation information and global features which harms the category transferability and classification performance, respectively. 2) In the visual-linguistic alignment stage, the spatial-spectral features and linguistic features can not be simply and directly aligned without the already well-aligned knowledge from pre-trained CLIP on the HSI data.

In this paper, for the former, we propose a novel image encoder, which equips a simple encoder-only transformer [18] that can effectively extract the global information and inter-class correlation, instead of the deep residual 3-D CNN network, which tends to fall short in capturing sufficient class correlation information and global spatial-spectral features. For the latter, the key is finding the approach to match the linguistic features and visual features with better alignment, where the visual features contain extra spectral information than traditional images. The processing of alignment in LDGnet has been reconsidered, which made the spatial image features and linguistic features entirely isolated during encoding and there is no bridge for inter-model information flow before the final matching, where inter-modality information means the interactions between visual features and linguistic features. It will work because generic visual and text representations have been obtained by CLIP, which is pre-trained by large-scale image-text pairs, so they can directly align via cosine similarities calculation, but this convenience can not be promoted to alignment between spatial-spectral features and linguistic features. According to this problem, A mechanism of the interaction between two modalities, which is composed of a special cross-attention, has been proposed. With this mechanism, the category information, which is derived from the spatial-spectral features, can make the text semantics become visual-aware and image-conditional, instead of remaining the same for the entire dataset. Correspondingly, the linguistic features make the more distinctive spatial-spectral features, which from more informative spatial regions and more instrumental bands of spectrums, be focused on. In addition, this cross-attention mechanism provides a moderate parameter that can play a role as an adapter to achieve efficient freshly learned knowledge learning via fine-tuning.

In particular, both two improved modules, the image encoder and cross-attention mechanism, have been designed as simply and effectively as possible to avoid conducting extra learnable parameters in supposed to tackle over-fitting caused by insufficient training data of the HSI dataset. The construction of the image encoder can be simply described as follows: a 3-D convolution layer, a specially designed DW convolution [19] layer, and a transformer encoder with single-head attention mechanism, which omits cls token. In addition, the cross-attention module for the Inter-Modality, which is abbreviated as Linguistic-Interact-with-Visual Engager (LIVE) for full text, only is implemented by several linear layers to generate queries, keys, and values in both the visual branch and the text branch. Extensive experiments demonstrate the superiority of our method.

In summary, our main contributions are as follows:

- We design a special image encoder in the visual branch of the CLIP for HSI data characteristics, where this encoder can extract the spatial-spectral features with global-ranged and category correlation.
- The LIVE module was proposed to improve the interaction between two modalities before the final matching. With this module, visual features will become more distinctive, and correspondingly linguistic features will be visual-features-based adaptive, instead of being invariable.
- Extensive experiments demonstrate the proposed method has a better performance compared with the *SOTA* work with a CLIP framework for HSI domain Generalization.

2. Materials and methods

2.1. Related Work

2.1.1. Domain Generalization (DG)

DG presents a more formidable challenge than DA, as it seeks to learn the model exclusively through SD data during the training phase without requiring access to TD. The model's extension to TD occurs in the zero-shot inference stage. Current DG methodologies are broadly classified into two categories: multi-source DG and single-source DG. Various methods explicitly address domain shift among multiple SD domain representations by minimizing differences in feature distributions, employing techniques such as MMD [20], second-order correlation [21], and Wasserstein distance [22]. For the single-source DG, the prevalent approach involves extensive data augmentation. These methods typically amplify or generate out-of-domain samples related to SD, subsequently incorporating these samples into model training to facilitate the transition from SD to TD. The overarching objective of data generation is to produce diverse and plentiful data to enhance generalization. Notably, techniques like variational auto-encoder (VAE) [11] and generative adversarial networks (GANs) [12] are frequently harnessed for this purpose.

2.1.2. CLIP

CLIP (Contrastive Language-Image Pre-training), introduced by [9], stands as a groundbreaking advancement in vision-language learning, demonstrating considerable potential in acquiring generic visual representations through contrastive pre-training. Leveraging its impressive transferability, the effective adaptation of CLIP to downstream tasks has been extensively explored, such as Context Optimization (CoOp), proposed by [23], introduces learnable prompts for textual inputs as opposed to handcrafted ones, drawing inspiration from prompt learning [24]. Adopting a strategy reminiscent of adapters [25], CLIP-Adapter [14] integrates a lightweight adapter module to generate adapted multi-modal features. Meanwhile, Tip-Adapter [26] substantially reduces training costs by implementing a key-value cache model. The pipeline of naive CLIP is shown in Figure 2.

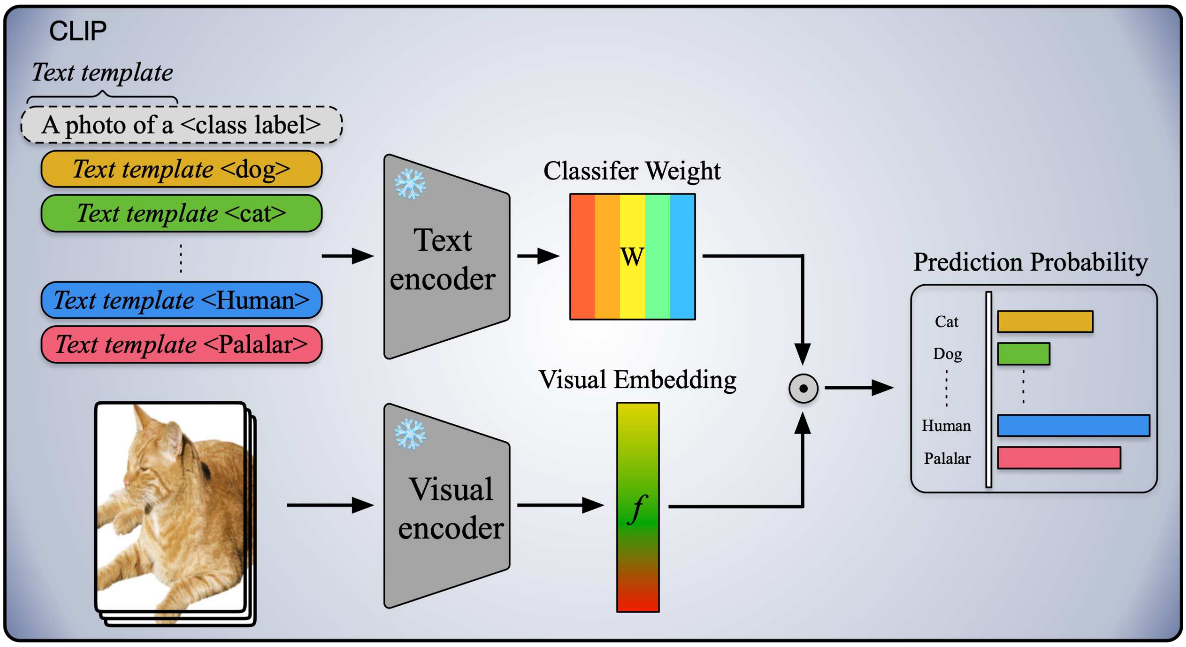

Figure 2. The pipeline of the CLIP for image classification.

2.1.3. Prompt Engineering

Inspired by the success of GPT-3, CLIP embarks on training a substantial contrastive learning model using a vast dataset comprising over 400 million image-text pairs. Notably, CLIP showcases the capability for prompt-based zero-shot visual classification. Serving as the foundation, subsequent advancements such as CoOp (Context Optimization) by [27]. and CPT (Continuous Prompt Tuning) by [28] extend these capabilities. These studies demonstrate that optimizing continuous prompts can significantly outperform manually designed discrete prompts in various vision tasks. In this paper, an HSI of {class name} is used as a template to construct coarse-grained text descriptions in a cloze way, which is proposed by [10]. Analogously, fine-grained text is described manually combined the prior knowledge.

2.2. Proposed Method

In this section, we first introduce the overall network, and then we present the details of the proposed model, including the image encoder and LIVE module. finally, we provide some variants of the proposed model.

2.2.1. Overall Network

Assume that $S = \{s_i\}_{i=1}^{N} \in \mathbb{R}^d$ and $Y = \{y_i\}_{i=1}^{N}$ is the data and labels from SD, respectively, where N denotes the number of source samples and d denote the channels of HSI data cube. Correspondingly, $T = t_{i_{i=1}}^{N} \in \mathbb{R}^d$ is reprensed as the data from TD. We do the same prompt engineering as [10] introduced. The text is made up of one coarse-grained prompt and two fine-grained prompts corresponding to each sample. The $VisEnc(\cdot)$ and $TexEnc(\cdot)$ denote the image (visual) encoder and text encoder, respectively. The HSI patches and text are respectively fed into the $VisEnc(\cdot)$ and $TexEnc(\cdot)$, to extract the spatial-spectral features (visual features F_v) and linguistic features (text features F_t). $TexEnc(\cdot)$ is a language-model transformer[33][34] (in LDGnet) and details of the $VisEnc(\cdot)$ will be elaborated in the subsection. Then the spatial-spectral features are delivered to the projection head (Proj) and the classification head (Cls), simulately. The Proj can project the visual features for interacting with linguistic features and visual-linguistic alignment. The Cls is used to implement assistant loss by calculating the cross-entropy. The pipeline of our proposed architecture is shown in Figure 3.

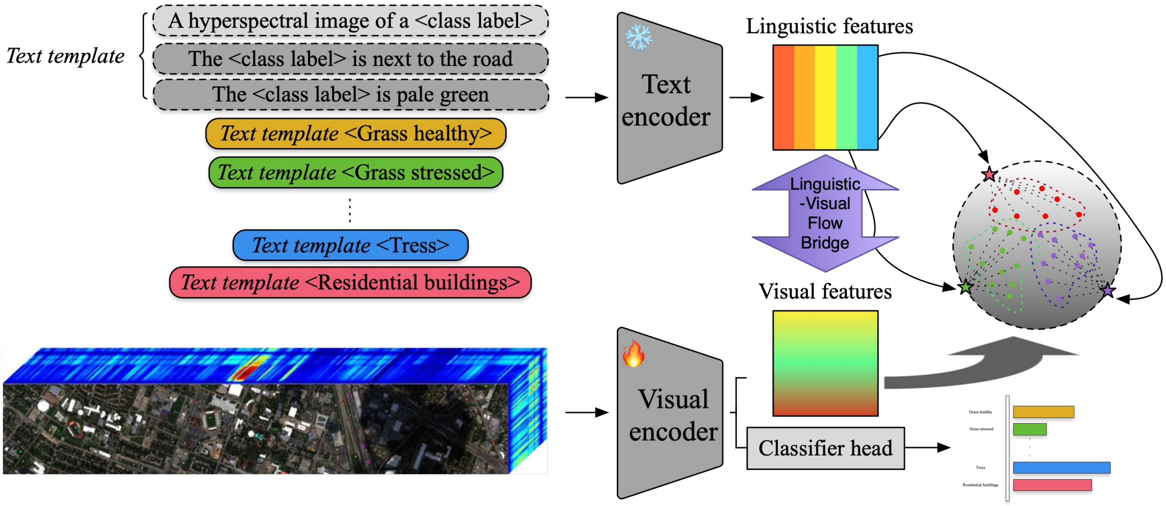
Text template
A hyperspectral image of a <class label>
The <class label> is next to the road
The <class label> is pale green
Text template <Grass healthy>
Text template <Grass stressed>
Text template <Tress>
Text template <Residential buildings>
Text encoder
Linguistic features
Linguistic -Visual Flow Bridge
Visual features
Visual encoder
Classifier head
Grass healthy
Grass stressed
Trees
Residential buildings

Figure 3. Pipeline of the proposed architecture. Firstly, the frozen text encoder and the visual encoder extract the linguistic and visual features, respectively. Then, the LIVE module makes the two modalities mutually communicate to achieve better alignment. Meanwhile, the visual features are fed into the classifier head to obtain the prediction probability.

2.3. Image Encoder

The image encoder consists of a CNN backbone and a simple transformer encoder, as shown in Figure 4 and the detail of the CNN backbone can be described in Figure 5. The subcubes X_s of $13 \times 13 \times d$ from SD are delivered to the CNN backbone for extracting the spatial-spectral features and reducing its spectral dimensions, then the feature embedding is used by the transformer to extract more distinctive features with contextual feature and inter-class correlation.

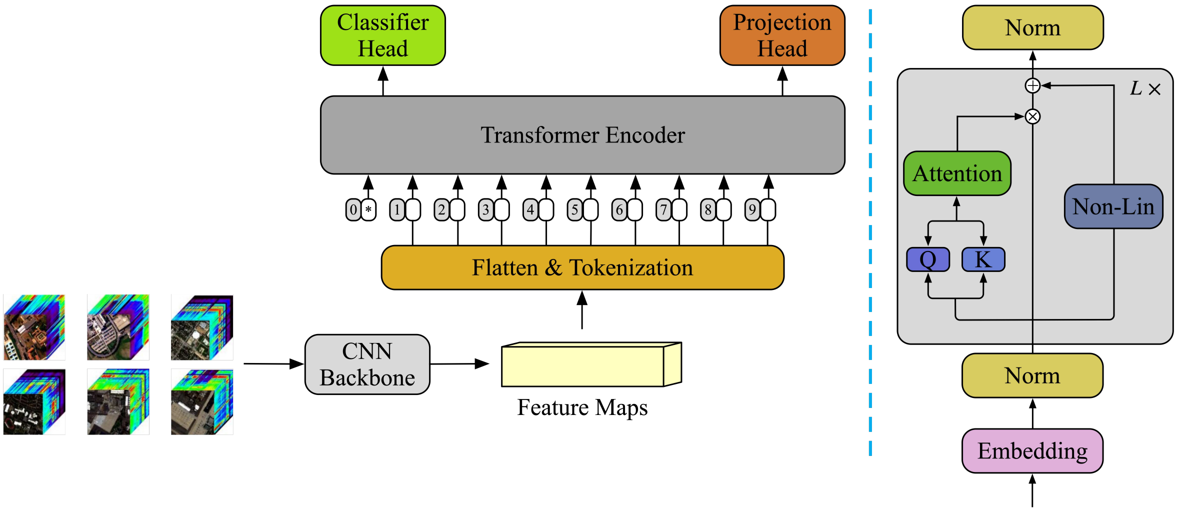

Figure 4. The structure of the image encoder. The Non-Lin denotes a non-linear layer for embedding feature projection.

The proposed CNN backbone consists of 3D-CNN with kernel size $(9, 3, 3)$ to extract the local features and spatial-spectral features, then the padding $(0, 1, 1)$ is used to keep the spatial dimensions the same and reduce the spectral dimensions for lightweight. Correspondingly, a DW block is designed by a bi-branch framework. The depth-wise convolution with kernel size $(3, 3)$ is adopted in one of the branches and a ReLU activation function also is introduced. Correspondingly, the point-wise convolution with kernel size of $(1, 1)$ is equipped in the other. Further, The ultimate feature maps are produced by multiplying the outputs of two branches and adding the output of 3D-CNN utilizing the residual connection. Finally, the feature maps are fed into a simple transformer encoder to extract the inter-class correlation and global contextual features, then the transformer outputs the F_v and Cls matrix. Here, the L is the number of transformer encoder blocks and is initialized with 3 for obtaining a more shallow transformer. In this case, the residual connection is not only unnecessary, because the outputs are low-level inlinear and identity-like, but increases redundancy and latency. To tackle this problem, we introduce a skipless module, which is composed of dual-branch attention without value.

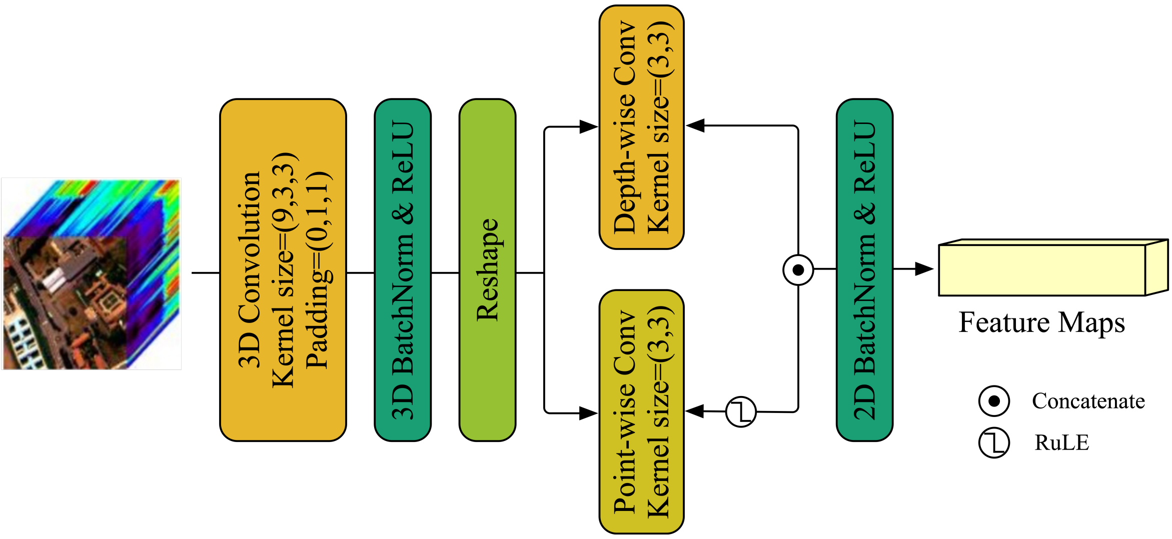

Figure 5. The framework of the CNN backbone.

2.4. LIVE Module

As shown in Figure 6, $F_t \in \mathbb{R}^{C \times D}$ and $F_v \in \mathbb{R}^{HW \times D}$ respectively denote the linguistic features and visual features, which outputted by the corresponding encoder. Here, C denotes the Classes number, and D denotes the feature Dimension. We propose projection layers to produce individual queries, keys, and values. Specifically, V_t, K_t, and Q_t are produced by the projection head using the linguistic features as input. Correspondingly, V_v, K_v, and Q_v denote value, key, and query projected from visual features.

$$Q_t, V_t, K_t = \text{Projection}(F_t) \tag{1}$$
$$Q_v, V_v, K_v = \text{Projection}(F_v) \tag{2}$$

Where Projection$(\cdot)$ denotes the projection head, which consists of a linear layer of $3 \times D$ dimension, and the outputs are split equally as Q,K,V on D dimension. Then two attention maps can be calculated, which are denoted as $A_v \in \mathbb{R}^{HW \times C}$ and $A_t \in \mathbb{R}^{C \times HW}$.

$$A_v = \text{SoftMax}(\frac{Q_t K_v^T}{\sqrt{D}}) \in \mathbb{R}^{C \times HW} \tag{3}$$
$$A_t = \text{SoftMax}(\frac{Q_v K_t^T}{\sqrt{D}}) \in \mathbb{R}^{HW \times C} \tag{4}$$

where A_t and A_v are respectively for linguistic and visual features update.

$$F_v^a = A_v V_t \tag{5}$$
$$F_t^a = A_t V_v, \tag{6}$$

where $F_v^a \in \mathbb{R}^{HW \times D}$ and $F_t^a \in \mathbb{R}^{C \times D}$ respectively denote that updated visual and textual features. With this cross-attention mechanism, the information communication and interaction bridge has been built between linguistic modality to visual modality to align the spatial-spectral and linguistic features with more reasonable and available. On the one hand, the high-level abstract visual features will be focused on with text guidance. On the other hand, this bridge makes the linguistic features flexible, image-conditional, and self-adapted instead of fixed.

In the ultimate stage, We proposed the top-class projection module to extract the linguistic feature vectors corresponding to the class with the highest probability for each sample in the text data. Subsequently, these selected feature vectors are projected into a new semantic space in preparation for domain alignment. Spa-Max Pooling is adopted to retain the distinguish characters and reduce its spatial dimension.

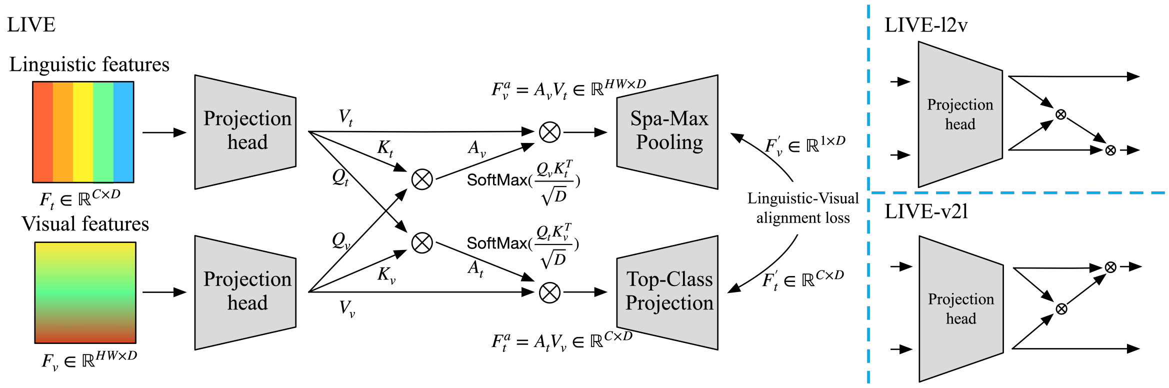
LIVE
Linguistic features
$F_t \in \mathbb{R}^{C\times D}$
Visual features
$F_v \in \mathbb{R}^{HW\times D}$
Projection head
Projection head
V_t
K_t
Q_t
Q_v
K_v
V_v
A_v
A_t
SoftMax($\frac{Q_v K_t^T}{\sqrt{D}}$)
SoftMax($\frac{Q_t K_v^T}{\sqrt{D}}$)
$F_v^a = A_v V_t \in \mathbb{R}^{HW\times D}$
$F_t^a = A_t V_v \in \mathbb{R}^{C\times D}$
Spa-Max Pooling
Top-Class Projection
$F_v^{'} \in \mathbb{R}^{1\times D}$
$F_t^{'} \in \mathbb{R}^{C\times D}$
Linguistic-Visual alignment loss
LIVE-l2v
Projection head
LIVE-v2l
Projection head

Figure 6. The architecture of the proposed LIVE module is depicted on the left side of the image, while on the right side are its two variants based on bidirectional or unidirectional cross-attention mechanisms, where the l2v and the v2l mean the text-guided visual and linguistic features alignment and image-conditional linguistic and visual features alignment framework, respectively.

2.5. Loss Function

Our loss function is composed of the classification loss and the linguistic-visual alignment loss [29]. Firstly, the classification loss can be denoted as

$$\hat{s}_i = \mathrm{Cls}(s_i) \tag{7}$$

$$L_{cls}(S, Y) = \frac{1}{N}\sum_i L_{CE}(\hat{s}_i, y_i), \tag{8}$$

where the L_{CE} is Cross-Entropy loss function and $\hat{y}_i$ is the output of classification branch of our model. In addition, the linguistic-visual alignment loss (L_{LVA}) is inspired by supervised contrastive learning [36](in LDGnet), which is defined as

$$\begin{aligned} L_{LVA} = -\sum_{i=0}^{N}\frac{1}{|P\,(i)\,|}\Bigg(&\sum_{p\in P_t(i)}\log\frac{\exp\left(v_i^T t_p^+/\tau\right)}{\sum_{a\in A_t(i)}\exp\left(v_i^T t_a^-/\tau\right)} \\ &+\sum_{p\in P_v(i)}\log\frac{\exp\left(t_i^T v_p^+/\tau\right)}{\sum_{a\in A_v(i)}\exp\left(t_i^T v_a^-/\tau\right)}\Bigg) \end{aligned} \tag{9}$$

where the v_i and t_i denote the visual feature and textual feature, respectively in minibatch, $P_v(i)$ and $A_v(i)$ means the positive and negative samples of visual feature. Correspondingly, $P_t(i)$ and $A_t(i)$ also are the positive and negative samples of two modalities features. Notably, this loss exists for both coarse-grained and fine-grained text prompts, denoted as L_{coarse} and L_{fine}.

Integrating the above loss functions, the total loss is represented as

$$L = L_{cls} + \alpha((1-\beta)L_{coarse} + \beta L_{fine}) \tag{10}$$

where the α and β is hyperparameters for balancing the L_{LVA} and control the contribution of both L_{coarse} and L_{fine}, respectively.

2.6. Variantes of Proposed Module

For the LIVE module, three versions can be selected, **ALL-LIVE** denotes that the module is used for aligning both coarse-grained text and fine-grained with visual features, **Coarse-LIVE** denotes that only is available in alignment between coarse-grained text and visual features, and similarly the **Fine-LIVE** is designed for only fine-grained text and visual features alignment. For each version, we explore and compare the performance of the unidirectional or bidirectional bridge that has been used for information interaction. For all descriptions and experiments, the bidirectional Coarse-LIVE module is the default, and more detailed discussions will be provided in Section 3.3.4.

3. Results

In this section, we display the details of experiments, including the dataset, evaluation metrics, experiment setting, and the results of comparative and ablation experiments.

3.1. Dataset and evaluation metrics

3.1.1. Dataset Description

1) Houston Dataset: The dataset includes Houston 2013 [30] and Houston 2018 [31]. Houston 2013 conducted research published by the IEEE Geoscience and Remote Sensing Society. The dataset used by UH was gathered in 2013 and involved the deployment of the CASI. Each image in the

dataset comprises 340 × 1905 pixels, encompassing 144 distinct spectral bands, covering a wavelength range spanning from 0.38 to 1.05 μ m. The spatial resolution is set at 2.5 meters per pixel (MPP). The ground truth for the images comprises 15 unique classes that correspond to various land cover types. Correspondingly, Houston 2018 has the same wavelength range but contains 48 spectral bands. The 48 spectral bands within the wavelength range 0.38-1.05μm are selected to correspond to the Houston 2018. The land-cover classes selected and the number of samples are exhibited in Table 1.

Table 1. Land-cover classes and the number of samples of the Houston13 and Houston18 datasets

Class No.	Land-cover Class	Number of Samples (Houston 13)	Number of Samples (Houston 18)
1	Stressed Grass	345	1353
2	Grass stressed	365	4888
3	Trees	365	2766
4	Water	285	22
5	Residential buildings	319	5347
6	Non-residential buildings	408	32459
7	Road	443	6365
	Total	2530	53200

2) Pavia dataset: University of Pavia (UP) has 103 spectral bands with 610×340 pixels. The Pavia Center (PC) has 1096×715 pixels and 102 bands. The last channel of UP is removed to ensure the same as PC on spectral dimension. The details are listed in Table 2.

Table 2. Land-cover classes and the number of samples of the UP and PC

Class No.	Land-cover Class	Number of Samples (UP)	Number of Samples (PC)
1	Tree	3064	7598
2	Asphalt	6631	9248
3	Brick	3682	2685
4	Bitumen	1330	7287
5	Shadow	947	2863
6	Meadow	18649	3090
7	Bare soil	5029	6584
	Total	39332	39355

Quantitative evaluation metrics of four are used to evaluate the effectiveness of the suggested technique, as well as other methods for comparison. These metrics include:

- The class-specific accuracy (CA): Measures the accuracy of the model on each class. It is calculated as the ratio of correctly classified samples for a specific class to the total number of samples in that class.
- Overall Accuracy (OA): A comprehensive assessment of the classification;
- Kappa Coefficient (κ): The kappa coefficient assesses the agreement between the anticipated classifications and the ground truth while taking into account any agreement that might happen by chance.

3.2. Experiment setting

We conducted all experiments using the PyTorch framework on the NVIDIA 3090 GPU. We trained our model with a batch size of 256. The AdamW [32] optimizer has been deployed for training all networks. On the UH and Pavia datasets, the learning rate of $1e-2$, using a gradient of loss function for the gradient descent updates with regularization parameter λ and weight α. Here the λ is 1e+0 for the UH dataset, and 1e-2 for the Pavia dataset, respectively. The data augmentation strategy has been adopted for the mitigation of overfitting, with the multiple ratio r of 5 for the UH dataset and

1 for the Pavia dataset. The training epoch is set to 200. During the experiment, patches with a size of 13 × 13 are taken from the HSIs and used as input to the model. For the UH dataset, all models have been trained on the source domain (Houston 13) and tested on the target domain (Houston 18). For the Pavia dataset, all experiments have been conducted regarding UP and PC as source domain and target domain, respectively.

3.3. Comparison experiments

3.3.1. Comparison method setting

The comparison methods contain DAAN[33], MRAN[34], DSAN[35], HTCNN[36], PDEN[8], LDSDG[37], SagNet[38], and LDGnet[10]. For DAAN, MRAN, DSAN, and HTCNN, which are considered DA techniques, the SD data with labels and TD data without labels have been used for training. Here, the SD data has been split randomly as 80% and 20% for training and validation, respectively. For the DG methods, such as PDEN, LDSDG, SagNet, LDGnet, and Ours, only the SD data has been used, under the same sample partition situation. In particular, the parch size of LDSDG and SagNet has been set to 32 × 32 to accommodate the input size of Resnet18. We adopted the random flip and random radiation noise in [10] to increase the UH dataset. The learning rate of all the comparison methods is 1e-5, 1e-4, 1e-3, 1e-2, 1e-1, 1e+0, respectively and the λ of all comparison models is 1e-3, 1e-2, 1e-1, 1e+0, 1e+1, 1e+2, 1e-2. The result of classification performance is shown in the following Table 3-4, and the best results already are emphasized in bold.

3.3.2. Classificaion performance analysis

The classification performance of different methods for the target Houston 18 and PC dataset is shown in Table 3 and Table 4, respectively. The following analyses can be obtained from classification performance on two datasets.

1. In all DA methods, The DSAN has the best performance on both two datasets. Compared with DG methods, the DSAN provides a 2% improvement in OA over PDEN on the Houston 18 dataset, but on the PC dataset, on the contrary, the PDEN provides a 2.6% improvement in OA over DSEN. This demonstrates that the superiority of DA and DG is equally matched and the key to cross-scene classification is the generalization ability of the model for various data.
2. In all DG methods, the LDGnet achieves the exceptive performance using the linguistic visual alignment, which demonstrates that linguistic prior knowledge is greatly helpful for cross-scene classification and can improve the generalization ability for different datasets.
3. Our LIVEnet outperforms the DG methods. Compared with the second best, ours provides a 3.7% and 1.7% in OA on the Houston 18, and PC, respectively. Because, for the hyperspectral image data, we specially design the image encoder to extract the global spatial-spectral features with good inter-class correlation. In addition, to tackle the prior knowledge of linguistics on HSI data, we propose a LIVE module to build a bridge between linguistic and visual semantic space. With this module, the linguistic visual alignment can be flexible and self-adapted.

We also design the visualization comparison on the UH dataset. The classification maps of the different methods on the Hoston 18 data are illustrated in Figure 8 and Figure 7. The visual outcomes generated by our approach exhibit heightened precision and feature more pronounced edge characteristics.

Table 3. CLASSIFICATION ACCURACY (%) OF DIFFERENT METHODS FOR THE TARGET HOUSTON 18 DATA

Class No.	DAAN	MRAN	DSAN	HTCNN	PDEN	LDSDG	SagNet	LDGnet	LIVEnet
1	68.29	41.02	62.31	11.83	46.49	10.13	25.79	61.71	44.42
2	77.80	76.94	77.50	70.11	77.60	62.97	62.79	77.45	77.41
3	67.50	65.91	74.55	54.99	59.73	60.81	48.66	62.08	62.33
4	100.0	100.0	100.0	54.55	100.0	81.82	81.82	95.45	81.82
5	47.69	36.90	73.39	55.60	49.62	45.65	59.57	69.53	84.65
6	79.49	82.68	86.84	92.85	84.98	89.22	89.28	91.57	92.03
7	45.12	56.43	46.33	46.47	64.21	44.15	34.99	45.42	60.25
OA	70.45±1.54	72.11±1.79	78.21±1.05	77.67±2.11	75.40±1.76	74.23±1.49	73.01±1.81	80.34±1.74	**83.39±1.47**
κ (×100)	53.86±2.14	55.03±2.86	64.10±1.83	60.18±3.07	55.87±2.92	55.42±2.27	55.29±3.50	65.80±2.59	**71.83±2.89**

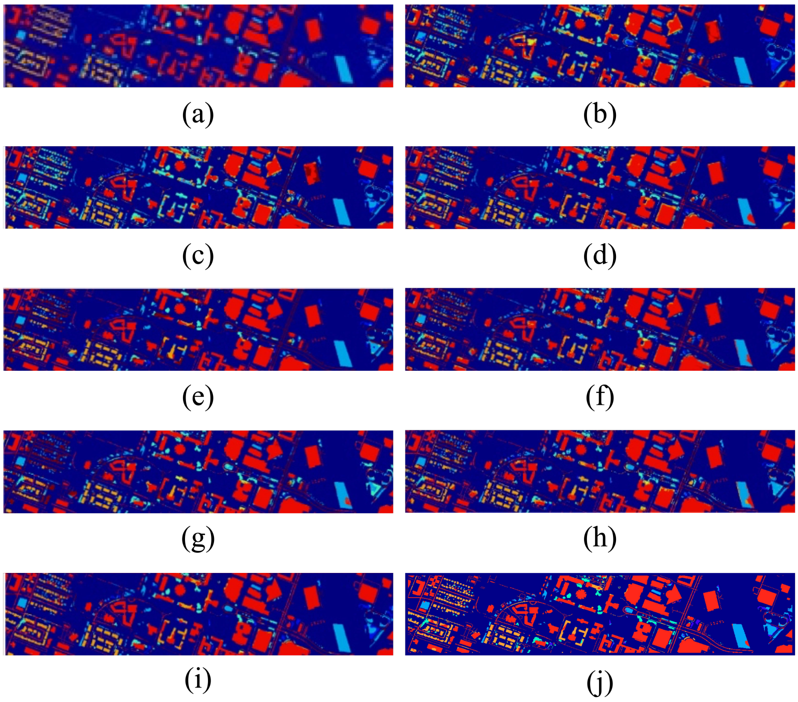
(a) (b) (c) (d) (e) (f) (g) (h) (i) (j)

Figure 7. The classification maps of the different methods on the Houston 18 dataset. (a) Ground truth. (b) DAAN. (c) MRAN. (d) DSAN. (e) HTCNN. (f) PDEN. (g) LDSDG. (h) SagNet. (i) LDGnet. (j) LIVEnet.

Table 4. CLASSIFICATION ACCURACY (%) OF DIFFERENT METHODS FOR THE TARGET PC DATA

Class No.	DAAN	MRAN	DSAN	HTCNN	PDEN	LDSDG	SagNet	LDGnet	LIVEnet
1	71.98	59.16	93.93	96.06	85.93	91.09	98.35	95.20	99.55
2	78.98	85.15	79.80	57.70	88.56	73.51	59.76	82.79	84.95
3	19.37	46.18	53.97	2.76	61.34	2.23	5.40	80.48	47.97
4	58.67	69.58	75.75	93.25	85.49	71.72	87.03	85.11	84.28
5	70.87	64.58	99.44	89.94	87.95	71.04	93.19	93.15	81.28
6	83.07	89.22	74.43	70.97	79.26	57.12	49.81	66.93	45.08
7	55.59	60.10	67.31	42.28	64.75	78.13	57.94	81.97	98.51
OA	66.71±2.15	69.01±1.21	78.21±1.10	68.31±1.62	80.27±1.38	70.88±1.55	69.28±1.11	82.53±1.24	**83.94±0.99**
κ (×100)	59.76±3.47	63.04±1.90	74.10±1.57	62.17±2.57	76.81±1.92	64.06±2.07	63.04±1.67	78.85±1.78	**80.51±1.52**

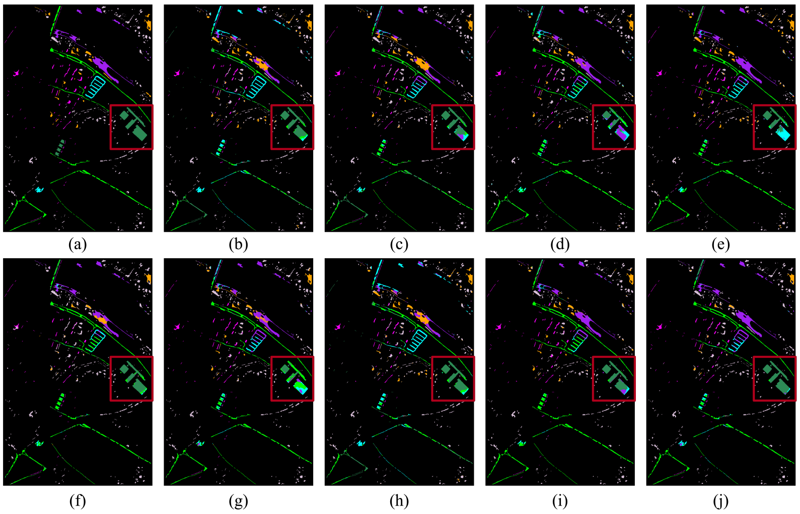

Figure 8. The classification maps of the different methods on the PC dataset. (a) Ground truth. (b) DAAN. (c) MRAN. (d) DSAN. (e) HTCNN. (f) PDEN. (g) LDSDG. (h) SagNet. (i) LDGnet. (j) LIVEnet.

3.3.3. Time complexity analysis

The results of execution time during one epoch training on all datasets are listed in Table 5. The time consumption of our LIVEnet is lower than the DAAN, DSAN, HTCNN, LDSDG, and LDGnet, because the single-head self-attention in Transformer is designed by a dual-branch framework with removing the value. In this Transformer, the residual connection is simplified, which improves the training speed. For the training parameters, excluding the parameters frozen text encoder, the parameters of our model (2.97M) are higher than the LDGnet (0.78M), because the extra cross attention provides the learnable parameters.

Table 5. EXECUTION TIME OF ONE EPOCH TRAINING IN COMPARISON METHODS ON THE UH AND PAVIA DATASET

Datasets	DAAN	MRAN	DSAN	HTCNN	PDEN	LDSDG	SagNet	LDGnet	LIVEnet
Houston 2013	16.04	14.01	15.94	32.22	4.79	40.28	7.63	18.19	15.61
UP	27.97	28.33	29.02	47.99	13.05	65.93	19.76	68.02	62.50

3.3.4. Variants of model analysis

In this section, we discuss various variants of the LIVE model, categorized into two groups based on the service object of the LIVE model and the framework of cross-attention.

1. Variants of the Service Object of LIVE: 1) The All-LIVE model integrates a complete across-attention mechanism for aligning linguistic features and visual features. Where the linguistic features are extracted from both coarse-grained text and fine-grained text. 2) The Fine-LIVE model is employed exclusively in the alignment of fine-grained linguistic features and visual features. 3) Correspondingly, the Coarse-LIVE model serves as a counterpart, focusing on coarse-grained text features.
2. Variants of the cross-attention framework: The Bidirection denotes the complete across-attention with the Q_t, K_t, V_t and Q_v, K_v, V_v. Those Q, K, and V are completely produced and correctly calculated to make attention maps A_t and A_v. With this bidirectional LIVE, the visual features can affect the linguistic features and also accept the feedback and guidance of linguistic features. The l2v and v2l are the unidirectional bridge to make the linguistic and visual features uniaxially communicate by removing the Q_v, K_v or Q_t, K_t, respectively. The LIVE, LIVE-l2v, and LIVE-v2l framework is described in Figure 6.

The classification performance of different service object variants for the target Houston 18 dataset is shown in Table 6. Taken as a whole, the Coarse-LIVE variant provides the best performance on the Houston 18 dataset, with 2.2% and 1.8% improvement compared with the Fine-LIVE and All-LIVE in OA, respectively. Notably, The Fine-LIVE has the damnedest 100.00% precision in class "Water" and the fine-grained text prompt is "The water has a smooth surface.", "The water appears dark blue or black". The emergence of this phenomenon is attributed to the distinct visual and semantic clarity present in the images and language descriptions of a particular category. Here, the language descriptions overtly incorporate specific visual attributes such as color and texture, exhibiting alignment with the linguistic domain priors acquired during CLIP training. Nonetheless, the manual effort required for prompt engineering for each category is both time-consuming and arduous, and may even be deemed impractical. Due to the potential redundancy or mutual conflicts in the information contained within fine-grained and coarse-grained textual features, which impact the performance of linguistic-visual feature alignment, resulting in lower AA (69.61%) and Kappa (69.22%) in the All-LIVE case. Above all, the Coares-LIVE module has been selected in this work, abbreviated as the "LIVE module".

The classification accuracy of the different frameworks of the Coarse-LIVE module for the target Houston 18 data is exhibited in Table 7. The bidirectional LIVE module has the best performance in OA (83.72%) and Kappa (72.10%). This demonstrates the mutual interaction between the linguistic and visual features is the key to alignment.

Table 6. CLASSIFICATION ACCURACY (%) OF DIFFERENT SERVICE OBJECT VARIANTS FOR THE TARGET HOUSTON 18 DATA

Class No.	All-LIVE	Fine-LIVE	Coarse-LIVE
1	41.24	47.89	44.23
2	69.46	71.62	77.59
3	60.88	59.58	62.57
4	81.82	**100.0**	81.89
5	83.60	76.30	84.38
6	92.05	92.58	92.89
7	58.24	56.62	61.11
OA	82.17	81.87	**83.72**
AA	69.61	72.09	**72.11**
κ ($\times$100)	69.22	68.11	**72.10**

Table 7. CLASSIFICATION ACCURACY (%) OF DIFFERENT FRAMEWORKS OF COARSE-LIVE MODULE FOR THE TARGET HOUSTON 18 DATA

Class No.	LIVE-l2v	LIVE-v2l	LIVE
1	15.23	60.90	44.23
2	80.16	65.22	77.59
3	62.62	60.30	62.57
4	81.82	81.82	81.89
5	75.56	87.94	84.38
6	92.64	89.98	92.89
7	61.60	63.33	61.11
OA	82.53	82.03	**83.72**
AA	67.09	72.78	**72.11**
κ ($\times$100)	69.86	69.99	**72.10**

3.3.5. Impact of the window size of the input HSI data analysis

The different window sizes are investigated thoroughly in this paper. Table 8 exhibits the impact of the window size of the input HSI data on the target Houston 18 data. With the window size expanding, both the execution time and the training parameters, excluding the frozen text encoder, are increasing. Additionally, the OA and AA initially ascend and subsequently decline, as shown in Figure 9. This phenomenon may be attributed to the increase in semantic information within each patch as the patch size grows. However, when the patch size becomes huge, it may lead to the loss of local details. After considering achieving a good trade-off between the three metrics, execution time, and training parameters, we ultimately chose a window size of 13 × 13 for all experiments.

Table 8. IMPACT OF THE DIFFERENT WINDOW SIZES ON THE TARGET HOUSTON 18 DATA

Window size	OA	AA	Kappa (%)	Execution time (s)	Training parameters (M)
11	79.94	60.88	63.30	15.16	2.08
13	83.72	72.11	72.10	15.61	2.97
15	80.55	67.95	64.41	19.09	3.73
17	79.98	69.97	66.31	21.29	4.75

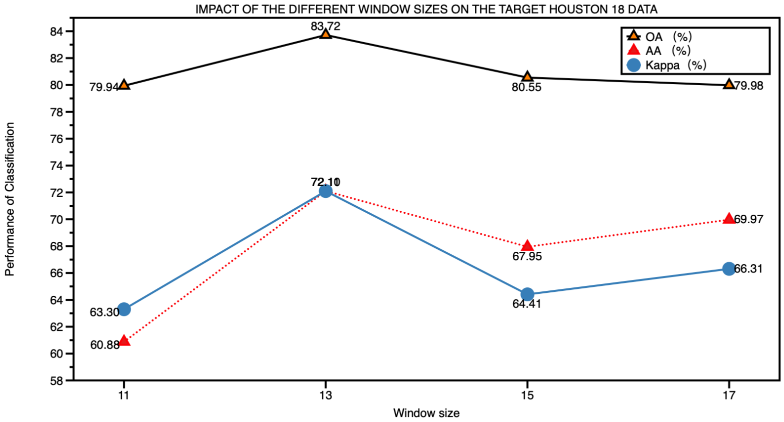

Figure 9. The impact of the different window sizes for classification on target Houston 18 data.

3.4. Ablation experiments

On the Houston dataset, we conducted ablation experiments to evaluate the superiority of the two modules we had introduced. Table 9 displays the experimental results for different cases. To make the baseline network, the Image encoder proposed has been removed in the allover network of LIVEnet, instead of a 3-D CNN residual network, and the LIVE module also has been discarded. For the w/o image encoder case, the 3-D CNN residual network has been instead of the proposed image encoder, and the LIVE module is still adopted. Where the w/o LIVE means the LIVEnet without the LIVE model and maintains the remaining architecture and condition the same. By analyzing the above results of ablation experiments, the effectiveness of the image encoder and the LIVE module is significantly reflected in the OA and Kappa metrics. This demonstrates the image encoder can effectively extract the global features and capture the inter-class correlation, which is important for the performance of classification and generalization of models. Correspondingly, the LIVE module is helpful for linguistic-visual alignment.

Table 9. Ablation experiment on the UH dataset

Cases	OA (%)	Kappa (%)
baseline	79.13	77.62
w/o image encoder	81.20	69.53
w/o LIVE	79.42	78.10
LIVEnet	83.39	71.83

4. Conclusions

In this paper, we proposed a LIVEnet, for cross-scene HSI classification. The LIVEnet is contributed by the CLIP framework and specially designed by a novel image encoder and LIVE module with cross attention. During the training, No samples of TD have been involved. For the target dataset, the LIVEnet is testing only using the Cls loss with low-level time consumption. With the LIVEnet, the local features, contributed by spectrums of HSI, can be extracted well and the linguistic features and visual features can be alignment without the extra prior knowledge. The LIVE module increases the information communication between linguistic and visual features with a simple cross-attention mechanism. The variants of the LIVE module have been fully discussed. The results of comprehensive experiments and analyses on UH and Pavia datasets demonstrate the proposed method's efficacy in improving performance in the domain generalization cross-scene classification.

Abbreviations

The following abbreviations are used in this manuscript:

HSIs	Hyperspectral Images
CNN	Convolution Neural Network
PU	Pavia University
CASI	Compact Airborne Spectrographic Imager
MPP	Meters Per Pixel
GSD	Groups Sampling Distance
IP	IndianPines
UH	University of Houston
POSIS	Reflective Optics System Imaging Spectromete
OA	Overall Accuracy
AA	Average Accurary
κ	Kappa Coefficient

References

1. Makantasis, K.; Karantzalos, K.; Doulamis, A.; Doulamis, N. Deep supervised learning for hyperspectral data classification through convolutional neural networks. 2015 IEEE international geoscience and remote sensing symposium (IGARSS). IEEE, 2015, pp. 4959–4962.
2. Sun, H.; Zheng, X.; Lu, X. A supervised segmentation network for hyperspectral image classification. *IEEE Transactions on Image Processing* **2021**, *30*, 2810–2825.
3. Liu, B.; Yu, X.; Zhang, P.; Yu, A.; Fu, Q.; Wei, X. Supervised deep feature extraction for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing* **2017**, *56*, 1909–1921.
4. Sun, L.; Wu, Z.; Liu, J.; Xiao, L.; Wei, Z. Supervised spectral–spatial hyperspectral image classification with weighted Markov random fields. *IEEE Transactions on Geoscience and Remote Sensing* **2014**, *53*, 1490–1503.
5. Condessa, F.; Bioucas-Dias, J.; Kovačević, J. Supervised hyperspectral image classification with rejection. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* **2016**, *9*, 2321–2332.
6. Wang, J.; Lan, C.; Liu, C.; Ouyang, Y.; Qin, T.; Lu, W.; Chen, Y.; Zeng, W.; Yu, P. Generalizing to unseen domains: A survey on domain generalization. *IEEE Transactions on Knowledge and Data Engineering* **2022**.
7. Zhou, K.; Yang, Y.; Hospedales, T.; Xiang, T. Deep domain-adversarial image generation for domain generalisation. Proceedings of the AAAI conference on artificial intelligence, 2020, Vol. 34, pp. 13025–13032.
8. Li, L.; Gao, K.; Cao, J.; Huang, Z.; Weng, Y.; Mi, X.; Yu, Z.; Li, X.; Xia, B. Progressive domain expansion network for single domain generalization. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 224–233.
9. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; others. Learning transferable visual models from natural language supervision. International conference on machine learning. PMLR, 2021, pp. 8748–8763.
10. Zhang, Y.; Zhang, M.; Li, W.; Wang, S.; Tao, R. Language-aware domain generalization network for cross-scene hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing* **2023**, *61*, 1–12.
11. Kingma, D.P.; Welling, M. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* **2013**.
12. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial nets. *Advances in neural information processing systems* **2014**, *27*.

13. Wei, Y.; Hu, H.; Xie, Z.; Liu, Z.; Zhang, Z.; Cao, Y.; Bao, J.; Chen, D.; Guo, B. Improving CLIP Fine-tuning Performance. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 5439–5449.
14. Gao, P.; Geng, S.; Zhang, R.; Ma, T.; Fang, R.; Zhang, Y.; Li, H.; Qiao, Y. Clip-adapter: Better vision-language models with feature adapters. *International Journal of Computer Vision* **2023**, pp. 1–15.
15. Li, L.H.; Yatskar, M.; Yin, D.; Hsieh, C.J.; Chang, K.W. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557* **2019**.
16. Li, W.; Gao, C.; Niu, G.; Xiao, X.; Liu, H.; Liu, J.; Wu, H.; Wang, H. Unimo: Towards unified-modal understanding and generation via cross-modal contrastive learning. *arXiv preprint arXiv:2012.15409* **2020**.
17. Gao, Y.; Liu, J.; Xu, Z.; Zhang, J.; Li, K.; Ji, R.; Shen, C. Pyramidclip: Hierarchical feature alignment for vision-language model pretraining. *Advances in neural information processing systems* **2022**, *35*, 35959–35970.
18. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; others. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* **2020**.
19. Chollet, F. Xception: Deep learning with depthwise separable convolutions. Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 1251–1258.
20. Wang, J.; Chen, Y.; Feng, W.; Yu, H.; Huang, M.; Yang, Q. Transfer learning with dynamic distribution adaptation. *ACM Transactions on Intelligent Systems and Technology (TIST)* **2020**, *11*, 1–25.
21. Sun, B.; Saenko, K. Deep coral: Correlation alignment for deep domain adaptation. Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part III 14. Springer, 2016, pp. 443–450.
22. Zhou, F.; Jiang, Z.; Shui, C.; Wang, B.; Chaib-draa, B. Domain generalization via optimal transport with metric similarity learning. *Neurocomputing* **2021**, *456*, 469–480.
23. Zhou, K.; Yang, J.; Loy, C.C.; Liu, Z. Learning to prompt for vision-language models. *International Journal of Computer Vision* **2022**, *130*, 2337–2348.
24. Li, X.L.; Liang, P. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190* **2021**.
25. Houlsby, N.; Giurgiu, A.; Jastrzebski, S.; Morrone, B.; De Laroussilhe, Q.; Gesmundo, A.; Attariyan, M.; Gelly, S. Parameter-efficient transfer learning for NLP. International Conference on Machine Learning. PMLR, 2019, pp. 2790–2799.
26. Zhang, R.; Fang, R.; Zhang, W.; Gao, P.; Li, K.; Dai, J.; Qiao, Y.; Li, H. Tip-adapter: Training-free clip-adapter for better vision-language modeling. *arXiv preprint arXiv:2111.03930* **2021**.
27. Yu, Z.; Yu, J.; Cui, Y.; Tao, D.; Tian, Q. Deep modular co-attention networks for visual question answering. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 6281–6290.
28. Yao, Y.; Zhang, A.; Zhang, Z.; Liu, Z.; Chua, T.S.; Sun, M. Cpt: Colorful prompt tuning for pre-trained vision-language models. *arXiv preprint arXiv:2109.11797* **2021**.
29. Khosla, P.; Teterwak, P.; Wang, C.; Sarna, A.; Tian, Y.; Isola, P.; Maschinot, A.; Liu, C.; Krishnan, D. Supervised contrastive learning. *Advances in neural information processing systems* **2020**, *33*, 18661–18673.
30. Debes, C.; Merentitis, A.; Heremans, R.; Hahn, J.; Frangiadakis, N.; van Kasteren, T.; Liao, W.; Bellens, R.; Pižurica, A.; Gautama, S.; others. Hyperspectral and LiDAR data fusion: Outcome of the 2013 GRSS data fusion contest. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* **2014**, *7*, 2405–2418.
31. Le Saux, B.; Yokoya, N.; Hänsch, R.; Prasad, S. 2018 IEEE GRSS data fusion contest: Multimodal land use classification [technical committees]. *IEEE geoscience and remote sensing magazine* **2018**, *6*, 52–54.
32. Dubey, S.R.; Chakraborty, S.; Roy, S.K.; Mukherjee, S.; Singh, S.K.; Chaudhuri, B.B. diffGrad: an optimization method for convolutional neural networks. *IEEE transactions on neural networks and learning systems* **2019**, *31*, 4500–4511.
33. Yu, C.; Wang, J.; Chen, Y.; Huang, M. Transfer learning with dynamic adversarial adaptation network. 2019 IEEE international conference on data mining (ICDM). IEEE, 2019, pp. 778–786.
34. Zhu, Y.; Zhuang, F.; Wang, J.; Chen, J.; Shi, Z.; Wu, W.; He, Q. Multi-representation adaptation network for cross-domain image classification. *Neural Networks* **2019**, *119*, 214–221.
35. Zhu, Y.; Zhuang, F.; Wang, J.; Ke, G.; Chen, J.; Bian, J.; Xiong, H.; He, Q. Deep subdomain adaptation network for image classification. *IEEE transactions on neural networks and learning systems* **2020**, *32*, 1713–1722.

36. He, X.; Chen, Y.; Ghamisi, P. Heterogeneous transfer learning for hyperspectral image classification based on convolutional neural network. *IEEE Transactions on Geoscience and Remote Sensing* **2019**, *58*, 3246–3263.
37. Wang, Z.; Luo, Y.; Qiu, R.; Huang, Z.; Baktashmotlagh, M. Learning to diversify for single domain generalization. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 834–843.
38. Nam, H.; Lee, H.; Park, J.; Yoon, W.; Yoo, D. Reducing domain gap by reducing style bias. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 8690–8699.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.