DiffCLIP: Few-shot Language-driven Multimodal Classifier

Jiaqing Zhang[1*], **Mingxiang Cao**[1*], **Xue Yang**[2], **Kai Jiang**[1], **Yunsong Li**[1†]

[1]The State Key Laboratory of Integrated Services Networks, Xidian University
[2]Shanghai AI Laboratory

Abstract

Visual language models like Contrastive Language-Image Pretraining (CLIP) have shown impressive performance in analyzing natural images with language information. However, these models often encounter challenges when applied to specialized domains such as remote sensing due to the limited availability of image-text pairs for training. To tackle this issue, we introduce DiffCLIP, a novel framework that extends CLIP to effectively convey comprehensive language-driven semantic information for accurate classification of high-dimensional multimodal remote sensing images. DiffCLIP is a few-shot learning method that leverages unlabeled images for pretraining. It employs unsupervised mask diffusion learning to capture the distribution of diverse modalities without requiring labels. The modality-shared image encoder maps multimodal data into a unified subspace, extracting shared features with consistent parameters across modalities. A well-trained image encoder further enhances learning by aligning visual representations with class-label text information from CLIP. By integrating these approaches, DiffCLIP significantly boosts CLIP performance using a minimal number of image-text pairs. We evaluate DiffCLIP on widely used high-dimensional multimodal datasets, demonstrating its effectiveness in addressing few-shot annotated classification tasks. DiffCLIP achieves an overall accuracy improvement of 10.65% across three remote sensing datasets compared with CLIP, while utilizing only 2-shot image-text pairs.

Code — https://github.com/icey-zhang/DiffCLIP

Introduction

Remote sensing images captured over the same geographic region by different sensors often provide complementary ground features (Zhang et al. 2024). Joint classification of multimodal remote sensing data leverages the integration of these complementary sources, enhancing classification accuracy. This approach has been extensively applied in various domains, including urban planning (Dong et al. 2023), natural resource management (Wu, Hong, and Chanussot 2021), and environmental monitoring (Dong et al. 2022), among others. A typical high-dimensional remote sensing image, hyperspectral imaging (HSI), provides rich high-dimensional spectral information, which can be used for material identification based on reflectance values, thus providing new multi-dimensional modal information for remote sensing image classification (Roy et al. 2023b). Effectively utilizing high-dimensional spectral information to integrate and learn features from different modalities to better understand and represent cross-modal features becomes the key to high-dimensional multimodal joint representation learning. Although each modality has unique features, they often share common information in the semantic space. For example, Hazarika *et al.* (Hazarika, Zimmermann, and Poria 2020) introduced a shared subspace to discover potential commonalities between different modalities, aiming to reduce the impact of the modality gap. Dutt *et al.* (Dutt, Zare, and Gader 2022) developed a universal shared manifold model, which can learn shared feature representations from hyperspectral and light detection and ranging (LiDAR) images. Subsequently, the transformer (Roy et al. 2023a) and diffusion model (Zhou et al. 2023) take the classification model to a larger scale. **However, these models only focus on seeking consistency within the latent semantic space in the visual image dimension, lacking joint exploration based on the visual-language view.**

Recently, the contrastive language image pre-training framework (CLIP) (Radford et al. 2021) has achieved remarkable success, providing a foundation for many subsequent tasks such as semantic segmentation (Rao et al. 2022), object detection (Shi and Yang 2023), and 3D point cloud understanding (Zhang et al. 2022). The latest progress in CLIP models mainly focuses on scaling the model size and data size (Cherti et al. 2023), combining self-supervision (Mu et al. 2022), improving pre-training efficiency (Chen et al. 2023), and few-shot adaptation (Zhou et al. 2022b). However, these models often struggle when applied to specific fields such as remote sensing images, especially high-dimensional data as shown in Figure 1. This is because these models are trained on natural images, which may not fully capture the diversity and complexity of specific fields. To address this issue, most research has focused on constructing large-scale pretraining datasets for each domain, with additional fine-tuning stages to adapt to downstream tasks in medical (Wang et al. 2022), e-commerce (Liu et al. 2023a) and remote sensing (Liu et al. 2023b) field. However, the requirement for pro-

[*]These authors contributed equally.

[†]Corresponding author

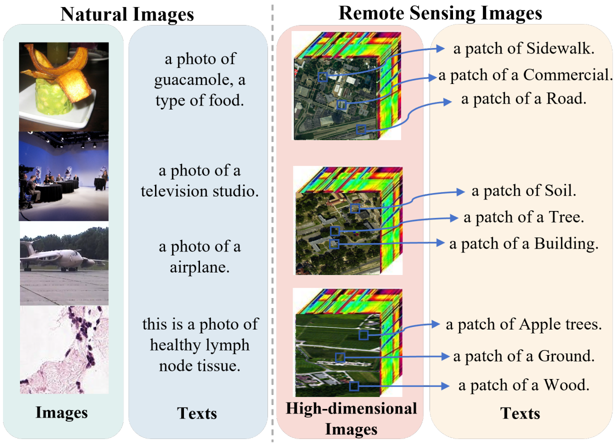


Figure 1: (a) The CLIP model is trained using four randomly sampled image-text pairs from natural image datasets, which are rich in labeled examples.(b) Remote sensing images are annotated in patch format, but there is a severe scarcity of annotated patch samples due to the specialized nature of remote sensing annotations, requiring professional expertise and efficient time management. This significant disparity between datasets makes it challenging to directly apply the CLIP model to remote sensing applications.

fessional availability for large-scale high-dimensional image datasets constrains supervised learning of high-dimensional images. Therefore, a natural question arises: **can we avoid the cost of collecting and labeling data and introduce CLIP into high-dimensional multimodal image classification with fewer labeled samples?** Unfortunately, the use of unsupervised learning knowledge to handle CLIP few-shot training has not been fully explored.

To overcome this issue, we propose DiffCLIP, a novel method that provides a few-shot training paradigm for high-dimensional multimodal remote sensing image classification. We begin by employing mask diffusion unsupervised learning to capture the distributions of various visual image modalities without relying on labels. To enhance the model's ability to extract semantic information, we introduce mask operations and diffusion processes. This operation facilitates sparse representation of input data, reducing interference from redundant information in multimodal remote sensing data, and accelerating training speed. By utilizing a visual image encoder with shared parameters, we map multimodal data into a shared subspace and conduct mask restoration through two modality-specific decoders. This reconstructs input visual image modalities to capture specific attributes of different modalities. The entire encoding and decoding process is integrated into a denoising diffusion model with strong implicit learning capabilities, helping to deal with the difficulty of reconstruction caused by the huge modal gap.

Language-driven few-shot classification is a text feature-driven supervised learning process, where a well-trained visual image encoder enables CLIP to effectively utilize few-shot class label text information for supervision. The semantic information of different modalities is aligned with the class label text information obtained from the text encoder, promoting consistent learning of visual representations across different modalities. Compared to discrete label values, it provides a more comprehensive semantic information representation. Leveraging language-driven methods helps the model capture rich inherent semantic details in complex data distributions, thereby enhancing classification performance.

Extensive experiments on several downstream tasks demonstrate the effectiveness of the proposed DiffCLIP. Specifically, on the Houston dataset, using unlabeled patch samples as the unsupervised mask diffusion pre-training dataset and utilizing the transformer as the image text encoder, DiffCLIP achieved an overall accuracy of 52.15% on 2-shot classification tasks, which is 16.32% higher than the baseline CLIP directly pre-trained with ViT-B-14. Our contributions are summarised as follows:

- We provide a few-shot training paradigm for a high-dimensional multimodal image classification framework, by exploring the potential of CLIP in specific few-shot learning domains.
- We design an unsupervised mask diffusion process to extract shared features across multiple modalities, providing a robust visual image encoder for CLIP to be introduced into the specialized domain of few-shot learning.
- We propose a language-driven framework that introduces class-label text information to enhance the extraction of semantic information of the multimodal visual image encoder, thereby encouraging consistent learning of visual representations across different modalities.

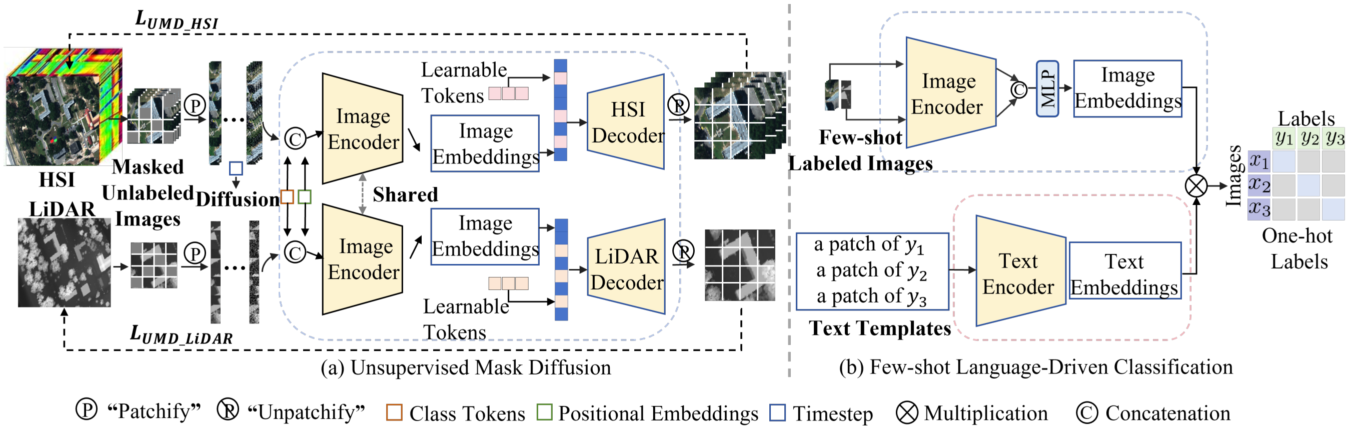
Related Work

Vision-Language Pre-Training (VLP)

Pre-trained visual language models (VLMs) have achieved significant success in various downstream tasks (Bommasani et al. 2021). Key approaches include masked input reconstruction (Kim, Son, and Kim 2021), joint visual-language embedding learning (Yu et al. 2022), language generation from images (Yu et al. 2022), and linking pre-trained unimodal models (Desai and Johnson 2021). CLIP (Radford et al. 2021), a prominent joint embedding model, has shown strong performance across diverse applications. However, VLMs trained on large-scale natural image datasets struggle with the diversity and complexity of specialized fields like remote sensing (Kim et al. 2023). These domains face challenges such as limited data availability, restricted accessibility, and the need for expert annotations, making it difficult to acquire the image-text pairs necessary for effective VLM training.

Few-Shot Learning

Few-shot learning (Li et al. 2020) enables models to adapt to new tasks with limited labeled samples. Traditional methods fall into meta-learning-based (Santoro et al. 2016) and transfer learning-based (Li, Liu, and Bilen 2022) approaches. While few-shot learning has shown success in remote sensing (Dai et al. 2024; Wang et al. 2024), most methods focus on single-modality tasks, with limited exploration of multimodal



(a) Unsupervised Mask Diffusion

(b) Few-shot Language-Driven Classification

Figure 2: The DiffCLIP framework consists of two main stages: a) Unsupervised Mask Diffusion: A modality-shared image encoder captures consistent features across two modalities, while two modality-specific decoders integrate semantic prompts and unique features. b) Few-shot Language-Driven Classification: DiffCLIP fine-tunes the modality-shared encoder and employs language methods to convey comprehensive semantic information. This approach helps capture rich semantic information inherent in complex data distributions.

applications. CLIP, pre-trained on extensive image-text pairs, excels in few-shot classification. Recent works (Zhu et al. 2023; Song et al. 2022) leverage CLIP's multimodal capabilities for few-shot tasks, such as CoOp (Zhou et al. 2022b) and CoCoOp (Zhou et al. 2022a), which automate prompt engineering to reduce reliance on domain expertise. However, adapting CLIP to specialized fields, particularly under data constraints in few-shot remote sensing, remains underexplored. This paper introduces a language-driven few-shot classification model leveraging diffusion-based unsupervised learning to address limited training data and high transfer barriers in domain-specific applications.

Method

DiffCLIP is a few-shot learning method designed to address high-dimensional multimodal remote sensing image classification. It integrates unsupervised training of an image encoder with mask diffusion and text-guided semantic supervision for feature alignment. As shown in Figure 2, the method includes two key stages: unsupervised learning and few-shot learning. The unsupervised learning stage involves forward mask diffusion and reverse denoising restoration. To reduce training costs, only a subset of the dataset undergoes forward mask diffusion, while the reverse process uses a shared image encoder to learn multimodal features, supported by two modality-specific decoders. The few-shot learning stage employs language-driven supervision, providing richer semantic context than discrete labels, enabling better capture of complex data nuances and improving classification. For simplicity, formulas are mainly presented for single-modal inputs, applicable to both modalities.

Unsupervised Mask Diffusion

Mask Diffusion Process Given a clean sample $x_0 \sim \mathcal{Q}(x_0)$ in the forward mask diffusion, DiffCLIP employs an asymmetric masking strategy (He et al. 2022) for each modality, to encourage the model to effectively capture shared features across different modalities. The x_0 is firstly divided into the non-overlapping masked region x_0^m according to a fixed masking ratio and treat the rest as visible patches x_0^v. Only the visible area x_0^v is gradually diffused, corrupted by recursively adding a small amount of Gaussian noise T times with variance $\beta_t \in (0, 1)$ to produce x_1^v, x_2^v, …, x_T^v following the Markov process below:

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

$$\mathcal{Q}(x_1^v, \dots, x_T^v|x_0^v) = \prod_{t=1}^{T} \mathcal{Q}(x_t^v|x_{t-1}^v), \tag{2}$$

where $t \in [1, 2, \dots, T]$ denotes the timestep, $\beta_{1:T}$ is predefined and undergoes a gradual linear decay, held as hyperparameters. The mask diffusion operation creates a task that cannot be easily solved by extrapolating visible neighboring patches (due to adding noise to visible patches), while also minimizing redundant information and resulting in highly sparse inputs. This helps reduce the computational cost of the diffusion model.

Denoising Restoration Process During the reverse process, DiffCLIP predicts input data x_0 based on the current sampling time t. This modification is based on the Bayesian theory as follows:

$$\mathcal{Q}(x_{t-1}|x_t, x_0) = \mathcal{N}(x_{t-1}; \tilde{\mu}_t(x_t, x_0), \tilde{\beta}_t I), \tag{3}$$

when the variance of noise β_t in each step t is small enough, $\mathcal{Q}(x_{t-1}|x_t, x_0)$ can also be considered Gaussian distributed, approximated by a deep network as follows:

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

where $\mu_\theta(x_t^v, t)$ is the mean of x_t^v, $\Sigma_\theta(x_t^v, t)$ is the variance of x_t^v, and the joint distribution $\mathcal{P}_\theta(x_{0:T}^v)$ is defined as a

Markov chain with learned Gaussian transitions starting at $\mathcal{P}(x_T^v) = \mathcal{N}(x_T^v; 0, I)$:

$$\mathcal{P}_\theta(x_{0:T}^v) = \mathcal{P}(x_T^v) \prod_{t=1}^{T} \mathcal{P}_\theta(x_{t-1}^v | x_t^v). \tag{5}$$

DiffCLIP acquires the clean input data x_0 through two parallel unsupervised learning constraints. Initially, it iteratively enhances the predicted scores of the unmasked x_0^v during denoising restoration. Additionally, it restores the masked x_0^m by formulating a set of learnable parameters equivalent to the number of masked patches. In the end, executing through a modality-shared image encoder and two modality-specific decoders, we update the DiffCLIP $f(\cdot)$ by minimizing the denoising restoration loss:

$$\mathbb{E}_{x_0 \sim \mathcal{Q}(x_0)} \mathbb{E}_{x_T \sim \mathcal{N}(0,I)} ||x_0 - f(x_T, t)||^2. \tag{6}$$

Encoder and Decoders The encoder uses standard ViT (Dosovitskiy et al. 2020) and embeds trainable parameters on the diffused visible patches before encoding:

$$\hat{x}^v = \mathcal{E}_\theta \left(\{p_{ct} + p_p + p_{dt}; x^v\}_1 \right), \tag{7}$$

where $\{;\}_1$ means performing concatenate operation in the first dimension. $\mathcal{E}_\theta(\cdot)$ denotes modality-shared image encoder, p_{ct}, p_p, and p_{dt} represent class tokens, positional embedding and diffusion timestep, respectively. It is noteworthy that each modality undergoes its linear projection layer after being mapped to the shared subspace to match dimensions.

To restore the real data of different modalities, we designed two decoders that, compared to the encoder, are more lightweight but fully capable of performing this task. Each decoder takes in visible tokens $\hat{x}^v$ and masked tokens $\hat{x}^m$ as the input. Each of the masked tokens is a learnable vector initialized to zeros. Similar to the encoder, we embed trainable positional embedding into the input data before decoding:

$$\hat{x} = \mathcal{D}_\theta \left(\{p_p^v + \hat{x}^v; p_p^m + \hat{x}^m\}_1 \right), \tag{8}$$

where $\hat{x}$ represents the reconstructed data restored to the original space and $\mathcal{D}_\theta(\cdot)$ means modality-specific decoder. The introduced diffusion model creates a challenging yet effective task to assist the decoder in self-restoration.

The unsupervised training process entails defining a hybrid optimization objective $\mathcal{L}_{\text{UMD}}$, which incorporates the denoising restoration loss of visible patches and the restoration loss of masked patches:

$$\mathcal{L}_{\text{UMD}} = \mathbb{E}_{x_0} \|x_0 - \hat{x}\|^2. \tag{9}$$

Few-shot Language-Driven Classification

CLIP is a widely used vision-language model that learns a joint embedding space between images x and texts y by training on N paired image-text data $\{x_i, y_i\}_{i=1}^N$ using contrastive learning. In this setup, positive pairs (matching images and texts) are aligned, while negative pairs (non-matching) are separated, enabling CLIP to perform zero-shot predictions based on text prompts. However, in the context of high-dimensional multimodal remote sensing, the lack of sufficient image-text pairs makes it difficult to train a robust model for accurate classification. DiffCLIP addresses this challenge by employing an unsupervised mask diffusion pre-training process, which enables the training of an effective image encoder using unlabeled samples. The training objective of DiffCLIP involves two classification tasks (Mo et al. 2023): predicting a text given an image $p(y \mid x)$ and predicting an image given a text $p(x \mid y)$. Each sample in the batch is assigned a label corresponding to its paired data as the target.

Text Description Generation In this section, we generate specific class text descriptions beyond simply appending semantically informative adjectives, and refrain from loading any pre-trained weights for the text encoder, to underscore the effectiveness of our method. Specifically, we give GPT-4o prior knowledge to generate text descriptions containing inherent attributes, inter-class relationships, and class names. Taking the class “Healthy Grass” as an example, we prompt that the grass is predominantly green in color, possesses a fine texture, and exhibits uniform distribution. We also suggest that healthy grass typically thrives under large trees or alongside roads, rooted in soil. These expanded text descriptions for specific classes can establish strong semantic associations between text and visual features to ensure a consistent representation. Before encoding, we obtain the tokenized representation of the text information through simple embedding. Subsequently, we utilize the transformer to encode the tokenized representation and produce text feature embedding, and the embedding is normalized to have a unit norm denoted as:

$$v_{\text{text}} = \mathcal{T}_\theta(y), z_{\text{text}} = v_{\text{text}}/\|v_{\text{text}}\|, \tag{10}$$

where y corresponds to tokenized text inputs and $\mathcal{T}_\theta(\cdot)$ denotes the transformer, these text feature vectors serve as semantic information for text and are utilized to align with the image feature space.

Visual Features Representation Our primary focus is on obtaining visual features. As previously mentioned, the modality-shared encoder, trained using unsupervised mask diffusion, serves as a replacement for the image encoder in CLIP. The parameters of this modality-shared encoder are loaded, with only the linear projection layer cascaded for dimensional reduction being randomly initialized, then they are fine-tuned together. We posit that the modality-shared encoder has effectively captured common information between the two modalities. Moreover, it has been imbued with robust semantic cues through the denoising restoration task during the decoding process, while preserving each modality's distinctive features. The process is formulated as follows:

$$v_{\text{image}} = \mathcal{E}_\theta(x), z_{\text{image}} = v_{\text{image}}/\|v_{\text{image}}\|, \tag{11}$$

where x corresponds to patched image inputs. v_{image} is the feature vector obtained by encoding the patched image, and it is normalized to get z_{image}.

DiffCLIP utilizes a softmax function with a temperature parameter $\tau > 0$ applied to the cosine similarity of embeddings to predict the label:

$$p(x \mid y) = \sigma_\tau \left(z_{\text{text}}, \left\{ z_{\text{image}}^i \right\}_{i=1}^N \right) \in \mathbb{R}^N, \tag{12}$$

$$p(y \mid x) = \sigma_\tau \left(z_{\text{image}}, \left\{ z_{\text{text}}^i \right\}_{i=1}^N \right) \in \mathbb{R}^N, \tag{13}$$

where the softmax classifier σ_τ represents a function of input and output embeddings, and probabilities are computed as follows:

$$p\left(x=x_i \mid y\right)=\frac{\exp\left(z_{\text{text}} \cdot z_{\text{image}}^{i} / \tau\right)}{\sum_{j=1}^{N} \exp\left(z_{\text{text}} \cdot z_{\text{image}}^{j} / \tau\right)}, \tag{14}$$

$$p\left(y=y_i \mid x\right)=\frac{\exp\left(z_{\text{image}} \cdot z_{\text{text}}^{i} / \tau\right)}{\sum_{j=1}^{N} \exp\left(z_{\text{image}} \cdot z_{\text{text}}^{j} / \tau\right)}. \tag{15}$$

The DiffCLIP model minimizes the following objectives:

$$\mathcal{L}_{\text{FLC}}=\frac{1}{2N}\sum_{i=1}^{N}\left(H\left(p\left(y \mid x_i\right), \mathbf{e}_i\right)+H\left(p\left(x \mid y_i\right), \mathbf{e}_i\right)\right), \tag{16}$$

where H denotes the cross-entropy loss and $\mathbf{e}_i \in \mathbb{R}^N$ is a one-hot vector with the i-th element being one.

Experiments

Experiments Setup

Datasets Description The experiments are conducted on four widely recognized benchmarks to assess the performance of our proposed method: Houston (Debes et al. 2014), Trento (Rasti, Ghamisi, and Gloaguen 2017), MUUFL (Gader et al. 2013) and MRNet dataset (Bien et al. 2018).

Evaluation Metric and Comparison Methods To evaluate the classification outcomes quantitatively, we employ three key metrics: overall accuracy (OA), average accuracy (AA), and Kappa coefficient. To assess the effectiveness of our method, we select four SOTA methods in multimodal remote sensing classification: GLT (Ding et al. 2022), CALC (Lu et al. 2023), MIViT (Zhang et al. 2024), LDS^{2}AE (Qu et al. 2024), five SOTA few-shot learning algorithms: RN-FSC (Gao et al. 2020), MFRN-ML (Dai et al. 2024), SC-Former (Li et al. 2024), HIPL (Yin et al. 2024), MMPR (Wu et al. 2024), and CLIP (Radford et al. 2021).

Implementation Details The experiments are conducted on a system with an NVIDIA GeForce RTX A100 GPU. During preprocessing, training samples are cropped into 11 × 11 patches. For optimization in both unsupervised and few-shot learning, the Adam optimizer is used with an initial learning rate of 1e-4 and weight decay of 1e-5. Two schedulers are employed: a cosine scheduler for unsupervised learning and a step scheduler for few-shot learning. The training consists of 100 epochs for unsupervised learning and 150 epochs for few-shot learning. To ensure optimal performance in comparative experiments, the batch size is set to 256 for unsupervised learning and 64 for few-shot learning, with consistent parameter settings across all datasets.

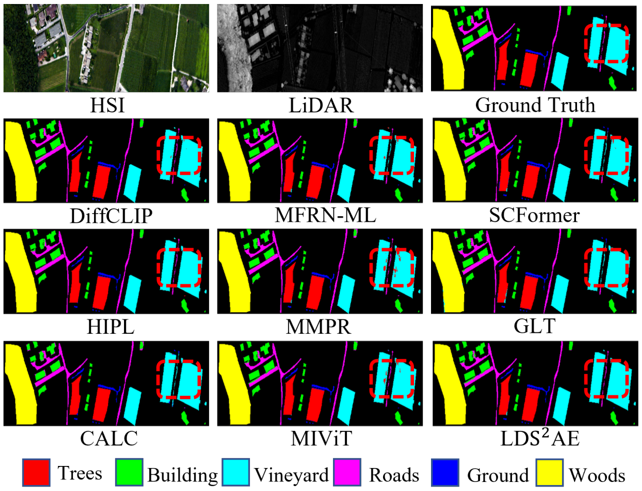


Figure 3: Classification maps of the Trento dataset.

Comparison Results

To evaluate our proposed method and compare it with current SOTA methods in the few-shot learning classification task, we conduct experiments on diverse datasets: Houston, Trento, and MUUFL. As shown in Table 1, DiffCLIP consistently outperforms other models in three datasets, showing improved performance as the number of training images increases. By leveraging unsupervised mask diffusion, our method effectively learns modality-specific and shared features across modalities while capturing individual modality data distributions. Moreover, through semantic supervision and modality-specific feature enhancement via denoising restoration tasks, our method effectively extracts multimodal information even with limited training samples and achieves an overall accuracy improvement of 10.65% across three datasets compared with CLIP, highlighting DiffCLIP's few-shot learning capabilities. We also conduct qualitative evaluations by visualizing the classification maps on the Trento dataset in Figure 3. DiffCLIP achieves optimal visual classification performance, effectively utilizing semantic information, particularly demonstrating perfect classification continuity in categories such as Roads.

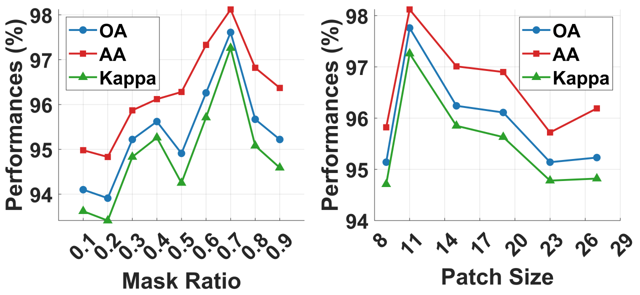


Figure 4: Classification performance of Houston dataset with different masking ratios and patch size.

Ablation Studies

Ablation of Sample Numbers The impact of sample size in the unsupervised mask diffusion stage is evaluated by varying the number of samples from 300 to 700 in Table 2. Results

Settings	Methods	Houston			MUUFL			Trento		
		OA(%)	AA(%)	Kappa(%)	OA(%)	AA(%)	Kappa(%)	OA(%)	AA(%)	Kappa(%)
2-shot	CLIP	35.83	42.75	32.44	51.62	38.26	38.91	85.33	78.63	82.47
	LDS^{2}AE	48.68	51.68	44.85	56.47	44.86	43.78	90.23	79.25	86.76
	MFRN-ML	48.73	51.55	43.79	57.16	44.54	43.12	90.81	79.33	86.04
	SCFormer	49.09	52.93	46.11	57.37	46.94	45.13	91.55	80.05	88.24
	HIPL	50.96	54.33	47.15	58.84	52.16	47.37	92.11	80.14	88.59
	MMPR	49.79	53.01	46.78	57.86	49.05	46.15	91.75	79.88	87.93
	DiffCLIP	**52.15**	**56.02**	**48.39**	**59.39**	**54.94**	**48.80**	**93.19**	**80.45**	**90.85**
8-shot	CLIP	52.11	56.37	52.86	67.56	66.78	63.21	92.76	91.66	92.53
	LDS^{2}AE	57.72	57.60	54.36	70.59	68.43	62.94	95.49	95.62	94.03
	MFRN-ML	58.03	60.59	55.71	72.36	70.28	65.39	95.54	95.78	94.02
	SCFormer	59.61	60.92	56.25	74.38	70.64	67.82	95.57	95.79	94.18
	HIPL	60.36	63.89	57.54	75.65	72.63	69.15	95.98	95.91	94.79
	MMPR	59.98	62.57	56.92	74.67	71.68	67.98	95.63	95.72	94.33
	DiffCLIP	**61.93**	**65.93**	**58.95**	**77.60**	**74.97**	**71.53**	**96.30**	**96.22**	**95.09**
20-shot	CLIP	56.31	59.98	54.32	67.30	71.79	65.10	91.28	93.11	92.43
	LDS^{2}AE	65.71	70.50	63.09	78.13	77.24	72.01	98.13	96.91	97.50
	MFRN-ML	74.49	77.81	73.22	78.83	77.45	72.32	98.25	96.98	97.43
	SCFormer	74.97	78.16	73.80	79.11	78.55	73.23	98.27	96.95	97.36
	HIPL	79.36	82.11	76.35	80.87	79.36	76.28	98.41	97.15	97.86
	MMPR	77.93	80.26	75.70	80.02	78.71	74.63	98.40	97.03	97.42
	DiffCLIP	**81.87**	**84.09**	**80.40**	**81.81**	**80.67**	**76.64**	**98.60**	**97.90**	**98.13**

Table 1: Comparison results on the three datasets under various few-shot settings.

Sample numbers	OA(%)	AA(%)	Kappa(%)
300	91.75	91.11	91.62
500	92.58	91.76	93.00
700	**94.80**	**94.24**	**93.28**

Table 2: Ablation of sample numbers used in unsupervised mask diffusion stage on three datasets.

show that increasing the sample size leads to notable improvements in OA, AA, and Kappa coefficient. Specifically, the model achieves its best performance with 700 samples. This indicates that a larger sample size enhances the model's classification accuracy and reliability, with 700 samples offering an optimal balance between performance and efficiency.

Ablation of Parameter Settings The study in Figure 4 on DiffCLIP's performance explores the effects of masking ratios and patch sizes. Increasing the masking ratio initially improves performance by reducing redundancy, but excessive masking leads to information loss. Results show a peak at 70%, which is selected for further experiments. Similarly, ablation experiments reveal that a patch size of 11 performs best, balancing category separation and receptive field size for optimal results.

Ablation of Model Architecture The study in Table 3 evaluates DiffCLIP by comparing it to four baseline approaches: i) replacing the text encoder with a fully connected layer, ii) removing the masking process, iii) removing the diffusion process, and iv) eliminating the entire unsupervised mask diffusion process. Results show that removing the text encoder reduces OA by 5.45%, underscoring the importance of textual information for semantics. Removing the diffusion and masking processes decreases OA by 5.13% and 4.02%, respectively, highlighting their role in model robustness. Eliminating the entire unsupervised mask diffusion leads to a significant 6.98% OA drop, affirming its critical role in adapting CLIP to remote sensing. Overall, DiffCLIP consistently outperforms the baselines.

Settings	OA(%)	AA(%)	Kappa(%)
w/o Text	89.35	89.66	89.01
w/o Diffusion	89.67	90.73	89.28
w/o Mask	90.78	91.72	89.83
w/o Unsupervised	87.82	88.94	87.76
DiffCLIP	**94.80**	**94.24**	**93.28**

Table 3: Ablation of model components on three datasets.

Ablation of Text Prompts DiffCLIP's classification performance benefits from language-driven methods. To assess the impact of text prompts, five crafted prompt sets are tested on the Houston dataset. The baseline, p1, is compared against extended prompts. As seen in Table 4, p2's longer text reduces accuracy due to weaker task relevance, while p3's minor adjustments with effective "fusion" text achieve minimal improvement. p4 demonstrates that longer, relevant prompts improve supervision. p5's specific class descriptions boosted OA accuracy by 0.54%. This highlights the importance of optimizing both prompt length and content for better performance in DiffCLIP.

	Context length	Text prompts	OA(%)	AA(%)	Kappa(%)
p1	5∼7	a patch of a {class name}.	97.61	98.12	97.26
p2	6∼8	a nice patch of a {class name}.	97.05	97.63	96.81
p3	7∼9	a fusion patch of a {class name}.	97.63	98.11	97.34
p4	11∼13	a multimodal fusion patch of a {class name} with strong semantic information.	97.93	98.51	97.64
p5	15∼30	specific class text descriptions.	**98.15**	**98.87**	**98.09**

Table 4: Ablation experiments of different text prompts and context length of Houston2013 dataset.

Methods		Houston			MUUFL			Trento		
		OA(%)	AA(%)	Kappa(%)	OA(%)	AA(%)	Kappa(%)	OA(%)	AA(%)	Kappa(%)
Supervised	GLT	90.13	90.42	89.42	82.75	75.70	78.67	98.19	97.75	98.04
	CALC	87.87	88.92	86.87	81.94	64.09	77.01	97.11	92.31	96.64
	MIViT	93.21	93.87	92.75	83.13	79.74	78.91	98.03	97.96	98.24
	LDS^{2}AE	94.88	95.31	94.46	84.82	82.19	80.42	98.77	98.11	98.42
Few-shot	RN-FSC	93.42	94.53	93.64	84.83	74.52	81.25	98.69	97.66	98.61
	MFRN-ML	94.50	95.38	93.99	84.78	77.03	81.07	98.37	97.71	98.49
	SCFormer	95.96	96.25	94.84	84.93	79.49	81.39	98.52	97.86	98.55
	HIPL	96.83	97.01	96.75	85.45	80.71	82.18	98.81	98.15	98.65
	MMPR	96.54	96.77	95.98	85.07	80.36	81.57	98.43	97.79	98.46
	DiffCLIP	**98.15**	**98.87**	**98.09**	**86.98**	**85.01**	**82.81**	**99.26**	**98.84**	**98.95**

Table 5: Comparison results of OA, AA, and Kappa on three datasets.

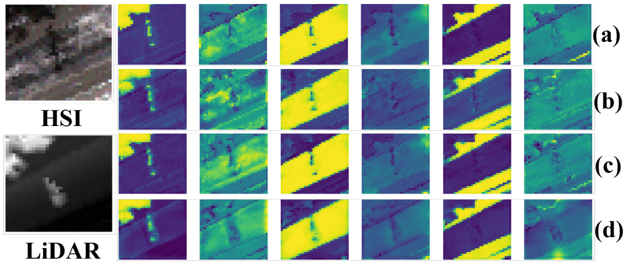


Figure 5: Feature visualization of the visual encoder.

Feature Visualization

The few-shot learning is based on the pre-trained visual encoder obtained from the unsupervised mask diffusion. On this basis, the text encoder is added, and a supervised learning process is conducted with a mere number of paired images and texts (*e.g.* 2 shot), which utilizes a limited number of labeled samples to train the model. We plot the visual encoder feature visualization of different modalities in Figure 5. (a) and (c) are from unsupervised mask diffusion, (b) and (d) are from few-shot language-driven classification. This similar visualization highlights the excellent performance of DiffCLIP for domain-invariant feature extraction and eliminates the domain shift.

Generalization Validation

We also compare with fully supervised algorithms in Tabel 5. For fair comparison, we randomly sample 40 samples per class for training with labels, and the remaining samples for evaluation. Our approach consistently achieves the highest scores across all datasets and evaluation metrics, including OA, AA, and Kappa. To validate the generalization of our method to the other fields, we compare the DiffCLIP with ELNet (Wu et al. 2021), TransMed-S (Dai, Gao, and Liu 2021), and SSL-DcGaR (Berenguer et al. 2024), using 10 samples of MRNet data to train and the rest to test. As shown in Table 6, even in the multimodal medical field, DiffCLIP still maintains better performance.

Method	ACC	AUC	SE	SP
ELNet	0.639	0.703	0.624	0.650
TransMed-S	0.667	0.705	0.635	0.664
SSL-DcGaR	0.731	0.758	0.723	0.734
DiffCLIP	**0.763**	**0.787**	**0.750**	**0.756**

Table 6: Comparison results of the MRNet dataset.

Conclusion

In summary, DiffCLIP introduces a novel approach to high-dimensional multimodal few-shot remote sensing image classification, addressing the challenge of limited training samples. It employs unsupervised mask diffusion pre-training to map multimodal data into a shared subspace, learning shared features while preserving modality-specific details through dedicated decoders. With minimal training samples, DiffCLIP provides a robust image feature encoder and reduces computational costs. Leveraging language-driven methods, it captures rich semantic details and integrates text with visual features for enhanced multimodal fusion. Experimental results on benchmark datasets validate the method's effectiveness and robustness in improving classification performance.

References

Berenguer, A. D.; Kvasnytsia, M.; Bossa, M. N.; Mukherjee, T.; Deligiannis, N.; and Sahli, H. 2024. Semi-supervised medical image classification via distance correlation minimization and graph attention regularization. *Medical Image Analysis (MIA)*, 94: 103107.

Bien, N.; Rajpurkar, P.; Ball, R. L.; Irvin, J.; Park, A.; Jones, E.; Bereket, M.; Patel, B. N.; Yeom, K. W.; Shpanskaya, K.; et al. 2018. Deep-learning-assisted diagnosis for knee magnetic resonance imaging: development and retrospective validation of MRNet. *PLoS medicine*, 15(11): e1002699.

Bommasani, R.; Hudson, D. A.; Adeli, E.; Altman, R.; Arora, S.; von Arx, S.; Bernstein, M. S.; Bohg, J.; Bosselut, A.; Brunskill, E.; et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Chen, D.; Wu, Z.; Liu, F.; Yang, Z.; Zheng, S.; Tan, Y.; and Zhou, E. 2023. ProtoCLIP: Prototypical Contrastive Language Image Pretraining. *IEEE Transactions on Neural Networks and Learning Systems (TNNLS)*.

Cherti, M.; Beaumont, R.; Wightman, R.; Wortsman, M.; Ilharco, G.; Gordon, C.; Schuhmann, C.; Schmidt, L.; and Jitsev, J. 2023. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2818–2829.

Dai, M.; Xing, S.; Xu, Q.; Wang, H.; Li, P.; Sun, Y.; Pan, J.; and Li, Y. 2024. Learning transferable cross-modality representations for few-shot hyperspectral and LiDAR collaborative classification. *International Journal of Applied Earth Observation and Geoinformation (JAG)*, 126: 103640.

Dai, Y.; Gao, Y.; and Liu, F. 2021. Transmed: Transformers advance multi-modal medical image classification. *Diagnostics*, 11(8): 1384.

Debes, C.; Merentitis, A.; Heremans, R.; Hahn, J.; Frangiadakis, N.; van Kasteren, T.; Liao, W.; Bellens, R.; Pižurica, A.; Gautama, S.; et al. 2014. 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 (JSTARS)*, 7(6): 2405–2418.

Desai, K.; and Johnson, J. 2021. VirTex: Learning Visual Representations from Textual Annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Ding, K.; Lu, T.; Fu, W.; Li, S.; and Ma, F. 2022. Global–local transformer network for HSI and LiDAR data joint classification. *IEEE Transactions on Geoscience and Remote Sensing (TGRS)*, 60: 1–13.

Dong, W.; Zhang, T.; Qu, J.; Xiao, S.; Zhang, T.; and Li, Y. 2022. Multibranch feature fusion network with self-and cross-guided attention for hyperspectral and LiDAR classification. *IEEE Transactions on Geoscience and Remote Sensing (TGRS)*, 60: 1–12.

Dong, W.; Zhao, J.; Qu, J.; Xiao, S.; Li, N.; Hou, S.; and Li, Y. 2023. Abundance matrix correlation analysis network based on hierarchical multihead self-cross-hybrid attention for hyperspectral change detection. *IEEE Transactions on Geoscience and Remote Sensing (TGRS)*, 61: 1–13.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *Proceedings of the IEEE/CVF International Conference on Learning Representations (ICLR)*.

Dutt, A.; Zare, A.; and Gader, P. 2022. Shared manifold learning using a triplet network for multiple sensor translation and fusion with missing data. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing (JSTARS)*, 15: 9439–9456.

Gader, P.; Zare, A.; Close, R.; Aitken, J.; and Tuell, G. 2013. MUUFL Gulfport hyperspectral and LiDAR airborne data set. *Univ. Florida, Gainesville, FL, USA, Tech. Rep. REP-2013-570*.

Gao, K.; Liu, B.; Yu, X.; Qin, J.; Zhang, P.; and Tan, X. 2020. Deep relation network for hyperspectral image few-shot classification. *Remote Sensing (RS)*, 12(6): 923.

Hazarika, D.; Zimmermann, R.; and Poria, S. 2020. Misa: Modality-invariant and-specific representations for multimodal sentiment analysis. In *Proceedings of the 28th ACM International Conference on Multimedia (ACMMM)*, 1122–1131.

He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; and Girshick, R. 2022. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 16000–16009.

Kim, W.; Son, B.; and Kim, I. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. In *Proceedings of the IEEE/CVF International Conference on Machine Learning (ICML)*, 5583–5594. PMLR.

Kim, Y.; Mo, S.; Kim, M.; Lee, K.; Lee, J.; and Shin, J. 2023. Bias-to-text: Debiasing unknown visual biases through language interpretation. *arXiv preprint arXiv:2301.11104*.

Li, A.; Huang, W.; Lan, X.; Feng, J.; Li, Z.; and Wang, L. 2020. Boosting few-shot learning with adaptive margin loss. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 12576–12584.

Li, J.; Zhang, Z.; Song, R.; Li, Y.; and Du, Q. 2024. SCFormer: Spectral coordinate transformer for cross-domain few-shot hyperspectral image classification. *IEEE Transactions on Image Processing (TIP)*.

Li, W.-H.; Liu, X.; and Bilen, H. 2022. Cross-domain few-shot learning with task-specific adapters. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 7161–7170.

Liu, F.; Chen, D.; Du, X.; Gao, R.; and Xu, F. 2023a. MEP-3M: A large-scale multi-modal E-commerce product dataset. *Pattern Recognition (PR)*, 140: 109519.

Liu, F.; Chen, D.; Guan, Z.; Zhou, X.; Zhu, J.; and Zhou, J. 2023b. Remoteclip: A vision language foundation model for remote sensing. *arXiv preprint arXiv:2306.11029*.

Lu, T.; Ding, K.; Fu, W.; Li, S.; and Guo, A. 2023. Coupled adversarial learning for fusion classification of hyperspectral and LiDAR data. *Information Fusion (IF)*, 93: 118–131.

Mo, S.; Kim, M.; Lee, K.; and Shin, J. 2023. S-clip: Semi-supervised vision-language learning using few specialist captions. *Advances in Neural Information Processing Systems (NeurIPS)*, 36: 61187–61212.

Mu, N.; Kirillov, A.; Wagner, D.; and Xie, S. 2022. Slip: Self-supervision meets language-image pre-training. In *Proceedings of the IEEE/CVF European Conference on Computer Vision (ECCV)*, 529–544. Springer.

Qu, J.; Yang, Y.; Dong, W.; and Yang, Y. 2024. LDS2AE: Local Diffusion Shared-Specific Autoencoder for Multimodal Remote Sensing Image Classification with Arbitrary Missing Modalities. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 38, 14731–14739.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the IEEE/CVF International Conference on Machine Learning (ICML)*, 8748–8763. PMLR.

Rao, Y.; Zhao, W.; Chen, G.; Tang, Y.; Zhu, Z.; Huang, G.; Zhou, J.; and Lu, J. 2022. Denseclip: Language-guided dense prediction with context-aware prompting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 18082–18091.

Rasti, B.; Ghamisi, P.; and Gloaguen, R. 2017. Hyperspectral and LiDAR fusion using extinction profiles and total variation component analysis. *IEEE Transactions on Geoscience and Remote Sensing (TGRS)*, 55(7): 3997–4007.

Roy, S. K.; Deria, A.; Hong, D.; Rasti, B.; Plaza, A.; and Chanussot, J. 2023a. Multimodal fusion transformer for remote sensing image classification. *IEEE Transactions on Geoscience and Remote Sensing (TGRS)*.

Roy, S. K.; Deria, A.; Shah, C.; Haut, J. M.; Du, Q.; and Plaza, A. 2023b. Spectral–spatial morphological attention transformer for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing (TGRS)*, 61: 1–15.

Santoro, A.; Bartunov, S.; Botvinick, M.; Wierstra, D.; and Lillicrap, T. 2016. Meta-learning with memory-augmented neural networks. In *Proceedings of the IEEE/CVF International Conference on Machine Learning (ICML)*, 1842–1850. PMLR.

Shi, C.; and Yang, S. 2023. Edadet: Open-vocabulary object detection using early dense alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 15724–15734.

Song, H.; Dong, L.; Zhang, W.-N.; Liu, T.; and Wei, F. 2022. Clip models are few-shot learners: Empirical studies on vqa and visual entailment. *arXiv preprint arXiv:2203.07190*.

Wang, Z.; Wu, Z.; Agarwal, D.; and Sun, J. 2022. MedCLIP: Contrastive Learning from Unpaired Medical Images and Text. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 3876–3887.

Wang, Z.; Zhao, S.; Zhao, G.; and Song, X. 2024. Dual-Branch Domain Adaptation Few-Shot Learning for Hyperspectral Image Classification. *IEEE Transactions on Geoscience and Remote Sensing (TGRS)*.

Wu, Q.; Qi, J.; Zhang, D.; Zhang, H.; and Tang, J. 2024. Fine-Tuning for Few-shot Image Classification by Multimodal Prototype Regularization. *IEEE Transactions on Multimedia (TMM)*.

Wu, X.; Hong, D.; and Chanussot, J. 2021. Convolutional neural networks for multimodal remote sensing data classification. *IEEE Transactions on Geoscience and Remote Sensing (TGRS)*, 60: 1–10.

Wu, Z.; Ge, R.; Wen, M.; Liu, G.; Chen, Y.; Zhang, P.; He, X.; Hua, J.; Luo, L.; and Li, S. 2021. ELNet: Automatic classification and segmentation for esophageal lesions using convolutional neural network. *Medical Image Analysis (MIA)*, 67: 101838.

Yin, X.; Wu, J.; Yang, W.; Zhou, X.; Zhang, S.; and Zhang, T. 2024. Hierarchy-Aware Interactive Prompt Learning for Few-Shot Classification. *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*.

Yu, J.; Wang, Z.; Vasudevan, V.; Yeung, L.; Seyedhosseini, M.; and Wu, Y. 2022. CoCa: Contrastive Captioners are Image-Text Foundation Models. *Transactions on Machine Learning Research (TMLR)*.

Zhang, J.; Lei, J.; Xie, W.; Yang, G.; Li, D.; and Li, Y. 2024. Multimodal Informative ViT: Information Aggregation and Distribution for Hyperspectral and LiDAR Classification. *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*.

Zhang, R.; Guo, Z.; Zhang, W.; Li, K.; Miao, X.; Cui, B.; Qiao, Y.; Gao, P.; and Li, H. 2022. Pointclip: Point cloud understanding by clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 8552–8562.

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

Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2022a. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 16816–16825.

Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2022b. Learning to prompt for vision-language models. *International Journal of Computer Vision (IJCV)*, 130(9): 2337–2348.

Zhu, X.; Zhang, R.; He, B.; Zhou, A.; Wang, D.; Zhao, B.; and Gao, P. 2023. Not all features matter: Enhancing few-shot clip with adaptive prior refinement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2605–2615.