Structure-Adaptive Multi-View Graph Clustering for Remote Sensing Data

Renxiang Guan[1], Wenxuan Tu[2], Siwei Wang[3], Jiyuan Liu[4], Dayu Hu[1], Chang Tang[5], Yu Feng[1], Junhong Li[5], Baili Xiao[1], Xinwang Liu[1]

[1]College of Computer Science and Technology, National University of Defense Technology, Changsha, China
[2]College of Computer Science and Technology, Hainan University, Haikou, China
[3]Intelligent Game and Decision Lab, Beijing, China
[4]College of Systems Engineering, National University of Defense Technology, Changsha, China
[5]School of Computer Science, China University of Geosciences, Wuhan, China
{renxiangguan, xinwangliu}@nudt.edu.cn

Abstract

Multi-view clustering (MVC) for remote sensing data is a critical and challenging task in Earth observation. Although recent advances in graph neural network (GNN)-based MVC have shown remarkable success, most prevalent approaches have two major limitations: 1) heavily relying on a predefined yet fixed graph, which limits the clustering performance because the large number of indistinguishable background samples contained in remote sensing data would introduce noise information and increase structure heterogeneity; 2) ignoring the effect of confusing samples on cluster structure compactness, which leads to fluffy cluster structure and decreases feature discriminability. To address these issues, we propose a **S**tructure-**A**daptive **M**ulti-**V**iew **G**raph **C**lustering method named **SAMVGC** on remote sensing data, which boosts the structure homogeneity and cluster compactness by adaptively learning the graph and cluster structures, respectively. Concretely, we use the geometric structure within the feature embedding space to refine adjacency matrices. The adjacency matrices are dynamically fused with the previous ones to improve the homogeneity and stability of structure information. Additionally, the samples are separated into two categories, including the central ones (i.e., intra-cluster center samples) and the confusing ones (i.e., inter-cluster boundary samples). On this basis, to improve the cluster compactness and consistency, we deploy the contrastive learning paradigm on the central samples within views and the consistent learning paradigm on the confusing samples between views. Finally, we conduct extensive experiments on four benchmarks and achieve promising results, well demonstrating the effectiveness and superiority of the proposed method.

Introduction

Benefiting from advancements in unsupervised techniques, remote sensing data clustering tasks have significantly alleviated the time-consuming and labor-intensive challenges associated with large-scale sample labeling (Zhai et al. 2021; Liu et al. 2025; Hu et al. 2024b). Mainstream clustering methods for remote sensing data focus on learning discriminative features to group each pixel into distinct clusters without relying on labels. These techniques have been systematically applied across various fields, including land use monitoring, urban planning, and agricultural production (Guan et al. 2024a, 2022; Ma et al. 2023; Liu et al. 2024a,c).

Over the past few decades, advancements in multimedia collection capabilities have significantly increased the availability of multi-view remote sensing data, which has proven beneficial for boosting the performance by utilizing complementary knowledge from various data views (Yang, Liu, and Liu 2022). Based on the differences in view information, existing multi-view clustering (MVC) methods for remote sensing data are primarily categorized into two types: multi-feature-based and multi-source-based methods. Multi-feature-based methods enhance the model's ability to obtain richer information by extracting various features from remote sensing data, such as texture features and contour features (Chen et al. 2022). However, these views fail to capture the intuitive and rich surface features provided by multi-source sensors (Cai et al. 2024; Zhang et al. 2024). For example, optical images, including hyperspectral (HS) and multispectral (MS) images, capture detailed spectral information about land covers, while synthetic aperture radar (SAR) and digital surface model (DSM) images tend to provide terrain roughness features and height information that describes the ground elevation of different objects. Therefore, leveraging the complementarity of these data types is beneficial for remote sensing data MVC (Shahi et al. 2022). However, the high resolution and large scale of remote sensing data present a challenge in effectively modeling the spatial relationships that are beneficial to the clustering performance.

Fortunately, multi-view graph clustering (MVGC) methods have emerged as a powerful approach to capture complex dependencies and explore the intrinsic connectivity of multi-view remote sensing data (Zhang et al. 2024; Liang et al. 2024). Existing MVGC methods for remote sensing data can be broadly categorized into classic graph-based methods and graph neural network (GNN)-based methods. Classic methods primarily learn the consensus graph with different regularization terms on view-specific graphs, followed by generating clustering results through algorithms like spectral clustering (Cai et al. 2024). To learn more robust and discriminative representation by deep learning, GNN-based MVGC methods have attracted increasing research enthusiasm. With the inherent ability to propagate and aggregate information from further nodes, the latter

group could capture a superior representation of both node attributes and graph topology in multi-view data (Wang et al. 2024; Li et al. 2024c, 2023). Despite the tremendous success of previous GNN-based MVGC methods, there still exist some inherent limitations: 1) the graph structure used in the existing remote sensing data MVC methods is constructed by original data, which may contain noisy connections of heterogeneous background samples. However, these methods fix the graph structure, which would result in erroneous edges that cannot be removed, introducing noise that interferes with model training; 2) existing methods treat central samples (i.e., intra-cluster center samples) and confusing samples (i.e., inter-cluster boundary samples) equally, ignoring the effect of confusing samples on cluster compactness, which leads to a fluffy cluster structure and increases the probability of selecting the erroneous samples for subsequent tasks oriented on clustering results.

To address the above challenges, we propose a **S**tructure-**A**daptive **M**ulti-**V**iew **G**raph **C**lustering framework named SAMVGC to adaptively learn graph structure and cluster structure, aiming to boost the structure homogeneity and cluster compactness. Specifically, we first fuse all the views and generate superpixels reflecting homogeneous regions as anchors to explore consistent spatial information and significantly reduce the data scale. Secondly, we use the geometric structure within the feature embedding space to refine adjacency matrices. The adjacency matrices are dynamically fused with the previous ones, which adaptively reduces noisy edges in graphs and enhances the homogeneity of graph structures. Subsequently, to learn cluster structure adaptively, the samples are separated into two categories, including the central and the confusing samples. For the central samples, we utilize intra-view samples for contrastive learning to improve clustering compactness. Meanwhile, we encourage each confusing sample and the central samples to produce consistency in both intra-view and inter-view perspectives. In summary, the main contributions are summarized as follows:

- We propose a novel multi-view graph clustering framework termed SAMVGC, which is the first attempt to learn the graph structure and clustering structure adaptively on remote sensing data.
- We design a new superpixel-level node spatial correlation optimization module, which dynamically learns graph structure to mitigate interference from noisy connections and enhances the homogeneity of graph structures.
- We design a novel clustering structure learning method, which treats the central and confusing samples differently to improve clustering compactness and consistency.
- Extensive experiments on four widely used multi-view remote sensing datasets have validated the effectiveness and superiority of the proposed method.

Related Work

Deep Clustering for Remote Sensing Data

Deep clustering for remote sensing data aims to learn discriminative features for grouping each pixel into distinct clusters without relying on labels. Deep learning-based clustering methods include autoencoder-based (Li et al. 2024b) and contrastive learning-based methods (Guan et al. 2024a,c). However, most remote sensing data clustering methods are designed to deal with single data, which has limitations in spatial resolution, spectral resolution, temporal resolution, and other factors that may not be sufficient to meet the requirements of all applications (Yang, Liu, and Liu 2022). To overcome these limitations, multi-view remote sensing data clustering has emerged in recent years. Compared with single-view data, multi-view data can be utilized to train a more comprehensive representation and achieve desirable results for clustering tasks. For example, Guan *el al.* (Guan et al. 2024b) propose a multi-view contrastive learning network to promote consistency of features between views. Liu et al. (Luo et al. 2024) design a self-supervised joint method to constrain the consensus representation of different views. Cai *et al.* (Cai et al. 2023) design the Transformer model combined with a prototype contrastive learning algorithm to extract deep features of multi-view remote sensing data. However, there is no doubt that some confusing samples would cause the position of the prototype to shift, which negatively impacts the compactness of the samples within the view. In contrast, we use central samples to compute the prototype to ensure its high quality. In addition, we encourage each confusing sample and the central samples to produce consistent similarity in both intra-view and inter-view perspectives to improve consistency.

Multi-View Graph Clustering

To better explore the intrinsic connectivity of multi-view data (Yuan et al. 2024a,b), graph-based methods are used for MVC (Hu et al. 2024a; Qu et al. 2024; Xie et al. 2024; Li et al. 2024d; Shen and Tang 2024; Shen et al. 2024). Early methods primarily learn the consensus graph with different regularization terms on view-specific graphs, followed by generating clustering results through algorithms like spectral clustering. Liu *et al.* (Liu et al. 2024b) construct a consistent anchor graph that captures inter-view commonality and filters out view-specific noise. Li *et al.* construct an essential similarity graph in a spectral embedding space to capture the global consistent information among multiple views. The former groups mainly focus on latent space learning by exploring pairwise local structures, but they are limited in their ability to leverage the global information embedded in data. Recently, GNN-based deep MVC methods have attracted increasing research enthusiasm, aiming to learn more discriminative representations. Xiao *et al.* (Xiao et al. 2023) propose a dual fusion-propagation GNN to capture multiple information among different views and then utilize them to refine the results of MVC. Chen *et al.* (Chen et al. 2024) incorporate the graph autoencoder with autoencoder into a unified feature learning framework that can effectively extract diverse representations to learn deep useful features. Nevertheless, these methods cannot be directly applied to complex remote sensing data, as the graph in each view remains fixed, making it difficult to correct erroneous edge connections in the original graph.

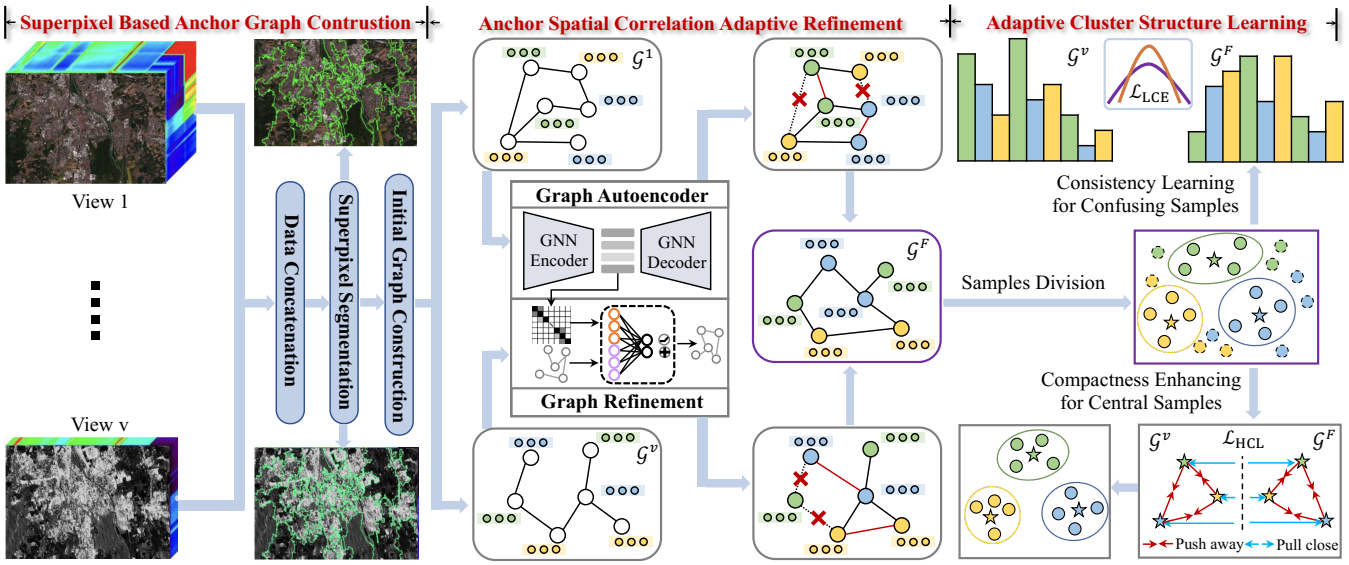

Figure 1: Illustration of the proposed SAMVGC. Initially, we select anchors based on the superpixel segmentation results and extract deep features of the anchors via graph autoencoders. Next, the geometric structures within the feature embedding space are utilized to refine adjacency matrices. Finally, the samples are categorized into central and confusing samples, enabling differential processing to enhance compactness and consistency.

Methods

In this part, we present the proposed SAMVGC in detail. An overview of SAMVGC is shown in Fig. 1.

Problem Definition

Given a set of multi-view remote sensing data $\mathcal{X} = \{\mathbf{X}^{(v)}\}_{v=1}^{V}$, where V is the number of views, and each $\mathbf{X}^{(v)} \in \mathbb{R}^{W\times H\times D^{(v)}}$ represents a remote sensing data consists of $N = W \times H$ pixels and $D^{(v)}$ feature dimensions. We aim to learn a mapping function $f : \mathbf{X}^{(v)} \mapsto \mathbf{Z}^{(v)} \in \mathbb{R}^{N\times d^{(v)}}$ for grouping each pixel into K distinct clusters without relying on labels, where $d^{(v)}$ is the number of dimension of the latent features.

Superpixel-based Anchor Graph Contrustion

Anchor Selection Based on Superpixel Segmentation The vast amount of data in remote sensing data demands significant computational resources, making the clustering task challenging. Existing MVC methods for handling large-scale datasets typically rely on random anchors or K-means methods, but the unlabeled nature of clustering often leads to inconsistent label assignments. To address this issue, we propose an anchor construction method based on superpixel segmentation. By leveraging the spatial homogeneity inherent in remote sensing data, we reduce computational complexity by using superpixels as anchors. Specifically, we fuse the features of all views, and then apply the efficient SLIC segmentation method $f_{\text{SLIC}}(\cdot)$ (Achanta et al. 2012) to generate superpixels with adaptive shape and size denoted as:

$$\hat{\mathbf{X}}^{(1)}, \hat{\mathbf{X}}^{(2)}, \ldots, \hat{\mathbf{X}}^{(v)} \leftarrow f_{\text{SLIC}}\left(\mathbf{X}^{(1)} \| \mathbf{X}^{(2)} \| \ldots \| \mathbf{X}^{(v)}\right), \quad (1)$$

where $\hat{\mathbf{X}}^{(v)} \in \mathbb{R}^{M\times D^{(v)}} = \bigcup_{i=1}^{M} \hat{\mathbf{x}}_i^{(v)}$ where $M \ll N$ is the number of superpixels, $\hat{\mathbf{x}}_i^{(v)}$ is the i -th superpixel and $\hat{\mathbf{x}}_i^{(v)} \cap \hat{\mathbf{x}}_j^{(v)} = \varnothing$ when $i \neq j$. Compared with other strategies, it is promising to select anchors with spatial information.

Initially Graph Construction GNN requires an adjacency matrix as input, so we capture the nearest samples of each anchor as neighbors and construct edges for them, with element $\mathbf{A}_{ij}^{(v)} \in \mathbb{R}^{M\times M}$ defined as follows:

$$\mathbf{A}_{ij}^{(v)} = \begin{cases} 1, & \text{if } \hat{\mathbf{x}}_j^{(v)} \in \Gamma(\hat{\mathbf{x}}_i^{(v)}) \\ 0, & \text{otherwise} \end{cases}, \quad (2)$$

where $\Gamma(\cdot)$ denotes the top k nearest samples. The graph is formulated as $\mathcal{G}^{(v)} = \{\mathbf{X}^{(v)}, \mathbf{A}^{(v)}\}$ for the v- th view.

Anchor Spatial Correlation Adaptive Refinement

Most existing methods assume $\mathbf{A}^{(v)}$ is fixed, adversely impacting the learning of representations in GNN. However, the graph structure is constructed by original data, which may contain noisy connections. Thus, we construct a local neighborhood graph to approximate the manifold construction in the embedding space, preserving the local geometry of neighborhoods. Specifically, we extract the feature $\mathbf{Z}^{(v)}$ and similarity matrix $\mathbf{S}_z^{(v)}$:

$$\mathbf{S}_z^{(v)} = \frac{\left\langle (\mathbf{Z}^{(v)})^{\top}, \mathbf{Z}^{(v)} \right\rangle}{\left\| (\mathbf{Z}^{(v)})^{\top} \right\| \cdot \left\| \mathbf{Z}^{(v)} \right\|}, \quad (3)$$

where $\langle \cdot, \cdot \rangle$ refers to an inner product operation. Based on $\mathbf{S}_z^{(v)}$, we construct graph $\mathbf{G}^{(v)}$ with its element:

$$g_{i,j}^{(v)} = \begin{cases} s_{i,j}^{(v)} & \text{if } s_{i,j} \in \Gamma\left(\mathbf{S}_i^{(v)}\right) \\ 0 & \text{otherwise} \end{cases}, \quad (4)$$

$$\text{s.t. } \mathbf{S}^{(v)} = \mathbf{S}_z^{(v)} - \text{diag}\left(\mathbf{S}_z^{(v)}\right),$$

where $s_{i,j}^{(v)}$ is the (i,j)-th element of $\mathbf{S}^{(v)}$, $\Gamma\left(\mathbf{S}_i^{(v)}\right)$ is the the top k value of $\mathbf{S}^{(v)}$ in the i-th row, diag $(\mathbf{S}_z^{(v)})$ is a diagonal matrix. Then, we apply the self-loop operation (Topping et al. 2022) to normalize $\mathbf{G}^{(v)}$, and then symmetrize it:

$$g_{i,j}^{(v)} = g_{j,i}^{(v)} = \begin{cases} \max\left\{g_{i,j}^{(v)}, g_{j,i}^{(v)}\right\} & \text{if } i \neq j \\ 1 & \text{if } i = j \end{cases}, \quad (5)$$

where $\max\left\{g_{i,j}^{(v)}, g_{j,i}^{(v)}\right\}$ selects the greater value, and we can obtain the adjacency matrix $\mathbf{A}_z^{(v)} = \mathbf{D}_z^{-1}\mathbf{G}^{(v)}$ with its degree matrix $\mathbf{D}_z^{(v)}$. Then, we build a multilayer perceptron layer parametrized by weight matrices $\mathbf{W}_{\mathbf{A}}^{(v)} \in \mathbb{R}^{2M\times 2}$ to capture the relationship among $\mathbf{A}^{(v)}$ and $\mathbf{A}_z^{(v)}$ with a normalization operation, which can be formulated as:

$$\mathbf{U}^{(v)} = \left[\mathbf{u}_z^{(v)} \| \mathbf{u}^{(v)}\right] = \ell_2\left(\text{Softmax}\left(\text{LReLU}\left(\left[\mathbf{A}_{\mathbf{z}}^{(v)} \| \mathbf{A}^{(v)}\right]\mathbf{W}_{\mathbf{A}}^{(v)}\right)\right)\right), \quad (6)$$

where $\mathbf{u}_z^{(v)}$ and $\mathbf{u}^{(v)}$ are the weight vectors with entries being greater than 0 for measuring the importance of $\mathbf{A}_z^{(v)}$ and $\mathbf{A}^{(v)}$, respectively. Meanwhile, we fuse them as below:

$$\mathbf{A}^{(v)} = (\mathbf{u}_z^{(v)}\mathbf{1}) \odot \mathbf{A}_z^{(v)} + (\mathbf{u}^{(v)}\mathbf{1}) \odot \mathbf{A}^{(v)}, \quad (7)$$

where $\odot$ is the Hadamard product. Finally, we iteratively conduct graph refinement to enhance the adjacency matrix, adaptively optimizing the spatial correlation of anchors. The optimization is performed every i_p epoch.

Graph Autoencoder

To fully explore the topology of the constructed graph and aggregate features from neighboring nodes, we deploy graph encoders to capture features. Concretely, we utilize multilayer graph convolutional networks (GCN) as graph encoders and extract deep features. The procedure could be defined as follows:

$$\mathbf{Z}_{(l)}^{(v)} = \sigma(\widetilde{\mathbf{A}}^{(v)}\mathbf{Z}_{(l-1)}^{(v)}\mathbf{W}_{(l)}^{(v)} + \mathbf{b}_{(l)}^{(v)}), \quad (8)$$

where $\widetilde{\mathbf{A}}^{(v)} \in \mathbb{R}^{M\times M}$ is the symmetric normalized Laplacian matrix derived from $\mathbf{A}^{(v)}$. Here, $\mathbf{W}_{(l)}^{(v)}$ and $\mathbf{b}_{(l)}^{(v)}$ refer to the learnable weight matrix and bias term in the l-th layer, respectively. $\mathbf{Z}_{(l)}^{(v)}$ denotes the hidden feature matrix of the l-th layer in the v-th view. For each view, we define the input feature in the first layer as $\mathbf{Z}_{(0)}^{(v)} = \hat{\mathbf{X}}^{(v)}$.

Within the graph auto-encoder module, we employ multilayer GCN as the decoder $\mathcal{D}(\cdot)$ to reconstruct the original features. The process of reconstruction is articulated as:

$$\bar{\mathbf{X}}^{(v)} = \mathcal{D}(\widetilde{\mathbf{A}}^{(v)}, \mathbf{Z}^{(v)} \mid \Theta_{\mathcal{D}}^{(v)}), \quad (9)$$

where $\bar{\mathbf{X}}^{(v)}$ is the reconstruction features, $\Theta_{\mathcal{D}}^{(v)}$ denotes the learnable parameters of graph encoders. Therefore, the superpixel-level reconstruction loss function could be defined as:

$$\mathcal{L}_{\text{REC}} = \sum_{v=1}^{(v)} \left\| \bar{\mathbf{X}}^{(v)} - \hat{\mathbf{X}}^{(v)} \right\|_F^2, \quad (10)$$

where $\|\cdot\|_F$ represents the Frobenius norm.

Algorithm 1: The learning procedure of SAMVGC

Require: Multi-view remote sensing data $\mathcal{X}$, the number of neighbors k and superpixels M, the rate of central samples α, and the training epochs I
Ensure: Clustering results Y

```
/* Pre-processing stage */
Initialize the parameters of the network
Contruct anchors by using superpixel segmentation
Construct initial adjacency matrix A^(v) by Eq. (2)
/* Model Training Stage */
for i = 1 to I do
    Obtain features Z^(v), Z^F by Eq. (8) and Eq. (11)
    Obtain C and Y from Z^F with K-means by Eq. (12)
    Divide central and confusing samples by Eq. (13)
    Calculate reconstruction loss L_REC by Eq. (10)
    Calculate contrastive learning loss L_HCL by Eq. (15)
    Calculate consistency learning loss L_LCE by Eq. (17)
    Update parameters of the network by miniziming Eq. (18)
    if i % 10 = 0 then
        Update the adjacency matrix A^(v) by Eq. (3)-Eq. (7)
    end if
end for
return Clutering result
```

Cluster Structure Adaptive Learning

To adaptively learn compact and consistenct cluster structures, we design a compactness and consistency learning module, which classifies the samples into central and confusing samples. For central samples we use a prototype-based contrastive learning method to enhance compactness within clusters and for confusing samples we use consistency learning between views to enhance consistency.

Compactness Enhancing for Central Samples To be specific, we firstly fuse the views of the node features as follows:

$$\mathbf{Z}^F = \mathbf{W}_1 \odot \mathbf{Z}^{(1)} + \mathbf{W}_2 \odot \mathbf{Z}^{(2)} + \cdots + \mathbf{W}_v \odot \mathbf{Z}^{(v)}, \quad (11)$$

where $\mathbf{W}_v \in \mathbb{R}^{M\times\hat{d}}$ are trainable weight matrices to adaptively control the importance of features. And we employ the K-means algorithm (Bauckhage 2015) to ascertain the clustering centroid of the fused feature:

$$\min_{\mathbf{P},\mathbf{C}} \|\mathbf{Z}^F - \mathbf{P}\mathbf{C}\|^2, \quad \text{s.t. } \mathbf{P}\mathbf{1} = \mathbf{1}, \mathbf{P} \geq \mathbf{0}, \quad (12)$$

where $\mathbf{C} = \mathbf{c}_1, \ldots, \mathbf{c}_K \in \mathbb{R}^{K\times\hat{d}}$ denotes the center matrix of clustering and $\mathbf{P} \in \mathbb{R}^{M\times K}$ is the cluster indicator matrix. To select central samples and extract more reliable clustering information, we calculate the distance between samples and their clustering centers:

$$d_i = \min_{1\leq k\leq K, 1\leq i\leq p(k)} \|\mathbf{z}_i^F - \mathbf{c}_k\|_2^2, \quad (13)$$

where d_i denotes the distance between the i-th superpixel embedding and the corresponding clustering center. The top α samples are selected as central samples and the remaining is the confusing sample. We then compute the prototype of each cluster:

$$\boldsymbol{\mu}_k^{(v)} = \frac{1}{M_k^h}\sum\nolimits_{i\in M_k^h} \mathbf{z}_i^{(v)}, \quad \boldsymbol{\mu}_k^F = \frac{1}{M_k^h}\sum\nolimits_{i\in M_k^h} \mathbf{z}_i^F, \quad (14)$$

where M_k^h is the number of central samples in the k-th cluster, $\boldsymbol{\mu}_k$ represent the cluster prototypes. Clustering prototypes are more accurate after removing confusing samples. We obtain a more compact cluster structure by using contrastive learning at the clustering prototype level:

$$\mathcal{L}_{\mathrm{HCL}} = \frac{1}{K}\sum_{k=1}^{K}\sum_{v=1}^{(v)} -log\frac{e^{\theta(\boldsymbol{\mu}_k^{(v)}\cdot\boldsymbol{\mu}_k^F)}}{e^{\theta(\boldsymbol{\mu}_k^{(v)}\cdot\boldsymbol{\mu}_k^F)}+\sum_{j=1}^{K}e^{\theta(\boldsymbol{\mu}_k^{(v)}\cdot\boldsymbol{\mu}_j^F)}}, \tag{15}$$

where θ is the cosine similarity. By optimizing $\mathcal{L}_{\mathrm{HCL}}$, we can make the samples near the cluster centers more compact.

Consistency Learning for Confusing Samples While confusing samples have a negative impact on contrastive learning (Yang et al. 2023), we can develop consistency learning to facilitate the exchange of knowledge between the intra- and inter-view perspectives. Specifically, we calculate the similarities between each central sample and these confusing samples. Mathematically, for the i-th node representations from view v and fused view, the similarity scores of the m-th confusing samples are

$$p_i^{(v)} = \frac{e^{\theta(\mathbf{z}_i^{(v)},\mathbf{z}_j^{(v)})}}{\sum_{j=1}^{M^h}e^{\theta(\mathbf{z}_i^{(v)},\mathbf{z}_j^{(v)})}}, \quad q_i^F = \frac{e^{\theta(\mathbf{z}_i^F,\mathbf{z}_j^F)}}{\sum_{j=1}^{M^h}e^{\theta(\mathbf{z}_i^F,\mathbf{z}_j^F)}}. \tag{16}$$

With the intra- and inter-view similarity distributions $\mathbf{p}_i^{(v)}$ and $\mathbf{q}_i^F$, we encourage their consistency to facilitate the knowledge transfer and mutually enhance the feature semantics. Formally, we define the consistency learning loss as

$$\mathcal{L}_{\mathrm{LCE}} = \frac{1}{2M^l}\sum_{v=1}^{(v)}\sum_{i=1}^{M^l}\left(\mathrm{KL}(\mathbf{p}_i^{(v)}||\mathbf{q}_i^F)+\mathrm{KL}(\mathbf{q}_i^F||\mathbf{p}_i^{(v)})\right), \tag{17}$$

where M^l is the number of confusing samples in each view and $\mathrm{KL}(\cdot||\cdot)$ is the Kullback-Leibler (KL) divergence. We thus combine intra- and inter-view information to improve the clustering consistency and compactness.

The Overall Loss Function

In summary, we introduce a novel structure-adaptive MVGC framework for remote sensing data. During the training stage, the graph encoders, the compactness enhancing part, and the consistency learning part are jointly optimized according to the objective function as follows,

$$\mathcal{L} = \mathcal{L}_{\mathrm{REC}} + \lambda_1 * \mathcal{L}_{\mathrm{HCL}} + \lambda_2 * \mathcal{L}_{\mathrm{LCE}}. \tag{18}$$

In the training phase, we minimize $\mathcal{L}$ to optimize the proposed SAMVGC. Finally, we take the K-means on $\mathbf{Z}^F$ to obtain the clustering results for all samples. The detailed learning procedure of SAMVGC is shown in Algorithm 1.

Experiments

Experiment Setup

Datasets Four open source datasets of remote sensing data are used in the experiments, including the Trento, MUULF, Augsburg, and MDAS datasets. The Trento and MUULF contain two views: HS and LiDAR. The Augsburg dataset contains three views: HS, SAR, and DSM. The MDAS dataset is composed of sub-scenes of the complete dataset, including four views: HS, MS, SAR, DSM. Table 1 briefly introduces the basic information of the datasets.

Datasets	Samples	Clusters	Views	Features
Trento	30,214	6	2	63/2
MUUFL	53,687	11	2	64/2
Augsburg	78,294	7	3	180/4/1
MDAS	88,026	14	4	242/2/1/12

Table 1: Statistical details of the datasets.

Baseline Methods To demonstrate the effectiveness of our proposed SAMVGC, we compare it with eleven state-of-the-art methods, including classic MVC methods: FMVACC (Wang et al. 2022), AWMVC (Wan et al. 2023), AMKSC (Cai et al. 2024) and deep MVC methods MFLVC (Xu et al. 2022), MDC (Shahi et al. 2022), CVCL (Chen et al. 2023), SDMVC (Xu et al. 2023), GCFAGG (Yan et al. 2023), TMPCC (Cai et al. 2023), MDFL (Li et al. 2024a), CMSCGC (Guan et al. 2024b).

Implement Details To ensure an equitable comparison between the proposed SAMVGC and these baselines, we compute the average results from ten iterative runs under identical experimental conditions, utilizing a 24GB RTX 3090 GPU and 64GB of RAM. We use a three-layer GCN as the encoder and decoder of the graph, with the number of hidden layers and output layers being 128, 256 and 512 respectively. We set λ_1 and λ_2 to 100 and 500 respectively making the three losses on the same scale. And i_p = 10 is used in our model. For the baselines, we use the optimal parameters reported in the papers to derive the final results. To clearly demonstrate the performance, we employ five evaluation metrics: accuracy (ACC), Kappa, normalized mutual information (NMI), adjusted rand index (ARI) and Purity (PUR) (Zhou et al. 2023, 2024; Tu et al. 2024, 2021; Xiao et al. 2024a,b; Yu et al. 2025).

Quantitative Results

As reported in Table 2, we present the quantitative results of the proposed SAMVGC in comparison with other competitive MVC baselines. From the table, we draw the following key observations: 1) MVC algorithms designed for remote sensing data are more accurate than generic MVC algorithms averaged over four datasets. This is due to the fact that remote sensing data need to take into account special but important spatial information, and illustrates the importance of the remote sensing MVC task; 2) taking the results of several algorithms specifically designed for remote sensing MVC tasks on the Trento dataset as an example, SAMVGC significantly outperforms MDC, TMPCC and MDFL by 11.65%, 6.47%, and 9.03% ACC, respectively. These results provides substantial evidence for the superiority of our approach and verify that SAMVGC can effectively learn clustering-friendly features in the remote sensing MVC

Datasets	Metric	MFLVC CVPR22	FMVACC NIPS22	MDC JSTARS22	CVCL ICCV23	SDMVC TKDE23	GCFAgg CVPR23	AWMVC AAAI23	TMPCC Inf.Sci23	MDFL AAAI24	CMSCGC TGRS24	AMKSC TNNLS24	SAMVGC Ours
Trento	ACC	51.15±2.26	75.03±0.11	83.39±0.14	58.36±5.73	63.46±1.14	56.05±4.64	58.74±0.76	88.57±4.06	86.01±2.17	88.74±2.14	93.90±1.97	**95.04±0.36**
	Kappa	30.23±2.17	25.96±0.02	78.54±0.18	48.23±6.34	48.34±1.26	42.25±4.99	18.30±0.15	87.09±6.65	69.45±1.19	88.59±2.16	91.85±2.58	**93.59±0.35**
	NMI	40.49±1.96	55.81±0.37	76.48±0.21	53.01±5.61	39.13±1.04	47.78±4.78	42.07±0.97	82.26±4.91	59.90±1.77	85.05±2.88	88.21±1.20	**93.36±0.49**
	ARI	28.24±1.18	60.86±0.17	71.20±0.17	40.78±6.13	38.22±0.11	32.92±5.95	37.31±0.59	90.24±3.33	60.23±1.55	82.45±2.19	92.83±1.94	**93.75±0.66**
	PUR	51.31±1.13	77.76±0.02	84.46±0.16	73.12±3.47	63.46±0.54	57.36±4.00	67.42±0.09	90.20±2.52	76.01±2.24	90.05±1.92	93.90±1.75	**95.52±0.37**
MUUFL	ACC	39.66±2.13	41.46±0.13	42.74±0.76	34.28±2.26	39.77±1.04	40.44±1.99	45.52±0.69	48.23±4.21	36.74±2.00	43.77±1.69	46.48±4.17	**50.50±1.50**
	Kappa	31.62±3.74	17.30±0.28	34.48±0.49	20.76±3.40	31.31±1.06	31.16±1.21	11.08±0.21	34.10±3.63	23.18±1.24	34.89±2.23	38.81±1.80	**42.48±1.04**
	NMI	32.61±4.63	46.39±0.37	43.90±0.26	25.13±2.84	44.22±0.59	33.03±1.36	38.42±0.52	40.29±4.28	27.37±1.21	43.50±1.82	42.62±3.53	**46.70±1.37**
	ARI	20.51±4.16	31.13±0.28	21.63±0.39	10.43±1.99	23.48±0.78	14.84±1.51	25.68±0.85	37.74±1.75	24.14±1.74	28.34±1.21	29.01±1.55	**43.72±0.98**
	PUR	53.47±1.80	69.14±0.07	62.98±0.20	48.75±1.25	65.64±1.26	60.10±1.85	65.02±0.19	70.70±2.80	47.48±1.01	66.78±2.89	68.70±3.76	**75.19±1.07**
Augsburg	ACC	51.30±2.60	42.47±0.05	52.03±1.42	41.98±4.48	45.86±0.19	41.94±0.30	36.08±0.58	52.70±2.61	41.54±1.22	50.60±2.54	37.96±2.15	**67.80±0.20**
	Kappa	34.45±3.09	8.92±0.04	25.60±2.16	17.13±9.88	16.92±1.29	28.57±1.25	15.18±0.27	43.27±1.82	14.36±1.51	35.37±2.14	23.93±3.36	**47.99±0.29**
	NMI	30.31±2.80	27.82±0.04	22.28±1.69	15.95±8.16	16.29±2.19	23.89±1.27	29.47±0.11	42.34±2.88	14.53±2.16	43.35±1.54	28.70±0.64	**55.21±0.24**
	ARI	21.43±2.62	22.61±0.06	19.72±2.18	24.35±6.07	26.12±0.12	16.31±0.46	17.21±0.82	48.07±1.54	6.64±1.69	44.33±1.66	17.77±1.11	**48.18±0.27**
	PUR	59.66±2.59	65.19±0.06	53.56±1.43	42.82±5.67	45.92±1.20	58.52±3.39	65.95±0.27	75.59±1.12	52.18±1.66	69.67±1.69	57.78±1.09	**78.88±0.99**
MDSA	ACC	40.56±3.84	25.68±0.53	21.95±0.13	34.69±4.22	29.73±0.12	23.70±0.55	21.31±0.36	23.97±0.81	25.03±1.69	40.24±1.43	27.85±1.83	**47.59±0.70**
	Kappa	13.21±6.92	1.46±0.02	12.50±0.27	4.81±2.87	11.74±0.04	14.17±0.11	11.28±2.68	6.71±0.54	11.96±2.04	20.99±2.10	18.62±1.43	**28.85±0.64**
	NMI	9.49±4.73	14.81±0.00	17.05±0.58	6.18±2.16	9.55±0.06	16.96±1.16	15.94±0.05	14.51±1.04	12.93±3.35	25.95±1.22	24.23±0.38	**35.09±0.87**
	ARI	7.03±5.01	2.93±0.01	4.59±0.22	1.33±0.68	2.89±0.19	5.05±0.57	4.00±0.15	19.60±1.33	5.97±3.10	25.95±1.77	12.47±0.97	**35.59±0.57**
	PUR	51.76±1.07	49.46±0.00	51.93±0.19	49.45±0.01	49.69±0.05	53.87±1.42	52.15±0.02	54.13±0.81	32.78±2.83	60.84±0.44	59.01±0.05	**66.30±0.55**

Table 2: The clustering performance on four benchmark datasets (mean%±std%). The best and the suboptimal results in all the methods are highlighted with **bold** and underline, respectively.

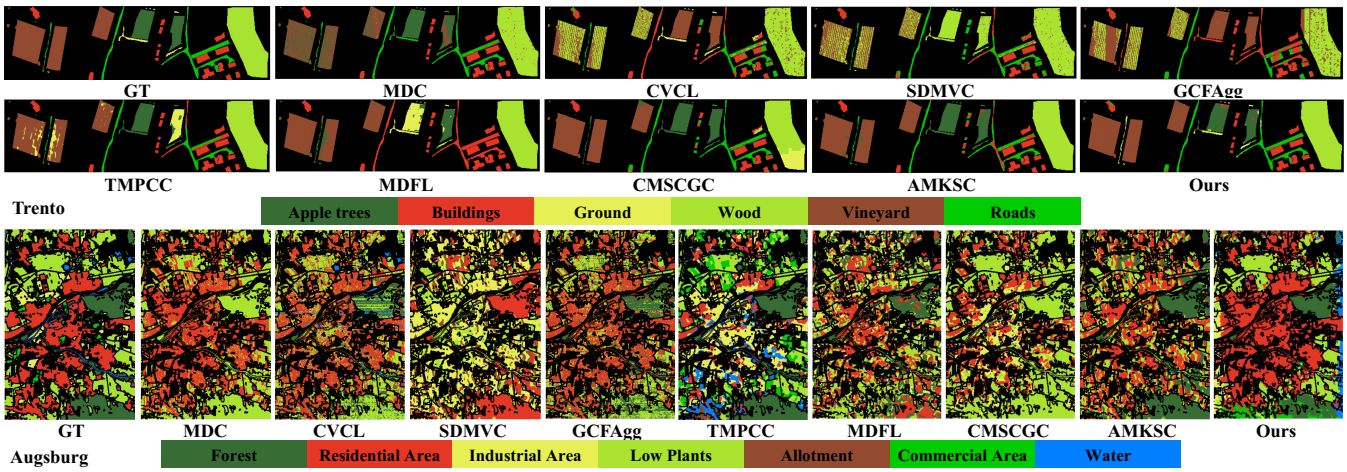

Figure 2: Clustering maps on the Trento and Augsburg datasets. GT represents the ground truth.

tasks; 3) Compared with the two graph-based algorithms CMSCGC and AMKSC, we surpass them by 6.3%/1.14%, 6.73%/4.02%, 17.2%/29.84%, and 7.35%/36.41% on four datasets. It can be seen that the performance of SAMVGC on the four datasets, especially datasets Augsburg and MDSA with complex scenes, far exceeds them, which illustrates the effectiveness of our proposed graph redefinition module and consistency compactness module.

Visualization

To visualize the advantages of SAMVGC, we compare it in terms of both clustering maps and T-SNE visualization.

Clustering Maps To provide a more intuitive comparison among different methods, Fig. 2 illustrate the clustering results on the Trento and Augsburg datasets. It can be observed that the common clustering methods have the noisiest clustering maps with massive misclustering pixels. In comparison, the remote sensing MVC methods produce smoother clustering maps by effectively utilizing spatial information. Further, the proposed SAMVGC obtains the smoothest clustering maps with least noise, verifying that SAMVGC can effectively learn clustering-friendly features.

T-SNE Visualization We also plot the distributions of the learned embeddings reduced by t-SNE on the Trento dataset in Fig. 3 to visually verify the validity of our proposed SAMVGC. We can see that SAMVGC exhibits better separability among different clusters, with improved aggregation and diversity within the same cluster and a larger gap between different clusters, showcasing its ability to learn more discriminative representations and effective cluster assignments compared with competitive methods.

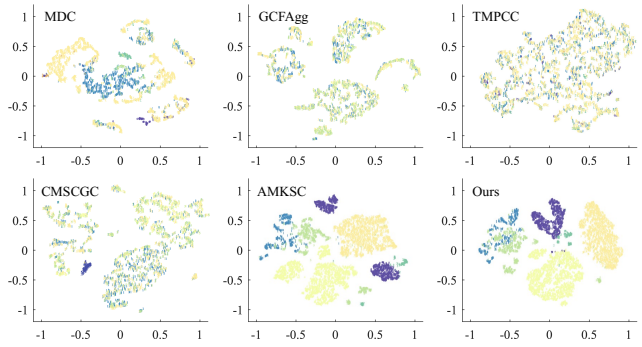

Figure 3: The t-SNE on the Trento dataset.

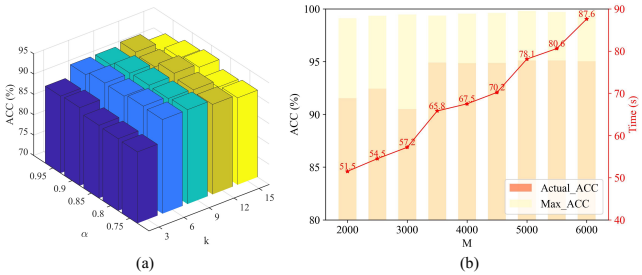

Figure 4: The model sensitivity analysis on Trento dataset. (a) The two hyper-parameters α and k; (b) impact of M on performance and running time.

Ablation Study

In this section, we conduct an ablation study to analyze the impact of various components within the proposed SAMVGC method. By selectively removing specific elements, we aim to evaluate their individual contributions and understand their influence on the performance. Concretely, "w/o GE", "w/o $\mathcal{L}_{\mathrm{HCL}}$" and "w/o $\mathcal{L}_{\mathrm{LCE}}$" denotes the two SAMVGC variants with the spatial correlation adaptive refinement module, the central samples contrastive learning loss and the confusing samples consistency learning loss being removed, respectively. As seen in Table 3, we can find that: 1) when compared to "w/o GE", SAMVGC achieves 2.15%, 3.92%, 2.69% and 2.01% ACC gains on four datasets, indicating the superiority of optimizing graph structure learning, which can help models learn more robust graph structures; 2) compared to "w/o $\mathcal{L}_{\mathrm{HCL}}$" and "w/o $\mathcal{L}_{\mathrm{LCE}}$", SAMVGC produces ACC performance gains of 7.97% / 1.88%, 7.91% / 4.33%, 3.46% / 0.91%, and 2.61% / 1.35%. These results illustrate that the joint learning of the two modules is better than that of each module alone.

Hyper-parameter Analysis

In this section, we study the sensitivity of hyper-parameters: the number of nearest neighbors k, the rate of central samples a, and the number of anchors M on the Trento dataset. As illustrated in Fig. 4(a), the performance exhibits a gradual increase when the value of k ranges from 3 to 15, followed by a slight decline. This indicates that a certain number of neighbors can increase the model's ability to capture spatial information, but too many neighbors can introduce a certain amount of noisy information leading to performance degradation. Similarly, when α is small, a large number of confusing samples affects the compactness of clustering. However, a very large α would incur excessive time and space costs, so we default to choosing 0.85 in our experiments to strike a balance. The impact of M on maximum theoretical accuracies of superpixel segmentation are also demonstrated in Fig. 4(b), which is calculated as the ratio of the number of dominant pixels within the superpixel to the total number of pixels within the superpixel. The results show that the maximum theoretical accuracies exceed 99% across the board. Considering both runtime efficiency and actual accuracy, we choose the values of M as 3500.

Datasets	Variants	ACC	Kappa	NMI	ARI	PUR
Trento	w/o GE	92.89	91.14	90.48	85.60	93.41
	w/o $\mathcal{L}_{\mathrm{HCL}}$	87.07	86.30	82.71	80.99	88.18
	w/o $\mathcal{L}_{\mathrm{LCE}}$	93.16	91.51	90.98	86.04	93.78
	Ours	**95.04**	**93.59**	**93.36**	**93.75**	**95.52**
MUULF	w/o GE	46.58	40.15	44.75	40.93	73.95
	w/o $\mathcal{L}_{\mathrm{HCL}}$	42.59	35.65	40.96	36.29	70.42
	w/o $\mathcal{L}_{\mathrm{LCE}}$	46.17	39.54	44.29	39.20	74.54
	Ours	**50.50**	**42.48**	**46.70**	**43.72**	**75.19**
Augsburg	w/o GE	65.11	47.14	52.71	45.19	76.92
	w/o $\mathcal{L}_{\mathrm{HCL}}$	64.34	45.36	51.79	46.62	75.67
	w/o $\mathcal{L}_{\mathrm{LCE}}$	66.89	46.72	54.22	45.38	77.47
	Ours	**67.80**	**47.99**	**55.21**	**48.18**	**78.88**
MDSA	w/o GE	45.58	26.48	34.21	34.25	66.14
	w/o $\mathcal{L}_{\mathrm{HCL}}$	44.98	25.84	33.35	35.25	65.66
	w/o $\mathcal{L}_{\mathrm{LCE}}$	46.24	28.22	34.28	35.14	**66.90**
	Ours	**47.59**	**28.85**	**35.09**	**35.59**	66.30

Table 3: Ablation results of different variants.

Conclusion

In this study, we propose a novel multi-view graph clustering framework named SAMVGC for remote sensing data. Specifically, SAMVGC uses superpixel segmentation results to select anchor points and uses graph autoencoder to capture deep features. Then, we utilize the geometric structure within the feature embedding space to refine adjacency matrices, which reduces noisy edges in graphs and enhances the homogeneity of graph structures. Subsequently, we divide the samples into central and confusing samples for differential treatment. For central samples, we utilize intra-view samples for contrastive learning to improve clustering compactness. For confusing samples, we encourage each sample and the central samples to produce consistent similarity in both intra-view and inter-view perspectives. Experiments show that SAMVGC can outperform other clustering methods on four multi-view remote sensing datasets.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (Grant No. 62325604, 62276271,

62306324, 62376279, U24A20333), the Science and Technology Innovation Program of Hunan Province (Grant No. 2024RC3128), National University of Defense Technology Research Foundation (No. ZK24-30), and the Natural Science Foundation of Hainan University (Grant No. XJ2400009401). Corresponding authors: Wenxuan Tu and Xinwang Liu.

References

Achanta, R.; Shaji, A.; Smith, K.; Lucchi, A.; Fua, P.; and Süsstrunk, S. 2012. SLIC superpixels compared to state-of-the-art superpixel methods. *IEEE transactions on pattern analysis and machine intelligence*, 34(11): 2274–2282.

Bauckhage, C. 2015. K-means clustering is matrix factorization. *arXiv preprint arXiv:1512.07548*.

Cai, Y.; Zhang, Z.; Ghamisi, P.; Rasti, B.; Liu, X.; and Cai, Z. 2023. Transformer-based contrastive prototypical clustering for multimodal remote sensing data. *Information Sciences*, 649: 119655.

Cai, Y.; Zhang, Z.; Liu, X.; Ding, Y.; Li, F.; and Tan, J. 2024. Learning Unified Anchor Graph for Joint Clustering of Hyperspectral and LiDAR Data. *IEEE Transactions on Neural Networks and Learning Systems*, 1–14.

Chen, J.; Mao, H.; Woo, W. L.; and Peng, X. 2023. Deep multiview clustering by contrasting cluster assignments. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 16752–16761.

Chen, R.; Tang, Y.; Cai, X.; Yuan, X.; and Feng, W. 2024. Graph Structure Aware Contrastive Multi-View Clustering. *IEEE Transactions on Big Data*, 10(3): 260–274.

Chen, Z.; Zhang, C.; Mu, T.; and He, Y. 2022. Tensorial Multiview Subspace Clustering for Polarimetric Hyperspectral Images. *IEEE Transactions on Geoscience and Remote Sensing*, 60: 1–13.

Guan, R.; Li, Z.; Li, T.; Li, X.; Yang, J.; and Chen, W. 2022. Classification of heterogeneous mining areas based on rescapsnet and gaofen-5 imagery. *Remote Sensing*, 14(13): 3216.

Guan, R.; Li, Z.; Li, X.; and Tang, C. 2024a. Pixel-superpixel contrastive learning and pseudo-label correction for hyperspectral image clustering. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6795–6799. IEEE.

Guan, R.; Li, Z.; Tu, W.; Wang, J.; Liu, Y.; Li, X.; Tang, C.; and Feng, R. 2024b. Contrastive Multiview Subspace Clustering of Hyperspectral Images Based on Graph Convolutional Networks. *IEEE Transactions on Geoscience and Remote Sensing*, 62: 1–14.

Guan, R.; Tu, W.; Li, Z.; Yu, H.; Hu, D.; Chen, Y.; Tang, C.; Yuan, Q.; and Liu, X. 2024c. Spatial-Spectral Graph Contrastive Clustering with Hard Sample Mining for Hyperspectral Images. *IEEE Transactions on Geoscience and Remote Sensing*, 1–16.

Hu, D.; Dong, Z.; Liang, K.; Yu, H.; Wang, S.; and Liu, X. 2024a. High-order Topology for Deep Single-Cell Multi-view Fuzzy Clustering. *IEEE Transactions on Fuzzy Systems*, 32(8): 4448–4459.

Hu, D.; Guan, R.; Liang, K.; Yu, H.; Quan, H.; Zhao, Y.; Liu, X.; and He, K. 2024b. scEGG: an exogenous gene-guided clustering method for single-cell transcriptomic data. *Briefings in Bioinformatics*, 25(6): bbae483.

Li, D.; Xie, W.; Zhang, J.; and Li, Y. 2024a. MDFL: Multi-Domain Diffusion-Driven Feature Learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 8653–8660.

Li, J.; Guan, R.; Han, Y.; Hu, Y.; Li, Z.; Wu, Y.; Xu, Z.; and Li, X. 2024b. Superpixel-based dual-neighborhood contrastive graph autoencoder for deep subspace clustering of hyperspectral image. In *International Conference on Intelligent Computing*, 181–192.

Li, J.; Lai, S.; Shuai, Z.; Tan, Y.; Jia, Y.; Yu, M.; Song, Z.; Peng, X.; Xu, Z.; Ni, Y.; Qiu, H.; Yang, J.; Liu, Y.; and Lu, Y. 2024c. A comprehensive review of community detection in graphs. *Neurocomputing*, 600: 128169.

Li, J.; Peng, X.; Hou, J.; Ke, W.; and Lu, Y. 2023. Community Detection Using Revised Medoid-Shift Based on KNN. In *International Conference on Intelligent Computing*, 345–353.

Li, Q.; Luo, T.; Jiang, M.; Liao, J.; and Jiang, Z. 2024d. Deep Incomplete Multi-View Network Semi-Supervised Multi-Label Learning with Unbiased Loss. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 9048–9056.

Liang, K.; Meng, L.; Liu, Y.; Liu, M.; Wei, W.; Liu, S.; Tu, W.; Wang, S.; and Liu, X. 2024. Simple Yet Effective: Structure Guided Pre-trained Transformer for Multi-modal Knowledge Graph Reasoning. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 1554–1563.

Liu, R.; Luo, T.; Huang, S.; Wu, Y.; Jiang, Z.; and Zhang, H. 2024a. CrossMatch: Cross-View Matching for Semi-Supervised Remote Sensing Image Segmentation. *IEEE Transactions on Geoscience and Remote Sensing*, 62: 1–15.

Liu, S.; Liao, Q.; Wang, S.; Liu, X.; and Zhu, E. 2024b. Robust and Consistent Anchor Graph Learning for Multi-View Clustering. *IEEE Transactions on Knowledge and Data Engineering*, 36(8): 4207–4219.

Liu, Y.; Diao, C.; Mei, W.; and Zhang, C. 2024c. CropSight: Towards a large-scale operational framework for object-based crop type ground truth retrieval using street view and PlanetScope satellite imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 216: 66–89.

Liu, Y.; Zhu, E.; Wang, Q.; Li, J.; Liu, S.; Hu, Y.; Han, Y.; Zhou, G.; and Guan, R. 2025. Spatial-Spectral Adaptive Graph Convolutional Subspace Clustering for Hyperspectral Image. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 18: 1139–1152.

Luo, F.; Liu, Y.; Gong, X.; Nan, Z.; and Guo, T. 2024. EMVCC: Enhanced Multi-View Contrastive Clustering for Hyperspectral Images. In *ACM Multimedia 2024*.

Ma, X.; Ma, M.; Hu, C.; Song, Z.; Zhao, Z.; Feng, T.; and Zhang, W. 2023. Log-can: local-global class-aware network for semantic segmentation of remote sensing images. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5.

Qu, Q.; Wan, X.; Liang, W.; Liu, J.; Feng, Y.; Xu, H.; Liu, X.; and Zhu, E. 2024. A Lightweight Anchor-Based Incremental Framework for Multi-view Clustering. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 8652–8661.

Shahi, K. R.; Ghamisi, P.; Rasti, B.; Scheunders, P.; and Gloaguen, R. 2022. Unsupervised Data Fusion With Deeper Perspective: A Novel Multisensor Deep Clustering Algorithm. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 15: 284–296.

Shen, F.; and Tang, J. 2024. IMAGPose: A Unified Conditional Framework for Pose-Guided Person Generation. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Shen, F.; Ye, H.; Zhang, J.; Wang, C.; Han, X.; and Wei, Y. 2024. Advancing Pose-Guided Image Synthesis with Progressive Conditional Diffusion Models. In *The Twelfth International Conference on Learning Representations*.

Topping, J.; Di Giovanni, F.; Chamberlain, B. P.; Dong, X.; and Bronstein, M. M. 2022. Understanding over-squashing and bottlenecks on graphs via curvature. In *International Conference on Learning Representations*.

Tu, W.; Guan, R.; Zhou, S.; Ma, C.; Peng, X.; Cai, Z.; Liu, Z.; Cheng, J.; and Liu, X. 2024. Attribute-missing graph clustering network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 15392–15401.

Tu, W.; Zhou, S.; Liu, X.; Guo, X.; Cai, Z.; Zhu, E.; and Cheng, J. 2021. Deep Fusion Clustering Network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 9978–9987.

Wan, X.; Liu, X.; Liu, J.; Wang, S.; Wen, Y.; Liang, W.; Zhu, E.; Liu, Z.; and Zhou, L. 2023. Auto-weighted multi-view clustering for large-scale data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 10078–10086.

Wang, J.; Guan, R.; Gao, K.; Li, Z.; Li, H.; Li, X.; and Tang, C. 2024. Multi-level Graph Subspace Contrastive Learning for Hyperspectral Image Clustering. In *2024 International Joint Conference on Neural Networks (IJCNN)*, 1–8.

Wang, S.; Liu, X.; Liu, S.; Jin, J.; Tu, W.; Zhu, X.; and Zhu, E. 2022. Align then fusion: Generalized large-scale multi-view clustering with anchor matching correspondences. *Advances in Neural Information Processing Systems*, 35: 5882–5895.

Xiao, S.; Du, S.; Chen, Z.; Zhang, Y.; and Wang, S. 2023. Dual Fusion-Propagation Graph Neural Network for Multi-View Clustering. *IEEE Transactions on Multimedia*, 25: 9203–9215.

Xiao, W.; Huang, Z.; Gan, L.; He, W.; Li, H.; Yu, Z.; Jiang, H.; Wu, F.; and Zhu, L. 2024a. Detecting and Mitigating Hallucination in Large Vision Language Models via Fine-Grained AI Feedback. arXiv:2404.14233.

Xiao, W.; Wang, Z.; Gan, L.; Zhao, S.; He, W.; Tuan, L. A.; Chen, L.; Jiang, H.; Zhao, Z.; and Wu, F. 2024b. A Comprehensive Survey of Direct Preference Optimization: Datasets, Theories, Variants, and Applications. arXiv:2410.15595.

Xie, W.; Lu, X.; Liu, Y.; Long, J.; Zhang, B.; Zhao, S.; and Wen, J. 2024. Uncertainty-aware pseudo-labeling and dual graph driven network for incomplete multi-view multi-label classification. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 6656–6665.

Xu, J.; Ren, Y.; Tang, H.; Yang, Z.; Pan, L.; Yang, Y.; Pu, X.; Yu, P. S.; and He, L. 2023. Self-Supervised Discriminative Feature Learning for Deep Multi-View Clustering. *IEEE Transactions on Knowledge and Data Engineering*, 35(7): 7470–7482.

Xu, J.; Tang, H.; Ren, Y.; Peng, L.; Zhu, X.; and He, L. 2022. Multi-level feature learning for contrastive multi-view clustering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16051–16060.

Yan, W.; Zhang, Y.; Lv, C.; Tang, C.; Yue, G.; Liao, L.; and Lin, W. 2023. Gcfagg: Global and cross-view feature aggregation for multi-view clustering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19863–19872.

Yang, X.; Liu, W.; and Liu, W. 2022. Tensor Canonical Correlation Analysis Networks for Multi-View Remote Sensing Scene Recognition. *IEEE Transactions on Knowledge and Data Engineering*, 34(6): 2948–2961.

Yang, X.; Liu, Y.; Zhou, S.; Wang, S.; Tu, W.; Zheng, Q.; Liu, X.; Fang, L.; and Zhu, E. 2023. Cluster-guided contrastive graph clustering network. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, 10834–10842.

Yu, H.; Ma, C.; Wan, X.; Wang, J.; Xiang, T.; Shen, M.; and Liu, X. 2025. DShield: Defending against Backdoor Attacks on Graph Neural Networks via Discrepancy Learning. *Network and Distributed System Security Symposium, NDSS*.

Yuan, Z.; Cao, J.; Li, Z.; Jiang, H.; and Wang, Z. 2024a. SD-MVS: Segmentation-Driven Deformation Multi-View Stereo with Spherical Refinement and EM Optimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 6871–6880.

Yuan, Z.; Cao, J.; Wang, Z.; and Li, Z. 2024b. Tsar-mvs: Textureless-aware segmentation and correlative refinement guided multi-view stereo. *Pattern Recognition*, 154: 110565.

Zhai, H.; Zhang, H.; Li, P.; and Zhang, L. 2021. Hyperspectral image clustering: Current achievements and future lines. *IEEE Geoscience and Remote Sensing Magazine*, 9(4): 35–67.

Zhang, Y.; Yan, S.; Jiang, X.; Zhang, L.; Cai, Z.; and Li, J. 2024. Dual Graph Learning Affinity Propagation for Multimodal Remote Sensing Image Clustering. *IEEE Transactions on Geoscience and Remote Sensing*, 62: 1–13.

Zhou, Y.; Liang, D.; Chen, S.; Huang, S.-J.; Yang, S.; and Li, C. 2023. Improving lens flare removal with general-purpose pipeline and multiple light sources recovery. In *Proceedings of the IEEE/CVF international conference on computer vision*, 12969–12979.

Zhou, Y.; Song, L.; Wang, B.; and Chen, W. 2024. MetaGPT: Merging Large Language Models Using Model Exclusive Task Arithmetic. *arXiv preprint arXiv:2406.11385*.