# Graph Masked Autoencoder for Multi-view Remote Sensing Data Clustering

**Renxiang Guan[1], Junhong Li[2], Siwei Wang[3], Tianrui Liu[1], Dayu Hu[4], Miaomiao Li[5], Xinwang Liu[1]**

[1]College of Computer Science and Technology, National University of Defense Technology, Changsha, China
[2]School of Computer Science, Northwestern Polytechnical University, Xi'an, China
[3]Intelligent Game and Decision Lab, Beijing, China
[4]College of Medicine and Biological Information Engineering, Northeastern University, Shenyang, China
[5]College of Electronic Information and Electrical Engineering, Changsha University, Changsha, China
renxiangguan@nudt.edu.cn

## Abstract

Multi-view graph clustering (MVGC) for remote sensing data has gained increasing attention due to its ability to integrate complementary information across modalities while capturing spatial dependencies in heterogeneous data. Although current methods based on graph contrastive learning achieve strong performance, they often misidentify intra-cluster samples as negatives, leading to class conflicts and reduced clustering accuracy. Graph masked autoencoders have recently shown promising potential in learning robust representations through masked reconstruction, but their application to remote sensing data remains underexplored. This challenge is especially notable in the multi-view remote sensing setting, where high heterogeneity and complex spatial structures increase the difficulty of effective representation learning. To address these issues, we propose Clustering-Guided graph Mask AutoEncoder (CG-MAE), the first framework to extend graph masked autoencoders to multi-view remote sensing clustering. We introduce a clustering-guided masking strategy that selectively masks nodes near cluster centers and intra-cluster edges, which are crucial for capturing key structural information. By reconstructing these masked components, the model is encouraged to focus on learning features that are highly relevant to clustering. To further improve training stability and efficiency, we design an easy-to-hard node masking strategy that enables the model to gradually learn from increasingly challenging patterns. Additionally, we propose a dual self-adaptive learning mechanism that encourages the model to align more closely with the underlying semantic distributions. Extensive experiments on four widely used multi-view remote sensing datasets demonstrate that CG-MAE consistently outperforms state-of-the-art methods in both clustering accuracy and representation quality.

## Introduction

Unsupervised clustering of remote sensing data has drawn increasing attention, as it offers a promising way to alleviate the high costs of manual annotation (Guan et al. 2024c). Early work in this field primarily focused on single-view data (Guan et al. 2024a), while advancements in satellite technologies have enabled mature multi-source Earth observation systems. Diverse data modalities (Guan et al. 2025a; Liu et al. 2023) such as multispectral (MS), hyperspectral (HS) imagery, light detection and ranging (LiDAR), and synthetic aperture radar (SAR) can now been acquired. These multi-view data provide complementary information that can significantly enhance clustering performance (Guan et al. 2025b). For instance, LiDAR data contributes crucial 3D geometric information, effectively disambiguating spectrally similar objects.

Existing multi-view clustering (MVC) methods for remote sensing data can be broadly categorized into traditional approaches (Zhang et al. 2025b) and deep learning-based techniques (Liu and Chang 2025; Peng et al. 2025). Traditional methods, such as distance-based, graph-based (Cai et al. 2024), and subspace-based approaches (Chen et al. 2022), often formulate clustering as an optimization problem to seek an optimal solution. However, these shallow-feature models struggle to capture the highly nonlinear and complex distributions inherent in data modalities like LiDAR and HS imagery. In contrast, deep learning-based methods have demonstrated great potential due to their powerful representation learning capabilities. Common architectures include convolutional neural networks (CNNs) (Shahi et al. 2022), Transformers (Cai et al. 2023), and graph convolutional networks (GCNs) (Guan et al. 2024b). Due to its inherent network architecture, CNNs are limited to capturing local spatial information and struggle to fully leverage global contextual dependencies (Guan et al. 2022). In contrast, Transformers have emerged as a powerful alternative by overcoming the restricted receptive fields of traditional CNNs and enabling long-range dependency modeling. However, the quadratic computational complexity of standard Transformers poses significant challenges for processing large-scale remote sensing data. On the other hand, GCNs have gained prominence for their ability to aggregate information from neighboring nodes, making them particularly suitable for addressing the strong heterogeneity of remote sensing data (Guan et al. 2025d).

Recent GCN-based MVC approaches often rely on contrastive learning (Guan et al. 2024b), which encourages positive samples to cluster closely while pushing apart negative samples. While effective, these methods face significant challenges in constructing reliable positive and negative pairs for complex remote sensing data. Mislabeling the sam-

ples within the same cluster as negative samples may lead to severe class conflicts. Graph autoencoders represent another mainstream paradigm, aiming to enhance feature representations by reconstructing input data. To further improve this, graph masked autoencoders (Hou et al. 2022) have been proposed, which mask a subset of nodes or edges and learn to reconstruct them, thereby improving representation robustness. However, applying such models to remote sensing clustering remains challenging. On one hand, remote sensing data often exhibit high heterogeneity and complex spatial structures, which make it difficult to construct graph representations that preserve meaningful information. On the other hand, clustering tasks lack explicit supervision signals, making it nontrivial for graph masked autoencoders to identify and reconstruct features that are truly relevant for clustering.

To address these challenges, we propose a clustering-guided graph masked autoencoder (CG-MAE) for multi-view remote sensing data clustering. Our approach first applies superpixel segmentation to partition remote sensing images into irregular regions, reducing computational overhead and constructing superpixels as graph nodes. We then introduce a novel clustering-guided masking strategy that selectively masks nodes near cluster centers and intra-cluster edges—structures most critical for clustering. By reconstructing these masked elements, the model learns to extract and emphasize key clustering-relevant features. Furthermore, we design an easy-to-hard masking curriculum for nodes, facilitating smoother model training. Finally, we propose a dual self-adaptive learning to encourage more accurate and semantically meaningful reconstructions. Extensive experiments and visualization results demonstrate the superior performance of our method across multiple remote sensing datasets. The main contributions of this work are summarized as follows:

- To the best of our knowledge, this is the first work to explore graph masked autoencoders for multi-view remote sensing data clustering.
- We propose a clustering-guided graph masked autoencoder that selectively masks key nodes and edges, encouraging the model to focus on extracting clustering-relevant features.
- Extensive experiments on four benchmark multi-view remote sensing datasets demonstrate the effectiveness and superior performance of the proposed approach.

## Related Work

### Deep MVGC for Remote Sensing Data

Multi-view information has attracted considerable attention from researchers (Zhou et al. 2025; Xiao et al. 2025; Feng et al. 2025; Yang et al. 2025; Hu et al. 2024a), as it enables complementary perspectives—such as terrain, spectral, and other modalities—to enhance clustering tasks (Wang et al. 2024a, 2025b; Guan et al. 2025c; Hu et al. 2024c). Existing graph-based MVC approaches can generally be categorized into traditional methods (Cao et al. 2025a,b) and deep learning-based methods (Guan et al. 2024b, 2025d). While traditional algorithms provide objective optimization for problem-solving, they are limited in their ability to extract deep representations from the data. To address this, deep MVGC techniques have emerged, particularly those leveraging contrastive learning. For instance, CMSCGC (Guan et al. 2024b) integrates contrastive learning with subspace clustering to make an initial attempt at deep MVC in this context. Wang *et al.* (Wang et al. 2024b) design multi-level graph contrastive learning to combine multi-view information for feature extraction. Guan *et al.* (Guan et al. 2025d) introduce the concept of confusing samples to design a contrastive learning module that optimizes the clustering structure for multi-view remote sensing data. Similarly, SEC-LSRM (Guan et al. 2025b) proposes a contrastive framework that employs block-diagonal and idempotent-constrained affinity matrices as sampling matrices to generate positive and negative sample pairs. However, contrastive learning-based methods often suffer from high computational complexity and difficulties in constructing accurate positive and negative sample pairs, which can easily lead to class conflict issues. We first implement the application of graph masked autoencoders on a multi-view remote sensing dataset and achieve excellent clustering performance.

### Graph Masked Autoencoders

Graph masked autoencoders are advanced forms of GAEs that boost the feature extraction capability and robustness by reconstructing masked data (Zhuo et al. 2024). Early graph autoencoder-based approaches are mainly categorized into two types in terms of reconstruction: either at the feature level (Hou et al. 2022; Shi et al. 2023; Wang et al. 2024c) or at the structure level (Li et al. 2023). The most notable ones among them are GraphMAE (Hou et al. 2022) and MaskGAE (Li et al. 2023). GraphMAE designs a re-masked strategy to further mask the encoded features, aiming to enhance the robustness of the decoder. Meanwhile, MaskGAE employs structure and degree dual autoencoders to make full use of structural information for guiding the reconstruction. Later methods usually design more sophisticated masking strategies to improve the representation ability of the model. For instance, AUG-MAE (Wang et al. 2024c) devises a masked learning method from easy to difficult, generating the mask probability of each node via the mask generator of the graph neural network to control the model learning process. Hi-GMAE (Liu et al. 2024) constructs a multi-scale graph hierarchy through graph pooling and conducts multi-level masking and unpooling. Duan *et al.* (Duan, Yu, and Xie 2024) design graph view, masked view, diffusion view, and masked diffusion view and promote consistency learning between views to improve the model's feature extraction capability. Moreover, there are also some methods that combine masked GAE and contrastive learning regularization. For example, UGMAE (Tian et al. 2024) masks the edges and features separately and employs the momentum encoder to extract the difference features for contrastive learning.

## Methods

In this section, we present the proposed Clustering-Guided Graph Mask Autoencoder (CG-MAE) in detail. An

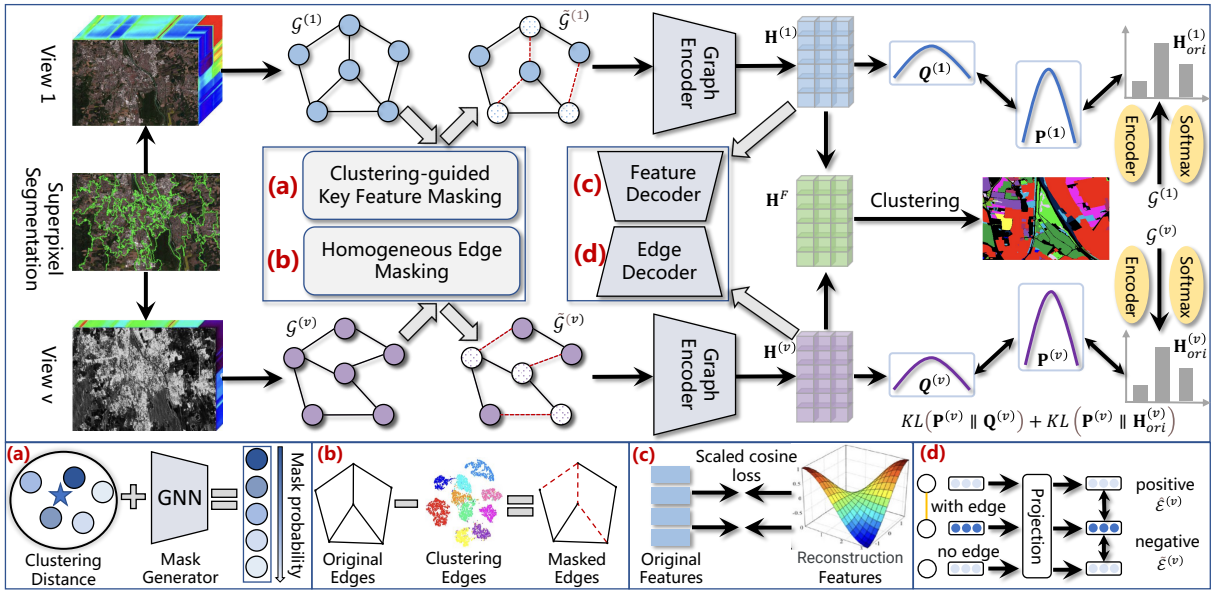


Figure 1: Illustration of the proposed CG-MAE. We propose a clustering-guided masking strategy that focuses on reconstructing nodes and edges critical to clustering, encouraging the model to learn more discriminative representations. To further enhance learning stability, a dual self-adaptive learning module is designed to guide the model toward better semantic alignment.

overview of the architecture is illustrated in Fig. 1.

## Notations

Let $\mathbf{X}^{(v)} \in \mathbb{R}^{W\times H\times D^{(v)}}$ denote the multi-view remote sensing data, where $W$ and $H$ represent the spatial dimensions, $D^{(v)}$ is the number of spectral bands in the $v$-th view, and $N = W \times H$ denotes the total number of pixels. To reduce computational overhead, we use the SLIC segmentation (Achanta et al. 2012) method on the fusion view $\mathbf{X}^F = \mathbf{X}^{(1)}\|\mathbf{X}^{(2)}\|\ldots\|\mathbf{X}^{(v)}$ to segment the HSI into superpixels $\mathbf{Z}^{(v)} = \bigcup_{i=1}^{M}\mathbf{z}_i^{(v)}$, where $\mathbf{z}_i^{(v)} = \frac{1}{N_i}\sum_{j=1}^{N_i}(\mathbf{x}^{(v)})_i^j$ represents the $i$-th superpixel in the $v$-th view, $(\mathbf{x}^{(v)})_i^j$ denotes the $j$-th element in $\mathbf{z}_i^{(v)}$, $M$ is the number of superpixels, $N_i$ is the number of pixel in $\hat{\mathbf{x}}_i^{(v)}$, where $\mathbf{z}_i^{(v)}\cap\mathbf{z}_j^{(v)} = \emptyset, \forall i \neq j$ and $\sum_{i=1}^{M} N_i = N$. The superpixel segmentation results are shared across each view. The number of superpixels $M$ after segmentation is much smaller than the number of pixels $N$.

Given that superpixels represent higher-level semantic units, and to effectively capture spatial dependencies, we construct a graph structure to model the relationships among them. Specifically, we define an undirected graph $\mathcal{G}^{(v)} = (\mathcal{V}^{(v)}, \mathcal{E}^{(v)})$, where $\mathcal{V}^{(v)} = \{\mathbf{v}_i\}_{i=1}^{M}$ denotes the node set, $\mathcal{E}^{(v)}$ the edge set, and $\mathbf{Z}^{(v)} \in \mathbb{R}^{M\times D^{(v)}}$ the node feature matrix. To build the adjacency matrix $\mathbf{A}^{(v)} \in \mathbb{R}^{M\times M}$, we compute the pairwise similarity between superpixel features and connect each node to its $k$ nearest neighbors. The similarity is computed using a Gaussian kernel as:

$$\mathbf{u}_{ij}^{(v)} = \exp\left(-\frac{\|\mathbf{z}_i^{(v)} - \mathbf{z}_j^{(v)}\|_2^2}{2\sigma^2}\right), \quad \forall i,j = 1,2,\ldots,M, \tag{1}$$

where $\sigma$ is a bandwidth parameter controlling the sensitivity of the kernel. The binary adjacency matrix $\mathbf{A}^{(v)}$ is then constructed such that $\mathbf{A}_{ij}^{(v)} = 1$ if an edge exists between nodes $\mathbf{v}_i^{(v)}$ and $\mathbf{v}_j^{(v)}$, and 0 otherwise. To prepare the graph for message passing, we perform symmetric normalization on the adjacency matrix as follows:

$$\tilde{\mathbf{A}}^{(v)} = \mathbf{D}^{(v)^{-\frac{1}{2}}}(\mathbf{A}^{(v)} + \mathbf{I})\mathbf{D}^{(v)^{-\frac{1}{2}}}, \tag{2}$$

where $\mathbf{I}$ is the identity matrix added to include self-loops, and $\mathbf{D}^{(v)}$ is the diagonal degree matrix of the $v$-th view.

## Clustering-Guided Key Feature Masking

Certain regions in remote sensing data contain crucial spatial structures that significantly influence clustering performance. Identifying and leveraging these key nodes is essential for learning effective representations. However, existing graph mask autoencoders often adopt random masking strategies, ignoring the structural importance of specific nodes, which leads to suboptimal clustering performance.

To address this issue, we propose a clustering-guided masking strategy that incorporates both structural and semantic information. Specifically, we introduce a mask generator $\mathcal{M}_\Phi(\cdot)$ based on graph neural networks, which estimates the importance of each node. The resulting mask probability for view $v$ is denoted as: $\text{prob}_{\text{gnn}}^{(v)} = \mathcal{M}_\Phi(\mathcal{G}^{(v)})$.

In parallel, we compute a clustering-based mask probability vector $prob_{\text{clu}}^{(v)}$ by evaluating the similarity between each node and its assigned cluster center $\mathcal{C}^{(v)} = \left[c_1^{(v)}, c_2^{(v)}, \ldots, c_K^{(v)}\right]^\top$. Nodes closer to the cluster center are

more likely to be masked:

$$p_{clu}{}_i^{(v)} = \frac{\exp\left(-\|\mathbf{z}_i^{(v)} - c_{y_i}^{(v)}\|^2/\hat{\sigma}^2\right)}{\sum_{\mathbf{z}_j \in \mathcal{C}_{y_i}^{(v)}} \exp\left(-\|\mathbf{z}_j^{(v)} - c_{y_i}^{(v)}\|^2/\hat{\sigma}^2\right)}, \quad (3)$$

where $y_i$ is the pseudo label of node $i$ and $\text{prob}_{\text{clu}}^{(v)} = \left[p_{clu}{}_1^{(v)}, p_{clu}{}_2^{(v)}, \ldots, p_{clu}{}_M^{(v)}\right]^\top$.

To stabilize training, we adopt an easy-to-hard curriculum strategy inspired by (Wang et al. 2024c). At the early training stage, the model primarily masks nodes near the cluster center to focus on easier samples. Over time, the masking gradually shifts toward harder edge nodes to enhance the model's discriminative power. The dynamic blending of the two masking strategies is controlled by:

$$\text{prob}^{(v)}(t) = (1 - \psi(t)) \cdot \text{prob}_{\text{clu}}^{(v)} + \psi(t) \cdot \text{prob}_{\text{gnn}}^{(v)}(t), \quad (4)$$

where $\text{prob}^{(v)}(t)$ is the final masking probability at epoch $t$, and $\psi(t)$ is the scheduling weight defined as:

$$\psi(t) = \psi_0 + \left(\frac{t}{T}\right)^{\eta} \cdot (\psi_T - \psi_0), \quad (5)$$

where $\psi_0$ and $\psi_T$ are the initial and final weights, $T$ is the total number of epochs, and $\eta$ controls the growth rate.

We then use the Gumbel-Softmax trick to generate a binary mask vector $\boldsymbol{m}^{(v)} \in \{0,1\}^M$:

$$\boldsymbol{m}_i^{(v)} = \text{Sigm}\left(\frac{1}{\tau}\left(\log\left(\frac{\text{prob}_i^{(v)}}{1 - \text{prob}_i^{(v)}}\right) + (\epsilon_0 - \epsilon_1)\right)\right), \quad (6)$$

where $\epsilon_0, \epsilon_1 \sim \text{Gumbel}(0,1)$, $\text{Sigm}(\cdot)$ is the sigmoid activation function and $\tau$ is the temperature parameter. The final masked node features are given by:

$$\tilde{\boldsymbol{z}}_i = \begin{cases} \boldsymbol{z}_{[\text{MASK}]} & \text{if } v_i \in \widetilde{\mathcal{V}}, \\ \boldsymbol{z}_i & \text{otherwise}, \end{cases} \quad (7)$$

where $\widetilde{\mathcal{V}}$ is the set of masked nodes. We then apply graph encoders to process the masked features:

$$\mathbf{H}_{(l)}^{(v)} = \sigma\left(\tilde{\mathbf{A}}^{(v)} \mathbf{H}_{(l-1)}^{(v)} \mathbf{W}_{(l)}^{(v)} + \mathbf{b}_{(l)}^{(v)}\right), \quad (8)$$

where $\mathbf{W}_{(l)}^{(v)}$ and $\mathbf{b}_{(l)}^{(v)}$ denote the learnable weight and bias at layer $l$, and $\mathbf{H}_{(0)}^{(v)} = \widetilde{\mathbf{Z}}^{(v)}$ is the input masked feature matrix.

Following the setup in GraphMAE (Hou et al. 2022), a decoder is used to reconstruct the masked node features. The reconstruction loss is defined as:

$$\mathcal{L}_{\text{FR}} = \frac{1}{|\widetilde{\mathcal{V}}|} \sum_{v=1}^{V} \sum_{i \in \widetilde{\mathcal{V}}} \left(1 - \frac{\mathbf{z}_i^{(v)\top} \hat{\mathbf{z}}_i^{(v)}}{\|\mathbf{z}_i^{(v)}\| \cdot \|\hat{\mathbf{z}}_i^{(v)}\|}\right)^{\gamma}, \quad \gamma \geq 1, \quad (9)$$

where $\gamma$ is a scaling factor that adjusts the sensitivity of the reconstruction loss and $\hat{\mathbf{Z}}^{(v)}$ represent the reconstructed features.

## Homogeneous Edge Masking

As discussed previously, random edge masking suffers from several limitations. In many graph-based remote sensing models, the adjacency structure is constructed purely based on nearest-neighbor relationships in feature space. This often leads to the inclusion of numerous *heterogeneous edges*—i.e., edges connecting samples from different semantic clusters—which may hinder meaningful message propagation and ultimately degrade clustering performance.

To mitigate this issue, we propose a **homogeneous edge masking strategy** aimed at enhancing structural consistency in the graph. Specifically, we first encode the input features $\mathbf{Z}^{(v)}$ to obtain latent representations $\mathbf{H}_{\text{ori}}^{(v)}$ and perform clustering to partition all nodes into $K$ clusters. We then define *homogeneous edges* as those connecting nodes within the same cluster, and use them to construct a homogeneous adjacency matrix $\mathbf{A}_c^{(v)}$. Since clustering outcomes can be unstable, we apply multiple K-Means clustering runs to improve robustness. The final homogeneous adjacency matrix is defined as the intersection of the individual clustering adjacency matrices:

$$\mathbf{A}_c^{(v)} = \mathbf{A}_{c1}^{(v)} \cap \mathbf{A}_{c2}^{(v)} \cap \cdots \cap \mathbf{A}_{cn}^{(v)}. \quad (10)$$

Based on the original adjacency matrix $\mathbf{A}^{(v)}$ and the refined homogeneous matrix $\mathbf{A}_c^{(v)}$, we partition the edge set into two disjoint subsets: homogeneous edges $\widetilde{\mathcal{E}}^{(v)}$ and heterogeneous edges $\widehat{\mathcal{E}}^{(v)}$. Our goal is to encourage reconstruction of the former while suppressing the latter. To this end, we introduce a single-layer multilayer perceptron decoder $f_D$ on masked representation $\mathbf{H}^{(v)}$ to reconstruct pairwise edge connections. The edge prediction score is defined as:

$$s_{n,n'}^{(v)} = \text{Sigm}\left(f_D\left(\mathbf{h}_n^{(v)}, \mathbf{h}_{n'}^{(v)}\right)\right), \quad (11)$$

where $s_{n,n'}^{(v)}$ denotes the predicted probability of an edge existing between node $n$ and node $n'$.

The edge reconstruction loss is then formulated as:

$$\mathcal{L}_{\text{ER}} = \sum_{v=1}^{V} \left(-\frac{1}{|\widetilde{\mathcal{E}}^{(v)}|} \sum_{e_{nn'}^{(v)} \in \widetilde{\mathcal{E}}^{(v)}} \log s_{n,n'}^{(v)} - \frac{1}{|\widehat{\mathcal{E}}^{(v)}|} \sum_{e_{ll'}^{(v)} \in \widehat{\mathcal{E}}^{(v)}} \log\left(1 - s_{l,l'}^{(v)}\right)\right). \quad (12)$$

This loss encourages the model to reconstruct edges between semantically similar nodes while penalizing the reconstruction of spurious heterogeneous connections, thereby improving the structural quality of learned graphs.

## Dual Self-Adaptive Learning

To further enhance feature consistency and improve the model's representation capability, we propose a *dual self-adaptive learning* module. Specifically, we encourage the feature distributions of the original graph $\mathcal{G}^{(v)}$ and the masked graph $\widetilde{\mathcal{G}}^{(v)}$ to remain consistent.

We begin by transforming the feature matrix $\mathbf{H}^{(v)}$ into a probability distribution using the Student's $t$-distribution:

$$q_{ij}^{(v)} = \frac{\left(1 + \left\|\mathbf{h}_i^{(v)} - \mathbf{c}_j^{(v)}\right\|^2/\zeta\right)^{-\frac{\zeta+1}{2}}}{\sum_{j'} \left(1 + \left\|\mathbf{h}_i^{(v)} - \mathbf{c}_{j'}^{(v)}\right\|^2/\zeta\right)^{-\frac{\zeta+1}{2}}}, \quad (13)$$

where $\zeta$ is the degrees of freedom parameter of the distribution. Let $\mathbf{Q}^{(v)}$ denote the matrix with entries $q_{ij}^{(v)}$. Then, we compute the target distribution $\mathbf{P}^{(v)}$ from $\mathbf{Q}^{(v)}$ as follows:

$$p_{ij}^{(v)} = \frac{{q_{ij}^{(v)}}^2 / \sum_i q_{ij}^{(v)}}{\sum_{j'} {q_{ij'}^{(v)}}^2 / \sum_i q_{ij'}^{(v)}}. \tag{14}$$

Additionally, we encode the unmasked graph $\mathcal{G}^{(v)}$ using a shared encoder $f_E$ followed by a softmax operation:

$$\mathbf{H}_{\text{ori}}^{(v)} = \text{Softmax}(f_E(\mathcal{G}^{(v)})). \tag{15}$$

To enforce dual consistency, we minimize the Kullback–Leibler divergence between the target distribution $\mathbf{P}^{(v)}$ and both the soft cluster assignment $\mathbf{Q}^{(v)}$ and the original output $\mathbf{H}_{\text{ori}}^{(v)}$:

$$\begin{aligned}\mathcal{L}_{\text{KL}} &= \sum_{v=1}^{V} \text{KL}\left(\mathbf{P}^{(v)} \| \mathbf{Q}^{(v)}\right) + \sum_{v=1}^{V} \text{KL}\left(\mathbf{P}^{(v)} \| \mathbf{H}_{\text{ori}}^{(v)}\right) \\ &= \sum_{v=1}^{V} \sum_i \sum_j \left( p_{ij}^{(v)} \log \frac{p_{ij}^{(v)}}{q_{ij}^{(v)}} + p_{ij}^{(v)} \log \frac{p_{ij}^{(v)}}{\mathbf{z}_{\text{ori}ij}^{(v)}} \right).\end{aligned} \tag{16}$$

## The Overall Loss Function

In summary, we introduce a novel masked MVGC framework for remote sensing data. During the training stage, the feature reconstruction loss, edge reconstruction loss and dual self-adaptive learning constraint are used to jointly train the model as follows,

$$\mathcal{L} = \mathcal{L}_{\text{FR}} + \lambda_1 * \mathcal{L}_{\text{ER}} + \lambda_2 * \mathcal{L}_{\text{KL}}. \tag{17}$$

In the training phase, we minimize $\mathcal{L}$ to optimize the proposed CG-MAE.

# Experiments

## Experiment Setup

**Datasets** To evaluate the proposed method, we conducted experiments on four multi-view remote sensing datasets: Trento, Salinas, XuZhou and MDAS. The Trento dataset contains LiDAR, DSM, and HS imagery, with a spatial size of 600 by 166 pixels, six land-cover classes, and 30,214 labeled samples. The Salinas dataset includes HS, EMP, and Gabor features, covering 512 by 217 pixels with 16 land-cover categories and 54,129 samples. The XuZhou dataset consists of HS, EMP, and Gabor modalities, with a size of 500 by 260 pixels, nine land-cover types, and 68,877 samples. The MDAS dataset provides SAR, MS, HS, and DSM data, all aligned at 300 by 360 pixels, containing nine categories and 88,026 samples.

To evaluate the effectiveness of our proposed CG-MAE, we compare it with fifteen state-of-the-art MVC methods, including MFLVC (Xu et al. 2022), FMVACC (Wang et al. 2022), MDC (Shahi et al. 2022), CVCL (Chen et al. 2023), SDMVC (Xu et al. 2023), GCFAgg (Yan et al. 2023), AWMVC (Wan et al. 2023), TMPCC (Cai et al. 2023), MDFL (Li et al. 2024), CMSCGC (Guan et al. 2024b), AMKSC (Cai et al. 2024), CDD (Liu and Chang 2025), DMAC (Wang et al. 2025a), ESTMC (Ji and Feng 2025) and SAMVGC (Guan et al. 2025d).

**Implement Details** To ensure a fair and objective comparison between the proposed CG-MAE and baseline algorithms, all experiments are conducted on a server equipped with an NVIDIA RTX 3090 GPU and 64 GB RAM. Each method is independently run ten times, and the average results are reported to mitigate the impact of randomness. The proposed model is trained using the Adam optimizer with a learning rate of $1 \times 10^{-3}$ for 200 epochs. Both the encoder and decoder adopt the GCN architecture, with the hidden layer dimension set to 512. After thorough hyperparameter tuning, $\alpha$ is set to 50. For the Trento, Salinas, XuZhou, and MDAS datasets, the number of superpixel neighbors $k$ is set to 40, 30, 10, and 10, respectively, and the number of superpixels $M$ is set to 7000, 3500, 1000, and 3500, respectively. In Eq. 17, the parameters $\lambda_1$ and $\lambda_2$ are assigned as follows: 0.01/1000, 10/1000, 0.1/0.1, and 0.01/1, respectively. To quantitatively evaluate clustering performance, five commonly used metrics are adopted: Accuracy (ACC), Kappa coefficient, Normalized Mutual Information (NMI), Adjusted Rand Index (ARI), and Purity (PUR) (Zhang, Zhang, and Yuan 2024; Zhang et al. 2025a; Fang et al. 2025; Hu et al. 2024b; Meng et al. 2024b,a).

## Quantitative Results

Table 1 reports the clustering performance of our CG-MAE and several representative comparison methods on four datasets. CG-MAE consistently achieves the best results, significantly outperforming all baselines. 1) Deep MVC methods generally surpass traditional ones. For instance, TMPCC, CMSCGC, and SAMVGC outperform FMVACC and AWMVC due to their capacity to capture complex nonlinear structures, enhancing robustness and discriminability. 2) Remote sensing-oriented models usually perform better than general-purpose approaches. TMPCC and CMSCGC achieve superior results by leveraging strong spatial correlations and handling complex noise patterns. 3) TMPCC, CMSCGC, and SAMVGC show relatively strong performance, with TMPCC and CMSCGC being remote sensing-specific deep models, and CMSCGC and SAMVGC adopting graph-based strategies. Despite their strengths, CG-MAE outperforms all of them across datasets. Notably, on Salinas, CG-MAE improves ACC by 37.09%, 12.55%, and 15.02% over TMPCC, CMSCGC, and SAMVGC, respectively. Similar gains are observed on Trento (7.02%/6.85%/0.55%), XuZhou (21.02%/11.36%/7.66%), and MDAS (25.12%/8.85%/1.50%). These results demonstrate the superior robustness of CG-MAE and validate the effectiveness of the proposed modules.

## Visualization

To visually demonstrate the effectiveness of CG-MAE, we present the clustering results of various algorithms on the Trento and XuZhou datasets, as illustrated in Fig. 2. It can be observed that several traditional MVC methods, such as

| Method | Trento | | | | | XuZhou | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | ACC | Kappa | NMI | ARI | PUR | ACC | Kappa | NMI | ARI | PUR |
| **MFLVC** | 51.15±2.26 | 30.23±2.17 | 40.49±1.96 | 28.24±1.18 | 51.31±1.13 | 39.66±2.13 | 31.62±3.74 | 32.61±4.63 | 20.51±4.16 | 53.47±1.80 |
| **FMVACC** | 75.03±0.11 | 25.96±0.02 | 55.81±0.37 | 60.86±0.17 | 77.76±0.02 | 50.20±0.48 | 31.74±0.58 | 30.42±0.67 | 35.48±0.49 | 52.71±0.85 |
| **MDC** | 83.39±0.14 | 78.54±0.18 | 76.48±0.21 | 71.20±0.17 | 84.46±0.16 | 56.90±2.50 | 47.14±2.93 | 50.07±1.45 | 48.32±2.61 | 66.12±1.74 |
| **CVCL** | 58.36±5.73 | 48.23±6.34 | 53.01±5.61 | 40.78±6.13 | 73.12±3.47 | 38.95±1.44 | 29.32±2.53 | 35.86±1.68 | 19.67±1.43 | 59.62±1.55 |
| **SDMVC** | 63.46±1.14 | 48.34±1.26 | 39.13±1.04 | 38.22±0.11 | 63.46±0.54 | 59.89±0.03 | 51.08±0.06 | 58.10±0.04 | 55.78±0.02 | 67.65±0.01 |
| **GCFAgg** | 56.05±4.64 | 42.25±4.99 | 47.78±4.78 | 32.92±5.95 | 57.36±4.00 | 68.84±0.45 | 61.53±0.52 | 63.92±1.49 | 63.25±0.75 | 77.46±0.67 |
| **AWMVC** | 58.74±0.76 | 18.30±0.15 | 42.07±0.97 | 37.31±0.59 | 67.42±0.09 | 34.29±1.16 | 6.72±2.16 | 30.64±0.76 | 17.25±0.81 | 59.76±0.89 |
| **TMPCC** | 88.57±4.06 | 87.09±6.65 | 82.26±4.91 | 90.24±3.33 | 90.20±2.52 | 62.40±1.01 | 59.33±6.54 | 66.95±5.15 | 62.58±6.07 | 72.10±5.20 |
| **MDFL** | 86.01±2.17 | 69.45±1.19 | 59.90±1.77 | 60.23±1.55 | 76.01±2.24 | 61.78±1.59 | 62.76±2.54 | 60.88±1.32 | 59.72±1.44 | 70.07±1.79 |
| **CMSCGC** | 88.74±2.14 | 88.59±2.16 | 85.05±2.88 | 82.45±2.19 | 90.05±1.92 | 72.06±2.35 | 66.62±1.48 | 66.01±1.32 | 72.36±1.83 | 81.54±2.10 |
| **AMKSC** | 93.90±1.97 | 91.85±2.58 | 88.21±1.20 | 92.83±1.94 | 93.90±1.75 | 71.92±5.16 | 66.26±5.49 | 69.77±1.74 | 61.20±7.79 | 80.29±0.05 |
| **CDD** | 70.83±6.38 | 62.43±7.81 | 64.45±4.69 | 60.93±6.89 | 79.28±4.41 | 56.65±4.45 | 48.15±4.65 | 52.30±1.75 | 47.59±6.50 | 67.17±1.69 |
| **DMAC** | 79.55±0.85 | 72.53±0.66 | 71.01±5.47 | 66.29±3.22 | 82.39±0.87 | 70.25±2.20 | 63.40±2.44 | 62.73±2.94 | 63.87±1.06 | 77.12±1.47 |
| **ESTMC** | 66.10±0.97 | 58.01±1.47 | 53.09±0.86 | 55.85±2.48 | 79.18±3.59 | 40.15±2.37 | 32.42±1.95 | 36.98±2.19 | 24.24±0.13 | 62.30±0.78 |
| **SAMVGC** | 95.04±0.36 | 93.59±0.35 | **93.36±0.49** | 93.75±0.66 | 95.52±0.37 | 75.76±0.58 | 70.58±0.64 | 72.50±0.79 | 61.97±1.08 | **85.30±0.77** |
| **Ours** | **95.59±0.35** | **94.13±0.47** | 91.33±0.40 | **96.14±0.24** | **96.02±0.24** | **83.42±1.57** | **79.28±1.95** | **74.26±0.71** | **73.73±3.77** | 83.91±1.18 |

| Method | Salinas | | | | | MDAS | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | ACC | Kappa | NMI | ARI | PUR | ACC | Kappa | NMI | ARI | PUR |
| **MFLVC** | 41.64±2.33 | 33.25±2.17 | 42.75±1.85 | 31.93±2.05 | 43.31±1.98 | 40.56±3.84 | 13.21±6.92 | 9.49±4.73 | 7.03±5.01 | 51.76±1.07 |
| **FMVACC** | 66.50±2.04 | 37.89±1.99 | 76.82±2.15 | 51.61±2.59 | 74.38±1.89 | 25.68±0.53 | 1.46±0.02 | 14.81±0.00 | 2.93±0.01 | 49.46±0.00 |
| **MDC** | 68.60±3.06 | 64.72±3.51 | 77.28±1.06 | 58.63±2.83 | 71.42±1.14 | 21.95±0.13 | 12.50±0.27 | 17.05±0.58 | 4.59±0.22 | 51.93±0.19 |
| **CVCL** | 34.47±1.87 | 25.76±2.53 | 41.38±1.75 | 24.87±2.38 | 38.50±1.90 | 34.69±4.22 | 4.81±2.87 | 6.18±2.16 | 1.33±0.68 | 49.45±0.01 |
| **SDMVC** | 70.85±0.01 | 67.71±0.01 | 77.45±0.03 | 56.97±0.02 | 77.10±0.04 | 29.73±0.12 | 11.74±0.04 | 9.55±0.06 | 2.89±0.19 | 49.69±0.05 |
| **GCFAgg** | 57.80±5.14 | 54.39±5.27 | 69.43±1.34 | 47.31±4.29 | 70.68±1.92 | 23.70±0.55 | 14.17±0.11 | 16.96±1.16 | 5.05±0.57 | 53.87±1.42 |
| **AWMVC** | 40.19±1.55 | 6.16±2.03 | 56.44±1.96 | 32.31±1.61 | 57.92±1.83 | 21.31±0.36 | 11.28±2.68 | 15.94±0.05 | 4.00±0.15 | 52.15±0.02 |
| **TMPCC** | 45.65±3.92 | 40.04±3.43 | 60.69±5.58 | 36.21±8.53 | 48.70±2.32 | 23.97±0.81 | 6.71±0.54 | 14.51±1.04 | 19.60±1.33 | 54.13±0.81 |
| **MDFL** | 38.19±2.33 | 32.85±1.59 | 54.37±2.17 | 30.77±1.76 | 55.30+2.79 | 25.03±1.69 | 11.96±2.04 | 12.93±3.35 | 5.97±3.10 | 32.78±2.83 |
| **CMSCGC** | 70.19±1.71 | 66.83±2.85 | 76.03±2.21 | 55.93±1.44 | 75.06±1.54 | 40.24±1.43 | 20.99±2.10 | 25.95±1.22 | 25.95±1.77 | 60.84±0.44 |
| **AMKSC** | 70.35±2.71 | 66.99±2.67 | 80.30±2.62 | 61.40±4.10 | 74.41±1.39 | 27.85±1.83 | 18.62±1.43 | 24.23±0.38 | 12.47±0.97 | 59.01±0.05 |
| **CDD** | 60.07±2.78 | 55.18±3.08 | 70.88±3.41 | 54.11±5.30 | 62.01+2.44 | 30.79±3.29 | 18.58±2.61 | 20.11±2.55 | 10.66±2.68 | 56.42±1.75 |
| **DMAC** | 74.53±0.85 | 71.67±1.00 | 78.30±0.60 | 60.21±2.26 | 76.46±1.48 | 21.96 ± 1.46 | 12.58±1.48 | 14.24±1.58 | 4.78±0.47 | 51.95±0.66 |
| **ESTMC** | 44.36±2.56 | 40.33±3.98 | 50.94±3.45 | 29.88±1.89 | 57.15±1.66 | 22.66±1.49 | 16.28±1.56 | 17.53±3.69 | 7.46±2.43 | 61.82±1.52 |
| **SAMVGC** | 67.72±0.30 | 64.30±0.29 | 69.45±0.38 | 54.81±0.80 | 69.78±0.52 | 47.59±0.70 | 28.85±0.64 | **35.09±0.87** | **35.59±0.57** | **66.30±0.55** |
| **Ours** | **82.74 ± 0.54** | **80.89±0.59** | **83.27±1.04** | **76.65±1.71** | **86.81±0.62** | **49.09±1.44** | **35.23±1.41** | 33.00±0.72 | 28.98±1.06 | 65.25±0.15 |

Table 1: Clustering performance of fifteen multi-view clustering (MVC) methods on four datasets.

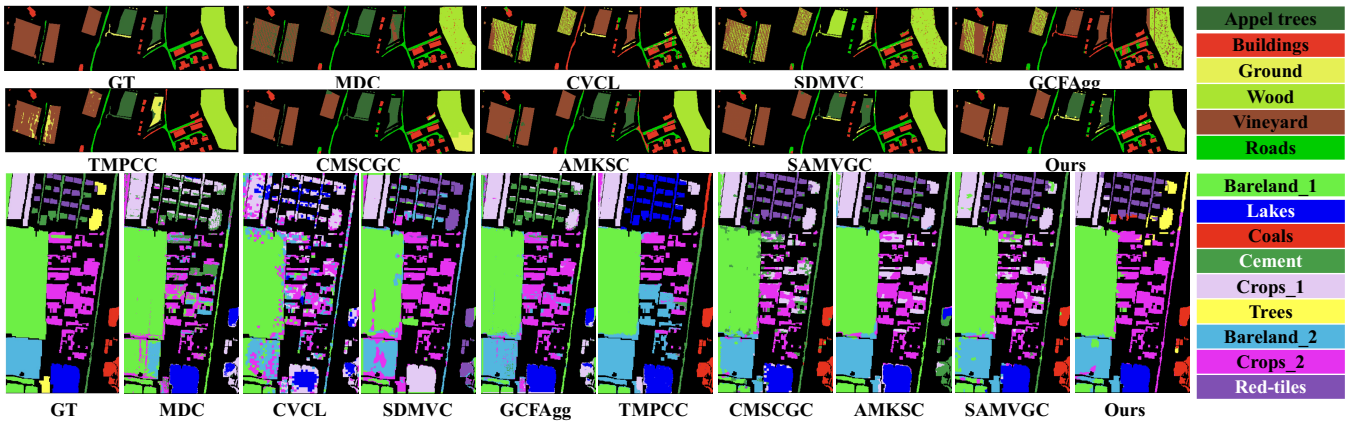


Figure 2: Clustering maps on the Trento and XuZhou datasets. GT represents the ground truth.

CVCL, SDMVC, and GCFAgg, although capable of integrating multi-modal information, often produce noisy and fragmented clustering maps. This is mainly due to their neglect of spatial contextual information, leading to scattered salt-and-pepper noise. In contrast, remote sensing-oriented models such as MDFL and TMPCC better capture spa-

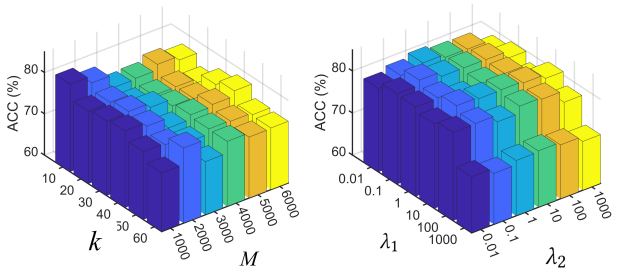


Figure 3: Sensitivity analysis with respect to hyper-parameters $k$, $M$, $\lambda_1$, and $\lambda_2$ on the XuZhou dataset.

tial dependencies, resulting in noticeably smoother clustering outputs. Moreover, graph-based methods including CM-SCGC, AMKSC, and SAMVGC are more effective in leveraging spatial structures and produce clearer results. Overall, the proposed CG-MAE achieves the most visually distinct and spatially consistent clustering maps across all datasets, demonstrating its strong capability in learning high-quality clustering representations.

## Ablation Study

To comprehensively evaluate the contribution of each module, we selectively remove specific components to assess their individual impact. The final ablation results are summarized in Table 2. Specifically, "w/o masking", "w/o $\mathcal{L}_{\mathrm{FR}}$", "w/o $\mathcal{L}_{\mathrm{ER}}$", and "w/o $\mathcal{L}_{\mathrm{KL}}$" denote CG-MAE variants without feature masking, reconstruction loss, edge refinement loss, and confused-sample consistency learning loss, respectively. As shown in the results, the ACC drops by 2.07% / 2.03% / 0.63% / 1.06%, 3.14% / 4.06% / 0.40% / 1.35%, 4.74% / 12.37% / 1.62% / 2.03%, and 2.74% / 17.36% / 0.73% / 0.96% on four datasets, respectively. These findings clearly indicate that all four components have a significant impact on the clustering performance of CG-MAE. Among them, the removal of masking and $\mathcal{L}_{\mathrm{FR}}$ leads to the most pronounced performance degradation, highlighting the importance of feature masking in promoting robust representation learning and the reconstruction loss in enhancing the model's learning capacity. Meanwhile, the edge refinement loss $\mathcal{L}_{\mathrm{ER}}$ improves graph structure quality by preserving semantically consistent edges while suppressing noisy connections, thereby enhancing feature propagation. The self-consistency loss $\mathcal{L}_{\mathrm{KL}}$ encourages alignment between the original and masked feature distributions, which stabilizes representation learning and improves clustering robustness. 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 analyze the sensitivity of hyper-parameters on the XuZhou dataset, including the number of superpixels $M$, the neighborhood size $k$, and the loss weights $\lambda_1$ and $\lambda_2$. As shown in Fig.3 (a), we fix other parameters and vary $k$ from 10 to 60 with a step of 10, and

| Datasets | Variants | ACC | Kappa | NMI | ARI | PUR |
|---|---|---|---|---|---|---|
| **Trento** | w/o masking | 93.52 | 91.41 | 89.71 | 93.2 | 95.86 |
| | w/o $\mathcal{L}_{\mathrm{FR}}$ | 93.56 | 91.37 | 87.44 | 92.71 | 94.62 |
| | w/o $\mathcal{L}_{\mathrm{ER}}$ | 94.96 | 93.29 | 90.52 | 95.72 | 95.46 |
| | w/o $\mathcal{L}_{\mathrm{KL}}$ | 94.53 | 92.76 | 90.08 | 94.48 | 95.50 |
| | Ours | **95.59** | **94.13** | **91.33** | **96.14** | **96.02** |
| **Salinas** | w/o masking | 79.60 | 77.52 | 82.14 | 69.64 | 83.75 |
| | w/o $\mathcal{L}_{\mathrm{FR}}$ | 78.68 | 76.39 | 80.39 | 65.73 | 79.33 |
| | w/o $\mathcal{L}_{\mathrm{ER}}$ | 82.34 | 80.44 | **83.28** | 75.17 | 86.52 |
| | w/o $\mathcal{L}_{\mathrm{KL}}$ | 81.39 | 79.43 | 82.18 | 74.08 | 85.20 |
| | Ours | **82.74** | **80.89** | 83.27 | **76.65** | **86.81** |
| **XuZhou** | w/o masking | 78.68 | 73.51 | 69.31 | 68.05 | 81.69 |
| | w/o $\mathcal{L}_{\mathrm{FR}}$ | 71.05 | 64.19 | 61.19 | 63.81 | 72.82 |
| | w/o $\mathcal{L}_{\mathrm{ER}}$ | 81.80 | 77.11 | 71.48 | 72.61 | 82.41 |
| | w/o $\mathcal{L}_{\mathrm{KL}}$ | 81.39 | 76.9 | 72.65 | 72.68 | 82.21 |
| | Ours | **83.42** | **79.28** | **74.26** | **73.73** | **83.91** |
| **MDSA** | w/o masking | 46.35 | 33.36 | 32.33 | 27.43 | 64.12 |
| | w/o $\mathcal{L}_{\mathrm{FR}}$ | 31.46 | 20.87 | 22.61 | 13.73 | 59.57 |
| | w/o $\mathcal{L}_{\mathrm{ER}}$ | 48.36 | 34.85 | 32.37 | 28.81 | 65.14 |
| | w/o $\mathcal{L}_{\mathrm{KL}}$ | 48.13 | 34.66 | 31.24 | 28.59 | 65.02 |
| | Ours | **49.09** | **35.23** | **33.00** | **28.98** | **65.25** |

Table 2: Ablation results of different variants.

$M$ from 1000 to 6000 with a step of 1000. The results indicate that small $k$ values may limit spatial context modeling, while large $k$ may cause over-smoothing. Similarly, too few superpixels may lose local details, whereas too many may introduce noise. Thus, a balance is required. Fig.3 (b) further shows the impact of $\lambda_1$ and $\lambda_2$ in Eq. 17, evaluated via grid search over $0.01, 0.1, 1, 10, 100, 1000$. Results reveal that improper settings can significantly affect performance, and the best results are obtained with $\lambda_1 = \lambda_2 = 0.1$.

## Conclusion

In this work, we introduced CG-MAE, a novel clustering-guided graph mask autoencoder designed for multi-view remote sensing clustering. To the best of our knowledge, this is the first attempt to adapt graph masked autoencoders to this domain. We proposed a clustering-guided masking strategy that focuses on reconstructing nodes and edges critical to clustering, encouraging the model to learn more discriminative representations. To further enhance learning stability and effectiveness, we incorporated an easy-to-hard masking curriculum and designed a dual self-adaptive learning to guide the model toward better semantic alignment. Comprehensive experiments on four benchmark datasets demonstrate that CG-MAE significantly outperforms existing state-of-the-art methods, both in terms of clustering performance and representation quality.

## Acknowledgments

This work was supported by the National Natural Science Foundation of China (NSFC) Youth Project (No. 62201604), the National Key R&D Program of China under Grant No.2022ZD0209103 and the National Natural Science Foundation of China (project no.62325604, 62276271,

62506369, 62406329, 62476280). Corresponding authors: Tianrui Liu and Miaomiao Li.

## 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.

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.

Cao, Z.; Lu, Y.; Xin, H.; Yu, C.; Wang, R.; and Nie, F. 2025a. Spatial–Spectral Bipartite Graph Clustering With Low-Frequency Tensor Regularization for Hyperspectral and LiDAR Data. *IEEE Transactions on Geoscience and Remote Sensing*, 63: 1–17.

Cao, Z.; Xin, H.; Wang, R.; and Nie, F. 2025b. Superpixel-Based Bipartite Graph Clustering Enriched With Spatial Information for Hyperspectral and LiDAR Data. *IEEE Transactions on Geoscience and Remote Sensing*, 63: 1–15.

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, 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.

Duan, H.; Yu, B.; and Xie, C. 2024. Cross-View Masked Model for Self-Supervised Graph Representation Learning. *IEEE Transactions on Artificial Intelligence*, 5540–5552.

Fang, R.; Li, B.; Zeng, Q.; Dashtbayaz, N. H.; Pu, R.; Wang, B.; Ling, C.; et al. 2025. On the Benefits of Attribute-Driven Graph Domain Adaptation. In *The Thirteenth International Conference on Learning Representations*.

Feng, Y.; Liang, W.; Wan, X.; Liu, J.; Liu, S.; Qu, Q.; Guan, R.; Xu, H.; and Liu, X. 2025. Incremental Nyström-based Multiple Kernel Clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 16613–16621.

Guan, R.; Li, J.; Wang, S.; Tu, W.; Li, M.; Zhu, E.; Liu, X.; and Chen, P. 2025a. Multi-view Graph Clustering with Dual Relation Optimization for Remote Sensing Data. In *Proceedings of the 33rd ACM International Conference on Multimedia*, 7346–7355.

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.; Liu, T.; Tu, W.; Tang, C.; Luo, W.; and Liu, X. 2025b. Sampling Enhanced Contrastive Multi-View Remote Sensing Data Clustering with Long-Short Range Information Mining. *IEEE Transactions on Knowledge and Data Engineering*, 1–15.

Guan, R.; Tu, W.; Hu, D.; Liang, W.; Liang, K.; Hu, Y.; Liu, Y.; and Liu, X. 2025c. Prototype-Driven Multi-View Attribute-Missing Graph Clustering. *IEEE Transactions on Multimedia*, 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.

Guan, R.; Tu, W.; Wang, S.; Liu, J.; Hu, D.; Tang, C.; Feng, Y.; Li, J.; Xiao, B.; and Liu, X. 2025d. Structure-Adaptive Multi-View Graph Clustering for Remote Sensing Data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 16933–16941.

Hou, Z.; Liu, X.; Cen, Y.; Dong, Y.; Yang, H.; Wang, C.; and Tang, J. 2022. Graphmae: Self-supervised masked graph autoencoders. In *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining*, 594–604.

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*.

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.

Hu, D.; Liu, S.; Wang, J.; Zhang, J.; Wang, S.; Hu, X.; Zhu, X.; Tang, C.; and Liu, X. 2024c. Reliable Attribute-missing Multi-view Clustering with Instance-level and feature-level Cooperative Imputation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 1456–1466.

Ji, J.; and Feng, S. 2025. Anchors crash tensor: efficient and scalable tensorial multi-view subspace clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

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

Li, J.; Wu, R.; Sun, W.; Chen, L.; Tian, S.; Zhu, L.; Meng, C.; Zheng, Z.; and Wang, W. 2023. What's behind the mask: Understanding masked graph modeling for graph autoencoders. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 1268–1279.

Liu, C.; Yao, Z.; Zhan, Y.; Ma, X.; Tao, D.; Wu, J.; Hu, W.; Pan, S.; and Du, B. 2024. Hi-GMAE: Hierarchical Graph Masked Autoencoders. *arXiv preprint arXiv:2405.10642*.

Liu, J.; Guan, R.; Li, Z.; Zhang, J.; Hu, Y.; and Wang, X. 2023. Adaptive Multi-Feature Fusion Graph Convolutional Network for Hyperspectral Image Classification. *Remote Sensing*, 15(23): 5483.

Liu, S.; and Chang, L. 2025. Conditional Dual Diffusion for Multimodal Clustering of Optical and SAR Images. *IEEE Transactions on Circuits and Systems for Video Technology*.

Meng, L.; Liang, K.; Xiao, B.; Zhou, S.; Liu, Y.; Liu, M.; Yang, X.; Liu, X.; and Li, J. 2024a. SARF: Aliasing relation–assisted self-supervised learning for few-shot relation reasoning. *IEEE Transactions on Neural Networks and Learning Systems*.

Meng, L.; Liang, K.; Yu, H.; Liu, Y.; Zhou, S.; Liu, M.; and Liu, X. 2024b. Fedean: Entity-aware adversarial negative sampling for federated knowledge graph reasoning. *IEEE Transactions on Knowledge and Data Engineering*.

Peng, X.; Liu, B.; Guan, R.; and Tu, W. 2025. Multi-view Graph Clustering with Dual Structure Awareness for Remote Sensing Data. In *Proceedings of the 33rd ACM International Conference on Multimedia*, 2313–2322.

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.

Shi, Y.; Dong, Y.; Tan, Q.; Li, J.; and Liu, N. 2023. Gigamae: Generalizable graph masked autoencoder via collaborative latent space reconstruction. In *Proceedings of the 32nd ACM international conference on information and knowledge management*, 2259–2269.

Tian, Y.; Zhang, C.; Kou, Z.; Liu, Z.; Zhang, X.; and Chawla, N. V. 2024. Ugmae: A unified framework for graph masked autoencoders. *arXiv preprint arXiv:2402.08023*.

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*, 10078–10086.

Wang, B.; Zeng, C.; Chen, M.; and Li, X. 2025a. Towards Learnable Anchor for Deep Multi-View Clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 21044–21052.

Wang, F.; Jin, J.; Hu, J.; Liu, S.; Yang, X.; Wang, S.; Liu, X.; and Zhu, E. 2024a. Evaluate then Cooperate: Shapley-based View Cooperation Enhancement for Multi-view Clustering. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

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

Wang, L.; Tao, X.; Liu, Q.; and Wu, S. 2024c. Rethinking graph masked autoencoders through alignment and uniformity. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 15528–15536.

Wang, S.; Liu, X.; Liao, Q.; Wen, Y.; Zhu, E.; and He, K. 2025b. Scalable multi-view graph clustering with cross-view corresponding anchor alignment. *IEEE Transactions on Knowledge and Data Engineering*.

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, B.; Dong, Z.; Liang, K.; Liu, S.; Wang, S.; Liu, T.; Hu, X.; Zhu, E.; and Liu, X. 2025. EASEMVC: Efficient Dual Selection Mechanism for Deep Multi-View Clustering. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 20716–20726.

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.; Wang, S.; Jin, J.; Wang, F.; Liu, T.; Jin, Y.; Liu, X.; Zhu, E.; and He, K. 2025. Generalized Deep Multi-view Clustering via Causal Learning with Partially Aligned Cross-view Correspondence. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

Zhang, G.; Yuan, G.; Cheng, D.; Liu, L.; Li, J.; and Zhang, S. 2025a. Mitigating propensity bias of large language models for recommender systems. *ACM Transactions on Information Systems*, 43(6): 1–26.

Zhang, G.; Zhang, S.; and Yuan, G. 2024. Bayesian graph local extrema convolution with long-tail strategy for misinformation detection. *ACM Transactions on Knowledge Discovery from Data*, 18(4): 1–21.

Zhang, Z.; Cai, Y.; Gong, W.; Liu, X.; Zeng, C.; and Yu, G. 2025b. MMAGL: Multiobjective Multiview Attributed Graph Learning for Joint Clustering of Hyperspectral and LiDAR Data. *IEEE Transactions on Geoscience and Remote Sensing*, 63: 1–14.

Zhou, T.; Dong, Z.; Wang, S.; Liang, K.; Li, M.; Liu, X.; Zhu, E.; and Dong, X. 2025. DPFMVC: Dynamic Progressive Fusion for Multi-view Clustering. In *Proceedings of the 33rd ACM International Conference on Multimedia*, 1102–1111.

Zhuo, J.; Qin, F.; Cui, C.; Fu, K.; Niu, B.; Wang, M.; Guo, Y.; Wang, C.; Wang, Z.; Cao, X.; et al. 2024. Improving graph contrastive learning via adaptive positive sampling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 23179–23187.