remote sensing

MDPI

Article

Multimodal Remote Sensing Image Clustering on Superpixel Manifolds

Shujun Liu [1,*], Yuhong Yao [1] and Luxi Xiao [2]

[1] School of Computer Engineering, Chengdu Technological University, Chengdu 611730, China; baigeil@outlook.com
[2] College of Geophysics, Chengdu University of Technology, Chengdu 610059, China; xiaoluxi3978@163.com
* Correspondence: suldier@outlook.com

Highlights

What are the main findings?

- Learning the unified cluster membership of superpixels captures complementary information of multimodal images.
- The proposed multimodal clustering method is equivalent to multiview low-rank subspace clustering.

What are the implications of the main findings?

- Learning the cluster membership graph directly on superpixel manifolds provides efficient optimization.
- The proposed method leads to strong interpretability for multimodal clustering.

Abstract

Despite offering rich complementary information, multimodal remote sensing images collected by diverse sensors increase the computational burden in clustering. To alleviate this issue, we devise an efficient multimodal clustering approach (MCSM) on superpixel manifolds formed by superpixel segmentation. The MCSM jointly learns cluster representation of all modalities and a consensus cluster membership graph that fuses the multimodal representation to yield clusters. To capture the local geometric structure of the superpixel manifolds, the optimization is constrained by manifold regularization of the consensus graph. In contrast to vanilla multiview subspace clustering techniques, the proposed approach does not rely on spectral clustering, and only involves element-wise product and multiplication on small-scale matrices. In addition, we prove that the MSCM is a special case of classic low-rank subspace clustering models, providing a perspective for understanding the learned cluster graphs. Extensive experiments are conducted on three popular multimodal remote sensing datasets, showing that the proposed method achieves competitive clustering performance compared to state-of-the-art methods, and significantly outperforms the latter in computational efficiency.

Keywords: multimodal clustering; multi-view subspace clustering; superpixel manifolds; remote sensing images

Check for updates

Academic Editor: Salah Bourennane

Received: 6 February 2026
Revised: 26 February 2026
Accepted: 18 March 2026
Published: 19 March 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

With the rapid development of remote sensing platforms and imaging sensors, the quantity and diversity of remote sensing images are becoming increasingly greater. However, due to the limitations of each sensor's imaging mechanism, single-modal images

are unable to depict comprehensive features of land covers [1,2]. For instance, hyperspectral (HS) and multispectral (MS) images capture the rich spectral reflectance of land covers [3]; Lidar data records the height information of objects; and SAR images reflect rough terrain features. Thus, fusing data and leveraging their complementary information is a promising development direction for downstream Earth observation tasks.

Due to the advantage of not requiring annotations, remote sensing image clustering has long been an important means for unsupervised land cover classification. Over the past decade, a vast number of clustering methods have emerged for HS images [4–6]. Among them, graph-based methods have attracted considerable attention. Specifically, subspace clustering is a typical graph-based method, which seeks subspace graphs through optimizing self-presentation of data constrained by sparse or low-rank norms [7–9]. To this end, ref. [10] first introduces sparse subspace clustering into the HS clustering community. After that, refs. [11,12] extend subspace clustering from a deep learning perspective, and ref. [13] applies a kernel trick to map HS data into the high-dimensional space, leading to the kernel counterpart. Instead of self-representation, ref. [14] learns probability coupling as a graph using optimal transport. Recently, in contrast to subspace clustering, refs. [15,16] directly optimize graph embedding by using a graph low-pass filter and graph adaptive filters on data. However, these approaches usually rely on additional clustering tools, such as spectral clustering and K-means, to segment graphs to yield resulting clusters. Although these single-modal methods attempt to fully exploit the spatial–spectral information of images [17,18], they still struggle to overcome the limitations of single modalities.

Multimodal clustering, equipped with fusion and clustering processes [19,20], essentially corresponds to multiview clustering (MVC) by treating each data modality as a separate view [21,22]. The seminal MVC approach introduced in [23] extended the classical single-view K-means algorithm to a multiview environment, thereby laying the groundwork for later multimodal clustering techniques. Following this, a variety of machine learning methods have been extended to their multiview versions, including subspace clustering [24], spectral clustering [25], and non-negative matrix factorization [26]. Of these, multiview subspace clustering (MVSC) has gained considerable attention because of its interpretability and robustness. MVSC approaches consider different views as related sets of subspaces and jointly optimize self-expression matrices from all views under specific norm constraints, so that each matrix reflects a unique subspace. These individual representations are then combined into an affinity matrix for spectral clustering (SC). A limitation of standard SC, however, is its dependence on a square matrix as input, which hinders its use on large-scale data. To mitigate this, ref. [27] implements MVSC on a set of representative anchors instead of the full dataset, resulting in a smaller anchor graph. Ref. [28] further lowers the computational expense by incorporating unified anchors into MVSC using multiview similarity measures. In contrast to earlier techniques that treated anchor selection and anchor graph construction as separate steps, more recent methods [29,30] unify the learning of anchors and the anchor graph within a single MVSC model. For handling high-dimensional and imbalanced data, fast MVSC [31] simultaneously learns anchors, anchor graphs, and cluster labels in a reduced embedding space. Along similar lines, ref. [32] presents a scalable MVSC method that also learns unified anchors and an anchor graph in an embedding space, with the anchor graph directly generating the cluster assignment matrix. Furthermore, ref. [33] advances MVSC by learning a unified bipartite graph subject to sparse and low-rank constraints, rather than learning a unified anchor graph. In general, the multiview-clustering-based methods fuse the multimodal information of data by integrating their individual clustering graphs into a unified one.

Despite their success in machine learning, relatively few studies have specifically addressed multimodal clustering for remote sensing imagery [34]. Recently, to enhance

sampling for contrastive learning, ref. [35] leveraged autoencoders on superpixels to capture short-range features and graph autoencoders on pixel patches to model long-range dependencies in multimodal data. Ref. [36] proposes a conditional dual diffusion model for optical and SAR image clustering. The method introduces a decoupling autoencoder conditioned on SAR images to model the noise distribution and reconstruct the original optical images, thereby preserving the underlying data manifold structure. Subsequently, the multimodal features learned by the autoencoder are utilized for clustering via K-means. Unlike the multiview-clustering-based methods, this approach directly fuses multimodal data to generate complementary features instead of clustering graphs. Using kernel embedding, ref. [37] maps both samples and anchors into high-dimensional kernel spaces and constructs an MVSC model to learn a unified anchor graph. Ref. [38] applies low-pass graph filtering to modality-specific attributed graphs and then solves for a fused graph under constraints imposed by four multi-objective clustering subproblems. Ref. [39] performs clustering by explicitly modeling sub-pixel information from mixed spectra to endmembers. Although these anchor-based methods have demonstrated promising performance, they commonly face the challenge of determining the optimal number of anchors, which remains a nontrivial design parameter. In addition, these approaches require spectral clustering or K-means to split the learned features into individual clusters, resulting in additional computational costs.

To eliminate the additional classifiers, recently, learning soft labels has become promising. Ref. [40] employs a multimodal transformer trained with contrastive and self-supervised losses to learn a discriminative shared representation, where stabilized prototypes serve as soft clustering labels. This approach, however, requires extensive iterative training and is sensitive to the choice of image-patch size. Ref. [41] first applies superpixel segmentation to multimodal data to generate anchors, then learns dual consensus anchor graphs from which clustering labels are derived. Similarly, based on anchors obtained via superpixel segmentation, ref. [42] simultaneously optimizes two fuzzy clustering objectives that, respectively, project samples to anchors and anchors to clusters, with the final label matrix obtained by multiplying the two projection matrices. Ref. [43] conducts superpixel segmentation and nonlinear neighborhood recovery on each modality to produce modality-specific spatial-aware anchors, which are then used to learn multi-scale spectral–spatial anchor graphs.

Inspired by the previous discussion, in this article, we aim to learn a consensus membership graph to indicate clusters on superpixel manifolds for multimodal remote sensing images. To do so, we apply superpixel segmentation of the SLIC algorithm [44] to multimodal data, forming superpixel manifolds. In contrast to K-means, SLIC adopts a local search strategy instead of a global search to cluster similar points, such that similar neighboring points tend to be grouped into a region as a superpixel. It is particularly suitable for remote sensing images because of their local spatial similarity, leading to fewer modeling units for the subsequent clustering process. For each modality, we employ an individual membership matrix to map associated manifolds to cluster centers, preserving the characteristics of each modality. Then, all the cluster centers are recovered to the superpixel manifolds via a unified cluster membership graph that captures the complementarity of multimodal data. To leverage the local geometric information of manifolds, the cluster membership matrix is constrained by a manifold regularization term. The proposed model is performed on superpixel manifolds without additional classifiers, which thus leads to efficient optimization. In addition, we provide a connection to low-rank subspace clustering for another perspective to understand the proposed model.In summary, the core contributions of this study are presented as the following highlights:

1. We devise an efficient mulimodal clustering method on superpixel manifolds by learning a consensus cluster membership graph.
2. The relationship between the proposed model and low-rank subspace clustering is given for a theoretical explanation.
3. Extensive experiments conducted on three datasets demonstrate the effectiveness and efficiency of the proposed approach.

The remainder of this paper is structured as follows. Section 2 introduces the preliminary background. Section 3 presents the proposed MCSM method. The optimization and clustering of the proposed model are provided in Section 4. Experiments and conclusions are described in Sections 5 and 6, respectively.

2. Preliminaries

2.1. Notations

Throughout this article, we write matrices in bold upper-case letters or bold upper-case Greek letters (e.g., $\mathbf{A}$, $\mathbf{\Phi}$), vectors in bold lower-case letters or bold lower-case Greek letters (e.g., $\mathbf{a}$, $\boldsymbol{\gamma}$), and scalars in letters or lower-case Greek letters (e.g., n, V, λ). In what follows, for a matrix $\mathbf{M}$, $rank(\mathbf{M})$ and $tr(\mathbf{M})$ read the rank and the trace of $\mathbf{M}$, respectively; $(\mathbf{M})_{ij}$ denotes its element; and $\|\mathbf{M}\|_F$ is the Frobenius norm of $\mathbf{M}$.

2.2. Multiview Subspace Clustering

Given data $\mathbf{X} \in \mathbb{R}^{D\times N}$ with samples of N and dimension of D lying in a union space of K subspaces, subspace clustering aims to split $\mathbf{X}$ into C subsets that span individual subspaces, where C indicates the number of clusters. Specifically, it pursues subspace self-representation $\mathbf{C} \in \mathbb{R}^{N\times N}$ of $\mathbf{X}$ by minimizing

$$\min_{\mathbf{C}} \|\mathbf{X} - \mathbf{X}\mathbf{C}\|_F^2 + \phi(\mathbf{C}), \tag{1}$$

where the regularization term $\phi(\mathbf{C})$ controls the property of $\mathbf{C}$ according to prior information of the data. For example, setting $\phi(\cdot) = \|\cdot\|_1$, Equation (1) becomes sparse subspace clustering [45], while it turns into low-rank subspace clustering [46] or low-rank representation [47] if $\phi(\cdot) = \|\cdot\|_*$.

Multiview subspace clustering (MVSC) extends the vanilla subspace clustering into a multimodal counterpart for fitting multimodal data, when regarding each modality as an individual view. Besides learning consensus representation $\mathbf{C}$ of all modalities, the MVSC methods also solve individual subspace representation $\mathbf{C}_v$ for each modality, which boils down to the following paradigm:

$$\min_{\mathbf{C}_v,\mathbf{C}} \sum_{v=1}^{V} \|\mathbf{X}_v - \mathbf{X}_v\mathbf{C}_v\|_F^2 + \phi(\mathbf{C}_v) + \psi(\mathbf{C}_v, \mathbf{C}), \tag{2}$$

where V is the number of modalities, and $\phi(\mathbf{C}_v, \mathbf{C})$ stands for the consensus regularization term that fuses the multimodal representation into a unified graph. However, solving Equation (2) requires heavy computational consumption, when processing the square matrix $\mathbf{C}_v$ with a large size. To alleviate this issue, anchor-based MVSC methods [32] employ anchors $\mathbf{A}_v \in \mathbb{R}^{D\times K}$ and anchor graph $\mathbf{Z}_v \in \mathbb{R}^{K\times N}$ to recover $\mathbf{X}_v$ by $\mathbf{A}_v\mathbf{Z}_v$ instead of $\mathbf{X}_v\mathbf{C}_v$. As a result, feeding the consensus graph $\mathbf{C}$ or $\mathbf{Z}$ to spectral clustering algorithms yields clusters [48].

3. Methodology

In this section, we first present an efficient multimodal clustering method that directly learns a unified membership graph as a cluster indicator on superpixel manifolds. Then,

we establish a connection between the proposed model and low-rank subspace clustering. The framework of the proposed method is illustrated in Figure 1.

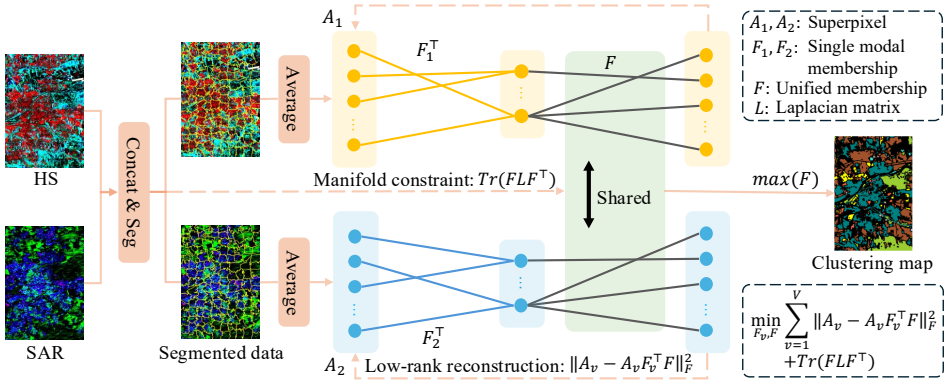


Figure 1. Framework of the proposed MCSM method for mulitimodal clustering. Averaging the segmented data by superpixel segmentation yields superpixels. Following the superpixel manifold structure, MCSM jointly optimizes cluster membership graphs of all modalities and the consensus membership graph that indicates clustering results.

3.1. Unified Membership Graph

Multiview subspace clustering methods and their anchor counterparts are dedicated to seeking sample graphs and anchor graphs. However, these graphs require a spectral process, resulting in additional computational costs. To address this issue, we aim to learn a unified membership graph of multimodal data as a label indicator. To this end, for each modality of $\mathbf{X}_v \in \mathbb{R}^{D\times N}, v = 1, \ldots, V$ with N samples, we introduce label projection $\mathbf{F}_v \in \mathbb{R}^{C\times N}$ to map $\mathbf{X}_v$ into individual label space with C cluster centers, i.e., $\mathbf{X}_v\mathbf{F}_v^\top$, preserving the diversity of multimodal data. Then, $\mathbf{X}_v$ is recovered from the label space via a unified membership graph $\mathbf{F} \in \mathbb{R}^{C\times N}$ that indicates consistent clustering results, i.e., $\mathbf{X}_v\mathbf{F}_v^\top\mathbf{F}$. Formally, the unified membership graph is achieved by minimizing

$$\begin{aligned} &\min_{\mathbf{F}_v,\mathbf{F}} \sum_{v=1}^{V} \|\mathbf{X}_v - \mathbf{X}_v\mathbf{F}_v^\top\mathbf{F}\|_F^2 \\ &s.t. \quad (\mathbf{F})_{ij}, (\mathbf{F}_v)_{ij} \geq 0, \mathbf{F}_v^\top\mathbf{1} = \mathbf{1}, \mathbf{F}^\top\mathbf{1} = \mathbf{1}, \end{aligned} \tag{3}$$

where the constraints ensure $\mathbf{F}_v$ and $\mathbf{F}$ are non-negative and are the probability matrices. Note that $\mathbf{F}_v$ is the cluster indicator of single modality, but $\mathbf{F}$ serves as a unified indicator that captures the complementarity of the multimodal data.

However, optimizing Equation (3) to achieve the best cluster indicator remains challenging due to the large feasible solution space. Thus, we apply the fuzzy c-means (FCM) [49] algorithm to each $\mathbf{X}_v$ and their concatenation to extract soft labels, which are used to initialize $\mathbf{F}_v$ and $\mathbf{F}$, respectively. Under the guidance of the initialization, the optimization pursues the optimal solution in the direction of the cluster semantic.

3.2. Multimodal Clustering on Superpixel Manifolds

Suffering from large-scale images, the membership graph $\mathbf{F}$ in Equation (3) is endowed with a big size, resulting in $N \gg C$, which leads to poor clustering performance as well as low optimizing efficiency. Hence, we adopt superpixels as modeling units instead of original sample pixels. To do so, the superpixel segmentation algorithm of SLIC is adopted to split images into many small regions with similar features. Each region is averaged as a superpixel $\mathbf{A} \in \mathbb{R}^{D\times K}$, where K denotes the number of the superpixels. The superpixel data form new low-dimensional manifolds but smaller samples. To capture the local geometric

structure of the superpixel manifolds, we introduce the manifold regularization on the membership graph as follows:

$$\min_{\mathbf{F}} tr(\mathbf{F}\mathbf{L}\mathbf{F}^\top), \tag{4}$$

where $\mathbf{L} = \mathbf{D} - \mathbf{M}$ denotes the Laplacian matrix; $\mathbf{M}$ is the affinity matrix calculated by RBF kernel $(\mathbf{M})_{ij} = e^{\frac{-\|\mathbf{a}_i - \mathbf{a}_j\|_F^2}{\sigma^2}}$ on the concatenation of the multimodal superpixels of $\{\mathbf{A}_v\}, v = 1, \ldots, V$, denoted as $\mathbf{A}$; and $\mathbf{D}$ is a diagonal matrix with $diag(\mathbf{D}) = \mathbf{M}\mathbf{1}$.

Substituting $\mathbf{X}_v = \mathbf{A}_v$ and adding Equation (4) with a weight λ into Equation (3), we have

$$\begin{aligned} &\min_{\mathbf{F}_v, \mathbf{F}} \sum_{v=1}^{V} \|\mathbf{A}_v - \mathbf{A}_v \mathbf{F}_v^\top \mathbf{F}\|_F^2 + \lambda tr(\mathbf{F}\mathbf{L}\mathbf{F}^\top) \\ &s.t. \quad (\mathbf{F})_{ij}, (\mathbf{F}_v)_{ij} \geq 0, \mathbf{F}_v^\top \mathbf{1} = \mathbf{1}, \mathbf{F}^\top \mathbf{1} = \mathbf{1}. \end{aligned} \tag{5}$$

Accordingly, $\mathbf{F}_v \in \mathbb{R}^{C \times K}$ and $\mathbf{F} \in \mathbb{R}^{C \times K}$ are initialized by applying the FCM algorithm to the superpixel data of $\mathbf{A}_v$ and their concatenation of $\{\mathbf{A}_v\}$. The manifold regularization enforces the superpixel reconstruction term to yield clusters with a manifold structure. Since $K \ll N$, solving Equation (5) is very fast, and will be discussed in detail in Section 4.

3.3. Connection to Low-Rank Subspace Clustering

We here build a bridge from the proposed MCSM model to the classic low-rank subspace clustering, which learns low-rank self-representation of data through subspace recovery and ℓ_* constraint, inspired by [50]. Their relationship provides a new perspective to understand the MCSM model following Theorem 1.

Theorem 1. *The proposed Equation (3) is equivalent to low-rank subspace clustering.*

Proof. Let $\mathbf{C}_v \in \mathbb{R}^{K \times K} = \mathbf{F}_v^\top \mathbf{F}$. Substituting it into Equation (3) yields

$$\min_{\mathbf{C}_v} \sum_{v=1}^{V} \|\mathbf{A}_v - \mathbf{A}_v \mathbf{C}_v\|_F^2, \tag{6}$$

which ignores the condition and regularization terms. Apparently, we have

$$rank(\mathbf{C}_v) \leq min\{rank(\mathbf{F}_v), rank(\mathbf{F})\} \leq C. \tag{7}$$

Since the number of clusters C is much smaller than that of superpixels k (i.e., $C \ll K$), $\mathbf{C}_v$ is a low-rank matrix. Thus, Equation (6) exactly satisfies the optimization objective of low-rank subspace clustering, as follows:

$$\min_{\mathbf{C}_v} \sum_{v=1}^{V} \|\mathbf{A}_v - \mathbf{A}_v \mathbf{C}_v\|_F^2 + \lambda \|\mathbf{C}_v\|_*. \tag{8}$$

Comparing Equations (6) and (8), we know that $\mathbf{F}_v^\top \mathbf{F}$ is a low-rank self-representation of $\mathbf{A}_v$ with subspace structure. □

Remark 1. *For a large number K, the proposed model of Equation (6) requires much less computation than the low-rank subspace clustering of Equation (8), owing to the matrix factorization* $\mathbf{C}_v = \mathbf{F}_v^\top \mathbf{F}$, *as well as avoiding the singular value decomposition (SVD) of* $\mathbf{C}_v$ *in solving the latter one.*

In contrast, based on Theorem 1, we can build the affinity matrix, with $\mathbf{B} = |\mathbf{F}^\top \mathbf{F}|$ showing natural symmetry. Similar to subspace clustering, the resulting clusters could be

achieved by applying spectral clustering to $\mathbf{B}$, but the clustering efficiency is much lower than directly extracting clusters from the indicator $\mathbf{F}$.

4. Optimization and Clustering

4.1. Optimization

Considering multiple variables and constraints, solving the problem of Equation (5) is challenging for close solution. Therefore, we adopt an alternating optimization algorithm to address this problem, updating each variable while fixing the others. The objective function is first reformulated into a trace form to facilitate derivation. For the sake of clarity, we define $\mathbf{C}_v = \mathbf{F}_v^\top \mathbf{F}$ and $\mathbf{S}_v = \mathbf{A}_v^\top \mathbf{A}_v$. By expanding the Frobenius norm, the loss function Equation (5) can be rewritten as

$$\begin{aligned} \mathcal{L} &= \sum_{v=1}^{V} (tr(\mathbf{A}_v - \mathbf{A}_v\mathbf{C}_v)^\top (\mathbf{A}_v - \mathbf{A}_v\mathbf{C}_v)) + \lambda tr(\mathbf{F}\mathbf{L}\mathbf{F}^\top) \\ &= \sum_{v=1}^{V} (tr(\mathbf{S}_v) - 2tr(\mathbf{S}_v\mathbf{C}_v) + tr(\mathbf{C}_v^\top \mathbf{S}_v\mathbf{C}_v)) + \lambda tr(\mathbf{F}\mathbf{L}\mathbf{F}^\top). \end{aligned} \tag{9}$$

(1) Updating $\mathbf{F}_v$: After fixing $\mathbf{F}$, rewriting Equation (9) yields

$$\mathcal{L}(\mathbf{F}_v) = tr(\mathbf{F}^\top \mathbf{F}_v \mathbf{S}_v \mathbf{F}_v^\top \mathbf{F}) - 2tr(\mathbf{S}_v \mathbf{F}_v^\top \mathbf{F}). \tag{10}$$

Let $\mathbf{P} = \mathbf{F}\mathbf{F}^\top$. To facilitate differentiation, we reformulate Equation (10) based on the properties of the trace as

$$\mathcal{L}(\mathbf{F}_v) = tr(\mathbf{F}_v \mathbf{S}_v \mathbf{F}_v^\top \mathbf{P}) - 2tr(\mathbf{S}_v \mathbf{F}_v^\top \mathbf{F}). \tag{11}$$

We further introduce the constraint $(\mathbf{F}_v)_{ij} \geq 0$ on $\mathbf{F}_v$ into Equation (11). To this end, we construct a Lagrangian function as follows:

$$\mathcal{L}(\mathbf{F}_v, \mathbf{\Phi}) = tr(\mathbf{F}_v \mathbf{S}_v \mathbf{F}_v^\top \mathbf{P}) - 2tr(\mathbf{S}_v \mathbf{F}_v^\top \mathbf{F}) + tr(\mathbf{\Phi}^\top \mathbf{F}_v), \tag{12}$$

where $\mathbf{\Phi} \in \mathbb{R}^{C \times N}, (\mathbf{\Phi})_{im} \geq 0$, denotes the Lagrange multiplier corresponding to the non-negativity constraint. Differentiating Equation (12) with respect to $\mathbf{F}_v$, we have

$$\frac{\partial \mathcal{L}(\mathbf{F}_v, \mathbf{\Phi})}{\partial \mathbf{F}_v} = -2\mathbf{F}\mathbf{S}_v + 2\mathbf{P}\mathbf{F}_v\mathbf{S}_v + \mathbf{\Phi}. \tag{13}$$

Combining the Karush–Kuhn–Tucker (KKT) conditions with Equation (13), we obtain

$$\begin{cases} \dfrac{\partial \mathcal{L}(\mathbf{F}_v, \mathbf{\Phi})}{\partial \mathbf{F}_v} = \mathbf{0} \\ \left(\dfrac{\partial \mathcal{L}(\mathbf{F}_v, \mathbf{\Phi})}{\partial \mathbf{F}_v} \right)_{ij} (\mathbf{F}_v)_{ij} = 0. \end{cases} \tag{14}$$

By adopting the multiplicative update strategy of non-negative matrix factorization (NMF) [51], we adopt an element-wise adaptive learning rate defined as $\eta = \frac{\mathbf{F}_v}{2\mathbf{P}\mathbf{F}_v\mathbf{S}_v}$, and the gradient descent-base update rule for $\mathbf{F}_v$ is given by

$$\mathbf{F}_v \longleftarrow \mathbf{F}_v - \eta \frac{\partial \mathcal{L}(\mathbf{F}_v, \mathbf{\Phi})}{\partial \mathbf{F}_v} = \mathbf{F}_v \odot \frac{\mathbf{F}\mathbf{S}_v}{\mathbf{P}\mathbf{F}_v\mathbf{S}_v}, \tag{15}$$

where $\odot$ denotes the Hadamard product, i.e., the element-wise multiplication operation, and the fraction represents the element-wise division operation. To ensure that the sum-to-

one constraint ($\mathbf{F}_v^\top \mathbf{1} = \mathbf{1}$) is satisfied, we further perform normalization on each row of the matrix, as follows:

$$(\mathbf{F}_v)_{:,j} \longleftarrow \frac{(\mathbf{F}_v)_{:,j}}{\sum_i (\mathbf{F}_v)_{i,j}}, \forall j = 1, 2, \cdots, K. \tag{16}$$

(2) Updating $\mathbf{F}$: For clarity, we define $\mathbf{Q}_v = \mathbf{F}_v \mathbf{S}_v \mathbf{F}_v^\top$. Then, Equation (9) can be rewritten as

$$\mathcal{L}(\mathbf{F}) = tr(\mathbf{F}^\top \mathbf{Q}_v \mathbf{F}) - 2tr(\mathbf{S}_v \mathbf{F}_v^\top \mathbf{F}) + \lambda tr(\mathbf{F}\mathbf{L}\mathbf{F}^\top). \tag{17}$$

To deal with $(\mathbf{F})_{ij} \geq 0$, we introduce a Lagrangian function as follows:

$$\begin{aligned} \mathcal{L}(\mathbf{F}, \mathbf{\Psi}) =& tr(\mathbf{F}^\top \mathbf{Q}_v \mathbf{F}) - 2tr(\mathbf{S}_v \mathbf{F}^\top \mathbf{F}) + \lambda tr(\mathbf{F}\mathbf{L}\mathbf{F}^\top) \\ &+ tr(\mathbf{\Psi}^\top \mathbf{F}), \end{aligned} \tag{18}$$

where $\mathbf{\Psi} \in \mathbb{R}^{C \times N}, \mathbf{\Psi} \geq 0$ denotes the Lagrange multiplier corresponding to the non-negativity constraint. We then differentiate Equation (18) with respect to $\mathbf{F}$, resulting in

$$\frac{\partial \mathcal{L}(\mathbf{F}, \mathbf{\Psi})}{\partial \mathbf{F}} = 2\mathbf{Q}_v \mathbf{F} + 2\lambda \mathbf{F}\mathbf{L} - 2\mathbf{F}_v \mathbf{S}_v^\top + \mathbf{\Psi}. \tag{19}$$

According to the KKT conditions, we have

$$\begin{cases} \dfrac{\partial \mathcal{L}(\mathbf{F}, \mathbf{\Psi})}{\partial \mathbf{F}} = \mathbf{0} \\ \left(\dfrac{\partial \mathcal{L}(\mathbf{F}, \mathbf{\Psi})}{\partial \mathbf{F}} \right)_{ij} \mathbf{F}_{ij} = 0. \end{cases} \tag{20}$$

Similar to Equation (16), we get the update formula for $\mathbf{F}$:

$$\mathbf{F} \longleftarrow \mathbf{F} \odot \frac{\mathbf{F}_v \mathbf{S}_v^\top}{\mathbf{Q}_v \mathbf{F} + \lambda \mathbf{F}\mathbf{L}}. \tag{21}$$

To ensure the sum-to-one constraint ($\mathbf{F}^\top \mathbf{1} = \mathbf{1}$) holds, we further normalize each column of $\mathbf{F}$ as follows:

$$(\mathbf{F})_{:,j} \longleftarrow \frac{(\mathbf{F})_{:,j}}{\sum_i (\mathbf{F})_{i,j}}, \forall j = 1, 2, \cdots, K. \tag{22}$$

By iteratively updating $\mathbf{F}_v$ and $\mathbf{F}$ using Equation (16) and Equation (22), respectively, we can efficiently find a set of local optimal solutions for the MCSM method.

4.2. Complexity Analysis

The proposed method benefits from the superpixel strategy and efficient cluster membership learning. The superpixels represent regions composed of a group of adjacent pixels, thereby achieving low complexity. In the optimization phase, the computational complexity mainly stems from variable updating. The intermediate variable $\mathbf{S}$ introduced by us incurs a time complexity of $\mathcal{O}\left(\sum_{v=1}^{V} D_v K^2\right)$, where V denotes the number of modalities, D_v denotes the dimension of the v-th modality, and K is the number of superpixels; updating $\mathbf{F}_v$ has a time complexity of $\mathcal{O}\left(VCK^2 + VC^2K\right)$, where C stands for the number of clusters; and the time complexity required for updating $\mathbf{F}$ is $\mathcal{O}\left(VCK^2 + VC^2K + CK\right)$. The overall time complexity is $\mathcal{O}\left(TVCK^2 + TVC^2K + TCK + \sum_{v=1}^{V} TD_vK^2\right)$, with T representing the number of iterations. Since $C \ll K$ holds in the proposed model, the overall time complexity can be simplified to $\mathcal{O}\left((C + \sum_{v=1}^{V} D_v)K^2\right)$. It can be observed that the time complexity of MCSM is almost linearly related to K.

4.3. MCSM for Multimodal Clustering

For multimodal remote sensing images $\mathbf{X}_v, v = 1, \cdots, V$, we first apply the superpixel segmentation algorithm to the concatenation of $\{\mathbf{X}\}_{v=1}^{V}$, and obtain locally similar regions for each modality. Then, we average the segmented $\mathbf{X}_v$ into superpixels $\mathbf{A}_v$, and calculate the Laplacian matrix $\mathbf{L}$ on the concatenation. To eliminate redundant spectral information, hyperspectral superpixels are further compressed into D-dimensional manifolds by using the principal component analysis (PCA) algorithm. Lastly, $\mathbf{F}_v$ and $\mathbf{F}$ are initialized, as discussed in Section 3.2, and, referring to Section 4.1, the MCSM approach optimizes Equation (5) to yield the consensus membership graph $\mathbf{F}$ that indicates the resulting clusters. In summary, the complete algorithm of MCSM is presented in Algorithm 1.

Algorithm 1: MCSM for multimodal clustering

Input: Multimodal data $\mathbf{X}_v, v = 1, \ldots, V$;
The number of superpixel segmentation, K;
The number of Land cover categories, C;
Manifold regularization coefficient, λ;
The dimension of PCA, D.
Initialize: $\mathbf{A}_v = SLIC(\mathbf{X}_v)$; $\mathbf{A}_v = PCA(\mathbf{A}_v)$ for HS; $\mathbf{F}_v = FCM(\mathbf{A}_v)$; $\mathbf{F} = FCM(concat(\{\mathbf{A}_v\}_1^V))$;
while *not converged* **do**
  for $v = 1$ **to** V **do**
    Updating $\mathbf{F}_v$ by Equation (16);
  end
  Updating $\mathbf{F}$ by Equation (22);
end
Output: Clustering results achieved by $\max(\mathbf{F})$.

5. Experiments

5.1. Experimental Settings

5.1.1. Multimodal Datasets

In the experiments, we adopt three multimodal remote sensing datasets, including Trento, MUUFL, and Augsburg. They are captured by several sensors over a rural area of the south of Trento, Italy, an urban area of Augsburg, Germany, and the University of Southern Mississippi, Long Beach, USA, respectively. The detailed imaging information is summarized in Table 1. Specifically, the Augsburg data is equipped with large spectral bands of HS and image size, while the MUUFL contains numerous categories of land covers, leading to potential challenges for clustering tasks.

Table 1. Parameters of the multimodal datasets and their sensors.

Property	Trento		MUUFL		Augsburg	
	HS	**LiDAR**	**HS**	**LiDAR**	**HS**	**SAR**
Sensors	AISA Eagle	ALTM 3100EA	CASI-1500	Gemini ALTM	HySpex	DLR3-3K
Wavelengths (μm)	0.4–0.98	–	0.375–1.05	–	0.4–2.5	–
Bands	63	2	64	2	180	4
Sizes	166 × 600		325 × 220		332 × 485	
Classes	6		11		7	

5.1.2. Comparison Methods

To verify model ability, we consider 12 competing methods related to our proposed method from the perspective of the clustering mechanism and state of the art (SOTA):

- BMVC [52] learns collaborative discrete representations and a binary clustering structure for binary multiview clustering.
- SMVSC [28] achieves scalable multiview subspace clustering through the use of unified anchors.
- LMVSC [27] handles large-scale multiview subspace clustering by employing anchor graphs as a substitute for full graphs.
- MDC [53] utilizes autoencoders to learn latent features for multisensor clustering.
- FPMVS [30] performs joint optimization of anchor selection and anchor graph construction in multiview subspace clustering.
- FMVACC [54] conducts multiview clustering by matching anchors across views.
- MSGL [55] attains scalable subspace clustering via the learning of a structured graph.
- SDMVC [56] employs deep self-supervised feature learning for multiview clustering.
- GCFagg [57] advances multiview clustering by fusing cross-sample and cross-view features.
- FPFC [42] formulates a dual multimodal fuzzy clustering framework based on sample–anchor and anchor–cluster associations.
- AMKSC [37] combines unified anchor learning and kernel techniques for multimodal subspace clustering.
- CDD [36] preserves data manifolds in multimodal clustering using a diffusion-based approach.

5.1.3. Evaluation Criteria

To gauge clustering performance, five widely used metrics were employed: accuracy (ACC), Kappa coefficient (Kappa), normalized mutual information (NMI), adjusted Rand index (ARI), and purity (Purity). Here, ACC, NMI, and Purity fall within the range of $[0, 1]$, while Kappa and ARI lie in $[-1, 1]$. These metrics are computed based on the optimal alignment of predicted labels with the ground truth, with higher values indicating better clustering performance. Additionally, the runtime of each clustering method was recorded in seconds (s) to assess computational efficiency.

5.1.4. Implementation Details

To ensure a fair comparison, all methods are executed using pixel-level features. For the TMPCC AMKSC method, the patch size is set to the minimum feasible value of 3 to mitigate the influence of patch-based feature extraction. The results for SDMVC and GCFAgg are taken directly from [35]. A consistent random seed of 42 is used across all experiments. The remaining parameters of all comparison methods are determined either according to the original references or through a systematic grid search. Accordingly, for the proposed MCSM method, the number of superpixels is set to 100, 400, and 150 on the Trento, MUUFL, and Augsburg datasets, respectively; we set D, the dimension of PCA, to be 9, 6 and 1, correspondingly; and λ is set to 1 for all datasets. All tests were conducted on an Ubuntu server with a Xeon Gold 5218 CPU, an RTX 2080Ti GPU, and 128 GB of RAM.

5.2. Clustering Results

5.2.1. Quantitative Performance

To evaluate multimodal clustering performance, Table 2 presents quantitative results in terms of 5 metrics, comparing the proposed MCSM method to the other methods on the Trento, MUUFL, and Augsburg datasets, respectively. The best values and the second best values are in bold and are underlined, respectively, and "N/A" means unavailable

scores. We see that the proposed MCSM consistently achieves superior or highly competitive performance across all datasets. On the Trento dataset, MCSM attains the best scores in ACC (0.896), Kappa (0.8611), NMI (0.8202), ARI (0.8976), and Purity (0.901), and significantly outperforms multiview subspace clustering methods such as SMVSC, LMVSC, and AMKSC, and the recent SOTA methods of FPFC and CDD. It has validated the effectiveness of the proposed MCSM. More notably, MCSM completes clustering in merely 0.55 s, which is an order of magnitude faster than most competitors. Similar trends are observed on the MUUFL dataset, where MCSM obtains the highest ACC (0.6554), Kappa (0.5679), and ARI (0.427), and competitive NMI and Purity. Its runtime (0.87 s) remains the lowest among all methods. For the Augsburg dataset, MCSM gains the second best scores on all metrics, achieved within 1.62 s—far more efficient than other high-performing methods such as FPFC and AMKSC.

5.2.2. Qualitative Performance

To examine the clustering characteristics of the proposed MCSM, we present cluster maps of ten comparative methods along with ground truth (GT) labels in Figures 2–4. These maps correspond to the clustering outcomes reported in Table 2. Unlabeled samples are included in the clustering process and are therefore forcibly assigned to annotated categories. As observed in the figures, the cluster maps produced by MCSM are visually closest to the GT labels across all three datasets, confirming the effectiveness of clustering on superpixel manifolds. In comparison with recent SOTA methods such as FPFC and AMKSC, MCSM yields smoother clustering maps that better preserve the local spatial consistency of land covers. From Figure 4k, we find that MCSM tends to produce large block maps, resulting in misclassification on boundaries of land covers. This is because the superpixel segmentation algorithm that the proposed MCSM is based on finds it difficult to handle interlaced land covers. According to Figure 2k, because MCSM relies on the superpixel segmentation that trends to produce contiguous regions, it fails to separate multiple subdivisions within a single region. Increasing the number of superpixels forces the segmentation algorithm to yield smaller segmented regions, potentially solving the above issues.

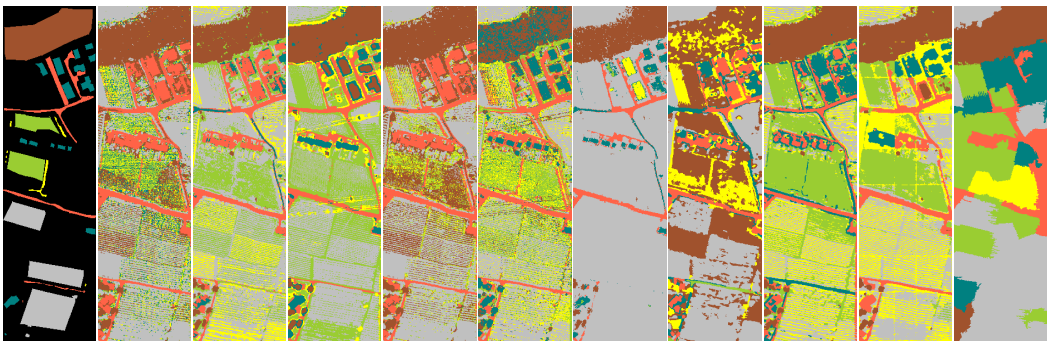
5.3. Ablation Study

5.3.1. Superpixel Manifolds

As discussed in Section 3.2, superpixels form data manifolds on which the proposed MCSM seeks the optimal membership matrix $\mathbf{F}$. The optimization is achieved by introducing manifold regularization. We now explore the effect of the regularization. To do so, ablation results of the superpixel manifolds are provided in Table 3, wherein MCSM-M denotes the MCSM model without the manifold regularization term. We observe from the table that MCSM achieves gains in ACC on the order of 48% and 53% over MCSM-M for MUUFL and Augsburg datasets, respectively, but obtain a slight improvement by absolute 0.049 ACC against the non-regularization counterpart for the Trento dataset. That is because superpixels of the Trento image tend to form a simpler manifold structure, which reduces impact of the regularization.

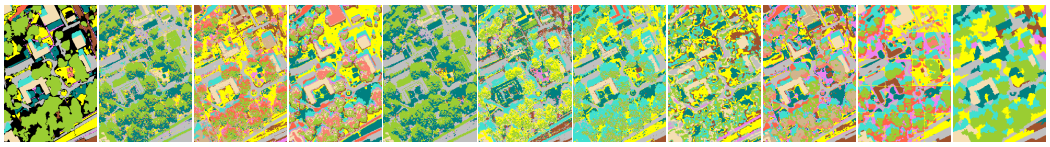
Table 2. Quantitative evaluation of fourteen multimodal clustering methods on the Trento, Augsburg, and MUUFL datasets in terms of the OA, Kappa, NMI, ARI, Purity, and time consumption metrics.

Datasets	Metrics	Methods												
		BMVC	**SMVSC**	**LMVSC**	**MDC**	**FPMVS**	**FMVACC**	**MSGL**	**SDMVC**	**GCFAgg**	**FPFC**	**AMKSC**	**CDD**	**MCSM**
Trento	ACC	0.2293	0.6749	0.6566	0.7788	0.6944	0.6148	0.7629	0.6346	0.5605	0.6418	0.7214	0.71	**0.896**
	Kappa	0.0803	0.5769	0.5533	0.7014	0.5912	0.5232	0.6664	0.4834	0.4225	0.5247	0.6541	0.6363	**0.8611**
	NMI	0.0264	0.5709	0.6221	0.6669	0.5375	0.5141	0.6859	0.3913	0.4778	0.6397	0.7453	0.6798	**0.8202**
	ARI	0.0166	0.5819	0.5661	0.6379	0.5337	0.4231	0.6051	0.3822	0.3292	0.5548	0.692	0.6563	**0.8976**
	Purity	0.3729	0.7377	0.7386	0.7863	0.7189	0.7201	0.792	0.6346	0.5736	0.7176	0.8724	0.8477	**0.901**
	Time (s ↓)	7.62	97.0	57.99	72.96	161.37	124.98	289.20	N/A	N/A	8.35	30.25	92.04	**0.55**
MUUFL	ACC	0.1655	0.379	0.4943	0.4391	0.3917	0.3671	0.5391	0.3977	0.4044	0.3928	0.4544	0.4913	**0.6554**
	Kappa	0.0795	0.2226	0.419	0.3643	0.23	0.2548	0.4624	0.3131	0.3116	0.2797	0.3872	0.4131	**0.5679**
	NMI	0.0721	0.3611	0.5045	0.4518	0.3553	0.3213	**0.5262**	0.4422	0.3303	0.2378	0.4855	0.4596	0.4888
	ARI	0.0237	0.2228	0.3173	0.2237	0.2055	0.179	0.3535	0.2348	0.1484	0.1532	0.2866	0.3155	**0.427**
	Purity	0.4532	0.5782	0.7195	0.6713	0.5562	0.5571	**0.7499**	0.6564	0.601	0.5415	0.7131	0.7374	0.7262
	Time (s ↓)	5.92	86.11	54.63	33.06	123.78	102.17	262.99	N/A	N/A	7.09	24.97	75.03	**0.87**
Augsburg	ACC	0.3571	0.3678	0.4771	0.5013	0.3961	0.2937	0.5823	0.4586	0.4194	**0.6987**	0.3789	0.4791	0.6465
	Kappa	0.2357	0.1501	0.3237	0.3834	0.1708	0.1672	0.4176	0.1692	0.2857	**0.5994**	0.2532	0.3184	0.4981
	NMI	0.2651	0.1514	0.2675	0.4307	0.1379	0.1438	0.3331	0.1629	0.2389	**0.6081**	0.2609	0.3362	0.3727
	ARI	0.1702	0.0802	0.1985	0.3095	0.0847	0.0701	0.3014	0.2612	0.1631	**0.5685**	0.165	0.2678	0.3467
	Purity	0.6468	0.5168	0.5919	0.7071	0.5057	0.5134	0.6233	0.4592	0.5852	**0.8598**	0.5853	0.6528	0.7062
	Time (s ↓)	18.78	160.19	126.18	73.90	302.86	262.76	504.32	N/A	N/A	16.83	54.45	142.48	**1.62**



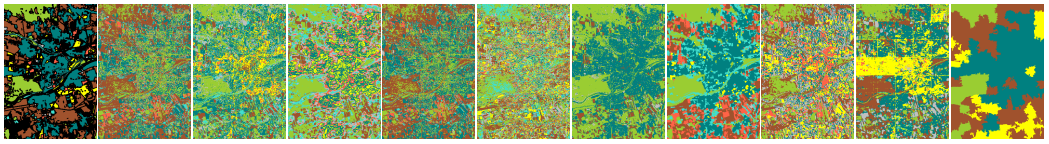
(**a**) (**b**) (**c**) (**d**) (**e**) (**f**) (**g**) (**h**) (**i**) (**j**) (**k**)

Figure 2. Cluster maps produced by different clustering models on the Trento dataset. (**a**) GT. (**b**) SMVSC. (**c**) LMVSC. (**d**) MDC. (**e**) FPMVS. (**f**) FMVACC. (**g**) MSGL. (**h**) FPFC. (**i**) AMKSC. (**j**) CDD. (**k**) MCSM.



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

Figure 3. Cluster maps produced by different clustering models on the MUUFL dataset. (**a**) GT. (**b**) SMVSC. (**c**) LMVSC. (**d**) MDC. (**e**) FPMVS. (**f**) FMVACC. (**g**) MSGL. (**h**) FPFC. (**i**) AMKSC. (**j**) CDD. (**k**) MCSM.



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

Figure 4. Cluster maps produced by different clustering models on the Augsburg dataset. (**a**) GT. (**b**) SMVSC. (**c**) LMVSC. (**d**) MDC. (**e**) FPMVS. (**f**) FMVACC. (**g**) MSGL. (**h**) FPFC. (**i**) AMKSC. (**j**) CDD. (**k**) MCSM.

Table 3. Clustering performance of MSCM and its manifold regularization-free version.

Datasets	Methods	ACC	Kappa	NMI	ARI	Purity
Trento	MCSM-M	0.8411	0.7905	0.7987	0.8835	0.9001
	MCSM	**0.896**	**0.8611**	**0.8202**	**0.8976**	**0.901**
MUUFL	MCSM-M	0.4417	0.3597	0.4226	0.2617	0.6878
	MCSM	**0.6554**	**0.5679**	**0.4888**	**0.427**	**0.7262**
Augsburg	MCSM-M	0.4207	0.2727	0.2511	0.1848	0.5888
	MCSM	**0.6465**	**0.4981**	**0.3727**	**0.3467**	**0.7062**

5.3.2. Multimodal Fusion

To investigate the fusion ability of MCSM, we conduct an ablation study on multimodal fusion. Specifically, we reform the MCSM model to a single modal version that is applied to individual modalities. As presented in Table 4, we find that the multimodal fusion version obtains the best clustering performance on all datasets except for one score on Augsburg, which verifies the effectiveness of the multimodal clustering mechanism. Specifically, the fusion version obtains significant improvement in all metrics on the MUUFL dataset. That is, MCSM can utilize the complementarity of multiple modalities to enhance the

clustering performance of single modalities. In addition, as can be seen from the HS term on Trento, its metric score is close to that of the associated fusion term, which indicates that even the single modal version of MCSM achieves high-quality clustering results, leading to a promising baseline system.

Table 4. Clustering performance with different modalities.

Datasets	Modalities	ACC	Kappa	NMI	ARI	Purity
Trento	HS	0.8835	0.8438	0.8016	0.8759	0.8835
	LiDAR	0.7808	0.7007	0.6233	0.6849	0.7915
	Fusion	**0.896**	**0.8611**	**0.8202**	**0.8976**	**0.901**
MUUFL	HS	0.4455	0.3452	0.3557	0.2362	0.5972
	LiDAR	0.5309	0.3707	0.3694	0.301	0.5713
	Fusion	**0.6554**	**0.5679**	**0.4888**	**0.427**	**0.7262**
Augsburg	HS	0.4643	0.1917	0.0844	0.0528	0.4839
	SAR	**0.6486**	0.4865	0.343	0.3201	0.6932
	Fusion	0.6465	**0.4981**	**0.3727**	**0.3467**	**0.7062**

5.4. Model Analysis

5.4.1. Impacts of Superpixels

The MCSM model is built on superpixels achieved by superpixel segmentation that involves an important parameter of K being the number of superpixels. To examine its impacts, we visualize the clustering performance of MCSM by plotting metric curves in a range of $[50, 500]$ with a step of 50, as shown in Figure 5. We can see from the figure that, for each dataset, all metric curves maintain consistent trends overall. Furthermore, the best performance in ACC appears at coordinates 100, 400, and 150 for the Trento, MUUFL, and Augsburg datasets, respectively. This result makes sense, because Trento has more concentrated land cover, while MUUFL contains more dispersed land cover, referring to Figures 2a and 3a. A smaller K tends to cause larger superpixels. Thus, the parameter setting is applicable to the characteristics of these land covers. In contrast, due to the interleaved land covers of Augsburg, the metric curves have certain fluctuations, as shown in Figure 4a. Thus, this phenomenon can guide us in determining parameters for real-world scenarios.

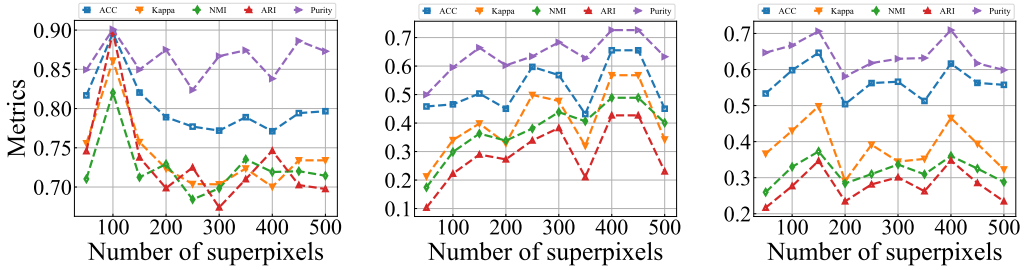


Figure 5. Effect of the number of superpixels on clustering performance. (**a**) Trento. (**b**) MUUFL. (**c**) Augsburg.

5.4.2. Convergence

Figure 6 presents the convergence of Algorithm 1 associated with the proposed MCSM. To do so, we examine residual curves of $\mathbf{F}^v, v = 1, 2$ and $\mathbf{F}$ being memberships, That is, for a $\mathbf{F}_t$ at iteration step of t, the residual is calculated by $\|\mathbf{F}_t - \mathbf{F}_{t-1}\|_\infty$—which we call loss in the figure. We see that the losses quickly start to converge from around iteration 20, 40, and 100 for Trento, MUUFL, and Augsburg datasets, respectively. On Trento, although $\mathbf{F}^2$ loss yields a peak at 5, which may be caused by the complex data distribution of LiDAR, it very

quickly converges together with $\mathbf{F}^1$ and $\mathbf{F}$. In addition, all losses converge to 0 within 50, 60, and 100 on the three datasets, respectively, demonstrating the efficiency of the optimization algorithm, thanks to the special gradient descent strategy. Further, each updating step only requires a few matrix operators, which also contributes to the low time consumption.

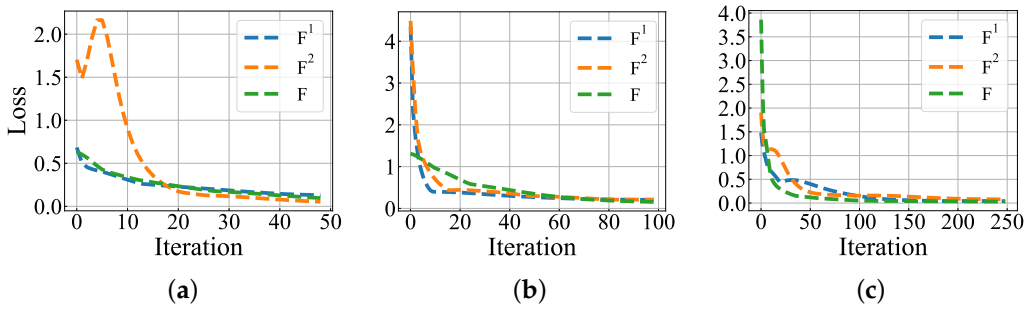


Figure 6. Convergence of the proposed optimization algorithm. (**a**) Trento. (**b**) MUUFL. (**c**) Augsburg.

5.4.3. Analysis of Parameters

The MCSM model comprises two parameters: λ is the manifold regularization coefficient and D is the dimension of PCA for HS data. They have a close relationship with manifold structure, so we aim to examine the sensitivity of the clustering performance in ACC to them. For this purpose, we adopt a grid search strategy by traversing λ and D in a range of $[1 \times 10^{-4}, 1 \times 10^{4}]$ with a step of 10 and a range of $[1, 10]$ with a step of 1, sequentially. As we can see from Figure 7, the clustering results in ACC show stable performance with respect to λ and D on Trento, while they are sensitive on MUUFL and Augsubrg when increasing D beyond 4. This reflects the necessity of eliminating redundant spectral information in HS images. Not surprisingly, taking a smaller D, the MCSM method achieves stable ACC scores with different λ on all datasets. This is because the low-dimensional features by PCA trend to provide a more intrinsic manifold structure to the manifold regularization. In addition, the best performance is achieved around $\lambda = 1$ on all datasets overall, which conveys the importance of balancing between the reconstruction error and the manifold regularization.

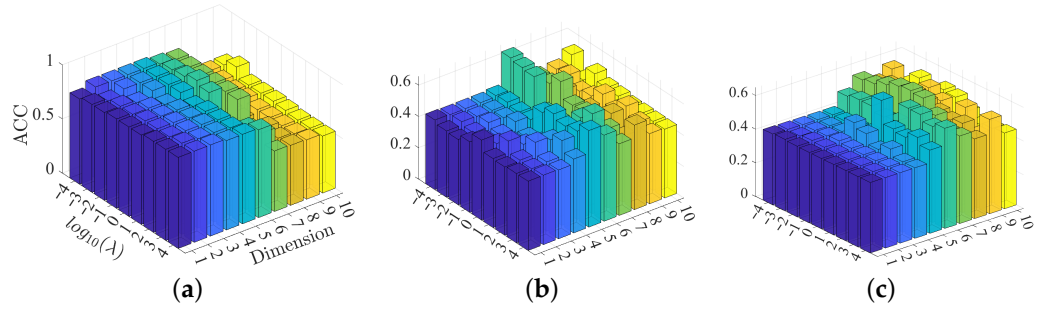


Figure 7. Sensitivity analysis of ACC of the MCSM method to λ and D. (**a**) Trento. (**b**) MUUFL. (**c**) Augsburg.

6. Conclusions

In this article, we devised an efficient multimodal clustering method on superpixel manifolds for large-scale remote sensing images. In contrast to traditional subspace clustering, the MCSM methods learn a unified membership matrix to indicate clusters on superpixel manifolds instead of the original pixel counterpart. We integrate manifold regularization into the framework, such that the membership matrix trends to capture the manifold structure of multimodal data. Indeed, we prove that the proposed model is equivalent to a special case of a multiview low-rank subspace clustering model. The pro-

posed method achieves significant clustering accuracy, with ACC scores of 0.896, 0.6554, and 0.6465 on the Trento, MUUFL, and Augsburg datasets, respectively. This quantitative advantage is further supported by qualitative results, as the method generates smoother clustering maps that demonstrate improved local spatial consistency. Furthermore, we verify that the optimization algorithm associated with MCSM converges very quickly, and the ablation study of superpixel manifolds verified the effectiveness of manifold guidance.

Author Contributions: Conceptualization, S.L. and L.X.; methodology, S.L.; software, S.L. and Y.Y.; validation, L.X.; formal analysis, S.L.; investigation, S.L.; resources, S.L.; data curation, S.L.; writing—original draft preparation, S.L. and Y.Y.; writing—review and editing, S.L.; visualization, S.L.; supervision, S.L.; project administration, S.L.; funding acquisition, S.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Sichuan Science and Technology Program under grant number 2026NSFSC1405.

Data Availability Statement: Data will be made available on request.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Kieu, N.; Nguyen, K.; Nazib, A.; Fernando, T.; Fookes, C.; Sridharan, S. Multimodal colearning meets remote sensing: Taxonomy, state of the art, and future works. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2024**, *17*, 7386–7409. [CrossRef]
2. Zhao, D.; Hu, B.; Jiang, W.; Zhong, W.; Arun, P.V.; Cheng, K.; Zhao, Z.; Zhou, H. Hyperspectral video tracker based on spectral difference matching reduction and deep spectral target perception features. *Opt. Lasers Eng.* **2025**, *194*, 109124. [CrossRef]
3. Zhao, D.; Zhang, H.; Arun, P.V.; Jiao, C.; Zhou, H.; Xiang, P.; Cheng, K. SiamSTU: Hyperspectral video tracker based on spectral spatial angle mapping enhancement and state aware template update. *Infrared Phys. Technol.* **2025**, *150*, 105919. [CrossRef]
4. Guan, R.; Tu, W.; Li, Z.; Yu, H.; Hu, D.; Chen, Y.; Tang, C.; Yuan, Q.; Liu, X. Spatial-spectral graph contrastive clustering with hard sample mining for hyperspectral images. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5535216. [CrossRef]
5. Luo, F.; Liu, Y.; Gong, X.; Nan, Z.; Guo, T. EMVCC: Enhanced multi-view contrastive clustering for hyperspectral images. In Proceedings of the 32nd ACM International Conference on Multimedia, Melbourne, VIC, Australia, 28 October–1 November 2024; pp. 6288–6296.
6. Huang, S.; Zeng, H.; Chen, H.; Zhang, H. Spatial and cluster structural prior-guided subspace clustering for hyperspectral image. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5511115. [CrossRef]
7. Zhai, H.; Zhang, H.; Zhang, L.; Li, P. Laplacian-regularized low-rank subspace clustering for hyperspectral image band selection. *IEEE Trans. Geosci. Remote Sens.* **2018**, *57*, 1723–1740. [CrossRef]
8. Huang, N.; Xiao, L.; Liu, J.; Chanussot, J. Graph convolutional sparse subspace coclustering with nonnegative orthogonal factorization for large hyperspectral images. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 5512016. [CrossRef]
9. Liu, S. Dual subspace clustering for spectral-spatial hyperspectral image clustering. *Image Vis. Comput.* **2024**, *150*, 105235. [CrossRef]
10. Zhai, H.; Zhang, H.; Zhang, L.; Li, P.; Plaza, A. A new sparse subspace clustering algorithm for hyperspectral remote sensing imagery. *IEEE Geosci. Remote Sens. Lett.* **2016**, *14*, 43–47. [CrossRef]
11. Lei, J.; Li, X.; Peng, B.; Fang, L.; Ling, N.; Huang, Q. Deep spatial-spectral subspace clustering for hyperspectral image. *IEEE Trans. Circuits Syst. Video Technol.* **2020**, *31*, 2686–2697. [CrossRef]
12. Zhai, H.; Yu, J. Local-Global Spectral-Spatial Dual Deep Subspace Clustering Network for Hyperspectral Images. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 5526520. [CrossRef]
13. Cai, Y.; Zhang, Z.; Cai, Z.; Liu, X.; Jiang, X.; Yan, Q. Graph Convolutional Subspace Clustering: A Robust Subspace Clustering Framework for Hyperspectral Image. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 4191–4202. [CrossRef]
14. Liu, S.; Wang, H. Graph Convolutional Optimal Transport for Hyperspectral Image Spectral Clustering. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 4414013. [CrossRef]
15. Ding, Y.; Zhang, Z.; Yang, A.; Cai, Y.; Xiao, X.; Hong, D.; Yuan, J. SLCGC: A lightweight Self-supervised Low-Pass Contrastive Graph Clustering Network for Hyperspectral Images. *IEEE Trans. Multimed.* **2025**, *27*, 8251–8262. [CrossRef]
16. Ding, Y.; Zhang, Z.; Kang, W.; Yang, A.; Zhao, J.; Feng, J.; Hong, D.; Zheng, Q. Adaptive homophily clustering: Structure homophily graph learning with adaptive filter for hyperspectral image. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 5510113. [CrossRef]

17. Zhao, D.; Wang, M.; Huang, K.; Zhong, W.; Arun, P.V.; Li, Y.; Asano, Y.; Wu, L.; Zhou, H. OCSCNet-tracker: Hyperspectral video tracker based on octave convolution and spatial–spectral capsule network. *Remote Sens.* **2025**, *17*, 693. [CrossRef]
18. Zhao, D.; Zhong, W.; Ge, M.; Jiang, W.; Zhu, X.; Arun, P.V.; Zhou, H. SiamBSI: Hyperspectral video tracker based on band correlation grouping and spatial-spectral information interaction. *Infrared Phys. Technol.* **2025**, *151*, 106063. [CrossRef]
19. Xiao, L.; Liu, S.; Liu, Y. Anchor-guided multi-view fuzzy clustering for hyperspectral and LiDAR images. *Sci. Rep.* **2026**. [CrossRef]
20. Zhao, D.; Xu, X.; You, M.; Arun, P.V.; Zhao, Z.; Ren, J.; Wu, L.; Zhou, H. Local sub-block contrast and spatial–spectral gradient feature fusion for hyperspectral anomaly detection. *Remote Sens.* **2025**, *17*, 695. [CrossRef]
21. Luo, F.; Liu, Y.; Guo, T.; Fu, C.; Duan, Y.; Shi, Q.; Du, B. SCMVC: Semantic Constraint-Based Spatial–Spectral Multiview Clustering for Hyperspectral Images. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 5533013. [CrossRef]
22. Guan, R.; Tu, W.; Hu, D.; Liang, W.; Liang, K.; Hu, Y.; Liu, Y.; Liu, X. Prototype-driven multi-view attribute-missing graph clustering. *IEEE Trans. Multimed.* **2025**, *27*, 9454–9466.
23. Bickel, S.; Scheffer, T. Multi-view clustering. In *ICDM'04: Proceedings of the Fourth IEEE International Conference on Data Mining, Brighton, UK, 1–4 November 2004*; IEEE Computer Society: Washington, DC, USA, 2004; Volume 4, pp. 19–26.
24. Zhang, P.; Liu, X.; Xiong, J.; Zhou, S.; Zhao, W.; Zhu, E.; Cai, Z. Consensus one-step multi-view subspace clustering. *IEEE Trans. Knowl. Data Eng.* **2020**, *34*, 4676–4689. [CrossRef]
25. Tang, C.; Li, Z.; Wang, J.; Liu, X.; Zhang, W.; Zhu, E. Unified one-step multi-view spectral clustering. *IEEE Trans. Knowl. Data Eng.* **2022**, *35*, 6449–6460. [CrossRef]
26. Peng, S.; Yin, J.; Yang, Z.; Chen, B.; Lin, Z. Multiview Clustering via Hypergraph Induced Semi-Supervised Symmetric Nonnegative Matrix Factorization. *IEEE Trans. Circuits Syst. Video Technol.* **2023**, *33*, 5510–5524. [CrossRef]
27. Large-scale multi-view subspace clustering in linear time. In *Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020*; AAAI Press: Washington, DC, USA, 2020; Volume 34, pp. 4412–4419.
28. Sun, M.; Zhang, P.; Wang, S.; Zhou, S.; Tu, W.; Liu, X.; Zhu, E.; Wang, C. Scalable multi-view subspace clustering with unified anchors. In *MM'21: Proceedings of the 29th ACM International Conference on Multimedia, Chengdu, China, 20–24 October 2021*; Association for Computing Machinery (ACM): New York, NY, USA, 2021; pp. 3528–3536.
29. Liu, S.; Wang, S.; Zhang, P.; Xu, K.; Liu, X.; Zhang, C.; Gao, F. Efficient one-pass multi-view subspace clustering with consensus anchors. In *Proceedings of the AAAI Conference on Artificial Intelligence, Virtual Conference, 22 February–1 March 2022*; AAAI Press: Washington, DC, USA, 2022; Volume 36, pp. 7576–7584.
30. Wang, S.; Liu, X.; Zhu, X.; Zhang, P.; Zhang, Y.; Gao, F.; Zhu, E. Fast Parameter-Free Multi-View Subspace Clustering with Consensus Anchor Guidance. *IEEE Trans. Image Process.* **2022**, *31*, 556–568. [CrossRef]
31. Mi, Y.; Chen, H.; Yuan, Z.; Luo, C.; Horng, S.J.; Li, T. Fast Multi-view Subspace Clustering with Balance Anchors Guidance. *Pattern Recognit.* **2024**, *145*, 109895.
32. Cheng, T.; Peng, J.; Li, H.; Wang, H. Large-scale multi-view subspace clustering via embedding space and partition matrix. *Neurocomputing* **2024**, *602*, 128266.
33. Yu, S.; Liu, S.; Wang, S.; Tang, C.; Luo, Z.; Liu, X.; Zhu, E. Sparse Low-Rank Multi-View Subspace Clustering with Consensus Anchors and Unified Bipartite Graph. *IEEE Trans. Neural Netw. Learn. Syst.* **2025**, *36*, 1438–1452. [CrossRef]
34. Guan, R.; Li, J.; Wang, S.; Tu, W.; Li, M.; Zhu, E.; Liu, X.; Chen, P.Multi-view Graph Clustering with Dual Relation Optimization for Remote Sensing Data. In *MM'25: Proceedings of the 33rd ACM International Conference on Multimedia, Dublin, Ireland, 27–31 October 2025*; Association for Computing Machinery (ACM): New York, NY, USA, 2025; pp. 7346–7355.
35. Guan, R.; Liu, T.; Tu, W.; Tang, C.; Luo, W.; Liu, X. Sampling Enhanced Contrastive Multi-View Remote Sensing Data Clustering with Long-Short Range Information Mining. *IEEE Trans. Knowl. Data Eng.* **2025**, *37*, 5598–5612. [CrossRef]
36. Liu, S.; Chang, L. Conditional Dual Diffusion for Multimodal Clustering of Optical and SAR Images. *IEEE Trans. Circuits Syst. Video Technol.* **2025**, *35*, 5318–5330. [CrossRef]
37. Cai, Y.; Zhang, Z.; Liu, X.; Ding, Y.; Li, F.; Tan, J. Learning Unified Anchor Graph for Joint Clustering of Hyperspectral and LiDAR Data. *IEEE Trans. Neural Netw. Learn. Syst.* **2025**, *36*, 6341–6354. [CrossRef] [PubMed]
38. Zhang, Z.; Cai, Y.; Gong, W.; Liu, X.; Zeng, C.; Yu, G. MMAGL: Multiobjective Multiview Attributed Graph Learning for Joint Clustering of Hyperspectral and LiDAR Data. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 5505314. [CrossRef]
39. Liu, Y.; Liu, S.; Xiao, L. Sub-pixel Anchor Graph for Multi-modal Clustering of Hyperspectral and LiDAR Data. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 5536915. [CrossRef]
40. Cai, Y.; Zhang, Z.; Ghamisi, P.; Rasti, B.; Liu, X.; Cai, Z. Transformer-based contrastive prototypical clustering for multimodal remote sensing data. *Inf. Sci.* **2023**, *649*, 119655. [CrossRef]
41. Zhang, Y.; Yan, S.; Jiang, X.; Zhang, L.; Cai, Z.; Li, J. Dual Graph Learning Affinity Propagation for Multimodal Remote Sensing Image Clustering. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5521713. [CrossRef]
42. Zhang, Y.; Yan, S.; Zhang, L.; Du, B. Fast Projected Fuzzy Clustering with Anchor Guidance for Multimodal Remote Sensing Imagery. *IEEE Trans. Image Process.* **2024**, *33*, 4640–4653. [CrossRef]

43. Wang, X.; Zhang, Y.; Zhou, Y. Multimodal Remote Sensing Image Clustering with Multiscale Spectral-Spatial Anchor Graphs. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 4405612. [CrossRef]
44. Achanta, R.; Shaji, A.; Smith, K.; Lucchi, A.; Fua, P.; Süsstrunk, S. SLIC superpixels compared to state-of-the-art superpixel methods. *IEEE Trans. Pattern Anal. Mach. Intell.* **2012**, *34*, 2274–2282.
45. Elhamifar, E.; Vidal, R. Sparse Subspace Clustering: Algorithm, Theory, and Applications. *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *35*, 2765–2781. [CrossRef]
46. Vidal, R.; Favaro, P. Low rank subspace clustering (LRSC). *Pattern Recognit. Lett.* **2014**, *43*, 47–61. [CrossRef]
47. Liu, G.; Lin, Z.; Yan, S.; Sun, J.; Yu, Y.; Ma, Y. Robust Recovery of Subspace Structures by Low-Rank Representation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *35*, 171–184. [CrossRef] [PubMed]
48. Wang, R.; Nie, F.; Yu, W. Fast Spectral Clustering with Anchor Graph for Large Hyperspectral Images. *IEEE Geosci. Remote Sens. Lett.* **2017**, *14*, 2003–2007. [CrossRef]
49. Bezdek, J.C.; Ehrlich, R.; Full, W. FCM: The fuzzy c-means clustering algorithm. *Comput. Geosci.* **1984**, *10*, 191–203. [CrossRef]
50. Kheirandishfard, M.; Zohrizadeh, F.; Kamangar, F. Deep low-rank subspace clustering. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Seattle, WA, USA, 14–19 June 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 864–865.
51. Lee, D.; Seung, H.S. Algorithms for non-negative matrix factorization. *Adv. Neural Inf. Process. Syst.* **2000**, *13*, 535–541.
52. Zhang, Z.; Liu, L.; Shen, F.; Shen, H.T.; Shao, L. Binary Multi-View Clustering. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *41*, 1774–1782. [CrossRef]
53. Shahi, K.R.; Ghamisi, P.; Rasti, B.; Scheunders, P.; Gloaguen, R. Unsupervised Data Fusion With Deeper Perspective: A Novel Multisensor Deep Clustering Algorithm. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2022**, *15*, 284–296. [CrossRef]
54. Wang, S.; Liu, X.; Liu, S.; Jin, J.; Tu, W.; Zhu, X.; Zhu, E. Align then fusion: Generalized large-scale multi-view clustering with anchor matching correspondences. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 5882–5895.
55. Kang, Z.; Lin, Z.; Zhu, X.; Xu, W. Structured Graph Learning for Scalable Subspace Clustering: From Single View to Multiview. *IEEE Trans. Cybern.* **2022**, *52*, 8976–8986. [CrossRef]
56. Xu, J.; Ren, Y.; Tang, H.; Yang, Z.; Pan, L.; Yang, Y.; Pu, X.; Yu, P.S.; He, L. Self-Supervised Discriminative Feature Learning for Deep Multi-View Clustering. *IEEE Trans. Knowl. Data Eng.* **2023**, *35*, 7470–7482. [CrossRef]
57. Yan, W.; Zhang, Y.; Lv, C.; Tang, C.; Yue, G.; Liao, L.; Lin, W. Gcfagg: Global and cross-view feature aggregation for multi-view clustering. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; IEEE: Piscataway, NJ, USA, 2023; pp. 19863–19872.

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