Human Communication-Inspired Semantic–View Collaborative Network for Multispectral Remote Sensing Image Retrieval

Nan Wu, *Student Member, IEEE*, Wei Jin, and Randi Fu

***Abstract*—Multispectral images (MSIs) have widespread applications, and efficiently managing these extensive MSIs via remote sensing image retrieval (RSIR) is key to boosting their practical value. While current deep learning-based methods offer strong image representation learning capabilities, adapting to complex and dynamic relationships between objects and spectral information in MSIs remains challenging. This difficulty arises due to the distinct attributes of different spectral bands and the lack of consideration of interactions among spectral combinations in MSIs, which limits their retrieval performance. For this purpose, we propose a dynamic learning system inspired by human communication named the semantic–view collaborative network (SVCNet), which actively promotes the interaction between spectral and semantic information. By linking multiview learning (MVL) with graph neural networks (GNNs) to simulate the three stages of human communication—understanding, communication, and collective consensus and reflection—SVCNet enhances RSIR with flexibility in representation extraction. Specifically, each spectral combination is processed to extract independent representations as view-specific knowledge. In the communication phase, we devise the graph attention-based multiround communication module (GACM), which uses GNN to perform graph-structured modeling and adaptive updating of views and semantics. Moreover, we achieve improved MSI representations by implementing novel objective functions that align learned semantics with category information, dynamically differentiating semantic similarities and disparities in MSIs, and flexibly weighting samples for enhanced adaptability in a multilabel RSIR environment. SVCNet surpasses current state-of-the-art methods in three MSI datasets for single and multilabel retrieval tasks. It effectively handles class imbalances and distinguishes challenging samples, highlighting its extensive applicability.**

***Index Terms*—Content-based image retrieval, graph neural networks (GNNs), multiview learning (MVL), multispectral image (MSI).**

I. Introduction

SATELLITE remote sensing technology captures various information about the Earth's surface, greatly enriching databases for geographical and temporal analysis in Earth observation [1]. Multispectral images (MSIs) generated by remote sensing sensors encompass multiple spectral bands, including but not limited to visible RGB channels. MSIs are widely applied in urban development [2], meteorology [3], [4], agriculture, and environmental monitoring research. Unprecedented advances in satellite remote sensing technology have led to the rapid expansion of high-resolution MSI archives [5]. However, efficiently managing these voluminous MSIs poses a challenge. Content-based remote sensing image retrieval (RSIR) uses remote sensing image (RSI) content to search and identify target images with similar semantics. The design of RSIR aligns with the way humans naturally recognize and categorize image content and improves the organization and application of MSIs [6], [7].

The efficacy of RSIR in returning relevant search results is based on its ability to extract robust and discriminative image representations. These are essential for accurately identifying the complex semantic content of MSI and for conducting effective image-to-image similarity assessments in subsequent retrieval processes. Traditional RSIR approaches have employed handcrafted features, such as texture, spectrum, and shape [8], as well as advanced global descriptor techniques including the scale-invariant feature transform (SIFT) for extracting bag-of-visual words representations [9] and local binary patterns (LBP) [10]. With the advancement and superior performance of deep learning (DL) technologies, DL-based image representation learning (IRL) has shifted the focus from handcrafted feature extraction to data-driven approaches, thereby revolutionizing the field of RSIR [5].

Many DL-based RSIR methods aim to improve IRL capabilities for RSI; for example, Hu et al. [3] adapted the ResNet [11] architecture to extract features from MSI, employing an attention mechanism to focus on regions of interest. Huang et al. [12] fused global and local features in RSI, achieving more discriminative representations of remote sensing images through contrastive learning. Kang et al. [13] utilized graph structures to link image representations and labels in RSI, developing a more discriminative metric space for multilabel RSIR tasks. Despite their effectiveness, existing DL-based RSIR methods typically extract a unified representation by concatenating channels of MSI or using synthesized RGB images from three bands. However, these methods do not fully capture the complexity and dynamism in the relationship between objects and spectral information, which is reflected in the way different spectral bands represent

Manuscript received 3 February 2024; revised 7 May 2024; accepted 28 May 2024. Date of publication 4 June 2024; date of current version 17 June 2024. This work was supported in part by the National Natural Science Foundation of China under Grant 42071323 and in part by the Joint Fund of the Zhejiang Provincial Natural Science Foundation of China under Grant LZJMZ24D050002. *(Corresponding authors: Wei Jin; Randi Fu.)*

The authors are with the Faculty of Electrical Engineering and Computer Science, Ningbo University, Ningbo 315211, China (e-mail: wunan@ieee.org; jinwei@nbu.edu.cn; furandi@nbu.edu.cn).

Digital Object Identifier 10.1109/JSTARS.2024.3409417

distinct properties of objects. This limitation hinders their ability to adequately capture interactions between spectral characteristics and object properties, potentially restricting adaptability in extracting contextual representations of spectral and spatial features.

Multiview learning (MVL) integrates and aligns information from various aspects into a unified representation space [14], suitable for handling the diverse spectral information of MSI due to its capability to process multiple aspects. Recent studies have explored DL-based methods, akin to the idea of MVL, to optimize representations in MSIs, enhancing adaptability in remote sensing tasks across a spectrum from representation extraction to downstream applications. For example, Xu et al. [15] computed weights for different spectral bands and encoded local features of the fused spectrum into a spatial graph structure for MSI representation generation. Zhao et al. [4] focused on extracting representations from various spectral combinations and integrating these into a Transformer [16] architecture for unified representation. Meanwhile, Cong et al. [1] developed a strategy involving distinct feature extractors for each spectral group, augmented with a novel spectral group position encoding to enhance training efficiency. While MVL is a promising tool for extracting more discriminative spectral representations, existing advances primarily concentrate on alignment and fusion across views, limiting their effectiveness in more complex and dynamic situations where spectral and semantic information interact intricately according to different contexts.

In our quest to enhance RSIR's flexibility in representation extraction, we recognize the critical role of graph neural networks (GNNs) [17] in capturing complex relational dynamics. Graphs, as abstract data types, are adept at capturing complex relationships, offering a robust framework for modeling dependencies between nodes in multispectral information and object properties. Recent advances have shown the efficacy of GNNs in RSI, effectively modeling relationships between samples [18], spatial structures [19], and various views [20], while integration with causal conditions (categories) remains a challenge. Contrasting with existing deep learning-based RSIR methods and GNNs, our work focuses on borrowing from the power of human thought and communication, which stems from the ability to gather, integrate, synthesize, and establish relationships from diverse information sources. Drawing inspiration from the natural autonomous learning and goal-oriented characteristics of human communication, we aim to refine existing GNNs in RSIR by incorporating these strategies for more robust IRL capabilities.

We propose a human communication-inspired dynamic learning system, named semantic–view collaborative network (SVCNet), actively promoting interaction between spectral and semantic information. By linking GNNs [17] to the three stages of human communication—understanding, communication, and collective consensus and reflection—we endow RSIR with flexibility in IRL. Additionally, the proposed loss functions are tailored to enhance representation optimization during the training of RSIR methods. Fig. 1 demonstrates how SVCNet differs from conventional DL-based RSIR algorithms. Specifically, SVCNet is composed of three components, each drawing inspiration from different facets of human communication.

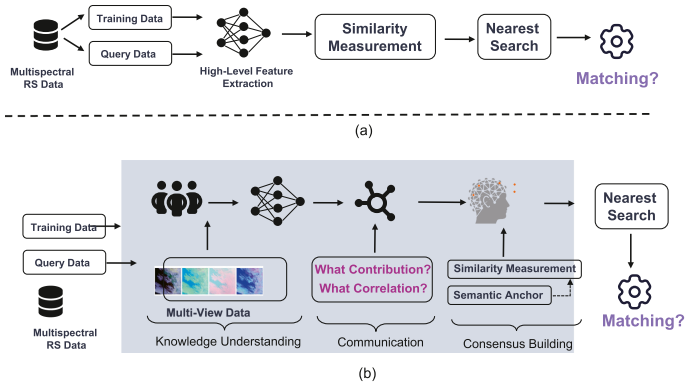

Fig. 1. Contrastive illustration between SVCNet and conventional multispectral retrieval networks. SVCNet serves as a MVL framework, aimed at extracting more discriminative representations through the collaborative interplay between views and semantics. (a) Generic representation extraction and retrieval in remote sensing image analysis. (b) SVCNet in multispectral remote sensing image retrieval.

View Understanding: We simulate how humans independently process and abstract knowledge before communication. This is achieved using a set of independent spectral representation extraction layers, each mimicking how humans individually process information sources to derive spectral representations from their respective physical properties.

Communication: Recognizing that effective communication among human information sources is key to collective decision-making [21] and affects the quality of decisions [22], Jia et al. [23] have shown that sharing information between views akin to human communication improves MVL performance. Benefiting from the capability of GNNs to dynamically capture the relationships between view and semantic aspects, we use GNNs to model and extract complex associations between objects and spectral information as representations to enhance RSIR. We devise the graph attention-based multiround communication module (GACM) to simulate the interactions between objects and semantics, where the semantics are modeled as learnable anchors connected with view representations in a graph structure. Utilizing GNNs, GACM extracts and updates complex associations among view–view, semantic–semantic, and view–semantic elements, thereby improving IRL for RSIR.

Collective Consensus and Reflection: After communication, the integration of human viewpoints, assigning importance, and reflective reevaluation during the post-communication phase form the basis for better decision-making and groundwork for future interactions [24], [25]. To this end, we integrate view–semantic association representations to achieve collective consensus. Leveraging machine learning's backpropagation process to simulate collective reflection, we introduce two novel objective functions in SVCNet: semantic alignment loss (SAL) function and tightened graph contrastive loss (TGCL) function. SAL function works by aligning semantic anchors with corresponding category information for supervision during the training of RSIR models, serving as semantic targets to promote the goal-oriented nature of human communication and ensuring adaptability to both single-label and multilabel RSIR

contexts. Meanwhile, TGCL function acts as an assessment and optimization of communication outcomes, effectively bringing semantic–view representations closer together in the metric space when they are semantically similar.

In summary, SVCNet advances beyond existing RSIR methods by representing content and view–semantic relationships. SVCNet offers two advantages over current DL-based RSIR methods. 1) It treats all categories as nodes within the GNN, ensuring equitable treatment within semantic anchors and addressing class imbalance challenges. 2) Its relationship-based modeling naturally accommodates both single-label and multilabel RSIR tasks. Our main contributions are as follows.

1) We propose SVCNet, a novel framework for RSIR using MSIs. Distinct from current DL-based RSIR methods, SVCNet draws inspiration from human communication, emphasizing the interactions, collaboration, and feedback between spectral data and semantics.
2) We devise GACM to advance MVL beyond its usual focus on view fusion and weighting. GACM integrates relationships among view–view, semantic–semantic, and semantic–view interactions, leading to more discriminative MSI representations for RSIR tasks.
3) We introduce two novel objective functions, the SAL and TGCL functions. These functions optimize MSI representations for RSIR tasks and enhance the versatility of the extracted representations across both single-label and multilabel scenarios.
4) Experimental results from three MSI datasets in both single-label and multilabel scenarios show that SVCNet is competitive with existing methods. Our model particularly excels in handling imbalanced class distributions and identifying challenging samples, indicating its broader applicability in RSIR tasks.

The rest of this article is organized as follows. We start by introducing related work in Section II, and then present our SVCNet in Section III. In Section IV, we conduct extensive experiments on three datasets and compare our approach with state-of-the-art methods. In Section V, we analyze and discuss key features of our approach. Finally, Section VI concludes this article.

II. Related Work

A. Content-Based Image Retrieval

In CBIR, technological implementations involve feature extraction and similarity measure [6], [7]. Before deep learning, feature extraction in CBIR methods relied on handcrafted descriptors for low-level visual features, such as color, texture, and shape [8]. Deep learning frameworks, like ResNet [11], DenseNet [26], and vision Transformer [16], are now widely adopted in CBIR for their ability to discover discriminative structures in image content and semantics [6], [7], [27], and to automatically learn optimal feature representations.

For the similarity measure in the CBIR, traditional methods utilized combinations of hand-crafted features. Doulamis et al. [28] introduced the maximum deviation class separation (MDCS) algorithm for video retrieval, using correlation coefficients between feature vectors as a distance metric. In DL-based methods, deep metric learning (DML) projects inputs into a metric space where representations of the same class are brought closer together, while those of different classes are separated, thus offering an end-to-end approach for extracting representations in deep learning. In DML implementations, Siamese networks learn in metric spaces by using comparisons of similar and dissimilar class pairs [5], [19]. However, the large volume of sample pairs resulting from extensive datasets in DML has led to the development of batch optimization methods [27] and selective pairing of samples [5], improving the performance of DL-based CBIR methods. In DL-based RSIR methods, a variety of advanced solutions have been proposed to enhance feature representation and similarity assessment. For example, Sumbul et al. [5] introduced a representative triplet selection scheme for a more effective training of DL RSIR models. Huang et al. [12] proposed a contrastive learning scheme to improve similarity comparison.

In early CBIR systems, feature representations and their weights were fixed. Given the limitations of computer-centered approaches, relevance feedback (RF) mechanisms emerged, incorporating humans as part of the retrieval process [29], [30]. RF automatically adjusts existing queries using user feedback about the relevance of previously retrieved objects, thereby better approximating the user's information needs and enabling interaction between the user and the CBIR system [29], [31]. RF integrates human communication and thinking capabilities into the CBIR system. The development of deep learning models has automated this process through supervised learning of model features before retrieval [29]. Building on and improving the RF concept, this article addresses the limitations of existing methods, which lack interactive integration of various spectral information and high-level category information. Building on the RF foundations [32], [33], SVCNet uniquely emphasizes the crucial interplay, collaboration, and feedback between spectral data and semantics, making it exceptionally well suited for MSI retrieval. In similarity assessment, SVCNet incorporates SAL and TGCL, specifically designed to optimize MSI representations for RSIR tasks. These components enhance performance and broaden the versatility of representations in both single-label and multilabel contexts.

B. Multiview Learning

MVL aims to harness the consistency and complementarity between different views for extracting more robust representations [14]. Current methods are categorized as connection-based, calibration-based, and hybrid [23]. Connection-based MVL focuses on extracting and integrating information from different views into a unified representation. While simple and effective, this approach may introduce redundant information. Calibration-based methods ensure alignment of representations from different views within the same semantic space through feature matching, typically using contrastive learning [34] or adversarial training [14], but may lose uncalibrated complementary information. Hybrid approaches value the complex relationships between views. For example, Zhu

et al. [35] introduced a gating aggregation mechanism to enhance information exchange and adaptive weighting between views. Shuai et al. [36] adaptively quantified latent view relationships using relative attention blocks and transformer modules, constructing richer feature representations. Inspired by human communication, Jia et al. [23] developed a model in which each view harnesses complementary information from other views to construct its representation, thus promoting knowledge interaction and integration among different views.

Compared to the above methods, our proposed SVCNet employs a hybrid strategy, promoting multilevel, multiround communication between views within the existing MVL framework. By leveraging GNN, SVCNet effectively bridges semantic information across different views, making it well suited for RSIR tasks.

C. Graph Neural Networks

GNNs offer a versatile and efficient framework for learning from graph-structured data. Their adaptability and potent representational capabilities make them apt for various domains, mainly when data embodies a collection of nodes with predictions contingent on internode relationships. The central premise of GNNs involves iterative state updates for each node through interactions with its neighbors. Distinct GNN variants, such as those in [37], [38], and [39], differ mainly in how nodes accumulate and assimilate information from neighbors.

Within a GNN layer, node representations are input as $\{\boldsymbol{h}_i \in \mathbb{R}^d \mid i \in \mathcal{V}\}$. The layer outputs new node representations $\{\boldsymbol{h}'_i \in \mathbb{R}^{d'} \mid i \in \mathcal{V}\}$. Every node i and its neighboring set $\mathcal{N}_i$ are updated using the same parameter function f_θ

$$\boldsymbol{h}'_i = f_\theta\left(\boldsymbol{h}_i, \operatorname{agg}\left(\{\boldsymbol{h}_j \mid j \in \mathcal{N}_i\}\right)\right). \tag{1}$$

Here, function f and the aggregation $\operatorname{agg}(\cdot)$ operation delineate different GNN variants. For instance, in GraphSAGE [40], the aggregation entails elementwise averaging, processed through a linear layer and an ReLU activation as f. In this context, graph attention networks (GAT) [41], [42], a prominent GNN architecture, employs attention mechanisms to compute weighted averages of neighboring nodes

$$\boldsymbol{h}'_i = f_\theta\left(\boldsymbol{h}_i, \mathcal{A}\left(\boldsymbol{h}_i, \{\boldsymbol{h}_j \mid j \in \mathcal{N}_i\}\right)\right) \tag{2}$$

where $\mathcal{A}$ is an attention function. GAT represents a more powerful variant of GNN in DL. Different GAT variants may have distinct attention functions $\mathcal{A}$ and mapping functions f, yet their essence remains to more effectively aggregate neighbor nodes.

III. Methodology

A. Overall Framework

For the RSIR task using MSI, we define a training dataset $\mathcal{D}^t = \{(\mathcal{I}_i^t, \boldsymbol{y}_i^t)\}_{i=1}^{N_t}$ comprising N_t images. Here, $\mathcal{I}_i^t$ represents the ith MSI, while its corresponding one-hot encoded semantic label is $\boldsymbol{y}_i^t \in \{0,1\}^C$, with C being the total number of classes. The query set, $\mathcal{D}^q = \{\mathcal{I}_i^q\}_{i=1}^{N_q}$, contains N_q images. These MSIs have multiple bands, each capturing features from different wavelength ranges. A given image $\mathcal{I}_i$ can be divided into M distinct views, each aggregating a subset of these bands. Details on this division can be found in Section IV-A and Tables I and II. The multiview set comprising M combinations of spectral bands for $\mathcal{I}_i$ is denoted as $\mathcal{V}_i = \{V_{i,1}, V_{i,2}, \ldots, V_{i,M}\}$.

TABLE I
Bands Covered by Sentinel-2's Multispectral Imager in the EuroSAT Dataset and Experimental View Partitioning; Data Sourced From [2]

Attributes	Bands	Center Wave Length (μm)	Group1	Group2	Group3	Group4
RGB	B02	0.49	1	1	1	1
	B03	0.56	1	1	1	1
	B04	0.67	1	1	1	1
Red edge	B05	0.70	1	1	2	2
	B06	0.74	1	1	2	2
	B07	0.78	1	1	2	2
	B08A	0.86	1	1	2	2
NIR	B08	0.84	1	2	3	3
Water-related	B01	0.44	—	—	—	—
	B09	0.94	—	—	—	—
	B10	1.38	—	—	—	—
SWIR	B11	1.61	1	2	3	4
	B12	2.19	1	2	3	4
Number of views			1	2	3	4

The aim of SVCNet is to construct a mapping function $\mathcal{F}: \mathcal{V}_i \mapsto \mathbb{R}^d$, transforming the set of views $\mathcal{V}_i$ derived from a MSI $\mathcal{I}_i$ into a d-dimensional unified representation $\mathbf{e}_i = \mathcal{F}(\mathcal{V}_i)$. By combining and adjusting these individual views, SVCNet captures the essence of the MSI. It ensures that semantically similar MSIs have pairwise feature similarity in a metric space, such as cosine space.

Fig. 2 provides an overview of SVCNet, structured in three steps inspired by human communication mechanisms. Initially, in the view understanding phase (Section III-B), each view, through neural networks, generates high-level feature representations. These features, combined with decorrelated semantic anchors, constitute the graph structures in GNN (Section III-C). Subsequently, the GACM (Section III-D) further processes these graph representations. Emulating the diversity and goal-oriented nature of human communication, GACM employs graph attention mechanisms for multiround, goal-oriented view–semantic interactions, achieving selective knowledge integration across views. In the consensus and reflection phase (Section III-E), the proposed SAL function aligns semantic anchors with corresponding category information for supervision, reinforcing the goal-oriented nature of human communication and adaptability in single-label and multilabel RSIR contexts. Concurrently, the TGCL function enhances communication outcomes, drawing semantically similar semantic–view representations closer in the metric space.

TABLE II
BAND CHARACTERISTICS AND VIEW PARTITIONING IN THE SEA FOG AND LSCIDMRV2 DATASETS; DATA SOURCED FROM [4]

Attributes	Bands	Center Wave Length (μm)	Group1	Group2	Group3	Group4†
Visible	albeo_1	0.46	1	1	1	1
	albeo_2	0.51	1	1	1	1
	albeo_3	0.64	1	1	1	2
Near-infrared	albeo_4	0.86	1	1	1	1
	albeo_5	1.6	1	1	1	2
	albeo_6	2.3	1	1	1	3
Infrared	tbb_7	3.9	1	2	2	2
	tbb_8	6.2	1	2	2	4
	tbb_9	7.0	1	2	2	4
	tbb_10	7.3	1	2	2	4
	tbb_11	8.6	1	2	2	2
	tbb_12	9.6	1	2	3	3
	tbb_13	10.4	1	2	3	2
	tbb_14	11.2	1	2	3	2
	tbb_15	12.3	1	2	3	2
	tbb_16	13.3	1	2	3	2
Number of Views			1	2	3	4

† According to [4], views in Group 4 are partitioned based on the discriminative characteristics of the bands.

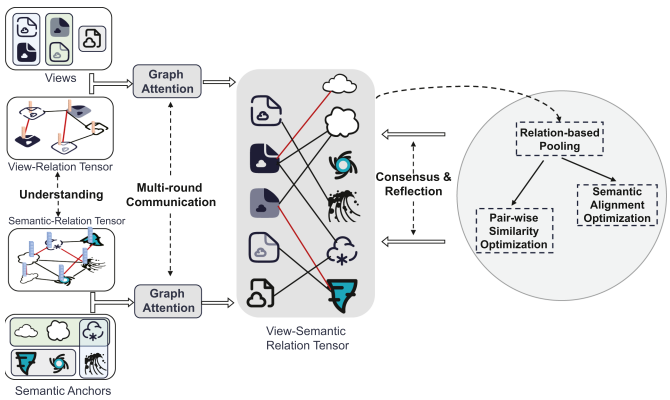

Fig. 2. Overview of the SVCNet architecture. Modeled after human communication principles, SVCNet integrates view-based knowledge and semantic objectives into the learning process. It employs three types of graph structures and utilizes graph attention mechanisms specifically to highlight view-to-semantic relationships. SVCNet is optimized for these relationships using two objective functions, both of which are grounded in the backpropagation mechanism of deep learning, enhancing consensus and reflection in RSIR.

B. View Knowledge Understanding

It is expensive and ineffective to represent each spectral band as an individual view for feature extraction in MSI analysis [4]. Feature extraction is performed on selected combinations of spectral bands using visual feature extractors. Specifically, these extractors are ResNet18 [11] models pretrained on ImageNet. The forward computation for a set of these view combinations, $\mathcal{V}_i$, can be defined as

$$\mathcal{H}\left(\mathcal{V}_i\right)=\left\{h_m\left(V_{i,m}\right)\right\}_{m=1}^{M}=\left\{\boldsymbol{v}_{i,1},\boldsymbol{v}_{i,2},\ldots,\boldsymbol{v}_{i,M}\right\} \tag{3}$$

where $\mathcal{H}=\{h_1,h_2,\ldots,h_M\}$ denotes a set of feature extractors. Each h_m extracts features from the corresponding view $V_{i,m}$. Here, $\boldsymbol{v}_{i,m}\in\mathbb{R}^d$ signifies the feature vector of the mth view for the ith image.

Inspired by works such as [43] and [44], we recognize the unique semantic characteristics often associated with different classes. To capture these attributes, we introduce a set of decorrelated semantic anchors to act as semantic representations. These representations are orthogonalized using the Gram–Schmidt process before training and are utilized in later stages to enrich the semantic understanding of our model. Let $\mathbf{A}\in\mathbb{R}^{C\times d}$ be a randomly initialized matrix of semantic anchors, where C denotes the number of classes. We utilize the Gram-Schmidt orthogonalization to produce a set of decorrelated semantic anchors, i.e.,

$$\overline{\boldsymbol{a}}_i=\begin{cases}\mathcal{N}\left(\boldsymbol{a}_i\right) & \text{if } i=1\\ \mathcal{N}\left(\boldsymbol{a}_i-\sum_{j=1}^{i-1}\left(\frac{\boldsymbol{a}_i\cdot\overline{\boldsymbol{a}}_j}{\overline{\boldsymbol{a}}_j\cdot\overline{\boldsymbol{a}}_j}\right)\overline{\boldsymbol{a}}_j\right) & \text{if } i>1\end{cases} \tag{4}$$

where $\boldsymbol{a}_i$ is the ith row of the matrix $\mathbf{A}$, representing the semantic anchor prior to orthogonalization. The function $\mathcal{N}(\cdot)$ normalizes vector to unit length. Following orthogonalization, the anchors are assembled into a new matrix $\overline{\mathbf{A}}=[\overline{\boldsymbol{a}}_1,\overline{\boldsymbol{a}}_2,\ldots,\overline{\boldsymbol{a}}_C]^T$, thereby generating a set of decorrelated semantic anchors.

C. Graph Network Construction

After obtaining the view and semantic anchor representations, we proceed to construct various forms of graphs to capture their relationships. To provide context for this construction process, we introduce some preliminary definitions.

Definition 1. Node-relation tensor: Consider a graph $\mathcal{G}=(\mathcal{V},\mathcal{E})$, where $\mathcal{V}=\{V_1,V_2,\ldots,V_K\}$ represents the set of K nodes, that is, $|\mathcal{V}|=K$. Each node V_i is associated with a d-dimensional feature vector $\mathbf{x}_i\in\mathbb{R}^d$. We define the function $\mathcal{R}:\mathcal{V}\to\mathbb{R}^{K\times K\times 2d}$ to construct the tensor capturing the relations between nodes as follows:

$$\mathcal{R}(\mathcal{V})=\begin{bmatrix}\mathcal{C}\left(\mathbf{x}_1,\mathbf{x}_1\right) & \mathcal{C}\left(\mathbf{x}_1,\mathbf{x}_2\right) & \ldots & \mathcal{C}\left(\mathbf{x}_1,\mathbf{x}_K\right)\\ \mathcal{C}\left(\mathbf{x}_2,\mathbf{x}_1\right) & \ddots & & \vdots\\ \vdots & & \ddots & \vdots\\ \mathcal{C}\left(\mathbf{x}_K,\mathbf{x}_1\right) & \ldots & \ldots & \mathcal{C}\left(\mathbf{x}_K,\mathbf{x}_K\right)\end{bmatrix}. \tag{5}$$

Here, $\mathcal{C}(\cdot)$ denotes the concatenation of two feature vectors, such that $\mathcal{C}(\mathbf{x}_i,\mathbf{x}_j)=[\mathbf{x}_i;\mathbf{x}_j]$

Definition 2. Multigraph connection tensor: Let two graphs $\mathcal{G}_1=(\mathcal{V}_1,\mathcal{E}_1)$ and $\mathcal{G}_2=(\mathcal{V}_2,\mathcal{E}_2)$ be given, where $\mathcal{V}_1=\{V_1^{(1)},\ldots,V_{K_1}^{(1)}\}$ and $\mathcal{V}_2=\{V_1^{(2)},\ldots,V_{K_2}^{(2)}\}$, with $|\mathcal{V}_1|=K_1$ and $|\mathcal{V}_2|=K_2$. Each node $V_i^{(1)}$ possesses a d_1 dimensional feature vector $\mathbf{x}_i^{(1)}\in\mathbb{R}^{d_1}$, and each node $V_j^{(2)}$ has a d_2-dimensional feature vector $\mathbf{x}_j^{(2)}\in\mathbb{R}^{d_2}$. We define a multigraph connection tensor constructor $\mathcal{Q}:\mathcal{V}_1,\mathcal{V}_2\to\mathbb{R}^{K_1\times K_2\times(d_1+d_2)}$ as follows:

$$\mathcal{Q}\left(\mathcal{V}_1,\mathcal{V}_2\right)$$

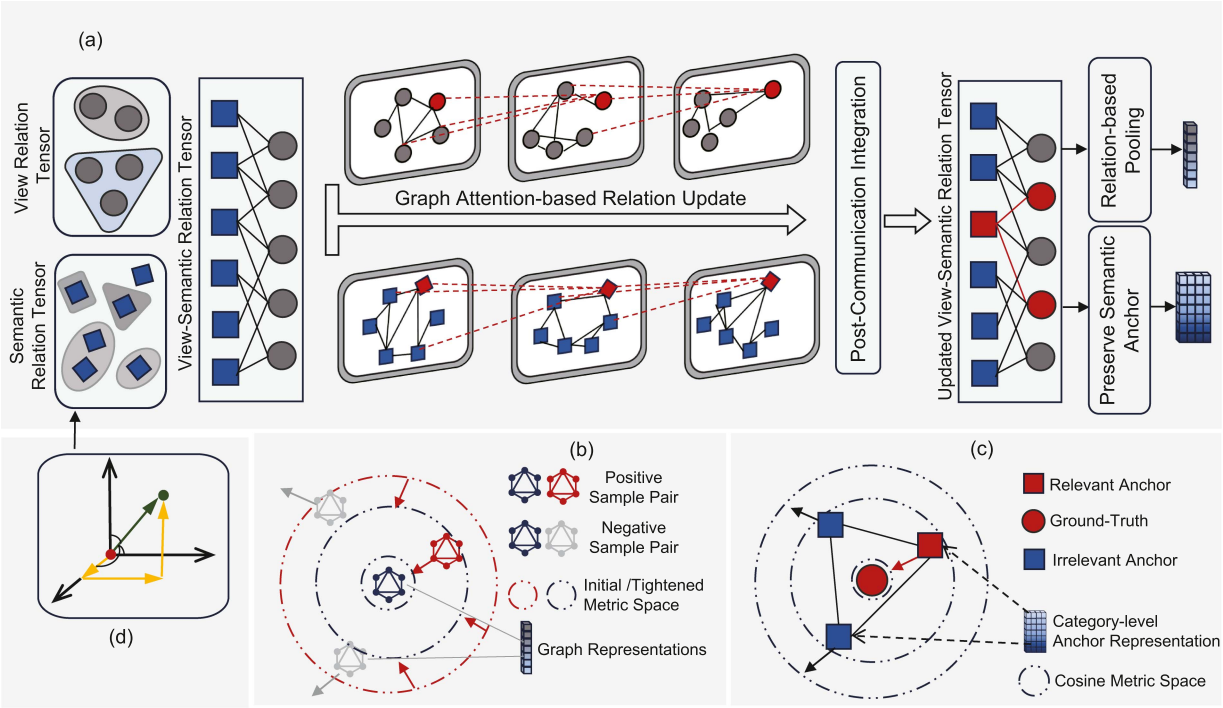

Fig. 3. Graphical representation of key components in the proposed SVCNet. (a) GACM, which updates the interrelationships among views, semantics, and view–semantics. (b) TGCL, a loss function designed for refined learning in multilabel settings. (c) SAL function for aligning labels with learned semantic signals. (d) Initialization procedure achieved via anchor decorrelation through Schmidt orthogonalization. (a) Graphical representation of graph attention-based multi-round communication module(GACM). (b) Diagram of tightened graph contrastive loss. (c) Diagram of semantic alignment loss. (d) Anchor de-correlation via Schmidt orthogonalization.

$$= \begin{bmatrix} \mathcal{C}(\mathbf{x}_1^{(1)}, \mathbf{x}_1^{(2)}) & \mathcal{C}(\mathbf{x}_1^{(1)}, \mathbf{x}_2^{(2)}) & \dots & \mathcal{C}(\mathbf{x}_1^{(1)}, \mathbf{x}_{K_2}^{(2)}) \\ \mathcal{C}(\mathbf{x}_2^{(1)}, \mathbf{x}_1^{(2)}) & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ \mathcal{C}(\mathbf{x}_{K_1}^{(1)}, \mathbf{x}_1^{(2)}) & \dots & \dots & \mathcal{C}(\mathbf{x}_{K_1}^{(1)}, \mathbf{x}_{K_2}^{(2)}) \end{bmatrix}. \quad (6)$$

Building on Definitions 1 and 2, we explore three distinct types of high-order connections within the graph: connections between views, connections between semantic anchors, and connections between views and semantic anchors, as illustrated in Figs. 2 and 3(a). Among these, the view relation tensor aims to highlight key views, while the semantic relation tensor emphasizes crucial semantics formed by contributions from multiple views. As for the connections between views and semantic anchors, this relationship primarily underscores the interdependence between views and semantic information. The construction details are as follows.

1) View Relation Tensor: The view relation graph $\mathcal{G}^v$ comprises nodes $\mathcal{V}^v = \{\boldsymbol{v}_1, \boldsymbol{v}_2, \dots, \boldsymbol{v}_M\}$, representing the multi-view set from Section III-A (we omit the sample index i for simplicity). With $|\mathcal{V}^v| = M$, each node in $\mathcal{G}^v$ signifies a distinct view. We define the view relation tensor as $\mathbf{E}_v \in \mathbb{R}^{M\times M\times 2d}$, which can be constructed using the node relation tensor function $\mathcal{R}$ from Definition 1 as $\mathbf{E}_v = \mathcal{R}(\mathcal{V}^v) \in \mathbb{R}^{M\times M\times 2d}$

2) Semantic Relation Tensor: Similar to the construction of the view relation tensor, the semantic relation graph $\mathcal{G}^s$ includes nodes $\mathcal{V}^s = \{\overline{\boldsymbol{a}}_1, \overline{\boldsymbol{a}}_2, \dots, \overline{\boldsymbol{a}}_C\}$, where $\overline{\boldsymbol{a}}$ are the orthogonalized semantic anchors formed according to (4). With $|\mathcal{V}^s| = C$, the semantic relation tensor $\mathbf{E}_s = \mathcal{R}(\mathcal{V}^s) \in \mathbb{R}^{C\times C\times 2d}$ aims to learn the relationships between different semantics and to foster connections that are semantically representative.

3) View–Semantic Relation Tensor: The view–semantic relation graph $\mathcal{G}^{sv}$ consists of nodes $\mathcal{V}^{sv} = \mathcal{V}^s \cup \mathcal{V}^v$. Based on (6) from Definition 2, the view–semantic relation tensor $\mathbf{E}_{sv} = \mathcal{Q}(\mathcal{V}^s, \mathcal{V}^v) \in \mathbb{R}^{C\times M\times 2d}$.

We opt for relation tensors instead of mere similarity scores to describe the connections between views and semantics. This approach enriches the relationship by concatenating features, offering a more comprehensive representation of the traits shared between two nodes. Compared to relying solely on a single similarity score, such feature concatenation captures more complex and multidimensional relationships and interactions among nodes. This not only facilitates the construction of intricate graph network structures but also enables more efficient learning of various characteristics of the relationships based on these structures.

D. Graph Attention-Based Multiround Communication Module

Fig. 3(a) illustrates the computational process of the GACM, which emulates the multiround human communication process in information acquisition, evaluation, and knowledge updating. To achieve this, we employ two parallel sets of graph attention networks to handle the encoding of view and semantic relation tensor, respectively. After the view understanding phase described in Section III-B and before commencing with the communication phase, we obtain initial view and semantic relation

tensors corresponding to multiview representations and initial semantic representations. These are denoted as $\mathbf{E}_v^{(0)}, \mathbf{E}_s^{(0)}$, and $\mathbf{E}_{sv}^{(0)}$, which are assembled according to the methods detailed in Section III-C, forming the view relation tensor, semantic relation tensor, and view–semantic relation tensor. The procedure for the rth round of communication is as follows.

1) *Computing communication weights:* Initially, attention weights $\mathbf{w}_{i,n}$ are calculated based on the relational associations among nodes. Here, i signifies the ith node. The set $\mathcal{N}(i)$ represents all nodes that are connected to node i. Thus, $n \in \mathcal{N}(i) \cup \{i\}$ refers to the index set of all nodes connected to node i, including i itself. Specifically, the attention weight $\mathbf{w}_{i,n}$ that governs the communication between the ith and nth nodes is computed as follows:

$$\mathbf{w}_{i,n} = \text{softmax}\left(\Psi_2\left(\sigma\left(\Psi_1\left(\mathbf{E}_{i,n}^{(r-1)}\right)\right)\right)\right) \tag{7}$$

where $\mathbf{E}_{i,n}^{(r-1)} \in \mathbb{R}^{2d}$ represents the relationship tensor between nodes after the $r-1$th round of communication. The functions $\Psi_1 : \mathbb{R}^{2d} \to \mathbb{R}^{d'}$ and $\Psi_2 : \mathbb{R}^{d'} \to \mathbb{R}^1$ serve as learnable linear mappings. The dimension d' is the hidden layer dimension post-linear layer mapping, and the activation function $\sigma(\cdot)$ is the LeakyReLU function.

2) *Node relationship update:* The updated node representation in a round of communication can be computed as

$$\mathbf{x}_i' = \sum_{j=1}^{n} \mathbf{w}_{i,j} \Psi_3\left(\mathbf{x}_j\right),$$

where $\Psi_3 : \mathbb{R}^d \to \mathbb{R}^d$ denotes a learnable linear mapping. This formulation applies to connections between both views and semantic anchors. Specifically, for the semantic relation graph, $\mathbf{x}_j = \overline{\boldsymbol{a}}_j$ and $\mathbf{x}_i' = \overline{\boldsymbol{a}}_i'$ with $n = C$. For the view relation graph $\mathcal{G}^v, \mathbf{x}_j = \boldsymbol{v}_j$ and $\mathbf{x}_i' = \boldsymbol{v}_i'$ with $n = M$. After computing all the updated representations, we sequentially assemble them into updated relationship tensors $\mathbf{E}_s^{(r)} = \mathcal{R}([\overline{\boldsymbol{a}}_1', \overline{\boldsymbol{a}}_2', \ldots, \overline{\boldsymbol{a}}_C'])$ and $\mathbf{E}_v^{(r)} = \mathcal{R}([\boldsymbol{v}_1', \boldsymbol{v}_2', \ldots, \boldsymbol{v}_M'])$.

3) *Postcommunication aggregated representation:* After multiple rounds of communication, we assemble the updated graph nodes, denoted as $\mathcal{V}_v' = \{\boldsymbol{v}_1', \boldsymbol{v}_2', \ldots, \boldsymbol{v}_M'\}$ for view nodes and $\mathcal{V}_s' = \{\overline{\boldsymbol{a}}_1', \overline{\boldsymbol{a}}_2', \ldots, \overline{\boldsymbol{a}}_C'\}$ for semantic nodes. These are then processed through a graph-based aggregation mechanism. Specifically, the aggregation operation is formulated as

$$\mathbf{E}_{sv}' \leftarrow \left(\Psi_{sv}^1 \mathbf{E}_{sv}^{(0)} + \Psi_{sv}^2 \mathcal{Q}\left(\mathcal{V}_s', \mathcal{V}_v'\right)\right)_{i,j} \tag{8}$$

where the notation $(\cdot)_{i,j}$ specifies elementwise updates for each row and column vector in $\mathbf{E}_{i,j}$, $\Psi_{sv}^1, \Psi_{sv}^2 : \mathbb{R}^{2d} \to \mathbb{R}^d$ are learnable linear mappings, $i = 1, \ldots, C$ and $j = 1, \ldots, M$, $\mathcal{Q}$ being defined in Definition 2, (6). We then proceed to a final pooling step to aggregate these relationships. The aggregated representation is calculated as: $\mathbf{e} = \frac{1}{C \cdot M} \sum_{i=1}^{C} \sum_{j=1}^{M} (\mathbf{E}_{sv}')_{i,j}$. This aggregated representation serves as a summarized yet comprehensive encapsulation of the relationships and features learned across the network. Furthermore, all the updated semantic anchors are retained for subsequent optimization, as discussed in Section III-E.

E. Objective Functions

In traditional retrieval methods, two challenges arise when dealing with semantically rich, multispectral, multilabel scenarios. First, conventional classification loss functions often force features to converge toward class centers, resulting in a lack of diversity. Second, these traditional approaches struggle with ambiguous boundaries between positive and negative samples in semantically rich, multilabel environments, leading to representations with limited discriminative power. To tackle these issues and achieve more precise semantic learning, as well as to foster enhanced collective consensus and collective reflection, we introduce two novel optimization techniques: SAL function and TGCL function.

We use the updated set of semantic anchors $\mathbf{A}' = [\overline{\boldsymbol{a}}_1', \ldots, \overline{\boldsymbol{a}}_C']^T \in \mathbb{R}^{C \times d}$ from Section III-D as representatives for specific labels or categories, thereby enriching the model with additional semantic information. Cosine similarity is used as a metric to compute the relationship between each d-dimensional sample feature $\mathbf{e} \in \mathbb{R}^d$ and each semantic anchor. Specifically, the category similarity vector $\boldsymbol{s} \in \mathbb{R}^C$ is computed as follows:

$$\boldsymbol{s}_c = \frac{\langle \mathbf{e}, \overline{\boldsymbol{a}}_c' \rangle}{\|\mathbf{e}\|_2 \cdot \|\overline{\boldsymbol{a}}_c'\|_2}. \tag{9}$$

Here, $\langle \cdot \rangle$ denotes the inner product, and $\boldsymbol{s}_c \in \boldsymbol{s} \quad \forall c \in \{1, \ldots, C\}$.

The proposed SAL aims to align the similarity between semantic anchors and the correct categories. It utilizes a weighted version of the cross-entropy loss $\mathcal{L}_c^{\text{CE}}$ to achieve this alignment. The standard cross-entropy loss $\mathcal{L}_c^{\text{CE}}$ is given by

$$\mathcal{L}_c^{\text{CE}}(\boldsymbol{s}_c, \boldsymbol{y}_c) = -\left[\boldsymbol{y}_c \log\left(\boldsymbol{s}_c\right) + \left(1 - \boldsymbol{y}_c\right) \log\left(1 - \boldsymbol{s}_c\right)\right] \tag{10}$$

and

$$\mathcal{L}^{SA} = \frac{1}{C} \sum_{c=1}^{C} \boldsymbol{\pi}_c \left(1 - \exp\left(\mathcal{L}_c^{\text{CE}}\right)\right)^{\gamma} \mathcal{L}_c^{\text{CE}} \tag{11}$$

where $\boldsymbol{\pi}_c$ and γ function as weighting coefficients that prioritize the contribution of rarer categories to the loss function.

In retrieval tasks, optimization methods based on contrastive learning between sample pairs are critical. Typically, samples sharing zero overlapping labels are treated as negative pairs. However, in semantically rich multilabel settings, the boundary between positive and negative samples becomes blurred due to partial label overlap. To tackle this, we introduce a tightening coefficient to fine-tune the delineation between these samples. In an optimization batch containing b samples, let the sample label matrix be denoted as $\mathbf{Y} = [\boldsymbol{y}_1, \ldots, \boldsymbol{y}_b]^T \in \mathbb{R}^{b \times C}$. In the context of a multilabel setting, the threshold vector $\boldsymbol{\delta}$ is computed as follows:

$$\mathbf{S} = \mathbf{Y}\mathbf{Y}^{\top} - \text{diag}\left(\mathbf{Y}\mathbf{Y}^{\top}\right)$$

$$\boldsymbol{\delta} = \frac{1}{b-1} \sum_{j=1}^{b} \left(\mathbf{S}_{i,j}\right), \quad i = 1, 2, \ldots, b \tag{12}$$

here, $\mathbf{S} \in \mathbb{R}^{b\times b}$ serves as the label sharing matrix, capturing the degree of label overlap between pairs of samples within an optimization batch. The threshold vector $\boldsymbol{\delta}$ further reflects the level of label sharing for each label within the batch. Samples surpassing this level of sharing are likely to belong to highly similar remote sensing scenes and should be pulled closer in the metric space. Conversely, samples falling below this threshold should be pushed apart. To maintain optimization flexibility, we employ cosine similarity as the metric space for feature optimization, an improvement based on [27]. We define sets of relatively positive and negative samples as $\mathcal{P}_i = \{j \mid \mathbf{S}'_{i,j} > \boldsymbol{\delta}_i, j = 1, 2, \ldots, b; j \neq i\}$ and $\mathcal{N}_i = \{j \mid \mathbf{S}'_{i,j} \leq \boldsymbol{\delta}_i, j = 1, 2, \ldots, b; j \neq i\}$. These sets include sample indices that display a notable level of semantic similarity (or dissimilarity) with sample i, as determined by the threshold $\boldsymbol{\delta}_i$, losses for relatively positive samples and relatively negative samples are then defined accordingly

$$\mathcal{L}_{\text{pos}}^{\text{TGC}} = \frac{1}{\alpha}\log\left[1 + \sum_{k\in\mathcal{P}_i} e^{-\alpha(\mathcal{S}_{ik}-1)}\right]$$
$$\mathcal{L}_{\text{neg}}^{\text{TGC}} = \frac{1}{\beta}\log\left[1 + \sum_{k\in\mathcal{N}_i} e^{\beta(\mathcal{S}_{ik}-1)}\right] \tag{13}$$

where $\mathcal{S}_{ik}$ represents the cosine similarity between the graph representations of the ith and kth samples. Specifically, it is calculated as: $\mathcal{S}_{ik} = \langle \mathbf{e}^i \cdot \mathbf{e}^k \rangle / (\| \mathbf{e}^i \|_2 \cdot \|\mathbf{e}^k\|_2)$, $\alpha = \sum_{j=1,j\neq i}^{b} \mathcal{I}(\mathbf{S}'_{i,j} > \boldsymbol{\delta}_i)$ and $\beta = \sum_{j=1,j\neq i}^{b} \mathcal{I}(\mathbf{S}'_{i,j} \leq \boldsymbol{\delta}_i)$ serve as weighting factors. These are calculated based on the number of relatively positive and negative samples, respectively, and are utilized to balance their respective influences, $\mathcal{I}$ denotes the indicator function. Note that in a single-label environment, TGCL degenerates into a standard metric loss in [27], as the number of shared labels for positive samples always exceeds the threshold, while for negative samples it always falls below the threshold. Ultimately, the aggregate loss $\mathcal{L}$ is computed as weighted sum of the SAL and the TGCL

$$\mathcal{L} = \underbrace{\mathcal{L}^{SA}}_{\text{Semantic}} + \lambda \underbrace{\left(\mathcal{L}_{\text{pos}}^{\text{TGC}} + \mathcal{L}_{\text{neg}}^{\text{TGC}}\right)}_{\text{Pairwise contrast}}. \tag{14}$$

Here, $\lambda = 0.5$ serves as a hyperparameter to balance the contributions of the two loss terms in the overall loss function $\mathcal{L}$.

IV. Experiments

A. Datasets

To evaluate SVCNet's retrieval performance in Earth observation datasets, experiments were conducted on three MSI datasets for single-label and multilabel RSIR tasks. These datasets support a range of satellite remote sensing applications, from remote sensing scene recognition to large-scale, multilabel cloud and weather system identification, and fine-grained meteorological disaster recognition.

EuroSAT [2]. Single-label land use and land cover dataset: The Sentinel-2-sourced EuroSAT dataset contains 27 000 labeled MSIs across 10 single-label land-use classes, spanning 13 spectral bands. View partitioning and band characteristics are detailed in Table I. RGB images are generated from Bands B2, B3, and B4; near-infrared (NIR) and red-edge features are captured by Bands B5 to B8; and short-wave infrared (SWIR) information is provided by Bands B11 and B12. Bands B1, B9, and B10 are discarded as they are primarily used for detecting aerosols or atmospheric water vapor. We randomly allocated 50% of the data for training and validation, while the remaining 50% was reserved for testing.

LSCIDMRv2 [4]. Multilabel land cover, cloud, and weather systems recognition dataset: The LSCIDNRv2 dataset, sourced from the Himawari-8 satellite, comprises 16 spectral bands and 17 diverse labels in a multilabel setting. However, the dataset exists in a condition of significant class imbalance; for instance, Ocean accounts for 90% and Cirrus for over 80%. Classes such as frontal surface (FS) and westerly jet (WJ) are minimally represented, making up less than 1%, which complicates model training. To mitigate this issue, we curated a subset that focuses on five critical weather systems: tropical cyclone (TC), extratropical cyclone (EC), FS, WJ, and stratus (St). Any sample containing at least one of these five labels was included in this subset. The label count statistics for this adjusted dataset are depicted in Fig. 4(a). Despite these adjustments, the low sample sizes for categories like WJ and FS continue to pose challenges for retrieval and recognition. Further details on view partitioning and band characteristics are provided in Table II. The division of the dataset into training, validation, and testing sets follows the guidelines provided by the dataset creators.

Sea Fog [3]. Single-label fine-grained meteorological hazard recognition dataset: The Sea Fog dataset, sourced from the Himawari-8 satellite, features 16 spectral bands and includes four types of image labels: Clear Sky, Low Clouds, Mid–High Clouds, and Sea Fog. The distribution of these labels is illustrated in Fig. 4(b). Notably, as shown in Fig. 4(c), the physical properties of Low Clouds and Sea Fog are remarkably similar, presenting a significant challenge for differentiation based on this dataset [3]. For dataset partitioning, we follow the methodology outlined in [3], dividing the data into training/validation and test sets. Fig. 4(b) provides further details on label distribution.

B. Implementation Details

Experimental Setup: We implemented the proposed model using the PyTorch framework and conducted all experiments on an NVIDIA 3080Ti GPU with 16 GB of memory. During the training phase, we employed the ADAM optimizer with a learning rate set at 0.0001 and a batch size of 100. We used random rotation and vertical flipping for data augmentation. In the SAL, denoted in (11), the weight $\boldsymbol{\pi}_c$ was specifically set to 20 for rare classes, such as FS and WJ, in the context of the LSCIDMRv2 dataset, with a γ value of 5. In other datasets, the weight vector $\boldsymbol{\pi}_c$ in (11) was set as an all-ones vector, and γ was set to 0.

The datasets were all subjected to fivefold cross-validation, wherein 20% of the training data was randomly partitioned into a validation set. This validation set was employed during the training process for hyperparameter tuning. The best-performing set

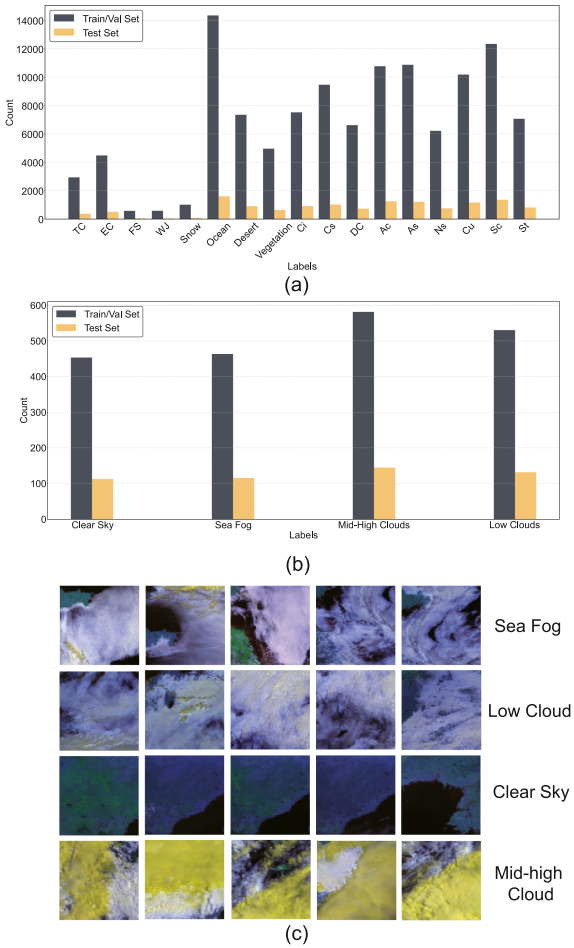

Fig. 4. Data label statistics and sample examples from the Sea Fog Dataset. (a) Label distribution in LSCIDMRv2 dataset. (b) Label distribution in Sea Fog dataset. (c) Examples of different categories in Sea Fog dataset. In (c), RGB color images were generated for display purposes using bands B03, B04, and B05 sourced from the Himawari-8 satellite.

of hyperparameters was subsequently chosen for the final model. Upon identifying the optimal hyperparameters, the model was retrained, and the version demonstrating the best performance on the validation set was saved for further evaluation on an independent test set.

Evaluation Protocol: In multilabel retrieval tasks, we employed three multilabel retrieval evaluation metrics: average cumulative gain (ACG), weighted mean average precision (wAP), and normalized discounted cumulative gain (NDCG) [45]. ACG measures the average number of shared labels between the query and the retrieved images. NDCG serves as a normalized measure for assessing the quality of the ranking in the retrieval results. wAP represents the average retrieval precision across all query images. For single-label retrieval tasks, we employed three metrics—mean average precision (mAP), precision (Prec), and NDCG—to evaluate performance.

To objectively assess the model's performance and application potential, we also report on the recognition (classification) performance based on retrieval results. Specifically, we consider the top-k retrieval results, where if a category appears more than $k/2$ times among the returned k MSIs, the sample is classified as belonging to that category. This method is applicable to both single-label and multilabel classification tasks, with k set to 100 in all our experiments. For multilabel classification based on retrieval results, we adopted classwise average F1-score (cF1), classwise average accuracy (cACC), and average precision (AP) as evaluation metrics, following [4]. For single-label classification, we used cACC, cF1, and Recall to assess performance.

Competing Methods: To comprehensively assess the performance of our proposed SVCNet, we selected a variety of state-of-the-art (SOTA) retrieval and classification algorithms from the literature as comparison benchmarks. To ensure the fairness of the evaluation, we meticulously replicated all the compared methods mentioned in the literature. These were retrained, and their results were reported within the same training, validation, and testing set environments across the three datasets. For hyperparameter selection, we followed the recommendations provided by the original authors of each competing method. Like SVCNet, we employed a fivefold cross-validation scheme to identify the optimal hyperparameters, reporting results obtained under these optimal settings. Below are detailed descriptions of the methods specifically chosen for comparison.

1) Baseline Models: To ensure a fair comparison, we employ ResNet50 [11], DenseNet121 [26], and Vision Transformer (ViT) [16] as our baseline models (we utilize the ViT-base variant based on experimental validation). All are pretrained on ImageNet. In these baselines, we concatenate all spectral information along the channel dimension. The models are trained using the same tightened contrastive loss as used in (13) for multilabel retrieval. Additionally, we incorporate appropriate classification losses for both single-label and multilabel settings to improve the baseline performance.

2) MS-GOGO [4]. Multilabel multiview satellite image classification method: It uses a set of convolutional layers for multiview feature extraction and integrates geographic and spectral data to improve classification robustness. We replicated this method and evaluated its performance on the LSCIDMRv2 dataset.

3) C-Tran [44]. Multilabel image classification framework: C-Tran uses Transformer architectures and semantic embeddings to handle complex multilabel image classification tasks. We tested this method's effectiveness on the LSCIDMRv2 dataset.

4) GRN [13]. Multilabel remote sensing image retrieval algorithm based on graph relation network: GRN employs graph structures in its network to model relationships between samples (or labels), and utilizes a scalable discriminative loss with binary cross-entropy to enhance training accuracy for retrieval tasks.

5) DMMVH [35]. Deep metric learning-based multiview retrieval method: This method employs gating mechanisms to learn the multiview interactions between multiview features (e.g., text and image). We have appropriately adapted this method, using features extracted by ResNet18 from different spectral combinations as view features and fusing them.

6) CSQ [43]: A method for both single-label and multilabel retrieval that employs metric learning and center loss to capture rich semantics.

TABLE III
COMPARATIVE PERFORMANCE OF CATEGORY-SPECIFIC F1 SCORES AND OVERALL RETRIEVAL AND CLASSIFICATION METRICS ON THE EUROSAT DATASET;THE BEST RESULTS ARE HIGHLIGHTED IN BOLD RED, WHILE THE SECOND-BEST ARE IN BOLD BLUE

Metric Type	Class/Metric	ResNet50	DenseNet121	ViT	CSQ	SCFR	DMMVH	AWNet	DAH	DGSSH	SVCNet
F1 per Class (%)	Annual Crop	87.72	92.28	94.20	92.78	90.38	95.43	93.82	94.79	**95.93**	**96.98**
	Forest	96.65	96.68	**98.15**	97.63	95.79	95.75	95.94	96.76	96.87	**99.13**
	Vegetation	85.34	88.62	**93.70**	92.00	90.57	91.05	90.65	93.49	93.07	**94.59**
	Highway	88.10	89.46	88.53	90.76	90.26	94.92	95.93	96.14	**96.46**	**97.06**
	Industrial	94.92	96.05	94.26	95.72	95.16	**98.10**	96.65	97.11	96.70	**98.57**
	Pasture	87.46	89.40	90.56	90.96	88.51	90.85	89.85	**92.26**	91.95	**94.94**
	Crop	80.74	87.26	87.71	89.55	87.72	90.30	90.01	92.21	**92.85**	**93.30**
	Residential	94.48	95.52	93.92	95.80	95.73	**98.66**	96.94	97.32	97.61	**98.97**
	River	97.07	97.21	96.17	**97.93**	97.76	96.86	96.53	97.63	**97.95**	**97.95**
	Sea Lake	**99.87**	99.73	99.63	99.83	**99.90**	99.47	99.46	99.40	99.57	99.73
Retrieval Metrics (%)	mAP	91.81	93.36	93.88	94.59	93.35	**94.75**	92.15	93.46	94.65	**97.35**
	Precision	90.08	92.53	92.99	93.80	92.45	**93.99**	91.50	93.42	93.94	**97.03**
	NDCG	95.54	96.11	96.57	96.64	96.09	**97.24**	95.10	95.95	96.45	**98.30**
Recognition Metrics(%)	cACC	91.35	93.13	93.79	94.14	93.20	95.09	94.61	95.64	**95.81**	**97.09**
	cF1	91.23	93.22	93.68	94.30	93.18	95.14	**96.57**	95.71	95.89	**97.12**

TABLE IV
COMPARATIVE PERFORMANCE OF CATEGORY-SPECIFIC F1 SCORES AND OVERALL RETRIEVAL AND CLASSIFICATION METRICS ON THE SEA FOG DATASET

Metric Type	Class/Metric	ResNet50	DenseNet121	ViT	CSQ	SCFR	DMMVH	AWNet	DAH	DGSSH	SVCNet(Ours)
F1 per Class (%)	Clear Sky	**99.56**	**100.00**	**99.56**	97.00	**99.56**	**100.00**	**99.56**	99.11	98.26	**100.00**
	Sea Fog	87.93	88.24	89.87	61.60	88.60	**90.76**	85.05	90.13	90.13	**92.11**
	Mid-high Cloud	90.73	91.95	91.76	85.62	**93.60**	90.14	90.66	92.36	93.06	**93.20**
	Low Cloud	82.07	84.00	86.12	64.98	**86.15**	83.33	82.98	85.39	85.82	**87.12**
Retrieval Metrics (%)	mAP	87.69	**90.25**	89.54	73.16	90.54	89.84	77.88	78.08	79.54	**91.32**
	Precision	85.92	**89.15**	88.25	70.40	89.62	88.09	76.86	74.87	79.13	**90.40**
	NDCG	94.33	95.28	94.96	88.00	94.61	**95.35**	88.74	90.65	89.66	**96.00**
Recognition Metrics(%)	cACC	90.11	91.13	91.74	77.72	**91.94**	91.17	89.35	91.70	**91.94**	**93.03**
	cF1	90.07	91.04	91.83	77.30	**91.98**	91.06	89.56	91.74	91.81	**93.10**

The highest results are highlighted in bold red, while the second-highest are in bold blue.

7) FDRL [46]. Multilabel remote sensing image retrieval method: This method employs feature disentanglement guided by label similarity and mutual learning. Using the BNInception network for multiscale feature extraction, it partitions features into high and low correlation categories to enhance multilabel retrieval robustness.

8) SCFR [12]. Supervised contrastive learning for single-label remote sensing image retrieval: This method uses supervised contrastive learning to refine feature distributions and incorporates a dynamic center loss to improve retrieval performance.

9) DGSSH [47]: A deep global semantic structure-preserving RSIR framework that utilizes the Swin Transformer architecture to extract global and multiscale features of RSI. DGSSH introduces a corrective triplet loss to learn discriminative and stable remote sensing image representations.

10) DAH [48]: A deep attention-based retrieval method with distance-adaptive ranking. The deep attention mechanism captures critical details of the remote sensing images and suppresses irrelevant regional responses. It also introduces a balanced pairwise weighted loss function with a distance-adaptive ranking strategy to fully leverage class information.

11) AWNet [49]: An adaptive weighted learning network for RSIR, AWNet dynamically adjusts the weighting parameters of the combined active–passive loss function based on two metrics, enhancing the discriminative ability and retrieval accuracy of RSI representations.

C. Experimental Results for Single-Label Retrieval Task

In this section, we evaluate the performance of our proposed SVCNet and compare it with other SOTA methods on the EuroSAT and Sea Fog datasets. Numerical results are detailed in Tables III and IV, and visual comparisons are illustrated in Fig. 5.

1) Main Results on EuroSAT:: Table III and Fig. 5 present the principal findings on the EuroSAT dataset. As indicated by these results, SVCNet performs better in overall performance compared to the other methods. Observations from Table III reveal that multiview methods, specifically DMMVH and SVCNet, hold a distinct advantage over traditional single-view models that merely concatenate multispectral data along the channel dimension. The marginal performance differences among single-view methods further emphasize the importance of MVL. Moreover, among the multiview approaches, SVCNet demonstrates the highest accuracy, exceeding DMMVH by 1.98% in F1 score. This suggests that SVCNet's integration of graph-based learning is effective in capturing both view-specific and semantic information, leading to more discriminative representations.

2) Main Results on Sea Fog:: In the Sea Fog dataset, sea fog acts as a natural hazard, making its accurate identification particularly crucial. Although the contrastive learning method SCFR is capable of extracting discriminative representations, its performance on identifying sea fog may leave something to be desired, closely resembling the baseline methods. Our proposed

TABLE V
COMPARISON OF MULTILABEL RETRIEVAL AND RECOGNITION PERFORMANCE ON THE LSCIDMRV2 DATASET

Algorithm Task Type	Method	Retrieval Metrics			Classification Metrics		
		ACG	NDCG(%)	wAP	ACC(%)	F1(%)	AP(%)
Classification	C-Tran [44]	—	—	—	87.92	74.86	87.41
	MS-GOGO [4]	—	—	—	**89.55**	**81.10**	**88.49**
Retrieval	ResNet50 [11]	6.52	82.28	6.58	86.01	75.69	84.44
	DenseNet121 [26]	6.64	83.58	6.70	87.71	79.47	86.58
	ViT [16]	6.70	84.35	6.77	87.96	79.49	85.95
	FDRL [46]	6.47	82.18	6.54	86.37	75.72	83.61
	DMMVH [35]	6.44	82.08	6.53	86.68	79.17	86.98
	GRN [13]	6.08	78.33	6.15	84.63	66.91	79.53
	CSQ [43]	6.22	80.22	6.30	83.28	67.39	74.75
	AWNet [49]	6.82	85.47	6.87	89.31	79.93	85.83
	DAH [48]	6.87	86.08	6.92	89.50	80.24	86.61
	DGSSH [47]	**6.91**	**86.76**	**6.97**	**89.71**	80.51	86.80
	SVCNet (Ours)	**7.00**	**86.86**	**7.03**	89.52	**82.73**	**89.02**

The highest results are highlighted in bold red, while the second-highest are in bold blue.

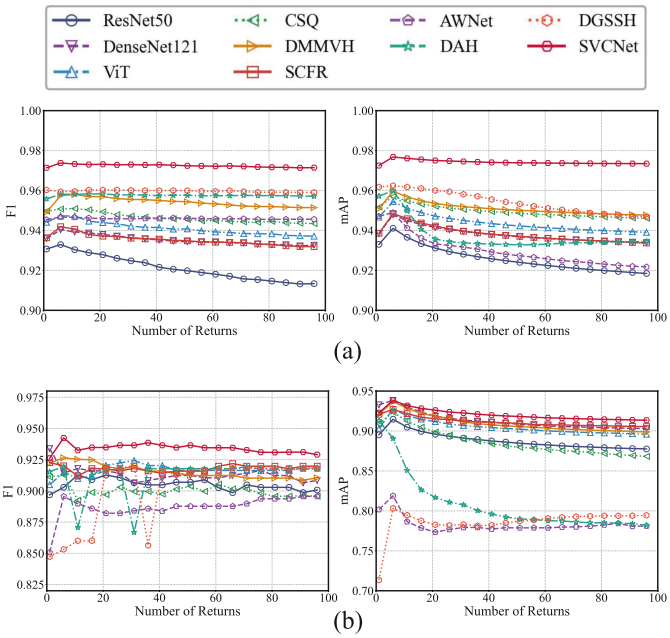

Fig. 5. Retrieval and classification performance across a range of returns on EuroSAT and Sea Fog datasets. (a) EuroSAT. (b) Sea Fog.

method outperforms the second-best method, SCFR, by 3.51% in sea fog identification accuracy and surpasses other multiview methods by 1.35% in accuracy. In terms of overall performance, our proposed method exceeds the second-best method by 1.12%, demonstrating its superior ability to learn discriminative semantics in fine-grained samples, thereby achieving improved retrieval and identification.

D. Experimental Results for Multilabel Retrieval Task

Table V and Fig. 6 show how SVCNet compares to benchmarked methods on the LSCIDMRv2 dataset across retrieval and classification tasks. Based on these results, the following main observations emerge.

1) Optimal performance: SVCNet achieves the best performance, highlighting the efficacy of incorporating human-communication-inspired MVL to effectively mine semantic information. This allows for more accurate multilabel retrieval in complex remote sensing environments. Compared to the second-best method, ViT, SVCNet improves the ACG by 0.30, indicating its ability to retrieve samples with more shared labels while maintaining multilabel accuracy. Additionally, it surpasses the advanced graph-relation-based retrieval method, GRN, by increasing the category-average F1 score by 15.82%, which suggests that the retrieved multilabel samples are both more similar and more accurate.

2) Multiview advantage: Methods that employ MVL, such as SVCNet, MS-GOGO, and DMMVH, stand out in classification metrics like F1 and multilabel AP, underscoring the advantages of MVL.

3) Discriminative feature representation: While SVCNet is not specifically tailored for classification, it either matches or outperforms specialized classification methods like MS-GOGO and C-Tran. This suggests robust discriminative feature representation within SVCNet.

E. Robustness Analysis in Multiview Settings

We conducted experiments to evaluate the robustness of SVCNet in varying MVL contexts. Spectral bands were classified based on their physical properties, as detailed in Tables I and II. Results for different view groupings on the Himawari dataset, corresponding to LSCIDMRv2 and Sea Fog datasets, are depicted in Fig. 8, while findings related to the Sentinel-2 satellite (corresponding to the EuroSAT dataset) are presented in Fig. 8. Key observations from Figs. 7 and 8 include the following. 1) Adding more views in multilabel retrieval enhances performance. Specifically, the ACG rises from approximately 6.6 in a single-view setting (Group 1) to around 6.9 in a multiview context (Group 2). 2) In single-label retrieval tasks, fewer views can occasionally result in reduced performance, particularly in complex scenarios such as sea fog detection. This limitation is mitigated by incorporating more views, leading to more accurate detection and corroborating our initial hypothesis that MVL offers advantages in handling complex tasks. 3) Overall, partitioning data into multiple spectral bands facilitates better

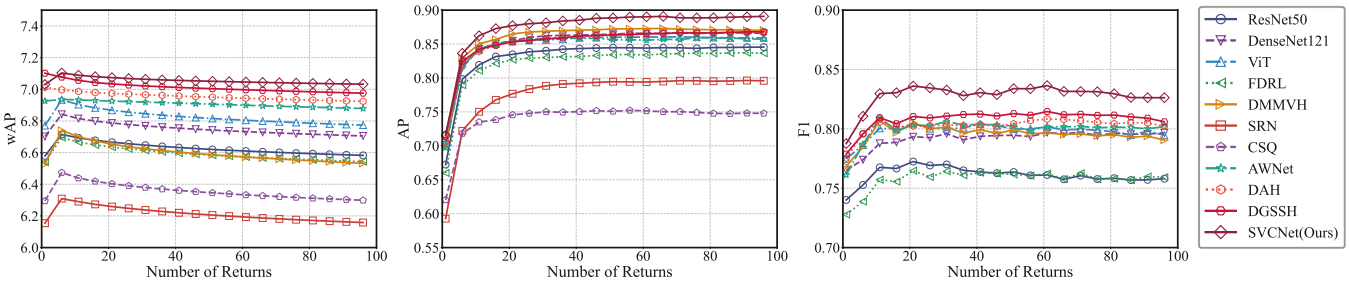

Fig. 6. Retrieval and classification performance across a range of returns on LSCIDMRv2 dataset.

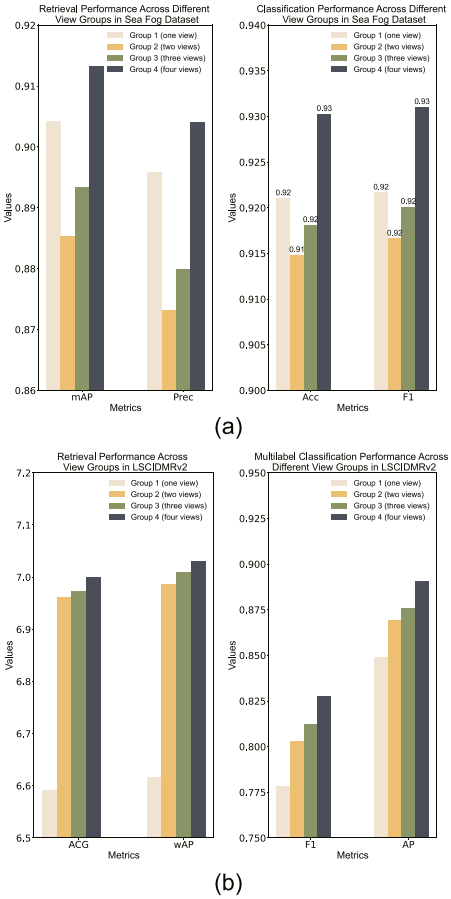

Fig. 7. Bar chart of experimental results on view groups from the Himawari-8 satellite. Definitions for Groups 1–4 can be found in Table II. (a) Retrieval performance and classification performance across different view groups in the Sea Fog dataset. (b) Retrieval performance and multilabel classification performance across different view groups in the LSCIDMRv2 dataset.

information exchange within graph networks, thereby boosting performance in satellite image recognition.

F. Module Functionality Analysis

To rigorously explore the effects of individual components and steps on model performance, we executed a set of ablation studies. These experiments scrutinize key elements: GACM, decorrelated semantic anchors (DSA), TGCL, and SAL. Let us first delineate the following core configurations.

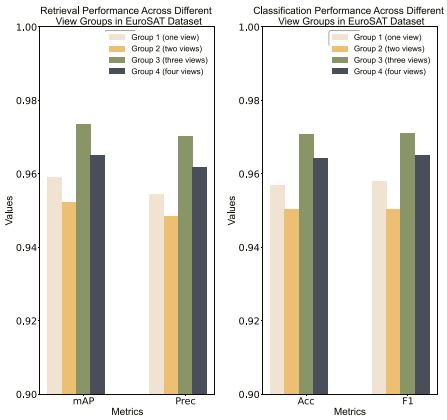

Fig. 8. Bar chart of experimental results on view groups from the Sentinel-2 satellite. Definitions for Groups 1–4 can be found in Table I.

1) Base: This setup includes only the foundational multiview feature extraction as described in Section III-B. ResNet18 is employed for feature extraction from each view, followed by concatenation and dimensionality reduction via a fully-connected layer. To ensure a fair comparison, the base model is optimized using metric loss [27] for retrieval and cross-entropy loss for semantic enhancement.

2) Base+GACM: Building on the "Base" configuration, we extend the model by incorporating the GACM module as outlined in Section III-D. Semantic anchors are initialized randomly and subjected to graph pooling to produce multiview sample representations. These are subsequently assessed for sample similarity using multisimilarity loss [27] and further enriched with cross-entropy loss for semantic guidance.

3) Base+GACM+DSA: Building on the "Base+GACM" setup, this configuration incorporates DSA as described in Section III-B, aiming to further enhance the discriminative power of class representations.

4) Base+GACM+DSA+SAL: Extending the above configuration, we replace the conventional classification loss with the SAL introduced in Section III-E.

5) Base+GACM+DSA+TGCL: In multilabel retrieval tasks, we only utilize TGCL as the metric loss for training our SVCNet.

As shown in Tables VI and VII, we analyzed the LSCIDMRv2 dataset in Table VI and the EuroSAT and Sea Fog datasets in Table VII. Our main findings are the following. 1) In multilabel

TABLE VI
ABLATION EXPERIMENT IN LSCIDMRV2

Base	GACM	DSA	SAL	TGCL	LSCIDMRv2			
					wAP	NDCG(%)	F1(%)	AP(%)
●	○	○	○	○	6.61	82.42	78.18	86.43
●	●	○	○	○	6.97	86.36	80.80	88.29
●	●	●	○	○	6.99	86.57	81.35	88.37
●	●	●	●	○	**7.02**	86.61	81.58	88.80
●	●	●	○	●	6.99	86.76	82.48	88.93
●	●	●	●	●	7.00	**86.86**	**82.73**	**89.02**

The highest results are highlighted in bold.

TABLE VII
ABLATION EXPERIMENT IN SEA FOG AND EUROSAT (%)

Base	GACM	DSA	SAL	Sea Fog			EuroSAT		
				mAP	NDCG	F1	mAP	NDCG	F1
●	○	○	○	89.10	94.41	91.51	94.78	97.03	94.99
●	●	○	○	87.54	94.23	90.93	97.05	98.11	96.85
●	●	●	○	89.32	94.92	92.91	97.29	98.28	97.07
●	●	●	●	**91.32**	**96.00**	**93.10**	**97.35**	**98.30**	**97.12**

The highest results are highlighted in bold.

image retrieval tasks, the employment of the GACM for multiview information communication notably enhanced the model's performance. Specifically, compared to the straightforward concatenation of multiview features, the wAP on the LSCIDMRv2 dataset for multilabel retrieval tasks increased from 6.61 to 6.97, and the classification F1-score rose from 78.18% to 80.80%. This validates the efficacy of our GACM approach in multilabel retrieval and identification tasks. 2) It is worth noting that the stand-alone application of GACM on the Sea Fog dataset led to a slight decline in performance. Interestingly, under the guidance of DSA, the model's performance improved relative to the "Base" model. This indicates that the use of initialized semantic anchors serves a positive role; it not only guides semantic learning more effectively but also adapts well to scenarios where the sample size is relatively limited. 3) The proposed loss functions, TGCL and SAL, have a positive impact on both single-label and multilabel retrieval tasks. Their combination outperforms the standard approach of metric loss combined with cross-entropy classification loss, enhancing retrieval performance. Notably, employing TGCL alone shows superior performance over this standard combination, highlighting TGCL's efficacy in multilabel retrieval tasks. 4) In summary, with the incremental implementation of the proposed components, performance across the three datasets was gradually enhanced in most scenarios, thereby substantiating the effectiveness of each individual component.

Moreover, to further assess the contribution of the metric loss, we tested the performance under different conditions by varying the weighting function coefficient λ in (14). The results reported in Table VIII indicate a performance decline for two datasets when $\lambda = 0$. Conversely, when $\lambda > 0.2$, there is a marked improvement in performance. This suggests that appropriately increasing the weight of the TGCL contributes positively to the enhancement of retrieval results.

TABLE VIII
RESULTS OF HYPERPARAMETER TUNING FOR THE BALANCE FACTOR λ IN (14), WITH THE BOLDED λ INDICATING THE SETTING ADOPTED IN THIS STUDY; THE BOLDED EXPERIMENTAL RESULTS REPRESENT THE OPTIMAL OUTCOMES

λ	LSCIDMRv2		Sea Fog	
	wAP	F1(%)	mAP(%)	F1(%)
0	6.80	79.09	87.79	90.55
0.1	6.90	80.00	85.65	91.50
0.2	6.96	81.13	87.95	92.13
0.3	6.98	80.80	88.65	92.19
0.4	7.00	82.05	90.89	92.18
0.5	7.00	**82.73**	**91.32**	93.10
0.6	7.00	82.50	90.19	93.05
0.7	**7.01**	81.61	81.05	**94.05**
0.8	6.98	81.58	89.86	93.53
0.9	6.99	81.56	85.79	92.79
1	**7.01**	82.05	89.31	92.90

The highest results are highlighted in bold.

TABLE IX
PERFORMANCE COMPARISON IN SELECTED CATEGORIES ON THE LSCIDMRV2 DATASET, WHERE BOTH RECALL AND AP REPRESENT THE RECALL AND AVERAGE PRECISION FOR EACH CATEGORY, RESPECTIVELY

Method	FS		WJ	
	Recall(%)	AP(%)	Recall(%)	AP(%)
MS-GOGO [4]	23.21	31.74	68.89	58.15
DMMVH [35]	21.43	39.42	62.22	59.96
ViT [16]	12.50	25.11	62.22	63.52
Ours	**76.69**	**44.15**	**77.07**	**64.86**

The highest results are highlighted in bold.

TABLE X
PERFORMANCE COMPARISON IN SELECTED CATEGORIES ON THE SEA FOG DATASET, WHERE BOTH RECALL AND AP REPRESENT THE RECALL AND AVERAGE PRECISION FOR EACH CATEGORY, RESPECTIVELY

Method	Sea Fog		Low Cloud	
	Recall(%)	F1(%)	Recall(%)	F1(%)
ViT [16]	87.93	89.86	**91.66**	86.12
DMMVH [35]	**93.10**	90.75	83.33	83.33
SCFR [12]	87.06	88.59	84.84	86.15
Ours	90.52	**92.10**	87.12	**87.12**

The highest results are highlighted in bold.

G. Quantitative and Qualitative Analysis of Challenging Samples

In the LSCIDMRv2 dataset, the categories FS and WJ, which have a limited number of samples, present particular challenges for accurate retrieval. In the Sea Fog dataset, the differentiation between low clouds and sea fog presents a challenge due to their similar physical properties. To demonstrate the effectiveness of our proposed method in these challenging identification tasks, we selected two models with superior overall performance for comparative experiments: a well-performing baseline model, ViT, and a MVL model, DMMVH. Qualitative results are presented in Figs. 9 and 10, while quantitative outcomes are detailed in Tables IX and X, the reported Recall, AP, and F1 in the tables are all category-specific values.

Figs. 9 and 10 demonstrate that when faced with multilabel query samples, our proposed SVCNet model exhibits superior retrieval performance. In contrast, alternative methods frequently misclassify FS and WJ as EC. In the task of

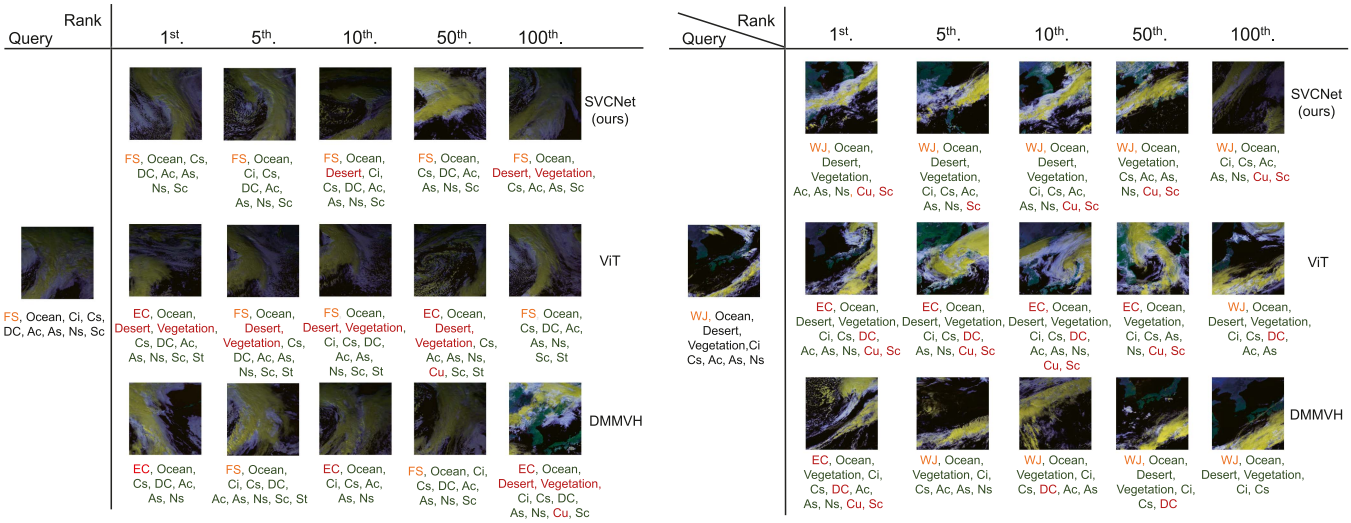

Fig. 9. Case study results for retrieval in classes with low sample sizes, comparing SVCNet with select methods.

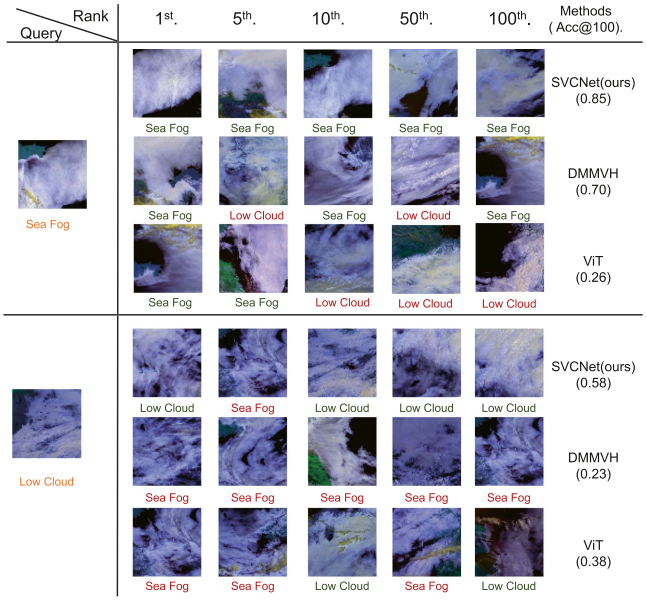

Fig. 10. Case study results for retrieval in hard-to-distinguish classes (low clouds and sea fog), comparing SVCNet with select methods.

distinguishing between sea fog and low clouds, SVCNet also manifests optimal performance, while other techniques are prone to conflating the two categories. Tables IX and X corroborate these findings by showing that, among all samples with specific labels, our proposed method achieves the highest recall rates, outperforming other techniques by over 10%. Notably, in the identification tasks involving low clouds and sea fog, our method exhibits the highest overall performance.

H. Model Efficiency Evaluation

We conducted experiments to evaluate the efficiency of the proposed SVCNet compared to all compared deep learning models, including model parameters, training time, and retrieval time. All models were trained and tested on an NVIDIA 3080Ti GPU with 16 GB of memory. The results are shown in Table XI. The training time refers to the time taken for one iteration with a batch size of 100, excluding data preprocessing time. Retrieval time refers to the time taken during testing to process a batch of 100 queries and return the database index results, measured in milliseconds.

TABLE XI
COMPARISON OF TRAINING AND RETRIEVAL TIMES IN DIFFERENT METHODS

Model	Parameters (M)	Training time (s/iter)	Retrieval time (ms/100 queries)
ResNet50 [11] + MS loss [27]	25.5	0.88	15
ViT [16] + MS loss [27]	27.1	2.18	130
DenseNet121 [26] + MS loss [27]	8.0	1.35	76
CSQ [43]	25.4	0.23	13
SCFR [12]	58.18	1.35	170
DMMVH [35]	51.0	2.64	110
GRN [13]	67.1	2.56	580
AWNet [49]	26.4	0.29	32
DAH [48]	25.9	0.55	85
DGSSH [47]	50.6	2.99	275
SVCNet (1 view)	12.9	1.26	65
SVCNet (2 views)	24.1	1.49	83
SVCNet (3 views)	35.3	1.64	102
SVCNet (4 views)	46.6	1.69	135

For the baseline methods trained with multisimilarity (MS) loss function [27], the parameter count and training/retrieval time of SVCNet increase with the number of views, as different view encoders are not shared. We evaluated the parameter count and retrieval time with different numbers of views. The results show that the proposed method has certain advantages compared to some advanced methods, but consumes more time compared to lightweight general retrieval methods, such as the baseline model, CSQ, and DMMVH.

V. DISCUSSION

We introduce a MVL framework called SVCNet, inspired by human communication mechanisms, designed to explore the intricate associations between various spectra and semantics for

more discriminative MSI representations. In this framework, multiview feature extraction serves as the understanding phase, while the GACM corresponds to the communication phase. Additionally, we incorporate two improved loss functions aimed at facilitating more precise semantic learning.

The overall effectiveness of the proposed method is substantiated by the results presented in Tables III–V. When compared to multiple SOTA approaches across three datasets, SVCNet achieved the best overall performance, demonstrating its capability to generate more discriminative representations than alternative methods. The efficacy of the individual components is further validated in Tables VI and VII. As various components are replaced by those proposed, a consistent improvement in model performance is observed. Notably, even without employing SAL and TGCL, our method achieved a wAP score of 6.97 when the GACM module was included. This surpasses many advanced remote sensing multilabel methods, such as GRN (wAP of 6.15) and FDRL (wAP of 6.54), highlighting the effectiveness of multiview graph learning in mining remote sensing images and further attesting to the utility of GACM. Moreover, the two loss function components we propose further promote a better consensus and reflection, especially the combination of SAL and GACM, which achieved F1 scores of 92.91% and 97.07% on the Sea Fog and EuroSAT datasets, respectively, outperforming all other methods in Table IV. Figs. 7–10, along with Tables IX and X, further demonstrate the model's robustness and additional gains, notably the robustness of MVL and the enhanced recognition of rare categories.

Despite achieving the promising results, this study has several limitations. While our model demonstrates strong performance in multispectral settings and performs well under the single-view conditions depicted in Figs. 7 and 8, we have not yet empirically validated it on datasets featuring only single-view RGB remote sensing images. Additionally, the LSCIDMRv2 dataset exhibits a long-tailed distribution of labels. Certain rare weather system categories, such as FS and WJ, appear with a frequency of less than 1%. This overrepresentation of head classes could potentially interfere with the training of the retrieval network. To mitigate this, we have selected a subset of the LSCIDMRv2 dataset for our experiments. Future research will introduce targeted algorithms, especially for those extremely scarce multilabel situations, and will employ the complete long-tailed dataset for validation.

VI. Conclusion

This article introduces SVCNet for RSIR tasks, an MVL framework inspired by human communication mechanisms. SVCNet represents image content and measures view–semantic relationships, characterizing the degree of association between views and categories for precise selective knowledge fusion. Linking MVL with GNNs, SVCNet simulates the three stages of human communication—understanding, communication, and collective consensus and reflection—thereby enhancing RSIR with flexible representation extraction. The versatility and efficacy of SVCNet are demonstrated through experiments on three multispectral remote sensing image datasets in both single-label and multilabel scenarios. Compared to existing SOTA methods, SVCNet exhibits superior performance, particularly in handling class imbalances and difficult-to-distinguish samples, highlighting its broad application potential in diverse RSIR contexts.

References

[1] Y. Cong et al., "SatMAE: Pre-training transformers for temporal and multi-spectral satellite imagery," in *Proc. Adv. Neural Inf. Process. Syst.*, 2022, pp. 197–211.

[2] P. Helber, B. Bischke, A. Dengel, and D. Borth, "EuroSAT: A novel dataset and deep learning benchmark for land use and land cover classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 7, pp. 2217–2226, Jul. 2019.

[3] T. Hu, Z. Jin, W. Yao, J. Lv, and W. Jin, "Cloud image retrieval for sea fog recognition (CIR-SFR) using double branch residual neural network," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 3174–3186, 2023.

[4] D. Zhao, Q. Wang, J. Zhang, and C. Bai, "Mine diversified contents of multi-spectral cloud images along with geographical information for multi-label classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 4102415.

[5] G. Sumbul, J. Kang, and B. Demir, *Deep Learning for Image Search and Retrieval in Large Remote Sensing Archives*. Hoboken, NJ, USA: Wiley, 2021, ch. 11, pp. 150–160, doi: 10.1002/9781119646181.ch11.

[6] Y. Li, J. Ma, and Y. Zhang, "Image retrieval from remote sensing Big Data: A survey," *Inf. Fusion*, vol. 67, pp. 94–115, 2021.

[7] X.-Y. Tong, G.-S. Xia, F. Hu, Y. Zhong, M. Datcu, and L. Zhang, "Exploiting deep features for remote sensing image retrieval: A systematic investigation," *IEEE Trans. Big Data*, vol. 6, no. 3, pp. 507–521, Sep. 2020.

[8] G. J. Scott, M. N. Klaric, C. H. Davis, and C.-R. Shyu, "Entropy-balanced bitmap tree for shape-based object retrieval from large-scale satellite imagery databases," *IEEE Trans. Geosci. Remote Sens.*, vol. 49, no. 5, pp. 1603–1616, May 2011.

[9] Y. Yang and S. Newsam, "Geographic image retrieval using local invariant features," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 2, pp. 818–832, Feb. 2013.

[10] I. Tekeste and B. Demir, "Advanced local binary patterns for remote sensing image retrieval," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2018, pp. 6855–6858.

[11] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[12] M. Huang, L. Dong, W. Dong, and G. Shi, "Supervised contrastive learning based on fusion of global and local features for remote sensing image retrieval," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5208513.

[13] J. Kang, R. Fernandez-Beltran, D. Hong, J. Chanussot, and A. Plaza, "Graph relation network: Modeling relations between scenes for multilabel remote-sensing image classification and retrieval," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 5, pp. 4355–4369, May 2021.

[14] J. Zhao, X. Xie, X. Xu, and S. Sun, "Multi-view learning overview: Recent progress and new challenges," *Inf. Fusion*, vol. 38, pp. 43–54, 2017.

[15] Y. Xu and M. Wei, "Multi-view clustering toward aerial images by combining spectral analysis and local refinement," *Future Gener. Comput. Syst.*, vol. 117, pp. 138–144, 2021.

[16] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Represent.*, 2021, pp. 1–21.

[17] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, and S. Y. Philip, "A comprehensive survey on graph neural networks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 1, pp. 4–24, Jan. 2021.

[18] L. Chen et al., "Multi-scale adaptive graph neural network for multivariate time series forecasting," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 10, pp. 10748–10761, Oct. 2023.

[19] J. Regan and M. Khodayar, "A triplet graph convolutional network with attention and similarity-driven dictionary learning for remote sensing image retrieval," *Expert Syst. Appl.*, vol. 232, 2023, Art. no. 120579.

[20] J. Cheng, Q. Wang, Z. Tao, D. Xie, and Q. Gao, "Multi-view attribute graph convolution networks for clustering," in *Proc. 29th Int. Conf. Int. Joint Conf. Artif. Intell.*, 2021, pp. 2973–2979.

[21] J. R. Larson, P. G. Foster-Fishman, and C. B. Keys, "Discussion of shared and unshared information in decision-making groups," *J. Pers. Social Psychol.*, vol. 67, no. 3, pp. 446–461, 1994.

[22] K. L. Kraemer and J. L. King, "Computer-based systems for cooperative work and group decision making," *ACM Comput. Surv.*, vol. 20, no. 2, pp. 115–146, 1988.
[23] X. Jia, X.-Y. Jing, Q. Sun, S. Chen, B. Du, and D. Zhang, "Human collective intelligence inspired multi-view representation learning–enabling view communication by simulating human communication mechanism," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 6, pp. 7412–7429, Jun. 2023.
[24] G. Marcu, A. K. Dey, and S. Kiesler, "Designing for collaborative reflection," in *Proc. 8th Int. Conf. Pervasive Comput. Technol. Healthcare*, 2014, pp. 9–16.
[25] M. J. Eppler and O. Sukowski, "Managing team knowledge: Core processes, tools and enabling factors," *Eur. Manage. J.*, vol. 18, no. 3, pp. 334–341, 2000.
[26] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 4700–4708.
[27] X. Wang, X. Han, W. Huang, D. Dong, and M. R. Scott, "Multi-similarity loss with general pair weighting for deep metric learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 5022–5030.
[28] A. D. Doulamis and N. D. Doulamis, "Optimal content-based video decomposition for interactive video navigation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 14, no. 6, pp. 757–775, Jun. 2004.
[29] Y. Rui, T. S. Huang, M. Ortega, and S. Mehrotra, "Relevance feedback: A power tool for interactive content-based image retrieval," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 8, no. 5, pp. 644–655, Sep. 1998.
[30] Y. Rui and T. S. Huang, "A novel relevance feedback technique in image retrieval," in *Proc. 7th ACM Int. Conf. Multimedia*, 1999, pp. 67–70.
[31] N. Doulamis and A. Doulamis, "Evaluation of relevance feedback schemes in content-based in retrieval systems," *Signal Process. Image Commun.*, vol. 21, no. 4, pp. 334–357, 2006.
[32] A. D. Doulamis and N. D. Doulamis, "Generalized nonlinear relevance feedback for interactive content-based retrieval and organization," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 14, no. 5, pp. 656–671, May 2004.
[33] N. D. Doulamis, A. D. Doulamis, and S. D. Kollias, "Nonlinear relevance feedback: Improving the performance of content-based retrieval systems," in *Proc. IEEE Int. Conf. Multimedia Expo. Proc. Latest Adv. Fast Changing World Multimedia*, 2000, pp. 331–334.
[34] Y. Tian, D. Krishnan, and P. Isola, "Contrastive multiview coding," in *Proc. 16th Eur. Conf. Comput. Vis.*, Glasgow, U.K., 2020, pp. 776–794.
[35] J. Zhu, X. Ruan, Y. Cheng, Z. Huang, Y. Cui, and L. Zeng, "Deep metric multi-view hashing for multimedia retrieval," in *Proc. IEEE Int. Conf. Multimedia Expo*, 2023, pp. 1955–1960.
[36] H. Shuai, L. Wu, and Q. Liu, "Adaptive multi-view and temporal fusing transformer for 3D human pose estimation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 4, pp. 4122–4135, Apr. 2023.
[37] F. Wu, A. Souza, T. Zhang, C. Fifty, T. Yu, and K. Weinberger, "Simplifying graph convolutional networks," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 6861–6871.
[38] K. Xu, W. Hu, J. Leskovec, and S. Jegelka, "How powerful are graph neural networks?," in *Proc. Int. Conf. Learn. Representations*, 2019, pp. 1–17. [Online]. Available: https://openreview.net/forum?id=ryGs6iA5Km
[39] Y. Li, D. Tarlow, M. Brockschmidt, and R. Zemel, "Gated graph sequence neural networks," 2015, *arXiv:1511.05493*.
[40] W. Hamilton, Z. Ying, and J. Leskovec, "Inductive representation learning on large graphs," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 1–11.
[41] S. Brody, U. Alon, and E. Yahav, "How attentive are graph attention networks?," in *Proc. Int. Conf. Learn. Representations*, 2022, pp. 1–26. [Online]. Available: https://openreview.net/forum?id=F72ximsx7C1
[42] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio, "Graph attention networks," in *Proc. Int. Conf. Learn. Represent.*, 2018, pp. 1–12.
[43] L. Yuan et al., "Central similarity quantization for efficient image and video retrieval," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 3083–3092.
[44] J. Lanchantin, T. Wang, V. Ordonez, and Y. Qi, "General multi-label image classification with transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 16478–16488.
[45] F. Zhao, Y. Huang, L. Wang, and T. Tan, "Deep semantic ranking based hashing for multi-label image retrieval," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1556–1564.
[46] Y. Dai, W. Song, Y. Li, and L. Di Stefano, "Feature disentangling and reciprocal learning with label-guided similarity for multi-label image retrieval," *Neurocomputing*, vol. 511, pp. 353–365, 2022.
[47] H. Zhou, Q. Qin, J. Hou, J. Dai, L. Huang, and W. Zhang, "Deep global semantic structure-preserving hashing via corrective triplet loss for remote sensing image retrieval," *Expert Syst. Appl.*, vol. 238, 2024, Art. no. 122105.
[48] Y. Zhang, X. Zheng, and X. Lu, "Remote sensing image retrieval by deep attention hashing with distance-adaptive ranking," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 4301–4311, 2023.
[49] X. Tian, D. Hou, S. Wang, X. Liu, and H. Xing, "An adaptive weighted method for remote sensing image retrieval with noisy labels," *Appl. Sci.*, vol. 14, no. 5, pp. 1–16, 2024.

Nan Wu (Student Member, IEEE) received the bachelor's degree in information management and information systems from Ningbo University, Ningbo, China, in 2020, where he is currently working toward the master's degree in computer science and technology with the Faculty of Electrical Engineering and Computer Science.

His research interests include remote sensing image and medical imaging representation learning, sparse representation, and deep neural networks.

Wei Jin received the Ph.D. degree in optical engineering from Chongqing University, Chongqing, China, in 2006.

He is currently a Professor with the Faculty of Electrical Engineering and Computer Science, Ningbo University, Ningbo, China. His research interests include remote sensing image processing, image retrieval, deep learning, and computer vision.

Randi Fu received the M.Eng. degree in photogrammetry and remote sensing from Information Engineering University, Zhengzhou, China, in 2001.

He is currently an Associate Professor with the Faculty of Electrical Engineering and Computer Science, Ningbo University, Ningbo, China. His research interests include artificial intelligence, pattern recognition, image retrieval, and remote sensing image interpretation.