Fine-Grained Information Supplementation and Value-Guided Learning for Remote Sensing Image-Text Retrieval

Zihui Zhou, Yong Feng, Agen Qiu, Guofan Duan, and Mingliang Zhou

***Abstract*—Remote sensing (RS) image-text retrieval is a practical and challenging task that has received considerable attention. Currently, most approaches rely on either convolutional neural networks or Transformers, which cannot effectively extract both global and fine-grained features simultaneously. Furthermore, the problem of high intramodal similarity in the RS domain poses a challenge for feature learning. In addition, the characteristics of model training at different stages seem to be neglected in most studies. In order to tackle these problems, we propose a fine-grained information supplementation (FGIS) and value-guided learning model that leverages prior knowledge in the RS domain for feature supplementation and employs a value-guided training approach to learn fine-grained, expressive, and robust feature representations. Specifically, we introduce the FGIS module to facilitate the supplementation of fine-grained visual features, thereby enhancing perceptual abilities for both global and local features. Furthermore, we mitigate the problem of high intra-modal similarity by proposing two loss functions: the weighted contrastive loss and the scene-adaptive fine-grained perceptual loss. Finally, we design a value-guided learning framework that focuses on the most important information at each stage of training. Extensive experiments on the remote sensing image captioning dataset (RSICD) and remote sensing image text match dataset (RSITMD) datasets verify the effectiveness and superiority of our model.**

***Index Terms*—Fine-grained feature, intramodal similarity, remote sensing image-text retrieval (RSITR), value-guided.**

Received 8 August 2024; revised 13 September 2024; accepted 9 October 2024. Date of publication 14 October 2024; date of current version 29 October 2024. This work was supported in part by the National Nature Science Foundation of China under Grant 62262006, in part by the Guangxi Science and Technology Key Research and Development Program under Grant AB24010112, in part by the Key Project of the Science and Technology Research Program of Chongqing Education Commission of China under Grant KJZD-K202402501, in part by the State Key Laboratory of Geo-Information Engineering and the Key Laboratory of Surveying and Mapping Science and Geospatial Information Technology of MNR, CASM under Grant 2023-04-03, and in part by the Technology Innovation and Application Development Key Project of Chongqing under Grant cstc2021jscx-gksbX0058. *(Corresponding author: Yong Feng.)*

Zihui Zhou is with the College of Computer Science, Chongqing University, Chongqing 401331, China, and also with the State Key Laboratory of Geo-Information Engineering and Key Laboratory of Surveying and Mapping Science and Geospatial Information Technology of MNR, CASM, Beijing 100036, China (e-mail: 20191694@cqu.edu.cn).

Yong Feng and Mingliang Zhou are with the College of Computer Science, Chongqing University, Chongqing 401331, China (e-mail: fengyong@cqu.edu.cn; zml-0913yy@163.com).

Agen Qiu is with the State Key Laboratory of Geo-Information Engineering and Key Laboratory of Surveying and Mapping Science and Geospatial Information Technology of MNR, CASM, Beijing 100036, China (e-mail: qiuag@casm.ac.cn).

Guofan Duan is with the Chongqing Metropolitan College of Science and Technology, Chongqing 402167, China (e-mail: wlm_bkesz@foxmail.com).

Digital Object Identifier 10.1109/JSTARS.2024.3480014

I. Introduction

With the development of remote sensing (RS) satellite technologies, there have been extensive applications in various fields, resulting in a significant amount of RS data being generated every day [1], [2]. Consequently, the effective extraction and management of valuable information from large-scale RS data represents a significant and promising area for further research. Artificial intelligence (AI) enhances the efficacy of RS data acquisition, analysis, and presentation [3], serving as a robust technique for facilitating RS data manipulation. As a technique in the field of AI, remote sensing image-text retrieval (RSITR) bridges knowledge understanding at a semantic level between image and text modalities. It facilitates the process of extracting the desired information from large amounts of RS data and has received considerable attention [4].

RSITR methods can be categorized into two types: convolutional neural network (CNN)-based methods [5], [6], [7] and Transformer-based methods [8], [9], [10]. CNN-based methods are likely to utilize ResNet for image feature extraction and the bidirectional gate recurrent unit for text feature extraction. Since different layers of the CNN could capture semantic information at different levels, such methods tend to leverage multiple layers of the CNN feature extractor to obtain multiscale features. GaLR [11] integrates features from different layers of the CNN feature extractor with the object-level features extracted by the object detector, allowing complementarity between them. Lightweight-multiscale crossmodal text-image retrieval (LW-MCR) [12] employs a feature extractor to extract multiscale features and uses a visual self-attention mechanism to filter salient features after the fusion of multiscale features. By enhancing and fusing the hierarchical representations, these approaches could capture fine-grained details and contextual information from images.

While CNN-based methods are excellent at capturing local features, they may be deficient at global perception. Therefore, some researchers [13], [14] set their sights on applying Transformer, which effectively captures global semantic information. However, Transformer-based methods fail to utilize the locality in spatial dimensions [15]. Furthermore, the hierarchical structure of convolutional kernels learns semantic information at different levels, which is a multiscale perception capability lacking in Transformer [16]. In addition, RS images tend to contain objects of various scales and numerous small targets,

© 2024 The Authors. This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

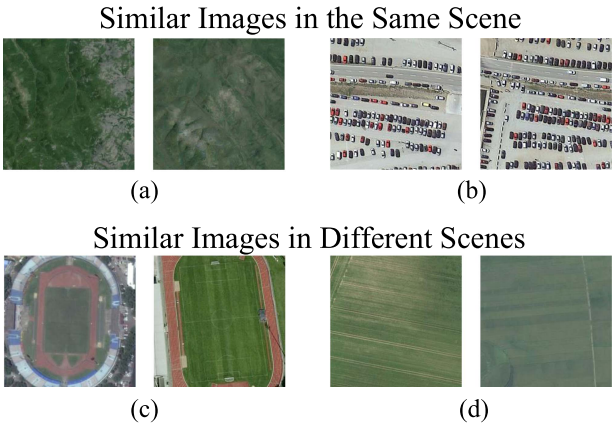

Fig. 1. High similarity of some RS images. Image pairs (a) and (b) depict two images of the same scene, while (c) and (d) represent pairs of images from different scenes.

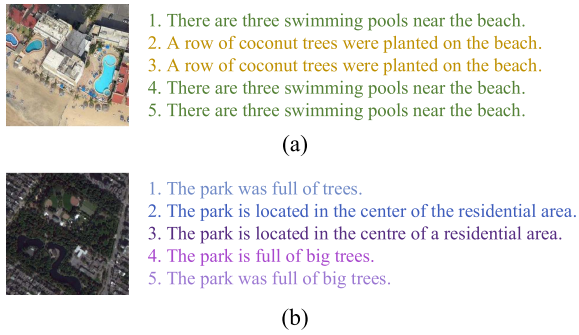

Fig. 2. High repetition and similarity rate of some RS sentences. Image pairs (a) and (b) show repeatability and high similarity between texts corresponding to a n image , respectively . The colors of repeated sentences are the same. Other sentences express very similar meanings, even if they are not the same.

making local and multiscale information perception crucial for RSITR. Therefore, to combine the advantages of both models, Wang et al. [10] designed a feature enhancement module, consisting of Transformer-based and CNN-based submodules, wherein the CNN-based submodule is employed for mining multiscale information. Pan et al. [9] also integrated the advantages of Transformer and CNN. They proposed PIR that includes a visual instruction representation module that filters the visual representation by applying prior knowledge of the RS field. However, PIR does not fully leverage the multiscale features extracted by the CNN feature extractor; instead, it just uses the global feature extracted by the instruction extractor. To address this issue, we propose a fine-grained information supplementation (FGIS) module that uses multiscale semantic information extracted by the CNN-based feature extractor to augment the features extracted by the Transformer-based extractor for a more thorough FGIS, thereby more thoroughly integrating the benefits of both CNN-based and Transformer-based methods.

In addition, in order to gain better performance, the data properties of the RS field need to be considered. In comparison to natural scenes, RS data demonstrates higher intramodal similarity. The images shown in Fig. 1 exhibit similar textures, colors, and arrangements of small objects within the same scene pairs, making them difficult to distinguish. In different scenes, the images also show a high degree of similarity. For example, in Fig. 1(c), both the "stadium" (left) and "playground" (right) scenes feature a large playground in the center. Similarly, in Fig. 1(d), the "meadow" (left) and "farmland" (right) scenes share similar background colors and textures. In addition, the visualization study conducted by Yuan et al. [17] demonstratesd the high similarity in the text of several RS datasets. Furthermore, due to the absence of original data, some sentences are replicated to ensure each image is associated with five sentences [18], further exacerbating the problem of high sentence similarity, as depicted in Fig. 2. The intramodal similarity issue in RS datasets may introduce overfitting risks and place increased demands on the model's ability for fine-grained awareness. Moreover, this problem can also cause the distribution of features of different semantic information to be overly concentrated in the common subspace, resulting in reduced robustness.

To address the above problem, we propose two losses. The weighted contrastive loss evaluates the value of each text sample from the perspective of uniqueness and richness and determines its influence on gradient updating based on this value. By using weighted contrastive loss, richer feature representation can be learned, and the negative effects of text repetition and high similarity are mitigated to a certain extent. In addition, a scene-adaptive fine-grained perceptual loss is introduced, which classifies intramodal sample pairs based on scene information and learns the fine-grained differences of intra-modal features from easy to difficult according to the curriculum learning (CL) theory. Scene-adaptive fine-grained perceptual loss enhances fine-grained awareness and encourages the construction of a more evenly distributed common subspace. By using these losses, the model is able to learn a more expressive and robust representation.

Furthermore, the unique features at different stages of training seem to be overlooked in most RSITR methods, which limits performance improvement. Therefore, we present a value-guided learning framework (VGLF) that takes into account the characteristics of model learning at each stage based on the proposed losses. Specifically, VGLF divides the model training process into different phases and selects an appropriate learning strategy that is consistent with the main goals and potential challenges of the current training phase. VGLF ensures that each phase is focused on the acquisition of the most valuable information, resulting in further performance improvement. In combination with the above methods, we present a fine-grained information supplementation and value-guided learning (FISVL) model for RSITR.

In summary, the main contributions of our study can be summarized as follows.

1) We propose FISVL, which utilizes prior RS knowledge for feature supplementation and adopts a value-guided training approach, allowing for learning more fine-grained, expressive, and robust feature representation.
2) The proposed FGIS module makes full use of fine-grained information for feature enhancement, thereby increasing the fine-grained awareness of the model.
3) The VGLF enables the model to focus on the most valuable information at different training stages, thereby adapting to the specific requirements of each phase.

4) Extensive experiments are conducted on RSITR benchmarks, and the results confirm the effectiveness of FISVL.

The rest of this article is organized as follows. In Section II, we present an overview of the related work. Then, in Section III, we provide a detailed explanation of the proposed method. In Section IV, we present extensive experiments and analyze the results to demonstrate the effectiveness of the proposed model. Finally, Section V concludes this article.

II. Related Work

A. Cross-Modal Image-Text Retrieval in the Natural Field

Cross-modal retrieval is a task that involves the retrieval of samples that are semantically related from a massive database based on a specified representation of another modality [19]. This task can be further divided into two subtasks: image-to-text retrieval and text-to-image retrieval [20]. Cross-modal image-text retrieval has found applications in various fields, including medicine, military, business, education, sports, and other domains [21].

The goal of a cross-modal image-text retrieval model is to learn a feature mapping that maps images and texts into a common subspace, where the semantic consistency between the two modalities can be evaluated based on their distances in that subspace. Existing methods can be categorized into two types: coarse-grained methods and fine-grained methods [22]. Coarse-grained methods aim to match the features of the entire image extracted by CNN or vision transformer with the textual information [23]. Huang et al. [24] proposed an end-to-end vision language pretraining model that learns cross-modal representations directly from data pairs, using dynamically updated visual dictionaries to align visual and linguistic features. Align the image and text representations before fusing (ALBEF) [25] uses a self-training method that learns from pseudolabels generated by a momentum model to enhance the learning capacity from noisy network data. In recent years, several methods have employed pre-training techniques to train models on large-scale datasets [26], [27], [28] to further improve the accuracy. Besides, some studies [29], [30] used generative adversarial networks (GANs) for feature embedding and alignment.

On the other hand, fine-grained methods focus on aligning local features from both images and sentences to infer the global alignment between them, which tend to have higher performance [23]. X-VLM [23] aims to locate visual concepts in images given relevant text and perform multigranularity alignment between text and visual concepts, while Zeng et al. [31] extended X-VLM further by unifying image-text pretraining and video–text pretraining in a single model. Kim et al. [32] presented a set-based embedding approach that includes smooth-chamfer similarity and a set prediction module based on the slot attention mechanism.

B. Cross-Modal Image-Text Retrieval in the RS Field

Similar to the natural field, RSITR also involves multimodal semantic alignment. However, RS images exhibit greater distributional variations while text distribution tends to have high similarity, which leads to more severe modality heterogeneity in the RS field, making semantic alignment more challenging [6]. To address this problem, Cheng et al. [33] proposed a semantic alignment module for cross-modal semantic alignment and designed a gating function to filter unnecessary information. Tang et al. [8] designed an information interaction-enhancing module that considers image and text features as a whole while simulating the internal relationships between RS image and text. Besides, some studies focused on applying loss functions or GAN for semantic alignment [6], [34], [35]. We introduce the weighted contrastive loss to align text and visual features by learning from samples of varying values.

In addition, RSITR faces some unique challenges. First, there is a problem of high intramodal similarity in the RS domain. While it is assumed that RS images have larger distributional variations [6], some RS images are similar, as explained in Section I, and this problem is also severe in the text modality. To solve this problem, Zhang et al. [6] introduced the feature uniformity strategy to optimize the distribution of image features. Mikriukov et al. [34] alleviated the problem of high intramodal consistency through intra-modal contrastive losses. Similarly, the proposed scene-adaptive fine-grained perceptual loss could assist the model in distinguishing fine-grained differences within the features in the same modality, thereby mitigating this problem.

In addition, the difficulties inherent in the collection of data and the operation of airborne platforms frequently result in the presence of greater levels of noise in RS images. Consequently, there is a heightened necessity for the algorithms to demonstrate enhanced robustness [36]. Moreover, RS images are more likely to contain objects at various scales, making object extraction challenging. Current methods tend to address this issue by fusing features from different layers of the CNN feature extractor [5], [11], [17]. Furthermore, the problem of semantic confusion zones in the RS domain has recently been mentioned [5], which refers to regions in the semantic space where there is semantic ambiguity due to the similarity of images from different scenes. Some studies [5], [9] have begun to address this issue, but further research is needed.

III. Proposed Method

A. Overall Framework

For RSITR, a dataset consists of a set of image-text pairs, denoted by the pair $(V, T) = \{(v_i, t_i)\}_{i=1}^{n}$, where $V = \{v_1, v_2, \ldots, v_n\}$ is the set of images and $T = \{t_1, t_2, \ldots, t_n\}$ is the set of texts, and n is the number of the pairs. The size of a single sample in V is $R^{H\times W\times C}$, where H, W, and C denote the height, width, and channel of the image, respectively. Sentence t_i in T can be expressed as $t_i = \{w_1, w_2, \ldots, w_j, \ldots, w_m\}$, where w_j stands for the jth word of t_i and the length of t_i is m. $v_i (i \in [1, n])$ matches $t_k (k \in [1, n])$ if $i = k$.

The overall framework of FISVL is depicted in Fig. 3. Given an image-text pair (v_i, t_i), FISVL passes the image and text through the vision feature extraction model and the language feature extraction model, respectively, to obtain d_{emb}-dimensional vision feature v_{emb}^i and language feature t_{emb}^i. The image is

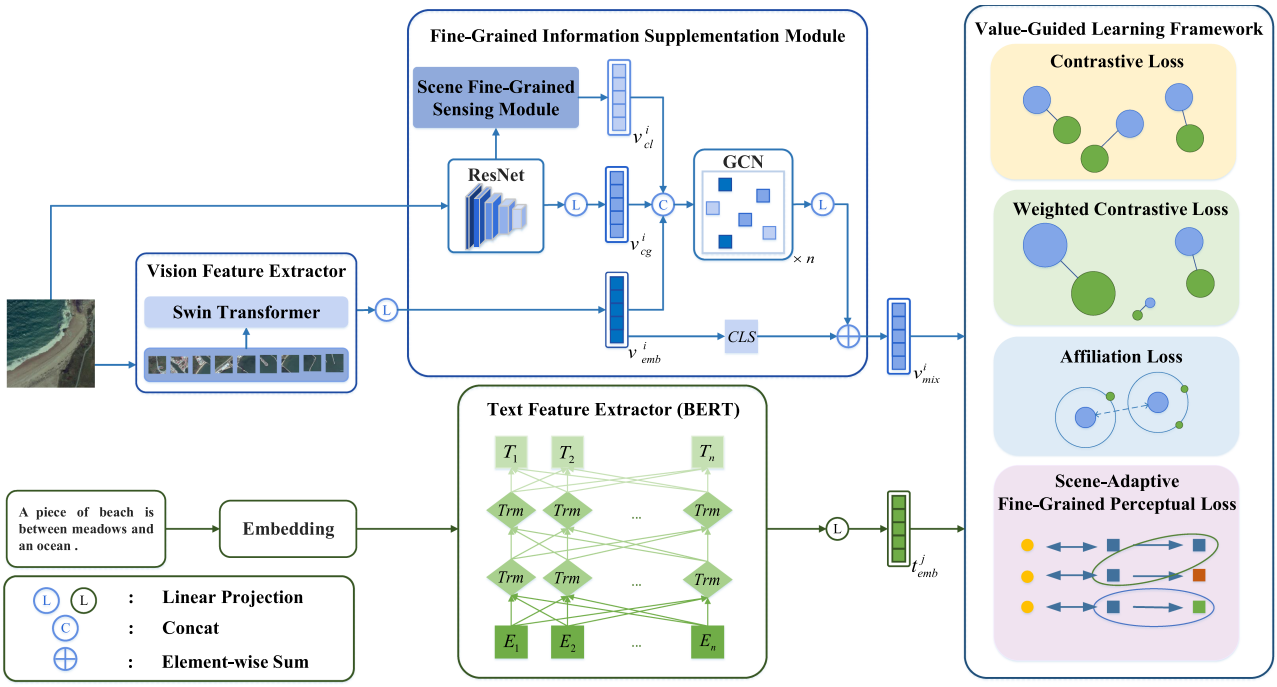

Fig. 3. Framework of the proposed method. The Swin Transformer is used for image feature extraction, and the FGIS module is employed for fine-grained perceptual enhancement. The text feature extraction module includes an embedding module, BERT, and a linear projector. The VGLF is employed for model training.

also fed into an instruction extractor in the FGIS module to get fine-grained awareness features. After that, all the vision features are employed for FGIS, and the final embedding of v_i is obtained, which can be expressed as $v^i_{\text{mix}} \in R^{1\times d_{\text{emb}}}$. Finally, the cosine similarity is calculated as the similarity score of the output vectors of both modalities, and the process can be described as

$$S(v_i, t_i) = {v^i_{\text{mix}}}^T t^i_{\text{emb}} \tag{1}$$

where $S(v_i, t_i)$ denotes the similarity score of the image-text pair (v_i, t_i). The model is trained using the VGLF, which includes contrastive loss, weighted contrastive loss, affiliation loss, and scene-adaptive fine-grained perceptual loss. Subsequently, the model is introduced from the following: the specifics of visual and textual feature extraction, the FGIS module, and the learning strategies.

B. Feature Extraction

1) Vision Feature Extraction: CNN-based feature extractors and Transformer-based feature extractors are commonly employed to extract vision features in RSITR. Given that Transformer is inclined to prioritize key information in images and effectively capture global context, we leverage Swin Transformer [37] to extract vision features. This process can be expressed as

$$v^i_{tv} = \{v^i_{cls}, v^1_p, v^2_p, \ldots, v^l_p\} = F_{tv}(v_i, \theta_{F_{tv}}) \tag{2}$$

where v^i_{tv} is the encoded feature of v_i, consisting of d_v-dimension $v^i_{cls} \in R^{1\times d_v}$ ([CLS]) and l patch features $v^j_p \in R^{1\times d_v}, j \in [1, l]$. $F_{tv}(\cdot,\cdot)$ is Swin Transformer and $\theta_{F_{tv}}$ denotes the fine-tuning weights of it. Subsequently, v^i_{emb}, which has a dimension of d_{emb}, is then obtained by passing v^i_{tv} through a linear projector which can be denoted as E_{tv}.

Although Transformer-based vision extractors demonstrate superior performance in numerous computer vision tasks, CNN-based extractors are better at encoding local features. Inspired by Pan et al. [9], we employ an instruction extractor F_{cv} to extract vision feature v^i_{cg}. In addition, we also utilize CNN-based modules to extract vision feature v^i_{cl}, which encompasses more fine-grained semantic information.

2) Text Feature Extraction: Since Transformer-based text feature extractors are more effective than those based on RNN [38] at modeling long-distance dependencies, we use Transformer-based BERT [39] to extract text features, and it can be expressed as

$$t^i_{tt} = \{t^i_{\text{cls}}, t^1_w, t^2_w, \ldots, t^m_w\} = F_{tt}(t_i, \theta_{F_{tt}}) \tag{3}$$

where t^i_{tt} is the extracted feature of t_i, consisting of d_t-dimension $t_{cls} \in R^{1\times d_t}$ ([CLS]) and m word features $t^j_w \in R^{1\times d_t}, j \in [1, m]$. $F_{tt}(\cdot,\cdot)$ denotes BERT and $\theta_{F_{tt}}$ denotes the fine-tuning weights of it. Similarly, we fed t^i_{tt} into a linear projector E_{tt} to obtain d_{emb}-dimension embedding t^i_{emb}.

It is noteworthy that, in contrast to other methods, we directly compute the similarity between t^i_{emb} and v^i_{mix} without the use of additional textual feature enhancement modules to enhance t^i_{emb}. Although there is a perception that there are distributional differences between text in the RS field and natural field, fine-tuning BERT on the RSITR datasets and utilizing a learnable linear projector for mapping could preliminarily enhance the adaptability of the text feature extraction module to the RS field. Furthermore, to address the redundancy issue in the text modality

of the RS field, the weighted contrastive loss is employed without increasing the number of parameters. The experimental results presented in Section IV demonstrate that FISVL is capable of achieving state-of-the-art performance on RSITR datasets even in the absence of complex text processing branches.

C. FGIS Module

While v^i_{emb} provides a comprehensive understanding of global relationships, it may lack some fine-grained information required to distinguish subtle differences between similar features and identify small objects. To mitigate this issue, we propose the FGIS module to enhance fine-grained awareness.

First, we obtain the output of the last layer of the instruction extractor as the global feature of v_i, which is a d_c-dimension vector. Then, we apply a learnable linear projection E_{cv} to the vector to obtain $v^i_{cg} \in R^{1\times d_{\text{emb}}}$. This can be described as follows:

$$v^i_{cg} = E_{cv}(F_{cv}(v_i, \theta_{F_{cv}})) \tag{4}$$

where $F_{cv}(\cdot,\cdot)$ denotes the instruction extractor and $\theta_{F_{cv}}$ denotes the weights of it. We froze $\theta_{F_{cv}}$ during model training. As the final layer output of the CNN-based encoder, v^i_{cg} contains the global information of v_i and, in comparison to v^i_{cls}, preserves a greater degree of local information.

The CNN-based feature extractor is superior to the Transformer-based feature extractor in its ability to capture fine-grained features. Therefore, relying solely on the global features extracted by the instruction extractor for FGIS may result in inadequate utilization of shallow features extracted from lower layers of the network, which are more likely to contain rich local information. Consequently, it is assumed that the incorporation of these shallow features into the existing framework could enhance the vision features and leverage the advantages of the instruction extractor to a greater extent. For this purpose, we utilize the scene fine-grained sensing module (SFGS) [5] to establish associations between semantic features of varying granularity and obtain a fine-grained local feature by fusing them. This can be represented as

$$v^i_{cl} = F_{\text{SFGS}}(g^i_s) \tag{5}$$

where $v^i_{cl} \in R^{1\times d_{\text{emb}}}$ is the obtained local feature, and $F_{\text{SFGS}}(\cdot,\cdot)$ denotes SFGS, g^i_s means the feature of v_i extracted by the first layer of the instruction extractor.

Subsequently, a graph convolution neural network (GCN) is applied to facilitate the transfer of information between v^i_{cg}, v^i_{cl} and v^i_{emb}. This process enables the refinement of features and enhances the extent of fine-grained awareness.

Given a set of features, GCN treats them as nodes and calculates the edges between the nodes to capture the relationship between them. This is followed by the use of the representations of the edges and adjacent nodes to update the node representations, with the objective of propagating information between nodes. This process is repeated n_g times, with each repetition following Diao et al. [40] to calculate the features. Specifically, we begin with concatenating v^i_{cg}, v^i_{cl}, and v^i_{emb} as input, which can be expressed as $v_{gin} = \text{concat}(v^i_{cg}, v^i_{\text{cl}}, v^i_{\text{emb}}) \in R^{(l+3)\times 1\times d_{\text{emb}}}$ and $\text{concat}(\cdot)$ means the concatenation operation. Each node has directed edges to other $l+2$ nodes. Then, the edge from node v_x to node v_y is updated each time as follows:

$$e^r(v^r_y, v^r_x; E^r_{\text{in}}, E^r_{\text{out}}) = \frac{\exp((E^r_{\text{in}} v^r_y)(E^r_{\text{out}} v^r_x))}{\sum_x \exp((E^r_{\text{in}} v^r_y)(E^r_{\text{out}} v^r_x))} \tag{6}$$

where v^r_x and v^r_y are the features of v_x and v_y at the rth step, E^r_{in} and E^r_{out} are the learnable linear projectors for incoming and outgoing nodes at the rth step, respectively, and $\exp(\cdot)$ stands for the exponential operation, $e^r(v^r_y, v^r_x; E^r_{\text{in}}, E^r_{\text{out}})$ denotes the edge from v_x to v_y at the rth step. After that, we update v_y as

$$v^{r+1}_y = \text{ReLU}(E^r_n \sum_x e^r(v^r_y, v^r_x; E^r_{\text{in}}, E^r_{\text{out}}) v^r_x) \tag{7}$$

where E^r_n is a learnable linear projector at the rth step, and $\text{ReLU}(\cdot)$ represents the ReLU activation function. v^{r+1}_y means the feature of v_y at the $r+1$th step, which replaces v^r_y at the step. The above process can be represented as follows:

$$\begin{aligned} {v^i_{cg}}^{r+1}, {v^i_{\text{cl}}}^{r+1}, {v^i_{\text{emb}}}^{r+1} &= GCN\left({v^i_{cg}}^{r}, {v^i_{\text{cl}}}^{r}, {v^i_{\text{emb}}}^{r}\right) \\ r &= 0, 1, \ldots, n_g - 1 \end{aligned} \tag{8}$$

where $\text{GCN}(\cdot)$ means an inference step of GCN. When $r=0$, ${v^i_{cg}}^{r}$, ${v^i_{cl}}^{r}$, and ${v^i_{\text{emb}}}^{r}$ mean the input v^i_{cg}, v^i_{cl}, and v^i_{emb}, respectively. We iteratively reason the features for n_gth steps, and the features obtained from the n_gth step are treated as the output of GCN, which can be expressed as $v_{g\text{out}} = \{{v^i_{cg}}^{n_g}, {v^i_{\text{cl}}}^{n_g}, {v^i_{\text{emb}}}^{n_g}\}$.

Finally, the feature of v^i_{cls} after passing through GCN, which is included in ${v^i_{\text{emb}}}^{n_g}$ and can be represented as ${v^i_{\text{cls}}}^{n_g}$, is fed into a learnable linear projector and added to v^i_{cls} to get the detail-aware embedding as the final vision feature of v_i. The formulation of these steps is

$$v^i_{\text{mix}} = E_v\left({v^i_{\text{cls}}}^{n_g}\right) + v^i_{cls} \tag{9}$$

where $E_v(\cdot)$ denotes the linear projection, and $v^i_{\text{mix}} \in R^{1\times d_{\text{emb}}}$ is the final vision feature.

D. Learning Strategies

1) Weighted Contrastive Loss: The contrastive loss [41] and the triplet loss [42] are commonly used loss functions in RSITR. The experiment of Pan et al. [9] demonstrated that the contrastive loss can optimize the model better than the triplet loss. Therefore, we follow the method [9] in applying the contrastive loss to optimize the model. Given a batch of samples, N_p positive pairs $\{(v_1,t_1),(v_2,t_2),\ldots,(v_{N_p},t_{N_p})\}$ are randomly sampled, and the similarity scores $S(v_i,t_j)$ are calculated as (1). After that, the vision-to-text contrastive loss $L^c_{v2t}(v_i,t_i)$ is calculated as

$$L^c_{v2t}(v_i,t_i) = -\log\left(\frac{\exp(S(v_i,t_i)/\tau)}{\sum_{j=1}^{N_p}\exp\left(S(v_i,t_j)/\tau\right)}\right) \tag{10}$$

where τ is the temperature parameter. Similarly, the text-to-vision contrastive loss $L^c_{t2v}(t_i,v_i)$ can be expressed as

$$L^c_{t2v}(t_i,v_i) = -\log\left(\frac{\exp\left(S(t_i,v_i)/\tau\right)}{\sum_{j=1}^{N_p}\exp\left(S(t_i,v_j)/\tau\right)}\right). \tag{11}$$

The contrastive loss L_c can be represented as

$$L_c = \frac{1}{2}\sum_{i=1}^{N_p}(L_{v2t}^c(v_i, t_i) + L_{t2v}^c(t_i, v_i)). \tag{12}$$

However, this method treats all image-text pairs equally, which exacerbates overfitting caused by the high similarity and repetition of text samples. Therefore, we propose a weighted contrastive loss function, whose core idea is to weight all image-text pairs corresponding to the text samples according to their values for model learning so as to optimize the model according to more valuable pairs and alleviate overfitting caused by similar or repeated text samples.

Specifically, we measure the value of a text sample to model optimization according to its uniqueness and information richness. In this case, the weight of text sample t_j consists of the uniqueness factor u_j and the richness factor r_j. The more unique or informative t_j is, the greater the value of u_j or r_j are. The weight w_j^{cw} of t_j is calculated by multiplying u_j and r_j, that is

$$w_j^{cw} = u_j \times r_j. \tag{13}$$

We weighted all pairs containing t_j as w_j^{cw}. In this case, the vision-to-text weighted contrastive loss $L_{v2t}^w(v_i, t_i)$ can be expressed as

$$L_{v2t}^w(v_i, t_i) = -\log\left(\frac{\exp\left(w_i^{cw} S(v_i, t_i)/\tau\right)}{\sum_{j=1}^{N_p}\exp\left(w_j^{cw} S(v_i, t_j)/\tau\right)}\right). \tag{14}$$

Similarly, the text-to-vision weighted contrastive loss $L_{t2v}^w(v_i, t_i)$ can be represented as

$$L_{t2v}^w(t_i, v_i) = -\log\left(\frac{\exp(w_i^{cw} S(t_i, v_i)/\tau)}{\sum_{j=1}^{N_p}\exp\left(w_i^{cw} S(t_i, v_j)/\tau\right)}\right). \tag{15}$$

The weighted contrastive loss L_w can be represented as

$$L_w = \frac{1}{2}\sum_{i=1}^{N_p}(L_{v2t}^w(v_i, t_i) + L_{t2v}^w(t_i, v_i)). \tag{16}$$

As mentioned in Section I, the text samples corresponding to an image sample may convey highly consistent information, in which case they are also likely to be the same or similar length. Thus, we can roughly assess the extent to which the information provided by a text sample overlaps with the information in other sentences by the length of the text sample, which is a simple yet effective method. Since sentences longer or shorter than the average length of these sentences are less likely to be repeated with others, we consider these sentences to be more unique. The greater the discrepancy between the length of these sentences and the average length, the more likely they are to diverge from others. Therefore, we define the uniqueness factor as

$$u_j = \begin{cases} 1 + \left|\frac{\text{len}(t_j)}{\frac{1}{N_s}\sum_{k=1}^{N_s}\text{len}(t_{j_k})} - 1\right|, \frac{\text{len}(t_j)}{\frac{1}{N_s}\sum_{k=1}^{N_s}\text{len}(t_{j_k})} \neq 1 \\ \alpha_{\text{avg}}, \text{else} \end{cases} \tag{17}$$

where N_s means the number of text samples corresponding to image v_j with which t_j matches, and t_{j_k} means the kth text of them, hyperparameter $\alpha_{\text{avg}}(\alpha_{\text{avg}} < 1)$ is the weight of texts in average length. $\text{len}(\cdot)$ represents the operation that calculates the length of the text samples, while the absolute value operation is represented by $|\cdot|$.

In addition, as shorter sentences may contain less information, the model is directed to focus its learning on longer sentences by setting the richness factor. Sentences shorter than the average length have a richness factor of $\alpha_s(\alpha_s < 1)$, while the remaining sentences have a richness factor of 1, that is

$$r_j = \begin{cases} \alpha_s, \text{len}(t_{j_k}) < \frac{1}{N_s}\sum_{k=1}^{N_s}\text{len}(t_{j_k}) \\ 1, \text{else.} \end{cases} \tag{18}$$

Note that α_s should be larger than α_{avg} to ensure that the model preferentially learns texts with lower repetition probabilities.

2) Affiliation Loss: To reduce semantic confusion zones [5], we apply affiliation loss [9] for optimization, which is a clustering loss [43] based on contrastive learning and scene category information. Given a batch of samples belonging to N_c classes, the image and text clustering center pairs $\{(v_1^*, t_1^*), (v_2^*, t_2^*), \ldots, (v_{N_c}^*, t_{N_c}^*)\}$ for each class are calculated, respectively. Subsequently, the clustering centers are employed to calculate the vision-to-text similarity score $S(v_i, t_j^*)$ and text-to-vision similarity score $S(t_i, v_j^*)$, where t_j^* and v_j^* represent the clustering centers of the category, which t_j and v_j belong to. Thereafter, the visual-to-text loss $L_{v2t}^*(v_i, t_i^*)$ and text-to-vision loss $L_{t2v}^*(t_i, v_i^*)$ can be expressed as

$$L_{v2t}^*(v_i, t_i^*) = -\log\left(\frac{\exp\left(S(v_i, t_i^*)/\tau\right)}{\sum_{j=1}^{N_p}\exp\left(S(v_i, t_j^*)/\tau\right)}\right) \tag{19}$$

$$L_{t2v}^*(t_i, v_i^*) = -\log\left(\frac{\exp\left(S(t_i, v_i^*)/\tau\right)}{\sum_{j=1}^{N_p}\exp\left(S(t_i, v_j^*)/\tau\right)}\right). \tag{20}$$

Then, the affiliation loss L_a can be denoted as

$$L_a = \frac{1}{2}\sum_{i=1}^{N_p}(L_{v2t}^*(v_i, t_i^*) + L_{t2v}^*(t_i, v_i^*)) \tag{21}$$

The affiliation loss facilitates the learning of common features of the same scene, reducing similarity between different categories, and thereby minimizing semantic confusion zones.

3) Scene-Adaptive Fine-Grained Perceptual Loss: While affiliation loss does bring features in the same scene closer and decrease semantic confusion zones, it can also make it more challenging to differentiate between features in the same scene. Besides, the majority of existing loss functions assign identical importance to all samples, which cannot flexibly adjust the limit of the loss function on model parameter updating. Therefore, we propose a scene-adaptive fine-grained perceptual loss based on the theory of CL, which enhances the fine-grained awareness of the model and the discrimination ability of similar samples by facilitating the learning of differences between similar features in an easy-to-difficult manner.

The original concept of CL was first proposed by Bengio et al. [44], which mimics the way humans learn. The basic idea of CL is "start small" [45], which is to train a machine learning model with simpler data subsets or easier subtasks, gradually increasing the difficulty of the data or subtasks during

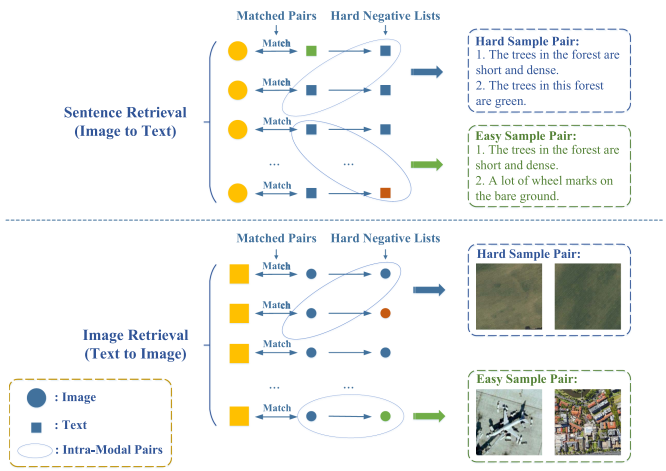

Fig. 4. Principle of scene-adaptive fine-grained perceptual loss. The colors of the images and texts belonging to the same scene are the same except for the queries, which are usually yellow. If the samples belong to the same scene, they form a hard sample pair. Otherwise, they form an easy sample pair. The scene-adaptive fine-grained perceptual loss pushes the samples in the same pair away to enhance the model's ability to discriminate subtle differences.

the training process until the entire training dataset is applied or the target task is achieved [46]. In the field of computer vision, CL has been employed for image classification [47], [48], [49], image segmentation [50], [51], [52], object detection [53], [54], and so on. However, CL is used less frequently in RSITR. To the best of the authors' knowledge, Zhang et al. [6] is the first to apply CL to RSITR, who improved hinge loss based on CL to learn inter-modal alignment from easy to difficult. Although we also introduce loss based on CL, scene-adaptive fine-grained perceptual loss learns differences between similar intramodal features, considering scene information and training epochs when assigning weights to sample pairs.

The principle of the loss is illustrated in Fig. 4. Since the hardest negative samples tend to be more difficult to distinguish, using sets of the hardest negative samples can further increase the semantic similarity of the samples within the pair, thus making it more challenging for the model to distinguish samples. Therefore, we denote $I^- = \{v_1^-, v_2^-, \ldots, v_{N_p}^-\}$ and $T^- = \{t_1^-, t_2^-, \ldots, t_{N_p}^-\}$ as the sets consisting of the hardest negative samples, where v_i^- is the hardest negative sample when t_i is used to retrieve v_i, and t_i^- has a similar meaning. For a pair of intra-modal samples (v_i, v_j) or (t_i, t_j) where $v_i \in I$, $t_i \in T$, $v_j \in I^-$ and $t_j \in T^-$, we calculate their contrastive loss as follows:

$$L_v^s(v_i, v_j) = \begin{cases} \frac{\exp(S(v_i,v_j)/\tau)}{\sum_{k=1}^{N_p} \exp(S(v_i,v_k)/\tau)}, S(v_i, v_j)/\tau \geq \alpha_m \\ 0, \text{else} \end{cases} \tag{22}$$

$$L_t^s(t_i, t_j) = \begin{cases} \frac{\exp(S(t_i,t_j)/\tau)}{\sum_{k=1}^{N_p} \exp(S(t_i,t_k)/\tau)}, S(t_i, t_j)/\tau \geq \alpha_m \\ 0, \text{else} \end{cases} \tag{23}$$

where α_m is the threshold to prevent underfitting due to excessive constraint of the loss. The loss prompts the model to differentiate between highly analogous intra-modal sample pairs, thereby enabling the acquisition of nuanced distinctions between analogous semantics.

Furthermore, as each sample pair encompasses varying amounts of the underlying information, we employ the CL theory to quantify and aggregate the associated losses for each sample pair. According to the CL theory, the difficulty of the knowledge is expected to increase with the progress of training. Therefore, we assign different weights to different sample pairs. Given that samples within the same scene tend to be more challenging to differentiate, we define the pairs as difficult samples. Conversely, pairs consisting of samples of different scenes are classified as easy samples. For easy samples, the weight w_i^s is calculated according to the following formula:

$$w_i^s = \exp(\beta_1 \cdot e_p) \tag{24}$$

where w_i^s is the weight, β_1 is a hyperparameter that affects the weight, and e_p is the number of epochs. Similarly, for a difficult pair, the weight w_i^s can be calculated as

$$w_i^s = \exp(\beta_2 \cdot e_p) \tag{25}$$

hyperparameter β_2 should be larger than β_1 to ensure that difficult samples gradually gain more attention than easy samples during training.

The scene-adaptive fine-grained perceptual loss L_s can be calculated as the sum of all the pairs' losses, that is

$$L_s = \frac{1}{2}\left(\sum_{i=1}^{N_p}\sum_{j=1}^{N_p} w_i^s L_v^s(v_i, v_j) + \sum_{i=1}^{N_p}\sum_{i=j}^{N_p} w_i^s L_t^s(t_i, t_j)\right). \tag{26}$$

In this way, the scene-adaptive fine-grained perceptual loss gradually learns to differentiate between subtle features through an easy-to-difficult strategy, thereby enhancing the fine-grained awareness and discrimination abilities of the model.

4) Value-Guided Learning Framework: At the beginning of training, the model is unable to extract features sufficiently, and it is challenging to mine the potential fine-grained information; thus, the primary problem of this stage is underfitting. With continuous training, the feature extraction ability of the model is gradually enhanced, but the risk of overfitting is also gradually increased. In addition, the RS datasets are likely to be smaller than those in the general scenario and contain a large number of repeated sentences, which exacerbates the problem of overfitting in the later stages of training.

Aiming to adapt to the characteristics of model training, we propose the VGLF, which focuses on learning the most valuable samples in every training stage. Specifically, model training is evenly divided into two phases according to the number of epochs. In the first stage, we apply the contrastive loss to align the features from different modalities and utilize the scene-adaptive fine-grained perceptual loss to encourage the model to learn subtle semantic distinctions in an easy-to-difficult way, thereby enhancing fine-grained awareness. In the second stage, the weighted contrastive loss is employed to focus on more valuable text samples for training, thereby avoiding the interference of low-value text samples and alleviating overfitting. Throughout the entire training process, affiliation loss is used to reduce

semantic confusion. This can be expressed as

$$L = \begin{cases} L_c + \mu_a L_a + \mu_s L_s, \text{the first stage} \\ L_w + \mu_a L_a, \text{the last stage} \end{cases} \tag{27}$$

where L denotes the total loss, μ_a and μ_s are the weights of affiliation loss and scene-adaptive fine-grained perceptual loss, respectively. By employing VGLF, the model is able to select and focus on the most valuable information at different stages of training.

The entire training procedure of FISVL is shown in Algorithm 1. Before training, the pretraining weights and image-text pairs should be provided, as well as the number of epochs. During training, image v_i is fed into the Swin Transformer to extract image feature v^i_{emb}. The image is also passed through the instruction extractor, with the output of the final layer mapped to v^i_{cg}, while the output of the first layer is sent to the SFGS model to obtain v^i_{cl}. Subsequently, all these features are fed into GCN modules for fine-grained perceptual enhancement, then ${v^i_{cls}}^{n_g}$ is added to the [CLS] to obtain the final image feature v^i_{mix}. For text modality, text t_j is embedded and encoded, then projected to t^j_{emb}. Thereafter, the similarity score between each image and text is calculated. Finally, the losses are selected and calculated at different training stages, and the weighted sum of the losses is employed to optimize the parameters. Compared to methods, such as PIR and IEFT [8], FISVL does not incorporate feature enhancement modules based on attention mechanisms, resulting in a more lightweight structure. In addition, VGLF serves as a training strategy that does not appear during the inference stage, thereby reducing the time required for inference.

Algorithm 1: FISVL Training Procedure.

Input: RS training dataset $(V, T) = \{(v_i, t_i)\}_{i=1}^n$, training epoch number e, pretrained weights of F_{tv}, F_{cv} and F_{tt}
Output: weights of FISVL
Initialization: F_{tv}, F_{cv}, F_{tt}, F_{SFGS}, GCN modules and linear projector E_{tv}, E_{cv}, E_v, E_{tt}, e_p

while $e_p < e$ **do**
  for each batch B in (V, T) **do**
    // image feature extraction
    for v_i in B **do**
      $v^i_{emb} = E_{tv}(F_{tv}(v_i, \theta_{F_{tv}}))$
    end for
    // fine-grained information supplementation
    for v_i in B **do**
      $v^i_{cg} = E_{cv}(F_{cv}(v_i, \theta_{F_{cv}}))$
      $v^i_{cl} = F_{SFGS}(g^i_s)$
      for $r = 1$ to $r = n_g - 1$
      $${v^i_{cg}}^{r+1}, {v^i_{cl}}^{r+1}, {v^i_{emb}}^{r+1} = GCN({v^i_{cg}}^{r}, {v^i_{cl}}^{r}, {v^i_{emb}}^{r})$$
      end for
      $v^i_{mix} = E_v({v^i_{cls}}^{n_g}) + v^i_{cls}$
    end for
    // text feature extraction
    for t_j in B **do**
      $t^j_{emb} = E_{tt}(F_{tt}(t_j, \theta_{F_{tt}}))$
    end for
    // calculate similarity scores
    for v_i in B
      for t_j in B
        $S(v_i, t_j) = {v^i_{mix}}^T t^j_{emb}$
      end for
    end for
    // value-guided learning
    if $e_p < 0.5e$ **do**
      $L = L_c + \mu_a L_a + \mu_s L_s$
    end if
    else do
      $L = L_w + \mu_a L_a$
    end else
  update weights of the modules **by** L
  end for
  $e_p = e_p + 1$
end while
return weights of FISVL

IV. Experiments and Discussion

A. Datasets and Evaluation Metrics

Remote sensing image captioning dataset (RSICD) [18] and remote sensing image text match dataset (RSITMD) [17] are two commonly used datasets for RSITR and they are also applied to evaluate our FISVL.

RSICD includes 10 921 images collected from Baidu Map, Google Earth [55], MapABC, and Tianditu. The images in RSICD are divided into 30 categories, including airports, forests, ponds, and so on. Each image is in the shape of 224 × 224 and matches with five corresponding sentences. Following [9], we divide the data into a training set, a validation set, and a test set, with the number of images being 7862, 1966, and 1093, respectively.

Images of RSITMD are collected from RSICD and Google Earth, containing 4743 in total with a size of 256 × 256. Each image contains five corresponding texts, for a total of 23 715. All images of RSITMD are divided into 32 categories. Compared to RSICD, RSITMD has more fine-grained text descriptions and richer scene changes. We follow [9] to divide the data into a training set with 3,435 images, a validation set with 856 images, and a test set with 452 images.

We employ $R@K(K = 1, 5, 10)$ and mR to evaluate the performance of the model. $R@K$ represents the proportion of retrieval results that match the query among the top K results, while mR is the average value of $R@K$, reflecting the overall accuracy of the model retrieval.

B. Implementation Details

In this article, ResNet-50, pretrained on the AID [55] dataset, is applied as the instruction extractor in our FISVL. We use Swin-Transformer (tiny version) pre-trained on ImageNet [56]

TABLE I
COMPARISONS OF DIFFERENT METHODS ON RSICD AND RSITMD TEST SETS

Method	RSICD							RSITMD						
	Sentence Retrieval (I→T)			**Image Retrieval (T→I)**				**Sentence Retrieval (I→T)**			**Image Retrieval (T→I)**			
	$R@1$	$R@5$	$R@10$	$R@1$	$R@5$	$R@10$	mR	$R@1$	$R@5$	$R@10$	$R@1$	$R@5$	$R@10$	mR
LW-MCR-u [12]	4.39	13.35	20.29	4.30	18.85	32.34	15.59	9.73	26.77	37.61	9.25	34.07	54.03	28.58
AMFMN-soft [17]	5.05	14.53	21.57	5.05	19.74	31.04	16.02	11.06	25.88	39.82	9.82	33.94	51.90	28.74
AMFMN-fusion [17]	5.39	15.08	23.40	4.90	18.28	31.44	16.42	11.06	29.20	38.72	9.96	34.03	52.96	29.32
AMFMN-sim [17]	5.21	14.72	21.57	4.08	17.00	30.60	15.53	10.63	24.78	41.81	11.51	34.69	54.87	29.72
GaLR [11]	6.59	19.85	31.04	4.69	19.48	32.13	18.96	14.82	31.64	42.48	11.15	36.68	51.68	31.41
HyperMatch [58]	6.08	20.37	33.82	7.14	20.04	31.02	19.75	9.16	32.31	46.64	11.73	28.10	38.05	27.66
KCR [59]	5.95	18.59	29.58	5.40	22.44	37.36	19.89	-	-	-	-	-	-	-
HVSA [6]	7.47	20.62	32.11	5.51	21.13	34.13	20.16	13.20	32.08	45.58	11.43	39.20	57.45	33.16
SWAN [5]	7.41	20.13	30.86	5.56	22.26	37.41	20.61	13.35	32.15	46.90	11.24	40.40	60.60	34.11
DOVE [60]	8.66	22.35	34.95	6.04	23.95	40.35	22.72	16.81	36.80	49.93	12.20	**44.13**	**66.50**	37.73
MSIT [10]	6.46	23.69	38.45	4.96	20.74	35.86	21.69	14.37	33.18	48.74	9.34	37.62	58.81	33.67
SMLGN [13]	8.87	25.53	37.24	**7.85**	**27.14**	42.58	24.87	17.26	39.38	51.55	**13.19**	43.94	60.40	37.62
PIR [9]	9.88	27.26	39.16	6.97	24.56	38.92	24.46	**18.14**	41.15	52.88	12.17	41.68	63.41	38.24
FISVL(Ours)	**11.49**	**28.75**	**41.56**	7.22	26.46	**42.88**	**26.39**	16.81	**41.33**	**56.24**	12.65	43.40	64.33	**39.13**

Bold indicates the best result under a certain metric, while underline indicates the second best result.

as a Transformer-based visual feature extractor. We employ BERT-B as the text encoder, with pretrained parameters provided by the official source. The output dimensions for the Transformer-based vision feature extractor and Transformer-based text feature extractor are set to 768 and then mapped to 512 dimensions. The dimension for the last layer's output of the CNN-based vision extractor is set to 2048 and is also mapped to 512 dimensions. Two GCN modules are employed to facilitate the exchange of information between visual features. The temperature parameter τ of all loss functions is set to 0.07, and the weights of the affiliation loss function and the scene-adaptive fine-grained perceptual loss are both set to 1. In the weighted contrastive loss, the value of α_{avg} is 0.7, and the value of α_s is 0.95. For the scene-adaptive fine-grained perceptual loss, the value of β_1 is -0.002, the value of β_2 is -0.001, and the value of α_m is 0.95 on the RSICD dataset, while β_1, β_2, and α_m are set to -0.0015, 0.02, and 1.15 on the RSITMD dataset. The experiments are executed in parallel on two RTX3090Ti GPUs with 24GB memory, utilizing the AdamW [57] optimizer with a learning rate of 4e-5 on the RSICD dataset and 6e-5 on the RSITMD dataset, and the weight decay is 0.01. The batch size is 135 on the RSICD dataset and 128 on the RSITMD dataset, with 20 epochs on both datasets. We also set the dropout rate to 0.2 to prevent overfitting. To ensure the reliability of the experimental results, we repeat the experiment five times and take the average as the final experimental result.

Before feeding data into the model, we perform a series of preprocessing operations on the texts and images, respectively. With regard to the texts, we first undertake cleaning and normalization, which involves the removal of punctuation. We then normalize the text length to 30 and pad the shorter texts. Furthermore, we find that randomly removing a portion of words during training can alleviate the overfitting problem and improve the model's accuracy. Meanwhile, the text repetition problem in the RSICD dataset is more pronounced than that in the RSITMD dataset. Therefore, during model training, each word is removed with a probability of 0.06 on the RSICD dataset and 0.03 on the RSITMD dataset. For the images, we first resize them to 224×224, which is followed by the applications of some image augmentation methods, including random flipping, Gaussian blurring, random erasing, and so on.

C. Comparisons With Other Approaches

We compare our proposed FISVL with the following methods for RSITR using the RSICD and RSITMD datasets to provide a more comprehensive evaluation. The methods included in the comparison are as follows: 1–8 represent traditional CNN-based RSITR approaches, while 9-11 are Transformer-based methods. The results are displayed in Table I.

1) LW-MCR [12]: LW-MCR is a lightweight model that combines multiscale visual features and performs dynamic filtering. In addition, a knowledge distillation-based hidden supervised optimization method is employed to learn relevant knowledge from a large-scale model. There are three versions of LW-MCR, and we use LW-MCR-u which has the highest mR for comparison.

2) AMFMN [17]: Asymmetric multimodal feature matching network (AMFMN) extracts salient features from multiscale features extracted by different layers of the vision encoder, with the output features subsequently utilized to guide the text representation. It is divided into three versions, namely AMFMN-soft, AMFMN-fusion, and AMFMN-sim, and we utilize all of them for comparison with our FISVL.

3) GaLR [11]: GaLR performs feature enhancement for small objects, then aggregates multiscale features and proposes a rerank algorithm to optimize similarity scores.

4) HVSA [6]: Hypersphere based visual semantic alignment (HVSA) is a hypersphere-based visual semantic alignment network that employs a contrastive learning-based adaptive alignment strategy to align data pairs from relatively simple to more challenging.

5) HyperMatch [58]: HyperMatch utilizes a multiscale RS image hypergraph network to model the intricate relationships between RS image features. It employs a method specifically designed for RS images to construct and update the hypergraph.

6) KCR [59]: Knowledge aware cross modal retrieval (KCR) addresses the issue of intramodal information asymmetry by

leveraging relevant information mined from an external knowledge graph to enhance the textual richness of the query. We use KCR with ResNet-101 for comparison.

7) SWAN [5]: SWAN states that the distinctive characteristics of RS images may give rise to semantic confusion, and reduce the semantic confusion by improving the model's fine-grained awareness of the scene.

8) DOVE [60]: DOVE employs a multiscale vision encoder and a region of interest encoder to encode visual features, a text encoder to encode text features, and a digging text genome assistant for text enhancement. The visual and text outputs are then fed into a region-oriented attention module for cross-modal modeling.

9) MSIT [10]: MSIT utilizes a mutually interactive learning module, comprising a Transformer encoder and a lightweight multiscale feature exploration submodule, to enhance text and visual features. As Wang et al. [10] does not conduct experiments on the RSICD and RSITMD datasets, we follow Hu et al. [61] to evaluate the performance of MSIT on the aforementioned two datasets.

10) SMLGN [13]: SMLGN proposes a multilevel guidance learning method, which introduces bidirectional guidance of intermodal fine-grained information and intramodal fusion guidance between local and global information.

11) PIR [9]: PIR employs prior knowledge to facilitate the adaptive learning of visual and textual features. It also improves the Transformer encoder layer to support RSITR.

On the RSICD dataset, most of the Transformer-based methods exhibit considerably higher overall accuracy than traditional approaches on the RSICD dataset, indicating that they are more effective at extracting vision and language features. Among these CNN-based and Transformer-based methods, FISVL achieves the highest accuracy in most cases, surpassing the highest overall accuracy of CNN-based and Transformer-based models by 16.15% and 6.11%, respectively, thereby substantiating the benefit of integrating the features extracted by Transformers with those derived from CNN-based feature encoding and enhancement modules. Specifically, in the sentence retrieval task, FISVL outperforms the second-ranked scores by 16.30%, 5.47%, 6.13% in $R@1$, $R@5$, and $R@10$, respectively. In the image retrieval task, FISVL surpasses the second-ranked score by 0.70% in $R@10$, while the $R@1$ and $R@5$ are lower than SMLGN.

On the RSITMD dataset, the Transformer-based PIR and FISVL still exhibit higher overall accuracy compared with traditional methods. FISVL consistently achieves the highest accuracy in mR, thereby demonstrating its superior performance. Specifically, FISVL outperforms the second-ranked scores in $R@5$ and $R@10$ for sentence retrieval, particularly with a 6.35% improvement in $R@10$ for sentence retrieval. While not achieving the highest accuracy in image retrieval, FISVL still attains the second-highest accuracy in $R@1$ and $R@10$, with a margin of only 1.24% from the second-highest accuracy in $R@5$. In comparison to the second-ranked PIR model in terms of overall accuracy, our model achieves a 2.33% improvement in mR. It is noteworthy that our model does not rely on attention-based modules with a large number of parameters to enhance image and text features. In fact, we only perform a simple mapping on the text features extracted by BERT in the text branch. This highlights the advantage of sufficient information exchange between global awareness and local awareness visual features and focusing on the most valuable information at different stages of training.

D. Ablation Studies

To demonstrate the effectiveness of the methods proposed in this article, we perform systematic ablation experiments. We conduct the experiments on both the RSICD and RSITMD datasets to ensure the reliability of the results. The experiments are divided into four groups: 1) removing the FGIS module; 2) using the contrastive loss throughout the training process instead of the weighted contrastive loss; 3) without the scene-adaptive fine-grained perceptual loss; 4) using the proposed FISVL. The results of these experiments are presented in Tables II and III. It can be observed from the tables that the proposed FISVL achieves the highest mR value on both datasets. This indicates that each module proposed has contributed to the improvement of the model's accuracy.

Comparing the results of experiment groups 1) and 4), it is easy to find that the inclusion of the FGIS module leads to a notable improvement in mR, with a 4.02% increase observed on the RSICD dataset and a 5.76% increase observed on the RSITMD dataset. The presence of the FGIS module demonstrates enhancements across all evaluation metrics, and the substantial disparity in accuracy between groups 1) and 4) indicates that the FGIS module could take advantage of the fine-grained features extracted from the instruction extractor to improve the features extracted from the Transformer-based vision extractor, thereby augmenting the fine-grained awareness of the model. Notably, the performance improvement on the RSITMD dataset surpasses that of the RSICD dataset. This is due to the finer granularity of the RSITMD dataset and the greater need to extract complex, fine-grained information.

With the introduction of the weighted contrastive loss, the model shows a remarkable overall performance improvement of 4.60% on the RSICD dataset, making it the most significant accuracy improvement among the experiments conducted on RSICD. The reason is that the text richness of the RSICD dataset is not high enough, and a large number of similar texts make the model prone to overfitting during training. The weighted contrastive loss measures the value of samples based on their uniqueness and information richness, allowing the model to focus on more crucial samples and alleviating the overfitting issue. Although the overall accuracy of FISVL on the RSITMD dataset does not show significant improvement compared with group 2) due to the lower repetition rate of RSITMD, it exhibits higher $R@1$ and $R@5$ scores in both sentence retrieval and image retrieval tasks. This suggests that the application of the weighted contrastive loss enables the model to improve the metrics that are more challenging without compromising the overall accuracy, thereby providing a better retrieval experience.

By incorporating the scene-adaptive fine-grained perceptual loss, the overall performances of the model on the RSICD and RSITMD datasets are increased by 0.21 and 0.19 points to 26.39 and 39.13, respectively. This highlights the effectiveness of the

TABLE II
ABLATION EXPERIMENTS ON THE RSICD TEST SET

Group	Method			Sentence Retrieval (I→T)			Image Retrieval (T→I)			mR
	$FGIS$	L_w	L_s	$R@1$	$R@5$	$R@10$	$R@1$	$R@5$	$R@10$	
1		✓	✓	10.59	27.15	39.69	7.04	25.87	41.90	25.37
2	✓		✓	10.17	27.01	38.85	**7.36**	26.37	41.64	25.23
3	✓	✓		**11.73**	28.55	40.95	7.31	25.98	42.59	26.18
4	✓	✓	✓	11.49	**28.75**	**41.56**	7.22	**26.46**	**42.88**	**26.39**

Bold indicates the best result under a certain metric.

TABLE III
ABLATION EXPERIMENTS ON THE RSITMD TEST SET

Group	Method			Sentence Retrieval (I→T)			Image Retrieval (T→I)			mR
	FGIS	L_w	L_s	$R@1$	$R@5$	$R@10$	$R@1$	$R@5$	$R@10$	
1		✓	✓	14.74	36.68	54.73	12.16	41.11	62.60	37.00
2	✓		✓	16.70	41.19	**56.64**	12.20	43.32	64.46	39.08
3	✓	✓		**17.66**	39.78	54.74	**12.91**	**43.92**	**64.65**	38.94
4	✓	✓	✓	16.81	**41.33**	56.24	12.65	43.40	64.33	**39.13**

Bold indicates the best result under a certain metric.

TABLE IV
COMPARISONS OF DIFFERENT LEARNING STRATEGIES ON THE RSITMD TEST SET

Method	Sentence Retrieval (I→T)			Image Retrieval (T→I)			mR
	$R@1$	$R@5$	$R@10$	$R@1$	$R@5$	$R@10$	
1) L_s used in the first stage	16.70	41.19	**56.64**	12.20	43.32	64.46	39.08
2) L_s used in all stages	17.13	40.58	55.93	12.52	43.55	64.01	38.95
3) L_w used in the last stage	**17.66**	39.78	54.74	**12.91**	**43.92**	**64.65**	38.94
4) L_w used in all stages	17.57	38.98	54.47	**12.91**	43.03	64.09	38.51
5) FISVL	16.81	**41.33**	56.24	12.65	43.40	64.33	**39.13**

Bold indicates the best result under a certain metric.

scene-adaptive fine-grained perceptual loss in improving the model's fine-grained representational ability by capturing the subtle differences between similar samples within each modality and learning from easy to difficult.

Combined with the experimental results on the RSICD and RSITMD datasets, the effect of the proposed methods can be confirmed. All of these methods contribute to improving the accuracy of the datasets. The FGIS module has a greater impact on the accuracy improvement of the RSITMD dataset, which can be attributed to the fine-grained nature of the dataset. On the other hand, the weighted contrastive loss is designed to focus on key samples, which alleviates the problem of high similarity within text modality. This is more beneficial for improving the accuracy of models trained on the RSICD dataset, which exhibits lower sample richness.

E. Learning Strategy Analysis

To verify the effectiveness of the proposed VGLF, we conduct five groups of experiments on the RSITMD dataset, the results of which are shown in Table IV. Among the experiments, group 5) is the proposed model, while group 1) represents the use of scene-adaptive fine-grained perceptual loss in the first stage of training. Compared with group 5), the weighted contrastive loss is not utilized in the second stage. Group 2) represents the application of scene-adaptive fine-grained perceptual loss throughout the training phase without weighted contrastive loss. Group 3) leverages weighted contrastive loss in the second stage but does not apply scene-adaptive fine-grained perceptual loss in the first stage compared with group 5). Group 4) applies weighted contrastive loss throughout the entire process of training without applying scene-adaptive fine-grained perceptual loss. Affiliation loss is used throughout training in all groups. For the training phases where weighted contrastive loss is not used, we use contrastive loss for intermodal feature alignment.

Table IV reveals that, in comparison to the overall performance of group 1), there is no improvement in group 2), but rather a slight decrease. This is because the scene-adaptive fine-grained perceptual loss amplifies the differences between similar samples. Although this encourages the model to learn the fine-grained representation in the early stage of model training, it will form too strong constraints when the feature extraction is more comprehensive, leading to overfitting the model and extracting some irrelevant information and noise.

Like the scene-adaptive fine-grained perceptual loss, using the weighted contrastive loss only in the second stage yields better results than using it throughout the training process. This may be attributed to the fact that while the weighted contrastive loss can mitigate overfitting issues, its early application during model training may introduce a risk of underfitting.

The combination of scene-adaptive fine-grained perceptual loss and weighted contrastive loss resulted in the highest average accuracy for group 5). This is due to the fact that the

TABLE V
SIMILARITY SCORES BETWEEN THE TOP FIVE SEARCH RESULTS AND THE QUERIES ON RSITMD TEST SET

Group	Rank 1	Rank 2	Rank 3	Rank 4	Rank 5
1	1.0000	0.9515	0.9194	0.8946	0.8741
2	1.0000	0.9519	0.9207	0.8963	0.8749
3	1.0000	0.9490	0.9171	0.8914	0.8698
4	1.0000	0.9337	0.8899	0.8567	0.8283
5	1.0000	0.9330	0.8902	0.8550	0.8247
6	1.0000	0.9418	0.9037	0.8729	0.8479

The scores are divided by the score of the first retrieval results.

scene-adaptive fine-grained perceptual loss enables the model to learn fine-grained features, thereby facilitating continuous fitting to the datasets. Furthermore, the weighted contrastive loss employed in the second stage prevents the model from overfitting. These findings demonstrate the efficacy of VGLF.

To further investigate the impact of different optimization strategies on FISVL's ability to discriminate similar features, we compute the average similarity scores of the top five retrieval results for models trained using different optimization strategies on the RSITMD dataset. Among the optimization strategies, group 1) does not utilize the scene-adaptive fine-grained perceptual loss or the weighted contrastive loss, while group 2) and group 3) apply the scene-adaptive fine-grained perceptual loss in the first stage and throughout training, respectively. Group 4) and group 5) apply the weighted contrastive loss in the second stage and throughout the entire training process, respectively. Group 6) corresponds to FISVL. Similarly to the above, affiliation loss is employed throughout training, and contrastive loss is utilized during the training phases without weighted contrastive loss. We divide the scores by the similarity between the first retrieval result and the query. The results are shown in Table V.

The data of group 1), group 2), and group 3) shown in Table V indicates that applying the scene-adaptive fine-grained perceptual loss only in the first half of training still struggles to discriminate between the top few retrieval results. However, employing the loss throughout the training process further increases the similarity difference between the most similar retrieval result and the rest. This may be because utilizing the scene-adaptive fine-grained perceptual loss only in the first stage of training tends to encourage the model to learn the differences between samples with relatively low similarity. Conversely, the continuous application of this loss enables the model to further discern the distinctions between the most similar samples. This also explains why in most cases, group 4) in Tables II and III has higher accuracy in $R@5$ and $R@10$ compared to group 3), but does not demonstrate an advantage in $R@1$.

Comparing group 1), group 4), and group 5), it can be observed that using the weighted contrastive loss either in the second stage or throughout the training process significantly increases the similarity scores difference between the retrieval results. Compared with only using the weighted contrastive loss in the second stage, applying the loss throughout the training process further widens the similarity gap between retrieval results in most cases.

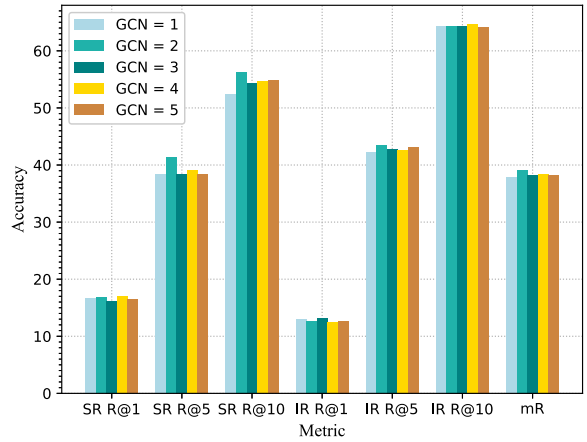

Fig. 5. Results of experiments of the number of GCN modules on the RSITMD test set. "SR" means sentence retrieval, and "IR" denotes image retrieval.

When using these two loss functions at different stages, the model still maintains a relatively high discriminative ability across different samples, although slightly lower than group 4). Compared to group 1), FISVL can distinguish the most similar features more clearly. Note that the significant differences in similarity scores are beneficial for the model to differentiate them but have no direct impact on accuracy. Nevertheless, as the similarity scores' difference is enlarged, the features are more likely to be evenly distributed in the common subspace, making the model more robust.

F. Sensitivity Analysis

In this section, we investigate the effect of some hyperparameter settings on the accuracy of the model. Specifically, we first examine the effect of the number of GCN modules in the FGIS on accuracy. Subsequently, we investigate the impact of varying the value of μ_s on accuracy. Finally, we analyze the effect of α_m on accuracy. All experiments are performed on the RSITMD dataset.

1) Number of GCN Module in FGIS: We examine the influence of using different numbers of GCN modules in the FGIS on accuracy. The number of GCN modules is set from 1 to 5, and the experimental results are presented in Fig. 5. It can be seen from the figure that the best performance is achieved by using two GCN modules in most cases.

When the number of GCN modules increases from 1 to 2, there are improvements in most metrics. This indicates that increasing the number of GCN modules allows for more comprehensive information exchange and feature enrichment when there are few modules. Nevertheless, when the number of GCN modules increases from 2 to 5, the overall performance of the model shows slight fluctuations but remains lower than the case with two modules. This may be because the excessive number of GCN layers leads to excessive aggregation of node features, resulting in smoothed features with reduced resolution [62]. Therefore, we apply two GCN modules in FGIS.

2) Influence of μ_s: The value of μ_s determines the importance of the scene-adaptive fine-grained perception loss. We conduct a series of experiments to study the influence of μ_s on

TABLE VI
RESULTS OF EXPERIMENTS OF μ_s ON THE RSITMD TEST SET

μ_s	Sentence Retrieval (I→T)			Image Retrieval (T→I)			mR
	$R@1$	$R@5$	$R@10$	$R@1$	$R@5$	$R@10$	
0.01	17.7	39.42	54.51	**12.78**	**43.73**	64.25	38.73
0.1	**17.83**	38.80	55.35	12.18	43.53	**64.49**	38.70
1	16.81	**41.33**	**56.24**	12.65	43.40	64.33	**39.13**
10	10.04	28.63	46.24	9.91	36.20	59.00	31.67
100	8.81	26.90	41.86	7.09	30.49	52.62	27.96

Bold indicates the best result under a certain metric.

TABLE VII
RESULTS OF EXPERIMENTS OF α_m ON THE RSITMD TEST SET

α_m	Sentence Retrieval (I→T)			Image Retrieval (T→I)			mR
	$R@1$	$R@5$	$R@10$	$R@1$	$R@5$	$R@10$	
0.75	16.99	40.26	56.46	12.55	43.51	**64.66**	39.07
0.95	**18.14**	40.75	55.35	12.31	43.58	64.50	39.11
1.15	16.81	**41.33**	56.24	12.65	43.40	64.33	**39.13**
1.35	16.77	39.96	**56.77**	12.60	**43.85**	64.58	39.09
1.55	17.66	40.58	55.88	12.66	43.66	64.28	39.12
1.75	17.57	39.69	55.26	**12.69**	43.40	64.31	38.82

Bold indicates the best result under a certain metric.

performance. We vary the value of μ_s from 0.01 to 100, which is shown in Table VI.

The results shown in the table reveal that, in general, the overall performance initially improves and then decreases as μ_s increases, with the optimal value being reached at $\mu_s = 1$. This is because when μ_s is less than 1, the influence of the loss function on the performance of the model is insufficient, so the model does not reach its optimal performance. However, when μ_s is set to 10 or 100, the model becomes overly focused on small differences between highly similar features within the same modality, neglecting cross-modal alignment. This also weakens the contribution of the affiliation loss to accuracy, resulting in a decrease in performance.

3) Influence of α_m: α_m affects the degree of limitation imposed by the scene-adaptive fine-grained perceptual loss on the model. We vary the value of α_m from 0.75 to 1.75 to observe the changes in accuracy for different values of α_m, and the experimental results are presented in Table VII.

In general, the slight variations in mR with differing α_m values suggest that FISVL is not susceptible to alterations in α_m. This signifies the model's resilience and its capacity to retain consistent performance across diverse α_m ranges.

The highest average accuracy is achieved when α_m is set to 1.15. When α_m is less than 1.15, the accuracy exhibits a slight decline as α_m decreases. This is due to the fact that when α_m is low, more samples meet the threshold limitation, and the scene-adaptive fine-grained perceptual loss exerts an undue constraint on FISVL. Consequently, FISVL becomes unduly inclined to differentiate similar semantic features, leading to overfitting and a subsequent decline in model accuracy.

When α_m exceeds 1.15 and continues to increase, the average accuracy exhibits minor fluctuations initially, followed by a significant decrease when α_m reaches 1.75. This suggests that when α_m reaches 1.75, fewer samples meet the threshold limitation. Consequently, the scene-adaptive fine-grained loss mitigates the constraints imposed by analogous semantics in the same modality of the model. Therefore, the model's fine-grained discrimination ability is diminished, leading to a decline in performance.

G. Examples of Semantic Localization (SeLo)

SeLo is the process of identifying the most pertinent objects within RS images through the utilization of semantic annotations at the textual level, such as captions. Some studies provide SeLo approaches to evaluate the cross-modal modeling capabilities of RS models [63], [64]. Yuan et al. [63] proposed a SeLo method that initially performs multiscale slicing on RS images, followed by calculating cross-modal similarities between queries and an array of image patches. The computed region probabilities are then subjected to pixel-level averaging and noise suppression, resulting in the generation of a SeLo map.

To assess the cross-modal semantic understanding capability of FISVL, we conducted SeLo on the AIR-SLT dataset [63]. The dataset comprises 59 testing samples, each comprising a large RS image, a query text, and one or more corresponding localization bounding boxes.

Fig. 6 illustrates some examples of SeLo. In Fig. 6(a), FISVL identifies the two white-roofed houses referenced in the query. As a consequence of the utilization of "gray parking lot" in the query text to delineate the location of the white-roofed houses, FISVL also assigns a degree of attention to the gray parking lot. Nevertheless, FISVL does not attend to the other white houses in the image, indicating that it is able to comprehend the spatial relationships between objects.

In Fig. 6(b), the objective is to identify the location of the planes parked at the airport, which are depicted in the image at a small scale. From the located image, it can be observed that FISVL focuses its attention on the airport. The probability map further reveals that the specific parking spots of the airplanes receive most of the attention. This indicates that FISVL is capable of perceiving small-scale objects like airplanes and exhibits superior fine-grained perception, which could be attributed to the utilization of the CNN-based instruction extractor.

In Fig. 6(c), the inquiry pertains to the localization of six tennis courts. Meanwhile, there is an athletic field that shares a similar visual appearance with the tennis courts in the image, as both are characterized by green sports surfaces and surrounding rubberized tracks. Furthermore, the "green forest" referenced in the query is situated in the vicinity of the athletic field. However, the focus of FISVL is not on the athletic field. This is evidenced by its ability to differentiate between the tennis court and the athletic field, as well as to comprehend the concept of "red land" mentioned in the query. This illustrates the capacity of FISVL to discern subtle distinctions and its proficiency in semantic understanding.

H. Qualitative Results

To provide a more intuitive demonstration of the effectiveness and mechanisms of FISVL, we present examples of sentence retrieval and image retrieval on the RSITMD dataset using FISVL and its variants in Figs. 7 and 8, respectively. In these figures, "FISVL-CNN" represents FISVL with the Transformer-based

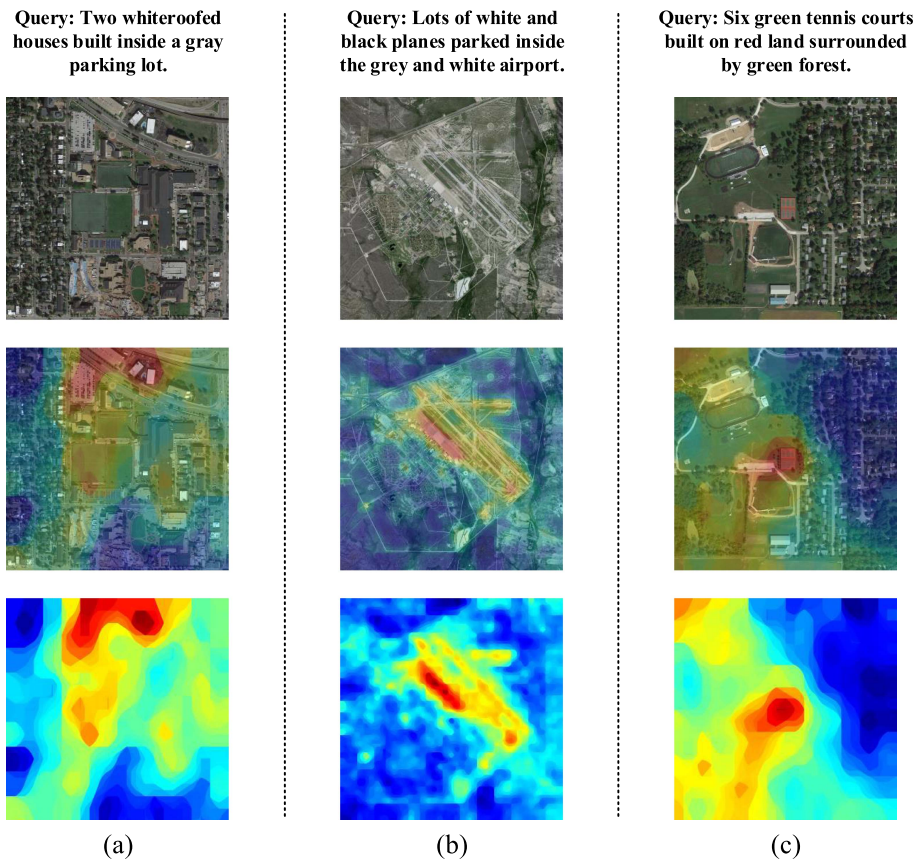

Fig. 6. Some examples of SeLo on the AIR SLT dataset. (a), (b) and (c) are three examples respectively, each comprises the query text, the original image, the located image, and the probability map, presented from top to bottom.

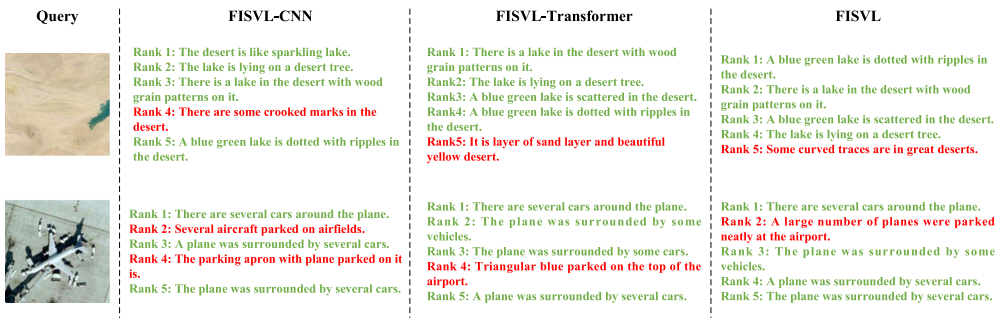

Fig. 7. Some examples of sentence retrieval on the RSITMD test set. Red fonts indicate unmatched search results, while green fonts indicate matched results.

visual feature extractor removed, and instead, it uses the features extracted by the instruction extractor as input to the GCNs. "FISVL-Transformer" denotes a network that only utilizes the Transformer-based feature extractor to extract visual features, which are then input to the GCNs. The retrieval results in the figures are sorted based on their similarity to the query.

From these retrieval examples, it can be observed that FISVL and its variants demonstrate good fine-grained awareness capabilities. For example, in the first sentence retrieval example, the majority of sentences retrieved by FISVL and its variants are found to match the image. Moreover, the remaining sentences exhibit characteristics that were present in the image, including "desert," "sand," "yellow," and "curved traces."

Furthermore, thanks to the excellent fine-grained perception and multiscale visual feature extraction capability of the instructor extractor and SFGS, FISVL exhibits fine-grained and multiscale visual feature perception. To be more precise, in the first example of image retrieval, there are multiple query terms such as "barelands,", "park," "building," and "swimming pool" represented at different scales. The results demonstrate that FISVL and FISVL-CNN retrieve images that match this query with higher ranks in comparison to FISVL-Transformer. This evidence suggests that the utilization of CNN for visual feature extraction enhances the model's multiscale visual perception capability. In the second example of image retrieval, the query terms "white building," "Russian bicycles," "square buildings,"

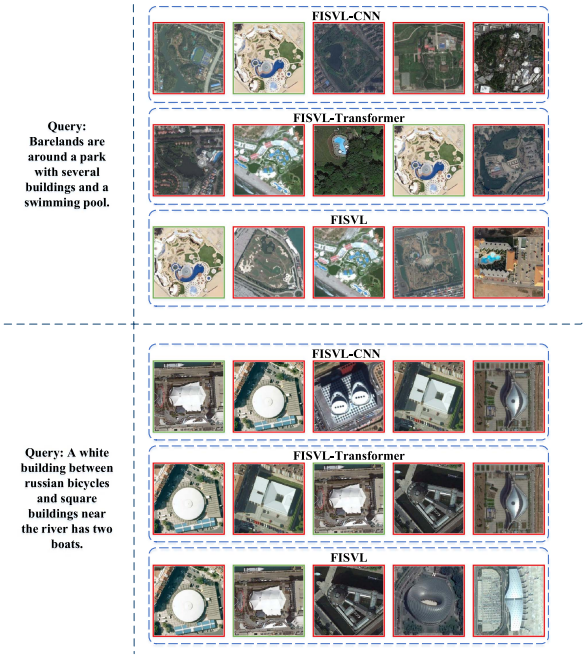

Fig. 8. Some examples of image retrieval on the RSITMD test set. Red boxes indicate unmatched search results, while green boxes indicate matched results.

"river," and "boats" exhibit disparate scales, with "Russian bicycles" and "boats" representing smaller-scale objects. In this instance, the FISVL-CNN model successfully retrieves the correct image with the highest rank. These two examples demonstrate that the CNN-based encoder and SFGS module, in comparison to the Transformer-based one, exhibit superior fine-grained and multiscale visual perception capabilities, thereby enhancing the model's retrieval performance.

However, it is worth noting that FISVL may not be sensitive to certain quantifiers, resulting in some retrieval results that do not match the query in terms of quantity. For example, in the second sentence retrieval example, both the second retrieval results of FISVL-CNN and FISVL confuse the quantities of airplanes and cars.

V. Conclusion

In this article, we propose an RSITR model that leverages prior knowledge in the RS domain for feature supplementation and adopts a value-guided training approach to learn more fine-grained, expressive, and robust feature representations. The proposed FGIS module effectively integrates local and global visual features, enhancing fine-grained perception while maintaining global perception. Besides, we introduce the weighted contrastive loss and the scene-adaptive fine-grained perceptual loss to address the problem of high intramodal similarity. In addition, the VGLF is designed according to the training characteristics of the model, focusing on high-value information at each stage of training. Extensive experiments are conducted to demonstrate the effectiveness of our proposed model.

References

[1] X. Sun et al., "From single-to multi-modal remote sensing imagery interpretation: A survey and taxonomy," *Sci. China Inf. Sci.*, vol. 66, no. 4, 2023, Art. no. 140301.

[2] R. Yang et al., "Multimodal fusion remote sensing image–audio retrieval," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 6220–6235, Aug. 2022.

[3] T. Zhao et al., "Artificial intelligence for geoscience: Progress, challenges and perspectives," *Innovation*, vol. 5, no. 5, pp. 1–25, Sep. 2024.

[4] W. Zhou, H. Guan, Z. Li, Z. Shao, and M. R. Delavar, "Remote sensing image retrieval in the past decade: Achievements, challenges, and future directions," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 1447–1473, Jan. 2023.

[5] J. Pan, Q. Ma, and C. Bai, "Reducing semantic confusion: Scene-aware aggregation network for remote sensing cross-modal retrieval," in *Proc. 2023 ACM Int. Conf. Multimedia Retrieval*, 2023, pp. 398–406.

[6] W. Zhang et al., "Hypersphere-based remote sensing cross-modal text–image retrieval via curriculum learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, Sep. 2023, Art. no. 3318227.

[7] Z. Ji, C. Meng, Y. Zhang, Y. Pang, and X. Li, "Knowledge-aided momentum contrastive learning for remote-sensing image text retrieval," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, Nov. 2023, Art. no. 5625213.

[8] X. Tang, Y. Wang, J. Ma, X. Zhang, F. Liu, and L. Jiao, "Interacting-enhancing feature transformer for cross-modal remote-sensing image and text retrieval," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, May 2023, Art. no. 5611715.

[9] J. Pan, Q. Ma, and C. Bai, "A prior instruction representation framework for remote sensing image-text retrieval," in *Proc. 31st ACM Int. Conf. Multimedia*, 2023, pp. 611–620.

[10] Y. Wang, J. Ma, M. Li, X. Tang, X. Han, and L. Jiao, "Multi-scale interactive transformer for remote sensing cross-modal image-text retrieval," in *Proc. IGARSS 2022-2022 IEEE Int. Geosci. Remote Sens. Symp.*, 2022, pp. 839–842.

[11] Z. Yuan et al., "Remote sensing cross-modal text-image retrieval based on global and local information," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Mar. 2022, Art. no. 5620616.

[12] Z. Yuan et al., "A lightweight multi-scale crossmodal text-image retrieval method in remote sensing," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Oct. 2021, Art. no. 5612819.

[13] Y. Chen, J. Huang, S. Xiong, and X. Lu, "Integrating multisubspace joint learning with multilevel guidance for cross-modal retrieval of remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, Feb. 2024, Art. no. 4702217.

[14] M. M. Al Rahhal, Y. Bazi, N. A. Alsharif, L. Bashmal, N. Alajlan, and F. Melgani, "Multilanguage transformer for improved text to remote sensing image retrieval," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 9115–9126, Oct. 2022.

[15] K. Yuan, S. Guo, Z. Liu, A. Zhou, F. Yu, and W. Wu, "Incorporating convolution designs into visual transformers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 579–588.

[16] H. Wu et al., "Cvt: Introducing convolutions to vision transformers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 22–31.

[17] Z. Yuan et al., "Exploring a fine-grained multiscale method for cross-modal remote sensing image retrieval," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, May 2021, Art. no. 4404119.

[18] X. Lu, B. Wang, X. Zheng, and X. Li, "Exploring models and data for remote sensing image caption generation," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 4, pp. 2183–2195, Apr. 2018.

[19] Y. Zhang, X. Zheng, and X. Lu, "Remote sensing cross-modal retrieval by deep image-voice hashing," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 9327–9338, Oct. 2022.

[20] M. Cao, S. Li, J. Li, L. Nie, and M. Zhang, "Image-text retrieval: A survey on recent research and development," 2022, *arXiv:2203.14713*.

[21] P. Kaur, H. S. Pannu, and A. K. Malhi, "Comparative analysis on cross-modal information retrieval: A review," *Comput. Sci. Rev.*, vol. 39, 2021, Art. no. 100336.

[22] H. Chen, G. Ding, X. Liu, Z. Lin, J. Liu, and J. Han, "Imram: Iterative matching with recurrent attention memory for cross-modal image-text retrieval," in *Proc. IEEE/CVF Conf. Comput. Vis. pattern Recognit.*, 2020, pp. 12655–12663.

[23] Y. Zeng, X. Zhang, and H. Li, "Multi-grained vision language pre-training: Aligning texts with visual concepts," in *Proc. Int. Conf. Mach. Learning. PMLR*, 2022, pp. 25994–26009.

[24] Z. Huang, Z. Zeng, Y. Huang, B. Liu, D. Fu, and J. Fu, "Seeing out of the box: End-to-end pre-training for vision-language representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 12976–12985.
[25] J. Li, R. Selvaraju, A. Gotmare, S. Joty, C. Xiong, and S. C. H. Hoi, "Align before fuse: Vision and language representation learning with momentum distillation," *Adv. Neural Inf. Process. Syst.*, vol. 34, pp. 9694–9705, 2021.
[26] W. Kim, B. Son, and I. Kim, "Vilt: Vision-and-language transformer without convolution or region supervision," in *Proc. Int. Conf. Mach. learning. PMLR*, 2021, pp. 5583–5594.
[27] L. Yuan et al., "Florence: A new foundation model for computer vision," 2021, *arXiv:2111.11432*.
[28] M. Cheng et al., "Vista: Vision and scene text aggregation for cross-modal retrieval," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 5184–5193.
[29] Z. Li, H. Lu, H. Fu, Z. Wang, and G. Gu, "Adaptive adversarial learning based cross-modal retrieval," *Eng. Appl. Artif. Intell.*, vol. 123, 2023, Art. no. 106439.
[30] Z. Li, H. Lu, H. Fu, and G. Gu, "Parallel learned generative adversarial network with multi-path subspaces for cross-modal retrieval," *Inf. Sci.*, vol. 620, pp. 84–104, 2023.
[31] Y. Zeng, X. Zhang, H. Li, J. Wang, J. Zhang, and W. Zhou, "X2-VLM: All-in-one pre-trained model for vision-language tasks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 5, pp. 3156–3168, May 2024.
[32] D. Kim, N. Kim, and S. Kwak, "Improving cross-modal retrieval with set of diverse embeddings," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 23422–23431.
[33] Q. Cheng, Y. Zhou, P. Fu, Y. Xu, and L. Zhang, "A deep semantic alignment network for the cross-modal image-text retrieval in remote sensing," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 4284–4297, Apr. 2021.
[34] G. Mikriukov, M. Ravanbakhsh, and B. Demir, "Unsupervised contrastive hashing for cross-modal retrieval in remote sensing," in *Proc. ICASSP 2022-2022 IEEE Int. Conf. Acoust., Speech Signal Process.*, 2022, pp. 4463–4467.
[35] Z. Yuan et al., "Mcrn: A multi-source cross-modal retrieval network for remote sensing," *Int. J. Appl. Earth Observ. Geoinformation*, vol. 115, 2022, Art. no. 103071.
[36] N. Kieu, K. Nguyen, A. Nazib, T. Fernando, C. Fookes, and S. Sridharan, "Multimodal co-learning meets remote sensing: Taxonomy, state of the art, and future works," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 17, pp. 7386–7409, Mar. 2024.
[37] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 10012–10022.
[38] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," in *Proc. NIPS 2014 Workshop Deep Learn.*, Dec. 2014, vol. 2014.
[39] J. D. M.-W. C. Kenton and L. K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL-HLT*, 2019, pp. 4171–4186.
[40] H. Diao, Y. Zhang, L. Ma, and H. Lu, "Similarity reasoning and filtration for image-text matching," in *Proc. AAAI Conf. Artif. Intell.*, 2021, pp. 1218–1226.
[41] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *Proc. 2006 IEEE Comput. Soc. Conf. Comput. Vis. pattern Recognit. (CVPR'06)*, 2006, vol. 2, pp. 1735–1742.
[42] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *Proc. IEEE Conf. Comput. Vis. pattern Recognit.*, 2015, pp. 3128–3137.
[43] G. Li et al., "Shapenet: A shapelet-neural network approach for multivariate time series classification," in *Proc. AAAI Conf. Artif. Intell.*, 2021, pp. 8375–8383.
[44] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, "Curriculum learning," in *Proc. 26th Annu. Int. Conf. Mach. Learn.*, 2009, pp. 41–48.
[45] J. L. Elman, "Learning and development in neural networks: The importance of starting small," *Cognition*, vol. 48, no. 1, pp. 71–99, 1993.
[46] X. Wang, Y. Chen, and W. Zhu, "A survey on curriculum learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 9, pp. 4555–4576, Sep. 2021.
[47] G. Hacohen and D. Weinshall, "On the power of curriculum learning in training deep networks," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 2535–2544.
[48] Ü. Dogan, A. A. Deshmukh, M. B. Machura, and C. Igel, "Label-similarity curriculum learning," in *Proc. 2020 16th Euro. Conf. Comput. Vision–ECCV*, Glasgow, U.K., Aug. 23–28, 2020, Proceedings, Part XXIX 16, 2020, pp. 174–190.
[49] B. Zhang et al., "Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling," *Adv. Neural Inf. Process. Syst.*, vol. 34, pp. 18408–18419, 2021.
[50] C. Sakaridis, D. Dai, S. Hecker, and L. Van Gool, "Model adaptation with synthetic and real data for semantic dense foggy scene understanding," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 687–704.
[51] C. Sakaridis, D. Dai, and L. V. Gool, "Guided curriculum model adaptation and uncertainty-aware evaluation for semantic nighttime image segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 7374–7383.
[52] W. Qin, Z. Hu, X. Liu, W. Fu, J. He, and R. Hong, "The balanced loss curriculum learning," *IEEE Access*, vol. 8, pp. 25990–26001, 2020.
[53] X. Yao, X. Feng, J. Han, G. Cheng, and L. Guo, "Automatic weakly supervised object detection from high spatial resolution remote sensing images via dynamic curriculum learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 1, pp. 675–685, Jan. 2020.
[54] Y. Huang et al., "Curricularface: Adaptive curriculum learning loss for deep face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 5901–5910.
[55] G.-S. Xia et al., "Aid: A benchmark data set for performance evaluation of aerial scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 7, pp. 3965–3981, Jul. 2017.
[56] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, 2017.
[57] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," 2017, *arXiv:1711.05101*.
[58] F. Yao et al., "Hypergraph-enhanced textual-visual matching network for cross-modal remote sensing image retrieval via dynamic hypergraph learning," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 688–701, Dec. 2022.
[59] L. Mi, S. Li, C. Chappuis, and D. Tuia, "Knowledge-aware cross-modal text-image retrieval for remote sensing images," in *Proc. 2nd Workshop Complex Data Challenges Earth Observation* (CDCEO 2022), 2022.
[60] Q. Ma, J. Pan, and C. Bai, "Direction-oriented visual-semantic embedding model for remote sensing image-text retrieval," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, Apr. 2024, Art. no. 4704014.
[61] G. Hu, Z. Wen, Y. Lv, J. Zhang, and Q. Wu, "Global-local information soft-alignment for cross-modal remote-sensing image-text retrieval," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, May 2024, Art. no. 5623915.
[62] U. A. Bhatti, H. Tang, G. Wu, S. Marjan, and A. Hussain, "Deep learning with graph convolutional networks: An overview and latest applications in computational intelligence," *Int. J. Intell. Syst.*, vol. 2023, pp. 1–28, 2023.
[63] Z. Yuan et al., "Learning to evaluate performance of multi-modal semantic localization," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–88, 2022.
[64] M. Yu et al., "Selo v2: Toward for higher and faster semantic localization," *IEEE Geosci. Remote Sens. Lett.*, vol. 20, Jun. 2023, Art. no. 6007005.

Zihui Zhou received the B.E. degree in computer science and technology in 2023 from the College of Computer Science, Chongqing University, Chongqing, China, where she is currently working toward the master's degree in computer science and technology.

Her main research interests include remote sensing cross-modal retrieval and remote sensing image processing.

Yong Feng received the Ph.D. degree in computer software and theory from the College of Computer Science, Chongqing University, Chongqing, China, in 2006.

Currently, he is a Professor with the College of Computer Science, Chongqing University. His research interest include multimodal artificial intelligence, cross-modal computer vision, and spatiotemporal big data intelligence.

Agen Qiu received the Ph.D. degree in cartography and geographic information engineering from Wuhan University, Wuhan, China, in 2017.

His research interests include spatially assisted decision making, spatiotemporal big data analysis, mining, the research and development of the geowindow spatial-temporal big data platform, and natural resources big data supercomputing cloud platform.

Guofan Duan received the master's degree in civil engineering from the College of Civil Engineering, Chongqing University, Chongqing, China, in 2014.

Currently, he is an Associate Professor with the Chongqing Metropolitan College of Science and Technology, Chongqing. His research interest include big data analysis, image processing, and artificial intelligence.

Mingliang Zhou received the Ph.D. degree in computer science and technology from Beihang University, Beijing, China, in 2017.

He is currently an Associate Professor with the College of Computer Science, Chongqing University, Chongqing, China. His research interests include image and video coding, perceptual image processing, multimedia signal processing, rate control, multimedia communication, machine learning, and optimization.