Reducing Semantic Confusion: Scene-aware Aggregation Network for Remote Sensing Cross-modal Retrieval

Jiancheng Pan
Zhejiang University of Technology
jianchengpan@zjut.edu.cn

Qing Ma*
Zhejiang University of Technology
maqing@zjut.edu.cn

Cong Bai
Zhejiang University of Technology
congbai@zjut.edu.cn

ABSTRACT

Recently, remote sensing cross-modal retrieval has received incredible attention from researchers. However, the unique nature of remote-sensing images leads to many semantic confusion zones in the semantic space, which greatly affects retrieval performance. We propose a novel scene-aware aggregation network (SWAN) to reduce semantic confusion by improving scene perception capability. In visual representation, a visual multiscale fusion module (VMSF) is presented to fuse visual features with different scales as a visual representation backbone. Meanwhile, a scene fine-grained sensing module (SFGS) is proposed to establish the associations of salient features at different granularity. A scene-aware visual aggregation representation is formed by the visual information generated by these two modules. In textual representation, a textual coarse-grained enhancement module (TCGE) is designed to enhance the semantics of text and to align visual information. Furthermore, as the diversity and differentiation of remote sensing scenes weaken the understanding of scenes, a new metric, namely, scene recall is proposed to measure the perception of scenes by evaluating scene-level retrieval performance, which can also verify the effectiveness of our approach in reducing semantic confusion. By performance comparisons, ablation studies and visualization analysis, we validated the effectiveness and superiority of our approach on two datasets, RSICD and RSITMD. The source code is available at https://github.com/kinshingpoon/SWAN-pytorch.

CCS CONCEPTS

• **Information systems** → **Information retrieval**; **Specialized information retrieval**; **Multimedia and multimodal retrieval**;

KEYWORDS

Cross-Modal Retrieval, Remote Sensing, Scene Perception

ACM Reference Format:
Jiancheng Pan, Qing Ma, and Cong Bai. 2023. Reducing Semantic Confusion: Scene-aware Aggregation Network for Remote Sensing Cross-modal Retrieval. In *ICMR '23: International Conference on Multimedia Retrieval, June 12–15, 2023, Thessaloniki, Greece.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3591106.3592236

*Corresponding authors

ICMR '23, June 12–15, 2023, Thessaloniki, Greece

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/10.1145/3591106.3592236

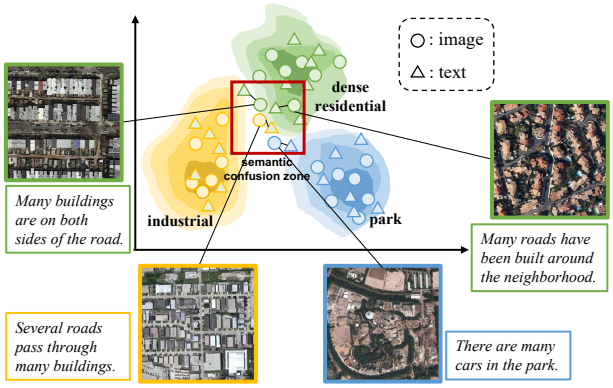


Figure 1: The distribution of image and text embeddings in the semantic space. The red box indicates the semantic confusion zone; the yellow boxes indicate image-text pairs of the industrial scene; the blue boxes indicate image-text pairs of the park scene; the green boxes indicate image-text pairs of the dense residential scene.

1 INTRODUCTION

Cross-modal retrieval is an important retrieval task that emerged in recent years, which aims to automatically mine valuable data of a given query from heterogeneous modalities. In remote sensing, cross-modal retrieval has recently attracted increasing attention, which has been widely used in disaster monitoring [7], land resource acquisition, and remote sensing image captioning [5].

Many efforts have been dedicated to remote sensing cross-modal retrieval over the past few years. Some early works [1, 17] directly mapped different modalities to the semantic space to achieve modality alignment. After which, Lv et al. and Cheng et al. [6, 16] started to use cross-modal interactions to enhance the semantic association between modalities. Although these approaches effectively improve the performance of remote sensing cross-modal retrieval, they ignore the substantial redundancy in the modalities. To reduce the impact of redundancy, Yuan et al. [29] proposed a multiscale visual self-attention module that fuses the visual features at different scales to filter redundancy. Might et al. [18] proposed a knowledge graph-based textual enhancement method to improve textual representation. To better represent the relationship between different objects, Yuan et al. [32] further designed a Multi-Level Information Dynamic Fusion module to fuse global and local visual features. This approach can connect objects but requires additional knowledge. The above works have contributed prominently to cross-modal retrieval in remote sensing. However, most of these approaches discard the fine-grained semantic information from visual features,

and they focus mainly on modality alignment with little attention to the semantic confusion caused by scene differences.

Unlike traditional cross-modal retrieval, remote sensing cross-modal retrieval is based on remote sensing images, with two main characteristics: small image foreground ratio and inter-class similarity. The small ratio of image foreground makes the semantic objects in remote sensing images insignificant, which is not conducive to understanding image semantics. There are diverse and differentiated scenes in remote sensing data. Inter-class similarity refers to the fact that similarities exist between different scene images, which affects the retrieval between modalities. As shown in Figure 1, there are remote sensing images belonging to dense residential scene, industrial scene and park scene respectively, but they have similar semantic contents. This characteristic easily lead to many semantic confusion zones in the final semantic space. In this semantic confusion zone, the images of different scenes have similar and confusing sentence descriptions, causing degradation of retrieval performance. Improving the scene perception capability can make the modalities of the same scene closer together and make the semantic confusion zones less in the semantic space.

To reduce these semantic confusion, we propose a **S**cene-a**W**are **A**ggregation **N**etwork (SWAN) to improve the fine-grained perception of the scene. Concretely, a Visual Multi-Scale Fusion module (VMSF) extracts the global semantic features with different scales to improve global scene perception. A Visual Fine-Grained Sensing module (SFGS) establishes associations between the semantic features of different granularity to get fine-grained local features. We aggregate global semantic and fine-grained local features to obtain the scene-aware visual representation. In addition, a Textual Coarse-Grained Enhancement module (TCGE) is designed to enhance text semantics and to align scene-aware visual information. The stronger scene perception capability, the better scene-level retrieval performance. A new metric, namely scene recall, is proposed to measure the perception of scenes by evaluating the scene-level retrieval performance, which can also verify our approach's effectiveness in reducing semantic confusion. We have conducted extensive experiments on RSICD [15] and RSITMD [29] datasets. The experimental results demonstrate the validity and superiority of our method compared to the state-of-the-art methods.

The main contributions of our work are as follows:

- We propose a novel scene-aware remote sensing cross-modal retrieval network SWAN to reduce semantic confusion by improving the fine-grained perception of the scene.
- VMSF module serves as a visual representation backbone to acquire global visual features, and the SFGS module establishes the relationship of salient features. A scene-aware visual representation is obtained by aggregating these two visual features. TCGE module is designed to enhance text representation and to align visual information.
- Scene recall is proposed to measure the perception of scenes, which can also verify our approach's effectiveness in reducing semantic confusion. Experiments have demonstrated that our method outperforms state-of-the-art methods on RSICD and RSITMD datasets.

2 RELATED WORK

2.1 Remote Sensing Cross-Modal Retrieval

Although traditional cross-modal retrieval methods based on natural images are increasingly mature, these methods cannot be directly applied to multiscale and redundant remote sensing images. There is still a long way to go in remote sensing cross-modal retrieval. According to different modal interaction methods [23], the remote sensing cross-modal retrieval methods can be divided into three categories: intra-modal interaction method, cross-modal interaction method, and intra-modal and cross-modal interaction method.

Intra-modal interaction method only performs self-interaction on the same modality. Generally, self-attention is used to obtain an enhanced modal semantic representation. In some works [1, 17], the information of different modalities was encoded and fused directly to obtain a joint semantic representation. In particular, some works [18, 32] introduced additional knowledge to enhance modality representation. Might et al. [18] proposed a knowledge-based method to solve the problem of coarse-grained textual descriptions. Yuan et al. [32] proposed a network structure based on global and local information, using attention-based modules to fuse multi-level information dynamically. Zhang et al. [33] designed a reconstruction module to reconstruct the decoupled features, which ensures the maximum retention of information in the features.

Cross-modal interaction method exchanges information between different modalities, which usually use information between different attention mechanisms or shared parameter networks for interactive learning. Lv et al. [16] propose a fusion-based association learning model to fuse image and text information across modalities to improve the semantic relevance of different modalities. Particularly, Yuan et al. [31] used a shared pattern transfer module to realize the interaction between modalities, solving the semantic heterogeneity problem between different modal data.

Intra-modal and cross-modal interaction method is an approach that combines the above two interaction methods. Cheng et al. [6] propose a semantic alignment module that uses an attention mechanism to enhance the correspondence between image and text. Yuan et al. [29] propose an asymmetric multimodal feature matching network to adapt multiscale feature inputs while using visual features to guide text representation.

2.2 Scene Recognition in Remote Sensing

Scene recognition is a basic work in the remote sensing field. Different from natural images, remote sensing images are more complex and diverse. Scene recognition facilitates instance-level understanding of semantic objects in remote-sensing images.

In the early days, some works mainly focused on solving remote sensing image classification problems with traditional methods. Penatti et al. [20] utilize some simple attributes in remote sensing images to distinguish different scenes, such as color, texture and contour. Nogueira et al. [19] utilize existing convolutional neural networks as feature extractors in different scenes and then directly perform classification. Aiming at the insensitivity of convolutional neural networks to the hierarchical structure and spatial information of entities, Zhang et al. [34] used capsule networks to replace traditional neural networks to encode the attributes and spatial information of features in remote sensing images. Later, researchers

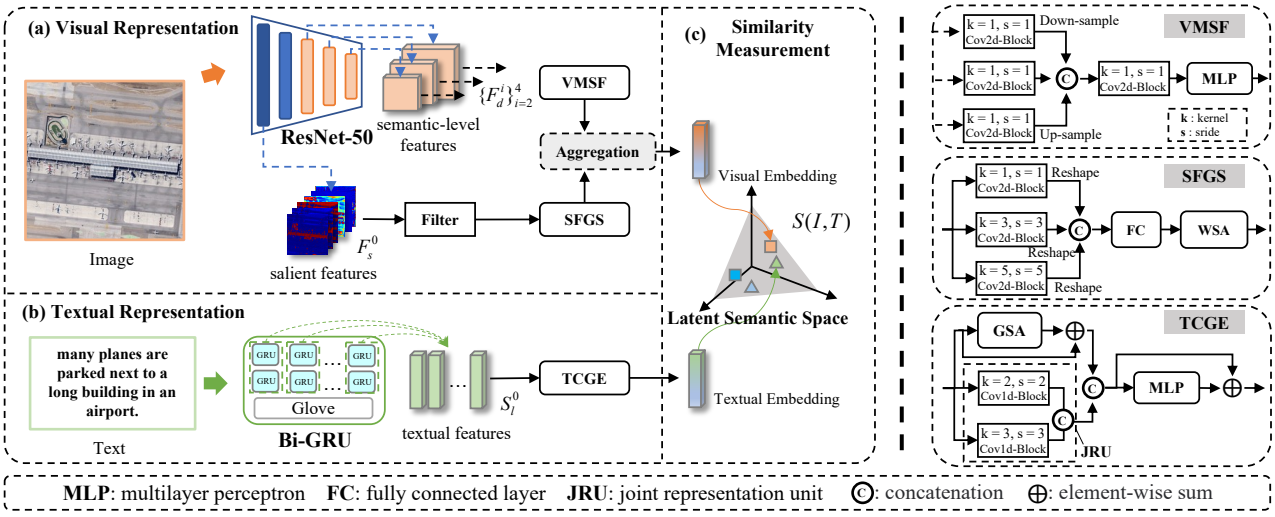


Figure 2: Schematic illustration of our proposed SWAN. It consists of three parts: (a) Visual Representation, (b) Textual Representation, and (c) Similarity Measurement.

began to pay more attention to extracting patch-level features of remote-sensing images. Xu et al. [28] introduced a model based on graph convolutional networks and used the patch-to-patch correlation of feature maps to obtain fine-grained visual features. Bi et al. [2] used the spatial attention weight matrix to describe the importance of each critical local region and obtain a more accurate local semantic representation. Chen et al. [4] proposed a scene classification method based on a multi-branch local attention network, which automatically suppresses the secondary features and extracts the critical information in the feature map. We take inspiration from these methods and propose a scene-aware cross-modal retrieval network to improve the scene perception capability, effectively reducing the semantic confusion zones in latent semantic space.

3 METHODOLOGY

This section will introduce our proposed method SWAN. We first introduce the self-attention of modality in Section 3.1, and then visual representation, textual representation and the loss function are described in detail in Section 3.2 , 3.3 and 3.4, which correspond to (a), (b), and (c) respectively in Figure 2.

3.1 Self-Attention Mechanism

Self-attention mechanism [26] is widely used in cross-modal retrieval. The use of self-attention mechanism can help the understanding of modalities.

Weighted Self-Attention (WSA). In the visual representation, we use the WSA module, an improvement on the general self-attention, for the fine-grained perception of visual features with different granularity. Figure 3(a) shows the structure of the WSA module, which consists of a multi-head attention module, a feedforward layer and a learnable weight matrix. Let $\boldsymbol{X} \in \mathbb{R}^{N\times d}$ denote the input of the WSA module through which we can make three copies of $\boldsymbol{Q} \in \mathbb{R}^{N\times d}$, $\boldsymbol{K} \in \mathbb{R}^{N\times d}$, and $\boldsymbol{V} \in \mathbb{R}^{N\times d}$, denoting query, key, and value, respectively. The multi-head attention layer can be expressed as

$$MultiHead(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = Concat\left(head_1,\ldots,head_h\right),$$
$$\text{where } head_i = softmax\left(\frac{\boldsymbol{Q}\boldsymbol{W}_i^Q(\boldsymbol{K}\boldsymbol{W}_i^K)^{\mathrm{T}}}{\sqrt{d}}\right)V\boldsymbol{W}_i^V, \tag{1}$$

where $\boldsymbol{W}_i^Q$, $\boldsymbol{W}_i^K$ and $\boldsymbol{W}_i^V$ are projection matrices, $softmax(\cdot)$ is the softmax function, and $Concat(\cdot)$ denotes the concatenation operation. Thus the overall module can be expressed as

$$\boldsymbol{X}' = LN(MutiHead(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) + \boldsymbol{X}), \tag{2}$$

$$\boldsymbol{X}'' = LN(\max\left(0, \boldsymbol{X}'\boldsymbol{W}_1 + \boldsymbol{b}_1\right)\boldsymbol{W}_2 + \boldsymbol{b}_2 + \boldsymbol{X}'), \tag{3}$$

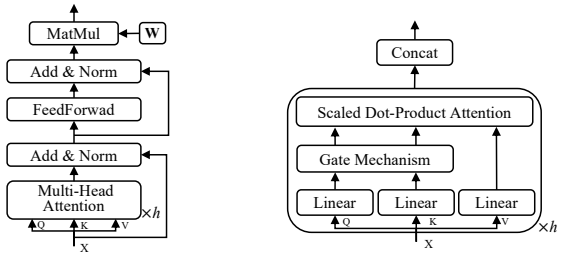
where $\boldsymbol{X}' \in \mathbb{R}^{N\times d}$ and $\boldsymbol{X}'' \in \mathbb{R}^{N\times d}$ respectively denote the output after the first and second *Add & Norm* layer in Figure 3(a), $LN(\cdot)$ means layer normlization, $\boldsymbol{W}_1(\boldsymbol{W}_2)$ and $\boldsymbol{b}_1(\boldsymbol{b}_2)$ are the weights and bias of the feedforward layer. To dynamically learn the importance of different features, we add a learnable weight matrix $\boldsymbol{W} \in \mathbb{R}^{N\times d}$ to get the final output as

$$\tilde{\boldsymbol{X}} = \boldsymbol{W}^{\mathrm{T}}\boldsymbol{X}'', \tag{4}$$

where $\tilde{\boldsymbol{X}} \in \mathbb{R}^{d\times d}$ represents the output of the WSA module.

Gating Self-Attention (GSA). The gating self-attention module was proposed to filter redundant information and enhance textual representation by Qu et al. [22]. We use the GSA module as a textual self-attention module to filter redundancy of the textual features, as shown in Figure 3(b). Similarly, we set the input of GSA module as $\boldsymbol{X} \in \mathbb{R}^{N\times d}$, and get $\boldsymbol{Q} \in \mathbb{R}^{N\times d}$, $\boldsymbol{K} \in \mathbb{R}^{N\times d}$, and $\boldsymbol{V} \in \mathbb{R}^{N\times d}$ copying from the input. Thus the GSA module can be expressed as

$$GSA(\boldsymbol{X}) = Concat\left(head_1,\ldots,head_h\right),$$
$$\text{where } head_i = \tag{5}$$
$$softmax\left(\frac{(\boldsymbol{M}_i^Q \odot (\boldsymbol{Q}\boldsymbol{W}_i^Q))(\boldsymbol{M}_i^K \odot (\boldsymbol{K}\boldsymbol{W}_i^K))^{\mathrm{T}}}{\sqrt{d}}\right)V\boldsymbol{W}_i^V,$$



(a) Weighted Self-Attention (b) Gating Self-Attention

Figure 3: Weighted Self-Attention and Gating Self-Attention.

where $\boldsymbol{W}_i^Q$, $\boldsymbol{W}_i^K$ and $\boldsymbol{W}_i^V \in \mathbb{R}^{d\times d}$ are projection matrices, $\boldsymbol{M}_i^Q \in \mathbb{R}^{N\times d}$ and $\boldsymbol{M}_i^K \in \mathbb{R}^{N\times d}$ mean the gating masks, and $\odot$ is the hadamard product. To calculate the gating masks, we first get the gated fusion value $\boldsymbol{G}_i$ of the Gate Mechanism as

$$\boldsymbol{G}_i = \left(\boldsymbol{Q}\boldsymbol{W}_i^Q\boldsymbol{W}_G^Q + \boldsymbol{b}_G^Q\right) \odot \left(\boldsymbol{K}\boldsymbol{W}_i^K\boldsymbol{W}_G^K + \boldsymbol{b}_G^K\right), \tag{6}$$

where $\boldsymbol{W}_G^Q(\boldsymbol{W}_G^K)$ and $\boldsymbol{b}_G^Q(\boldsymbol{b}_G^K)$ represent the weights and bias. Finally, the gating masks can be obtained as

$$\boldsymbol{M}_i^Q = \sigma_1\left(\boldsymbol{G}_i\boldsymbol{W}_M^Q + \boldsymbol{b}_M^Q\right), \tag{7}$$

$$\boldsymbol{M}_i^K = \sigma_1\left(\boldsymbol{G}_i\boldsymbol{W}_M^K + \boldsymbol{b}_M^K\right), \tag{8}$$

where $\boldsymbol{W}_M^Q(\boldsymbol{W}_M^K)$ and $\boldsymbol{b}_M^Q(\boldsymbol{b}_M^K)$ represent the weights and bias, and $\sigma_1(\cdot)$ denotes the sigmoid function.

3.2 Visual Representation

Image Feature Extraction. Like some of the previously mentioned remote sensing cross-modal retrieval methods [6, 18, 29, 32], we use ResNet as image feature extractors. A bit different is that we use ResNet pre-trained on AID dataset [27], which has a more detailed edge feature perception capability for remote sensing images than using the model pre-trained on ImageNet [9]. Shallow features of CNNs with small receptive fields contain fine-grained semantic information (texture, color, etc.). The use of this information can improve the fine-grained perception of the scene.

Given an input image $\boldsymbol{I} \in \mathbb{R}^{H\times W\times C}$, where (H, W) is the resolution of the original image, we can obtain object-aware and semantic-level features using the pre-trained image encoder. To get these features, we first utilize the salient features before $layer1$ of the image encoder, denoted as $\boldsymbol{F}_s^0 \in \mathbb{R}^{N_0\times H_0\times W_0}$, which represents the shallow feature of the image encoder. Similarly, we apply the semantic-level features from $layer2$, $layer3$, $layer4$, denoted as $\boldsymbol{F}_d^i \in \mathbb{R}^{N_i\times H_i\times W_i}(i = 2, 3, 4)$.

Visual Multi-Scale Fusion Module (VMSF). Traditional cross-modal retrieval methods [3, 11, 12] generally map visual features into target-sized feature vectors using a fully connected layer. To better represent global semantic features, we propose the VMSF module to fuse multi-scale semantic information, as shown in Figure 2. We use three 1×1 convolution layers to process the depth features of different sizes as

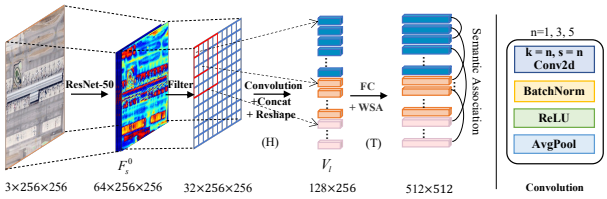


Figure 4: Schematic diagram of SFGS module.

$$\hat{\boldsymbol{F}}_d^i = Conv_{2d}^i(\boldsymbol{F}_d^i),\ i = 2, 3, 4, \tag{9}$$

where $Conv_{2d}^i(\cdot)$ denotes 2D convolution for the i-th deep feature. We sample these features of different sizes and integrate them as

$$\boldsymbol{V}_t = Concat(Downsample(\hat{\boldsymbol{F}}_d^2), \hat{\boldsymbol{F}}_d^3, Upsample(\hat{\boldsymbol{F}}_d^4)), \tag{10}$$

where $\boldsymbol{V}_t \in \mathbb{R}^{2048\times H_t\times W_t}$ indicates the multiscale visual features, $Downsample(\cdot)$ and $Upsample(\cdot)$ denote down-sampling and up-sampling operations. Immediately after, we process the features using a 1×1 convolution block in order to activate the critical information as

$$\boldsymbol{V}_h = \sigma_2\left(BN\left(Conv_{2d}(\boldsymbol{V}_t)\right)\right), \tag{11}$$

where $\boldsymbol{V}_h \in \mathbb{R}^{512\times H_c\times W_c}$ denotes the activated features after convolutional activation, $\sigma_2(\cdot)$ and $BN(\cdot)$ respectively represent the $ReLU$ function and batch normalization. Finally, we use the homogenization process for these features and enhance their nonlinear representation using multilayer perceptron (MLP) as

$$\boldsymbol{V}_d = MLP\left(Avg_{(H,W)}(\boldsymbol{V}_h)\right), \tag{12}$$

where $\boldsymbol{V}_d \in \mathbb{R}^{512}$ represents the global multiscale features, $Avg_{(H,W)}$ indicates averaging of the (H, W) dimension.

Scene Fine-Grained Sensing Module (SFGS). We find that the salient features of the shallow network contain not only the semantic objects feature but also the insignificant features. Therefore, we use a CNN-based filter (1×1 convolution block) to screen the salient features and get the filtered features $\boldsymbol{F}_s^{0\prime} \in \mathbb{R}^{N_0\times\frac{H_0}{2}\times W_0}$ as

$$\boldsymbol{F}_s^{0\prime} = \sigma_2\left(BN\left(Conv_{2d}\left(\boldsymbol{F}_s^0\right)\right)\right). \tag{13}$$

To reduce semantic confusion zones, we propose the SFGS module that can perceive salient objects at a fine-grained level. The SFGS module uses different sizes of convolution (1×1, 3×3, 5×5) to perform secondary feature extraction for these salient features, which can obtain semantic object features of different granularity, as shown in Figure 2. To facilitate the description, we divide the SFGS into two parts, the multiscale convolutional head (H) and the semantic associative attention tail (T), as shown in Figure 4. The H part can be expressed as follows:

$$\boldsymbol{\mathcal{H}}_l^j = Reshape\left(\sigma_2\left(BN\left(Conv_{2d}^j\left(\boldsymbol{F}_s^{0\prime}\right)\right)\right)\right),\ j = 1, 2, 3, \tag{14}$$

where $\boldsymbol{\mathcal{H}}_l^j \in \mathbb{R}^{L_j\times 256}$ denotes the fused local feature, $Reshape(\cdot)$ means to reshape (H, W) two dimensions into one dimension, $Conv_{2d}^j(\cdot)$ denotes the j-th convolution with different sizes. We directly merge salient features at different scales as

$$\boldsymbol{V}_l = Concat(\boldsymbol{\mathcal{H}}_l^1, \boldsymbol{\mathcal{H}}_l^2, \boldsymbol{\mathcal{H}}_l^3), \tag{15}$$

where $\boldsymbol{V}_l \in \mathbb{R}^{\sum_{i=1}^{3} L_j \times 256}$ represents the multi-granularity perception features. To enhance the perception of the scene, we also need to establish connections between them. We apply the WSA module to establish links between different granularity features, which can be expressed as

$$\boldsymbol{V}_s = Avg_{(1)}\left(WSA(\boldsymbol{V}_l \boldsymbol{W}_l + \boldsymbol{b}_l)\right), \tag{16}$$

where $\boldsymbol{V}_s \in \mathbb{R}^{512}$ represents the fine-grained local features, $\boldsymbol{W}_l \in \mathbb{R}^{256\times512}$ and $\boldsymbol{b}_l \in \mathbb{R}^{\sum_{i=1}^{3} L_j \times 512}$ mean the weight and bias of the fully connected layer, $Avg_{(1)}$ indicates averaging of the first dimension.

Visual Aggregation Representation. Visual aggregation representation can perceive fine-grained differences in different scenes. On the one hand, redundant information is filtered, and on the other hand, the perception of local features is enhanced. To obtain the scene-aware visual representation, we fuse $\boldsymbol{V}_d$ and $\boldsymbol{V}_s$ as

$$\boldsymbol{V}_p = MLP\left(Concat(\boldsymbol{V}_d, \boldsymbol{V}_s)\right), \tag{17}$$

where $\boldsymbol{V}_p \in \mathbb{R}^{512}$ represents the final visual embedding.

3.3 Textual Representation

Text Feature Extraction. To achieve image-text alignment, we extract text features from the sentences. We first obtain word vectors by pre-training Glove [21] and then use Bi-GRU [8] as the text encoder to learn the contextual relationships between words. Given a text input $\boldsymbol{T}$ to get the word vector $\{\boldsymbol{\tau}_1, \boldsymbol{\tau}_2, ..., \boldsymbol{\tau}_L\}$, then we embed into a vector by the Golve as $\boldsymbol{e}_i = \boldsymbol{W}_e \boldsymbol{\tau}_i$. Following it, we use the Bi-GRU to get the output of different hidden layers as follows:

$$\begin{aligned} \overrightarrow{\boldsymbol{h}_i} &= \overrightarrow{GRU}\left(\boldsymbol{e}_i, \overrightarrow{\boldsymbol{h}_{i-1}}\right), \\ \overleftarrow{\boldsymbol{h}_i} &= \overleftarrow{GRU}\left(\boldsymbol{e}_i, \overleftarrow{\boldsymbol{h}_{i+1}}\right). \end{aligned} \tag{18}$$

where $\overrightarrow{\boldsymbol{h}_i}$ and $\overleftarrow{\boldsymbol{h}_i}$ represent the output of i-th hidden layer. Finally, we can get the textual features as

$$\boldsymbol{h}_i = \frac{\overrightarrow{\boldsymbol{h}_i} + \overleftarrow{\boldsymbol{h}_i}}{2}, \tag{19}$$

where $\boldsymbol{h}_i$ denotes the mean value of the bi-directional output of the i-th layer. Thus we can get the initial textual features $\boldsymbol{S}_l^0 = [\boldsymbol{h}_1, \boldsymbol{h}_2, ..., \boldsymbol{h}_L]^{\mathrm{T}} \in \mathbb{R}^{L\times d}$.

Textual Coarse-Grained Enhancement Module (TCGE). We find that learning contextual relationships using the GRU model alone is insufficient for complex and variable remote sensing environments, where the differences in sentence descriptions are minor, resulting in the text not aligning the image well. The textual features obtained directly using the text encoder are semantically related to the word-level context. However, there is no strong correlation between adjacent words, which often makes it difficult for the model to detect subtle semantic differences. For example, "*an airport runway **extends** to a black ocean*" and "*an airport is **near** a large piece of black water*", as shown in Figure 5(a). The former emphasizes spatial extension, while the latter emphasizes spatial adjacency. We propose the TCGE module to improve the fine-grained textual representation. More subtle semantic differences can be perceived through the joint representation unit (JRU) as shown in Figure 5(b). The calculation formula is shown below:

$$\boldsymbol{S}_m^1 = \sigma_2\left(BN\left(Conv_{1d}^m\left(\boldsymbol{S}_l^0\right)\right)\right), \tag{20}$$

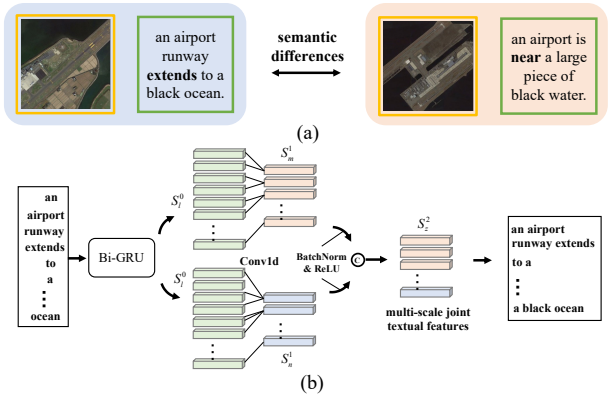


Figure 5: (a): Subtle semantic differences in remote-sensing image-text pairs. (b): The illustration of JRU module, which strengthen the connection between adjacent words, using two different sizes of 1-dimensional convolution.

$$\boldsymbol{S}_n^1 = \sigma_2\left(BN\left(Conv_{1d}^n\left(\boldsymbol{S}_l^0\right)\right)\right), \tag{21}$$

where $\boldsymbol{S}_m^1 \in \mathbb{R}^{M\times d}$ and $\boldsymbol{S}_n^1 \in \mathbb{R}^{N\times d}$ represents textual features obtained by two 1-dimensional convolution blocks, $Conv_{1d}^m(\cdot)$ and $Conv_{1d}^n(\cdot)$ denote the 1D convolution with different sizes. Finally, the multiscale joint features are obtained by channel concatenation as

$$\boldsymbol{S}_z^2 = Concat(\boldsymbol{S}_m^1, \boldsymbol{S}_n^1), \tag{22}$$

where $\boldsymbol{S}_z^2 \in \mathbb{R}^{(M+N)\times d}$ represents multiscale joint features. In addition, to establish deep semantic connections and focus on more important word-level textual features, we use the GSA module to adaptively filter significant textual features, which exploit the intra-modal interactions by calculating the dot-product similarities between queries and keys. The calculation is shown as follows:

$$\boldsymbol{S}_g^2 = GSA(\boldsymbol{S}_l^0) + \boldsymbol{S}_l^0, \tag{23}$$

where $\boldsymbol{S}_g^2 \in \mathbb{R}^{L\times d}$ represents significant textual feature, $GSA(\cdot)$ denotes the Gatting Self-Attention, as shown in Figure 3 (b). Finally, the enhanced textual features and multiscale joint textual features are fused as

$$\boldsymbol{S}_t^2 = Concat(\boldsymbol{S}_z^2, \boldsymbol{S}_g^2), \tag{24}$$

where $\boldsymbol{S}_t^2 \in \mathbb{R}^{(L+M+N)\times d}$ represents the fused textual features. In order to ensure that no important textual information is lost, residual concatenation is used to get the final textual embedding as

$$\boldsymbol{C}_t = Avg_{(1)}\left(MLP\left(\boldsymbol{S}_t^2\right) + \boldsymbol{S}_t^2\right), \tag{25}$$

where $\boldsymbol{C}_t \in \mathbb{R}^d$ denotes the final textual representation vector.

3.4 Loss Function

In remote sensing cross-modal retrieval, the bidirectional triplet ranking loss [12] is often used. We also apply this loss function to achieve image and text alignment. We compute the inner product of the final visual and textual embeddings to obtain the similarity of the image $\boldsymbol{I}$ and text $\boldsymbol{T}$ as

$$S(\boldsymbol{I}, \boldsymbol{T}) = \boldsymbol{V}_p(\boldsymbol{C}_t)^{\mathrm{T}}, \tag{26}$$

Table 1: Comparison results of the cross-modal retrieval on RSICD and RSITMD.

		RSICD dataset							RSITMD dataset						
		Image-query-Text			Text-query-Image				Image-query-Text			Text-query-Image			
Type	**Method**	**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**
Traditional Cross-modal Retrieval	VSE++	4.56	16.73	22.94	4.37	15.37	25.35	14.89	9.07	21.61	31.78	7.73	27.80	41.00	23.17
	SCAN i2t	4.82	13.66	21.99	3.93	15.20	25.53	14.19	8.92	22.12	33.78	7.43	25.71	39.03	22.83
	SCAN t2i	4.79	16.19	24.86	3.82	15.70	28.28	15.61	7.01	20.58	30.90	7.06	26.49	42.21	22.37
	CAMP	4.64	14.61	24.09	4.25	15.82	27.82	15.20	8.11	23.67	34.07	6.24	26.37	42.37	23.47
	CAMERA	4.57	13.08	21.77	4.00	15.93	26.97	14.39	8.33	21.83	33.11	7.52	26.19	40.72	22.95
Remote Sensing Cross-modal Retrieval	LW-MCR	3.29	12.52	19.93	4.66	17.51	30.02	14.66	10.18	28.98	39.82	7.79	30.18	49.78	27.79
	AMFMN	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	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
	KCR	5.95	18.59	29.58	5.40	**22.44**	37.36	19.89	-			-			-
	SWAN(Ours)	**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**

and get the ranking loss function as

$$\mathcal{L}_R = \sum_{\hat{T}} [\alpha - S(\boldsymbol{I}, \boldsymbol{T}) + S(\boldsymbol{I}, \hat{\boldsymbol{T}})]_+ + \sum_{\hat{I}} [\alpha - S(\boldsymbol{I}, \boldsymbol{T}) + S(\hat{\boldsymbol{I}}, \boldsymbol{T})]_+, \quad (27)$$

where α represents a margin parameter, and $[x]_+ \equiv \max(x, 0)$, $\hat{\boldsymbol{I}}$ and $\hat{\boldsymbol{T}}$ are the hardest negatives in mini-batch. The loss function consists of two parts; the first part is given a query image $\boldsymbol{I}$ that maintains the distance as far as possible from each hardest negative text $\hat{\boldsymbol{T}}$. Similarly, the second is given a query text $\boldsymbol{T}$ that controls the distance from each hardest negative image $\hat{\boldsymbol{I}}$ as far as possible.

4 EXPERIMENTS

4.1 Datasets

Two benchmark datasets, RSICD and RSITMD, are used in our experiments, in which each image with five sentences. RSICD [15] contains 10,921 images, and RSITMD [29] contains 4743 images with more fine-grained sentences than RSICD. We follow [29, 32] in dividing the two datasets, 7,862 training images, 1,966 validation images, and 1,093 test images on RSICD and 3,435 training images, 452 validation images, and 856 test images on RSITMD. The training and validation sets are not fixed in the comparison experiments, and the final experimental results are obtained by averaging three cross-validation experiments. However, in the ablation experiments, we fix the training and validation sets to reduce the impact of data distribution on model performance.

4.2 Metrics

Following the previous cross-modal retrieval methods [29, 32], we use $R@K (K = 1, 5, 10)$ and mR to evaluate the cross-modal retrieval methods. $R@K$ represents the percentage of ground-truth matched pairs among the top K ranked results. mR is the mean value of $R@K$, which denotes the overall retrieval performance.

The diversity and differentiation of remote sensing scenes make the understanding of scenes weaker. We propose Scene Recall (SR) to measure the perception of scenes by evaluating scene-level retrieval performance. The scene-level retrieval performance can quantitatively represent the proportion of retrieval results that belong to the same scenes as the query. $SR@K (K = 1, 5, 10)$ represents the proportion of retrieval results with the same scenes as the query to the top K retrieval results, which could be defined as

$$SR@K = \frac{\sum_{i=1}^{N} \frac{\sum_{j=1}^{K} S_i^j}{K}}{N}, \quad (28)$$

where N represents the number of queries, and S_i^j is a boolean function of the i-th query as

$$S_i^j = \begin{cases} 1, & \text{the } j\text{-th result is the same scene as the } i\text{-th query,} \\ 0, & \text{the } j\text{-th result is not the same scene as the } i\text{-th query.} \end{cases} \quad (29)$$

We also use this metric to verify the effectiveness of our approach in reducing semantic confusion. Larger $SR@K$ values indicate better scene perception and fewer semantic confusion regions.

4.3 Implementation Details

All our experiments are conducted on a single NVIDIA RTX A6000 GPU. For the hyperparameter settings, we fix the same random number seed as [24] to ensure the model is reproducible and set 50 and 70 epochs on RSICD and RSITMD, respectively. The Adam [13] is used as the model optimizer, and we set the initial learning rate to 0.0002 with decays of 0.7 every 20 epochs. we freeze the pre-trained ResNet and set the margin α in Eqn.(27) to 0.2. For the parameter of self-attention modules, we set the number of parallel processing layers of the WSA and GSA modules to 8 and 2, respectively.

4.4 Performance Comparisons

To verify the effectiveness of our proposed SWAN model, we compare it with the state-of-the-art methods on RSICD and RSITMD datasets. These methods can be divided into two groups: traditional cross-modal retrieval methods, VSE++ [11], SCAN [14], CAMP [30], and CAMERA [22], and remote sensing cross-modal retrieval methods, LW-MCR [30], AMFMN [29], GaLR [32], and KCR [18]. The former experimental results are obtained by averaging three sets of repeated experiments using the official source code provided. Same image and text encoders[1] as our model are used to extract initial features for experimental fairness. The latter directly cites the original paper for the best results. VSE++ adopts the bidirectional triplet ranking loss with hard negatives to align CNN-based visual features and textual features. SCAN, CAMP, and CAMERA employ the

[1]For SCAN, CAMP and CAMERA, we use ROI encoder [10] to extract regional features.

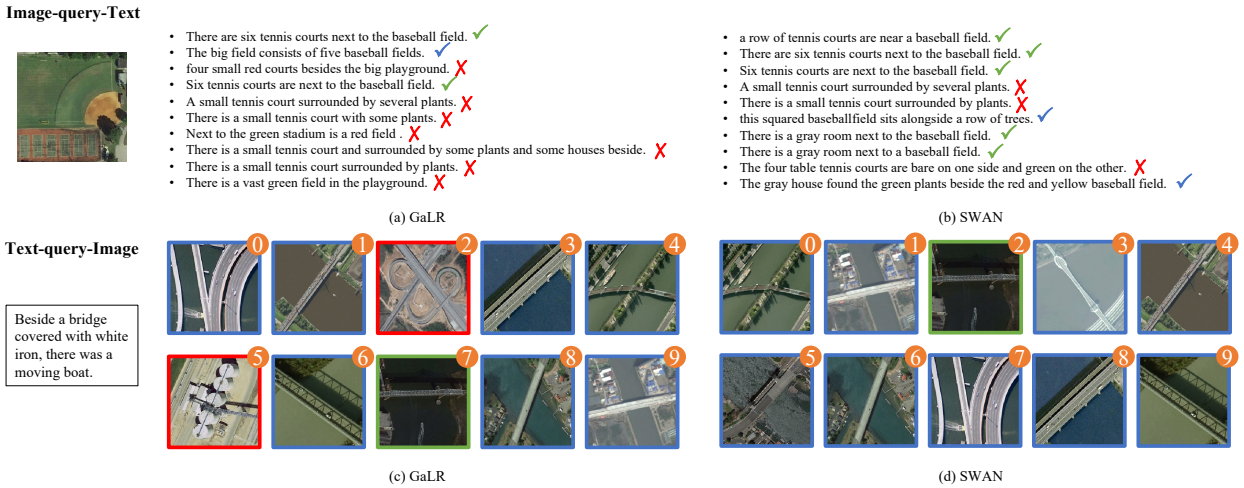


Figure 6: Qualitative results of bidirectional retrieval on RSITMD dataset. Green check marks and boxes indicate right retrieval results, blue check marks and boxes indicate retrieval results of the same scenes, and red crosses and boxes indicate retrieval results of the different scenes.

attention module to align salient visual features and text features. LW-MCR is a lightweight cross-modal retrieval model obtained by using knowledge distillation. AMFMN and GaLR enhance visual representations using attention-based methods to achieve image and text alignment. KCR utilizes knowledge graphs to enhance text representation and thus improve overall retrieval performance.

Table 2: Comparison results of the scene-level retrieval on RSITMD test set.

	Image-query-Text			Text-query-Image		
Method	SR@1	SR@5	SR@10	SR@1	SR@5	SR@10
VSE++	56.42	55.00	53.92	55.27	50.59	43.85
SCAN i2t	51.11	49.56	48.96	52.08	45.88	40.23
SCAN t2i	54.87	52.57	52.26	58.85	52.85	47.40
CAMP	56.19	54.60	53.52	52.70	48.22	44.04
CAMERA	55.31	52.96	51.62	58.10	52.31	46.75
AMFMN	73.45	71.42	70.38	70.44	64.67	57.70
GaLR	76.55	75.27	73.01	70.09	66.10	59.38
SWAN(Ours)	**82.08**	**81.68**	**79.31**	**74.65**	**70.63**	**64.92**

Comparison Results of Bidirectional Retrieval. Table 1 shows the comparison results on the RSICD and RSITMD datasets. We can find that the traditional cross-modal methods are lower in overall performance than the remote sensing cross-modal methods, which shows that a direct transfer of traditional methods to the remote sensing field is not feasible. Our approach has significant advantages compared to other state-of-the-art methods, especially the *R*@1 of image-query-text on the RSICD dataset is improved by 12.4% comparing to GaLR, and *mR* is improved by 3.6% and 8.7% comparing to the best competitor on the RSICD and RSITMD datasets, respectively. The above comparison can prove the effectiveness and superiority of our method. To be more intuitive, we compare the top 10 ranking retrieval results with the GaLR [32] method, which is the best method in open-source remote sensing cross-modal retrieval algorithms, as shown in Figure 6. It can be found that our method significantly outperforms the GaLR method in terms of bidirectional retrieval performance.

Comparison Results of Scene-level Retrieval. Table 2 demonstrates the comparison results of scene-level retrieval on the RSITMD dataset. Scene-level retrieval evaluates the scene perception capability, with higher values indicating stronger scene perception and fewer semantic confusion zones. The proposed SWAN outperforms the state-of-the-art methods in scene-level retrieval. *SR*@1 improves 7.2% and 6.0% on image-query-text and text-query-image comparing to the best competitor, respectively. From Figure 6, we can see that SWAN has significantly more occurrences of the same scene retrieval results (blue check marks and boxes) than GaLR, which shows that our method has a better scene perception than GaLR. The above comprehensive results show that our method has a significant advantage over state-of-the-art methods in scene-level retrieval and reducing the semantic confusion zones.

4.5 Ablation Studies

To better demonstrate the role of our different modules, we set up different ablation experiments under the assessment of two metrics on the RSITMD dataset. These experiments are divided into four main groups: 1) w/o VMSF indicates removing the VMSF module, 2) w/o SFGS indicates removing the SFGS module, 3) w/o TCGE indicates removing the TCGE module, and 4) w/o JRU indicates removing the JRU module.

Ablation Experiment with Bidirectional Retrieval. Table 3 demonstrates the results of the ablation experiments for the general bidirectional retrieval on the RSITMD dataset. Looking at the results of w/o VMSF, we can find that VMSF as a visual representation backbone can substantially improve the overall performance. Looking at the results of w/o SFGS, it can be concluded that the

SFGS module can significantly improve the performance of text-query-image retrieval, which is improved by 10.8% on *R*@1. We investigate the results of w/o TCGE and find that the TCGE module can significantly improve bi-directional retrieval performance, resulting in an overall performance improvement of 7.8%. Combining the results of w/o TCGE and w/o JRU, we could know that the JRU module mainly improves the performance of text-query-image retrieval. The above ablation experiments demonstrate the effectiveness of our approach, with different modules acting as different roles in the retrieval performance.

Ablation Experiment with Scene-level Retrieval. The ablation experiments' results for the different modules in scene-level retrieval are shown in Table 4. We can find that each module contributes to the improvement of scene perception. Among them, the VMSF module as the visual representation backbone improves the scene perception capability the most. In addition to VMSF, we can find that SFGS can more substantially improve the later ranking results of text-query-image *SR*@10. In conclusion, our method achieves optimal bidirectional retrieval while outperforming other state-of-the-art methods in scene perception.

Table 3: Ablation experiment results with different modules on RSITMD test set.

	Image-query-Text			Text-query-Image			
Method	R@1	R@5	R@10	R@1	R@5	R@10	mR
w/o VMSF	4.28	13.86	22.05	5.27	22.76	36.76	17.50
w/o SFGS	13.49	30.09	44.69	10.70	38.67	56.84	32.41
w/o TCGE	12.61	30.16	43.07	9.96	37.39	56.06	31.54
w/o JRU	13.35	32.23	44.25	10.72	38.39	58.44	32.90
SWAN(full)	**13.57**	**32.89**	**46.53**	**11.86**	**39.72**	**59.35**	**33.99**

Table 4: Ablation experiment results with different modules in scene-level retrieval on RSITMD test set.

	Image-query-Text			Text-query-Image		
Method	SR@1	SR@5	SR@10	SR@1	SR@5	SR@10
w/o VMSF	48.23	48.23	47.37	54.03	51.23	45.77
w/o SFGS	74.78	72.74	71.22	67.21	64.10	58.77
w/o TCGE	71.24	71.55	70.02	67.61	64.69	60.75
w/o JRU	72.79	73.27	70.73	66.90	64.13	59.51
SWAN(full)	**82.08**	**81.68**	**79.31**	**74.65**	**70.63**	**64.92**

4.6 Visualization Analysis

To visually demonstrate the role of our method in reducing the semantic confusion zones, we use t-SNE [25] to visualize the embedding space according to modal type and scene category, as shown in Figure 7. We still select GaLR [32] as a comparison and visualize it with the same experimental conditions. From the modality type visualization, the image and text embeddings are relatively dispersed under the action of SWAN compared to GaLR, which reflects that SWAN has a more fine-grained perceptual capability for modality. From scene category visualization, both SWAN and GaLR enable image and text embeddings to be distributed in clusters according

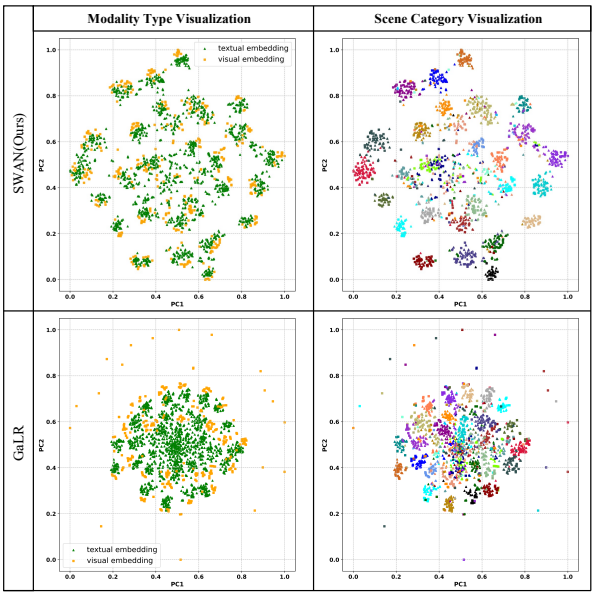


Figure 7: Visualization of latent embedding space. The results are visualized by modality type (left) and scene category (right). The different colors of the scene category visualization represent different categories of scenes.

to scene categories. Compared to GaLR, SWAN has fewer semantic confusion zones, which reflects the effectiveness of SWAN in reducing semantic confusion. However, there are still a few areas of semantic confusion under the SWAN method, which is an area that can be improved in the future. The visualization demonstrates the effectiveness of our method in reducing the semantic confusion region and the superiority of the scene perception capability.

5 CONCLUSION

In this paper, we propose the SWAN network to reduce semantic confusion in remote sensing cross modal retrieval. Firstly, the VMSF module is presented to fuse multiscale visual features to enhance the visual representation. Meanwhile, we propose the SFGS module to improve the fine-grained scene perception by establishing the relationship of multi-granularity visual features. A scene-aware visual representation is obtained by aggregating the visual information obtained from these two visual modules. To align visual modality, we propose a TCGE module to enhance the fine-grained textual representation. Finally, scene recall is designed to measure the perception of scenes. Visualization analysis shows that SWAN can optimize the distribution of embeddings in the latent semantic space to reduce semantic confusion zones. Extensive experiments demonstrate that our approach outperforms the state-of-the-art methods on the RSICD and RSITMD datasets.

ACKNOWLEDGMENTS

This work is partially supported by Natural Science Foundation of China under Grant No. 61976192, 62102365 and Zhejiang Provincial Natural Science Foundation of China under Grant No. LR21F020002.

REFERENCES

[1] Taghreed Abdullah, Yakoub Bazi, Mohamad M Al Rahhal, Mohamed L Mekhalfi, Lalitha Rangarajan, and Mansour Zuair. 2020. TextRS: Deep bidirectional triplet network for matching text to remote sensing images. *Remote Sensing* 12, 3 (2020), 405.

[2] Qi Bi, Kun Qin, Zhili Li, Han Zhang, Kai Xu, and Gui-Song Xia. 2020. A multiple-instance densely-connected ConvNet for aerial scene classification. *IEEE Transactions on Image Processing* 29 (2020), 4911–4926.

[3] Jianan Chen, Lu Zhang, Qiong Wang, Cong Bai, and Kidiyo Kpalma. 2022. Intra-Modal Constraint Loss for Image-Text Retrieval. In *2022 IEEE International Conference on Image Processing (ICIP)*. IEEE, 4023–4027.

[4] Si-Bao Chen, Qing-Song Wei, Wen-Zhong Wang, Jin Tang, Bin Luo, and Zu-Yuan Wang. 2021. Remote sensing scene classification via multi-branch local attention network. *IEEE Transactions on Image Processing* 31 (2021), 99–109.

[5] Qimin Cheng, Haiyan Huang, Yuan Xu, Yuzhuo Zhou, Huanying Li, and Zhongyuan Wang. 2022. NWPU-Captions dataset and MLCA-Net for remote sensing image captioning. *IEEE Transactions on Geoscience and Remote Sensing* 60 (2022), 1–19.

[6] Qimin Cheng, Yuzhuo Zhou, Peng Fu, Yuan Xu, and Liang Zhang. 2021. A deep semantic alignment network for the cross-modal image-text retrieval in remote sensing. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 14 (2021), 4284–4297.

[7] Mingmin Chi, Antonio Plaza, Jon Atli Benediktsson, Zhongyi Sun, Jinsheng Shen, and Yangyong Zhu. 2016. Big data for remote sensing: Challenges and opportunities. *Proc. IEEE* 104, 11 (2016), 2207–2219.

[8] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555* (2014).

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 248–255.

[10] Jian Ding, Nan Xue, Yang Long, Gui-Song Xia, and Qikai Lu. 2019. Learning RoI transformer for oriented object detection in aerial images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2849–2858.

[11] Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. 2017. Vse++: Improving visual-semantic embeddings with hard negatives. *arXiv preprint arXiv:1707.05612* (2017).

[12] Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3128–3137.

[13] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[14] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. 2018. Stacked cross attention for image-text matching. In *Proceedings of the European conference on computer vision (ECCV)*. 201–216.

[15] Xiaoqiang Lu, Binqiang Wang, Xiangtao Zheng, and Xuelong Li. 2017. Exploring models and data for remote sensing image caption generation. *IEEE Transactions on Geoscience and Remote Sensing* 56, 4 (2017), 2183–2195.

[16] Yafei Lv, Wei Xiong, Xiaohan Zhang, and Yaqi Cui. 2021. Fusion-based correlation learning model for cross-modal remote sensing image retrieval. *IEEE Geoscience and Remote Sensing Letters* 19 (2021), 1–5.

[17] Guo Mao, Yuan Yuan, and Lu Xiaoqiang. 2018. Deep cross-modal retrieval for remote sensing image and audio. In *2018 10th IAPR Workshop on Pattern Recognition in Remote Sensing (PRRS)*. IEEE, 1–7.

[18] Li Mi, Siran Li, Christel Chappuis, and Devis Tuia. 2022. Knowledge-Aware Cross-Modal Text-Image Retrieval for Remote Sensing Images. (2022).

[19] Keiller Nogueira, Otávio AB Penatti, and Jefersson A Dos Santos. 2017. Towards better exploiting convolutional neural networks for remote sensing scene classification. *Pattern Recognition* 61 (2017), 539–556.

[20] Otávio AB Penatti, Keiller Nogueira, and Jefersson A Dos Santos. 2015. Do deep features generalize from everyday objects to remote sensing and aerial scenes domains?. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*. 44–51.

[21] Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 1532–1543.

[22] Leigang Qu, Meng Liu, Da Cao, Liqiang Nie, and Qi Tian. 2020. Context-aware multi-view summarization network for image-text matching. In *Proceedings of the 28th ACM International Conference on Multimedia*. 1047–1055.

[23] Leigang Qu, Meng Liu, Jianlong Wu, Zan Gao, and Liqiang Nie. 2021. Dynamic modality interaction modeling for image-text retrieval. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1104–1113.

[24] Jun Rao, Fei Wang, Liang Ding, Shuhan Qi, Yibing Zhan, Weifeng Liu, and Dacheng Tao. 2022. Where Does the Performance Improvement Come From?-A Reproducibility Concern about Image-Text Retrieval. *arXiv preprint arXiv:2203.03853* (2022).

[25] Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of machine learning research* 9, 11 (2008).

[26] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

[27] Gui-Song Xia, Jingwen Hu, Fan Hu, Baoguang Shi, Xiang Bai, Yanfei Zhong, Liangpei Zhang, and Xiaoqiang Lu. 2017. AID: A benchmark data set for performance evaluation of aerial scene classification. *IEEE Transactions on Geoscience and Remote Sensing* 55, 7 (2017), 3965–3981.

[28] Kejie Xu, Hong Huang, Peifang Deng, and Yuan Li. 2021. Deep feature aggregation framework driven by graph convolutional network for scene classification in remote sensing. *IEEE Transactions on Neural Networks and Learning Systems* (2021).

[29] Zhiqiang Yuan, Wenkai Zhang, Kun Fu, Xuan Li, Chubo Deng, Hongqi Wang, and Xian Sun. 2022. Exploring a fine-grained multiscale method for cross-modal remote sensing image retrieval. *arXiv preprint arXiv:2204.09868* (2022).

[30] Zhiqiang Yuan, Wenkai Zhang, Xuee Rong, Xuan Li, Jialiang Chen, Hongqi Wang, Kun Fu, and Xian Sun. 2021. A lightweight multi-scale crossmodal text-image retrieval method in remote sensing. *IEEE Transactions on Geoscience and Remote Sensing* 60 (2021), 1–19.

[31] Zhiqiang Yuan, Wenkai Zhang, Changyuan Tian, Yongqiang Mao, Ruixue Zhou, Hongqi Wang, Kun Fu, and Xian Sun. 2022. MCRN: A Multi-source Cross-modal Retrieval Network for remote sensing. *International Journal of Applied Earth Observation and Geoinformation* 115 (2022), 103071.

[32] Zhiqiang Yuan, Wenkai Zhang, Changyuan Tian, Xuee Rong, Zhengyuan Zhang, Hongqi Wang, Kun Fu, and Xian Sun. 2022. Remote Sensing Cross-Modal Text-Image Retrieval Based on Global and Local Information. *IEEE Transactions on Geoscience and Remote Sensing* 60 (2022), 1–16.

[33] Huan Zhang, Yingzhi Sun, Yu Liao, SiYuan Xu, Rui Yang, Shuang Wang, Biao Hou, and Licheng Jiao. 2022. A Transformer-Based Cross-Modal Image-Text Retrieval Method using Feature Decoupling and Reconstruction. In *IGARSS 2022-2022 IEEE International Geoscience and Remote Sensing Symposium*. IEEE, 1796–1799.

[34] Wei Zhang, Ping Tang, and Lijun Zhao. 2019. Remote sensing image scene classification using CNN-CapsNet. *Remote Sensing* 11, 5 (2019), 494.