Efficient Reconstruction of Spatial Features for Remote Sensing Image-Text Retrieval

ZHANG Weihang[1,2,3], *CHEN Jialiang*[1,2*], *ZHANG Wenkai*[1,2], *LI Xinming*[4], *GAO Xin*[1,3], *SUN Xian*[1,2,3]

1. Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100190, P. R. China;
2. Key Laboratory of Target Cognition and Application Technology (TCAT), Chinese Academy of Sciences, Beijing 100190, P. R. China;
3. School of Electronic, Electrical and Communication Engineering, University of Chinese Academy of Sciences, Beijing 100190, P. R. China;
4. School of Computer Science and Artificial Intelligence, Aerospace Information Technology University, Jinan 250299, P. R. China

(Received 23 April 2024; revised 18 November 2024; accepted 15 December 2024)

Abstract: Remote sensing cross-modal image-text retrieval (RSCIR) can flexibly and subjectively retrieve remote sensing images utilizing query text, which has received more researchers' attention recently. However, with the increasing volume of visual-language pre-training model parameters, direct transfer learning consumes a substantial amount of computational and storage resources. Moreover, recently proposed parameter-efficient transfer learning methods mainly focus on the reconstruction of channel features, ignoring the spatial features which are vital for modeling key entity relationships. To address these issues, we design an efficient transfer learning framework for RSCIR, which is based on spatial feature efficient reconstruction (SPER). A concise and efficient spatial adapter is introduced to enhance the extraction of spatial relationships. The spatial adapter is able to spatially reconstruct the features in the backbone with few parameters while incorporating the prior information from the channel dimension. We conduct quantitative and qualitative experiments on two different commonly used RSCIR datasets. Compared with traditional methods, our approach achieves an improvement of 3%—11% in sumR metric. Compared with methods finetuning all parameters, our proposed method only trains less than 1% of the parameters, while maintaining an overall performance of about 96%. The relevant code and files are released at https://github.com/AICyberTeam/SPER.

Key words: remote sensing cross-modal image-text retrieval (RSCIR) ; spatial features; channel features; contrastive learning; parameter effective transfer learning

CLC number: TP391.3 **Document code:** A **Article ID:** 1005-1120(2025)01-0101-11

0 Introduction

In recent years, the exponential growth of remote sensing (RS) data and progressive processing techniques have greatly expanded human perceptual capabilities and prospected for many applications, such as ecological monitoring, land planning, and disaster prediction[1-2]. However, it is still challenging to process and retrieve valuable RS data efficiently. Remote sensing cross-modal image-text retrieval (RSCIR) aims to retrieve RS images utilizing text that describes the content of the image. This content-based retrieval approach has gradually become a research hotspot in recent years[3].

*Corresponding author, E-mail address: chenjl@aircas.ac.cn.

How to cite this article: ZHANG Weihang, CHEN Jialiang, ZHANG Wenkai, et al. Efficient reconstruction of spatial features for remote sensing image-text retrieval[J]. Transactions of Nanjing University of Aeronautics and Astronautics, 2025, 42(1):101-111.

http://dx.doi.org/10.16356/j.1005-1120.2025.01.008

The current mainstream in RSCIR is the end-to-end retrieval method based on embedding vectors[4]. Specifically, in order to directly measure similarity, the end-to-end retrieval approach utilizes the powerful representation capability of neural networks to map data from different modalities into a common hypersphere space. The cross-modal features are aligned through contrastive learning.

According to the different ways of interacting between multimodal features, there are two main representative categories: Dual-stream and single-stream. Dual-stream methods refer to independently encoding multimodal features without interaction. Representative methods include VSE++[5], HVSA[6], etc. Single-stream methods involve the fusion and guidance of cross-modal features during the encoding process. Representative methods include AMFMN[7], SWAN[8], etc.

Meanwhile, the emergence of large-scale visual-language pre-training (VLP) models has provided new insights for RSCIR[9]. Recently, there has been rapid development in large-scale multimodal pre-training models, such as CLIP[10], ALBEF[11] and BLIP-2[12]. Instead of training all parameters in VLP, the parameter-efficient transfer learning method is designed to train a fraction of the parameters, which significantly reduce computational consumption while maintaining reliable performance[13].

Although RSCIR has had some achievements, it still confronts some challenges. Firstly, as for the RS domain, training a VLP model from scratch requires considerable computational resources and annotated data[14]. The initial CLIP method, for example, was trained on 400 million image-text pairs collected from the internet, which has been upgraded to 2 billion. Captions in the RS domain mostly rely on manual annotation by professionals thus it is quite challenging to train a VLP from scratch for the RS domain. Therefore, how to efficiently transfer the prior knowledge of the natural domain to the complex RS domains is worth further exploration.

Moreover, recently proposed parameter-efficient transfer learning methods mainly reconstruct features in the channel dimension by up-sampling and down-sampling[14-15]. This is because most of them tend to transfer to downstream tasks in the same domain as VLP[15]. However, there is an inherent domain gap between RS scenes and natural scenes. RS scenes are complex and targets can vary greatly in scale. Merely reconstructing channel features is insufficient to explore the spatial relationships of instances, making it suboptimal for image-text retrieval.

To address these issues, an efficient transfer learning framework for RSCIR is proposed, which is based on spatial feature efficient reconstruction (SPER). First, to enhance spatial relationship extraction and reduce computational consumption, we introduce a concise and efficient spatial adapter that reconstructs image-text features in the spatial dimension and integrates prior information from the channel dimension. By partitioning the cross-modal features in the channel dimension, we can obtain features that contain both spatial and a priori channel information. Differing from traditional methods, SPER reduces the volume of additional parameters introduced by the down-sampling and up-sampling processes. Then the proposed spatial adapters are inserted into the backbone of the VLP model. During training, SPER freezes the parameters of the backbone and only updates the parameters of the inserted spatial adapters. The main process of our method is illustrated in Fig.1, where SPA represents spatial adapter, LN the layer norm, MHA the multi-head attention, and FFN the feed-forward network. The contributions of this paper can be summarized as follows:

(1) Different from traditional methods based on fine-tuning all parameters, we propose an innovative and efficient transfer learning framework for RSCIR, which reduces the consumption of computational and storage resources.

(2) To bridge the gap between different domains, we design the spatial adapter to efficiently reconstruct multimodal features in the spatial dimension and achieve superior performance.

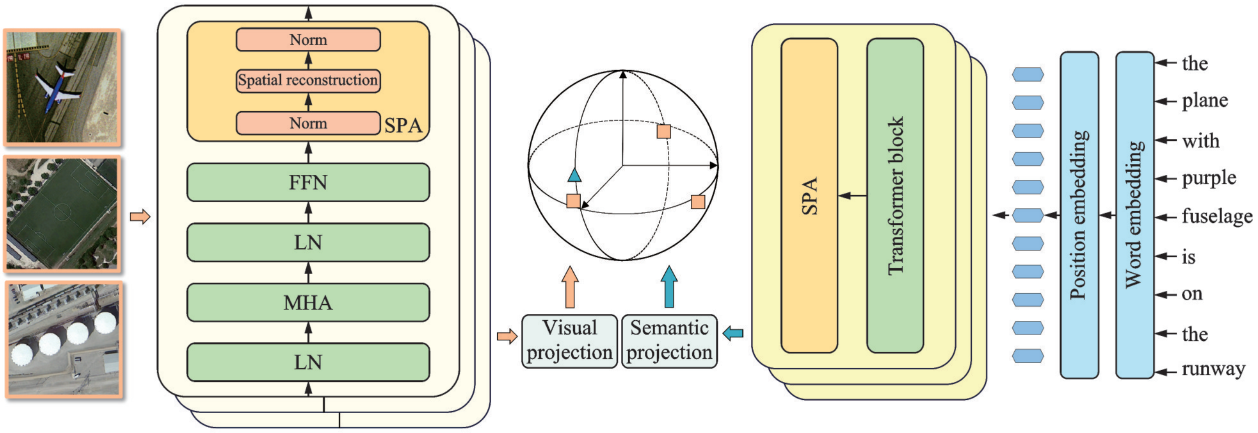


Fig.1 Pipeline of the proposed SPER

(3) We have conducted quantitative and qualitative experiments on different publicly available datasets, demonstrating the effectiveness of our approach.

1 Methods

1.1 Cross-modal feature representation

Consistent with the idea of contrastive learning[16], SPER constrains positive pairs to be as close as possible and negative pairs to be as far away as possible. The overall process is illustrated in Fig.1.

For simplicity, residual connections are ignored in Fig.1. We denote the RS image and query text as $\boldsymbol{I}\in\mathbf{R}^{H\times W\times 3}$ and $\boldsymbol{C}=\{\boldsymbol{w}_m\}_{m=0}^{M}$, respectively, where $H\times W$ is the size of the RS image and $\boldsymbol{w}_m$ is the mth word in the query text. For RSCIR, we first encode the RS image $\boldsymbol{I}$ and the corresponding caption $\boldsymbol{C}$ with a multimodal encoder to obtain the visual embedding vector $\boldsymbol{v}$ and the semantic embedding vector $\boldsymbol{s}$. Then, we map the visual embedding vector $\boldsymbol{v}\in\mathbf{R}^{d_v}$ and semantic embedding vector $\boldsymbol{s}\in\mathbf{R}^{d_s}$ to the common hypersphere space and measure the similarity $\mathcal{S}(\boldsymbol{I},\boldsymbol{C})$ by the inner product.

The vision transformers (ViTs)[17] are utilized to extract visual features initially. Specifically, we divide the RS image into $N\times N$ patches and add the class token as an aggregate representation of the image, which can be defined as

$$\hat{\boldsymbol{I}}=[\boldsymbol{I}_c,\boldsymbol{I}_0,\boldsymbol{I}_1,\cdots,\boldsymbol{I}_{N^2-1}]+\boldsymbol{I}_p \tag{1}$$

where $\hat{\boldsymbol{I}}\in\mathbf{R}^{(N^2+1)\times d_v}$ is the input to the ViT, $\boldsymbol{I}_c\in\mathbf{R}^{d_v}$ the class token for the image, $\boldsymbol{I}_n\in\mathbf{R}^{(H/N\cdot W/N)\times d_v}$ the nth image patch, and $\boldsymbol{I}_p\in\mathbf{R}^{(N^2+1)\times d_v}$ the position embedding added to each token. And one fundamental ViT block is modeled as follows

$$\boldsymbol{v}_h=\mathrm{SA}(\mathrm{Norm}(\hat{\boldsymbol{I}}))+\hat{\boldsymbol{I}} \tag{2}$$

$$\boldsymbol{v}_o=\mathrm{MLP}(\mathrm{Norm}(\boldsymbol{v}_h))+\boldsymbol{v}_h \tag{3}$$

where Norm(•) represents the layer normalization, SA(•) the self-attention module in ViT, and MLP(•) the multi-layer perceptron, $\boldsymbol{v}_h\in\mathbf{R}^{(N^2+1)\times d_v}$ the hidden feature obtained by SA(•), and $\boldsymbol{v}_o\in\mathbf{R}^{(N^2+1)\times d_v}$ the output feature of the ViT block.

Similar to the visual feature $\boldsymbol{v}$, the semantic feature $\boldsymbol{s}$ is extracted with BERT[18]. The query caption is first preprocessed and then a sequence of tokens $[\boldsymbol{c}_{\mathrm{bos}},\boldsymbol{c}_0,\boldsymbol{c}_1,\cdots,\boldsymbol{c}_M,\boldsymbol{c}_{\mathrm{eos}}]$ is obtained as

$$\boldsymbol{c}_m=\boldsymbol{w}_m\boldsymbol{M}_e+\boldsymbol{w}_p \tag{4}$$

$$\boldsymbol{s}_o=\mathrm{Trans}(\boldsymbol{c}_{\mathrm{bos}},\boldsymbol{c}_0,\boldsymbol{c}_1,\cdots,\boldsymbol{c}_M,\boldsymbol{c}_{\mathrm{eos}})\in\mathbf{R}^{L\times d_s} \tag{5}$$

where $\boldsymbol{w}_m\in\mathbf{R}^{|\mathbb{V}|}$ represents the mth words in the caption, and $|\mathbb{V}|$ the vocabulary size of the BERT; $\boldsymbol{w}_p\in\mathbf{R}^{d_s}$ is the positional embedding vector, $\boldsymbol{M}_e\in\mathbf{R}^{|\mathbb{V}|\times d_s}$ the word embedding matrix, $\boldsymbol{c}_{\mathrm{bos}}\in\mathbf{R}^{d_s}$ the beginning of sentence token, and $\boldsymbol{c}_{\mathrm{eos}}\in\mathbf{R}^{d_s}$ the end of the sentence token; L represents the length of tokens, and the semantic encoder is denoted by Trans(•).

To address the challenges presented above, the innovative aspects of our proposed method are: (1) Compared with recent efficient transfer learning methods, our approach enhances the extraction of spatial relationships in RS images. It leverages fewer parameters for the efficient reconstruction of multimodal features in the spatial dimension. (2) Compared with traditional RSCIR methods, our SPER framework is more concise and efficient. We only specify a limited number of parameters to be involved in backpropagation and updates. Further details are provided below.

1.2 Spatial adapter

Compared with natural scenes, RS scenes are characterized by greater complexity and variability in scale. Traditional parameter-efficient transfer learning methods[14-15] that rely solely on channel feature reconstruction are insufficient to capture the relationships between instances in RS images. These methods are designed to transfer prior knowledge to downstream tasks within the same domain as the pre-training task, ignoring transfer learning across domains, e.g., from the nature domain to the RS domain. Thus, they mostly focus on reconstructing channel features, as shown in Fig.2(a). They perform the reconstruction of features in the channel dimension by up-sampling and down-sampling, which can be expressed as

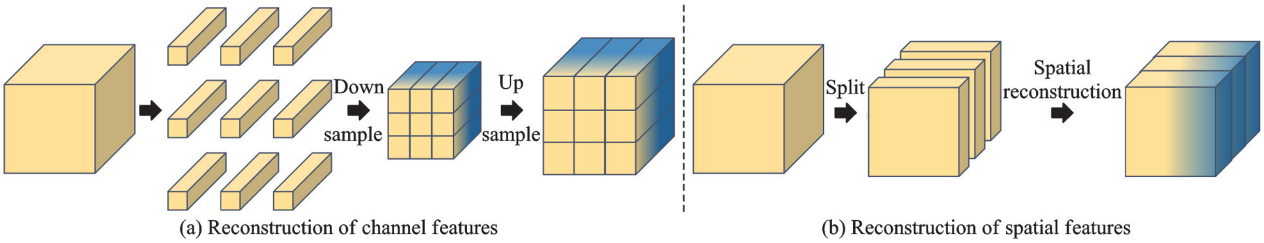
$$\tilde{\boldsymbol{v}}_{\mathrm{o}}=\phi(\boldsymbol{v}_{\mathrm{o}}\boldsymbol{W}_{\mathrm{down}})\boldsymbol{\cdot}\boldsymbol{W}_{\mathrm{up}} \tag{6}$$

where $\boldsymbol{v}_{\mathrm{o}}$ represents the original visual feature, $\phi(\bullet)$ the activation function, and $\tilde{\boldsymbol{v}}_{\mathrm{o}}\in\mathbf{R}^{(N^2+1)\times d_{\mathrm{v}}}$ the reconstructed feature; $\boldsymbol{W}_{\mathrm{down}}\in\mathbf{R}^{d_{\mathrm{v}}\times h}$ and $\boldsymbol{W}_{\mathrm{up}}\in\mathbf{R}^{h\times d_{\mathrm{v}}}$ represent the down-sampling matrix and the up-sampling matrix, respectively.

To address the issue mentioned above, we propose the SPA which can better handle the complexity of RS scenes and the scale variability of valuable targets. The details are shown as SPA in Fig.1.

Specifically, the process of spatial reconstruction is shown in Fig.2(b). In contrast to existing methods that primarily reconstruct channel features, our spatial adapter explicitly focuses on spatial feature reconstruction, a crucial aspect for handling RS images. It enhances the ability to model and extract spatial relationships while effectively incorporating prior channel information. We innovatively partition the visual feature $\boldsymbol{v}_{\mathrm{o}}$ in the channel dimension, obtaining a sequence of features containing spatial information as

$$\boldsymbol{v}_{\mathrm{o}}=[\boldsymbol{v}^0,\boldsymbol{v}^1,\cdots,\boldsymbol{v}^N] \tag{7}$$



(a) Reconstruction of channel features

(b) Reconstruction of spatial features

Fig.2 Comparison of feature reconstruction between traditional methods and the spatial adapter

Different from traditional methods, we do not simply employ down-sampling and up-sampling for feature reconstruction. We complete the reconstruction by applying cross-correlation between the features containing spatial information and the reconstruction matrix. By utilizing cross-correlation, our method efficiently captures spatial dependencies while reducing the need for excessive parameter overhead as

$$\boldsymbol{i}_m=r(\boldsymbol{v}_{\mathrm{o}})=b_m+\sum_{n=0}^{N-1}\boldsymbol{W}_{m,n}\otimes\boldsymbol{v}^n \tag{8}$$

where $\boldsymbol{v}^n\in\mathbf{R}^{H\times W\times(d_{\mathrm{v}}/N)}$ is the nth visual feature after partitioning, $r(\bullet)$ the spatial reconstruction function, $\otimes$ the valid cross-correlation operator, $\boldsymbol{W}\in\mathbf{R}^{H\times W\times(d_{\mathrm{v}}/N)}$ the reconstruction weight matrix in the spatial adapter, and b the bias parameter. $\boldsymbol{i}_m\in\mathbf{R}^{d_{\mathrm{v}}}$ is the visual feature obtained by the mth spatial reconstruction matrix.

In order to decrease the gap between different domains and reduce the difficulty of cross-modal alignment, we similarly perform global reconstruction for semantic features. Similar to the visual feature $\boldsymbol{v}_{\mathrm{o}}$, the semantic feature $\boldsymbol{s}_{\mathrm{o}}$ is first partitioned in the channel dimension to obtain the global feature sequence $[\boldsymbol{s}^0, \boldsymbol{s}^1, \cdots, \boldsymbol{s}^N]$. After efficient reconstruction, semantic features with global information are finally obtained, which can be expressed as

$$t_m = r(\boldsymbol{s}_{\mathrm{o}}) = b_m + \sum_{n=0}^{N-1} \boldsymbol{W}_{m,n} \otimes \boldsymbol{s}^n \tag{9}$$

where $\boldsymbol{s}^n \in \mathbf{R}^{L \times (d_s/N)}$ is the nth global semantic feature after partitioning, and $t_m \in \mathbf{R}^{d_s}$ the mth token after reconstruction.

Finally, we employ the class token from the last visual encoder as the visual embedding vector $\boldsymbol{v}$ and the begin of sentence (BOS) token from the last semantic encoder as the semantic embedding vector $\boldsymbol{s}$. They are mapped to the d-dimensional hypersphere space after L_2 normalization.

1.3 Efficient transfer learning

Moreover, to reduce the consumption of computational and storage resources, SPER freezes the parameters of the backbone and only updates the parameters of the proposed SPAs during transfer learning. Following the same procedure as previous methods[12], the ViT and BERT in SPER are initialized by the pre-training weights of CLIP and encode images and query text as backbone networks, respectively, shown as

$$\theta_{\mathrm{b}}^{(n+1)} = \theta_{\mathrm{b}}^{(n)} \tag{10}$$

$$\theta_{\mathrm{s}}^{(n+1)} = \theta_{\mathrm{s}}^{(n)} - \eta \frac{\delta L}{\delta \theta_{\mathrm{s}}^{(n)}} \tag{11}$$

where $\theta_{\mathrm{b}}^{(n)}$ denotes the parameter of backbone at the iteration n, $\theta_{\mathrm{s}}^{(n)}$ the parameter of spatial adapter at iteration n, η the step size of the parameter update, and $\frac{\delta L}{\delta \theta_{\mathrm{s}}^{(n)}}$ the derivative of the loss function L with respect to parameter $\theta_{\mathrm{s}}^{(n)}$.

Our objective is to restrict positive paired samples as close as possible and negative paired samples as far away as possible. Instead of the traditional triplet loss, we employ the InfoNCE loss in contrastive learning[13]. In a batch of training data containing N paired samples, the alignment loss of one positive pair $(\boldsymbol{I}, \boldsymbol{C})$ can be expressed as

$$\begin{aligned} L = -\frac{1}{2}\Bigg[& \log_2 \frac{\exp(\mathcal{S}(\boldsymbol{I}, \boldsymbol{C})/\tau)}{\sum_{n=1}^{N} \exp(\mathcal{S}(\boldsymbol{I}, \boldsymbol{C}_n)/\tau)} + \\ & \log_2 \frac{\exp(\mathcal{S}(\boldsymbol{I}, \boldsymbol{C})/\tau)}{\sum_{n=1}^{N} \exp(\mathcal{S}(\boldsymbol{I}_n, \boldsymbol{C})/\tau)} \Bigg] \end{aligned} \tag{12}$$

where $\mathcal{S}(\bullet)$ is the cosine similarity between image $\boldsymbol{I}$ and caption $\boldsymbol{C}$, and τ the temperature coefficient; $\boldsymbol{I}_n$ and $\boldsymbol{C}_n$ represent the nth image and caption in the current batch, respectively. During backpropagation, only the parameters of the proposed spatial adapter are updated.

2 Experimentation and Analysis

2.1 Experimental datasets

RSICD and RSITMD are two commonly used RSTIR datasets. The RSICD dataset contains 10 921 RS images of various resolutions and 54 605 query texts. The image size is 224 pixel × 224 pixel. The RSITMD dataset includes 4 743 images of different resolutions and 23 715 query texts. The image size is 256 × 256. The pixel resolution is about 0.5 m to 20 m.

The UCM Captions and Sydney datasets were not considered due to their small sample sizes and single resolutions, with only 2 100 and 613 samples, respectively, and resolutions of 0.3 m and 0.5 m.

2.2 Experimental implementation details

We conducted qualitative and quantitative experiments on the RSICD and RSITMD datasets. To ensure the fairness and reproducibility of the experiments, we follow the same dataset partitioning as in Ref.[13]. Two metrics, R@K (K = 1, 5, 10) and sumR, are used to evaluate retrieval performance quantitatively. The R@K metric denotes the percentage of ground truth in the first K recalled results. The sumR metric reflects the overall performance of the retrieval and can be calculated by Eq.(13). The optimization algorithm is AdamW.

The initial learning rate is set to 5×10^{-4}, with the linear warm-up strategy for the first four epochs, and a total of 20 epochs for training. The dimension of the multimodal features is 512. All experiments are conducted on one NVIDIA Telsa V100 GPU.

$$\mathrm{sumR}=\sum_{K\in[1,5,10]}\mathrm{R@}K \tag{13}$$

2.3 Retrieval performance comparison

We compare our approach with previous excellent methods, including traditional RSCIR methods and methods transferred from VLP (CLIP). Traditional methods include HVSA[6], AMFMN[7], SWAN[8], etc. Transfer learning-based methods include Full fine-tuning[10], Adapter[14], Cross-Modal Adapter[15], etc. To demonstrate the effectiveness of the proposed framework, we choose the Full fine-tuning method and the Adapter algorithm as our baselines. We also report the zero-shot capability of the CLIP model in RSCIR, with 0.00 million training parameters.

Table 1 demonstrates the experimental results, where "R" denotes traditional methods and "T" CLIP methods. If not specified, the architecture of the visual encoder always adopts ViT-B-32. With the exception of Singe Language and Full fine-tuning methods, the best results are bolded to provide a better illustration of comparisons between similar methods. Concretely, the Full fine-tuning method for direct transfer learning is based on the CLIP model, all 151.00 million parameters are involved in backpropagation and gradient updates. Methods like the Adapter[13], Cross-Modal Adapter[14], etc. are employed to efficiently transfer the pre-trained CLIP's prior knowledge to RSCIR.

Table 1 Comparison of cross-modal retrieval performance on RSICD

Type	Method	Training parameter/10^6	RSICD						
			Text retrieval			Image retrieval			sumR
			R@1	R@5	R@10	R@1	R@5	R@10	
R	LW-MCR-u[19]	1.65	4.39	13.35	20.29	4.30	18.85	32.34	93.52
	AMFMN-sim[7]	35.94	5.21	14.72	21.57	4.08	17.00	30.60	93.18
	MCRN[20]	52.35	6.59	19.40	30.28	5.03	19.38	32.99	113.67
	SWAN[8]	—	7.41	20.13	30.86	5.56	22.26	37.41	123.63
	GaLR with MR [21]	46.89	6.59	19.85	31.04	4.69	19.48	32.13	113.78
T	Single Language[22]	151.00	10.70	29.64	41.53	9.14	28.96	44.59	164.56
	Linear probe[10]	0.53	8.46	24.41	37.72	7.81	25.89	42.47	146.76
	RS-light[23]	9.20	6.67	18.92	28.42	8.94	26.45	41.06	130.46
	TGKT[24]	4.70	8.69	24.52	37.15	6.61	24.74	39.71	141.42
	Cross-Modal Adapter[15]	0.16	11.18	27.31	40.62	9.57	30.74	**48.36**	167.78
	Full fine-tuning[10]	151.00	13.54	30.83	43.46	11.55	33.14	49.83	182.35
	Adapter[14]	2.57	12.99	28.63	42.54	9.84	30.74	45.92	170.66
	CLIP(ViT-B-16)[10]	0.00	6.67	17.65	26.44	7.33	22.15	33.57	113.81
	Adapter (ViT-B-16)	2.57	14.36	31.65	44.46	11.60	**32.68**	48.32	183.07
	Ours	0.18	14.36	30.19	43.73	10.57	30.52	46.03	175.40
	Ours (ViT-B-16)	0.60	**16.01**	**33.57**	**46.11**	**11.82**	31.94	47.77	**187.22**

Firstly, compared with traditional methods on the RSICD, we have achieved a significant performance lead, which we believe is due to the powerful visual-semantic extraction capability of the pre-trained model. Additionally, compared with CLIP methods, our approach requires fewer training parameters and exhibits superior overall performance. Compared with the baseline method Adapter, SPER only needs to train 0.18 million parameters, while the Adapter method requires training 2.57 million parameters. Importantly, the sumR metric of SPER leads the Adapter method by 4.74 points on the RSICD dataset. In our opinion, the advancement of SPER lies in its ability to model and extract spatial relationships, whereas the Adapter method mainly focuses on channel features. Furthermore, compared to the Full fine-tuning method, SPER only needs to train less than 1% of the parameters to achieve 96% of its performance, demonstrating the efficiency of our proposed approach.

Our method also performs well on the RSITMD, as shown in Table 2. Compared with similar methods, our approach makes a better trade-off between the volume of training parameters and retrieval performance. It is worth noting that the performance of SPER (ViT-B-32) is comparable to the Adapter (ViT-B-16), which demonstrates the validity of SPER for efficient reconstruction of spatial features. The rest of the experimental results are more or less the same as RSICD and will not be repeated.

Table 2 Comparison of cross-modal retrieval performance on RSITMD

Type	Method	Training parameter/ 10^6	RSITMD						
			Text retrieval			Image retrieval			sumR
			R@1	R@5	R@10	R@1	R@5	R@10	
R	LW-MCR-u[19]	1.65	9.73	26.77	37.61	9.25	34.07	54.03	171.46
	AMFMN-sim[7]	35.94	10.63	24.78	41.81	11.51	34.69	54.87	178.29
	MCRN[20]	52.35	13.27	29.42	41.59	9.42	35.53	52.74	181.97
	SWAN[8]	—	13.35	32.15	46.90	11.24	40.40	60.60	204.64
	GaLR with MR[21]	46.89	14.82	31.64	42.48	11.15	36.68	51.68	188.45
T	Single Language[22]	151.00	19.69	40.26	54.42	17.61	49.73	66.59	248.30
	Linear probe[10]	0.53	13.71	33.41	48.01	10.97	36.85	56.15	199.10
	RS-light[23]	9.20	12.61	31.85	46.23	12.92	38.98	60.08	202.67
	TGKT[24]	4.70	17.92	36.95	52.88	12.83	43.14	62.48	226.20
	Cross-Modal Adapter[15]	0.16	18.16	36.08	48.72	16.31	44.33	64.75	228.35
	Full fine-tuning[10]	151.00	24.16	47.12	61.28	20.40	50.53	68.54	272.03
	Adapter[14]	2.57	21.01	41.59	53.76	**16.94**	46.19	64.02	243.51
	CLIP(ViT-B-16)[10]	0.00	8.84	23.45	36.28	9.86	34.38	49.38	162.19
	Adapter (ViT-B-16)	2.57	**23.67**	40.92	52.65	15.35	46.72	65.35	244.66
	Ours	0.18	21.46	43.36	**54.42**	16.81	45.88	62.96	244.89
	Ours (ViT-B-16)	0.60	23.45	**42.47**	52.87	15.48	**47.38**	**65.84**	**247.49**

2.4 Ablation study

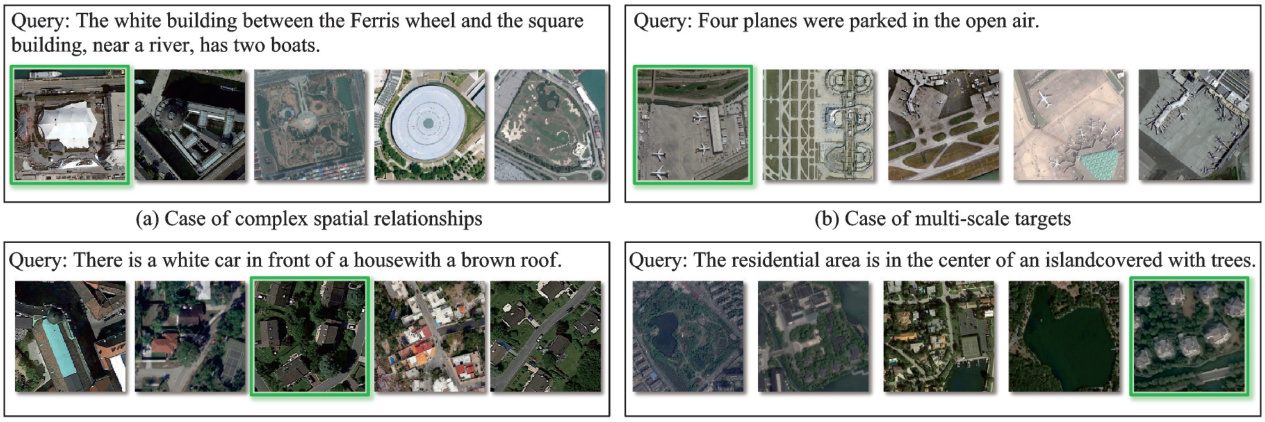
We explored the effect of the channel division step on the proposed SPER and the experimental results are shown in Table 3. When multimodal features are partitioned in the channel dimension, different division steps can be adopted, which is an important scientific hyperparameter. The division step size affects the quantity of spatial information as well as the volume of parameters required for reconstruction, which in turn affects the retrieval performance of the SPER. The best overall results are achieved when the division step takes 1. A longer division step length brings about an improvement in image retrieval metrics and has little effect on the overall performance. Therefore, we believe that SPER can efficiently perform feature reconstruction when the division step takes.

Table 3 Comparison of different division steps on RSITMD

Division step	Training parameter/10^6	RSITMD						
		Text retrieval			Image retrieval			sumR
		R@1	R@5	R@10	R@1	R@5	R@10	
1	0.18	**21.46**	**43.36**	**54.42**	16.81	45.88	62.96	**244.89**
3	0.43	20.79	42.03	53.54	**18.23**	45.93	63.32	243.84
5	0.67	20.57	42.69	53.53	17.96	**46.01**	**63.67**	244.43

2.5 Case study

Fig.3 shows some of the SPER retrieval results in different RS scenarios, and the retrieved images are arranged in order of similarity from left to right. The ground truth is indicated by the green box. Benefiting from the efficient reconstruction of spatial features, SPER is able to better extract valuable information and enhance the spatial relationships in RS images which is bolded in the query text. As shown in Fig.3(a), the performance of the proposed SPER remains reliable even in the presence of many entities and complex spatial relationships. Besides the ground truth, the retrieved images also contain the white building or boats in the river that are relevant to the query text. As shown in Fig.3(b), SPER could also align the image content and query semantics well when dealing with multi-scale targets. However, SPER is not accurate enough in retrieving RS images based on the number of entities described in the queries, as shown in Fig. 3(c) and Fig. 3(d). The SPER should be further optimized for the ability to extract the quantity of valuable targets.


Query: The white building between the Ferris wheel and the square building, near a river, has two boats.
Query: Four planes were parked in the open air.
Query: There is a white car in front of a housewith a brown roof.
Query: The residential area is in the center of an islandcovered with trees.

(a) Case of complex spatial relationships

(b) Case of multi-scale targets

(c) Case of similar localized information

(d) Case of similar global information

Fig.3 Retrieval cases of SPER on the RSITMD test set

2.6 Analysis of time consumption

Table 4 presents a comparison of the retrieval time consumption between different methods, where TT denotes the training time for one pass through the training set, ET the evaluation time for the test set, and IT the inference time for a single cross-modal retrieval. The computing platform consists of a 2.50 GHz Intel Xeon Gold 6133 CPU and a single NVIDIA 32 GB V100 GPU. The experimental dataset is RSITMD. The recorded results are the average of three runs.

Compared with the traditional GaLR method, SPER's TT increases by 26.22 s, and IT increases

Table 4 Retrieval time consumption of different methods

Type	Method	TT/s	ET/s	IT/ms
R	AMFMN	47.83	4.79	1.76
	GaLR	50.18	4.85	1.78
T	CLIP	101.04	5.66	2.08
	Adapter	79.39	5.86	2.16
	SPER	76.40	5.71	2.10

by 0.32 ms. However, considering the significant improvement in SPER's retrieval performance, we believe the additional time consumption is acceptable. Compared with the CLIP method, SPER reduces the TT by 24.3%, while the ET and IT are at the same level. Compared with the baseline method Adapter, SPER benefits from the efficient reconstruction of spatial features, leading to superior retrieval performance along with improved training and inference efficiency.

2.7 Limitations of SPER

One potential limitation of SPER is its performance in aligning fine-grained information, particularly regarding quantities. As demonstrated in subsection 2.5, SPER is not always accurate when retrieving RS images based on the number of entities described in the queries. This suggests that while SPER performs well in general retrieval tasks, there is room for improvement in its ability to model and retrieve precise numerical or quantity-based details.

Another limitation is SPER's efficiency when processing high-resolution remote sensing images (e.g., 10 000 pixel × 10 000 pixel). The high-resolution RS images were sliced to accommodate retrieval. As discussed in subsection 2.6, compared with traditional CNN-based approaches, i. e., approaches for instance AMFMN and GaLR, SPER could require more computational resources, potentially affecting inference speed. Thus, the trade-off between retrieval performance and efficiency is an important area for further exploration.

3 Conclusions

(1) We propose an efficient spatial feature reconstruction framework for RSCIR, which apparently reduces the consumption of computational and storage resources. Compared with the baseline method of fine-tuning all parameters in the VLP, our framework requires training only 0.18 million parameters (<1%) to achieve 96% of the baseline performance, reducing the training time by 24.3%.

(2) To bridge the gap between different domains, our designed spatial adapter efficiently models and extracts spatial relationships from multimodal features. In terms of retrieval performance, SPER leads similar methods by at least 2.7%.

(3) As discussed in the limitations section, our future research will focus on addressing the challenges of improving fine-grained retrieval and reducing computational demands during inference.

References

[1] ZHOU W, GUAN H, LI Z, et al. Remote sensing image retrieval in the past decade: Achievements, challenges, and future directions[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2023, 16: 1447-1473.

[2] YAN J, YU L, XIA C, et al. Super-resolution inversion and reconstruction of remote sensing image of unknown infrared band of interest[J]. Transactions of Nanjing University of Aeronautics & Astronautics, 2023, 40(4): 472-486.

[3] CAO M, LI S, LI J, et al. Image-text retrieval: A survey on recent research and development[EB/OL]. (2022-03-18). https://arxiv.org/abs/2203.14713.

[4] TANG X, WANG Y, MA J, et al. Interacting-enhancing feature transformer for cross-modal remote-sensing image and text retrieval[J]. IEEE Transactions on Geoscience and Remote Sensing, 2023, 61: 1-15.

[5] FAGHRI F, FLEET D J, KIROS J R, et al. VSE++: Improving visual-semantic embeddings with hard negatives[EB/OL]. (2018-07-18). https://arxiv.org/abs/1707.05612.

[6] ZHANG W, LI J, LI S, et al. Hypersphere-based remote sensing cross-modal text-image retrieval via curriculum learning[J]. IEEE Transactions on Geoscience and Remote Sensing, 2023, 61: 1-15.

[7] YUAN Z, ZHANG W, FU K, et al. Exploring a fine-grained multiscale method for cross-modal remote sensing image retrieval[J]. IEEE Transactions on Geoscience and Remote Sensing, 2022, 60: 4404119.

[8] PAN J, MA Q, BAI C. Reducing semantic confusion: Scene-aware aggregation network for remote sensing cross-modal retrieval[C]//Proceedings of the 2023 ACM International Conference on Multimedia Retrieval. [S.l.]: ACM, 2023: 398-406.

[9] WEN C, HU Y, LI X, et al. Vision-language models in remote sensing: Current progress and future trends[J]. IEEE Geoscience and Remote Sensing Magazine, 2024, 12(2): 32-66.

[10] RADFORD A, KIM J W, HALLACY C, et al. Learning transferable visual models from natural language supervision[C]//Proceedings of International Conference on Machine Learning. [S.l.]: PMLR, 2021: 8748-8763.

[11] LI J, SELVARAJU R, GOTMARE A, et al. Align before fuse: Vision and language representation learning with momentum distillation[J]. Advances in Neural Information Processing Systems, 2021, 34: 9694-9705.

[12] LI J, LI D, SAVARESE S, et al. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models[C]//Proceedings of International Conference on Machine Learning. [S.l.]: PMLR, 2023: 19730-19742.

[13] YUAN Y, ZHAN Y, XIONG Z. Parameter-efficient transfer learning for remote sensing image-text retrieval[J]. IEEE Transactions on Geoscience and Remote Sensing, 2023, 61: 1-14.

[14] HOULSBY N, GIURGIU A, JASTRZEBSKI S, et al. Parameter-efficient transfer learning for NLP[C]// Proceedings of International Conference on Machine Learning. [S.l.]: PMLR, 2019: 2790-2799.

[15] JIANG H, ZHANG J, HUANG R, et al. Cross-modal adapter for text-video retrieval[EB/OL]. (2022-11-17). https://arxiv.org/abs/2211.09623.

[16] HE K, FAN H, WU Y, et al. Momentum contrast for unsupervised visual representation learning[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, WA, USA: IEEE, 2020: 9729-9738.

[17] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16x16 words: Transformers for image recognition at scale[EB/OL]. (2021-06-03). https://arxiv.org/abs/2010.11929.

[18] DEVLIN J. BERT: Pre-training of deep bidirectional transformers for language understanding[EB/OL]. (2018-10-24). https://arxiv.org/abs/1810.04805.

[19] YUAN Z, ZHANG W, RONG X, et al. A lightweight multi-scale cross modal text-image retrieval method in remote sensing[J]. IEEE Transactions on Geoscience and Remote Sensing, 2021, 60: 1-19.

[20] YUAN Z, ZHANG W, TIAN C, et al. MCRN: A multi-source cross-modal retrieval network for remote sensing[J]. International Journal of Applied Earth Observation and Geoinformation, 2022, 115: 103071.

[21] YUAN Z, ZHANG W, TIAN C, et al. Remote sensing cross-modal text-image retrieval based on global and local information[J]. IEEE Transactions on Geoscience and Remote Sensing, 2022, 60: 1-16.

[22] AL RAHHAL M M, BAZI Y, ALSHARIF N A, et al. Multilanguage transformer for improved text to remote sensing image retrieval[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2022, 15: 9115-9126.

[23] LIAO Y, YANG R, XIE T, et al. A fast and accurate method for remote sensing image-text retrieval based on large model knowledge distillation[C]//Proceedings of IGARSS 2023—2023 IEEE International Geoscience and Remote Sensing Symposium. Pasadena, CA, USA: IEEE, 2023: 5077-5080.

[24] LIU A A, YANG B, LI W, et al. Text-guided knowledge transfer for remote sensing image-text retrieval[J]. IEEE Geoscience and Remote Sensing Letters, 2024, 21: 3504005.

Acknowledgement This work was supported by the National Key R&D Program of China (No.2022ZD0118402).

Authors

The first author Mr. **ZHANG Weihang** received the B.S. degree from Xidian University, Xi'an, China, in 2021. He is pursuing the Ph.D. degree with Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing, China. His research interests include remote sensing image retrieval and multi-modal remote sensing image interpretation.

The corresponding author Mr. **CHEN Jialiang** received the B.S. degree from Zhengzhou University, Zhengzhou, China in 2012 and the M.S. degree from Beijing Institute of Technology, Beijing, China, in 2016. He is currently an assistant professor in Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing, China. His research interest includes remote sensing image processing.

Author contributions Mr. **ZHANG Weihang** designed the study, developed the methodology, interpreted the results, and wrote the manuscript. Mr. **CHEN Jialiang** conducted validation and contributed to the writing, review, and editing of the manuscript. Prof. **ZHANG Wenkai** managed

project administration, organized resources, and supported the study's progress. Prof. **LI Xinming** conducted validation and contributed to the manuscript through review and editing. Prof. **GAO Xin** supervised the research process and reviewed the manuscript. Prof. **SUN Xian** provided funding support and critical insights into the methodology. All authors commented on the manuscript draft and approved the submission.

Competing interests The authors declare no competing interests.

(Production Editor: ZHANG Huangqun)

基于空间特征高效重构的遥感图文检索方法

张伟航[1,2,3]，陈佳良[1,2]，张文凯[1,2]，李新明[4]，高 鑫[1,3]，孙 显[1,2,3]

(1.中国科学院空天信息创新研究院，北京 100190，中国；
2.中国科学院目标认知与应用技术重点实验室，北京 100190，中国；
3.中国科学院大学电子电气与通信工程学院，北京 100190，中国；
4.空天信息大学计算机与人工智能学院，济南 250299，中国)

摘要：遥感跨模态图文检索(Remote sensing cross-modal image-text retrieval, RSCIR)旨在利用查询文本灵活、主观地检索遥感图像，近年来受到了越来越多研究者的关注。然而，随着预训练模型参数的不断增加，直接迁移学习的方法需要消耗大量的计算和存储资源。此外，最近提出的参数高效迁移学习方法主要聚焦于通道特征的重建，忽略了对关键实体关系建模至关重要的空间特征。为了解决这些问题，本文提出了一种基于空间特征高效重构(Spatial feature efficient reconstruction, SPER)的遥感跨模态图文检索方法，设计了一个简洁高效的空间适配器，以增强空间关系的提取能力。空间适配器只需通过少量参数即可对骨干网络中的特征进行空间重构，同时结合通道维度的先验信息。在两个常用的遥感图文检索数据集上进行的定量和定性实验表明，本文方法在sumR指标上相比传统方法提升了3%～11%。此外，与全参数训练方法相比，本文方法仅训练不到1%的参数，但整体性能仍保持在96%左右。本文相关代码和文件将开源在：https://github.com/AICyberTeam/SPER。

关键词：遥感跨模态图文检索；空间特征；通道特征；对比学习；参数高效迁移